

STATISTICS & PROBABILITY

for Senior High School

Revised Edition



Christian Paul O. Chan Shio
Maria Angeli T. Reyes

STATISTICS & PROBABILITY

for Senior High School

Revised Edition

STATISTICS & PROBABILITY

for Senior High School

Revised Edition

Christian Paul O. Chan Shio
Maria Angeli T. Reyes



C & E Publishing, Inc.
2021



C & E
Publishing, Inc.

*C & E Publishing, Inc. was
established in 1993 and is a
member of ABAP, PBAI, NBDB,
and PEPA.*

Statistics & Probability
for Senior High School
Revised Edition
Published by C & E Publishing, Inc.
839 EDSA, South Triangle, Quezon City
Tel. No.: (02) 8929-5088
E-mail: info@cebookshop.com

Copyright © 2021 by C & E Publishing, Inc.,
Christian Paul O. Chan Shio, and Maria Angeli T. Reyes

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted in any form or by any
means—electronic, mechanical, photocopying, recording, or
otherwise—without the prior written permission of the publisher.

Editing: Jose Dante Dela Merced
Jericho Gatbonton
Jaymie Guillermo

Cover Design: Darwin M. Tolentino

Layout: Maricar L. Sanchez

eISBN: 978-971-98-1579-2

Table of Contents

Preface	ix
----------------	----

Chapter 1 A Review of Probability

Lesson

1	Sample Space and Events	2
2	Fundamental Principles of Counting	11
3	Permutations and Combinations	17
4	Probability	25
5	Some Rules of Probability	31

Chapter Review	54
----------------	----

Chapter Performance Tasks	56
---------------------------	----

Chapter Exercises	58
-------------------	----

Chapter 2 Random Variables and Probability Distributions

Lesson

1	Concept of a Random Variable	62
2	Discrete Random Variable and Its Probability Mass Function	67
3	Continuous Random Variable and Its Probability Density Function	76
4	Mean and Variance of a Discrete Random Variable	83
5	Applications of Expected Value	93
6	Some Discrete Random Variables with Special Names	100

Chapter Review	122
----------------	-----

Chapter Performance Tasks	126
---------------------------	-----

Chapter Exercises	127
-------------------	-----

Chapter 3**The Normal Distribution****Lesson**

1	Introduction to the Normal Distribution	136
2	The Standard Normal Distribution.....	140
3	Areas under the Normal Curve.....	148
4	Applications of the Normal Distribution.....	153

Chapter Review.....	162
---------------------	-----

Chapter Performance Tasks.....	164
--------------------------------	-----

Chapter Exercises.....	166
------------------------	-----

Chapter 4**Sampling and Sampling Distributions****Lesson**

1	Random Sampling	170
2	The Sampling Distribution of the Sample Mean.....	181
3	The Central Limit Theorem.....	190

Chapter Review.....	200
---------------------	-----

Chapter Performance Tasks.....	201
--------------------------------	-----

Chapter Exercises.....	202
------------------------	-----

Chapter 5 Estimation of Parameters

Lesson

1	Point Estimators	206
2	Interval Estimation for a Mean.....	212
3	Interval Estimation for a Proportion.....	220
4	The t -distribution.....	225
5	Finding the Sample Size for Estimating Population Parameters.....	236

Chapter Review.....	245
---------------------	-----

Chapter Performance Tasks.....	248
--------------------------------	-----

Chapter Exercises.....	250
------------------------	-----

Chapter 6 Tests of Hypotheses

Lesson

1	The Hypothesis Testing Procedure	254
2	Tests Involving the Population Mean.....	263
3	p -values in Hypothesis Testing.....	275
4	Tests Involving the Population Proportion.....	280
5	Errors in Hypothesis Testing.....	287

Chapter Review.....	295
---------------------	-----

Chapter Performance Tasks.....	298
--------------------------------	-----

Chapter Exercises.....	300
------------------------	-----

Chapter 7**Linear Correlation and Simple Linear Regression****Lesson**

- 1** Linear Correlation304
- 2** Simple Linear Regression.....321
- 3** Model Adequacy and Inference on the Slope β_1332

Chapter Review.....339

Chapter Performance Tasks.....342

Chapter Exercises.....344

Appendices349

Glossary354

Index356

Bibliography358

About the Authors

Preface

Background

An article in Forbes magazine last 2016 reports that most of the top jobs for 2016 involve statistics.* Far from being just a statement about employment, this serves to emphasize the importance of statistics in the modern world. According to computer giant IBM, 2.5 exabytes (or 2.5 billion gigabytes) of data were generated every day in 2012. It is of practical importance for every Filipino to know how to collect, process, analyze, understand, and interpret all these data which are available to us. The main medium that allows us to do these is *statistics*.

However, it is a sad reality that many Filipinos nowadays are not statistically literate. This is especially evident in articles or comments that have come out recently in the media or on the Internet. Many comments in prominent web sites lash out at the credibility of surveys done by reputable firms simply because of the “small” sample size. Blogs and web sites conduct surveys or post statistical results which are either just plainly biased or lacking scientific merit. Even certain academics that come out in the media make statements based on inappropriate and misused statistical methods. These problems are symptoms of most Filipinos’ lack of knowledge and understanding of statistics.

Outline of the Material

Based on the years of experience of the authors in teaching both basic and mathematical statistics, this book attempts to provide the students with the necessary theoretical background along with the practical applications of statistics.

The book covers the relevant topics in the new DepEd Statistics and Probability curriculum for Senior High School, as well as some additional topics for enrichment. While the current K-12 curriculum includes some statistics in grades 1–8 as well as in grade 10, the authors are aware that the depth and preparation of students on these topics may vary. It is for this reason that we have included a preliminary chapter (chapter 1) which reviews the basic concepts of probability and counting techniques. Then the topics beginning with *random variables* and ending with *regression and correlation* follow in chapters 2 to 7.

We believe that students will be more motivated to learn the material when they are relevant to their own context. As such, we have included a large number of real-life

* Karsten Strauss, The Best Jobs in 2016. Forbes Magazine, April 14, 2016. Retrieved Oct. 24, 2016. www.forbes.com/sites/karstenstrauss/2016/04/14/the-best-jobs-in-2016/

applications which aim to show how the topics they learn are applicable to their daily lives. Furthermore, as most analyses involving statistics involve more than one concept, we believe that it is important for the students to see how the topics in each chapter are interrelated. As such, the authors have chosen to defer the exercises to the end of each chapter, allowing the students to see the “big picture” before doing any computational work.

The following are some of the main features of the book:

1. *Points to Remember*. Found mostly in chapters with more content and theory, these summarize the main points that the students are expected to know in the section.
2. *Software Tutorial in MS Excel*. As most statistical computations involve the use of software, these include instructions on how to do specific tasks or analyses in MS Excel.
3. *Chapter Review*. This serves to review the concepts and formulas introduced in the chapter.
4. *Chapter Performance Tasks*. These tasks allow the students to integrate everything that they have learned in the chapter in an applied problem. From data collection to data processing and analysis, these challenge the students to perform an actual statistical study.
5. *Chapter Exercises*. These include exercises of varying levels of difficulty, from the routine exercises to strengthen basic proficiencies, to the more challenging items to stimulate the problem solving skills of fast learners.

With all these features, it is our hope that the students will be able to experience and appreciate the beauty and insights from an understanding of statistics.

What’s new in this edition?

Aside from the revisions due to typos and errors in the previous edition, here are some of the changes in the following edition:

- Additional examples in Chapter 1 concerning tree diagrams and Venn diagrams.
- A discussion on the Negative Binomial distribution and an Excel tutorial on how it is solved has been added to Lesson 6 of Chapter 2.
- Additional discussion on the coefficient of determination, residuals, and assumptions of the simple linear regression model in Chapter 7.
- Additional exercises in Chapter 1, 2, 3, and 7.
- Additional figures to illustrate areas and to obtain critical values in Chapters 4 to 6.

The Authors

Chapter 1

A Review of Probability



The roots of probability can be traced back as far as the 17th century and was largely associated with gambling. A professional gambler, Chevalier de Méré (1607–1684), in his correspondence with Blaise Pascal (1623–1662) and Pierre de Fermat (1601–1665), wanted to model some gambling odds. It is no wonder that probability is prominent in games of chance, and experiments such as dice-rolling, card-picking, coin-tossing, and roulette-spinning are commonly used to illustrate the concept. In the present day, probability is widely used in various fields such as business, politics, psychology, medicine, and education.

The purpose of this chapter is to help you review the concepts of random experiments and events. You will reinforce your knowledge of the basic notions and laws of probability. A lesson on counting techniques is included, as some examples of combinatorial probability are discussed.

Lesson 1

Sample Space and Events

Learning Outcomes

- At the end of this lesson, you should be able to
 - define and give examples of statistical experiments;
 - define the sample space of an experiment;
 - compute the union, intersection, and complement of two events; and
 - illustrate operations on two events using a Venn diagram.

Introduction

In probability and statistics, we are interested in outcomes from statistical experiments. When we hear the word *experiment*, we usually associate it with a laboratory and with scientific research. While some statistical experiments may also be done inside a laboratory, they actually deal with something more general.

Definition 1

An **experiment** is any procedure that

- can be repeated, theoretically, an infinite number of times; and
- has a well-defined set of possible outcomes.

Consider rolling a die as a statistical experiment. Theoretically, one can roll a die infinitely many times, and in each roll, we know the possible results: the number of spots will be from 1 to 6.

Some experiments can be even more practical. For instance, selecting a random student from a class in the school and measuring his or her blood pressure can be considered as an experiment. In this case, the possible outcomes consist of pairs of numbers corresponding to the student's blood pressure.

As we can see from the given examples, an important detail of each experiment is its set of possible outcomes.

Definition 2

Each possible result of an experiment is referred to as a **sample outcome**. The set of all possible outcomes is called the **sample space**, and is usually denoted by S .

The nature of the sample space of an experiment can be quite varied. The sample space can be finite or countably or uncountably infinite. An *infinite* but *countable* set has elements which can be put into one-to-one correspondence with the natural numbers whereas an *uncountable* set has a larger number of elements than the natural numbers. The following examples illustrate these varied sample spaces.

Example 1

Identify the sample space S in the following experiments:

1. Roll a die and observe the number that comes up.
2. Toss a coin repeatedly until the first head appears.
3. Turn on a lightbulb and measure its life span.
4. Flip two coins and observe the sequence of heads and tails.
5. Choose real numbers a , b , and c such that the quadratic equation $ax^2 + bx + c = 0$ has imaginary roots.

Solution:

1. In this case, the sample space consists of a *finite* number of outcomes. In particular, $S = \{1, 2, 3, 4, 5, 6\}$.
2. In each toss of a coin, either a head (H) or a tail (T) will appear. If the first toss is a head, then the experiment is over, and we have the sample outcome H . Otherwise, if the first toss is a tail, then the second toss may either be a head or a tail. In the first case where the second toss is a head, we have the outcome TH . If the case where the second toss results in a tail, then the sequence of tosses continues, and we can have the outcomes TTH , $TTTH$, and so on. Here, the sample space is *infinite* but *countable*: $S = \{H, TH, TTH, TTTH, \dots\}$.
3. Some lightbulbs do not even turn on, and so its life span is 0 hours. While lightbulbs eventually die out, any bulb that lights up can theoretically last any positive real number of hours. In this case, the sample space is *infinite* but *uncountable*, consisting of all nonnegative real numbers: $S = [0, \infty)$.

4. Each outcome is an ordered pair of results from each coin. Letting H and T denote a head and a tail respectively, it is easy to see by listing that the sample space has exactly four outcomes: $S = \{(H, H), (H, T), (T, H), (T, T)\}$.
5. In this case, the outcomes consist of ordered triples of real numbers (a, b, c) . It is impossible to list down all the outcomes in the sample space, although it is possible to describe which outcomes are in S . Recall from algebra that a quadratic equation $ax^2 + bx + c = 0$ only has imaginary solutions if its *discriminant* $b^2 - 4ac$ is negative. Thus, the sample space for this experiment is $S = \{(a, b, c) \mid b^2 - 4ac < 0\}$.

A sample space that has a finite or countably infinite number of outcomes is said to be *discrete*. On the other hand, a sample space with an uncountably infinite number of outcomes is said to be *continuous*. In example 1, the experiments in numbers 1, 2, and 4 have discrete sample spaces, while those in numbers 3 and 5 are both continuous.

For most experiments, we are not interested in the entire sample space, but rather only a subset of it. That is, we wish to study only an *event*.

Definition 3

A subset of the sample space S of an experiment is called an **event**.

An event is usually denoted by a capital letter, followed by either a description or a list of all the outcomes which are included.

Example 2

Suppose we roll a die and observe the number that comes up. Two possible events can be defined as follows:

A : The outcome is odd.

B : The outcome is at least 4.

These can be viewed as subsets of the sample space S , with $A = \{1, 3, 5\}$ and $B = \{4, 5, 6\}$.

Example 3

Performing the same experiment in example 2, consider the following events:

C : The outcome is at least 7.

D : The outcome is either odd or even.

Then, $C = \emptyset$ and $D = \{1, 2, 3, 4, 5, 6\}$. This means that an event may also contain no elements.

As events are simply subsets of the sample space S , we can perform set operations on one or more events. Some of the more common operations include taking the intersection and the union of two events, as well as taking the complement of an event.

Definition 4

Let A and B be two events defined over the same sample space S . Then

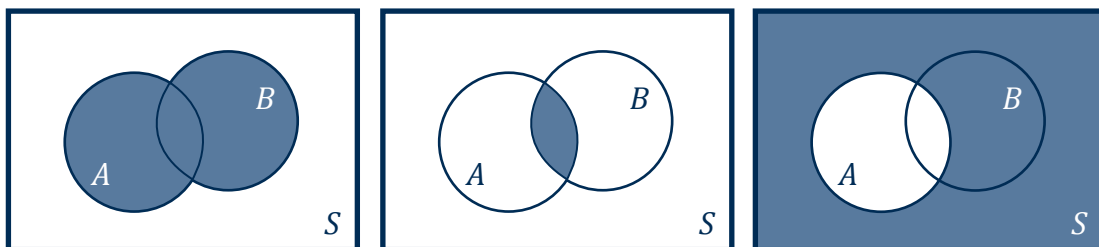
- the **intersection** of A and B , denoted by $A \cap B$, is the event whose outcomes belong to both A and B .
- the **union** of A and B , denoted by $A \cup B$, is the set of all outcomes in A or B (or both).
- if $A \cap B = \emptyset$, then the events A and B are said to be **mutually exclusive**.

We can also take the complement of an event A .

Definition 5

Let A be any event defined on a same sample space S . The **complement** of A , denoted by A' or A^c , is the set of outcomes in S which are not in A .

Like sets, we can also represent operations on events using Venn diagrams.



S represents the entire sample space. The first two diagrams represent the union and intersection of events A and B , respectively. The rightmost figure shows the complement of A .

Example 4

Consider again the experiment of tossing two fair coins. Let events A and B be defined as follows:

A : Two heads come up.

B : Two tails come up.

Then A consists of the single outcome (H, H) while B only contains the outcome (T, T) .

- The union of A and B , $A \cup B$, is the event where either two heads or two tails turn up. In set notation, we can write $A \cup B = \{(H, H), (T, T)\}$.
- The intersection of A and B , $A \cap B$, is the event where both two heads and two tails come up in the same trial of the experiment. Since we cannot have two heads and two tails at the same time, $A \cap B = \emptyset$. We can also deduce this by seeing that the events A and B have no elements in common.
- Since A and B have no elements in their intersection, by definition, they are mutually exclusive.

Example 5

Using the same experiment and definition of A and B as in example 4, A' refers to the event for which the result is not two heads. It therefore consists of all elements in S except for (H, H) . Therefore, we have

$$A' = \{(H, T), (T, H), (T, T)\}.$$

Although A and B are mutually exclusive, they are not complements. That is, if the result in the experiment is not two heads, it does not mean that it must be two tails—it could be that we have one tail and one head instead. In particular, B does not contain all the elements in S which are not in A .

Sometimes, we might encounter problems which involve the complement of a union or an intersection of two events. For these cases, the following *De Morgan's Laws* may be useful:

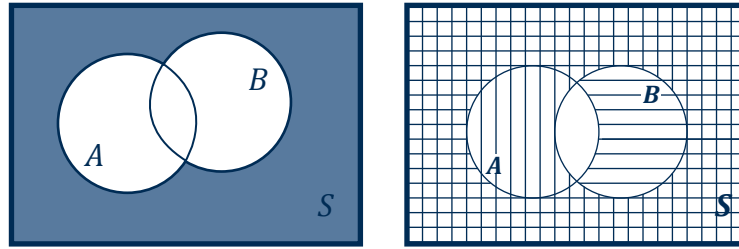
Definition 6

De Morgan's Laws: For any two events A and B ,

- $(A \cup B)' = A' \cap B'$.
- $(A \cap B)' = A' \cup B'$.

Intuitively, this means that to “distribute” the complement within the parentheses, one needs to *reverse* the operation between A' and B' , switching a union to an intersection, and vice versa.

One way to see why these rules make sense is to illustrate them using a Venn diagram. For example, for the first rule, $(A \cup B)'$ refers to the set of outcomes which are outside both circles as seen in the first diagram below. On the other hand, A' and B' refer to the outcomes outside circles A and B , respectively. These are illustrated in the second diagram below by the regions shaded with horizontal lines (for A') and vertical lines (for B'), respectively. The portion which contains lines of both directions is outside both A and B .



Venn diagrams illustrating the first De Morgan's Law: $(A \cup B)' = A' \cap B'$

Example 6

Using the same definition of A and B as in example 4, recall that $A \cap B = \emptyset$, so $(A \cap B)'$ is the entire sample space S . By (the second) De Morgan's Law, this is the same as $A' \cup B'$. To verify this, note that A' is the event of not getting two heads, while B' is the event of not getting two tails. Therefore,

$$A' = \{(H, T), (T, H), (T, T)\}, \text{ and}$$

$$B' = \{(H, H), (H, T), (T, H)\}.$$

Thus,

$$A' \cup B' = \{(H, H), (H, T), (T, H), (T, T)\},$$

which also corresponds to the entire sample space.

Example 7

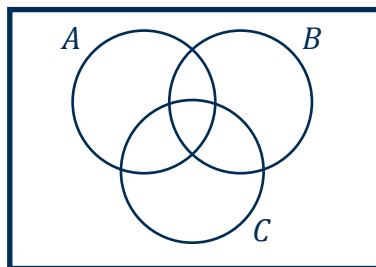
From a Junior High School, 250 students were surveyed on their preferred Senior High School strands. The following results were obtained:

150 chose STEM (event A)	68 chose both ABM and HumSS
109 chose ABM (event B)	95 chose HumSS and STEM
127 chose HumSS (event C)	48 chose the three Strands
84 chose both STEM and ABM	

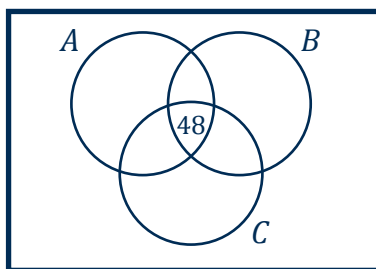
1. Construct a Venn diagram to represent the three events.
2. How many students chose STEM only? ABM only? HumSS only?
3. Represent using set operations those who chose all three strands.
4. How many elements are there in the event $(A \cup B \cup C)$?

Solution:

1. Represent the entire sample space S with a rectangle and draw circles for sets A , B , and C as shown below:

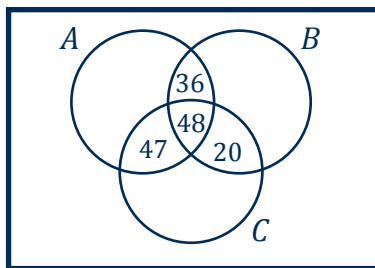


Fill in the common intersection of the 3 circles with 48, as this is the number of students who chose the three strands.



Consider now the number of students who chose exactly two strands. After removing the 48 students who chose all three strands, we see that:

- $84 - 48 = 36$ chose both STEM and ABM only
- $68 - 48 = 20$ chose both ABM and HumSS only
- $95 - 48 = 47$ chose HumSS and STEM only



Next is to fill up the number of students who chose only one strand by subtracting the numbers that chose the two and three strands from the following:

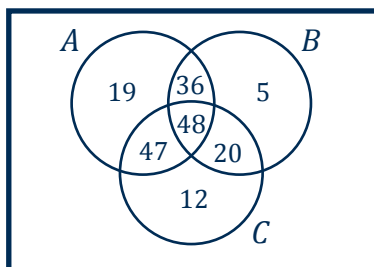
150 chose STEM (event A)

109 chose ABM (event B)

127 chose HumSS (event C)

This means that:

- $150 - (48 + 36 + 47) = 19$ chose STEM only
- $109 - (48 + 36 + 20) = 5$ chose ABM only
- $150 - (48 + 47 + 20) = 12$ chose HumSS only



2. As indicated in the solution for the Venn diagram, 19 chose STEM only, 5 chose ABM only, and 12 chose HumSS only.
3. The students who chose all three strands represent the intersection of all three events, which is $A \cap B \cap C$.
4. $(A \cup B \cup C)' = 250 - (19 + 47 + 48 + 36 + 5 + 20 + 12) = 63$

This means that 63 students are undecided on their preferred Senior High School strand.

Let's Practice

- In each of the following experiments, compute the number of elements in the sample space and list the elements of the given event A .
 - A card is drawn from a standard deck of playing cards. Let A be the event where the card drawn is a red face card.
 - Four coins are tossed. Let A be the event where the result has fewer heads than tails.
 - Two fair dice are rolled. Let A be the event where the sum of the dice is 8.
 - Two chips are drawn (with replacement) from an urn with chips numbered 1, 2, 3, and 4. Let A be the event where the two chips are of the same number.
 - A chip is drawn (with replacement) from an urn with chips numbered 1, 2, 3, and 4 until a 1 is drawn. Let A be the event where three or fewer chips are drawn.
- A deck consists of three black cards, numbered 1, 2, and 3, and two red cards numbered 1 and 2. Chris draws a card from the deck. Then Graham draws a card from the remaining cards. Let A be the event where Chris's card has a higher number than that of Graham's card, and let B be the event where Graham's card has a higher number than that of Chris's card.
 - List down the outcomes in A and B .
 - Are A and B mutually exclusive? Are they complements? Justify your answers.
- Let A and B be any two events defined on a sample space S . Which of the following sets are necessarily subsets of which other sets?

A	B	$A \cup B$	$A \cap B$
$A' \cap B$	$A \cap B'$	$(A' \cap B')'$	$(A' \cup B')'$

- A public health researcher examines the medical records of a group of 937 men who died in 1999 and discovers that 210 of the men died of causes related to heart diseases. Moreover, 312 of the 937 men had at least one parent who suffered from heart disease, and of these 312 men, 102 died of causes related to heart disease. Suppose A is the event where a selected man from this group died of causes related to heart diseases, and B is the event where at least one parent suffered from heart disease.
 - Construct a Venn diagram to represent the given scenario.
 - How many elements are there in A ? in B ? in $A \cap B$?
- From a small town, 120 persons were selected at random and asked the following question: "Which of the three brands of shampoo—A, B, and C—do you use? The following results were obtained: 20 use A and C, 10 use A and B but not C, 15 use all three, 30 use only C, 35 use B but not C, 25 use B and C, and 10 use none of the three.
 - Construct a Venn diagram to represent the given events.
 - How many people use only brand A? How many use neither brand A nor brand B?

Lesson 2

Fundamental Principles of Counting

Learning Outcomes

- At the end of this lesson, you should be able to
 - differentiate between the rule of sum and the rule of product; and
 - use the rules of sum and product to solve counting problems.

Introduction

An important part of probability is counting the number of outcomes of events or the sample space that results from an experiment. For sample spaces or events with a small number of outcomes, this can be done by listing down or directly counting all the elements in the sample space. But for larger sample spaces or events, we will need to apply some *counting techniques*.

In this lesson, we look at two basic principles of counting: the rule of sum and the rule of product. As these properties are central in most counting problems, they are sometimes called the *fundamental principles of counting*.

The rule of sum is also sometimes called the *addition principle*.

Rule of Sum (Addition Principle)

If a particular action can be done in m ways and another in n ways, and the two actions cannot be done at the same time, then there are $m + n$ ways of doing exactly one of these actions.

Example 1

Jobelle wants to travel from Cebu to Bacolod. She finds out that there are two possible flights and five possible ferries for today. Assume that she wants to travel today and that she has no restrictions on which mode of transport and schedule to take. As it is not possible to take both a flight and a ferry at the same time, by the rule of sum, there are $2 + 5 = 7$ ways for Jobelle to get to Bacolod from Cebu.

We can extend the rule of sum to more than two actions. Suppose there are a total of k actions, and there are n_1 ways to do the first action, n_2 ways to do the second, and so on, with n_k ways to do the k th action. If no two of these actions can be done at the same time, then there are $n_1 + n_2 + \cdots + n_k$ ways to do exactly one of the k actions.

Example 2

Mark is choosing which template to use for his presentation. Suppose that he has three folders of templates, containing 12, 23, and 30 different designs. Assuming he can only use one template design for his presentation, by the rule of sum, he would have $12 + 23 + 30$ or 65 possible choices.

Some counting problems can be more complicated than the ones which we have already discussed. Suppose, for instance, that a milk tea shop allows you to make your own milk tea combination. You can select one tea base and one order of sinkers. The choices are given as follows:

Tea Base	Sinkers
Black tea	Pearls
Green tea	Nata de coco
Earl grey tea	Pudding
Red tea	Rainbow jelly
	Aloe

To do this, there are two actions that you must do: (1) choose a tea base and (2) choose a sinker. If you choose black tea as the tea base, for example, notice that you will have *five* possible drinks, depending on which of the five sinkers you select. This is the same for each of the *four* choices of tea base. Therefore, the total number of possible milk tea combinations is $4 \times 5 = 20$.

Another way of illustrating such number of combinations is through a *tree diagram*.

Tea Base	Sinkers	Combinations
Black tea	Pearls	Black tea and pearls
	Nata de coco	Black tea and nata de coco
	Pudding	Black tea and pudding
	Rainbow jelly	Black tea and rainbow jelly
	Aloe	Black tea and aloe

Green tea	Pearls	Green tea and pearls
	Nata de coco	Green tea and nata de coco
	Pudding	Green tea and pudding
	Rainbow jelly	Green tea and rainbow jelly
	Aloe	Green tea and aloe
Earl grey tea	Pearls	Earl grey tea and pearls
	Nata de coco	Earl grey tea and nata de coco
	Pudding	Earl grey tea and pudding
	Rainbow jelly	Earl grey tea and rainbow jelly
	Aloe	Earl grey tea and aloe
Red tea	Pearls	Red tea and pearls
	Nata de coco	Red tea and nata de coco
	Pudding	Red tea and pudding
	Rainbow jelly	Red tea and rainbow jelly
	Aloe	Red tea and aloe

Generalizing this to any two actions, we obtain the second fundamental principle of counting known as the *rule of product*, which is also sometimes called the *multiplication principle*.

Rule of Product (Multiplication Principle)

If a particular action can be done in m ways and another in n ways, then there are $m \times n$ ways of doing both actions (one after the other).

This can be further generalized to the case of three or more actions. For example, in the case of the milk tea problem, if one would also consider the size of the drink and there are two sizes, then there would be a total of $4 \times 5 \times 2 = 40$ possible milk tea combinations. Mathematically, we can state the rule as follows:

Rule of Product (General Case)

Given k actions, if the first action can be done in n_1 ways, the second action in n_2 ways, and so on, then there are $n_1 \times n_2 \times \cdots \times n_k$ ways of doing all the k actions.

Example 3

Currently, the format for license plates for vehicles in our country is three letters followed by four numbers.

1. How many different license plates can be made?
2. How many different license plates are there if letters can be repeated but no two numbers can be the same?
3. How many different license plates can be made if repetition of numbers and letters are allowed except that no plate can have four zeros?

Solution:

1. There are 26 choices for each letter and 10 choices (from 0 to 9) for each of the numbers. Thus, there are

$$\underbrace{26 \times 26 \times 26}_{\text{for the three letters}} \times \underbrace{10 \times 10 \times 10 \times 10}_{\text{for the four numbers}} = 175,760,000$$

distinct license plates that can be made.

2. We still have 26 choices for each of the three letters. Since numbers cannot be repeated, there are 10 choices for the first number, 9 choices for the second, 8 choices for the third, and 7 choices for the fourth. This gives us

$$\underbrace{26 \times 26 \times 26}_{\text{for the three letters}} \times \underbrace{10 \times 9 \times 8 \times 7}_{\text{for the four numbers}} = 88,583,040$$

possible license plates.

3. Here, it is easier to just subtract the license plates with four zeros from the grand total of plates as computed in item (1). To compute the number of license plates with four zeros, note that there are still 26 choices for each letter, but only 1 choice for each number (it must be equal to 0). Hence, we have

$$\underbrace{26 \times 26 \times 26}_{\text{choices for the 3 letters}} \times \underbrace{1 \times 1 \times 1 \times 1}_{\text{each of these must be 0}} = 17,576$$

plates with four zeros. Subtracting this from the total number of license plates without restrictions gives us our final answer of

$$175,760,000 - 17,576 = 175,742,424$$

possible license plates.

Example 4

A restaurant offers a choice of three appetizers, twelve entrées, five desserts, and four beverages. How many different meals are possible if a customer intends to order only three courses? (Consider the beverage to be a “course.”)

Solution:

This problem requires the use of both the rule of sum and the rule of product.

There are four possible cases, depending on which three courses the customer orders. We can use the rule of product to count the number of meals in each case:

Courses	Number of possible meals
Appetizer, entrée, dessert	$3 \times 12 \times 5 = 180$
Appetizer, entrée, beverage	$3 \times 12 \times 4 = 144$
Appetizer, dessert, beverage	$3 \times 5 \times 4 = 60$
Entrée, dessert, beverage	$12 \times 5 \times 4 = 240$

By the rule of sum, there are $180 + 144 + 60 + 240 = 624$ possible three-course meals.

Points to Remember

1. The rule of sum is used for counting problems which involve several possibilities or actions, only one of which must occur at any given time.
2. The rule of product is used for tasks which involve several actions, all of which must occur one after the other.

Let's Practice

1. Five trails lead to the top of a mountain. In how many ways can a mountaineer climb up and down the mountain if the mountaineer would prefer not to retrace his steps?
2. There are five action, three drama, and four comedy films showing at a local mall. If Kristel goes to this mall to watch a single movie, how many possible choices does she have?
3. Leonard would like to change his profile picture in his social media account. He looks at his currently uploaded photos, and finds 6 photos in one folder and 12 more photos in a second folder as possible profile pictures. If he wishes to change his profile picture every day, how many days can he go without repeating a picture?
4. An experiment consists of rolling a die and then selecting a letter at random from the English alphabet. How many outcomes are there in the sample space for this experiment?
5. A school has 71 freshmen, 108 sophomores, 93 juniors, and 89 seniors. In how many ways can one student be chosen to represent the school in a youth convention?
6. In a fuel economy study, each of three car models is tested using 5 different brands of gasoline at 8 test sites located in different locations in the country. If there are 3 test drivers and it is desired to have one test under each distinct set of conditions, how many test runs are needed?
7. How many three-digit numbers can be formed from the digits 0, 1, 2, 3, 4, and 5 if
 - a. repetition is allowed?
 - b. repetition is not allowed?
 - c. repetition is not allowed, and the number must be odd?
8. An exam consists of 10 multiple-choice questions, each of which has 4 choices.
 - a. In how many ways can a student answer the test?
 - b. In how many ways can a student answer the test and get all the answers wrong?
9. A *palindromic number* is a positive number that remains the same when its digits are reversed. For example, 16,361, 2,552, and 777 are palindromes, but 1,778 is not. Assume that single-digit numbers are also considered palindromic.
 - a. How many palindromic numbers less than 100,000 are there?
 - b. How many six-digit palindromic numbers greater than 250,000 are there?
10. A tennis tournament begins with a field of 32 players. After five rounds of play, the player who remains unbeaten is declared the champion. How many different configurations of winners and losers are possible, starting with the first round? Assume that the initial pairing of the 32 players into 16 first-round matches has already been done.

Lesson 3

Permutations and Combinations

Learning Outcomes

- At the end of this lesson, you should be able to
 - distinguish between a permutation and a combination; and
 - solve counting problems involving permutations and combinations.

Introduction

In many cases, we are interested in solving problems involving all possible orders or arrangements of objects. For example, we may wish to know the number of ways to arrange distinct objects in a row, or we may want to compute the number of ways to select a committee head and assistant head from a committee of 10 people. In this case, we obtain a permutation.

Definition 1

An ordered selection of r objects from a set A containing n objects ($0 \leq r \leq n$) is called a **permutation** of A taken r at a time, and is denoted by ${}_nP_r$.

Example 1

Suppose you have 10 books that you wish to place on a shelf. In how many ways can you arrange the books?

Solution:

This is a permutation of all 10 elements of a set A with 10 objects, or ${}_{10}P_{10}$. To find its numerical equivalent, we fill up each of the ten slots in the shelf one by one.

There are 10 choices for which book to place on the first slot. Since one of the books is already on the first slot, there are now just 9 choices left for the second slot. Continuing in this manner, there are 8 possible books remaining to fill the third slot, and so on, with the last slot having just 1 possible book. By the rule of product, there are

$${}_{10}P_{10} = 10 \times 9 \times 8 \times \dots \times 1 = 3,628,800$$

ways to arrange the 10 books in the shelf.

Definition 2

For any integer $n \geq 0$, **n factorial**, denoted by $n!$ is defined by

$$n! = n \times (n - 1) \times \dots \times 3 \times 2 \times 1, n \geq 1.$$

$$0! = 1$$

Using this definition in the answer to the previous example, we can say that there are $10! = 3,628,800$ ways to arrange the books.

Example 2

Suppose you draw out 3 books from the shelf with 10 books in example 1 and place them in another shelf. In how many ways can you now arrange the books?

Solution:

This problem involves a permutation of 3 objects taken from a set containing 10 objects. We can use the same argument as in the previous example, except that we only need to fill up three slots in the new shelf. There are 10 possible books to fill the first slot, 9 choices for the second slot, and 8 choices left for the third slot. Therefore, there are

$${}_{10}P_3 = 10 \times 9 \times 8 = 720$$

possible ways to arrange the books.

We now try to find a general formula for ${}_nP_r$. Consider the problem of selecting, and then placing distinct objects into slots, where the order is important. Using the same reasoning as the previous two examples, we have n choices for the first slot. Since one object has already been chosen, there are now just $n - 1$ choices left for the second slot. Continuing in this manner, there are $n - 2$ choices for the third slot, and so on, with $(n - r + 1)$ for the r th slot. Therefore, we have:

$$\begin{aligned} {}_nP_r &= n \times (n - 1) \times (n - 2) \times \dots \times (n - r + 1) \\ &= \frac{n \times (n - 1) \times (n - 2) \times \dots \times (n - r + 1) \times (n - r)!}{(n - r)!} \\ &= \frac{n!}{(n - r)!} \end{aligned}$$

Theorem 1

The number of permutations of r elements that can be formed from a set of n distinct elements is

$${}_nP_r = \frac{n!}{(n - r)!}.$$

In cases where we wish to arrange or permute an entire set of n objects, this corresponds to $n = r$ in the previous theorem. Since $(n - n)! = 0! = 1$, we obtain the following corollary.

Corollary 1

The number of ways to permute an entire set of n distinct objects is

$${}_nP_n = n!.$$

Example 3

Five boys and four girls are to be arranged in a row. In how many ways can this be done if

1. there are no restrictions?
2. the boys and girls must alternate?
3. two particular boys, Rudy and Ali, insist on sitting next to each other?

Solution:

1. Since there are five boys and four girls and there are no restrictions, we are to arrange all nine people in a row. This can be done in ${}_9P_9 = 9! = 362,880$ ways.
2. There are two parts to this problem: (1) arranging the boys and (2) arranging the girls. The boys can be arranged in ${}_5P_5 = 5! = 120$ ways, while the girls can be arranged in ${}_4P_4 = 4! = 24$ ways. By the rule of product, the number of arrangements of all nine people is $(120)(24) = 2,880$ ways.
3. Since Rudy and Ali must be seated next to each other, we consider them as one block. Then the seven other people and this block can be arranged in ${}_8P_8 = 8! = 40,320$ ways. However, within the block, Rudy and Ali can be arranged in ${}_2P_2 = 2! = 2$ ways. By the rule of product, the required number of arrangements is $40,320 \times 2 = 80,640$.

Consider now the problem of arranging n distinct objects in a row where order is not important. In this case, we are looking at a *combination* of the objects.

Definition 3

A selection of r objects from a set A containing n objects ($0 \leq r \leq n$) without regard to order is called a **combination** of the elements of A taken r at a time, and is denoted by $\binom{n}{r}$ or ${}_nC_r$.

To count the number of combinations of a set with n elements taken r at a time, we will need to remove the ordering of the objects by dividing by the number of arrangements of the slots.

For example, suppose that in example 2, we were only interested in choosing three books without regard to the order of selection. If the selected books are A, B, and C, then all the $3! = 6$ permutations ABC, ACB, BAC, BCA, CAB, CBA all count as one combination only. That is, to count the number of combinations of three books from the ten books, we can just count the number of permutations of three objects from the ten objects, then divide it by the number of ways to arrange any choice of three books, which is $3!$. That is,

$$\binom{10}{3} = \frac{{}_{10}P_3}{3!} = \frac{\frac{10!}{(10-3)!}}{3!} = \frac{10!}{3!(10-3)!}.$$

Generalizing this to the case of a combination of the elements of a set with n objects taken r at a time, we have the following result:

Theorem 2

The number of combinations of length r that can be formed from a set of n distinct elements is

$${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Note that Theorem 2 is equivalent to

$$\binom{n}{r} = {}_nC_r = \frac{{}_nP_r}{r!},$$

where ${}_nP_r$ is the number of permutations of r objects taken from a set of n objects.

Example 4

A 10-person committee is to be selected out of 15 equally capable candidates.

1. In how many ways can this be done if there are no restrictions?
2. Suppose we know that 9 of the candidates are females. If we wish our committee to be equally represented by both genders, how many committees are possible?

Solution:

1. Since order is irrelevant when selecting members of a committee, direct use of combinations gives us $\binom{15}{10} = 3,003$ possible committees.
2. Since the committee must have the same number of males and females, we must select five persons among the 9 females and five persons from among the $15 - 9 = 6$ males. We can choose the 5 females in $\binom{9}{5} = 126$ ways, while we can choose the 5 males in $\binom{6}{5} = 6$ ways. By the rule of product, there are $126 \times 6 = 756$ ways to form the committee.

Example 5

A poker hand consists of 5 cards from a standard deck of 52 cards. Find the number of different poker hands which have two pairs. (Note: A poker hand of *two pairs* consists of two different pairs of cards where the cards in each pair are of the same rank and a fifth card of unmatched rank. An example is given on the right.)



Solution:

We first select which of the 13 ranks comprise the two pairs. Note that the order of the two ranks does not matter. For example, a hand of two 8's and two jacks is the same as a hand of two jacks and two 8's. Therefore, there are $\binom{13}{2}$ possible ways to choose the ranks.

For each of the two ranks, we need to choose which two of the four suits are included in the hand. This gives $\binom{4}{2}$ ways for the first rank, and another $\binom{4}{2}$ for the second rank.

Finally, we complete the poker hand by selecting the fifth card in the hand. This should come from any of the 44 cards which are not of the two chosen ranks for the pairs. This gives $\binom{44}{1}$ ways to choose the fifth card.

Since all of the operations above must be performed to produce the poker hand, by the rule of product, the number of possible poker hands is

$$\binom{13}{2} \binom{4}{2} \binom{4}{2} \binom{44}{1} = 123,522.$$

Some combinatorial problems can be solved in more than one way. The main difference in these methods involve how the outcomes are characterized, as we shall see in the next example.

Example 6

A pizza parlor offers customers an option to make their own pizza. From a base of tomato sauce and mozzarella cheese, customers may choose to add one or more (or none) of the following eight toppings: beef, onions, bell pepper, anchovies, pepperoni, olives, tuna, and tomatoes. How many possible pizzas can be made?

Solution:

We solve this problem in two different ways.

Method 1: We consider cases based on the number of toppings that will be added.

In general, for each $i = 0, 1, 2, \dots, 8$, there are $\binom{8}{i}$ ways to select i toppings to include in the pizza. Since at any given time, we can only have one of these cases as the number of toppings, by the rule of sum, there are

$$\sum_{i=0}^8 \binom{8}{i} = \binom{8}{0} + \binom{8}{1} + \dots + \binom{8}{8} = 256$$

possible pizzas that can be made.

Method 2: For each topping, we may choose to include it or exclude it in the pizza. This gives 2 possible choices for each of the 8 toppings. By the rule of product, we have $2^8 = 256$ possible pizzas, which is, of course, the same as in method 1.

Remark: The previous problem shows that

$$\sum_{i=0}^8 \binom{8}{i} = \binom{8}{0} + \binom{8}{1} + \dots + \binom{8}{8} = 2^8.$$

In general, we can show that for any positive integer n ,

$$\sum_{i=0}^n \binom{n}{i} = \binom{n}{0} + \binom{n}{1} + \dots + \binom{n}{n} = 2^n.$$

This implies that the total number of subsets of a finite set S having n elements is 2^n .

Let's Practice

1. Compute the exact values of the following without the use of calculators:
 - a. ${}_6P_4$
 - b. ${}_5P_5$
 - c. ${}_9C_2$
 - d. ${}_7C_3$

2. A freshman has 4 examinations to take and there are 10 examination periods available. How many possible arrangements of his examination schedule are there?
3. How many ways are there to select the starting 5 basketball players from a pool of 15 men who can play all of the positions?
4. A classroom consists of three rows of five chairs each. How many possible ways can a teacher assign seats to her eight students?
5. Assuming the animals are all distinguishable, in how many ways can 4 zebras, 5 lions, and 2 hippos be arranged in a row if
 - a. there are no restrictions?
 - b. animals of the same kind must be together?
 - c. the hippos are not together?
6. In how many ways can a research team of 5 members be formed from a group consisting of 3 chemists and 7 physicists if
 - a. there is no restriction in the selection?
 - b. the team must include exactly 2 chemists?
 - c. the team must include at most 3 physicists?
7. A math class consists of 15 boys and 10 girls. In how many ways can a committee of 8 members be formed if
 - a. the number of boys is more than twice the number of girls?
 - b. Mary and Anne cannot be in the committee simultaneously?
8. In how many ways can five couples be seated in a row if
 - a. there are no restrictions?
 - b. each couple must be seated together?
 - c. each husband must be seated to the left of his wife?
9. A comedian is planning to make his show more appealing by telling three jokes at the beginning of each show. He has been booked in a bar for the next three months. If he gives one performance a night, and he never wants to repeat the same set of jokes on any two nights, what is the minimum number of jokes he needs to prepare?
10. In how many ways can three people be selected from 5 Japanese, 5 Koreans, and 5 Chinese if each nationality must be represented?

Lesson 4

Probability

Learning Outcomes

- At the end of this lesson, you should be able to
 - define probability;
 - differentiate between the classical (*a priori*) approach and the relative frequency (*a posteriori*) approach of assigning probability; and
 - compute probabilities of events.

Introduction

In the previous lesson, we have defined an event to be a subset of the sample space of an experiment. In practice, we may want to determine the chance of an event happening. When we do, we are interested in finding the *probability* of an event.

Definition 1

Probability is a measure of the likelihood of occurrence of an event.

Suppose an experiment has a sample space S , where A is an event defined on S . The probability of event A is denoted by $P(A)$, and we formulate the following axioms of probability:

Axiom 1: The probability of event A is a real number between 0 and 1; that is $0 \leq P(A) \leq 1$.

Axiom 2: The probability that the outcome of an experiment will be an element of the sample space S is 1; that is, $P(S) = 1$.

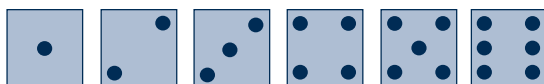
Axiom 3: Suppose that A_1, A_2, A_3, \dots is a sequence of mutually exclusive events defined on S , then

$P(A_1 \cup A_2 \cup A_3 \cup \dots) = \sum_{i=1}^{\infty} P(A_i)$, meaning, the probability that at least one of the mutually exclusive events A_1, A_2, A_3, \dots will occur is merely the sum of their respective individual probabilities.

The null event, \emptyset , does not contain any sample point and it has a probability 0 of occurring; that is, $P(\emptyset) = 0$.

One approach in defining the probability of an event is in terms of the **classical probability** or the **a priori** approach. In the classical probability approach, the experiment is not performed but the probability of an event may be calculated in advance. This approach is based on the assumption that the elements of the sample space are given equal weights of occurrence, thus the events are said to be equally likely to occur. Under this approach, if an experiment can result in any one of N equally likely outcomes of which exactly n are attributed to the event A , then the probability of event A is $P(A) = \frac{n}{N}$.

Example 1



In a single roll of a fair six-sided die, what is the probability of obtaining a number that is a perfect square?

Solution:

Using the classical probability approach, all six outcomes $\{1, 2, 3, 4, 5, 6\}$ are equally likely to occur, thus, $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$. Among these possible outcomes, only 1 and 4 are perfect squares. Hence, by axiom 3, the desired probability is $P(\{1, 4\}) = P(1) + P(4) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$.

Example 2

A raffle draw has eight possible outcomes labeled as letters S, H, O, U, L, D, E, and R. These letters are written on chips and placed in a box. These chips are loaded in such a way that each vowel is twice as likely to occur as the consonant. What is the probability of drawing a vowel?

Solution:

If we assign a probability w to each consonant and probability $2w$ to each vowel, then we have $w + w + 2w + 2w + w + w + 2w + w = 1$, in accordance with axiom 2. Accordingly,

$11w = 1$ and $w = \frac{1}{11}$. Then the probability of drawing a vowel is

$$P(\{O, U, E\}) = \frac{2}{11} + \frac{2}{11} + \frac{2}{11} = \frac{6}{11},$$

and the probability of drawing a consonant is

$$P(\{S, H, L, D, R\}) = \frac{1}{11} + \frac{1}{11} + \frac{1}{11} + \frac{1}{11} + \frac{1}{11} = \frac{5}{11}.$$

Example 3

The sample space S consists of the first M positive integers; that is, $S = \{1, 2, \dots, M\}$. A number is selected at random from S . What is the probability that it is divisible by an integer m , where $1 \leq m \leq M$?

Solution:

Using the classical probability approach, each of the elements in S is given equal weight of $\frac{1}{M}$ of being selected. Let A be the event where the selected number is divisible by integer m , where $1 \leq m \leq M$. The number of elements in S that are attributed to event A is actually $\left\lfloor \frac{M}{m} \right\rfloor$, which is the greatest integer less than or equal to $\frac{M}{m}$.

Therefore, the probability of selecting a number divisible by m from the set of the first M positive integers $S = \{1, 2, \dots, M\}$, where $1 \leq m \leq M$, is $P(A) = \frac{\left\lfloor \frac{M}{m} \right\rfloor}{M}$.

As a specific example, if $M = 1,000$, and $m = 3$, and we want to find the probability of selecting a number divisible by 3 from the first 1,000 positive integers, then the desired probability is

$$P(A) = \frac{\left\lfloor \frac{1,000}{3} \right\rfloor}{1,000} = \frac{333}{1,000} = 0.333.$$

Example 4

Two cards are drawn in succession and without replacement from a standard deck of 52 cards. What is the probability that both cards are face cards?

Solution:

The total number of equally likely pair of cards that may be drawn is $N = {}_{52}P_2$. There are four suits in a standard deck, namely the clover (\clubsuit), diamond (\diamond), heart (\heartsuit), and spade (\spadesuit), and there are three face cards per suit, namely the king, the queen, and the jack. Thus the total number of face cards in a standard deck is $4 \times 3 = 12$.



The number of ways in which we can draw a pair of face cards in succession and without replacement is $n = {}_{12}P_2$. Using the classical probability approach, the probability of drawing two face cards is

$$P(A) = \frac{n}{N} = \frac{{}_{12}P_2}{{}_{52}P_2} = \frac{132}{2,652} = \frac{11}{221}.$$

Another approach in defining the probability of an event is in terms of the **relative frequency** or the **a posteriori approach**. Unlike the *a priori* approach where the experiment is not performed, in the relative frequency approach, the experiment is performed over and over again under exactly the same conditions. For every event A defined on the sample space S , under the relative frequency approach, if the experiment is performed repeatedly, then the probability of event A is

$$P(A) = \frac{\text{number of occurrences of event } A}{\text{number of experiment repetitions}}.$$

Probabilities of events computed using the classical probability approach and the relative frequency approach may not necessarily be the same. However, the motivation for the relative frequency approach of probability is that for repeated trials of an experiment, the probability $P(A)$ converges to a constant limiting value that is equal to the theoretical probability of event A .

In the succeeding sections, probabilities will be calculated using the classical probability approach.

Let's Practice

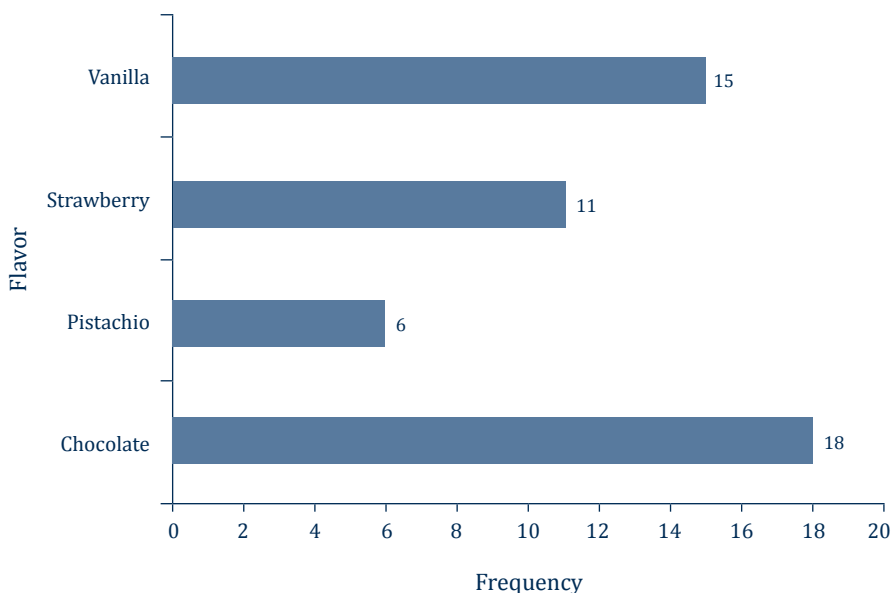
I. Write the letter that corresponds to the correct answer. Write X if your answer is not among the choices.

- _____ 1. What approach is used when the experimental outcomes are assigned with probabilities that are equally likely to occur?
 - a. classical approach
 - b. objective approach
 - c. relative frequency approach
 - d. subjective approach
- _____ 2. What approach is used when the assigned probabilities are the results of repeated trials of an experiment or historical data?
 - a. classical approach
 - b. objective approach
 - c. relative frequency approach
 - d. subjective approach
- _____ 3. What approach is used when the assigned probabilities are based on a person's judgment?
 - a. classical approach
 - b. objective approach
 - c. relative frequency approach
 - d. subjective approach

- _____ 4. Which statement is true regarding the probability assigned to an experimental outcome?
- It can be any positive value.
 - It can be less than zero.
 - It is between zero and one.
 - It is equal to one.
- _____ 5. An urn contains g green balls, v violet balls, and o orange balls. If the classical method for computing probability is used, what is the probability that a randomly drawn ball from the urn is green?
- $\frac{1}{3}$
 - $\frac{g}{3(v+o)}$
 - $\frac{g}{g+v+o}$
 - $\frac{g}{3}$

II. Analyze and solve each problem.

1. The following bar graph shows the distribution of the flavours of ice cream ordered by school children. What is the probability that a randomly selected child from the group has ordered chocolate ice cream?



2. A number is selected at random from the set of natural numbers $\{1, 2, 3, 4, \dots, 1000\}$. What is the probability that the number selected is
- an even number?
 - divisible by 7?
 - divisible by both 2 and 7?

3. A cabinet contains 6 distinct pairs of socks. If you select 4 pieces of socks at random from the cabinet, what is the probability that you get 2 matching pairs of socks?
4. If you select a random arrangement of the letters of the word LEXICOGRAPHY, what is the probability that there are exactly three consonants in the last five letters of the random arrangement?
5. A pair of dice is tossed, one brown (b) and one crème (c). The outcomes determine the coefficients of the quadratic equation $x^2 + bx + c = 0$. What is the probability that the equation will have no real roots?
6. If six fair dice are tossed simultaneously, find the probability that
 - a. each of the possible numbers 1 through 6 will occur?
 - b. exactly 3 pairs of numbers will occur?
7. If 13 cards are dealt at random from a standard deck to a player, what is the probability of getting
 - a. exactly 8 cards of one particular suit?
 - b. exactly two sets of 6 cards each, all of the same suit?
8. Consider all the possible subsets of the set $S = \{B, E, A, U, T, I, F, Y\}$. One of these subsets is selected at random. What is the probability that the selected subset
 - a. contains the letter Y?
 - b. contains exactly 2 vowels and 2 consonants?
 - c. contains all the vowels?
9. Trina is a philatelist, and she has just received special edition stamps from the Philippines, Sweden, Japan, Denmark, Nigeria, Colombia, Thailand, Argentina, Uganda, France, Singapore, and Germany. She intends to display the stamps lined up on a frame. What is the probability that she has arranged them so that
 - a. the countries of the same continent are next to each other?
 - b. no 2 Asian flags are adjacent to each other?
10. There are 3 accountants, 2 professors, and 1 lawyer that line up at a service counter to renew their driver's licenses. What is the probability that the arrangement in the line is such that no 2 persons with the same profession are adjacent to each other?

Lesson 5

Some Rules of Probability

Learning Outcomes

- At the end of this lesson, you should be able to
 - know the basic rules of probability;
 - demonstrate an understanding of the concept of probabilities involving a complement, union, and intersection of events;
 - apply the appropriate rule of probability when solving a problem;
 - solve problems using the various rules of probability such as conditional probability, multiplication, independence, and law of total probability; and
 - compute probabilities using Bayes's formula.

Introduction

From the three axioms of probability mentioned in the previous lesson, other probability rules may be derived, which have important applications.

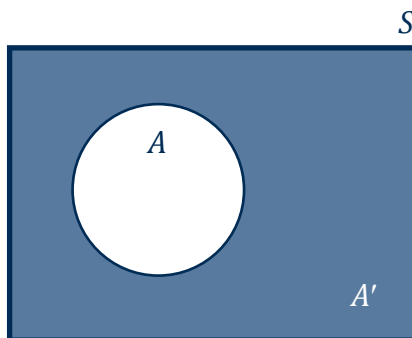
Rule of Probability 1

Rule of Complement

Suppose that an event A and its complement A' are defined on a sample space S . Then

$$P(A') = 1 - P(A)$$

Refer to the figure below. The sample space S is partitioned as A and A' , which are two mutually exclusive events, that is, $A \cap A' = \emptyset$. Then it follows from axioms 2 and 3 that $P(A \cup A') = P(S) = 1 \Rightarrow P(A) + P(A') = 1$. Therefore, $P(A') = 1 - P(A)$.



Example 1

A three-digit number that does not begin with zero must be formed from the numbers 0, 1, 2, 3, 4, and 5. Repetition of digits is allowed. What is the probability that the formed three-digit number contains the number 2 at least once?

Solution:

The sample space S consists of all possible three-digit numbers that do not begin with 0, formed from the numbers 0, 1, 2, 3, 4, and 5. There are 180 such three-digit numbers, by the rule of product.

$$\begin{array}{ccccc} 5 & \times & 6 & \times & 6 & = 180 \\ \text{for the} & & \text{for the} & & \text{for the} & \\ \text{hundreds digit} & & \text{tens digit} & & \text{units digit} & \end{array}$$

Let A be the event that the number formed contains the digit 2 at least once. Then its complement A' is the event that the number formed does not contain the digit 2 at all. There are 100 such three-digit numbers. The digit 2 is eliminated from the number of choices for each place.

$$\begin{array}{ccccc} 4 & \times & 5 & \times & 5 & = 100 \\ \text{for the} & & \text{for the} & & \text{for the} & \\ \text{hundreds digit} & & \text{tens digit} & & \text{units digit} & \end{array}$$

Thus, the probability that the number formed contains the number 2 at least once is

$$\begin{aligned} P(A) &= 1 - P(A') \\ &= 1 - \frac{100}{180} \\ &= 1 - \frac{5}{9} \\ &= \frac{4}{9} \end{aligned}$$

Example 2

Three letters are randomly selected to form a word (meaningful or not) from the letters R, A, I, N, B, O , and W . What is the probability that the word formed contains the letter W ?

Solution:

The sample space S consists of all possible three-letter words that can be formed from the letters R, A, I, N, B, O , and W . There are ${}_7P_3 = 210$ elements in the sample space.

Let A be the event where the word formed contains the letter W . Then its complement A' is the event that the word formed does not contain the letter W at all. There are ${}_6P_3 = 120$ such words. The letter W is removed from the possible choices of letters.

Thus, the probability that the word formed contains the letter W is

$$\begin{aligned} P(A) &= 1 - P(A') \\ &= 1 - \frac{120}{210} \\ &= 1 - \frac{4}{7} \\ &= \frac{3}{7}. \end{aligned}$$

Example 3

A group of seven students bought seven seats in a row for a concert. Among the students are Anne and Kim, who do not seem to get along with each other and refuse to sit beside each other. What is the probability that a seating arrangement will be favorable to these two feuding students?

Solution:

The sample space S consists of all the possible linear arrangements of seven students. The sample space S has $7! = 5,040$ elements. Let A be the event where Anne and Kim are separated by at least one student, that is, they are not seated beside each other. Then its complement A' is the event where Anne and Kim are seated beside each other. There are 1,440 such arrangements in A' .

$$\begin{array}{ccccc} 2! & \times & 6! & = & 1,440 \\ \text{Anne and Kim} & & \text{arranging 5 students} & & \\ \text{beside each} & & \text{+ the block of Anne} & & \\ \text{other} & & \text{and Kim} & & \end{array}$$

Thus, the probability that Anne and Kim are seated apart by at least one student is

$$\begin{aligned} P(A) &= 1 - P(A') \\ &= 1 - \frac{1440}{5040} \\ &= 1 - \frac{2}{7} \\ &= \frac{5}{7}. \end{aligned}$$

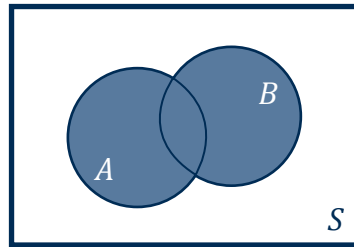
Rule of Probability 2

Rule of Union

Suppose that A and B are two events defined on a sample space S . Then

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

Refer to the figure below. To find the probability of the union $A \cup B$, we add their individual probabilities as $P(A) + P(B)$, but in doing so, we have added the middle section, which is $A \cap B$, two times. Therefore, to compute $P(A \cup B)$, we get the sum of $P(A)$ and $P(B)$, and then subtract the probability of the intersection $P(A \cap B)$ once, obtaining $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.



Special Case of Rule of Union

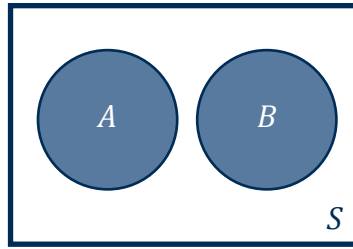
Rule of Union for Mutually Exclusive Events

Suppose that A and B are two mutually exclusive events defined on a sample space S . Then

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

Refer to the figure below. To find the probability of the union $A \cup B$, we add their individual probabilities. Since there are no common events for sets A and B , $P(A \cap B) = 0$, therefore, using the Rule of Union,

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - 0 \\ &= P(A) + P(B) \end{aligned}$$



This rule follows from axiom 3 of probabilities.

Example 4

A survey was conducted among 600 men. Each of them was asked for their opinion regarding the issue of the passing of a bill allowing divorce in the country. The men were classified as college students, working professionals, and retired people. The observed responses, whether they were in favor of, against, or undecided, are summarized as follows:

Divorce Bill	Occupation			Total
	College Student (D)	Working Professional (E)	Retired Person (R)	
In Favor (A)	74	76	5	155
Against (B)	158	207	40	405
Undecided (C)	18	17	5	40
Total	250	300	50	600

A man is randomly selected among the 600 men. Find the probability that he is:

1. is in favor of the passing of the divorce bill.
2. is a working professional.
3. is retired and against the passing of the divorce bill.
4. is a college student or undecided about the passing of the divorce bill.

Solution:

Using the classical probability approach, where $N = 600$, we solve these problems as follows:

- a. Let A be the event where a randomly selected man among the 600 men is in favor of the passing of the divorce bill. Then $P(A) = \frac{155}{600} = 0.258\bar{3}$.
- b. Let E be the event where a randomly selected man among the 600 men is a working professional. Then $P(E) = \frac{300}{600} = 0.5$

The probabilities $P(A)$ and $P(E)$ are examples of *marginal probabilities*, since they are located along the margins of the table. Marginal probabilities are for *simple events* or *singular events*.

- c. Let F be the event where a randomly selected man among the 600 men is retired and B be the event that he is against the passing of the divorce bill. The desired probability is $P(F \cap B) = \frac{40}{600} = 0.06\bar{6}$. This is an example of a *joint probability*, which involves the probability of the intersection of two singular events.
- d. Let C be the event that a randomly selected man among the 600 men is undecided about the passing of the divorce bill and D be the event that he is a college student. The desired probability can be obtained by using the rule of union, that is:

$$\begin{aligned} P(C \cup D) &= P(C) + P(D) - P(C \cap D) \\ &= \frac{40}{600} + \frac{250}{600} - \frac{18}{600} \\ &= \frac{272}{600} = 0.45\bar{3} \end{aligned}$$

Example 5

Six letters are randomly selected to form a word (meaningful or not) from the letters H, Y, S, T, E, R, I, C, A, and L. What is the probability that the word formed contains either the sequence (intact, as a block) AIR or the sequence SHY?

Solution:

The sample space S consists of all possible six-letter words that can be formed from the letters H, Y, S, T, E, R, I, C, A, and L, and there are $_{10}P_6 = 151,200$ elements in the sample space. Let A be the event where the word formed contains the sequence “AIR”. How many elements are attributed to event A ? By the rule of product, there are 840 such words formed.

$$1 \times {}_7C_3 \times 4! = 1 \times 35 \times 24 = 840$$

1 way to
form the
sequence
AIR

Select 3 more
letters from
the remaining
7 letters
(excluding A,
I, R)

Arrange the
3 letters and
the block with
the sequence
AIR

Let B be the event where the word formed contains the sequence SHY. Similarly, there are 840 such words formed. However, included among these 840 words are those that contain both the sequences AIR and SHY (i.e., the words AIRSHY and SHYAIR). The desired

probability may be obtained using the rule of union. The probability that the formed word contains either the sequence AIR or the sequence SHY is:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{840}{151200} + \frac{840}{151200} - \frac{2}{151200} \\ &= \frac{1678}{151200} \approx 0.0111 \end{aligned}$$

Example 6

At a state university, $\frac{5}{9}$ of the students enrolled in the Master of Business Administration (MBA) program are under 25 years old. Moreover, $\frac{3}{5}$ of the students are females, and $\frac{4}{9}$ of the students are females under 25 years old. A student enrolled in the MBA program of the state university is randomly selected. What is the probability that the selected student is male and at least 25 years old?

Solution:

Let A be the event where a randomly selected student enrolled in the MBA program is under 25 years old, and let B be the event where a student is female. The probability of the union, $P(A \cup B)$, is the probability that the selected student is under 25 years old or is female. It is computed as

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{5}{9} + \frac{3}{5} - \frac{4}{9} \\ &= \frac{32}{45} \end{aligned}$$

Then the probability that the selected student is male and at least 25 years old may be obtained by De Morgan's law.

$$\begin{aligned} P(A' \cap B') &= P(A \cup B)' \\ &= 1 - P(A \cup B) \\ &= 1 - \frac{32}{45} \\ &= \frac{13}{45} \end{aligned}$$

Example 7

Two numbers a and b are *relatively prime* if 1 is their only common positive divisor. Hence, 6 and 7 are relatively prime but 6 and 9 are not. Let the sample space consist of the first 45 positive integers; that is, $S = \{1, 2, \dots, 44, 45\}$. Suppose a number is randomly selected from S . What is the probability that it is relatively prime to 45?

Solution:

The number 45 may be factored into prime numbers as $45 = 3^2 \cdot 5$. If we denote A as the event where the selected number is divisible by 3, and B is the event where the selected number is divisible by 5, then the union $P(A \cup B)$ is the probability that the selected number is divisible by either 3 or 5. By the rule of union,

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\&= \frac{\left\lfloor \frac{45}{3} \right\rfloor}{45} + \frac{\left\lfloor \frac{45}{5} \right\rfloor}{45} - \frac{\left\lfloor \frac{45}{15} \right\rfloor}{45} \\&= \frac{15 + 9 - 3}{45} \\&= \frac{21}{45} \text{ or } \frac{7}{15}.\end{aligned}$$

Note that the probability of the intersection $P(A \cap B)$ is the probability that the selected number is divisible by both 3 and 5, or divisible by 15. To find the probability that the selected number is relatively prime to 45, we use De Morgan's law to find $P(A \cup B)'$, because the numbers that are relatively prime to 45 are those that are neither divisible by 3 nor by 5. Hence, the desired probability of selecting a number at random from S that is relatively prime to 45 is

$$\begin{aligned}P(A \cup B)' &= P(A' \cap B') \\&= 1 - P(A \cup B) \\&= 1 - \frac{21}{45} \\&= \frac{24}{45} \text{ or } \frac{8}{15}.\end{aligned}$$

In fact, by enumeration, the 24 numbers in S that are relatively prime to 45 are $\{1, 2, 4, 7, 8, 11, 13, 14, 16, 17, 19, 22, 23, 26, 28, 29, 31, 32, 34, 37, 38, 41, 43, 44\}$.

Suppose that A and B are defined on a sample space S , and that we have knowledge that the event A has already occurred. We are now faced with the knowledge that A has occurred, and the original sample space S has “shrunk” in effect. If we are now interested in finding the probability that event B will occur, with the additional information that event A has already occurred, we need to revise the sample space S , which has been reduced as a result of occurrence. We introduce the next rule of probability, the concept of conditional probability.

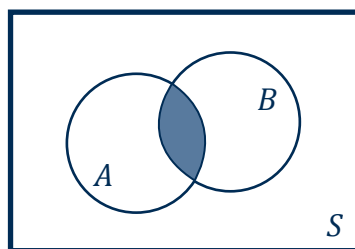
Rule of Probability 3

Rule of Conditional Probability

Let A and B be two events defined on a sample space S . Suppose that A has already occurred, and that $P(A) > 0$. The conditional probability of B given that A has already occurred is denoted by $P(B | A)$ and defined as

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

Refer to the figure below. To find the conditional probability of B given that A has occurred, $P(B | A)$, we reduce the sample space to A . The shaded area is $A \cap B$. Therefore, to compute $P(B | A)$, we restrict the outcomes to A , which now serves as the sample space, and compute the quotient of $P(A \cap B)$ and $P(A)$, giving us the formula $P(B | A) = \frac{P(A \cap B)}{P(A)}$.



Example 8

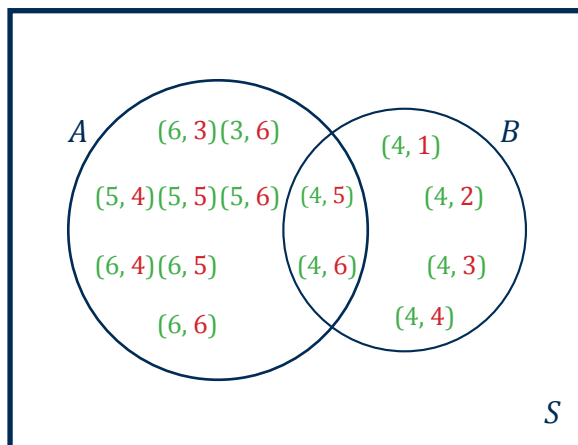
An experiment involves rolling a green and a red dice. The sample space would consist of 36 outcomes that are all equally likely to occur. If we write the outcomes in the form (g, r) , where g is the outcome on the green die and r is the outcome on the red die, then we enumerate the 36 outcomes as follows:

		Red Die					
		1	2	3	4	5	6
Green Die	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)

What is the probability that the green die shows a 4 if it is known that the sum of the spots on the two dice is at least 9?

Solution:

The Venn diagram below shows the elements of A , B , and $A \cap B$.



Let A be the event where the sum of the spots on the two dice is at least 9; that is, $g + r \geq 9$. Then the elements of A are $\{(3, 6), (4, 5), (5, 4), (6, 3), (6, 4), (5, 5), (4, 6), (5, 6), (6, 5), (6, 6)\}$.

Therefore, $P(A) = \frac{10}{36} = \frac{5}{18}$. Now, let B be the event where the green die shows a 4. The elements of B are $\{(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6)\}$. The only elements that are common to A and B comprise the intersection $A \cap B$, and these elements are $\{(4, 5), (4, 6)\}$.

Hence, $P(A \cap B) = \frac{2}{36} = \frac{1}{18}$. Therefore, the conditional probability that the green die shows a 4 given that the sum of the spots is $g + r \geq 9$ is

$$P(B|A) = \frac{\frac{1}{18}}{\frac{5}{18}} = \frac{1}{5}.$$

Example 9

Revisit example 4. Suppose it is known that a randomly selected man among the 600 men is retired. What is the probability that he is in favor of the passing of the divorce bill?

Solution:

Refer to the table below to determine the given.

	Occupation			
Divorce Bill	College Student (D)	Working Professional (E)	Retired Person (R)	Total
In Favor (A)	74	76	5	155
Against (B)	158	207	40	405
Undecided (C)	18	17	5	40
Total	250	300	50	600

Let F be the event where the randomly selected man is retired. According to the table, 50 out of the 600 men are retired; therefore, $P(F) = \frac{50}{600}$. Now, let A be the event where the randomly selected man is in favor of the passing of the divorce bill. To find the conditional probability $P(A|F)$, we need the intersection $A \cap F$, which is for the retired men who are in favor of the passing of the divorce bill. From the table, we see that there are 5 such men; hence, $P(A \cap F) = \frac{5}{600}$. Finally, the conditional probability that the randomly selected man is in favor of the passing of the divorce bill, given that he is retired is

$$P(A|F) = \frac{\frac{5}{600}}{\frac{50}{600}} = \frac{5}{50} = 0.10.$$

Example 10

In a small community, 40 percent of the homeowners have at least two vehicles. Among these homeowners who have at least two vehicles, 70 percent own a pet dog. Furthermore, 30 percent of the homeowners have a pet dog. Suppose a homeowner is randomly selected from this small community.

1. What is the probability that he or she owns both a pet dog and at least two vehicles?
2. What is the conditional probability that he or she owns at least two vehicles, given that he or she owns a pet dog?

Solution:

Let A be the event where a randomly selected homeowner has at least two vehicles. Then $P(A) = 0.40$. Let B be the event where a randomly selected homeowner has a pet dog. Then $P(B) = 0.30$. The conditional probability $P(B | A) = 0.70$ is from the given information. To answer item 1, we can substitute the information into the formula for conditional

probability, $P(B | A) = \frac{P(A \cap B)}{P(A)}$. Therefore, the probability that a randomly selected

homeowner has at least two vehicles and owns a pet dog is

$$\begin{aligned}P(A \cap B) &= P(A) \cdot P(B | A) \\&= 0.4 \times 0.7 \\&= 0.28.\end{aligned}$$

To answer item 2, we find the conditional probability $P(A | B) = \frac{P(A \cap B)}{P(B)}$. Substituting the values into this equation, we find that

$$\begin{aligned}P(A | B) &= \frac{0.28}{0.30} \\&= \frac{14}{15}.\end{aligned}$$

In the previous example, we have seen that from the definition of conditional probability, $P(B | A) = \frac{P(A \cap B)}{P(A)}$, we can determine the joint probability $P(A \cap B)$ by multiplying $P(A)$ with the conditional probability $P(B | A)$.

Rule of Probability 4

Rule of Multiplication

Let A and B be two events defined on a sample space S . Then if $P(A) > 0$, then

$$P(A \cap B) = P(A) \cdot P(B | A).$$

Since the operation of the intersection of events is commutative, that is, $A \cap B = B \cap A$, alternatively, for $P(B) > 0$, we can write $P(A \cap B) = P(B) \cdot P(A | B)$.

Example 11

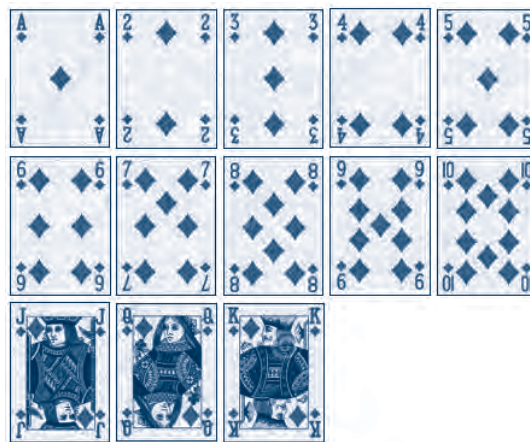
Two cards are drawn in succession and without replacement from a standard deck of 52 cards. What is the probability that both cards drawn are diamond cards?

Solution:

Let D_i be the event where the i th drawn card is a diamond, for $i = 1, 2$. We use the rule of multiplication to find the probability that both cards drawn are diamonds, $P(D_1 \cap D_2)$.

$$P(D_1 \cap D_2) = P(D_1) \cdot P(D_2 | D_1)$$

There are 13 diamond cards in a deck as shown below.



Hence, $P(D_1) = \frac{13}{52}$. Since the cards are drawn in succession, without replacement, $P(D_2 | D_1) = \frac{12}{51}$. Therefore,

$$\begin{aligned} P(D_1 \cap D_2) &= P(D_1) \cdot P(D_2 | D_1) \\ &= \frac{13}{52} \times \frac{12}{51} \\ &= \frac{1}{17}. \end{aligned}$$

It is also interesting to note that we can use counting techniques to solve for the unknown. The sample space consists of all possible selections of two cards from the deck containing 52 cards, or ${}_{52}C_2 = \binom{52}{2}$. Since we are interested in both selected cards being diamonds, there are ${}_{13}C_2 = \binom{13}{2}$ such selections. Accordingly, the probability that both selected cards are diamonds is

$$\begin{aligned} P(D_1 \cap D_2) &= \frac{{}_{13}C_2}{{}_{52}C_2} \\ &= \frac{78}{1326} \\ &= \frac{1}{17}. \end{aligned}$$

Suppose that A and B are two events defined on a sample space S . A and B are said to be **independent** if A and B do not influence each other; that is, if the occurrence of event A does not in any way affect the occurrence of event B , and vice versa. Consequently, if A and B are independent, then the conditional probability $P(B | A)$ is equal to $P(B)$ since the occurrence of event A does not affect the occurrence of event B . Alternatively, if A and B are independent, then $P(A | B)$ is equal to $P(A)$.

Rule of Probability 5

Rule of Independence

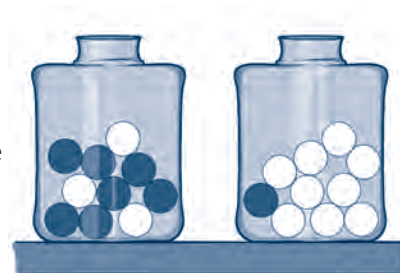
Let A and B be two events defined on a sample space S . Then A and B are **independent** if and only if

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

In contrast, if A and B are not independent, they are said to affect each other's occurrence, and considered to be dependent. When A and B are dependent events, the rule of multiplication applies for computing the joint probability $P(A \cap B)$; that is, $P(A \cap B) = P(A) \cdot P(B | A)$. The rule of independence is a special case of the rule of multiplication and applies when $P(B | A) = P(B)$. In the case when A and B are independent, $P(A \cap B) = P(A) \cdot P(B)$.

Example 12

Consider two urns. Urn A has seven blue marbles and three white marbles. Urn B has nine white marbles and one blue marble. A marble is drawn at random from each urn. What is the probability that both marbles are white?



Solution:

Let A be the event where the marble drawn from Urn A is white, and B be the event where the marble drawn from Urn B is also white. The desired probability is $P(A \cap B)$. We use the rule of independence because the probability of drawing a white marble from one urn does not affect the probability of drawing a white marble from the other urn. Now,

$P(A) = \frac{3}{10}$ and $P(B) = \frac{9}{10}$. Therefore,

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ &= \frac{3}{10} \times \frac{9}{10} \\ &= \frac{27}{100}. \end{aligned}$$

Example 13

A building with nine floors has an elevator that services all floors. Two passengers, who do not know each other, get on the elevator at the ground floor (level 0). Assuming that each of these passengers is equally likely to get off at any of the nine floors, what is the probability that at least one of them gets off at the seventh floor?

Solution:

Let A be the event where one of the passengers (passenger A) gets off at the seventh floor, and let B be the event where the other passenger (passenger B) also gets off at the seventh floor. Since there are nine floors above the ground floor (designated as level 0), and each passenger is equally likely to get off at any of those floors, $P(A) = P(B) = \frac{1}{9}$. To find the probability that at least one of the passengers gets off at the seventh floor, we use the rule of union to find the probability that either passenger A, or passenger B, or both of them get off at the seventh floor.

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

But what is $P(A \cap B)$? Since passengers A and B do not know each other, it is reasonable to assume that they choose their destination floor independently of each other. We can use the rule of independence to find the probability that they both get off at the seventh floor.

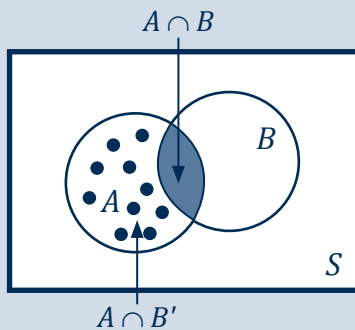
Hence, $P(A \cap B) = P(A) \cdot P(B) = \frac{1}{9} \times \frac{1}{9} = \frac{1}{81}$. Finally, the probability that at least one of the passengers get off at the seventh floor is

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= \frac{1}{9} + \frac{1}{9} - \frac{1}{81} \\ &= \frac{17}{81}. \end{aligned}$$

Rule of Probability 6

Law of Total Probability

Consider two events A and B defined on a sample space S , as shown in the figure below.



Suppose that the probability of the occurrence of event A cannot be calculated directly. However, for event B where $P(B) > 0$ and $P(B') > 0$, the conditional probabilities $P(A | B)$ and $P(A | B')$ can be determined. Observe from the figure that the event A can be written as the union of two mutually exclusive events, $A \cap B$ (the shaded region) and $A \cap B'$ (the dotted region). Then, by the axioms of probability, we may write the probability of event A as

$$\begin{aligned} P(A) &= P[(A \cap B) \cup (A \cap B')] \\ &= P(A \cap B) + P(A \cap B'). \end{aligned}$$

By the multiplication rule, the last expression can be written as the **law of total probability**; that is,

$$P(A) = P(A | B) \cdot P(B) + P(A | B') \cdot P(B').$$

This shows that the probability of event A is the “weighted average” of the probability of event A given that B has occurred and the probability of event A given that B has not occurred.

Example 14

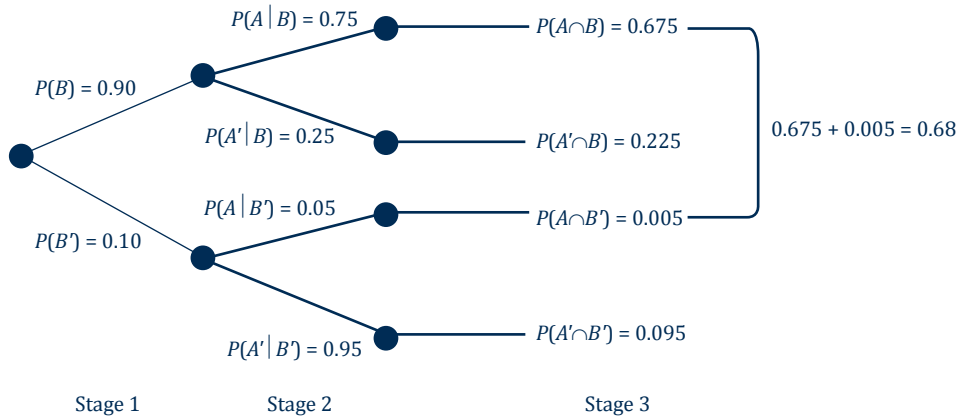
Bobby is fond of collecting refrigerator magnets, the ever popular travel souvenir. Ninety percent of the magnets in his collection are those that he purchased during his personal travels, and the rest are gifts from friends who know that he collects them. Among the magnets that he purchased himself, 75% are from the European cities he visited, whereas 5% of the magnets that are gifts from friends are from European cities as well. His niece Monica marveled at the numerous magnets in his collection and randomly chose one of them. What is the probability that she chose a magnet from a European city?

Solution:

Let A be the event where Monica selected a magnet from a European city. This chosen magnet could either be one from Bobby’s personal travels or a gift from friends. Let B be the event where a randomly chosen magnet is from Bobby’s personal travels and B' be the event where the magnet is a gift from friends. By the law of total probability,

$$\begin{aligned} P(A) &= P(A | B) \cdot P(B) + P(A | B') \cdot P(B') \\ &= (0.75) \cdot (0.90) + (0.05) \cdot (0.10) \\ &= 0.68. \end{aligned}$$

Tree diagrams are useful tools that facilitate the solutions to problems of this type. Consider the following tree diagram. Stage 1 of the tree is for events B and B' , where the randomly chosen refrigerator magnet is from Bobby's personal travels and is a gift from friends, respectively. In this stage, the simple event probabilities $P(B)$ and $P(B')$ are indicated. Stage 2 of the tree is for the conditional probabilities, whereas stage 3 is for the joint probabilities, which are obtained by multiplying the respective probabilities in stages 1 and 2, as a result of the multiplication rule.



Example 15

Urn 1 has 3 red balls and 5 blue balls. Urn 2 has 4 red balls and 4 blue balls. A ball is chosen at random from urn 1 and discarded. All the remaining balls in urn 1 are then transferred to urn 2. Now, a ball is drawn from urn 2. What is the probability that it is blue?

Solution:

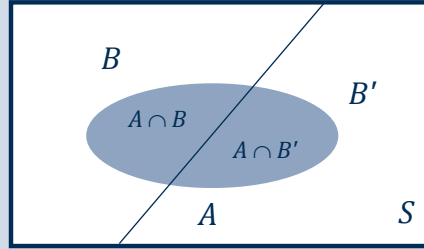
Let A be the event where the ball drawn from urn 2 is blue. The probability of event A depends on the color of the discarded ball from urn 1. Let B be the event where a blue ball was discarded from urn 1. This results in transferring 3 red balls and 4 blue balls to Urn 2, raising the total number of balls in urn 2 to 15 balls, of which 7 are red and 8 are blue. For this case, the conditional probability is $P(A|B) = \frac{8}{15}$. Likewise, if B' is the event where a red ball was discarded from urn 1, then this results in transferring 2 red balls and 5 blue balls to urn 2, and so it will contain a total of 15 balls, of which 6 are red and 9 are blue. Then the conditional probability is $P(A|B') = \frac{9}{15}$. The probability of event A is computed as follows using the law of total probability:

$$\begin{aligned}
 P(A) &= P(A|B) \cdot P(B) + P(A|B') \cdot P(B') \\
 &= \left(\frac{8}{15}\right) \cdot \left(\frac{5}{8}\right) + \left(\frac{9}{15}\right) \cdot \left(\frac{3}{8}\right) = \frac{67}{120}
 \end{aligned}$$

Rule of Probability 7

Bayes's Formula

Suppose that the sample space S is partitioned as $\{B, B'\}$ such that $B \cup B' = S$ and $B \cap B' = \emptyset$, as shown in the figure below.



If $P(B) > 0$ and $P(B') > 0$, then for an event A defined on S with $P(A) > 0$, the conditional probability $P(B|A)$ can be written as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B') \cdot P(B')}$$

The numerator $P(A \cap B) = P(A) = P(A|B) \cdot P(B)$ is due to the multiplication rule and the denominator $P(A) = P(A|B) \cdot P(B) + P(A|B') \cdot P(B')$ is due to the law of total probability.

The conditional probability $P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B') \cdot P(B')}$ is called **Bayes's formula**.

Example 16

In the field of drug testing, *sensitivity* pertains to the percentage of persons who are truly ill which the drug indicates have the illness, whereas *specificity* pertains to the percentage of well persons which the drug indicates do not have the illness. Suppose that a new brand of pregnancy test kit advertises itself as “99.5% sensitive” and “99.9% specific.” In other words, the test kit will give a positive result for a woman who is truly pregnant 99.5% of the time, and will yield a negative result 99.9% of the time for tested women who are not pregnant. In a group of 100 women of which only 12 are truly pregnant but not yet aware of their condition, a woman is selected at random and the test kit is administered. Her test result is positive. What is the probability that she is truly pregnant?

Solution:

Let A be the event where the pregnancy test result is positive, and let B be the event where the woman is truly pregnant. Then $P(B) = 0.12$ and $P(B') = 0.88$. The sensitivity of the test, which is 0.995, is the conditional probability $P(A|B)$. On the other hand, the complement of the specificity, 0.001, is the conditional probability $P(A|B')$. This means that the test will indicate a false positive result in 0.1% of the time. The desired conditional probability $P(B|A)$ is computed by Bayes's Formula, as follows:

$$\begin{aligned}P(B|A) &= \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B') \cdot P(B')} \\&= \frac{(0.995) \cdot (0.12)}{(0.995) \cdot (0.12) + (0.001) \cdot (0.88)} \\&= 0.9927\end{aligned}$$

Points to Remember

In example 16, the conditional probability $P(B|A)$ is referred to in medical studies as the positive predictive value of the test. The sensitivity of a test indicates the “true positive” rate, meaning, the percentage of persons with a certain condition that is correctly diagnosed by the test as having the condition. On the other hand, the positive predictive value is the percentage of persons with a diagnostic test result of positive that truly have the condition. A helpful method for solving problems similar to example 16 is to construct a table such as this:

		Woman is Pregnant?		
		Yes (B)	No (B')	Totals
Test Result	Positive (A)	w	x	$w + x$
	Negative (A')	y	z	$y + z$
	Totals	$w + y$	$x + z$	1.0

Here,

$$w = P(A \cap B) = P(A|B) \cdot P(B) = (0.995)(0.12) = 0.1194$$

$$z = P(A' \cap B') = P(A'|B') \cdot P(B') = (0.999)(0.8) = 0.87912$$

The completed table is

		Woman is Pregnant?		
		Yes (B)	No (B')	Totals
Test Result	Positive (A)	0.1194	0.00088	0.12028
	Negative (A')	0.0006	0.87912	0.87972
	Totals	0.12	0.88	1.000000

Therefore, the positive predicted value of the test, which is the conditional probability $P(B|A)$ is

$$P(B|A) = \frac{w}{w+x} = \frac{0.1194}{0.12028} = 0.9927$$

Example 17

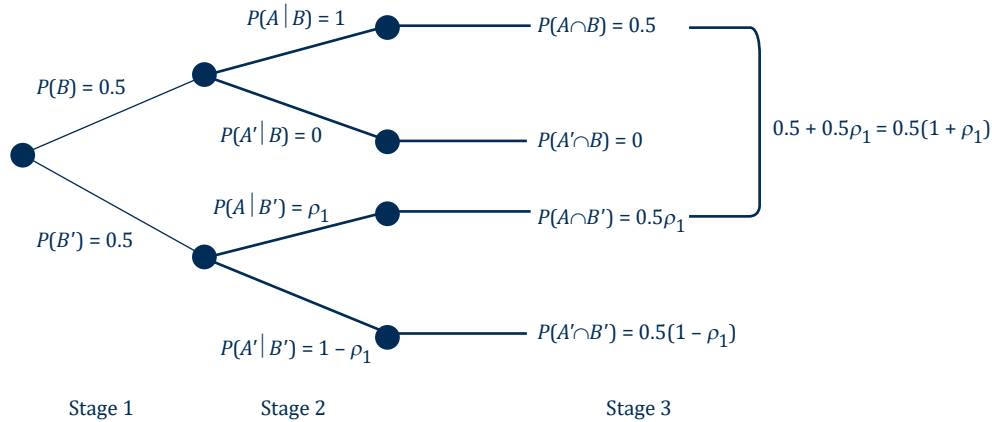
A clerk is searching for a document. The document is equally likely to be in one of two department offices, Purchasing Department and Accounting Department, each with several filing cabinets. Let ρ_1 be the probability that a quick search of the cabinets at the Purchasing Department will be unsuccessful in locating the document, if indeed, the document is there. Similarly, let ρ_2 be the probability that a quick search of the cabinets at the Accounting Department will be unsuccessful in locating the document, if indeed, the document is there. What is the conditional probability that the document is in the Accounting Department given that a search of the filing cabinets in the Purchasing Department is unsuccessful?

Solution:

Let A be the event of not finding the document after searching the Purchasing Department and let B be the event where the document is located in the Accounting Department. Since the document is equally likely to be in one of the two departments, $P(B) = 0.5$ and $P(B') = 0.5$. The overlook probability ρ_1 is actually the conditional probability $P(A|B')$, whereas the conditional probability $P(A|B)$ is equal to 1, because if the document is actually in the Accounting Department, a search of the Purchasing Department will surely not come up with the document. The conditional probability $P(B|A)$ is computed by Bayes's formula as follows:

$$\begin{aligned}
 P(B|A) &= \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B') \cdot P(B')} \\
 &= \frac{(1) \cdot (0.5)}{(1) \cdot (0.5) + (\rho_1) \cdot (0.5)} \\
 &= \frac{1}{1 + \rho_1}.
 \end{aligned}$$

Below is the corresponding tree diagram.



Let's Practice

I. Write True if the statement is *always* true; otherwise, write False.

- _____ 1. Given that event A has a probability of 0.35, the probability of the complement of event A must be 0.65.
- _____ 2. The union of events A and B is the event containing all the elements common to both A and B only.
- _____ 3. If A and B are mutually exclusive events with $P(A) = 0.3$ and $P(B) = 0.5$, then $P(A \cap B) = 0.8$.
- _____ 4. Two events with nonzero probabilities can be both mutually exclusive and independent.
- _____ 5. If a fair coin is tossed three times and comes up heads all three times, the probability of getting a head on the fourth toss is 1.
- _____ 6. Given two events A and B defined on a sample space S , if $P(A|B) = 0.30$, then $P(A|B') = 0.70$.
- _____ 7. Given two events A and B defined on a sample space S , if $P(A|B) = 0.30$, then $P(A'|B) = 0.70$.
- _____ 8. If two events A and B are independent, then their complements A' and B' are also independent.
- _____ 9. Suppose that A and B are independent events where $P(A) = 0.5$ and $P(B) = 0.4$. Then $P(A \cup B) = 0.70$.
- _____ 10. If $P(A \cap B) = 1$, then A and B must be mutually exclusive.

II. Analyze and solve the following problems.

1. If six fair dice are tossed simultaneously, find the probability that at least one 6 occurs.
2. A commercial bank made a survey to 500 of its clients and asked them how many credit cards they own. The bank summarized its findings in the table below:

	Number of Credit Cards Owned			
Gender	1	2	3	Total
Male	70	110	90	270
Female	25	175	30	230
Total	95	285	120	500

If a client is selected at random, what is the probability that the client

- a. is female?
 - b. owns one credit card only?
 - c. is male and owns 2 credit cards?
 - d. is female or owns 3 credit cards?
 - e. is male, if it is known he owns only one credit card?
3. Swirl Dishwashing Detergent has a 15% market share. It can be said that there is a 0.15 probability that a randomly selected buyer of dishwashing detergent will be a buyer of Swirl. Recently, Swirl aired a commercial on TV showing a famous comedian as endorser. A survey showed that 70% of Swirl users remember the new commercial, whereas 20% of non-Swirl users remember the new commercial.
 - a. What percentage of dishwashing detergent buyers remember the new Swirl commercial?
 - b. Suppose that a randomly selected dishwashing detergent buyer is selected at random, and it is known that she remembers the new Swirl commercial. What is the probability that she is a Swirl user?
 4. Consider the sample space $S = \{1, 2, 3, \dots, 89, 90\}$, and suppose a number is selected at random from S . Let A be the event where the selected number is divisible by 2. Let B be the event where the number is divisible by 3. And let C be the event where the number is divisible by 5.
 - a. Find the elements of event A . What is $P(A)$?
 - b. Find the elements of event B . What is $P(B)$?
 - c. Find the elements of event C . What is $P(C)$?
 - d. Find the elements of $A \cap B$, $A \cap C$, and $B \cap C$. Then find $P(A \cap B)$, $P(A \cap C)$, and $P(B \cap C)$.
 - e. Are events A and B independent? Are events A and C independent? Are events B and C independent?

5. Marla has just graduated from college and is applying for admission into medical school. There is a 50% chance she will be admitted to University A and a 40% chance she will be admitted to University B. However, there is a 25% chance that she will be denied admission at both universities.
 - a. What is the probability that Marla will be accepted at both universities?
 - b. Suppose we know that Marla was accepted at University B. What is the conditional probability that she was also accepted at University A?
6. A fitness center held a program similar in concept to the show “The Biggest Loser.” The objective of the program is to promote exercise and maintain proper weight. Of those who signed up for the program, 57% were females and 43% were males. A total of 40% of the women and 28% of the men who enrolled in the program were successful in losing weight and were able to maintain ideal weight for at least 6 months after completing the program. These people attended a reunion party to celebrate their success after one year. What percentage of those who attended the reunion are males?

Chapter Review

- An **experiment** is any procedure that can be repeated, theoretically, an infinite number of times and has a well-defined set of possible outcomes.
- Each possible result of an experiment is referred to as a **sample outcome**. The set of all possible outcomes is called the **sample space**, and is usually denoted by S .
- A sample space which has a finite or countably infinite number of outcomes is said to be **discrete**. On the other hand, a sample space with an uncountably infinite number of outcomes is said to be **continuous**.
- The **intersection** of A and B , denoted by $A \cap B$, is the event whose outcomes belong to both A and B .
- The **union** of A and B , denoted by $A \cup B$, is the set of all outcomes in A or B (or both).
- If $A \cap B = \emptyset$, then the events A and B are said to be **mutually exclusive**.
- The **complement** of A , denoted by A' or A^c , is the set of outcomes in S which are not in A .
- **Rule of Sum:** If a particular action can be done in m ways and another in n ways, and the two actions cannot be done at the same time, then there are $m + n$ ways of doing exactly one of these actions.
- **Rule of Product:** If a particular action can be done in m ways and another in n ways, then there are $m \times n$ ways of doing both actions (one after the other).

- **Rule of Product (General Case):** Given k actions, if the first action can be done in n_1 , the second action in n_2 ways, and so on, then there are $n_1 \times n_2 \times \dots \times n_k$ ways of doing all the actions.
- An ordered selection of r objects from a set A containing n objects (where $0 \leq r \leq n$) is called a **permutation** of the elements of A taken r at a time, denoted by ${}_nP_r$.
- For an integer $n \geq 0$, **n factorial**, denoted by $n!$, is defined by

$$n! = \begin{cases} n \times (n-1) \times \dots \times 3 \times 2 \times 1 & \text{if } n \geq 1 \\ 1 & \text{if } n = 0 \end{cases}$$

- The **number of permutations** of r elements that can be formed from a set of n distinct elements is given by ${}_nP_r = \frac{n!}{(n-r)!}$.
- The number of ways to arrange an entire set of n distinct objects is ${}_nP_n = n!$.
- A selection of r objects from a set A containing n objects (where $0 \leq r \leq n$) without regard to order is called a **combination** of the elements of A taken r at a time, denoted by $\binom{n}{r}$ or ${}_nC_r$.
- The **number of combinations** of r elements that can be formed from a set of n distinct elements is given by ${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$.
- **Probability** is a measure of the likelihood of occurrence of an event.
- **Classical probability approach:** If an experiment can result in any one of N equally likely outcomes of which exactly n are attributed to the event A , then the probability of event A is $P(A) = \frac{n}{N}$.
- **Relative frequency approach:** For every event A defined on the sample space S , under the relative frequency approach, if the experiment is performed repeatedly, then the probability of event A is $P(A) = \frac{\text{number of occurrences of event } A}{\text{number of experiment repetitions}}$.
- **Rule of Complement:** Suppose that an event A and its complement A' are defined on a sample space S . Then $P(A') = 1 - P(A)$.
- **Rule of Union:** Suppose A and B are two events defined on a sample space S . Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- **Rule of Union for Mutually Exclusive Events:** Suppose A and B are two mutually exclusive events defined on a sample space S . Then $P(A \cup B) = P(A) + P(B)$.
- **Rule of Conditional Probability:** Let A and B be two events defined on a sample space S . Suppose that A has already occurred, and that $P(A) > 0$. The conditional probability of B given that A has already occurred is denoted by $P(B|A)$, defined as $P(B|A) = \frac{P(A \cap B)}{P(A)}$.

- **Rule of Multiplication:** Let A and B be two events defined on a sample space S . Then if $P(A) > 0$, $P(A \cap B) = P(A) \cdot P(B|A)$.
- **Rule of Independence:** Let A and B be two events defined on a sample space S . Then A and B are **independent** if and only if $P(A \cap B) = P(A) \cdot P(B)$.
- **Law of Total Probability:**

$$P(A) = P(A|B) \cdot P(B) + P(A|B') \cdot P(B')$$

- **Bayes's Formula:** If $P(B) > 0$ and $P(B') > 0$, then for an event A defined on S with $P(A) > 0$, the conditional probability $P(B|A)$ can be written as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B') \cdot P(B')}$$

Chapter Performance Tasks

1. The Birthday Paradox

Imagine that you are invited to give a 30-minute talk to a group of senior high school students during their Math Week celebration. To catch students' interest, you planned to conduct an activity where each student will be asked his or her birthday. But before doing so, you ask students to form group(s) consisting of 23 members each. You are to ask the members of the group (or each group) to compare their birthdays. The highlight of the activity is to prove to them that there is a small number (n) of people that needs to be in a group so that the chance of having at least two persons sharing the same birthday is at least 50%. This is by the assumption that none of them were born on February 29 and that all the days in a year are equally likely to be a person's birthday. Regardless of the findings in the group (or each group), **show your proof on finding the smallest number (n) of people in a group, wherein the probability of obtaining at least two persons having the same birthday is 50%, through a PowerPoint presentation.** Make sure your proof is correct, neat, detailed, and organized. The organizers of the event will hand out evaluation forms to students to assess how much they learned from the talk and draw out suggestions for improvement, if any.



2. Revisiting the Monty Hall Problem

Monty Hall is the host of a game show called "Let's Make a Deal." A contestant is presented with three doors. Behind one door is a brand new car, and behind the two other doors are goats. The contestant, hoping to win the car, is asked to choose a door. Monty Hall then opens

one of the two doors that the contestant did not choose; and each time, he reveals a goat. After this, he offers the contestant the option to change his mind and switch doors.



Organize a math game show. This will be watched by invited teachers from selected class sections or grade levels. One student in your class will be assigned to play the role of Monty Hall. Two assistants will be assigned to make props (from cardboard or other materials) that consist of three doors numbered 1, 2, and 3, and a toy car and two toy goats that will be placed behind the doors. (*Note:* Only the host and the two assistants should know where the toy car is placed.) Each of the rest of the students in your class will be a contestant in the game show. The host will ask each contestant, after revealing the door with the goat, if he or she wishes to change his or her mind to switch to one of the two doors unopened. All contestants who say they will switch doors will be in group 1. All contestants who say they will stick with their original choice will be in group 2. In each group, the host will count how many of the contestants won the car and announce which group had a higher number of car winners.

The highlight of the activity is to prove that the chance of winning the car increases when a contestant switches doors. After revealing where the toy car is placed, group 1 contestants will present a video presentation of the proof that the probability of winning the car if the contestant switches doors is not $\frac{1}{2}$ but $\frac{2}{3}$ using Bayes's formula. The proof must be correct, neat, detailed, and organized. Afterwards, group 2 contestants will present a video presentation of the proof that there will be a 50-50 chance of winning the car (again, using Bayes's formula) if the assumption of group 1 is changed; that is, instead of the host knowing where the car is, he has forgotten where the car is, puts on a straight face, opens a door at random, and luckily, the door reveals a goat. The proof must be correct, neat, detailed, and organized as well.

Evaluation forms will be handed out to the teachers who watched the game show to rate the activity and give comments and suggestions for improvement, if any.

Chapter Exercises

- For each of the following experiments, determine the number of outcomes in the sample space. Then find the probability of the given event.
 - Three fair coins are tossed; a tail comes up on the second coin.
 - Three fair dice (one red, one blue, one green) are tossed. Let A be the event where the sum of the three faces showing equals 5.
 - Three people are chosen at random from six equally capable candidates to form a committee; neither James nor Nadine (two of the six) is chosen.
- Let A and B be any two events. Use Venn diagrams to determine whether the event $A' \cup B'$ is equal to $(A' \cap B) \cup (A \cap B')$.
- How many four-digit numbers can be formed from the numbers 0, 1, 2, 3, 4, and 5 if
 - digits may be repeated?
 - all the digits must be distinct?
 - digits may be repeated, and the number must be an even number less than 2000?
- The officers of the statistics club of Bersales High School consist of four boys and five girls. They are to be seated in a row for an official photograph. In how many ways can this be done if
 - there are no restrictions?
 - Pedro, Juana, and Maria, being the top 3 officers of the club, must be seated together?
 - the boys and the girls must alternate?
 - Joanna and Edith insist on sitting two seats apart?
- In the game of *bridge*, each player is dealt 13 cards from a standard deck of 52 playing cards.
 - How many possible bridge hands are there?
 - How many bridge hands are there containing 4 hearts, 4 diamonds, 4 clubs, and 1 spade?
 - Define the events “the hand contains four aces” and “the hand contains four kings.” Find $P(A \cup B)$.
- An urn contains 5 red balls, 5 white balls, and 5 blue balls. Two balls are drawn simultaneously from the urn. What is the probability that they are of the same color?

7. Melissa rolls a die three times. Find the probability that
 - a. the outcomes of the second and third rolls are larger than the outcome of the first roll.
 - b. a pair has occurred, that is, exactly two rolls produced the same outcome.
8. A standard deck of 52 cards is shuffled and then placed as a pile on a desk. What is the probability that
 - a. the top twelve cards on the pile are face cards?
 - b. the bottom thirteen cards on the pile are all spades?
9. Suppose that 9 people, including you and your arch business rival, line up for a group picture. What is the probability that the photographer will arrange the line in such a way that there is at least one person between you and your arch business rival?
10. The following cross tabulation is for a symphony orchestra with 100 student members. Each member is categorized by the type of instrument that he or she plays: woodwind, brass, string, or percussion, and the school where he or she studies.

	Instrument Category				
School	Woodwind	Brass	String	Percussion	Total
Jacinto High School	11	7	9	10	37
Pagdilao Regional High School	9	6	12	6	33
Hilaga Rural High School	10	7	9	4	30
Total	30	20	30	20	100

- If a member of this orchestra is selected at random, find the probability that he or she
- a. is a student from Jacinto High School.
 - b. plays a brass instrument.
 - c. does not play a string instrument.
 - d. either is from Pagdilao Regional High School or plays a woodwind instrument.
 - e. plays a percussion instrument, given that the member is a student from Hilaga Rural High School.
11. In item (10), determine if the events “school” and “instrument category” are independent.
 12. If $P(A) = 0.32$ and $P(B) = 0.58$, find
 - a. $P(A \cup B)$ if A and B are mutually exclusive.
 - b. $P(A \cup B)$ if A and B are independent.
 - c. $P(A | B)$ if A and B are independent.

13. An experiment involves rolling a green die and a red die. Suppose the outcomes are written in the form (g, r) where g is the outcome on the green die and r is the outcome on the red die. Let A be the event where the green die shows a “6.” Let B be the event where the sum of the spots $g + r$ on the two dice is 8. Let C be the event where the sum of the spots $g + r$ on the two dice is 7.
- Find the elements of event A . What is $P(A)$?
 - Find the elements of event B . What is $P(B)$?
 - Find the elements of event C . What is $P(C)$?
 - Find the elements of $A \cap B$ and $A \cap C$. Then find $P(A \cap B)$ and $P(A \cap C)$.
 - Are events A and B independent? Are events A and C independent?
14. An oil prospector claims that there is a 25% chance of oil beneath a piece of land. He could obtain more information from a seismic test, however, such a test is 92% reliable in providing a favorable forecast when there is actually oil, but only 84% reliable in providing an unfavorable forecast when a site is actually dry. Suppose that the oil prospector has decided to order the seismic test and the result shows a favorable outcome. What is the probability that he will strike oil when he drills?
15. Urn 1 has 5 white balls and 6 black balls. Urn 2 has 4 white balls and 7 black balls. A ball is drawn from urn 1 and placed unseen in urn 2. Then a ball is drawn from urn 2. It is black. What is the probability that a black ball was transferred from urn 1 to urn 2?
16. A man’s pocket contains a fair coin and a biased coin that comes up head 75% of the time. He chooses a coin at random and flips it twice. Both tosses showed heads. What is the probability that he chose the fair coin?
17. Suppose it is known that among the people in a certain town, 1% have a certain type of cancer. Since it is expensive to undergo a medical procedure (but it is more accurate), a company has come up with a less expensive blood test that detects the cancer, but it is not foolproof. Among people who do have the condition, this test correctly diagnoses the person 85% of the time, whereas for those who do not have the condition, the test correctly diagnoses the person 90% of the time. Suppose that a person is randomly selected from the town and his or her test result is negative. What is the probability that he or she does not have the disease? (This probability is called the **negative predictive value** of the test.)
18. A standard poker deck of cards is shuffled. One card is removed from the deck, and then the deck is shuffled again. A card is now drawn from the reduced deck and it is found to be a Queen. What is the probability that the removed card was not a Queen?

Chapter 2

Random Variables and Probability Distributions



Every day, several babies are born in hospitals. The number of babies born varies from day to day. What determines this quantity is something that cannot be controlled and can be attributed to random elements or factors. Other quantities of interest could be the number of babies that weigh less than 2 kilograms or had jaundice at birth among 20 randomly selected babies. These quantities and several others that are similar in nature lead to the topic of *random variables* and their *probability distributions*, which is the focus of this chapter. This chapter also discusses the concept of *expected values* of a random variable and various applications of the concept.

Lesson 1

Concept of a Random Variable

Learning Outcomes

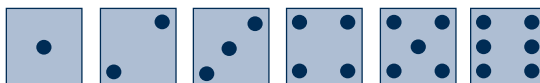
- At the end of this lesson, you should be able to
 - define and illustrate a random variable; and
 - determine the values of a random variable.

Introduction

Throughout previous introductory courses in probability, the term *random statistical experiment*, or simply *random experiment*, has been used to refer to a procedure that generates well-defined outcomes where the outcomes cannot be predicted. However, the list of possible outcomes in a random experiment is known, and these elementary outcomes constitute the random experiment's *sample space*, denoted by S .

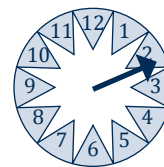
Example 1

A roll of a single six-sided die is a random experiment. The outcome of the roll cannot be predicted in advance. If our interest is the number of spots on the topmost face of the die when it is rolled, then the sample space S is:



Example 2

A wheel is divided evenly into 12 sectors, each numbered from 1 to 12. An arrow fastened at the center of the wheel is spun once. The outcome of the spin cannot be predicted in advance. Since we are only assuming that the arrow will land on a numbered sector when it is spun, this is a random experiment and the sample space S is



$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.

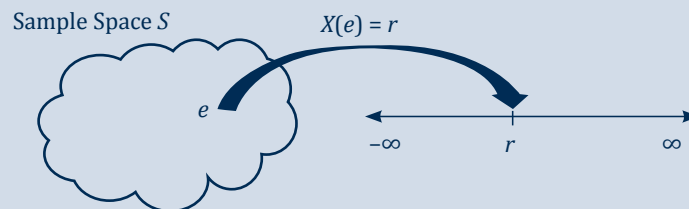
In the aforementioned examples, we enumerated the elements of the sample space. In many situations, we may not be concerned with the details of the sample space, but instead on a defined quantity or a certain numerical description of the outcome of the random experiment. These quantities do not have fixed values and their values vary from one random experiment to another. To illustrate, the number of accidents at a given road intersection in a month is not a fixed quantity. Its values depend on many random factors that change from month to month. We present the following examples, wherein the values of the defined quantity are not fixed:

1. the sum of the spots when a pair of dice is rolled
2. the number of customers that enter a department store in a minute
3. the amount of rainfall in Marikina in a year

In the study of probability, the quantities mentioned in the aforementioned examples, as well as similar quantities defined in random experiments, are called *random variables*. The term *random* is used to emphasize that the value of the quantity arises from a random experiment, the outcome of which cannot be predicted with certainty. The term *variable* refers to the value of the quantity that is not fixed.

Definition 1

A **random variable** is a function that assigns a unique real number to each element in the sample space. In other words, a **random variable** is a real-valued function whose domain is the sample space S . This is illustrated in the figure below, where the random variable X maps the sample point e from the sample space S to the real number line.



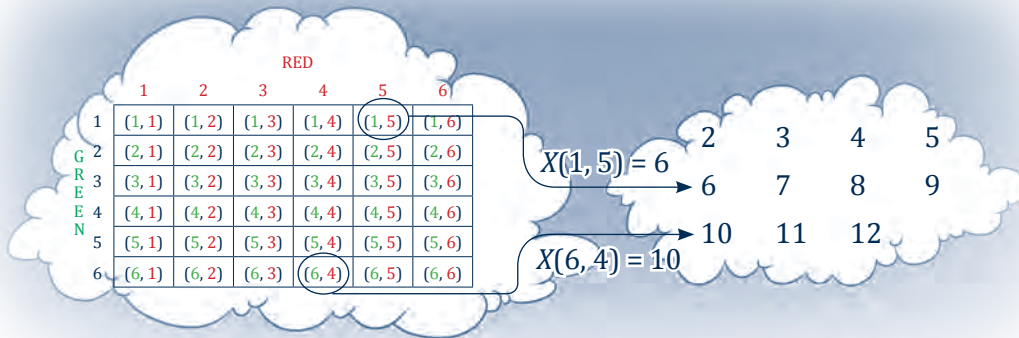
A random variable is usually denoted by upper case letters, oftentimes X or Y . A particular value of a random variable X is usually denoted by a lower case letter, say x or k .

Example 3

In rolling a pair of fair dice, say a green die and a red die, the sample space would consist of 36 outcomes that are all equally likely to occur. If we write the outcomes in the form (g, r) , where g is the outcome on the green die and r is the outcome on the red die, then we enumerate the 36 outcomes as follows:

		Red Die					
		1	2	3	4	5	6
Green Die	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)

Suppose that we define the random variable X as the sum of the spots on the two dice. Then X can only assume the values 2, 3, 4, 5, 6, ..., 11, 12. This is illustrated below.



Example 4

An absent-minded security guard at a bank randomly returns 3 umbrellas to 3 clients named Alice, Belle, and Carla, who have previously checked them in.



Alice's Umbrella



Belle's Umbrella



Carla's Umbrella

If Alice, Belle, and Carla, *in that particular order*, each receives one of the 3 umbrellas, then the sample space consists of the 6 possible orders that the security guard may return the umbrellas. If we denote Alice's umbrella as A, Belle's umbrella as B, and Carla's umbrella as C, then the 6 elements of the sample space S are {ABC, ACB, BAC, BCA, CAB, CBA}. Suppose that we define the random variable X as the number of clients who received their own umbrellas. Then X can only assume the values 0, 1, and 3.

Elements of S		x	remarks
ABC	$X(ABC) = 3$	3	All 3 received their own umbrellas
ACB		1	Only Alice received her own umbrella
BAC		1	Only Carla received her own umbrella
BCA		0	None received her own umbrella
CAB	$X(CBA) = 1$	0	None received her own umbrella
CBA		1	Only Belle received her own umbrella

Example 5

A card is drawn with replacement from a standard poker deck of 52 cards until an ace is drawn. This means that a card that has been drawn is placed back to the deck and the deck is reshuffled before the next card is picked. The card-picking is to be done until any one of the following four ace cards is picked.



Let a be the event where an ace card is drawn, and let n be the event where a non-ace card is drawn. The sample space would consist of an unending sequence of elements $S = \{a, na, nna, nnna, \dots\}$, and the random variable X defined as the number of draws needed to pick an ace would assume the values of positive integers 1, 2, 3,

Elements of S		x
a	$X(a) = 1$	1
na		2
nna	$X(nna) = 3$	3
$nnna$		4
\vdots		\vdots

Let's Practice

I. Write the letter that corresponds to your answer. Write X if your answer is not among the choices.

- _____ 1. Which statement is true about a random variable in a random experiment?
- It computes the probabilities of the outcomes.
 - It measures the outcomes.
 - It records the frequency of the outcomes.
 - It represents all possible outcomes.
- _____ 2. Which statement is false?
- The value of a random variable may be negative.
 - The number of values of a random variable may be finite or countable.
 - A random variable may take on any value in some interval of real numbers.
 - A random variable is a function that maps the set of real numbers to the set of elements in a sample space.
- _____ 3. You are making a study on families with three children. Your interest is the number of females among the three children. Which statement is true?
- The number of females among the three children in a family is not a random variable.
 - The number of females among the three children in a family is a random variable and its values are 1, 2, and 3.
 - The number of females among the three children in a family is a random variable and its values are 0, 1, 2, and 3.
 - The number of females among the three children in a family is a random variable and its values are 0, 1, 2, 3, and 4, since the mother is part of the family.

II. Determine the values of the random variables defined in the given scenarios.

- When a faculty member at a university is invited to the reception of commencement exercises, he or she may choose not to attend, to attend alone, or to attend with a guest. The random variable X is the number of event attendees for a faculty member.
- A green die and a red die are rolled. The random variable X is the absolute value of difference between the number of spots on the green die and the number of spots on the red die.
- A green die and a red die are rolled. The random variable X is the number of rolls required until a sum of seven spots is attained for the first time.
- A box contains one ₱20 bill, one ₱50 bill, one ₱100 bill, one ₱200 bill, one ₱500 bill, and one ₱1,000 bill. Two bills are drawn in succession and without replacement from the box. The random variable T is the total monetary value of the two bills.

Lesson 2

Discrete Random Variable and Its Probability Mass Function

Learning Outcomes

- At the end of this lesson, you should be able to
 - demonstrate a clear understanding of a discrete random variable by being able to illustrate it and find its values; and
 - determine the probability mass function of a discrete random variable and construct its histogram.

Introduction

We have defined a random variable as a real-valued function whose domain is the sample space S . Any random variable maps the elements of S to the real number line. If the range of the mapping consists of finite or countably infinite number of possible values, then the random variable is said to be *discrete*.

Definition 1

If the sample space of a random experiment consists of a finite number of elements or has an unending sequence with as many elements as there are counting numbers, then it is called a **discrete sample space**. A random variable defined over a discrete sample space is called a **discrete random variable**.

The values of a discrete random variable arise from a counting process, thus assuming clearly separated values. The following are examples of discrete random variables:

- the sum of the spots when a pair of dice is rolled
- the number of correct umbrella-owner matches when 3 umbrellas are randomly returned
- the number of card draws needed to draw an ace card
- the number of typhoons that enter the Philippine Area of Responsibility (PAR) in a year
- the number of customers that enter a department store in a minute

Definition 2

For a discrete random variable X , the **probability mass function** (or **PMF**) of X , denoted by $p(k)$, is defined as a function $p(k) = P(X = k)$. A PMF of a discrete random variable may be a table or a formula that lists down all the possible values that the random variable can assume, along with the corresponding probabilities of those values.

Thus, a tabular PMF of the discrete random variable X with associated values x_1, x_2, \dots will look like this:

k	x_1	x_2	\dots
$P(X = k)$	$p(x_1)$	$p(x_2)$	\dots

A PMF in the form of a formula will look like this:

$$P(X = k) = \begin{cases} p(x_k) & \text{if } k = x_1, x_2, \dots \\ 0 & \text{otherwise} \end{cases}$$

The probability mass function $p(x)$ is nonnegative for at most a countable number of values of x . If X can take the values x_1, x_2, \dots , then $p(x_k) \geq 0$, $k = 1, 2, \dots$, and $p(x) = 0$ for all other values of x . This is called the *nonnegativity property* of a PMF. In addition, the sum of all the probabilities in a PMF for all possible values of x is equal to 1. This is called the *norming property* of a PMF.

Points to Remember

If X is a discrete random variable with probability mass function $p(x)$, then

1. $p(x_k) \geq 0$ for $k = 1, 2, \dots$ and $p(x) = 0$ for all other values of x .

(Nonnegativity Property)

2. $\sum_{k=1}^{\infty} p(x_k) = 1$ **(Norming Property)**

3. the values x_1, x_2, \dots of X for which $p(x) > 0$ are called the **mass points** of X .

4. the probabilities involving events associated with any value or values of X may be computed by taking the sum of the probabilities of the mass points. Therefore,

$$\text{a. } P(X \leq a) = \sum_{\{k: x_k \leq a\}} p(x_k) \qquad \text{c. } P(X \geq b) = \sum_{\{k: x_k \geq b\}} p(x_k)$$

$$\text{b. } P(a \leq X \leq b) = \sum_{\{k: a \leq x_k \leq b\}} p(x_k)$$

Example 1

In example 3 of lesson 1, the random variable X is the sum of the spots when a pair of dice is rolled. X assumes the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 with the following probabilities, where the 36 outcomes are given equal weights of $\frac{1}{36}$ of occurrence:

$$P(X = 2) = P[(1,1)] = \frac{1}{36}$$

$$P(X = 3) = P[(1,2), (2,1)] = \frac{2}{36}$$

$$P(X = 4) = P[(1,3), (2,2), (3,1)] = \frac{3}{36}$$

$$P(X = 5) = P[(1,4), (2,3), (3,2), (4,1)] = \frac{4}{36}$$

$$P(X = 6) = P[(1,5), (2,4), (3,3), (4,2), (5,1)] = \frac{5}{36}$$

$$P(X = 7) = P[(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)] = \frac{6}{36}$$

⋮

$$P(X = 12) = P[(6, 6)] = \frac{1}{36}.$$

The complete PMF of X in tabular form is

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

It can be seen that $p(x)$ is zero for all values that do not belong to the set

$\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$ and $\sum_{k=2}^{12} P(X = k) = 1$. The PMF of X may also be written in the

form of a formula as $P(X = k) = \frac{6 - |7 - k|}{36}$ for $k = 2, 3, \dots, 12$.

Example 2

Three fair coins, a one-peso coin, a five-peso coin, and a ten-peso coin, are tossed simultaneously.



Define the random variable X as the number of coins that show head. If an outcome of head is denoted by h and an outcome of tail is denoted by t , we define an element of the sample space S as a triplet (o, f, n) where o is the outcome on the one-peso coin, f is the outcome on the five-peso coin, and n is the outcome on the ten-peso coin. Then the sample space is

$$S = \{(h, h, h), (h, h, t), (h, t, h), (t, h, h), (h, t, t), (t, h, t), (t, t, h), (t, t, t)\}$$

As such, $X(h, h, h) = 3$, $X(h, h, t) = 2$, and so on. It can be seen that X can take the values 0, 1, 2, and 3. Let us consider the case of $P(X = 0)$ or the probability that none of the 3 coins shows a head. The rule of statistical independence applies because the outcome of a coin does not influence the outcome of the other coins. Then

$$\begin{aligned} P(X = 0) &= P[(t, t, t)] \\ &= P[(o=t) \cap (f=t) \cap (n=t)] \\ &= P(o=t) \times P(f=t) \times P(n=t) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}. \end{aligned}$$

Similar computations apply for $P(X = 1)$, $P(X = 2)$, and $P(X = 3)$.

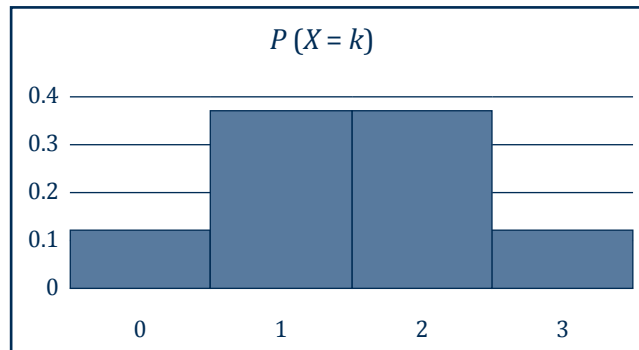
$$\begin{aligned} P(X = 1) &= P[(t, t, h), (t, h, t), (h, t, t)] = \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) = \frac{3}{8} \\ P(X = 2) &= P[(t, h, h), (h, h, t), (h, t, h)] = \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) + \left(\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}\right) = \frac{3}{8} \\ P(X = 3) &= P[(h, h, h)] = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8} \end{aligned}$$

The PMF of X in tabular form is

outcome	(t, t, t)	(h, t, t) (t, h, t) (t, t, h)	(h, h, t) (t, h, h) (h, t, h)	(h, h, h)
k	0	1	2	3
$P(X = k)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

and the PMF of X in formula form is $P(X = k) = \frac{{}_3C_k}{8} = \frac{3!}{k!(3-k)!}$ for $k = 0, 1, 2, 3$. This follows from the counting technique that there are ${}_3C_k$ ways to arrange 3 objects of which k are of the first kind (heads or h) and the remaining $3 - k$ are of the second kind (tails or t).

The graph of the probability mass function of the random variable X for the number of heads when 3 coins are tossed simultaneously is shown in the figure below. This graph is called a **probability histogram**.



Example 3

In example 5 of lesson 1, the random variable X was defined as the number of card draws with replacement from a standard deck needed to draw an ace. The sample space consists of an unending sequence of elements $S = \{a, na, nna, nnna, \dots\}$. Since there are 4 aces in a standard deck, the probability of obtaining an ace is $\frac{4}{52}$ or $\frac{1}{13}$, and that of a non-ace card is $\frac{12}{13}$. The values of X are 1, 2, 3,

Let us look at $P(X = 1)$, the probability that an ace is observed in just one card draw. Then $P(X = 1) = P(a) = \frac{1}{13}$. Now, if 2 draws are needed to observe an ace, it means that a non-ace card occurred on the first draw, followed by an ace. Since the outcome of a particular card draw does not influence the outcome of other draws (draws are made with replacement, and the deck is always intact before the next card is drawn), the rule of statistical independence can be used.

$$P(X = 2) = P(na) = P(n \cap a) = P(n) \times P(a) = \frac{12}{13} \times \frac{1}{13} = \frac{12}{169}$$

Similar computations follow, as

$$P(X = 3) = P(nna) = P(n \cap n \cap a) = P(n) \times P(n) \times P(a) = \left(\frac{12}{13}\right)^2 \times \frac{1}{13} = \frac{144}{2,197}$$

$$P(X = 4) = P(nnna) = P(n \cap n \cap n \cap a) = P(n) \times P(n) \times P(n) \times P(a) = \left(\frac{12}{13}\right)^3 \times \frac{1}{13} = \frac{1,728}{28,561}$$

⋮

and so on. Identifying a pattern in the computation, we see that there are $k - 1$ draws of non-ace cards that occur before an ace card is drawn. The PMF of X in formula form is

$$P(X = k) = \left(\frac{12}{13}\right)^{k-1} \left(\frac{1}{13}\right) \text{ for } k = 1, 2, 3, \dots$$

Clearly, this formula satisfies the properties of a PMF, namely, $p(x_k) \geq 0$, $k = 1, 2, \dots$ and the sum of all the probabilities in this PMF is equal to 1.

$$\begin{aligned} \sum_{k=1}^{\infty} p(x_k) &= \sum_{k=1}^{\infty} \left(\frac{12}{13}\right)^{k-1} \left(\frac{1}{13}\right) \\ &= \left(\frac{1}{13}\right) \sum_{k=1}^{\infty} \left(\frac{12}{13}\right)^{k-1} \\ &= \left(\frac{1}{13}\right) \left[1 + \frac{12}{13} + \left(\frac{12}{13}\right)^2 + \dots \right] \\ &= \left(\frac{1}{13}\right) \left[\frac{1}{1 - \frac{12}{13}} \right] = 1 \end{aligned}$$

The last step follows from the sum of an infinite geometric progression,

$$S = \sum_{i=0}^{\infty} r^i = \frac{1}{1-r}, \text{ provided that the common ratio } r \text{ is such that } |r| < 1.$$

Example 4

The following probability mass function is for the random variable X , in this case the number of chocolate chips in a chocolate chip cookie baked in large commercial quantities.



k	4	5	6	7	8
$P(X = k)$	0.30	0.45	0.15	0.08	0.02

1. What is the probability that a randomly selected chocolate chip cookie has at least 5 chocolate chips?
2. What is the probability that a randomly selected chocolate chip cookie has fewer than 6 chocolate chips?

Solution:

1. The probability that a randomly selected chocolate chip cookie has at least 5 chocolate chips is

$$\begin{aligned} P(X \geq 5) &= \sum_{k=5}^8 p(x_k) \\ &= P(X=5) + P(X=6) + P(X=7) + P(X=8) \\ &= 0.45 + 0.15 + 0.08 + 0.02 \\ &= 0.70 \end{aligned}$$

Therefore, the probability that a randomly selected chocolate chip cookie has at least 5 chocolate chips is 70%.

2. The probability that a randomly selected chocolate chip cookie has fewer than 6 chocolate chips is

$$\begin{aligned} P(X < 6) &= \sum_{k=4}^5 p(x_k) \\ &= P(X=4) + P(X=5) \\ &= 0.30 + 0.45 \\ &= 0.75 \end{aligned}$$

Therefore, the probability that a randomly selected chocolate chip cookie has fewer than 6 chocolate chips is 75%.

Let's Practice

I. Write the letter that corresponds to your answer. Write X if your answer is not among the choices.

- _____ 1. Which is *not* a discrete random variable?
- a. a student's score in a true-or-false quiz consisting of ten questions
 - b. the height of a randomly selected student in a group
 - c. the number of grammatical errors in a 300-word essay composition of a senior high school student
 - d. the number of tattoos that a randomly selected person has
- _____ 2. Given below is the probability mass function for the random variable X defined as the number of courses for which graduate school students at a university are enrolled in a semester.

x	1	2	3	4
$P(X = x)$	0.3	0.4	c	0.1

What is the value of c ?

- a. 0
 - b. 0.1
 - c. 0.2
 - d. 0.3
- _____ 3. Which is a property of the probabilities in the probability mass function for a discrete random variable X ?
- a. The probabilities for all possible values of X must be the same.
 - b. The probabilities for the values of X may be negative or may exceed 1, as long as the probabilities for all possible values sum up to 1.00.
 - c. The probabilities for the values of X are nonnegative but at most 1, and the probabilities for all possible values must sum up to 1.00.
 - d. The probabilities for all values of X must be less than 1.00.

II. Solve each problem.

1. Construct the probability mass functions of the random variables defined in the following exercises.
 - a. A green die and a red die are rolled. The random variable X is the absolute value of the difference between the number of spots of the green die and the number of spots on the red die.
 - b. A green die and a red die are rolled. The random variable X is the number of rolls required until a sum of seven spots is attained for the first time.
 - c. A box contains one ₱20 bill, one ₱50 bill, one ₱100 bill, one ₱200 bill, one ₱500 bill, and one ₱1,000 bill. Two bills are drawn in succession and without replacement from the box. The random variable T is the total monetary value of the two bills.
 - d. Four students are applying for a summer job in a fastfood restaurant. The result of the application is that a student may be hired or rejected. Assume that a student is equally likely to be hired or rejected for the job. The random variable X is the number of students among the four who are hired.
2. The following probability mass function is for the random variable X .

k	0	1	2	3	4	5
$P(X = k)$	0.05	0.05	0.15	0.20	0.25	0.30

- a. Find the probability that X takes a value less than 2.
 - b. Find the probability that X takes a value at least as large as 4.
3. Show that $P(X = k) = \frac{k+3}{15}$, for $k = -2, -1, 0, 1, 2$, and zero elsewhere, is a valid probability mass function for the discrete random variable X .
4. Find the value of the constant c so that the following is a probability mass function for the random variable X .

k	-3	-2	1	2	6
$P(X = k)$	c	$2c$	$3c$	$2c$	c

Lesson 3

Continuous Random Variable and Its Probability Density Function

Learning Outcomes

- At the end of this lesson, you should be able to
 - demonstrate a clear understanding of a continuous random variable and its probability density function (PDF) by being able to illustrate them and determine probabilities from a PDF; and
 - distinguish between a discrete random variable and a continuous random variable.

Introduction

There are random experiments where the outcomes are neither finite nor countably infinite. As an illustration, if we were to record the length of time that elapses between two successive customer arrivals at a department store, then the possible elapsed times, measured to any degree of accuracy, that make up the sample space is infinite in number. We could also consider the amount of unleaded gasoline (in liters) sold per day at a particular gasoline station. Again, if we assume that the variable amount is measured to any degree of accuracy, then this also shows that the sample space cannot just assume whole numbers, rendering the sample space to have an infinite number of possible amounts.

Definition 1

If the sample space of a random experiment consists of uncountably infinite number of outcomes, then it is called a **continuous sample space**. A random variable defined over a continuous sample space is called a **continuous random variable**.

The values of a continuous random variable arise from a measuring process. The following are examples of continuous random variables:

1. temperature, in degrees Celsius, of a patient admitted in a hospital
2. amount of rainfall in Marikina in a year
3. weight of a student, in kilograms
4. actual arrival time of the 5:00 am bus

Consider the first example. The random variable X is the temperature of a selected patient admitted in a hospital (measured in degrees Celsius). It is virtually impossible to find a patient whose temperature is *exactly* 38 degrees Celsius. When we say that the temperature of a patient is 38 degrees Celsius, we are actually giving an approximate measurement of the patient's temperature. This is due to the fact that the thermometer used to measure body temperature has a limitation in precision, and thus gives results usually rounded off to the nearest point on the scale. The patient's temperature could be shown as 38 degrees Celsius but could actually measure 37.9997685 degrees or 38.000234 degrees. For this reason, the probability $P(X = 38)$ is zero. This event almost never happens.

Points to Remember

Suppose that X is a continuous random variable.

1. The probability that X will assume a particular value k is practically zero; thus, $P(X = k) = 0$.
2. Unlike in the discrete case, a continuous random variable has no mass points. Thus, calculation of probabilities of the type $P(X = k)$ are not done. Instead, probabilities of X taking on a value on an interval such as $P(a < X < b)$, $P(X < a)$, or $P(X > b)$ is calculated.
3. Unlike in the discrete case, when computing the probability that X takes a value on the interval (a, b) , it does not matter whether an endpoint of the interval is included or not. Therefore,

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

Definition 2

For a continuous random variable X , the **probability density function** (or **PDF**) of X , denoted by $f(x)$, is a real-valued function defined on a continuous sample space S with an uncountable number of outcomes. A PDF may not be expressed in the form of a table, unlike a PMF of a discrete random variable. However, $f(x)$ is expressed in the form of a formula, and it satisfies the following properties:

1. $f(x) \geq 0$ for all x in S .
2. The total area under the whole curve of $f(x)$ above the x -axis is always equal to 1.
3. The probability that X assumes a value within an interval (a, b) is the area bounded by the curve of $f(x)$, the x -axis, and the lines $x = a$ and $x = b$. This is $P(a < X < b)$.

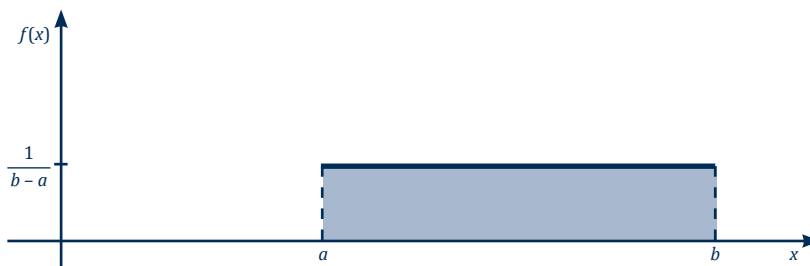
As an illustration, consider the simplest probability density function f for a continuous random variable X defined over an uncountably infinite sample space, which is the *uniform probability density function*, as shown below:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{elsewhere} \end{cases}$$

The graph of the uniform PDF is presented in the figure below. Notice that $f(x)$ satisfies the properties of a continuous probability density function because

1. $f(x) \geq 0$ for all x in the real number line; and
2. the total area bounded by the graph above the x -axis over the interval $[a, b]$ is the area of a rectangle. We use the formula for the area of a rectangle which is $Area = Base \times Height$.

In this case, $Area = (b-a) \times \left(\frac{1}{b-a} \right) = 1$.

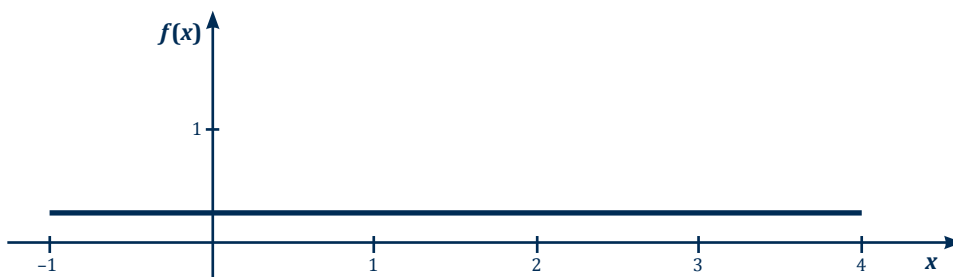


Example 1

To illustrate the uniform PDF, suppose that the random variable X is uniformly distributed between $x = -1$ and $x = 4$. Then the PDF of X is given by:

$$f(x) = \begin{cases} \frac{1}{5} & \text{for } -1 \leq x \leq 4 \\ 0 & \text{elsewhere} \end{cases}$$

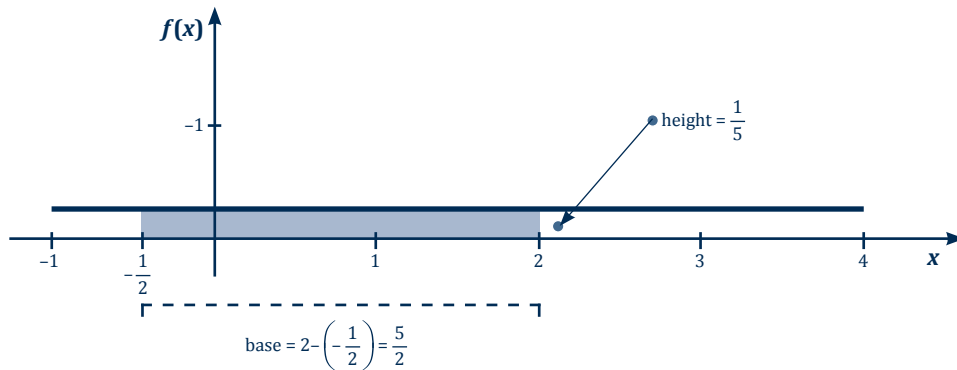
The graph of $f(x)$ is shown in the figure below.



To calculate the probability $P\left(-\frac{1}{2} < X < 2\right)$, we get the area of the rectangle formed by the graph of $f(x)$ above the x -axis, bounded by the graph of $x = -\frac{1}{2}$ and $x = 2$, which is

$$P\left(-\frac{1}{2} < X < 2\right) = \text{Base} \times \text{Height} = \left(2 - \left(-\frac{1}{2}\right)\right) \times \frac{1}{5} = \frac{5}{2} \times \frac{1}{5} = \frac{1}{2}$$

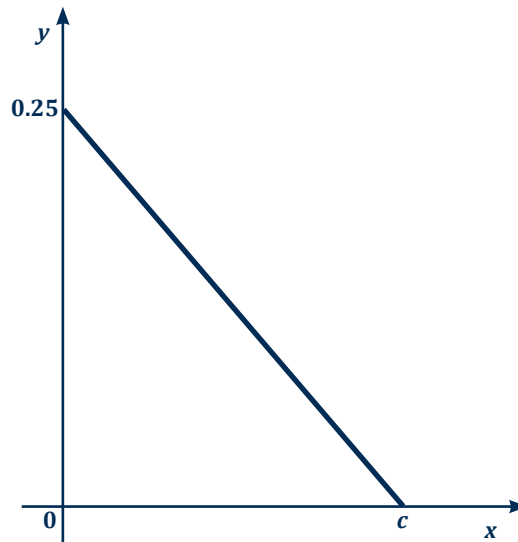
Illustrate the area corresponding to $P\left(-\frac{1}{2} < X < 2\right)$ to better picture the numerical calculation.



Example 2

Suppose that the graph below is for the PDF of a continuous random variable X .

1. What is the value of c ? Write down the PDF of X .
2. What is the probability that X exceeds 2?



Solution:

1. We know that the graph in example 1 is for a probability density function; hence, it must satisfy the properties of a PDF. Clearly, $f(x) \geq 0$ for all x in $[0, c]$. Since the total area under its curve above the x -axis is equal to 1, and the figure forms a right triangle, we can use the formula for the area of a triangle to find the value of c . The area of a triangle is $Area = \frac{1}{2} \times Base \times Height$.

$$1 = \frac{1}{2} \times (c - 0) \times \frac{1}{4}$$

$$1 = c \times \frac{1}{8}$$

$$c = 8$$

The value of c is 8. Thus, the PDF of X can be determined using the equation of a straight line. Since the line passes through the points $\left(0, \frac{1}{4}\right)$ and $(8, 0)$, the slope of

the line is $m = \frac{0 - \frac{1}{4}}{8 - 0} = -\frac{1}{32}$. Using the slope-intercept form of the straight line, we

determine the PDF of X to be $f(x) = \begin{cases} -\frac{1}{32}x + \frac{1}{4} & \text{for } 0 \leq x \leq 8 \\ 0 & \text{elsewhere} \end{cases}$

2. We first evaluate $f(x)$ at $x = 2$, to determine the point on the line whose abscissa is 2.

We see that $f(2) = -\frac{1}{32}(2) + \frac{1}{4} = \frac{3}{16}$. This means that the line passes through the point $\left(2, \frac{3}{16}\right)$. To find the probability $P(X > 2)$, we use the formula for the area of a

triangle again, this time with the height of the triangle as $\frac{3}{16}$ and the base as $8 - 2 = 6$.

$$P(X > 2) = \frac{1}{2} \times Base \times Height = \frac{1}{2} \times 6 \times \frac{3}{16} = \frac{9}{16}$$

Let's Practice

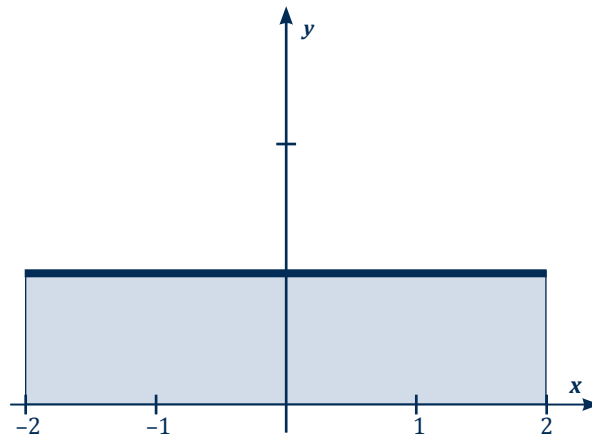
I. Write the letter that corresponds to your answer. Write X if your answer is not among the choices.

- _____ 1. Which is *not* a continuous random variable?
- a. the hand span of a grade 4 student
 - b. the liquid volume of soft drink in a can marked 300 ml
 - c. the number of persons in a group of twenty with type O blood
 - d. the time it takes a job applicant to complete the reading comprehension part of an exam
- _____ 2. Which is **not** a valid uniform probability density function for a random variable X ?
- a. $f(x) = \begin{cases} \frac{1}{60} & \text{for } 0 \leq x \leq 60 \\ 0 & \text{elsewhere} \end{cases}$
 - b. $f(x) = \begin{cases} \frac{1}{4} & \text{for } -2 \leq x \leq 1 \\ 0 & \text{elsewhere} \end{cases}$
 - c. $f(x) = \begin{cases} 2 & \text{for } 4 \leq x \leq 4.5 \\ 0 & \text{elsewhere} \end{cases}$
 - d. $f(x) = \begin{cases} \frac{1}{2\pi} & \text{for } 0 \leq x \leq 2\pi \\ 0 & \text{elsewhere} \end{cases}$
- _____ 3. Which of the following is true about a probability density function for a continuous random variable X ?
- a. the values of X are limited to some values only.
 - b. the values of X have to fall within a certain range of real numbers.
 - c. the probabilities for the values of X are nonnegative but at most 1, and the probabilities for all possible values must sum up to 1.00.
 - d. the frequency distribution of X is shown in the PDF.

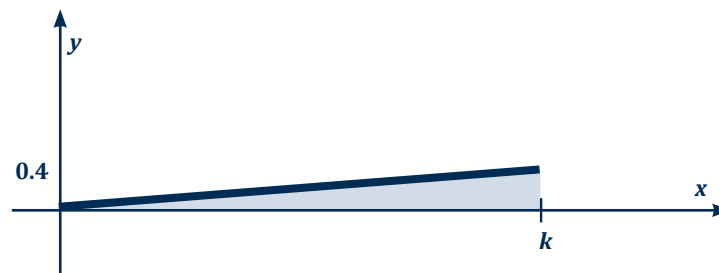
II. Solve each problem.

1. Write down the probability density function of the uniformly distributed random variable X defined on the interval $[0,5]$ and sketch the graph of the PDF of X .

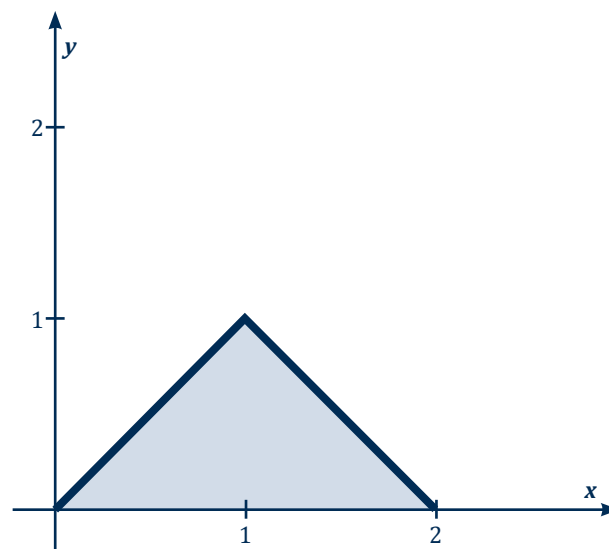
2. Write down the probability density function of the random variable X whose graph is as follows:



3. Determine the value of k so that the following graph is for the probability density function of the continuous random variable X . Then write down the PDF of X .



4. Determine the probability density function of the continuous random variable X whose graph appears here.



Lesson 4

Mean and Variance of a Discrete Random Variable

Learning Outcomes

- At the end of this lesson, you should be able to
 - demonstrate a clear understanding of the mean and variance of a discrete random variable by being able to illustrate them; and
 - compute and interpret the mean and variance of a discrete random variable.

Introduction

So far, we have studied discrete and continuous random variables and their probability distribution functions. The next step is to examine some features of probability distributions and quantities that would describe those features.

The first feature of a random variable's probability distribution that we want to examine is *central tendency*. Central tendency is a term that refers to the characteristic of centering at the average of the random variable. The most frequently used measure of central tendency in a random variable is the *expected value*.

Definition 1

For a discrete random variable X with probability mass function

x	x_1	x_2	\dots	x_n
$P(X = x)$	$p(x_1)$	$p(x_2)$	\dots	$p(x_n)$

the **expected value** of X is defined as

$$E(X) = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \dots + x_n \cdot p(x_n) = \sum_{k=1}^n x_k \cdot p(x_k)$$

The expected value of a random variable X is also referred to as the **mean of X** , denoted by μ_X .

Looking closely at the formula, we see that the expected value of X is a *weighted average* whereby the values of the random variable are weighted by their corresponding probabilities of occurrence. To visualize this, suppose we write the numbers x_1, x_2, \dots, x_n on n identical marbles and mix them in an urn. Then a marble is selected at random from the urn. The set of possible values of the random variable X ,

the numbers written on the marbles selected, is $\{x_1, x_2, \dots, x_n\}$. The marbles are equally likely to be selected, and the probability mass function of X is

x	x_1	x_2	\dots	x_n
$P(X = x)$	$\frac{1}{n}$	$\frac{1}{n}$	\dots	$\frac{1}{n}$

Thus, the expected value of X is

$$E(X) = x_1 \cdot \frac{1}{n} + x_2 \cdot \frac{1}{n} + \dots + x_n \cdot \frac{1}{n} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{k=1}^n x_k}{n}.$$

The expected value of X coincides with the mean of its values $\{x_1, x_2, \dots, x_n\}$. If the procedure of drawing a marble is done repeatedly with replacement, and each time we would record the number on the selected marble and then find the average of these numbers, the result obtained is the mean of the numbers $\{x_1, x_2, \dots, x_n\}$.

Points to Remember

For the discrete random variable X ,

1. its expected value doesn't have to be one of the values of X , since it is just an average value. For example, a university student's Grade Point Average (GPA) is not likely to be equal to any one of the valid grades in the university grading system; and
2. its expected value is the long-run average value of X . This means that if a random experiment is done repeatedly under the same conditions, then the expected value of a random variable X defined on the random experiment is the average value of X in the long run.

Example 1

In example 4 of lesson 1, we defined the random variable X as the number of correct umbrella-owner matches. The PMF of X in tabular form is

x	0	1	3
$P(X = x)$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{1}{6}$

The mean or expected value of X is

$$\mu_X = E(X) = \sum_{k=1}^n x_k \cdot p(x_k) = (0)\left(\frac{2}{6}\right) + (1)\left(\frac{3}{6}\right) + (3)\left(\frac{1}{6}\right) = 1$$

This means that if the security guard were to randomly return 3 umbrellas repeatedly (under the same condition that he remains absent-minded), and each time the number of clients who got their own umbrellas is observed, then the average of all these values, which is the expected number of correct umbrella-owner matches, is 1. From here, we see that 1 correct umbrella-owner match is the long-run average.

Example 2

Suppose that a certain die is distorted in such a way that the probability corresponding to a given number of spots is directly proportional to the number of spots. We define the random variable X as the number of spots in the topmost face when this distorted die is rolled. Then the six outcomes of the random experiment are not equally likely, and the PMF of X will be, for some constant c ,

x	1	2	3	4	5	6
$P(X = x)$	c	$2c$	$3c$	$4c$	$5c$	$6c$

To find the value of c , we use the norming property of a PMF, so that $\sum_{k=1}^{\infty} p(x_k) = 1$.

$$\begin{aligned} \sum_{k=1}^6 p(x_k) &= c + 2c + 3c + 4c + 5c + 6c = 1 \\ 21c &= 1 \\ c &= \frac{1}{21} \end{aligned}$$

The complete PMF of X is now

x	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{21}$	$\frac{2}{21}$	$\frac{3}{21}$	$\frac{4}{21}$	$\frac{5}{21}$	$\frac{6}{21}$

The mean or expected value of X is

$$\mu_x = E(X) = (1)\left(\frac{1}{21}\right) + (2)\left(\frac{2}{21}\right) + (3)\left(\frac{3}{21}\right) + (4)\left(\frac{4}{21}\right) + (5)\left(\frac{5}{21}\right) + (6)\left(\frac{6}{21}\right) = \frac{91}{21} \text{ or } \frac{13}{3}$$

This is an illustration where the expected value of a random variable doesn't have to be one of the values of the random variable. The possible values of X are 1, 2, 3, 4, 5, and 6. The mean of X is $\mu_x = \frac{13}{3} = 4.\bar{3}$, which is not one of the values of X .

The mean is a commonly used summary measure for a discrete probability distribution, but it does not describe the amount of spread or dispersion in the distribution. The second feature of a random variable's probability distribution that must be examined is *variability*. Variability is a term that refers to the characteristic of dispersing from the average of the random variable. A common measure of variability in a random variable is the *variance*.

Definition 2

For a discrete random variable X with probability mass function

x	x_1	x_2	\dots	x_n
$P(X = x)$	$p(x_1)$	$p(x_2)$	\dots	$p(x_n)$

the **variance** of x , denoted by σ_x^2 is defined as

$$\begin{aligned}\sigma_x^2 &= (x_1 - \mu_x)^2 \cdot p(x_1) + (x_2 - \mu_x)^2 \cdot p(x_2) + \dots + (x_n - \mu_x)^2 \cdot p(x_n) \\ &= \sum_{k=1}^n (x_k - \mu_x)^2 \cdot p(x_k)\end{aligned}$$

Definition 3

The positive square root of the variance is called the **standard deviation** of the random variable X , and is denoted by σ_X .

The standard deviation makes use of the same unit as the values of X , whereas the variance makes use of the square of the unit of the values of X . Thus, if the unit of the random variable X is in meters, then the unit of the variance σ_X^2 is in square meters (m^2), and the unit of the standard deviation σ_X is in meters (m).

Example 3

In connection with the random variable X , which is the number of correct umbrella-owner matches in example 4 of lesson 1, we found the mean μ_X , or expected value of X , to be 1. To find the variance of X , we write

x_k	$p(x_k)$	$(x_k - \mu_X)$	$(x_k - \mu_X)^2$	$(x_k - \mu_X)^2 \cdot p(x_k)$
0	$\frac{2}{6}$	$0 - 1 = -1$	1	$\frac{2}{6}$
1	$\frac{3}{6}$	$1 - 1 = 0$	0	0
3	$\frac{1}{6}$	$3 - 1 = 2$	4	$\frac{4}{6}$
Total	1			1

Therefore,

$$\sigma_X^2 = \sum_{k=1}^n (x_k - \mu_X)^2 \cdot p(x_k) = (0-1)^2 \cdot \left(\frac{2}{6}\right) + (1-1)^2 \cdot \left(\frac{3}{6}\right) + (3-1)^2 \cdot \left(\frac{1}{6}\right) = 1.$$

Since the variance of X is $\sigma_X^2 = 1$, the standard deviation of X is $\sigma_X = \sqrt{\sigma_X^2} = 1$.

Points to Remember

For the case of discrete random variables,

1. an alternative solution for finding the variance is

$$\sigma_X^2 = [x_1^2 \cdot p(x_1) + x_2^2 \cdot p(x_2) + \cdots + x_n^2 \cdot p(x_n)] - \mu_X^2 = \sum_{k=1}^n x_k^2 \cdot p(x_k) - \mu_X^2.$$

2. given two random variables X and Y , the expected value of the sum of X and Y is equal to the sum of the individual expected values, that is,
 $E(X + Y) = \mu_X + \mu_Y$.
3. given two random variables X and Y that are independent, the variance of the sum of X and Y is equal to the sum of the individual variances, that is,
 $Var(X + Y) = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$.

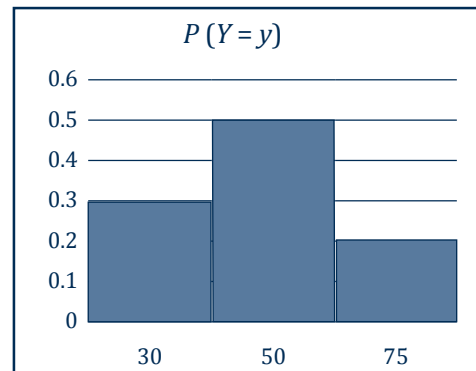
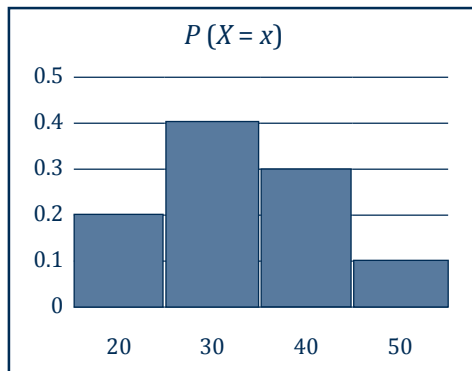
Example 4

Mr. Jacinto works as a part-time lecturer in two universities namely University X and University Y . Next academic year, he will get a raise in his hourly rate of X pesos from University X and a raise in his hourly rate of Y pesos from University Y . The increase in hourly rate in both universities depends on several factors, and thus can be regarded as random variables. Suppose that X and Y are independent random variables with probability mass functions, respectively:

x	20	30	40	50
$P(X = x)$	0.2	0.4	0.3	0.1

y	30	50	75
$P(Y = y)$	0.3	0.5	0.2

1. Sketch the graphs of the probability histograms of the PMF's of X and Y .



2. To find the expected total raise in hourly rate that Mr. Jacinto will get next academic year, first find the individual expected values, μ_X and μ_Y , that is,

$$\mu_X = E(X) = (20)(0.2) + (30)(0.4) + (40)(0.3) + (50)(0.1) = 33$$

$$\mu_Y = E(Y) = (30)(0.3) + (50)(0.5) + (75)(0.2) = 49$$

Therefore, Mr. Jacinto can expect a total raise in hourly rate from the two universities of $33 + 49 = 82$ pesos.

3. To find the standard deviation of the total raise in hourly rate that Mr. Jacinto will get next academic year, first find the individual variances, σ_X^2 and σ_Y^2 . Now demonstrate the use of the alternative method of finding variances,

$$\sigma_X^2 = \sum_{k=1}^n x_k^2 \cdot p(x_k) - \mu_X^2 = (20)^2(0.2) + (30)^2(0.4) + (40)^2(0.3) + (50)^2(0.1) - (33)^2 = 81$$

$$\sigma_Y^2 = \sum_{k=1}^n y_k^2 \cdot p(y_k) - \mu_Y^2 = (30)^2(0.3) + (50)^2(0.5) + (75)^2(0.2) - (49)^2 = 244$$

Since X and Y are independent, the variance of the total raise in hourly rate $\sigma_X^2 + \sigma_Y^2 = 81 + 244 = 325$. Therefore, the standard deviation of the total raise in hourly rate is the square root of 325, which is

$$\sigma_{X+Y} = \sqrt{325} = 5\sqrt{13} \text{ or } 18.03 \text{ pesos.}$$

Definition 4

The **coefficient of variation**, or CV , is a measure of relative dispersion that expresses the standard deviation as a percentage of the mean. The coefficient of variation of a random variable X is computed as

$$CV = \frac{\sigma_X}{\mu_X} \times 100.$$

Points to Remember

1. The coefficient of variation is unitless; thus it enables the comparison of the variability of two random variables even if they have different units of measurement.
2. The coefficient of variation is not applicable when the mean of a random variable is zero or negative.
3. When the standard deviation is large relative to the size of the mean, the coefficient of variation will be large, indicating high variability for a random variable and its probability distribution. When the standard deviation is small relative to the size of the mean, the coefficient of variation will be small, indicating low variability for a random variable and its probability distribution.

Example 5

In example 4 of this lesson, we compared the variability of the increase in hourly rate of Mr. Jacinto in University X and University Y using the coefficient of variation. The CV of the hourly rate in University X is

$$CV_X = \frac{\sigma_X}{\mu_X} = \frac{\sqrt{81}}{33} = \frac{9}{33} = 0.27273,$$

whereas the CV of the hourly rate in University Y is

$$CV_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\sqrt{244}}{49} = \frac{2\sqrt{61}}{49} = 0.31879.$$

Expressing these results in percentage, the CV of X is 27.273% and the CV of Y is 31.879%. Based on the computation, University X has less variability in the increase in hourly rate of a lecturer than University Y .

Let's Practice

I. Write the letter that corresponds to your answer. Write X if your answer is not among the choices.

- _____ 1. Which is true about the expected value for a random variable X with a given probability distribution function?
- It is the arithmetic mean of all the possible values of X .
 - It is the most common value of X in repeated trials of the experiment over which X is defined.
 - It is the value of X that has the highest probability of occurring.
 - It is the weighted mean over all the possible values of X .
- _____ 2. It was determined that the expected value of a discrete random variable X is 3.5. Which is the correct interpretation of this value?
- The value of the random variable X with the highest probability value is 3.5.
 - The most common value over many repeats of the experiment is 3.5.
 - Two of the values of X are 3 and 4, with equal probabilities of occurrence.
 - The weighted average value for X over many repeats of the experiment is 3.5.
- _____ 3. Which is true about the variance of a discrete random variable X ?
- It is the weighted average of the squares of the deviations from the mean of X .
 - It is the weighted average of the square roots of the deviations from the mean of X .
 - It is the sum of the squares of the deviations from the mean of X .
 - It is the sum of the square roots of the deviations from the mean of X .

II. Solve each problem.

1. Find the mean and variance of the random variable X whose PMF is given below:

k	0	1	2	3	4	5
$P(X = k)$	0.05	0.05	0.15	0.20	0.25	0.30

2. Find the mean and variance of the random variable X , or the number of chocolate chips in a cookie, as given in example 4 of lesson 2.
3. A box contains one ₱20 bill, one ₱50 bill, one ₱100 bill, one ₱200 bill, one ₱500 bill, and one ₱1,000 bill. Two bills are drawn in succession and without replacement from the box. The random variable T is the total monetary value of the two bills. Find the mean and variance of T .

4. An urn contains 10 chips numbered 1 to 10. Two chips are drawn at random without replacement from the urn. The random variable X is the larger number among the two chips. Determine the PMF of X , and compute its mean and variance.
5. Compute the coefficient of variation of the random variable in T , the total monetary value of the two bills in item 3, and the random variable X , the larger number among the two chips in item 4. Which has a larger coefficient of variation, T or X ?
6. Find the value of the constant c so that the following is a probability mass function for the random variable X , then find the mean and variance of X .

k	1	2	3	4	5	6
$P(X = k)$	$3c$	$3c$	$2c$	$2c$	c	c

7. A standard poker deck of cards is shuffled and placed as a single pile on a table. Find the probability mass function of the random variable X , which is the number of face cards that appear in the top five cards of the pile. Find the mean and variance of X .
8. A ball is drawn at random from an urn containing balls numbered 1 to 45. Define a random variable X that assumes a value of -2 if the number on the drawn ball is divisible by 3 but not by 5; a value of 2 if the number is divisible by 5 but not by 3; a value of 8 if the number is divisible by both 3 and 5; and a value of 15 if the number is neither divisible by 3 nor by 5. Find the probability mass function of X , and then calculate its mean and variance.
9. Mariel has a box of chocolates of which 8 are dark chocolates and 12 are milk chocolates. She chooses 4 of the chocolates at random. The random variable T is defined as the number of dark chocolates (denoted by X) minus the number of milk chocolates (denoted by Y) in the selection. Find the probability mass function of the random variable T , and find its mean and variance.

Lesson 5

Applications of Expected Value

Learning Outcomes

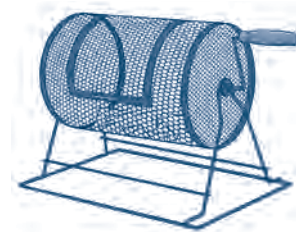
- At the end of this lesson, you should be able to
 - apply the concepts of mean and variance of a discrete probability distribution in solving real-life problems; and
 - solve similar and related problems to the various applications presented in this lesson

Introduction

The concept of expected value is commonly used in many fields of study such as statistics, finance, and economics, to name a few. In this section, we consider various applications of expected value. Originally, the concept of expected value emerged from games of chance. Today, expected values are used in common applications such as lotteries, gambling, insurance, and evaluation of investment options.

Example 1

At an office anniversary raffle, you hold one of 5,000 tickets for which the grand prize is a trip for 2 worth ₱100,000, the second prize is an LCD TV worth ₱50,000, and the third prize is a mobile phone worth ₱25,000. What is your expected winning?



Solution:

We define the random variable X to be the net winning.

If the raffle is repeated many times, you would win (for each of the three prizes) with probability

$$P(\text{winning for each draw}) = \frac{1}{5,000} = 0.0002$$

and lose with probability

$$P(\text{losing}) = 1 - 3(0.0002) = 0.9994$$

The probability mass function of X would be

x	₹100,000	₹50,000	₹25,000	0
$P(X = x)$	0.0002	0.0002	0.0002	0.9994

The expected value of X is

$$E(X) = (100,000)(0.0002) + (50,000)(0.0002) + (25,000)(0.0002) + (0)(0.9994) = 35.$$

Therefore, on the average, you would win ₹35. The expected value ₹35 is interpreted in the sense of an average. It was obtained by summing the products of each amount and the corresponding probability. We could look at it this way: collectively, the 5,000 tickets have a payout of ₹100,000 + ₹50,000 + ₹25,000 = ₹175,000. If we divide this amount by 5,000 (the total number of tickets), we have $\frac{₹175,000}{5000} = ₹35$ per ticket.

Definition 4

A **fair game** is one where the expected amount of the winnings is equal to the ante, the bet, or the amount paid out.

In a fair game, there is neither gain nor loss. Thus, a game with zero expectation is defined as fair.

Example 2

In a gambling game, two dice are tossed simultaneously. If the sum of the spots on the two dice is 5, 7, or 9, you win ₹900. If both dice show the same number of spots (a “double dice”), you win ₹1,800. For all other outcomes, you lose. How much should you bet to play if the game is fair?

Solution:

Let b be the amount of money to bet and the random variable X is the net winning. The event of getting a sum of 5, 7, or 9 can occur in these outcomes: $\{(1, 4), (2, 3), (3, 2), (4, 1), (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1), (3, 6), (6, 3), (4, 5), (5, 4)\}$. The “double dice” can occur in these outcomes: $\{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$. Thus, the probability mass function of X is:

x	₹900 – b	₹1,800 – b	– b
$P(X = x)$	$\frac{14}{36}$	$\frac{6}{36}$	$\frac{16}{36}$



To find the value of b , we equate the expected value of X to be zero, as a fair game is one with zero expectation.

$$E(X) = 0 = (900 - b)\left(\frac{14}{36}\right) + (1800 - b)\left(\frac{6}{36}\right) + (-b)\left(\frac{16}{36}\right)$$

$$0 = 350 - \frac{14b}{36} + 300 - \frac{6b}{36} - \frac{16b}{36}$$

$$0 = 650 - b$$

Thus, the value of the bet for the gambling game to be fair is $b = ₱650$.

Example 3

You purchase a fire insurance policy for your townhouse. The insurance company charges you ₱2,000 premium per annum. In the unfortunate incident that your townhouse catches fire, the insurance company pays you ₱1.5 million. The insurance company estimates that the probability of fire in your area is 0.0002. What is the insurance company's expected net gain from such fire insurance policy?

Solution:

Suppose X is the random variable for the company's net gain. Then X has two possible values: in case there's a fire, it incurs a net loss of ₱2,000 – ₱1,500,000 = –₱1,498,000, and in case there's no fire, a premium of ₱2,000 is received. The probability mass function of X is

x	Fire	No Fire
	–₱1,498,000	₱2,000
$P(X = x)$	0.0002	0.9998



The expected value of X is $E(X) = (-1,498,000)(0.0002) + (2,000)(0.9998) = 1,700$.

This means that for each such fire insurance policy of this type, the company expects to gain ₱1,700 per year. Of course, the company hopes that it will never have to pay out ₱1.5 million to you, inasmuch as you hope you never have to collect fire insurance money from them for your townhouse. In addition, if the company sells 100,000 such fire insurance policies, they expect to gross ₱200 million in premiums. Based on the probability of fire, which is 0.0002, the company can expect to pay out ₱1.5 million to 20 policy holders. Nevertheless, the company still stands to have a net profit of ₱170 million.

Example 4

Mr. Mina wants to buy a stock and keep it for one year in anticipation of capital gain.

He has narrowed down his choices to Company A and Company B. Both stocks currently sell for ₱200 per share and yield ₱10 dividends. The following probability mass functions for next year's price have been judgmentally assessed for each stock, where A is the random variable for the selling price of Company A stock and B is the random variable for the selling price of Company B stock.

a	₱150	₱200	₱250	₱300	₱350
$P(A = a)$	0.40	0.15	0.15	0.15	0.15

b	₱195	₱200	₱205	₱210
$P(B = b)$	0.10	0.20	0.45	0.25

We compute the expected values, variances, and coefficients of variation for the selling price of both companies, and make a comparison.

$$E(A) = (150)(0.4) + (200)(0.15) + (250)(0.15) + (300)(0.15) + (350)(0.15) = 225$$

$$E(B) = (195)(0.10) + (200)(0.20) + (205)(0.45) + (210)(0.25) = 204.25$$

Company A's expected selling price is ₱225 per share, whereas Company B's expected selling price is ₱204.25 per share. Company A has a higher expected selling price.

$$\sigma_A^2 = \left[(150)^2(0.4) + ((200)^2 + (250)^2 + (300)^2 + (350)^2)(0.15) \right] - (225)^2 = 5,625$$

$$\sigma_B^2 = \left[(195)^2(0.10) + (200)^2(0.20) + (205)^2(0.45) + (210)^2(0.25) \right] - (204.25)^2 = 20.6875$$

The selling price variance of Company A is 5,625, while that of Company B is 20.6875 per share. Company A has a higher variance.

$$CV_A = \frac{\sqrt{5,625}}{225} \times 100 = 33\frac{1}{3}\%$$

$$CV_B = \frac{\sqrt{20.6875}}{204.25} \times 100 = 2.2269\%$$

The coefficient of variation for Company A is much higher at 33.3333% compared to Company B. This illustrates the potential gain and the risk of the investment. We could regard Company A as a *speculative investment*, one with a higher risk of loss. The expected selling price of the speculative investment is ₱225, which is a higher return on investment. If Mr. Mina makes a large number of such investments, he would have a high chance of making a profit. But if he makes such investment, there is a 40% risk of the price going down, resulting in a loss.

Let's Practice

I. Write the letter that corresponds to your answer. Write X if your answer is not among the choices.

- _____ 1. What does the “expected net winning” from playing a game of chance mean?
- It is the amount you need to bet to break-even on many plays of the game.
 - It is the net amount you expect to win or lose in the long run on many plays of the game.
 - It is the net amount you expect to win or lose on a single play.
 - It is the net amount you should expect to win if you are lucky.

For items 2-4, consider the following situation:

You buy one ₱10 raffle ticket where the prize is a new mobile phone valued at ₱15,000. Two thousand tickets are sold.

- _____ 2. Which gives the values of the random variable X = net winning in the raffle?
- $X = 15,000$ (win); $X = 0$ (lose)
 - $X = 15,000$ (win); $X = -10$ (lose)
 - $X = 14,990$ (win); $X = 0$ (lose)
 - $X = 14,990$ (win); $X = -10$ (lose)

- _____ 3. Which gives the probability mass function for the random variable X = net winning in the raffle?

a.

x	15,000	0
$P(X = x)$	$\frac{1}{2,000}$	$\frac{1,999}{2,000}$

b.

x	14,990	-10
$P(X = x)$	$\frac{1}{2,000}$	$\frac{1,999}{2,000}$

c.

x	15,000	-10
$P(X = x)$	$\frac{1}{2,000}$	$\frac{1,999}{2,000}$

d.

x	14,990	0
$P(X = x)$	$\frac{1}{2,000}$	$\frac{1,999}{2,000}$

- _____ 4. Which gives the expected value of the random variable X = net winning in the raffle?
- ₱2.50
 - ₱7.50
 - ₱2.495
 - ₱7.495

II. Analyze and solve each problem.

- In a gambling game, a gambler pays ₱50 to play. A fair coin is tossed twice. If the two tosses result in the same outcomes, that is, either both heads or both tails, then he wins ₱100. But if the two tosses have different outcomes, he loses his bet. What is his expected net winning?
- In a gambling game, you draw a card from a standard deck of 52 cards. If the outcome is a numbered card from 6 to 10, you win ₱500. If the outcome is an ace or any of the face cards, you win ₱1,000. But for outcomes where the number on the card is 2 to 5, you lose. How much should you pay to play if the game is fair?

3. Maritess works for Acme Assurance Company and sold a life insurance policy worth ₱1 million to a 38-year old father for an annual premium of ₱24,000. Actuarial research shows that given the man's age, health background, and other factors, the probability of death next year is 0.0001. What is Acme Assurance Company's net gain for life insurance policies of this type?
4. In a Christmas party, lottery tickets are sold at ₱10 each, and the lone jackpot prize is ₱7,000. A total of 800 tickets was sold. Suppose you purchased 5 tickets. What is your expected net winning?
5. A gambler pays ₱50 to play a game where a ball is drawn at random from an urn containing balls numbered 1 to 20. If the number drawn is perfectly divisible by 2 but not by 5, he wins ₱60. If it is perfectly divisible by 5 but not by 2, he wins ₱70. If it is perfectly divisible by both 2 and 5, he wins ₱90. For all other outcomes, he loses his bet. What is the gambler's expected net winnings for playing this game?
6. A gambling game consists of two stages. The first stage is tossing a die, and if the die reveals 1, 2, 3, or 4, the gambler wins ₱100. For all other outcomes, he loses his bet. The second stage is flipping a coin. A head doubles the net amount obtained in the first stage, while a tail retains the net amount obtained in the first stage. How much should a gambler pay to play if the game is fair?

Lesson 6

Some Discrete Random Variables with Special Names

Learning Outcomes

- At the end of this lesson, you should be able to
 - familiarize yourself with some discrete random variables with special names such as the binomial, hypergeometric, Poisson, geometric, and negative binomial random variables and their various applications; and
 - solve similar and related problems presented in this lesson.

Introduction

In this section, we look at some special discrete random variables and their probability mass functions. There are many applications of these random variables with special names that frequently appear in probability and statistics, as well as in science and engineering.

The Bernoulli Random Variable

Definition 1

The **Bernoulli random variable** is a random variable whose only two values are 1, with probability p , and 0, with probability $1 - p$. The probability mass function of the Bernoulli random variable X is

x	1	0
$P(X = x)$	p	$1 - p$

The Bernoulli random variable, attributed to Swiss mathematician James Bernoulli, is associated with a *Bernoulli trial*, the outcomes of which are classified as a success ($X = 1$) or a failure ($X = 0$).

Example 1

Some examples of a single Bernoulli trial are as follows:

- a. Toss a fair coin once. If the occurrence of the head is the interest, then we may consider the head as “success” and the tail as “failure.” The PMF of X is:

x	1	0
$P(X = x)$	$\frac{1}{2}$	$\frac{1}{2}$

- b. Roll a single die once. If we are interested in the number 3, then we may consider the 3 as “success” and the non-3 outcomes as “failure”. The PMF of X is:

x	1	0
$P(X = x)$	$\frac{1}{6}$	$\frac{5}{6}$

- c. Inspect a single bottle randomly selected from a production line. If we are on the look-out for a defective bottle, then we may consider a defective bottle as “success” and a non-defective bottle as “failure”. Suppose it has been observed that the machine that makes the bottles produces defective bottles 5% of the time. Then the PMF of X is:

x	1	0
$P(X = x)$	0.05	0.95

The expected value of the Bernoulli random variable X , with probability of success p is

$$\mu_x = (1)(p) + (0)(1 - p) = p.$$

The variance of the Bernoulli random variable is

$$\sigma_x^2 = E[(X - \mu_x)^2] = (1 - p)^2(p) + (0 - p)^2(1 - p) = p(1 - p).$$

Now, suppose we consider an experiment where Bernoulli trials are performed n times and independently. In a sequence of n independent Bernoulli trials, the probability of success remains constant throughout all the n trials.

As an illustration, in example 1(b), a single roll of a fair die is a single Bernoulli trial. If the die is rolled four times, then $n = 4$, and the four rolls constitute a sequence of four independent Bernoulli trials, since the outcome of a roll does not affect the outcomes of the other rolls.

Moreover, the probability of success, $p = \frac{1}{6}$, remains constant throughout all the four rolls. Our

interest is now the number of times the 3 (labeled as “success”) occurs in the four rolls. The sample space S of this experiment is the set of different sequences of length 4 with y successes (T, for the 3) and $4 - y$ failures (N, for the non-3). In particular, the sample space is

$$S = \{TTTT, TTTN, TTNT, TNTT, NTTT, TTNN, TNTN, NNTT, \\ NTNT, TNNT, NTTN, NNNT, NNTN, NTNN, TNNN, NNNN\}$$

Clearly, the values of the random variable Y are 0, 1, 2, 3, and 4. The random variable Y is called a binomial random variable.

The Binomial Random Variable

Definition 2

The random variable Y that counts the number of successes in n independent Bernoulli trials, with probability of success p , is called a **binomial random variable**. The probability mass function of Y is $P(Y = y) = {}_nC_y p^y (1 - p)^{n-y}$ for $y = 0, 1, 2, \dots, n$ and zero elsewhere.

Here, ${}_nC_y = \frac{n!}{y!(n-y)!}$ is the number of ways to order a sequence of length n , partitioned as y successes and $n - y$ failures. The expected value of the binomial random variable Y is $\mu_Y = np$ and the variance of Y is $\sigma_Y^2 = np(1 - p)$.

Points to Remember

1. The binomial distribution applies when taking a sample with replacement from a small (finite) population.
2. The binomial distribution also applies when taking a sample, even without replacement, from a very large population.

In both of the aforementioned cases, the samples constitute independent Bernoulli trials, wherein the probability of “success” remains constant throughout all the trials.

Example 2

Consider example 1(b) of this lesson. If we let Y be the number of 3's in a sequence of four independent rolls of a fair die, where the occurrence of the 3 is "success" and the non-3 is "failure," the PMF of Y with probability $p = \frac{1}{6}$ is

$$P(Y = y) = {}_4C_y \left(\frac{1}{6}\right)^y \left(\frac{5}{6}\right)^{4-y} \text{ for } y = 0, 1, 2, 3, 4.$$

Thus,

$$\text{for } y = 0: P(Y = 0) = {}_4C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^{4-0} = \frac{625}{1296}$$

$$\text{for } y = 1: P(Y = 1) = {}_4C_1 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^{4-1} = \frac{500}{1296}$$

$$\text{for } y = 2: P(Y = 2) = {}_4C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^{4-2} = \frac{150}{1296}$$

$$\text{for } y = 3: P(Y = 3) = {}_4C_3 \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^{4-3} = \frac{20}{1296}$$

$$\text{for } y = 4: P(Y = 4) = {}_4C_4 \left(\frac{1}{6}\right)^4 \left(\frac{1}{6}\right)^{4-4} = \frac{1}{1296}$$

If we write the PMF in tabular form, we can clearly see that the sum of all the probabilities in the PMF is equal to 1.

y	0	1	2	3	4
$P(Y = y)$	$\frac{625}{1296}$	$\frac{500}{1296}$	$\frac{150}{1296}$	$\frac{20}{1296}$	$\frac{1}{1296}$

Example 3

In example 1(c) of this lesson, if we randomly select 3 bottles from the production line, without replacement, would the 3 bottles constitute a set of 3 independent Bernoulli trials? Suppose that the machine can make a large number of bottles so that when a bottle is sampled, the machine replaces it by making a new bottle. If the machine produces defective bottles independently from bottle to bottle, then the 3 bottles are considered 3 independent Bernoulli trials. If we let the random variable Y be the number of defective bottles, then Y is a binomial random variable with $n = 3$ and $p = 0.05$. The PMF of Y is

$$P(Y = y) = {}_3C_y (0.05)^y (0.95)^{3-y} \text{ for } y = 0, 1, 2, 3 \text{ and zero elsewhere.}$$

To compute the probability that at least one of the 3 bottles is defective:

$$\begin{aligned} P(Y \geq 1) &= P(Y=1) + P(Y=2) + P(Y=3) \\ &= {}_3C_1 (0.05)^1 (0.95)^{3-1} + {}_3C_2 (0.05)^2 (0.95)^{3-2} + {}_3C_3 (0.05)^3 (0.95)^{3-3} \\ &= 0.135375 + 0.007125 + 0.000125 \\ &= 0.1426 \end{aligned}$$

An alternative solution would be to apply the rule of complement.

$$P(Y \geq 1) = 1 - P(Y=0) = 1 - {}_3C_0 (0.05)^0 (0.95)^3 = 1 - 0.8574 = 0.1426.$$

Example 4

A multiple-choice exam consists of 15 questions. Each question has 5 choices with only one correct answer. An unprepared student takes this exam and answers the questions on sheer guesswork. The 15 questions can be regarded as 15 independent Bernoulli trials, and if a correct answer is the “success,” then the random variable Y , which is the number of questions that the student answers correctly, is binomial with $n = 15$ and $p = \frac{1}{5}$.

- a. What is the probability that the student gets exactly 7 correct answers?

$$P(Y=7) = {}_{15}C_7 (0.20)^7 (0.80)^8 = 0.01382$$

- b. What is the probability that the student gets anywhere from 4 to 6 correct answers?

$$P(4 \leq Y \leq 6) = \sum_{y=4}^6 {}_{15}C_y (0.20)^y (0.80)^{15-y} = 0.187604 + 0.103182 + 0.042993 = 0.33378$$

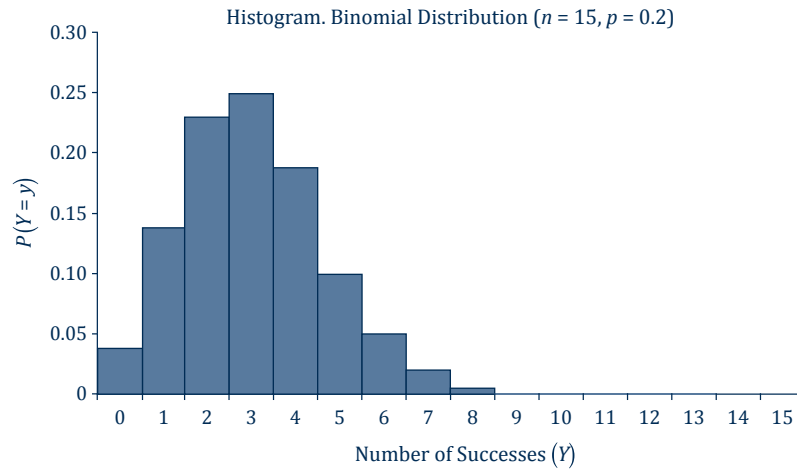
- c. What is the expected number of correct answers that the student gets? What is the standard deviation of the number of correct answers that the student gets?

$$\mu_Y = np = (15)(0.2) = 3$$

$$\sigma_Y^2 = np(1-p) = (15)(0.2)(0.8) = 2.4 \quad \text{and} \quad \sigma_Y = \sqrt{np(1-p)} = \sqrt{2.4} = 1.5492$$

Therefore, the mean number of correct answers that the student gets is 3, with a standard deviation of 1.5492.

- d. The histogram for the binomial distribution with $n = 15$ and $p = \frac{1}{5}$ is shown here, and we can see that the shape of the distribution for these values of n and p is right-skewed.



For the binomial distribution with a fixed value of n , as the probability of success p approaches 0.5, the distribution becomes symmetrical.

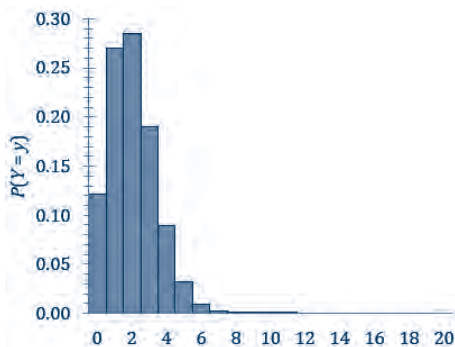
Let's Practice

I. Write the letter that corresponds to the correct answer. Write X if your answer is not among the given choices.

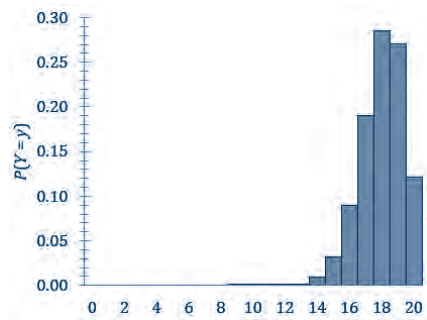
- _____ 1. Which is *not* a characteristic of the binomial distribution?
- The outcome of a trial can be classified in two ways.
 - The probability of success remains constant for all trials.
 - The outcome of a trial depends on the outcomes of the previous trials.
 - There is a fixed number of trials.
- _____ 2. In which random variable does the binomial distribution *not* apply?
- Three balls are drawn with replacement from an urn containing four white balls and six black balls. The random variable X is the number of black balls drawn.
 - Two books are selected from a shelf that contains five statistics books and five algebra books. The random variable X is the number of statistics books in the selection.
 - The daily movement in the price of a given stock may be up by one point or down by one point. Assume that these daily price fluctuations are independent. The random variable X is the number of upward movements in the price of the stock in five trading days.
 - A pair of dice is rolled ten times. The random variable X is the number of times the sum of 12 spots on the two dice appears.

_____ 3. Which graph is the histogram for the binomial distribution with $n = 20$ and $p = 0.5$?

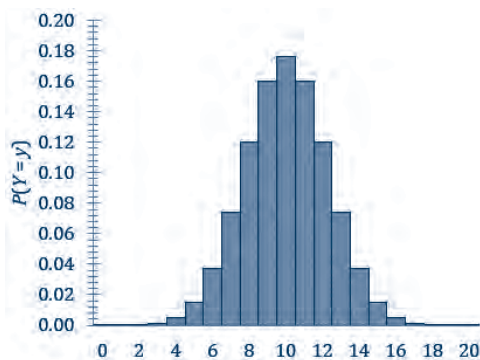
a.



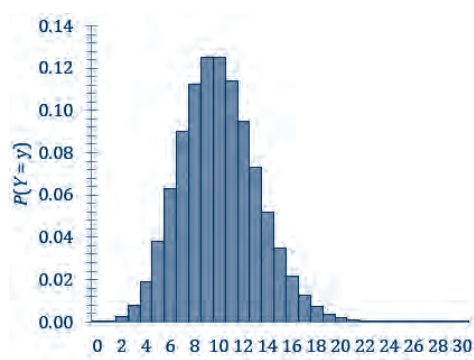
b.



c.



d.



II. Analyze and solve each problem.

- Suppose that in a certain university, it is known that 20% of students shift to another degree program after one year in their initial program. In a random sample of 15 freshman students at the university, find the probability that
 - exactly 8 students shift to another degree program.
 - anywhere from 5 to 7 students, shift to another degree program.
 - at least 3 students shift to another degree program.
- Five cards are drawn with replacement from a standard deck of 52 cards.
 - What is the probability that at least 3 cards are diamonds?
 - What is the expected number of diamond cards among the 5 cards?
 - What is the variance of the number of diamond cards among the 5 cards?

3. Answer the following problem, which was presented by the English diarist Samuel Pepys to Sir Isaac Newton: *Which is most likely to occur: at least one 6 when six dice are rolled; at least two 6's when twelve dice are rolled; or at least three 6's when eighteen dice are rolled?*
4. Suppose that 75% of adults who take a certain brand of antibiotic for upper respiratory infection report that they experience hyperacidity as a side effect. If the antibiotic is given to 10 new patients with upper respiratory infection, what is the probability that at least seven will not experience hyperacidity as a side effect?
5. Miguel has a pet dog named Chubibo. Miguel has 4 empty boxes, and he ordered Chubibo to place each of 6 identical balls into any of the boxes. Assume that each box can accommodate 6 balls. What is the expected value of the random variable T , which is the total number of boxes that are empty?

The Hypergeometric Random Variable

The binomial distribution does not apply in cases when one is counting the number of successes in a random sample of size n from a small population, without replacement. The sample does not constitute a sequence of independent Bernoulli trials, and the probability of success does not remain constant throughout the trials. We need a probability distribution function that will help us calculate probabilities of obtaining a given number of successes in the case where the binomial distribution does not apply.

Definition 3

The random variable X that counts the number of successes in a random sample of size n drawn from a small (finite) population of size N , of which m are labeled as “successes” and $N - m$ are labeled as “failures” is called a **hypergeometric random variable**. If we assume that the sample size n does not exceed the number of “successes” m or the number of “failures” $N - m$, then the probability mass function of X is

$$P(X=x) = \frac{{}_m C_x \cdot {}_{N-m} C_{n-x}}{{}_N C_n}, \quad x = 0, 1, 2, \dots, n, \quad \text{where } n \leq \min(m, N-m) \text{ and}$$

zero elsewhere.

There may be situations where the smallest possible value of x is not zero. In general, the possible values of the hypergeometric random variable X are $\{\max(0, m + n - N) \leq x \leq \min(m, n)\}$ where x is a nonnegative integer.

The expected value of the hypergeometric random variable X is $\mu_x = \frac{nm}{N}$ and the variance of X is $\sigma_x^2 = n \left(\frac{m}{N} \right) \left(1 - \frac{m}{N} \right) \left(\frac{N-m}{N-1} \right)$.

Example 5

A box of candies contains 12 chocolate-covered walnuts and 8 truffles. A child selects 5 candies at random from the box. If the random variable X is the number of truffles in the child's selection, then X is a hypergeometric random variable with $m = 8$, $n = 5$, and $N = 20$.

The PMF of X is $P(X = x) = \frac{{}_8C_x \cdot {}_{12}C_{5-x}}{{}_{20}C_5}$, $x = 0, 1, 2, 3, 4, 5$ and zero elsewhere.

To compute the probability that the child selects fewer than 2 truffles, we have:

$$\begin{aligned} P(X < 2) &= P(X = 0) + P(X = 1) \\ &= \sum_{x=0}^1 \frac{{}_8C_x \cdot {}_{12}C_{5-x}}{{}_{20}C_5} \\ &= \frac{{}_8C_0 \cdot {}_{12}C_{5-0}}{{}_{20}C_5} + \frac{{}_8C_1 \cdot {}_{12}C_{5-1}}{{}_{20}C_5} \\ &= 0.051084 + 0.255418 \\ &= 0.306502 \end{aligned}$$

Example 6

An urn contains 12 balls numbered 1 to 12. In a gambling game, you are asked to take 3 balls at random, without replacement from the urn. You place a wager (bet) that the largest numbered ball in your selection is a number that is at least as large as the number 8.

- a. If we let the random variable X be the number of selected balls whose number label is at least as large as the number 8, then X is a hypergeometric random variable with $m = 5$ (balls numbered 8, 9, 10, 11, and 12) and $N = 12$. Suppose you want to compute the probability that you will win in the wager. The desired probability is $P(X \geq 1) = 1 - P(X = 0)$. This is the complement of the event where none of the 3 selected balls has a number that is at least as large as 8. As such, $1 - P(X = 0) = 1 - \frac{{}_5C_0 \cdot {}_7C_3}{{}_{12}C_3} = \frac{37}{44}$. This is the probability that you will win the game.

- b. If you play this gambling game a total of 10 times, what is the probability that you will win anywhere from 5 to 7 times?

We can think of the 10 times you played the gambling game as 10 independent Bernoulli trials, with winning the game as “success” and using the probability in part (a),

$p = \frac{37}{44}$. If we let the random variable Y be the number of times out of 10 games that you win, then Y is a binomial random variable. The desired probability is

$$\begin{aligned}
 P(5 \leq Y \leq 7) &= \sum_{y=5}^7 {}_{10}C_y \left(\frac{37}{44}\right)^y \left(\frac{7}{44}\right)^{10-y} \\
 &= P(Y=5) + P(Y=6) + P(Y=7) \\
 &= {}_{10}C_5 \left(\frac{37}{44}\right)^5 \left(\frac{7}{44}\right)^5 + {}_{10}C_6 \left(\frac{37}{44}\right)^6 \left(\frac{7}{44}\right)^4 + {}_{10}C_7 \left(\frac{37}{44}\right)^7 \left(\frac{7}{44}\right)^3 \\
 &= 0.010799 + 0.047566 + 0.143668 \\
 &= 0.202033
 \end{aligned}$$

Let's Practice

I. Write the letter that corresponds to the correct answer. Write X if your answer is not among the given choices.

- _____ 1. Which statement is *false*?
 - a. When sampling without replacement from a small population, the probability of obtaining a certain number of successes is best given by a hypergeometric distribution.
 - b. The hypergeometric distribution is defined by three parameters, namely, the finite population size N , the sample size n , and the probability of success, p .
 - c. The hypergeometric distribution is a discrete probability distribution.
 - d. In a hypergeometric distribution, the outcomes are dependent on the outcomes of the previous trials.
- _____ 2. Which is *true* about the probability of success in a hypergeometric distribution?
 - a. It is equal to the probability of failure.
 - b. It is larger than the probability of failure.
 - c. It remains constant from trial to trial.
 - d. It varies from trial to trial.
- _____ 3. For what condition does the hypergeometric distribution “converge” to the binomial distribution?
 - a. When the population size $N \rightarrow \infty$.
 - b. When the population size N equals the sample size n .
 - c. When the sample size $n \rightarrow \infty$.
 - d. When the sample size n exceeds the population size N .

- _____ 4. Which is true about the mean of a hypergeometric distribution?
- The mean is equal to the variance.
 - The mean is half of the variance.
 - The mean is larger than the variance.
 - The mean is smaller than the variance.
- _____ 5. In a hypergeometric distribution, suppose the parameters are $N = 10$, $n = 3$, and $m = 5$. What is the range of values of the distribution?
- | | |
|-----------|------------|
| a. 0 to 3 | c. 0 to 10 |
| b. 0 to 5 | d. 0 to 15 |
- _____ 6. In a hypergeometric distribution, suppose the parameters are $N = 7$, $n = 3$, and $m = 5$. What is the range of values of the distribution?
- | | |
|-----------|-----------|
| a. 0 to 3 | c. 1 to 3 |
| b. 0 to 5 | d. 1 to 5 |

II. Analyze and solve each problem.

- An urn contains 4 white balls and 6 red balls. Three balls are drawn in succession, without replacement from the urn. Find the probability that
 - there are 2 red balls and 1 white ball in the selection.
 - at least one ball in the selection is red.
 - at most two balls in the selection are red.
- The Philippine Lotto 6/42 for Luzon started in 1995, and at present, it is played three times a week. A ticket for a single play costs ₱24. Balls of equal weights are numbered 1 to 42 and placed in a machine to be mixed up. Six balls are ejected from the machine at random to form the winning combination for the draw. A player wins the day's jackpot prize if he or she matches all the six winning numbers (in no particular order). For a standard 6/42 lottery,
 - a pay out of ₱24 is made if a player matches exactly 3 of the six winning numbers. What is the probability of this event happening?
 - a pay out of ₱1,000 is made if a player matches exactly 4 of the six winning numbers. What is the probability of this event happening?
 - a pay out of ₱25,000 is made if a player matches exactly 5 of the six winning numbers. What is the probability of this event happening?
- A batch of 20 circuit boards contains 3 boards that are defective. Five of these boards are to be selected at random, without replacement, for function testing. If at least one circuit board in the sample is found defective, the entire batch is to be returned.
 - What is the probability that the entire batch will be deemed acceptable?
 - What is the expected number of defective circuit boards in the batch?

4. A box contains balls that are numbered 1 to 50. Two balls are selected at random and without replacement from the box. What is the probability of selecting two numbers such that their sum is an odd number?

The Poisson Random Variable

In 1837, French mathematician Simeon-Denis Poisson published his work *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile* involving a study made on the computation of binomial probabilities when the number of independent Bernoulli trials is large. His procedure obtained the formula that approximates the binomial distribution for large values of n , with a small probability of success p . The importance of this approximation in the field of probability theory and its applications became known in 1889, when German-Russian mathematician Ladislaus Bortkiewicz proved that the approximation proposed by Poisson is itself a probability mass function and is now regarded as one of the most important probability distribution functions. It was Bortkiewicz who made a study on the first application of the Poisson distribution, where he used the distribution to model the number of Prussian army cavalier casualties arising from the horse kicks.

Definition 4

A **Poisson experiment** is a procedure that possesses the following properties:

1. The number of successes, which will be referred to as *event occurrences*, in two disjoint time intervals or two disjoint regions of space are independent.
2. The number of event occurrences in a given small time interval or region of space is proportional to the entire length of the time interval or region of space.
3. The probability that events occur at exactly the same instant is virtually zero.
4. The mean rate of event occurrences per time interval or region of space is a constant.

If we denote the constant mean rate of event occurrences per time interval as λ , then the random variable X that counts the number of event occurrences resulting from a Poisson experiment is called a **Poisson random variable** with probability mass function

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ for } x = 0, 1, 2, \dots \text{ and zero elsewhere}$$

where e is the Euler's number approximately equal to 2.71828.

The expected value of the Poisson distribution is λ , and its variance is also λ .

Some examples of random variables that can be modeled with the Poisson distribution:

1. *The number of customer arrivals in a minute at a grocery store.* A customer arrival is an event occurrence. If we assume that at no time during the day will there be a higher or lower rate of customer arrival, then the Poisson distribution applies.
2. *The number of typographical errors per page of a manuscript typed by a clerk.* A typographical error is an event occurrence; and it is reasonable to assume that the average rate of such errors is constant. Thus, the Poisson distribution applies.
3. *The number of meteors that hit the earth in a year.* A meteor hit is an event occurrence and can be assumed to be a Poisson event.
4. *The number of checked-in pieces of luggage during commuter flights that are lost per week.* A lost piece of luggage is an event occurrence and the Poisson distribution applies.

Example 7

The number of misprints on a single page of a textbook can be regarded as a Poisson random variable, say X . Suppose that, on the average, in every five pages of a textbook there are two misprints, then the mean rate of event occurrence is $\lambda = \frac{2}{5}$.

- a. What is the probability that there is exactly one error on a specific page of a textbook?

$$P(X=1) = \frac{e^{-2/5} \left(\frac{2}{5}\right)^1}{1!} = 0.26813$$

- b. What is the probability that there are at least two errors on a specific page of a textbook?

$$\begin{aligned} P(X \geq 2) &= 1 - [P(X=0) + P(X=1)] \\ &= 1 - \sum_{x=0}^1 \frac{e^{-2/5} \left(\frac{2}{5}\right)^x}{x!} \\ &= 1 - \left[\frac{e^{-2/5} \left(\frac{2}{5}\right)^0}{0!} + \frac{e^{-2/5} \left(\frac{2}{5}\right)^1}{1!} \right] \\ &= 1 - (0.67032 + 0.26813) \\ &= 1 - 0.93845 \\ &= 0.06155 \end{aligned}$$

Example 8

The number of blueberries in a muffin baked in commercial quantities by a popular pastry shop is a random variable X , believed to follow a Poisson distribution with a mean rate of $\lambda = 5.3$ blueberries per muffin.

- a. What is the probability that a given muffin has exactly 4 blueberries?

$$P(X = 4) = \frac{e^{-5.3} (5.3)^4}{4!} = 0.16411$$

- b. If a customer purchases a dozen blueberry muffins from the shop, what is the probability that anywhere from 6 to 8 of the muffins will contain exactly 4 blueberries each?

The twelve blueberry muffins in the purchase may be regarded as 12 independent Bernoulli trials, with a muffin having exactly 4 blueberries as the “success.” Let the random variable Y be the number of blueberry muffins in a dozen muffins that have exactly 4 blueberries. Then Y has a binomial distribution with $n = 12$ and probability of success $p = P(X = 4) = 0.16411$. The desired probability is

$$P(6 \leq Y \leq 8) = \sum_{y=6}^8 {}_{12}C_y (0.16411)^y (0.83589)^{12-y} = 0.006157 + 0.001036 + 0.000127 = 0.00732.$$

Example 9

The number of babies born at a government-operated hospital is a Poisson random variable with a mean rate of $\lambda = 8$ babies per day.

- a. What is the probability that at least 3 babies are born during the next twelve hours?

If we let X be the number of babies born per day at the hospital, then X has a Poisson distribution with $\lambda = 8/\text{day}$, which is equivalent to $\lambda = 4/\text{half-day or 12 hours}$. The desired probability is $P(Y \geq 3)$ where Y has a Poisson distribution with $\lambda = 4$.

$$P(Y \geq 3) = 1 - P(Y \leq 2) = 1 - \sum_{y=0}^2 \frac{e^{-4} 4^y}{y!} = 1 - 0.2381 = 0.7619$$

- b. What is the probability that exactly six babies are born during the next two days at the hospital?

If a mean rate of $\lambda = 8/\text{day}$ is equivalent to $\lambda = 16/\text{two days}$, then the desired probability is $P(W = 6)$ where W has a Poisson distribution with $\lambda = 16$.

$$P(W = 6) = \frac{e^{-16} (16)^6}{6!} = 0.002622$$

Points to Remember

1. The Poisson distribution applies when it is desired to calculate the probability that a number of events will occur in a given period of time or region.
2. The Poisson distribution has the following assumptions:
 - a. Events occur at a constant rate, λ , in a given period of time or region.
 - b. Events occur independently.
 - c. Events occur at random in a given period of time or region.
 - d. In an instant, there is either no occurrence or one occurrence of an event.
 - e. The probability that an event occurs in a given interval is proportional to the length of the interval.
3. In the binomial distribution with n independent trials and probability of success p , as n increases without bound ($n \rightarrow \infty$), with probability of success p approaching zero, ($p \rightarrow 0$), and if the mean of the binomial distribution, np is equated to the mean of the Poisson distribution, λ (a constant) such that $np = \lambda$, then the binomial distribution converges to the Poisson distribution.

Let's Practice

I. Write the letter that corresponds to the correct answer. Write X if your answer is not among the given choices.

- _____ 1. Which is *true* of Poisson distribution?
- a. The mean is equal to the standard deviation.
 - b. The mean is equal to the variance.
 - c. The median is equal to the standard deviation.
 - d. The median is equal to the variance.
- _____ 2. Cars arrive at a gasoline station at the rate of 3 per minute following the Poisson distribution. What is the probability of 4 arrivals in a one-minute interval?
- | | |
|---------------------------|---------------------------|
| a. $\frac{e^{-3}3^4}{4!}$ | c. $\frac{e^{-4}3^4}{3!}$ |
| b. $\frac{e^{-4}4^3}{3!}$ | d. $\frac{e^{-3}4^3}{3!}$ |

- _____ 3. Which statement is true regarding a Poisson distribution?
- The rate at which events occur may be higher in some intervals or lower in other intervals.
 - Two or more events can occur at exactly the same instant.
 - The occurrence of an event does not influence the probability of the occurrence of another event.
 - The number of times an event can occur in a given period can only be a positive integer.

II. Analyze and solve each problem.

- The number of employees in an office who are absent on Monday is assumed to follow a Poisson distribution with a mean of $\lambda = 3.2$. On a given Monday, what is the probability that
 - there are no absent employees?
 - at least 3 employees are absent?
- The number of calls received per minute by a call center agent for a credit card company is assumed to follow a Poisson distribution with a mean of $\lambda = 4$.
 - What is the probability that in a given minute, the call center agent receives exactly 2 calls?
 - What is the probability that in a given two-minute period, the call center agent receives at least 3 calls?
- Suppose that a large lake is contaminated with bacteria. It has been found that for every 10 mL of water from the lake, the average number of bacteria is assumed to follow a Poisson distribution with an average of $\lambda = 6$ bacteria. What is the probability that there are exactly 24 bacteria in a 50-mL sample of water from the lake?
- The job-stream of a computer system receives an average of $\lambda = 2$ jobs per minute. Assuming it follows a Poisson distribution,
 - what is the probability that exactly 6 jobs are received in a 5-minute period?
 - what is the probability that fewer than 3 jobs are received in a 2-minute period?
 - what is the largest number of jobs received in a 10-minute period so that the probability that the system receives this number of jobs is at least 0.80?

The Geometric Random Variable

Consider independent Bernoulli trials with probability of success, p , remaining constant throughout the trials. Our interest is the number of trials needed to observe the first “success.” If we let s denote the occurrence of a success and f denote the “failure,” then the sample space will be $S = \{s, fs, ffs, fffs, \dots\}$ where the first “success” happens on the x th trial for $x = 1, 2, 3, \dots$. It follows that previously, there have been $x - 1$ trials that all resulted in “failures.” The trials are independent and the probability of “success” remains constant.

Definition 13

The random variable X that counts the number of independent trials of an experiment until the **first success** occurs is called a **geometric random variable**, with probability of success p and probability of failure $1 - p$. Its probability mass function is

$$P(X = x) = (1 - p)^{x-1} \cdot p \text{ for } x = 1, 2, 3, \dots \text{ and zero elsewhere.}$$

The expected value of the geometric random variable X is $\mu_x = \frac{1}{p}$ and its variance is $\sigma_x^2 = \frac{1 - p}{p^2}$.

Example 10

A mother has three daughters named Alice, Belle, and Carla, who have yet to decide who gets to wash the dishes after dinner. One of the girls suggested that each of them flips a fair coin, and the “odd girl” (the girl who gets a different side of the coin from the other two) will wash the dishes. If all three outcomes turn up the same, they flip their coins again.

- a. What is the probability that fewer than 4 flips are required to reach a decision?

If we denote the occurrence of a head as h and tail as t , we can refer to sample space specified in Example 2 of lesson 2, where $S = \{(h, h, h), (h, h, t), (h, t, h), (t, h, h), (h, t, t), (t, h, t), (t, t, h), (t, t, t)\}$ wherein Alice flips the one-peso coin, Belle flips the five-peso coin, and Carla flips the ten-peso coin. Since all three coins are fair, these eight outcomes are equally likely. The “success” is when they reach a decision, that is, when there is an “odd girl,” meaning, not all the three tosses come up with the same result. The probability of success in this case is $p = \frac{6}{8}$ or 0.75. The number

of trials needed to reach a decision is a geometric random variable and the desired probability is

$$\begin{aligned}
 P(X < 4) &= \sum_{x=1}^3 (1-p)^{x-1} \cdot p \\
 &= \sum_{x=1}^3 \left(\frac{1}{4}\right)^{x-1} \cdot \frac{3}{4} \\
 &= \left(\frac{1}{4}\right)^{1-1} \cdot \frac{3}{4} + \left(\frac{1}{4}\right)^{2-1} \cdot \frac{3}{4} + \left(\frac{1}{4}\right)^{3-1} \cdot \frac{3}{4} \\
 &= \frac{3}{4} + \frac{3}{16} + \frac{3}{64} \\
 &= \frac{63}{64}
 \end{aligned}$$

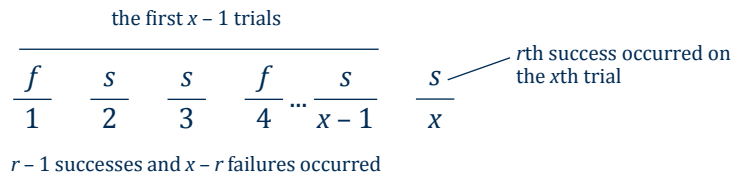
- b. What is the expected number of tosses needed to reach a decision? What is the variance of the number of tosses needed to reach a decision?

Since X has a geometric distribution with probability of success as 0.75, the mean number of tosses need to reach a decision is $\frac{1}{p} = \frac{4}{3}$ and the variance is

$$\frac{1-p}{p^2} = \frac{0.25}{(0.75)^2} = \frac{4}{9}.$$

The Negative Binomial Distribution

The last discrete probability distribution with a special name to be discussed in this chapter also involves independent Bernoulli trials with probability of success, p , remaining constant throughout the trials. Our interest is the number of trials needed to accumulate a total of r “successes.” As usual, we let s denote the occurrence of a success and f denote the failure. Suppose that independent trials are performed and the r th success occurs on the x th trial. Then it follows that on the previous $x - 1$ trials, there were exactly $r - 1$ successes and $x - 1 - (r - 1) = x - r$ failures. The figure below illustrates this.



For a sequence of $x - 1$ characters of which $r - 1$ are “ s ” and $x - r$ are “ f ”, the number of possible arrangements of the characters is ${}_{x-1}C_{r-1}$, and with independent trials and p as the probability of success and $1 - p$ as probability of failure, we now define the following probability mass function.

Definition 14

The random variable X that counts the number of independent trials of an experiment until the observance of r “successes” is called a **negative binomial random variable**, with probability of success p and probability of failure $1 - p$. Its probability mass function is

$$P(X = x) = {}_{x-1}C_{r-1} p^r (1-p)^{x-r} \text{ for } x = r, r+1, r+2, \dots \text{ and zero elsewhere.}$$

Remark: The negative binomial distribution is a generalization of the geometric distribution. Note that if $r = 1$ in the negative binomial PMF, it becomes the PMF of the geometric distribution.

The expected value of the negative binomial distribution is $\mu_X = \frac{r}{p}$ and its variance is $\sigma_X^2 = \frac{r(1-p)}{p^2}$.

Example 11

Cody and Joe want to play a series of chess games until one of them wins seven games. Assume that the games are independent and the probability that Cody wins a game is 0.60.

- What is the probability that the series ends in nine games?
- Given that the series ends in nine games, find the conditional probability that Cody is the winner.



Solution:

To solve (a), let X be the number of chess games needed for Cody to win seven games, and let Y be the number of games needed for Joe to win seven games. Both random variables X and Y are negative binomial (NB) distributed with $X \sim \text{NB}(r = 7, p = 0.6)$ and $Y \sim \text{NB}(r = 7, p = 0.4)$. Suppose that we let A be the event that the series ends in nine games. The desired probability is $P(A)$.

$$\begin{aligned} P(A) &= P(X = 9) + P(Y = 9) \\ &= {}_8C_6 (0.6)^7 (0.4)^2 + {}_8C_6 (0.4)^7 (0.6)^2 \\ &= 0.125411 + 0.016515 \\ &= 0.141926 \end{aligned}$$

To solve (b), let B be the event that Cody wins the series. The conditional probability $P(B|A)$ is

$$\begin{aligned} P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{P(X=9)}{P(A)} \\ &= \frac{0.125411}{0.141926} \\ &= 0.88364 \end{aligned}$$

Let's Practice

I. Write True if the statement is *always* true. Otherwise, write False.

- _____ 1. In a geometric distribution, if the random variable X counts the number of failures that occur before a success is finally observed, then the possible values of X are $x = 1, 2, 3, \dots$
- _____ 2. The geometric distribution is similar to the binomial distribution in the sense that the trials are independent and the probability of success must remain constant throughout all the trials.
- _____ 3. The sample space of the geometric distribution has a finite number of sample points.
- _____ 4. In the negative binomial distribution where $r = 7$, the possible values of X are $x = 1, 2, 3, 4, 5, 6$, and 7 .
- _____ 5. It is possible to remodel the negative binomial PDF where the random variable X counts the number of failures (instead of successes) until r successes are observed.

II. Analyze and solve each problem.

1. A man has 12 keys in his key chain, only one of which will open his apartment door. One night, he came home drunk and his doorway is too dark for him to identify the key to his apartment door. He tries to open the door by testing the keys one at a time, without tossing out the unsuccessful keys he has tried.
 - a. What is the probability that he will find the key to his apartment door in exactly 5 tries?
 - b. What is the probability that he will find the key in at most 4 tries?
 - c. What is the expected number of keys he would have to try until he finds the key to his apartment door?

2. Cards are drawn from a standard poker deck with replacement, until a face card (king, queen, or jack) is found. What is the probability that at least 8 draws are needed?
3. A newly wed couple is starting a family and they will not stop having children until they have a son. Assuming that the genders of children are equally likely, what is the probability that they will have a son who has three older sisters?
4. A fair coin is tossed until we observe 4 heads. What is the probability that this event happens in exactly 9 tosses?
5. From a standard deck of 52 cards, we draw cards at random in succession with replacement until we observe 5 face cards. What is the probability that at least 10 draws are needed?

Software Tutorial in MS Excel

1. Probabilities involving the binomial distribution may be calculated using the function “=BINOM.DIST(y, n, p, cumulative),” where y is the desired number of successes, n is the number of Bernoulli trials, p is the probability of success, and cumulative is a logical command. The values of cumulative are 0 (FALSE) if one is solving for the probability of a single value of Y , and 1 (TRUE) if one is accumulating the probabilities of several values of Y . For example, to find $P(Y = 7)$ when Y is binomial with $n = 10$ and $p = \frac{3}{4}$, we type “=BINOM.DIST(7, 10, 0.75, 0)” on the command line, and obtain the answer 0.2503.

	A	B	C	D
1	0.250282	=BINOM.DIST(7,10,0.75,0)		

For the same example, to find $P(4 \leq Y \leq 6)$, we type “=BINOM.DIST(6, 10, 0.75, 1) – BINOM.DIST(3, 10, 0.75, 1)” on the command line, and obtain the answer of 0.2206.

	A	B	C	D	E	F
1	0.220619	=BINOM.DIST(6,10,0.75,1)-BINOM.DIST(3,10,0.75,1)				

For the same example, to find $P(Y \geq 8) = 1 - P(Y \leq 7)$, we type “= 1 – BINOM.DIST(7, 10, 0.75, 1)” on the command line and obtain the answer 0.5256.

	A	B	C	D
1	0.525593	=1-BINOM.DIST(7,10,0.75,1)		

2. Probabilities involving the hypergeometric distribution may be calculated using the function “=HYPGEOM.DIST(x, n, m, N, cumulative),” where x is the desired number of successes, n is the sample size drawn without replacement, m is the number of items in the population labeled as success, N is the (finite) population size, and cumulative is the same logical command presented in the binomial distribution illustration in part (1). For example, to find $P(X = 1)$ when X is hypergeometric with $n = 5$, $m = 8$, and $N = 20$, we type “=HYPGEOM.DIST(1, 5, 8, 20, 0),” on the command line, and obtain the answer 0.2554.

	A	B	C	D
1	0.255418	=HYPGEOM.DIST(1,5,8,20,0)		

For the same example, to find $P(X < 2)$, we type “=HYPGEOM.DIST(1, 5, 8, 20, 1)” on the command line and obtain the answer 0.3065.

	A	B	C	D
1	0.306502	=HYPGEOM.DIST(1,5,8,20,1)		

3. Probabilities involving the Poisson distribution may be calculated using the function “=POISSON.DIST(x, λ , cumulative),” where x is the desired number of event occurrences, λ is the Poisson parameter, and cumulative is the same logical command presented in the binomial distribution illustration in parts (1) and (2). For example, to find $P(X = 0)$ when X is Poisson distributed with $\lambda = 1$, we type “=POISSON.DIST(0, 1, 0)” on the command line, and obtain the answer 0.36788.

	A	B	C	D
1	0.367879	=POISSON.DIST(0,1,0)		

To find $P(X \geq 2)$ when X is Poisson distributed with $\lambda = 2.5$, we type “=1-POISSON.DIST(1, 2.5, 1)” on the command line and obtain the answer 0.7127.

	A	B	C	D
1	0.712703	=1-POISSON.DIST(1,2.5,1)		

4. Probabilities involving the negative binomial distribution and the geometric distribution may be calculated using the function “=NEGBINOM.DIST($x - r$, r , p , cumulative),” where x is the number of independent Bernoulli trials needed to accumulate r successes, with probability of success p , and cumulative is the same logical command presented in the previous probability distributions. For example, if X has a negative binomial distribution with $r = 7$ and $p = 0.6$ and we want to compute $P(X = 9)$, we type “=NEGBINOM.DIST(2,7,0.6,0)” on the command line, and obtain the answer 0.125411.

	A	B	C	D
1	0.125411	=NEGBINOM.DIST(2,7,0.6,0)		

As another example, if X has a geometric distribution, then we can use the same NEGBINOM.DIST function by just setting $r = 1$. Thus, if X has a geometric distribution with $p = 0.25$ and we want to compute $P(X \geq 10)$, which is equal to $1 - P(X \leq 9)$, we type “=1-NEGBINOM.DIST(8,1,0.25,1)” on the command line, and obtain the answer 0.075085.

	A	B	C	D
1	0.07508469	=1-NEGBINOM.DIST(8,1,0.25,1)		

Chapter Review

- A **random variable** is a function that assigns a unique real number to each element in the sample space. In other words, a **random variable** is a real-valued function whose domain is the sample space S .
- If the sample space of a random experiment consists of a finite number of elements or has an unending sequence with as many elements as there are counting numbers, then it is called a **discrete sample space**.
- A random variable defined over a discrete sample space is called a **discrete random variable**.
- For a discrete random variable X , the **probability mass function** (or **PMF**) of X , denoted by $p(k)$, is defined as a function $p(k) = P(X = k)$. A PMF in the form of a formula will look like this:

$$P(X = k) = \begin{cases} p(x_k) & \text{if } k = x_1, x_2, \dots \\ 0 & \text{otherwise} \end{cases}$$

- If X is a discrete random variable with probability mass function $p(x)$, then
 1. $p(x_k) \geq 0$ for $k = 1, 2, \dots$ and $p(x) = 0$ for all other values of x . (**Nonnegativity Property**)
 2. $\sum_{k=1}^{\infty} p(x_k) = 1$ (**Norming Property**)
 3. the values x_1, x_2, \dots of X for which $p(x) > 0$ are called the **mass points** of X .
 4. the probabilities involving events associated with any value(s) of X may be computed by taking the sum of the probabilities of the mass points. Therefore,
 - a. $P(X \leq a) = \sum_{\{k: x_k \leq a\}} p(x_k)$
 - b. $P(a \leq X \leq b) = \sum_{\{k: a \leq x_k \leq b\}} p(x_k)$
 - c. $P(X \geq b) = \sum_{\{k: x_k \geq b\}} p(x_k)$

- The graph of the probability mass function of the random variable X is called a **probability histogram**.
- If the sample space of a random experiment consists of uncountably infinite number of outcomes, then it is called a **continuous sample space**.
- A random variable defined over a continuous sample space is called a **continuous random variable**.
- Suppose that X is a continuous random variable.
 1. The probability that X will assume a particular value k is practically zero; thus, $P(X = k) = 0$.
 2. Unlike in the discrete case, a continuous random variable has no mass points. Thus, calculation of probabilities of the type $P(X = k)$ are not done. Instead, we will calculate probabilities of X taking on a value on an interval such as $P(a < X < b)$, $P(X < a)$, or $P(X > b)$.
 3. Unlike in the discrete case, when computing the probability that X takes a value on the interval (a, b) , it does not matter whether we include an endpoint of the interval or not. Therefore, $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$.
- For a continuous random variable X , the **probability density function** (or **PDF**) of X , denoted by $f(x)$, is a real-valued function defined on a continuous sample space S with an uncountable number of outcomes.
 1. $f(x) \geq 0$ for all x in S .
 2. The total area under the whole curve of $f(x)$ above the x -axis is always equal to 1
 3. The probability that X assumes a value within an interval (a, b) is the area bounded by the curve of $f(x)$, the x -axis, and the lines $x = a$ and $x = b$. This is **$P(a < X < b)$** .
- For a discrete random variable X with probability mass function

x	x_1	x_2	...	x_n
$P(X = x)$	$p(x_1)$	$p(x_2)$...	$p(x_n)$

the **expected value** of X is defined as

$$E(X) = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \cdots + x_n \cdot p(x_n) = \sum_{k=1}^n x_k \cdot p(x_k)$$

The expected value of a random variable X is also referred to as the **mean** of X , denoted by μ_X .

- For a discrete random variable X with probability mass function

x	x_1	x_2	\dots	x_n
$P(X = x)$	$p(x_1)$	$p(x_2)$	\dots	$p(x_n)$

the **variance** of X , denoted by σ_X^2 is defined as

$$\sigma_X^2 = (x_1 - \mu_X)^2 \cdot p(x_1) + (x_2 - \mu_X)^2 \cdot p(x_2) + \dots + (x_n - \mu_X)^2 \cdot p(x_n) = \sum_{k=1}^n (x_k - \mu_X)^2 \cdot p(x_k).$$

- The positive square root of the variance is called the **standard deviation** of the random variable X , denoted by σ_X .
- For the case of discrete random variables,
 - an alternative solution for finding the variance is

$$\sigma_X^2 = [x_1^2 \cdot p(x_1) + x_2^2 \cdot p(x_2) + \dots + x_n^2 \cdot p(x_n)] - \mu_X^2 = \sum_{k=1}^n x_k^2 \cdot p(x_k) - \mu_X^2.$$
 - given two random variables X and Y , the expected value of the sum of X and Y is equal to the sum of the individual expected values, that is, $E(X + Y) = \mu_X + \mu_Y$.
 - given two random variables X and Y that are independent, the variance of the sum of X and Y is equal to the sum of the individual variances, that is, $\text{Var}(X + Y) = \sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$.
- The **coefficient of variation**, or CV , is a measure of relative dispersion that expresses the standard deviation as a percentage of the mean. The coefficient of variation of a random variable X is computed as $CV = \frac{\sigma_X}{\mu_X} \times 100$.
- A **fair game** is one where the expected amount of the winnings is equal to the ante, the bet, or the amount paid out. In a fair game, there is neither gain nor loss. Thus, a game with zero expectation is defined as fair.
- The **Bernoulli random variable** is a random variable whose only two values are 1 with probability p and 0 with probability $1 - p$. The probability mass function of the Bernoulli random variable X is

x	1	0
$P(X = x)$	p	$1 - p$

The expected value of the Bernoulli random variable X , with probability of success p is $\mu_X = p$.

The variance of the Bernoulli random variable is $\sigma_X^2 = p(1 - p)$.

- The random variable Y that counts the number of successes in n independent Bernoulli trials, with probability of success p , is called a **binomial random variable**. The probability mass function of Y is $P(Y = y) = {}_n C_y p^y (1 - p)^{n-y}$ for $y = 0, 1, 2, \dots, n$ and zero elsewhere. The expected value of the binomial random variable Y is $\mu_Y = np$ and the variance of Y is $\sigma_Y^2 = np(1 - p)$.

- The random variable X that counts the number of successes in a random sample of size n drawn from a small (finite) population of size N , of which m are labeled as “successes” and $N - m$ are labeled as “failures,” is called a **hypergeometric random variable**. If we assume that the sample size n does not exceed the number of “successes” m or the number of “failures” $N - m$, then the probability mass function of X is $P(X = x) = \frac{{}^m C_x \cdot {}^{N-m} C_{n-x}}{{}^N C_n}$ for $x = 0, 1, 2, \dots, n$, where $n \leq \min(m, N - m)$ and zero elsewhere. In general, the possible values of the hypergeometric random variable X are $\{\max(0, m + n - N) \leq x \leq \min(m, n)\}$ where x is a non-negative integer. The expected value of the hypergeometric random variable X is $\mu_x = \frac{nm}{N}$ and the variance of X is $\sigma_x^2 = n \left(\frac{m}{N} \right) \left(1 - \frac{m}{N} \right) \left(\frac{N-m}{N-1} \right)$.
- A **Poisson experiment** is a procedure that possesses the following properties:
 1. The number of successes, which will be referred to as *event occurrences*, in two disjoint time intervals or two disjoint regions of space are independent.
 2. The number of event occurrences in a given small time interval or region of space is proportional to the entire length of the time interval or region of space.
 3. The probability that events occur at exactly the same instant is virtually zero.
 4. The mean rate of event occurrences per time interval or region of space is a constant.

If we denote the constant mean rate of event occurrences per time interval as λ , then the random variable X that counts the number of event occurrences resulting from a Poisson experiment is called a **Poisson random variable** with probability mass function $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$ for $x = 0, 1, 2, \dots$ and zero elsewhere. The expected value of the Poisson distribution is λ , and its variance is also λ .

- The random variable X that counts the number of independent trials of an experiment until the first “success” occurs is called a **geometric random variable**, with probability of success p and probability of failure $1 - p$. Its probability mass function is $P(X = x) = (1 - p)^{x-1} \cdot p$ for $x = 1, 2, 3, \dots$ and zero elsewhere. The expected value of the geometric distribution is $\frac{1}{p}$ and its variance is $\frac{1-p}{p^2}$.
- The random variable X that counts the number of independent trials of an experiment until the observance of r “successes” is called a **negative binomial random variable**, with probability of success p and probability of failure $1 - p$. Its probability mass function is

$$P(X = x) = {}_{x-1} C_{r-1} p^r (1-p)^{x-r} \text{ for } x = r, r+1, r+2, \dots \text{ and zero elsewhere.}$$

The negative binomial distribution is a generalization of the geometric distribution. Note that if $r = 1$ in the negative binomial PMF, it becomes the PMF of the geometric distribution. The expected value of the negative binomial distribution is $\mu_x = \frac{r}{p}$ and its variance is $\sigma_x^2 = \frac{r(1-p)}{p^2}$.

Chapter Performance Tasks

1. Mobile Phone Number

Conduct a simple study. Gather data by asking every student in class the last digit of his or her mobile phone number. Write down the answers then tally the results in a frequency distribution. Let the random variable X be the last digit of the mobile phone number of a student from the class. Then, using MS Excel, construct the probability mass function of X . Using the Chart Wizard tool, construct the probability histogram of X . Based on the probability mass function of X , find the following probabilities involving X : $P(X < 4)$, $P(X \leq 7)$, and $P(3 \leq X \leq 6)$.



2. How Large is Your Family?

Imagine that you are a sociology student making a study on the size of the Filipino family today compared to the size of the Filipino family 15 years ago. Collect data by asking each student in the class how many children there are in his or her family (the number of siblings in the family, including the student) through a survey. Write down the answers and tally the results in a frequency distribution. Let the random variable X be the number of children in the family of a student from the class. Afterwards, using MS Excel, construct the probability mass function of X . Using the Chart Wizard tool, construct the probability histogram of X . Based on the probability mass function of X , find the expected value of X . Then research the average number of children in a Filipino family 15 years ago.



Your final task is to write a term paper. Make the cover page of your term paper and design it using a collage of the family pictures of some of your classmates. In the body of your paper, include the results of your survey. Then compare the average number of children in a Filipino family 15 years ago to the expected value of X that you obtained from your probability mass function. Is it higher or lower? Discuss the possible suggested reasons for the difference (if any).

Chapter Exercises

1. Classify each of the following random variables as discrete or continuous.
 - a. the number of spelling errors per page in a manuscript
 - b. the number of tourist visa applicants at a consulate office in a day
 - c. the amount of soda contained in a bottle in mL
 - d. the time it takes to download an image file from the Internet
 - e. the height of a senior high school student
 - f. the number of complaint calls received by a customer service assistant in a day
 - g. the temperature in Tuguegarao at noon tomorrow in degrees Celsius
 - h. the number of personal computers in a household

2. Which is a valid probability mass function for a discrete random variable X ? Justify your answer.

a.

x	1	2	3	4
$P(X=x)$	$\frac{5}{19}$	$\frac{2}{19}$	$\frac{10}{19}$	$\frac{1}{19}$

b.

x	-1	0	1	2.5
$P(X=x)$	0.28	0.15	0.30	0.27

c.

x	4	5	6	7
$P(X=x)$	0.80	-0.15	0.20	0.15

d. $P(X=k) = \frac{k^2}{30}$ for $k = 0, 1, 2, 3, 4$

e.

x	2	4	6	8
$P(X=x)$	0.38	0.25	1.06	0.17

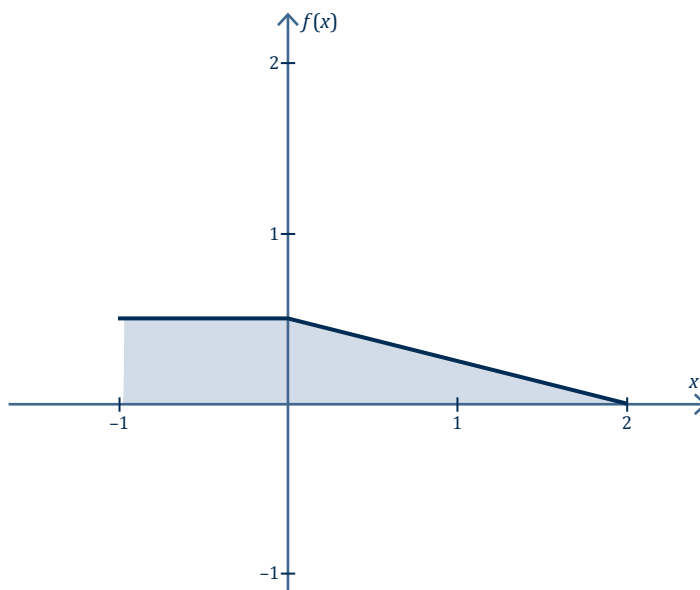
f. $P(X=k) = \left(\frac{1}{4}\right)\left(\frac{3}{4}\right)^k$ for $k = 0, 1, 2, \dots$

3. Suppose that the given function serves as the probability mass function of the discrete random variable X . Find the value of c .

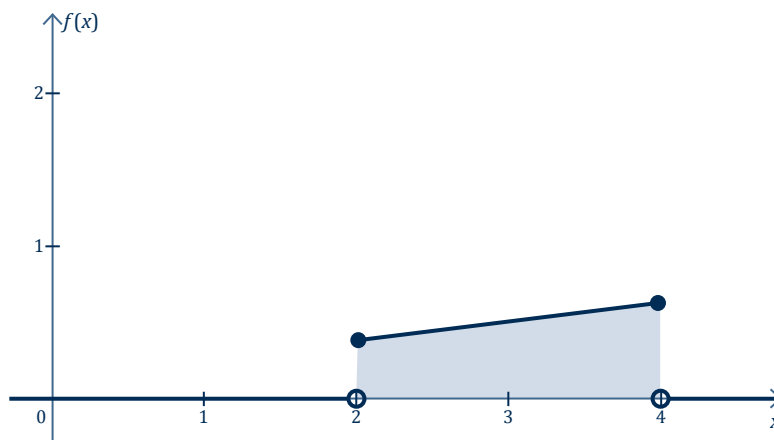
x	0	1	2	3	4
$P(X=x)$	0.035	0.078	0.193	c	0.360

Find $P(X \geq 2)$.

4. Prove that $P(X = k) = \frac{2k}{n(n+1)}$ for $k = 1, 2, \dots, n$ is a valid probability mass function of the discrete random variable X .
5. Suppose that $P(X = k) = c \left(\frac{2}{3} \right)^k$ for $k = 1, 2, 3, \dots$ serves as the probability mass function of the discrete random variable X .
 - a. Find the value of c .
 - b. What is the probability that X is an even number?
6. The following figure shows the graph of a function $f(x)$. Prove that $f(x)$ can serve as the probability density function of a continuous random variable X . Write down the PDF of X .

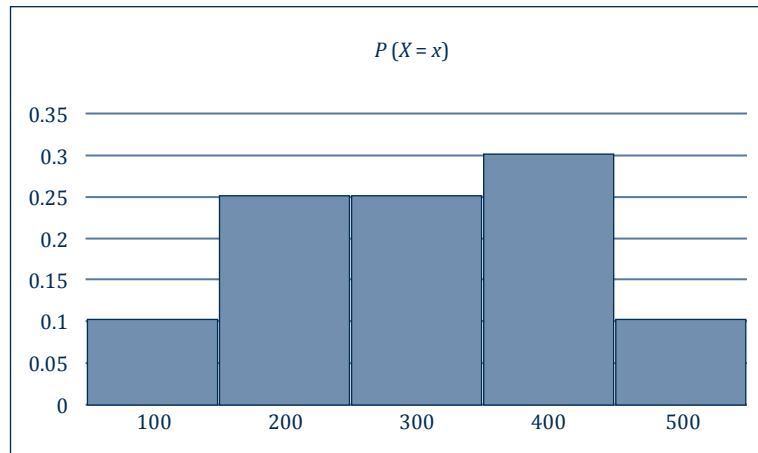


7. The graph of $f(x) = \begin{cases} \frac{1}{10}(x+2) & 2 \leq x \leq 4 \\ 0 & \text{elsewhere} \end{cases}$ is given in the following figure.



- a. Use the formula for the area of a trapezoid $A = \frac{(h_1 + h_2)}{2} b$, where $h_1 + h_2$ is the sum of the heights of the parallel sides and b is the length of the base, to prove that $f(x)$ can serve as the probability density function of a continuous random variable X .
 - b. Find $P(2.5 < X < 3.2)$.
8. Find the value of c so that $f(x) = \begin{cases} \frac{1}{9} & -2 \leq x \leq c \\ 0 & \text{elsewhere} \end{cases}$ can serve as the probability density function of the continuous random variable X . What is $P(X > 0)$?
9. In a famous quote by US President John F. Kennedy,
- “Ask not what your country can do for you, ask what you can do for your country.”
- Let the random variable X be the number of letters in a word chosen at random from this famous quote. Write down the probability mass function of X in tabular form.
10. In the experiment of tossing a green die and a red die, let the random variable X be the larger of the two numbers obtained. Determine the probability mass function of X and sketch its probability histogram.
11. From a standard deck of cards, you draw 3 cards with replacement. Let the random variable X be the number of spades among the 3 cards. Determine the probability mass function of X .

12. A box contains four 5-peso coins and two 10-peso coins. Three coins are drawn without replacement from the box. Determine the probability mass function of X , which is the total monetary value of the 3 coins.
13. A small rent-a-car business has 10 sedans numbered 1 to 10. Three of these sedans are selected at random and dispatched to the airport to convey passengers. Determine the probability mass function of X , which is the smallest number among the 3 sedans. Write your PMF both in tabular and formula forms.
14. Four crime suspects, which include Hanz and Mitch, are called for a police line-up. The random variable X is the number of persons standing between Hanz and Mitch in the line-up. Find the probability mass function of X .
15. The following figure is the histogram for the probability mass function (PMF) of the discrete random variable X .
- Write down the PMF of X .
 - Determine the expected value of X .
 - Determine the variance of X .



16. An urn contains 4 red chips, 4 white chips, and 4 blue chips. Suppose you draw 2 chips from the urn without replacement. For every red chip drawn you win ₱3, for every white chip drawn you win ₱2, but for every blue chip drawn you lose ₱1.
- Determine the probability mass function of X , which is the amount of your winnings.
 - Find the expected value of X . Is this a fair game?

17. A child is asking for money from his parents. His father offers him to flip a fair coin. If it comes up head, he gives his son ₱200, but if it comes up tails, his son gets nothing. His mother offers a roll of a single die, and will give him an amount equivalent to 30 times whatever number on the topmost face that will appear.
- Which offer has a larger expected value?
 - Should the son choose the offer with the larger expected value? Explain.
18. Suppose that four fair dice are tossed simultaneously. The random variable X is the number of different topmost faces that occur. The possible values of X are 1, 2, 3, and 4.
- Determine the probability mass function of X .
 - Find the expected value of X .
 - Find the variance of X .
19. If you have ₱100,000 that you want to invest, which investment option would you prefer? Explain.

Investment Option X	Probability
Total Loss	0.10
Breakeven	0.25
100% gain	0.35
200% gain	0.20
500% gain	0.10

Investment Option Y	Probability
Total Loss	0.05
Breakeven	0.35
50% gain	0.50
100% gain	0.10

Make a table such as this for both investment options to facilitate your computation:

	A	B	C	D
1	INVESTMENT OPTION X			
2	Outcome	Payoff	Probability	Expected Value
3	Total Loss	-100,000	0.10	
4	Breakeven	0	0.25	
5	100% gain	100,000	0.35	
6	200% gain	200,000	0.20	
7	500% gain	500,000	0.10	
8	TOTAL		1.00	

20. It costs ₱10 to play Suertres Lotto. A player chooses a three-digit number between 000 and 999 inclusive. If the player's chosen number matches the selected number for the day, the player wins ₱4,500. Let the random variable X be the player's net winnings in a game of Suertres Lotto.
- Write down the probability mass function of X .
 - What is the expected value of X ?
 - How much should the player pay to play if this game is to be fair?

21. The owner of a bakeshop knows that the number of pecan pies she can sell on any given day is a random variable having the following probability mass function.

k	0	1	2	3	4	5
$P(X = k)$	0.05	0.05	0.20	0.35	0.25	0.1

A pie costs ₱600 to make and sells at ₱700 each. Pies can only be sold on the day they are made, and any unsold pie is discarded. Find the expected profit for a day on which she bakes

- one pie
- two pies
- three pies
- four pies
- five pies

How many pies should she bake in order to maximize her expected profit?

Hint: Profit = Sales – Cost

22. For each of the following situations, specify if the binomial, hypergeometric, Poisson, geometric, or negative binomial distribution is to be used.
- It is known that at a domestic airport, 15% of the flights are delayed. You want to know the probability with which if 20 flights are chosen at random, fewer than four flights are delayed.
 - A tutor receives, on average, 18.2 e-mails from students per day. You want to compute the probability of the tutor receiving at least 9 e-mails in a day.
 - A shipment of 30 computers contains 3 that are defective. You want to know the probability that if a sample of 5 computers is taken from the shipment, none of them is defective.
 - Suppose that it is known that one in a hundred births in a country is a twin birth. You want to know the probability that you must observe at least six births before a twin birth occurs.
 - You shuffle a standard deck of cards and draw cards with replacement. You want to know the probability that you need at least 5 draws to draw 3 aces.
 - During a semester, a university's information technology office received 20 service requests for software installation, of which 14 were statistical software and 6 were graphics software. A sample of 5 of these requests was taken, and you want to know the probability that exactly 4 were for statistical software.
 - You roll a pair of dice and you want to know the probability that you need at least 4 rolls to get the first double 6's.

- h. Medical records indicate that among patients suffering from a certain disease caused by a vicious virus, about 0.5% die of it. You want to know the probability that among 25 randomly selected patients afflicted with this disease, exactly one of them will die.
 - i. The number of unique visitors to a certain website averages 12 per day. You want to know that probability that in a given day, the website gets at least 8 unique visitors.
23. Nick is a popular basketball player in a university. Based on his performance record, he makes a successful attempt from the foul line with probability 0.75. On a particular practice day, Nick attempts to shoot from the foul line 20 times.
- a. What is the random variable in this experiment? What is its probability distribution?
 - b. What is the probability that Nick will be able to shoot from the foul line in exactly half of the total number of attempts?
 - c. What is the probability that he will make at least 16 successful shoots from the foul line?
 - d. What is the expected number of successful shoots from the foul line?
 - e. What is the standard deviation of the number of successful shoots from the foul line?
24. An urn contains 45 balls numbered 1 to 45. Suppose that five balls are drawn at random, one at a time, with replacement. What is the probability that on exactly two of those draws, the balls have numbers that are relatively prime to 45? Recall that two integers r and s are relatively prime if 1 is their only common positive divisor. Thus, 4 and 3 are relatively prime, but 4 and 6 are not.
25. An electrical machine has rotors that operate independently of each other and are in good working condition with probability p . Suppose that the machine will function properly if at least half of its rotors is in good working condition.
- a. If $p = \frac{3}{4}$, which among a 5-rotor electrical machine or a 3-rotor electrical machine has a better chance of functioning properly?
 - b. Answer part (a) if $p = \frac{1}{3}$.
26. A committee consisting of 7 members must be formed from 10 doctors and 15 lawyers.
- a. What is the probability that the committee has exactly 4 doctors?
 - b. What is the probability that the committee will have more lawyers than doctors?
27. Among 20 applicants for the job of a call-center agent, 8 have bachelor's degrees. If the employer calls three of the job applicants randomly for an interview, what is the probability that at least one applicant has a bachelor's degree?

28. The jewelry collection of Madam Bizeau has 5 genuine diamond rings and 9 cubic zirconia rings, which look very similar in appearance to a real diamond. A thief, who is unknowledgeable in such distinctions, plans to make off with Madam's collection. However, the burglar alarm at Madam's mansion has sounded off, and the thief has to be contented with stashing away 6 rings from the collection. What is the probability the thief gets not a single genuine diamond ring?
29. Statistics from the Philippine Atmospheric, Geophysical, and Astronomical Administration (PAGASA) indicate that that an average of 19 tropical cyclones or storms enter the Philippine Area of Responsibility (PAR) in a year. If we assume that the number of storms that enter the PAR every year follows a Poisson distribution, find the probability that
- exactly 12 storms will enter the PAR next year.
 - anywhere from 15 to 18 storms will enter the PAR next year.
 - at least one typhoon will enter the PAR in a month.
30. The number of phone calls arriving at a call center is believed to follow a Poisson distribution with a mean rate of $\lambda = 5$ per minute. Find the probability that
- exactly six phone calls occur in a given minute.
 - at least two phone calls occur in half a minute.
31. A green die and a red die are tossed simultaneously until a sum of 7 spots is observed.
- What is the probability that at least 5 tosses are required?
 - What is the expected number of tosses needed to observe the sum of 7 spots?
 - What is the variance of the number of tosses needed to attain the sum of 7 spots?
32. The probability that an applicant for driver's license will pass the road test on any given attempt is 0.80. What is the probability that an applicant for driver's license will finally pass the road test on the fifth attempt?
33. Nick is a popular basketball player in a university. Based on his performance record, he makes a successful attempt from the foul line with probability 0.75. On a particular practice day, Nick attempts to shoot from the foul line.
- What is the probability that he makes his first successful shot on the third attempt?
 - Nick wants to be able to make ten successful shots. What is the probability that fifteen attempts are needed?
34. A cosmetics salesperson has a 25% chance of being received into a home and making a sale during door-to-door visits. Suppose that each home she visits is an independent trial. What is the probability that she requires at most ten homes to achieve her fourth success?

Chapter 3

The Normal Distribution



If you observe the students in your school, you will probably notice that a few are short, a few are rather tall, and most of them have “average” height. The same is true if you also observe their weight. Moreover, if you ask your classmates about their final grades in a particular subject, it is likely that the bulk of them will have an “average” mark of C, fewer students will have marks of B and D, and an even smaller percentage of students will have marks of A and F. In other words, for a certain characteristic, many would be clustered on the group with average characteristic, whereas fewer would be clustered on the group of those with extreme characteristic. These sets of data from the previous scenarios lead to a distribution that resembles a bell-shaped curve, and such a distribution is called a *normal distribution*. In this chapter, you will learn about the normal distribution and its properties as well as how to compute probabilities using the normal table. You will also learn the applications of normal distribution in real-life situations.

Lesson 1

Introduction to the Normal Distribution

Learning Outcomes

- At the end of this lesson, you should be able to
 - illustrate the normal distribution and its characteristics; and
 - describe and construct a normal curve.

Introduction

The normal distribution is perhaps the most widely used and known probability distribution. In many ways, it is considered as the cornerstone of modern statistics. It describes many sets of quantitative data such as some of the human features which include height, weight, and IQ level. Living things in nature have characteristics that can also be modeled with the normal distribution such as the life span of insects and the growth of crops. Because of its many applications, the normal distribution is regarded as the most important probability distribution.

Tradition has it that initial studies of the normal distribution were made by the French mathematician Abraham de Moivre (1667–1754), when he determined with impressive accuracy that the binomial distribution converged to the normal distribution as a limit. Later on in scientific studies, errors in repeated measurements of objects were noticed to have occurred in regular patterns that could be approximated closely by continuous curves. This curve was referred to as the *normal curve of error* or the *Gaussian distribution*, named after German mathematician Karl Friedrich Gauss (1777–1855). It was Gauss who derived the equation of the normal probability distribution function. To some extent, some credit for discovering the normal distribution is also given to the French mathematician Pierre Simon de Laplace (1749–1827). Laplace worked on the normal distribution independently but around the same time as Gauss.



Karl Friedrich Gauss
(1777-1855)

Definition

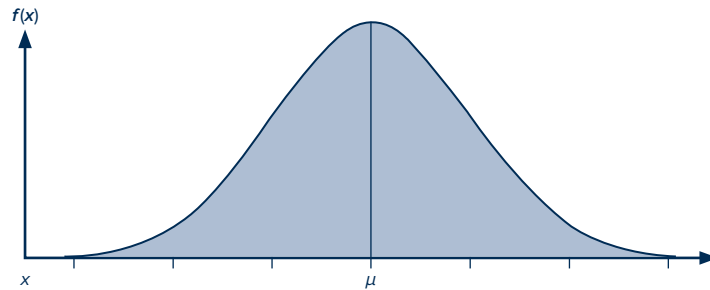
A continuous random variable X has a **normal distribution** with a mean μ and standard deviation σ if its probability distribution function is of the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

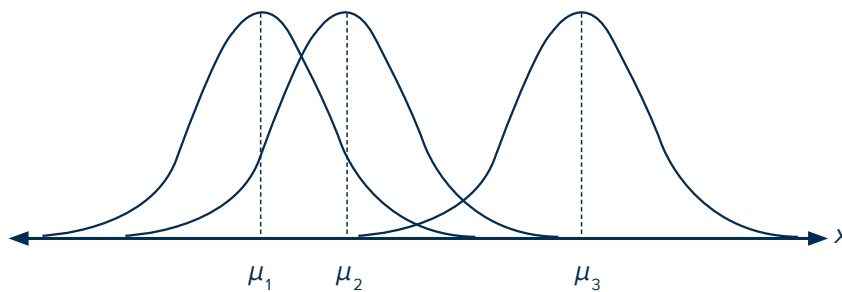
where $X, \mu \in \mathbb{R}$ and $\sigma > 0$.

The mean and the standard deviation are called the *parameters* of the normal distribution which specify the form of the normal distribution. The mean μ is a location parameter whereas the standard deviation σ is a scale parameter.

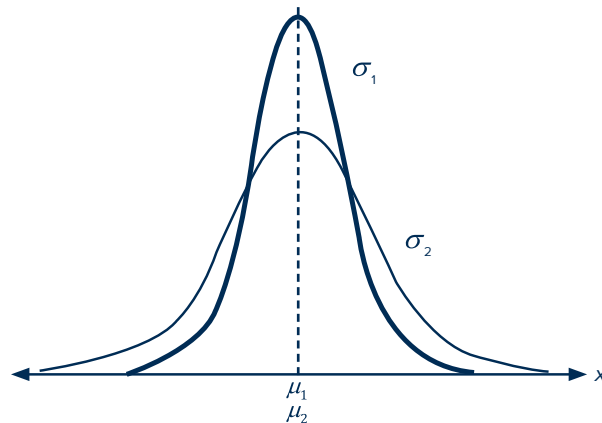
The graph of the normal distribution, also called the normal curve, is shown below.



The **normal curve** is shaped like the cross section of a bell and is centered at the mean. It gives the location of the normal curve. Shown in the next figure are three identical-looking normal curves with equal standard deviations ($\sigma_1 = \sigma_2 = \sigma_3$) but positioned on different points on the x -axis. Here, we have $\mu_1 < \mu_2 < \mu_3$.



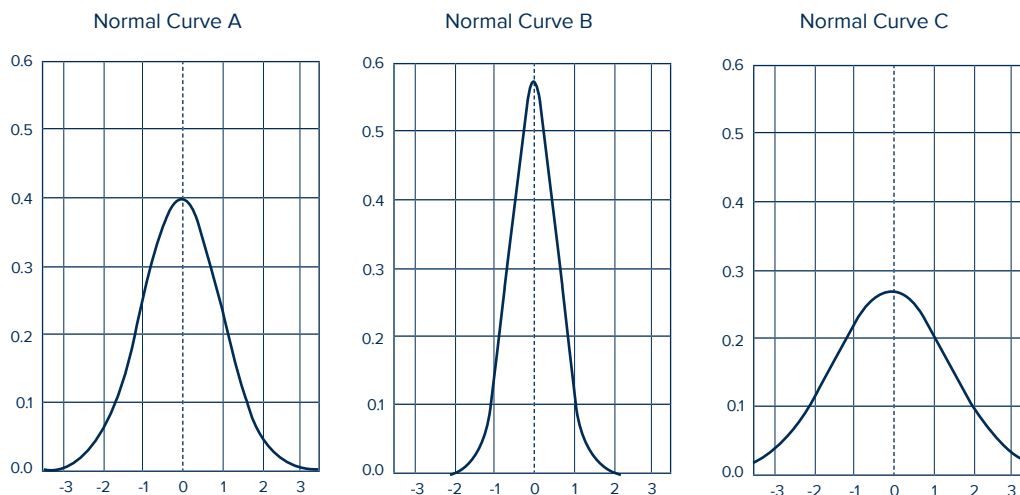
In the figure below, two nonidentical normal curves are superimposed on each other. Here, $\mu_1 = \mu_2$ and $\sigma_1 < \sigma_2$.



The standard deviation is a scale parameter that determines the dispersion in the normal distribution. The greater the value of the standard deviation, the wider the curve of the distribution.

Example 1

Compare the following normal curves whose means are equal to 0 but with different standard deviations. The standard deviations of curves A, B, and C are 1, 0.7, and 1.5, respectively.



Since the mean of each curve is 0, they are all centered at 0 as seen above. On the other hand, comparing their standard deviations, we have $\sigma_B < \sigma_A < \sigma_C$. Thus, curve B is narrower than curve A and curve C is wider than curve A.

Points to Remember

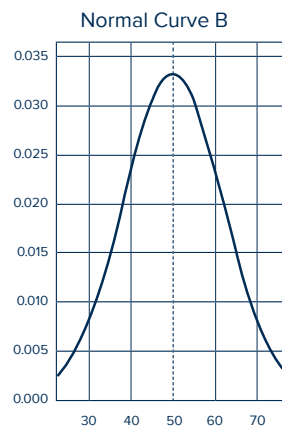
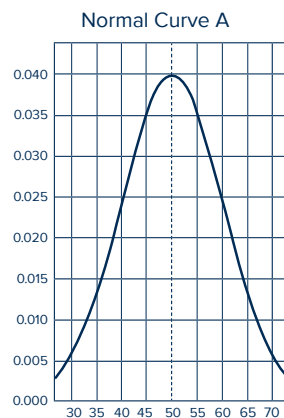
The normal curve exhibits the following properties:

1. The curve is symmetric about the mean μ .
2. The curve is unimodal. This means that it has a unique mode, which is at the point on the horizontal axis where the center of the curve lies.
3. The curve is asymptotic to the horizontal axis.
4. The total area under the curve, bounded by the horizontal axis, is equal to 1.

Let's Practice

Write True if the statement is *always* true. Otherwise, write False.

- _____ 1. The normal distribution is a discrete probability distribution.
- _____ 2. The mean, median, and mode of a normal distribution are all equal.
- _____ 3. In the normal curve, the highest point occurs at the standard deviation.
- _____ 4. If the mean of the normal distribution is negative, then the standard deviation must also be negative.
- _____ 5. If two normal curves have equal means but different standard deviations, then the curve with a smaller standard deviation is narrower and more peaked than the other curve.
- _____ 6. If two normal curves have different means, then the curve with a larger mean is located to the right of the other curve.
- _____ 7. Based on the figures below, normal curve A has a larger standard deviation than normal curve B.



Lesson 2

The Standard Normal Distribution

Learning Outcomes

- At the end of this lesson, you should be able to
 - identify regions under the normal curve corresponding to a standard normal value; and
 - compute probabilities and percentiles involving the standard normal distribution.

Introduction

It was mentioned in the previous lesson that the parameters of the normal distribution are its mean and standard deviation, which define the form of the normal distribution. Thus, every unique pair of μ and σ corresponds to a unique normal distribution. In this lesson, you will learn a special case of the normal distribution called the *standard normal distribution*.

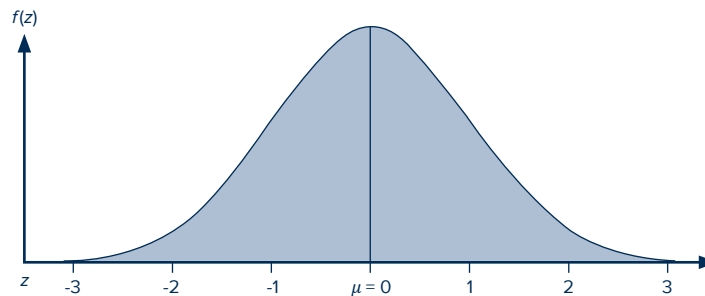
Definition 1

The **standard normal distribution** is a normal distribution whose mean is 0 and whose standard deviation is 1. A random variable Z that has a standard normal distribution has a probability distribution function given by

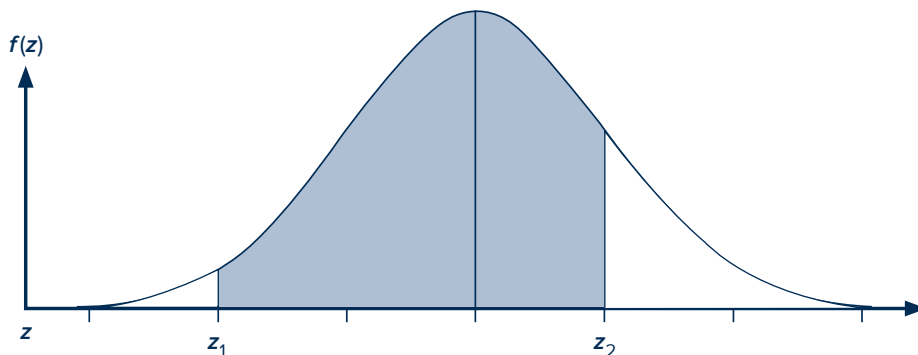
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

where $z \in \mathbb{R}$.

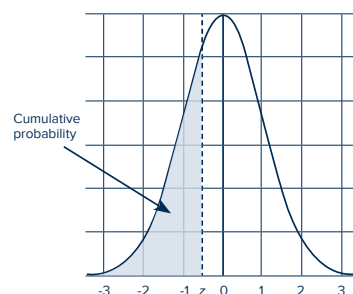
The graph of the standard normal distribution is shown below.



To calculate probabilities involving the normal distribution, we must calculate areas under the curve. From the example, to find the probability that Z lies in the interval z_1 to z_2 , we have to find the area under the standard normal Z -curve bounded by the two ordinates $z = z_1$ and $z = z_2$ as shown below.



The cumulative left-tail area of a value z for the standard normal distribution is denoted by $\Phi(z) = P(Z < z)$. A portion of the tabulated cumulative left-tail areas is shown in the next table. The first column of the Z -table denotes the first two digits of the number z . The first row of the Z -table denotes the hundredths (second decimal) place of the number z . The intersection of the row and the column is the cumulative left-tail area of z . The complete table is found in *Appendix B*.



Entries in this table give the area under the curve to the left of the z -value.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545

Example 1

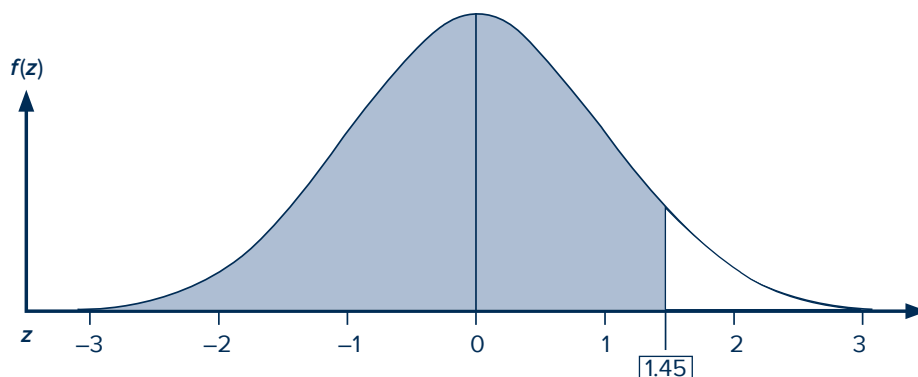
Given the standard normal distribution, find $\Phi(1.45)$.

Solution:

From the Z-table, we locate the first two digits, 1.4, on the first column and the second decimal place, 0.05, on the first row. Then their intersection is the cumulative left-tail area of 1.45 on the Z-curve. Thus, $\Phi(1.45) = P(Z < 1.45) = 0.9265$ as seen in the table below.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545

Shown below is the cumulative left-tail area of 1.45 under the Z curve, which is equal to 0.9265.



Example 2

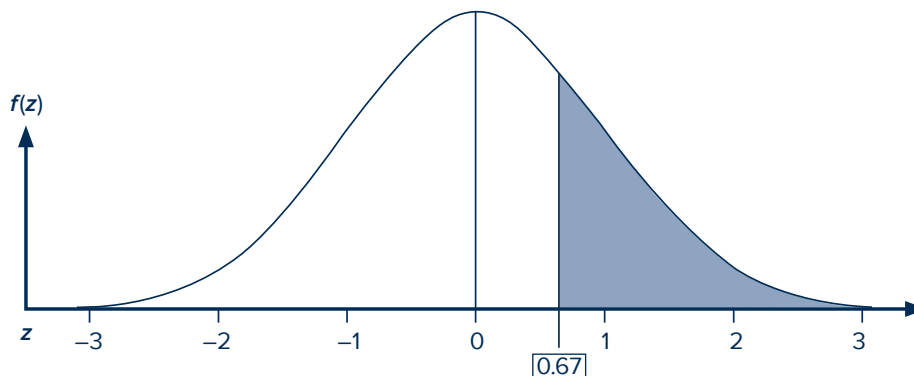
Given the standard normal distribution, find $P(Z > 0.67)$.

Solution:

The unknown value is an example of upper-tail probability for the Z-distribution. First, from the Z-table, we locate the first two digits, 0.6, on the first column and the second decimal place, 0.07, on the first row. Then their intersection is the cumulative left-tail area of 0.67 on the Z-curve. Thus, $\Phi(0.67) = P(Z < 0.67) = 0.7486$. In the figure below, the shaded region under the curve corresponds to $P(Z > 0.67)$ while the unshaded part corresponds

to $\Phi(0.67)$ which is equal to 0.7486. Recall that the total area under the normal curve is 1. Thus, to find $P(Z > 0.67)$, we have

$$P(Z > 0.67) = 1 - \Phi(0.67) = 1 - 0.7486 = 0.2514.$$

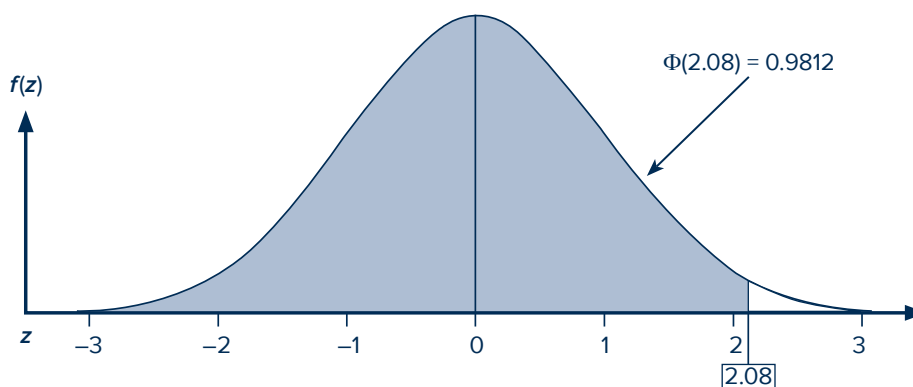


Example 3

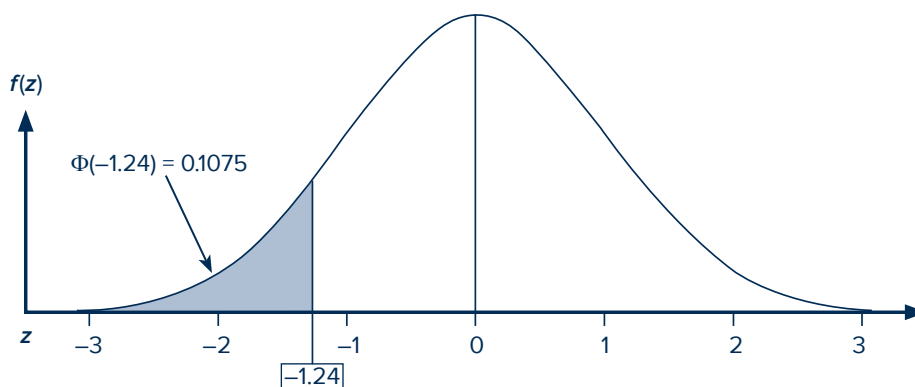
Given the standard normal distribution, find $P(-1.24 < Z < 2.08)$.

Solution:

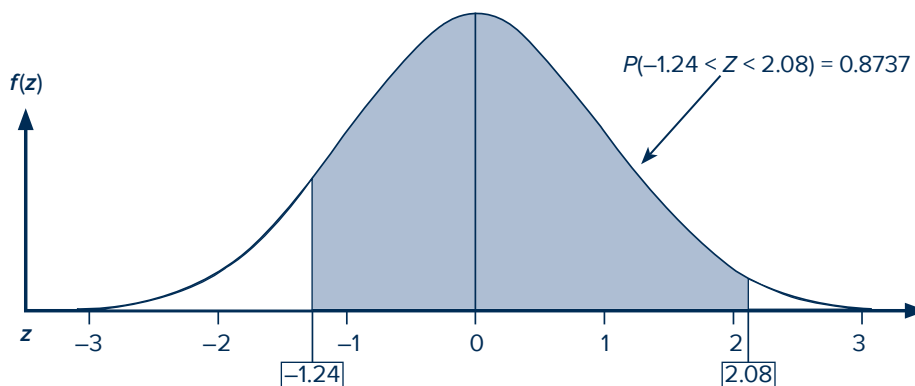
This is an example of in-between probability for the Z-distribution. First, we locate the cumulative left-tail area of the bigger number, 2.08 on the Z-curve, and this is the probability $\Phi(2.08) = P(Z < 2.08) = 0.9812$, shown in the figure below.



Next, we locate the cumulative left-tail area of the smaller number, -1.24, and this probability is $\Phi(-1.24) = P(Z < -1.24) = 0.1075$, shown in the figure below.



Finally, we subtract 0.1075 from 0.9812. Therefore, the desired probability is $P(-1.24 < Z < 2.08) = \Phi(2.08) - \Phi(-1.24) = 0.9812 - 0.1075 = 0.8737$. This is illustrated below.



Example 4

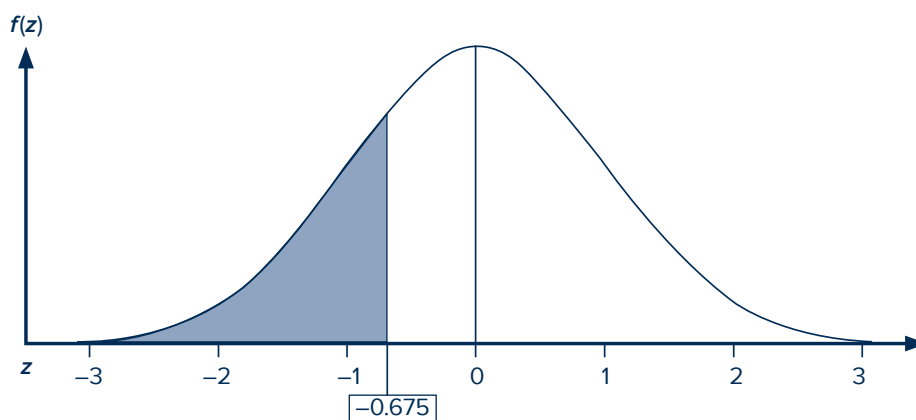
Find the first quartile of the standard normal distribution.

Solution:

The first quartile of a distribution is a number such that one-fourth or 25% of the values in the distribution are less than or equal to this number. The first quartile is also called the **lower quartile, Q_1** or the **25th percentile, P_{25}** .

The lower quartile of the standard normal distribution is a number z such that $\Phi(z) = P(Z < z) = 0.25$. This cumulative left-tail area does not appear exactly in the Z -table. For convenience, we choose the z -value that corresponds to the tabular value closest to the specified probability. It is a common practice that if a given probability falls approximately in the middle of two tabular values, the average of the two z -values corresponding to the two tabular values is used. In the case of the first quartile of the Z -distribution, which corresponds to a probability of 0.25, the tabular value falls between 0.2483 and 0.2514. These correspond to the Z -values -0.68 and -0.67 respectively. Then we get the average of -0.68 and -0.67 as the first quartile of the Z -distribution, which is -0.675 . Thus, Q_1 or $P_{25} = -0.675$.

The shaded region in the curve below is 0.25, which corresponds to the left-tail area of Q_1 or P_{25} .



Definition 2

The k th percentile of a random variable X , denoted by P_k , is defined as the smallest value c that satisfies the condition that

$$P(X \leq c) \geq \frac{k}{100} \text{ or simply } P(X \leq c) = \frac{k}{100}.$$

Points to Remember

When using the standard normal Z-distribution table, one must take note of the following:

1. The tabular values correspond to the cumulative left-tail area of z -value in the distribution.
2. The cumulative left-tail area of z , denoted by $\Phi(z)$, is the intersection of the row-value corresponding to the first 2 digits of z , and the column-value corresponding to the second decimal place of z .
3. The Z-distribution is a continuous probability distribution; therefore, the total area under the Z-curve is 1. To find the upper-tailed area of a z -value, use the formula $P(Z > z) = 1 - \Phi(z)$.
4. To find in-between probabilities of the form $P(z_1 < Z < z_2)$, use the formula $P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1)$.
5. To compute tail-end probabilities of the form $P(Z < z_1)$ or $P(Z > z_2)$, simply add these two areas as $\Phi(z_1) + 1 - \Phi(z_2)$.

Let's Practice

I. Write True if the statement is *always* true. Otherwise, write False.

- _____ 1. The mean of the standard normal distribution is 0.
- _____ 2. The standard deviation of the standard normal distribution can be any positive real number.
- _____ 3. For the standard normal distribution, $P(Z < 0) = 0.5$.
- _____ 4. For the standard normal distribution, $P(Z > 2.12) = P(Z < -2.12)$
- _____ 5. For the standard normal distribution, the value of k so that $P(Z < k) = 0.1401$ is -1.08 .
- _____ 6. For the standard normal distribution, the value of k so that $P(-k < Z < k) = 0.754$ is 2.16 .

II. Given the normal Z-distribution, find the following and provide a sketch of the bounded area.

1. the area below -0.71
2. the area below -1.65
3. the area below 3.14
4. the area below 0.82
5. the area between -1.06 and 3.10
6. the area between 0.75 and 2.69
7. the area between -2.33 and -0.56
8. the area below -2.51 or above 1.09
9. the area below 1.30 or above 3.03
10. the Z-score so that the area to the left of the Z-score is 0.4090
11. the Z-score so that the area to the right of the Z-score is 0.8888
12. the 90th percentile, P_{90}
13. the value of z so that the area between 0 and z is 0.1480

Lesson 3

Areas under the Normal Curve

Learning Outcomes

- At the end of this lesson, you should be able to
 - transform a normal random variable into a standard normal variable; and
 - find areas under the normal curve.

Introduction

It would be a very tedious task to construct tables of all normal distributions—one for each different combination of the mean μ and the standard deviation σ . With this concern, you will learn in this lesson how to transform all of the observations of any normal distribution with specified values of μ and σ into the standard normal distribution, Z .

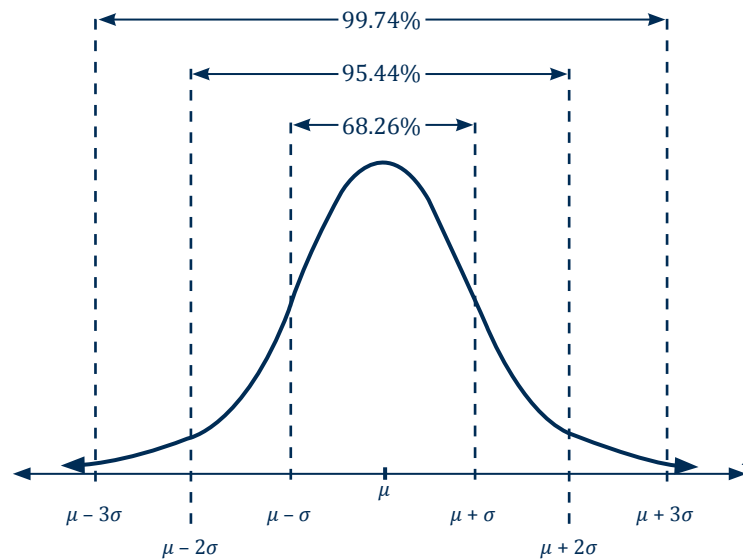
Definition

Let the random variable X be normally distributed with mean μ and standard deviation σ . The **z-score** of a value x is defined as

$$z = \frac{x - \mu}{\sigma}.$$

The formula converts any x -value of a normal distribution into a value in the standard normal Z -distribution. The z -score of a number x determines the number of standard deviations that x is above or below the mean. If the z -score of x is negative, then the value of x is less than the mean. If the z -score of x is positive, then the value of x is greater than the mean. If the z -score of x is zero, then x equals the mean. Thus, the z -score transformation formula enables the conversion of the distance of any x -value from its mean in terms of standard deviation units.

The **empirical rule** is a widely-used rule of thumb for cases involving normally distributed values. It specifies that in any normal distribution, approximately 68.26% of the values lie within one standard deviation away from the mean; 95.44% of the values lie within two standard deviations away from the mean; and 99.74% of the values lie within three standard deviations away from the mean. The empirical rule is illustrated in the next figure.



Suppose that the random variable X is normally distributed with mean μ and standard deviation σ . The probability that X assumes a value within 1 standard deviation from the mean, or $P(\mu - \sigma < X < \mu + \sigma)$, is equal to the area under the standard normal curve corresponding to $P(z_1 < Z < z_2)$ where z_1 and z_2 are the z-scores of $\mu - \sigma$ and $\mu + \sigma$, respectively.

$$z_1 = \frac{(\mu - \sigma) - \mu}{\sigma} = -1 \quad \text{and} \quad z_2 = \frac{(\mu + \sigma) - \mu}{\sigma} = 1$$

Therefore,

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= P(-1 < Z < 1) \\ &= \Phi(1) - \Phi(-1) \\ &= 0.8413 - 0.1587 \\ &= 0.6826 \text{ or } 68.26\%. \end{aligned}$$

The probabilities that X assumes a value within two or three standard deviations from the mean are computed similarly.

Points to Remember

1. The **empirical rule**, which applies to normally distributed values, specifies that approximately 68.26% of the values lie within the interval $\mu \pm \sigma$; 95.44% of the values lie within the interval $\mu \pm 2\sigma$; and 99.74% of the values lie within the interval $\mu \pm 3\sigma$.
2. A value that has a z-score lower than -3 or larger than 3 is considered an **outlier**. An **outlier** is an observation that is unusually different (higher or lower) from the rest of the observations.

Example

Given that the normally distributed random variable X has a mean of 40 and a standard deviation of 5, find the area below 39.1.

Solution:

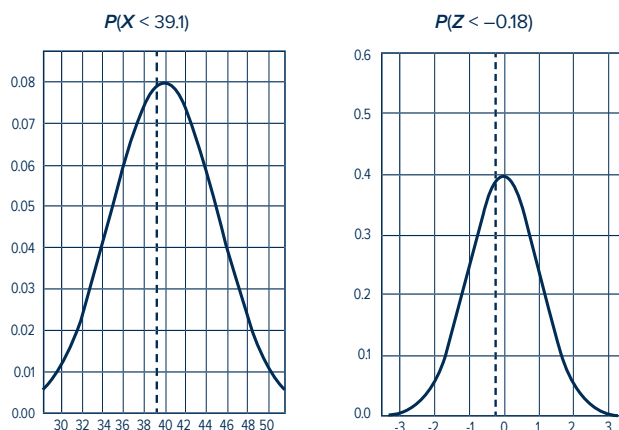
It is given that $\mu = 40$ and $\sigma = 5$. As such, the z-score of 39.1 is

$$z = \frac{x - \mu}{\sigma} = \frac{39.1 - 40}{5} = -0.18.$$

The required probability is $P(X < 39.1)$. Therefore,

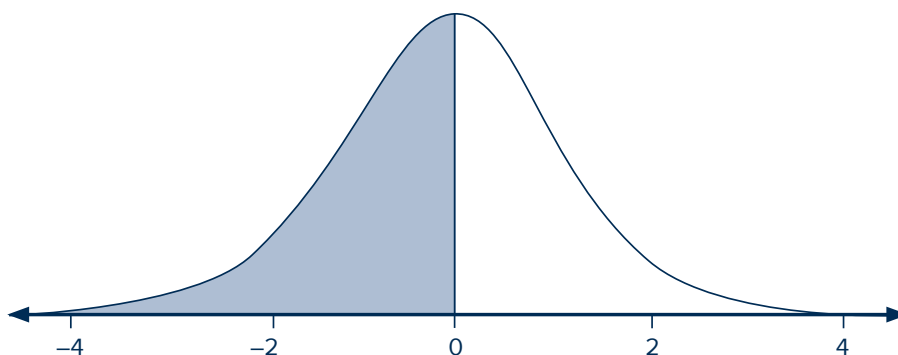
$$\begin{aligned} P(X < 39.1) &= P\left(Z < \frac{39.1 - 40}{5}\right) \\ &= P(Z < -0.18) \\ &= \Phi(-0.18) \\ &= 0.4286 \end{aligned}$$

The area below 39.1 on the normal X -curve is equivalent to the area below -0.18 on the standard normal Z -curve as shown in the following figures:



As a prelude to the next lesson, we may think of the above example as a way to model the weight in kilograms of Grade 7 students in Junior High School, which we may assume to be normally distributed with a mean of 40 kilograms and a standard deviation of 5 kilograms. As such, $P(X < 39.1)$ is the probability of randomly selecting a Grade 7 student whose weight is under 39.1 kilograms. If the mean of 40 kilograms is transformed into a z-score, then

$$z = \frac{40 - 40}{5} = 0$$



This means that $P(X < 40) = \Phi(0) = 0.5$. Thus, there is a 50% chance of randomly selecting a Grade 7 student whose weight is under (or over) 40 kilograms.

Let's Practice

I. Write True if the statement is *always* true. Otherwise, write False.

- _____ 1. If the incomes of school teachers in a city are such that there are more teachers with low incomes and very few with high incomes, then the incomes cannot be assumed to have a normal distribution.
- _____ 2. Given a normal distribution with a mean of 18 and a standard deviation of 4, then approximately 68% of the values of the distribution are between 14 and 22.
- _____ 3. Given a normal distribution where $\mu = 65$ and $\sigma = 2.5$, then approximately 95% of the values of the distribution are between 60 and 70.
- _____ 4. Given a normal distribution where $\mu = 65$ and $\sigma = 2.5$, $P(X > 72) = 0.9974$.
- _____ 5. The 75th percentile of the standard normal distribution has a corresponding Z score of approximately 0.675.

II. Given that the random variable X has a normal distribution with $\mu = 60$ and $\sigma = 10$, find the following:

1. $P(X < 48.4)$
2. $P(X > 81.6)$
3. $P(59.5 < X < 73.2)$
4. the 95th percentile, P_{95}
5. the upper quartile, Q_3 (also the 75th percentile)
6. two values x_1 and x_2 such that $P(x_1 < X < x_2)$ has a symmetrically centered area of 0.9000
7. $P(X < 52.6)$ or $P(X > 84.8)$

III. Given that the random variable X has a normal distribution where $\mu = 3.25$ and $\sigma = 0.6$, find the following:

1. the area between 2.7 and 3.92
2. the area above 2.65
3. the area below 4.10
4. the 48th percentile, P_{48}
5. the value of k such that $P(X > k) = 0.1335$

IV. Bernadette has obtained measurements on the errors, X , in her scientific experiment, where it is believed that such errors (can be \pm) follow a normal distribution with a mean of $\mu = 0.0005$ with a standard deviation of $\sigma = 0.10$. Find $P(|X| < 0.001)$.

Lesson 4

Applications of the Normal Distribution

Learning Outcome

- At the end of this lesson, you should be able to apply the concepts of normal distribution in real-life situations.

Introduction

In this lesson, you will be exposed to problems that can be modeled by normal distribution, and you will learn how to solve such problems using the standard normal Z-distribution.

Example 1

A coffee-vending machine is set to dispense amounts of coffee per cup that follows a normal distribution with a mean of 200 mL and a standard deviation of 10 mL. Let the random variable X be the amount of coffee (in mL) per cup.

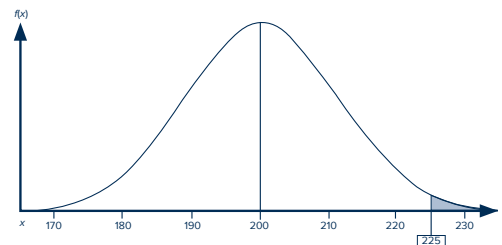


- What is the probability that a randomly selected cup will contain more than 225 mL of coffee as dispensed by the machine?

Solution:

The problem required us to find $P(X > 225)$ which corresponds to the shaded region in the figure below. Then we have

$$\begin{aligned} P(X > 225) &= 1 - P(X < 225) \\ &= 1 - P\left(Z < \frac{225 - 200}{10}\right) \\ &= 1 - \Phi(2.5) \\ &= 1 - 0.9938 \\ &= 0.0062. \end{aligned}$$

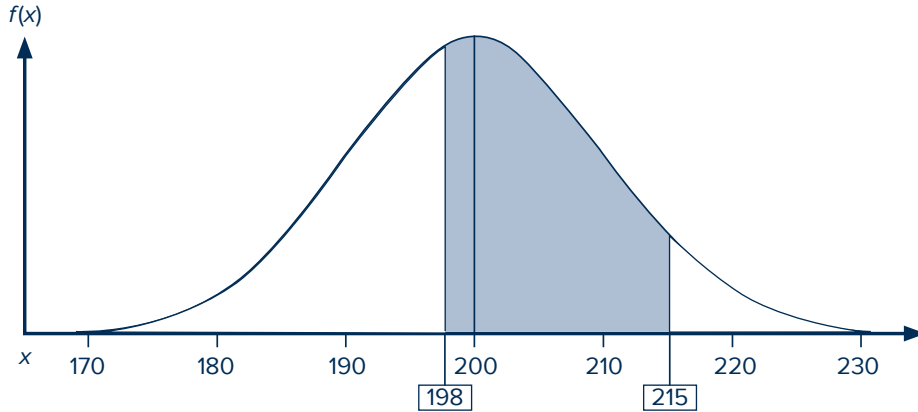


Therefore, there is a 0.0062 probability that a cup will contain more than 225 mL of coffee, as dispensed by the machine.

2. What proportion of the cups will contain anywhere between 198 and 215 mL of coffee?

Solution:

The problem required us to find $P(198 < X < 215)$ which corresponds to the shaded region in the figure below.



Then we have

$$\begin{aligned}
 P(198 < X < 215) &= P(X < 215) - P(X < 198) \\
 &= P\left(Z < \frac{215 - 200}{10}\right) - P\left(Z < \frac{198 - 200}{10}\right) \\
 &= \Phi(1.5) - \Phi(-0.2) \\
 &= 0.9332 - 0.4207 \\
 &= 0.5125.
 \end{aligned}$$

Therefore, 51.25% of the cups will contain anywhere between 198 and 215 mL of coffee.

3. Above what value do we get the largest filled 9.01% of the cups of coffee?

Solution:

The problem required us to find the value of k so that $P(X > k) = 0.0901$. Since 0.0901 is an upper-tailed probability and the Z -table we have is for cumulative lower-tailed probability, we first compute $P(X < k) = 1 - P(X > k) = 1 - 0.0901 = 0.9099$. Hence, we have:

$$\begin{aligned}
 0.9099 &= P(X < k) \\
 &= P\left(Z < \frac{k - 200}{10}\right)
 \end{aligned}$$

We now find from the Z-table the z-value that corresponds to a cumulative left-tail area of 0.9099, which is $Z = 1.34$. To find the value of k , we equate 1.34 to its corresponding z-score, that is,

$$\begin{aligned} 1.34 &= \frac{k - 200}{10} \\ k &= (1.34)(10) + 200 \\ &= 213.4. \end{aligned}$$

Therefore, 9.01% of the cups contain at least 213.4 mL of coffee, as dispensed by the machine.

Example 2

The *Stanford-Binet Intelligence Scales* is a cognitive ability and intelligence test designed to determine the intellectual and developmental attributes or deficiencies in young children. Assume that IQ scores determined by the Stanford-Binet test are normally distributed with a mean of 100 and a standard deviation of 16.

Stanford-Binet Fifth Edition (SB5) classification

IQ Range ("deviation IQ")	IQ Classification
145–160	Very gifted or highly advanced
130–144	Gifted or very advanced
120–129	Superior
110–119	High average
90–109	Average
80–89	Low average
70–79	Borderline impaired or delayed
55–69	Mildly impaired or delayed
40–54	Moderately impaired or delayed

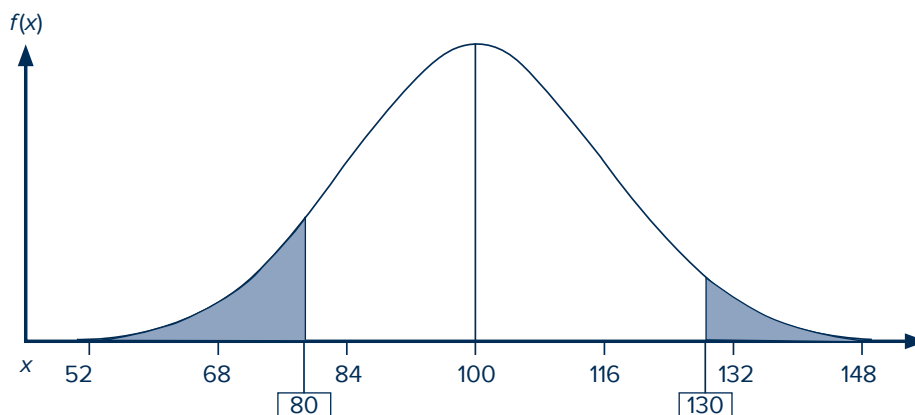
Adapted from Bain, S.K., and Allin, J. D. (2005). Book Review: *Stanford-Binet Intelligence Scales*, 5th ed. *Journal of Psychoeducational Assessment*, 23, 87-95.

- Children who are deemed as requiring special education are those whose IQ scores are under 80 points (classified as “impaired or delayed”) or those with an IQ score of at least 130 points (classified as “gifted or highly advanced”). If the Stanford-Binet test is administered to 1,500 students, about how many of them will be deemed as requiring special education?

Solution:

The problem requires us to find the two-tailed probability $P(X < 80) + P(X > 130)$, where X is the IQ score of a student in the Stanford-Binet test.

The two-tailed probability that we are looking for is shown in the figure below.



Since X is normally distributed with $\mu = 100$ and $\sigma = 16$, we have

$$\begin{aligned} P(X < 80) + P(X > 130) &= P\left(Z < \frac{80 - 100}{16}\right) + 1 - P\left(Z < \frac{130 - 100}{16}\right) \\ &= \Phi(-1.25) + 1 - \Phi(1.88) \\ &= 0.1056 + 1 - 0.9699 \\ &= 0.1357. \end{aligned}$$

Thus, 13.57% of the 1,500 students who took the Stanford-Binet test, or approximately 204 students, will be deemed as requiring special education.

2. A school formed an organization, *The Millennial Mind Mover*, with members whose IQ scores, based on the administered Stanford-Binet test, belong to the top 5%. What must be the minimum IQ score of a student for him or her to qualify as a member of the said organization?

Solution:

Let k be the minimum IQ score that the school organization required of its members. This value k is called the 95th percentile of the IQ scores, and at least 5% of the IQ scores must be at least as large as k . Thus, we are looking for k such that $P(X < k) = 0.95$.

Looking at the Z -table, the z -values with areas closest to 0.95 are 1.64 and 1.65, that is, $P(Z < 1.64) = 0.9495$ and $P(Z < 1.65) = 0.9505$. Then we get the average of the two z -values, which is 1.645. Solving for k , we have

$$\begin{aligned} 1.645 &= \frac{k - 100}{16} \\ k &= (1.645)(16) + 100 \\ &= 126.32. \end{aligned}$$

Thus, the minimum IQ score among the members of the school organization is 127. Approximately 75 of the 1,500 students who took the test are qualified as members.

3. Kendra is among the 1,500 students of the school. If Kendra's IQ score based on the Stanford-Binet test is 104, what is her percentile rank?

Solution:

The percentile rank of a Stanford-Binet test IQ score is the percentage of the scores in the distribution that is less than or equal to it. Since it is assumed that the IQ scores are normally distributed with a mean of 100 and a standard deviation of 16, Kendra's percentile rank can be determined by computing the z-score of 104 and finding the cumulative left-tail area of the resulting z-score. That is,

$$\begin{aligned} P(X < 104) &= P\left(Z < \frac{104 - 100}{16}\right) \\ &= \Phi(0.25) \\ &= 0.5987. \end{aligned}$$

Hence, Kendra's IQ score is approximately the 60th percentile. This implies that 59.87% of the students who took the Stanford-Binet test have a score of at most 104. Moreover, Kendra is classified as an "average" student based on the Stanford-Binet test.

Example 3

In a certain province, the set of recorded midday temperature readings for the month of April is believed to follow a normal distribution. It is known that 10.03% of recorded temperature readings were under 32.32°C, and that 33% were above 38.34°C. Determine the mean and standard deviation of this normal distribution.

Solution:

Let X denote the recorded temperature reading. From the problem, we know that $P(X < 32.32) = 0.1003$ and $P(X > 38.34) = 0.33$. Now, we can write two equations in terms of μ and σ of the normal distribution. Since $\Phi(-1.28) = 0.1003$ and $1 - \Phi(0.44) = 0.33$, we have

$$-1.28 = \frac{32.32 - \mu}{\sigma} \quad \text{and} \quad 0.44 = \frac{38.34 - \mu}{\sigma}.$$



First, we rewrite the equations as follows:

$$\mu - 1.28\sigma = 32.32$$

$$\mu + 0.44\sigma = 38.34$$

To find the values of μ and σ , we use the algebraic method of elimination.

Get the difference of the two equations to eliminate μ and solve for σ , that is,

$$\begin{array}{r} \mu - 1.28\sigma = 32.32 \\ (-) \mu + 0.44\sigma = 38.34 \\ \hline -1.72\sigma = -6.02 \\ \sigma = 3.5 \end{array}$$

Substitute the obtained value of σ into any of the equations to solve for μ , that is,

$$\begin{array}{r} \mu - 1.28\sigma = 32.32 \\ \mu - 1.28(3.5) = 32.32 \\ \mu = 36.8 \end{array}$$

Therefore, the set of recorded midday temperature readings has a mean of 36.8°C and a standard deviation of 3.5°C.

Example 4

Dr. Quintana, a college professor, uses standardized tests as an assessment tool. Based on experience, she believes that the scores of her students in the tests can be modeled by a normal distribution with a mean of 68 and a standard deviation of 10. Moreover, she believes that using the standardized tests, only 5% of her students will get A's, 25% will get B's, 46% will get C's, 14% will get D's, and the rest will get F's. If this is the case, what should be the cut-off score between the B's and the C's?

Solution:

Let X denote the test score of a student. Assume that X is normally distributed with $\mu = 68$ and $\sigma = 10$. Suppose we let k be the lowest B score which is the cut-off score between the B's and the C's. Since the top 5% of the students will get A's and 25% of the students will get B's, then $P(X > k) = 0.30$, or equivalently, $P(X < k) = 0.70$. From the Z-table, we see that $\Phi(0.52) = 0.6985$ and $\Phi(0.53) = 0.7019$. We get the z-value midway between 0.52 and 0.53, which is 0.525. Solving for k , we have

$$\begin{aligned} 0.525 &= \frac{k - 68}{10} \\ k &= 0.525(10) + 68 \\ k &= 73.25. \end{aligned}$$

Therefore, Dr. Quintana has set the lowest B score at 74 and the highest C score at 73.

Example 5

A certain university administers a standardized entrance examination to student applicants. Assume that the scores in this exam are normally distributed with a mean of 600 and a standard deviation of 50. Applicants must obtain a score higher than at least 80% of the examinees for them to be admitted to the university. Suppose that Leon, Marcia, Nancy, and Oscar took the examination.

1. What is the probability that none of the four applicants will be admitted to the university?

Solution:

Let L be the event that Leon qualifies for admission, and let M , N , and O be similarly defined. Then their respective complements, L' , M' , N' , and O' are the events that they do not qualify for admission.

Recall that to qualify for college admission, an applicant's score must be higher than the 80th percentile; that is, he or she must belong to the top 20% of examinees. Therefore, $P(L) = P(M) = P(N) = P(O) = 0.20$. The desired probability, that none of the four applicants qualify for admission, is $P(L' \cap M' \cap N' \cap O')$. It is reasonable to assume that these four applicants take the exam independently of each other. Hence, by the rule of statistical independence,

$$P(L' \cap M' \cap N' \cap O') = P(L') \cdot P(M') \cdot P(N') \cdot P(O') = (0.8)^4 = 0.4096.$$

2. Suppose that the exam scores of Leon, Marcia, Nancy, and Oscar are 628, 652, 636, and 660, respectively. Who among them qualify for admission?

Solution:

We transform their exam scores to z-scores and find their percentile ranks. If his or her percentile rank is at least as high as 80, then he or she qualifies for admission.

For Leon, $z = \frac{628 - 600}{50} = 0.56$ and $\Phi(0.56) = 0.7123$. This means that his percentile rank is 71.23; therefore, he is not qualified for admission.

For Marcia, $z = \frac{652 - 600}{50} = 1.04$ and $\Phi(1.04) = 0.8508$. This means that her percentile rank is 85.08; therefore, she is qualified for admission.

For Nancy, $z = \frac{636 - 600}{50} = 0.72$ and $\Phi(0.72) = 0.7642$. This means that her percentile rank is 76.42; therefore, she is not qualified for admission.

For Oscar, $z = \frac{660 - 600}{50} = 1.20$ and $\Phi(1.20) = 0.8849$. This means that his percentile rank is 88.49; therefore, he is qualified for admission.

Solve each problem.

1. Suppose that Glenda got a final grade of 84 in Physics and 89 in Chemistry. It is known that the final grades in Physics in her class are approximately normally distributed with a mean of 79 and a standard deviation of 8, whereas their final grades in Chemistry are also normally distributed with a mean of 82 and a standard deviation of 11. In which subject did Glenda have a more impressive academic performance?
2. The gasoline consumption of a certain model of an automobile is normally distributed with a mean of 12 kilometers per liter (km/L) and a standard deviation of 3 km/L.
 - a. What is the probability that an automobile of this model will consume between 11.35 and 14.1 km/L?
 - b. What is the probability that an automobile of this model will consume more than 19.6 km/L?
 - c. What is the probability that an automobile of this model will consume less than 15.9 km/L?
3. The life expectancy of a certain brand of television is normally distributed with a mean of 5 years and a standard deviation of 1.5 years.
 - a. What is the probability that a randomly selected television of this brand will be in working condition for more than seven years?
 - b. The company has a two-year warranty period on their products. What percentage of the televisions will still be in working condition after the warranty period?
 - c. Ninety percent of the televisions will have a life expectancy of at least how many years?
4. In a large pharmaceutical manufacturer, it is observed that the bottles of cough syrup marked 300 mL vary in content with a standard deviation of 5 mL. Assume the content of the bottles are normally distributed.
 - a. What percentage of the bottles of liquid cough syrup contains between 300.4 mL and 302.9 mL?
 - b. Ninety-nine percent of the bottles of cough syrup will not exceed what amount (in mL)?

5. In an annual marathon sponsored by a popular energy drink company, women in the 18–34 age group must run the official qualifying marathon in 4 hours and 30 minutes (270 minutes) or less in order to compete in the 42.195 km National Finals. Rhonda is 30 years old. Assume that the times at which she can run a marathon are normally distributed with a mean of 280 minutes and a standard deviation of 10 minutes.
 - a. If Rhonda participates in one marathon, what is the probability that her time will qualify her for the National Finals?
 - b. Above what time do we find the slowest 6% of Rhonda's marathon running?
 - c. Rhonda intends to run 3 marathons next year. Suppose her finishing times are independent from one marathon to another. Find the probability that she has at least one running completion time that qualifies her for the National Finals.
6. Keith owns an orchard where he grows carabao mangoes. The mangoes that are classified as medium-sized have a mean weight of 265 grams. Assume the weight of the mangoes to be normally distributed with a standard deviation of 12 grams.
 - a. What percentage of the medium-sized carabao mangoes weigh anywhere from 250 to 270 grams?
 - b. What percentage of the medium-sized carabao mangoes weigh under 255 grams?
 - c. What percentage of the medium-sized carabao mangoes weigh over 280 grams?
 - d. Below what weight in grams will you find the lightest 80% of all the medium-sized carabao mangoes?

Software Tutorial in MS Excel

Aside from looking at statistical tables to manually find probabilities, you can use the MS Excel to make your task easier and faster. Following are tutorials on how to use MS Excel in finding some of the statistical values of the discussed in this chapter.

1. Cumulative left-tail areas in the standard normal distribution may be obtained using the command "`=NORMSDIST(z)`." For example, to find $\Phi(-1.45) = P(Z < -1.45)$, we type "`=NORMSDIST(-1.45)`" on the command line and obtain the answer of 0.0735.

	A	B	C
1	0.073529	=NORMSDIST(-1.45)	

2. To find the corresponding z-value in the standard normal distribution that leaves a cumulative left-tail area of size a , we use the command "`=NORMSINV(a)`." For example, to find the 45th percentile of the Z-distribution, we type in "`=NORMSINV(0.45)`" on the command line, and obtain the answer of -0.1257.

	A	B	C
1	-0.12566	=NORMSINV(0.45)	

3. Probabilities involving a normal distribution with a specified value of μ and σ may be obtained using the command “=NORM.DIST(x, mean, standard_deviation, cumulative).” “Cumulative” is a logical command, and since we are getting cumulative left-tail probabilities, we set its value to 1 (TRUE). For example, if we have a normal distribution with a mean of 200 and a standard deviation of 10, to find $P(X < 225)$, we type in “=NORM.DIST”(225,200,10,1) on the command line, and obtain the answer of 0.9938.

	A	B	C	D
1	0.99379	=NORM.DIST(225,200,10,1)		

4. We can use the “=NORM.INV(k/100, mean, standard_deviation)” command to obtain the k th percentile of a normal distribution with a specified value of μ and σ . For example, to find the upper quartile (or P_{75} , the 75th percentile) of a normal distribution with a mean of 200 and a standard deviation of 10, we type “=NORM.INV(0.75,200,10)” on the command line, and obtain the answer of 206.749.

	A	B	C
1	206.7449	=NORM.INV(0.75,200,10)	

Chapter Review

- A continuous random variable X has a **normal distribution** with a mean μ and standard deviation σ if its probability distribution function is of the form $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ where $X, \mu \in \mathbb{R}$ and $\sigma > 0$.
- The mean and the standard deviation are called the *parameters* of the normal distribution which specify the form of the normal distribution. The mean μ is a *location parameter* whereas the standard deviation σ is a *scale parameter*.
- The graph of the normal distribution, also called the *normal curve*, is shaped like the cross section of a bell and is centered at the mean.
- The **standard normal distribution** is a normal distribution whose mean is 0 and whose standard deviation is 1. A random variable Z that has a standard normal distribution has a probability distribution function given by $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ where $z \in \mathbb{R}$.

- When using the standard normal Z-distribution table, one must take note of the following:
 1. The tabular values correspond to the cumulative left-tail area of z-value in the distribution.
 2. The cumulative left-tail area of z, denoted by $\Phi(z)$, is the intersection of the row-value corresponding to the first 2 digits of z, and the column-value corresponding to the second decimal place of z.
 3. The Z-distribution is a continuous probability distribution; therefore, the total area under the Z-curve is 1. To find the upper-tailed area of a z-value, use the formula $P(Z > z) = 1 - \Phi(z)$.
 4. To find in-between probabilities of the form $P(z_1 < Z < z_2)$, use the formula $P(z_1 < Z < z_2) = \Phi(z_2) - \Phi(z_1)$.
 5. To compute tail-end probabilities of the form $P(Z < z_1)$ or $P(Z > z_2)$, simply add these two areas as $\Phi(z_1) + 1 - \Phi(z_2)$.
- Let the random variable X be normally distributed with mean μ and standard deviation σ . The **z-score** of a value x is defined as

$$z = \frac{x - \mu}{\sigma}.$$

- The **empirical rule** specifies that in any normal distribution, approximately 68.26% of the values lie within one standard deviation away from the mean; 95.44% of the values lie within two standard deviations away from the mean; and 99.74% of the values lie within three standard deviations away from the mean.
- A value that has a z-score lower than -3 or larger than 3 is considered an outlier. An **outlier** is an observation that is unusually different (higher or lower) from the rest of the observations.

Chapter Performance Tasks

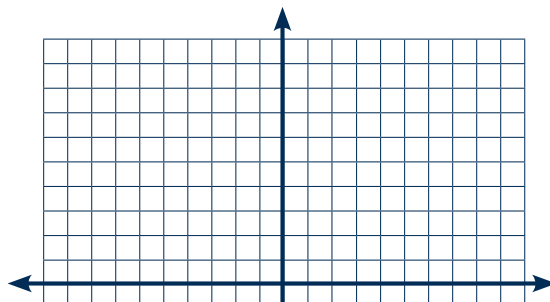
1. How Long in Mommy's Tummy

Imagine that you are a journalist for the science and mathematics section of your school's organ, and applications of the normal distribution have caught your attention. You would like to determine if the length of a woman's pregnancy can be modeled by the normal distribution. Collect data from every student in the class of at least 3 sections of your grade level. Tell each of the students to ask his or her mother for how long she carried him or her in her womb, that is, the



length of the mother's pregnancy in days. (It is expected that some mothers will just give an approximate answer as she may not remember the exact length of her pregnancy.) Have the students report their mothers' responses to you. Then, suppose that you will write a journal article for publication in the school organ where you will indicate the findings of your survey. Include in your article the following:

- a. A calculation of the mean and standard deviation of the data collected from students of at least 3 sections of your grade level. Draw a normal curve, marking the mean and ± 1 , ± 2 , and ± 3 standard deviations on a Cartesian plane template such as this one:



- b. A calculation of the percentage of students whose mothers' length of pregnancy is within 1, 2, and 3 standard deviations from the mean.
- c. Your observations and comments if the results in part (b) are consistent with the empirical rule, that is, 68.26% of the values lie within 1 standard deviation from the mean, 95.44% of the values lie within 2 standard deviations from the mean, and 99.74% of the values lie within 3 standard deviations from the mean.
- d. A report if there are any outliers in the data set.

2. My Favorite Filipino Athlete

Collect data from every member of at least 4 athletic teams in your school. Ask each team member to write down his or her height in centimeters. Then, ask five random team members to name their favorite Filipino athlete. Research the heights of the athletes named, also in centimeters. Afterwards, make a video presentation of your findings from the data collected from all the teams. Include the following:



- a. A calculation of the mean and standard deviation of the heights of all team members. Draw a normal curve, marking the mean and ± 1 , ± 2 , and ± 3 standard deviations from the mean.
- b. The pictures of the five favorite Filipino named and their respective heights in centimeters.
- c. The probabilities of a team member having a height that exceeds each of the five height measurements in part (b).
- d. A report if the heights of the athletes in part (b) are outliers.

Chapter Exercises

1. Given the normal Z-distribution, find the following and provide a sketch of the bounded area:
 - a. the area below -1.32
 - b. the area above -0.46
 - c. the area below 2.71
 - d. the area above 1.89
 - e. the area between -0.06 and 3.01
 - f. the z-score so that the area to the left of the z-score is 0.3050
 - g. the z-score so that the area to the right of the z-score is 0.5199
 - h. the 97th percentile, P_{97}
 - i. the value of z so that the area between 0 and z is 0.4750
2. Given that the random variable X has a normal distribution with $\mu = 95$ and $\sigma = 12$, find the following:
 - a. $P(X < 98)$
 - b. $P(X > 65)$
 - c. $P(85 < X < 100)$
 - d. the 90th percentile, P_{90}
 - e. the lower quartile, Q_1
 - f. two values x_1 and x_2 such that $P(x_1 < X < x_2)$ has a symmetrically centered area of 0.9500
3. The number of hours a person sleeps in a day may depend on several factors such as age and lifestyle. Suppose that the length of sleep follows a normal distribution with an average of 7 hours and a standard deviation of 1.25 hours.
 - a. What percentage of people sleep between 6.5 and 8 hours?
 - b. What is the probability that a person sleeps more than 8.5 hours?
 - c. What percentage of people have less than 5.75 hours of sleep?
 - d. Above what value do we find the longest 4.01% number of hours of sleep?
4. The mean cholesterol content of a certain brand of eggs is 215 mg with a standard deviation of 15 mg. Assume the cholesterol content of such eggs is normally distributed.
 - a. What is the probability that a randomly selected egg of this brand will contain under 206 mg of cholesterol?
 - b. What percentage of eggs will contain anywhere between 210 and 225 mg of cholesterol?
 - c. Below what value do we find the lowest 1.5% of cholesterol content in eggs?

5. In a certain study, it is found out that the time spent by children aged 2–5 years watching television in a week is normally distributed. Suppose that 2.28% of the children in this age group watch for more than 20 hours in a week, and 10.75% of them watch for less than 10.28 hours. Find the mean and standard deviation of this normal distribution.
6. In Trisha’s class, the mean and standard deviation of their grades in different subjects are shown in the table below. Assume that their grades in each subject are normally distributed.

Subject	Mean	Standard Deviation
Calculus	70.5	8.6
Chemistry	75.8	6.3
English	84.9	4.1
Philosophy	88.2	3.5

Trisha’s grades are as follows:

Calculus: 83

Chemistry: 79

English: 86

Philosophy: 90

- a. In which subject did Trisha perform best, relative to the other students’ grades? Explain.
- b. What was her percentile rank on the subject identified in item (a)?
7. The life span of a certain brand of microwave oven is normally distributed with a mean of 10 years and a standard deviation of 2 years.
- a. What percentage of this certain brand will fail during the first 3 years of use?
- b. If the manufacturer of this appliance is willing to replace only 3% of the malfunctioning microwave ovens, for how long should the warranty period be offered?
8. Suppose that the minimum speed on the expressway for cars is 60 kilometers per hour (kph) and the maximum speed is 100 kph. Cars with speed lower than the minimum and exceeding the maximum are apprehended. Suppose that the speed of cars on the expressway are normally distributed with a mean of 80 kph and a standard deviation of 10 kph. What percentage of cars traveling on the expressway will be apprehended for going under the minimum speed or for overspeeding?
9. Suppose that the random variable X is the score in a Physics exam of a randomly selected student from Professor Santos’s class. Based on his grading system (which is based on a normal curve), the students whose scores exceed 3 standard deviations from the mean will get A’s, and the students whose scores are below 2 standard deviations from the mean will

get F's. Moreover, scores in the range $[\mu + 2\sigma, \mu + 3\sigma]$ will get B+'s; scores in the range $[\mu + \sigma, \mu + 2\sigma]$ will get B's; scores in the range $[\mu, \mu + \sigma]$ will get C+'s; scores in the range $[\mu - \sigma, \mu]$ will get C's; and scores in the range $[\mu - 2\sigma, \mu - \sigma]$ will get D's.

- a. Determine the percentage of his students who will get A, B+, B, C+, C, D, and F.
 - b. Suppose that the average score was 65 with a standard deviation of 7, what is the cut-off score between D and F? between A and B+?
10. The time it takes to make a transaction from an automated teller machine (ATM) is said to be normally distributed with a mean of 60 seconds with a standard deviation of 12 seconds. If 3 clients made a transaction in the ATM independently of each other, what is the probability that at least one of them took more than 70 seconds?
 11. Among the employees of a downtown office are an accountant and an auditor. Each day, they commute from their suburban homes to the office independently of each other. For the accountant, the time it takes for the one-way trip is normally distributed with an average of 30 minutes with a standard deviation of 3.8 minutes, whereas for the auditor, the one-way trip is normally distributed with a mean of 34 minutes with a standard deviation of 4.5 minutes. Suppose that work at the office starts at 8:00 a.m., and both employees leave their homes at 7:25 a.m. What is the probability that both of them will be late for work?
 12. The scores on a nationwide mathematics aptitude exam are normally distributed with a mean $\mu = 80$ and standard deviation $\sigma = 12$. Suppose that 10 students who took the exam are selected at random. What is the probability that exactly half of those students had exam scores between 75 and 90?
 13. In each of the following cases, find the value of the constant c so that the given is the probability distribution function of a normally distributed random variable X . Identify the mean and variance of X in each case.
 - a. $f(x) = ce^{\frac{-x^2}{40} + \frac{3x}{10} - \frac{9}{10}}$
 - b. $f(x) = ce^{-2x^2 - 6x - \frac{9}{2}}$
 14. Raymond has one analog clock in his kitchen and another clock in the study room. Each clock requires the use of one AA (penlight) battery. He bought batteries from Company X and its competitor Company Y. He installed the Company X battery on the kitchen clock and the Company Y battery on the study room clock at the same time. Assume that the useful life of batteries from Company X is normally distributed with a mean of 5,000 hours and a standard deviation of 250 hours, whereas for Company Y, the useful life is normally distributed with a mean of 5,200 hours with a standard deviation of 300 hours. What is the probability that at least one of the clocks will run for at least 4,850 hours until the battery needs replacement?

Chapter 4

Sampling and Sampling Distributions



Quality control is an important part of any manufacturing process. It involves ensuring that all the components of the product and all steps of the production process conform to the highest standards of quality. For example, in producing smartphone batteries, battery life is tested whether it lasts as long as what is advertised by the manufacturer.

Despite the proliferation of machines and automation, there still remains some variation in the capacity of the produced battery. To verify whether a fully charged battery lasts a number of hours within a small range around the advertised value, *quality assurance specialists* take a *sample* of several batteries and subject them to rigorous tests. These tests allow the manufacturer to gauge whether the machines produce the batteries as programmed, or whether they need to be recalibrated.

In this chapter, you will learn several methods on how samples can be chosen. You will also be introduced to the related concepts of the *sampling distribution of the sample mean* and the *central limit theorem*, both of which will be crucial in understanding why such samples can be used to say something about the corresponding population.

Lesson 1

Random Sampling

Learning Outcomes

- At the end of this lesson, you should be able to
 - define and illustrate the different methods of obtaining a random sample; and
 - distinguish between parameter and statistic.

Introduction

The field of inferential statistics is mainly concerned with generalizations and predictions. For example, consider a survey trying to determine the public's preferred candidate for president. Based on the opinion of several people interviewed, the researcher might say that 30% of the votes in the coming election will go to a particular candidate.

In this example, we are using the value of some characteristic of a *sample* to make a statement about the *population* which may or may not be true. In this case, we are using the value of a sample *statistic* to infer something about the corresponding population *parameter*.

Definition 1

A **population** is the totality of items, things, or people under consideration. A **sample** is a subset of the population.

Any measurable characteristic of a population is called a **parameter**. Any measurable characteristic of a sample is called a **statistic**.

It would be ideal if a researcher would be able to study the entire population to determine the value of a population parameter or some characteristic of the population. However, this is generally not feasible. Instead, the researcher can usually study only a sample of a population.

Here are some of the reasons why a researcher would rather take a sample than study the entire population.

1. *It saves time and money.* During the months before an election, a researcher might be interested in determining what percentage of registered voters

would vote for a particular candidate. Since the population of registered voters is located all over the entire country, it would be too expensive and also time-consuming to attempt to ask the preference of every single member of the population.

2. *It may be physically impossible to get the entire population.* An ecologist might be interested in determining the average length of a certain type of fish in the Pacific Ocean. Clearly, due to the vastness of the ocean, it is physically impossible for him to gather up the entire population of fish to collect their sizes.
3. *Some tests are destructive in nature.* For a wine connoisseur to verify the quality of a Chardonnay wine from a vineyard, he cannot do this by tasting a small amount of wine from each bottle. If he does, there will be no more wine to be sold. It is enough for him to pick a sample of one or two bottles, and use it to infer the quality of the Chardonnay produced by the wine producer. A related situation occurs when testing the tensile strength of wires. In this case, the tests involve pulling the wire with a gradually increasing amount of force until it breaks. Clearly, one cannot do it for all the wires a company produces.
4. *The sample results are sufficient.* Even if we had the time and money for conducting a study of the entire population, the additional accuracy provided by studying the entire population is usually not significant. In fact, statistical theory shows that under certain assumptions, the results obtained from a sample would usually not differ substantially from that of the population.

When selecting a sample from a population, it is important that the sample is an appropriate representation of the population. That is, we wish that the sample is not *biased*, and does not systematically select certain elements more often than the others. One way to ensure this is to take a random sample from the relevant population.

There are several ways to take a random sample. The simplest and most widely used way is known as *simple random sampling*.

Definition 2

Simple random sampling is a selection of a subset of a population where each element has an equal chance of being selected.

To select a simple random sample, it is necessary to have a complete list of all the members of the population that we will be sampling from, known as the *sampling frame*. This ensures that every member of the population has a chance to be selected in our sample.

To illustrate simple random sampling, suppose that a school has a total of 200 students. A sample of 30 students is to be taken to determine the average number of hours a student spends playing computer games. One way to ensure that each student has an equal chance of being selected is to write the name of each of the 200 students on a slip of paper, and then place all the slips in a box. After mixing the slips thoroughly, a slip is drawn from the box. This will correspond to the first member of our sample. After discarding the slip drawn, we can repeat this process until the entire sample size of 30 is selected.

However, there are more convenient ways to obtain the sample. One common way is to first label the students from 001 to 200, and then choose a random student using a *table of random numbers*. Such a table can be found in *Appendix A*. This table typically contains five-digit numbers. Each of the digits of the numbers in this table was chosen so that each number from 0 to 9 has an equal chance of being selected. A portion of this table is given below.

	1	2	3	4	5
1	10480	15011	01536	02011	81647
	22368	46573	25595	85393	30995
	24130	48360	22527	97265	76393
	42167	93093	06243	61680	07856
5	37570	39975	81837	16656	06121
	77921	06907	11008	42751	27756
	99562	72905	56420	69994	98872
	96301	91977	05463	69994	98872
	89579	14342	63661	10281	17453
10	85475	69857	53342	53988	53060

For easier reference, the rows and the columns in the table have been labeled.

To use this table to select the sample of students, we begin by choosing an arbitrary starting point. For example, we could start from the number in the first row of the third column; that is, 01536. Since the students are labeled using three-digit numbers, we could select the student with the same first three digits as this number, which happens to be 015. We can then read downwards, skipping those numbers which are greater than 200. For example, looking at the first three digits of the numbers on the second and third rows, we have 255 and 225, which are both greater than 200. We simply skip these numbers. Continuing down the third column, we obtain the next three samples as students 062, 110, and 054.

Most statistical software packages also have commands to select a simple random sample. The next example uses MS Excel to select a sample.

Example 1

Listed below are the semifinalists in the 100-meter dash event of the 2012 Olympics, arranged alphabetically by their first name.

No.	Name	Country
1	Adam Gemili	GBR
2	Antoine Adams	SKN
3	Asafa Powell	JAM
4	Ben Youssef Meite	CIV
5	Bingtian Su	CHN
6	Churandy Martina	NED
7	Daniel Bailey	ANT
8	Derrick Atkins	BAH
9	Dwain Chambers	GBR
10	Gerald Phiri	ZAM
11	James Dasaolu	GBR
12	Jimmy Vicaut	FRA

No.	Name	Country
13	Justin Gatlin	USA
14	Justyn Warner	CAN
15	Kemar Hyman	CAY
16	Keston Bledman	TTO
17	Richard Thompson	TTO
18	Rondel Sorrillo	TTO
19	Ryan Bailey	USA
20	Ryota Yamagata	JPN
21	Suwaibou Sanneh	GAM
22	Tyson Gay	USA
23	Usain Bolt	JAM
24	Yohan Blake	JAM

Source: <https://www.olympic.org/london-2012/athletics/100m-men>, retrieved on 11 August 2016.

Suppose we wish to use MS Excel to select a simple random sample of four of these athletes to undergo random drug testing. One possible output is given below.

No.	Name	Country	Sample
1	Adam Gemili	GBR	14
2	Antoine Adams	SKN	23
3	Asafa Powell	JAM	17
4	Ben Youssef Meite	CIV	4
5	Bingtian Su	CHN	
6	Churandy Martina	NED	
7	Daniel Bailey	ANT	
8	Derrick Atkins	BAH	
9	Dwain Chambers	GBR	
10	Gerald Phiri	ZAM	
11	James Dasaolu	GBR	
12	Jimmy Vicaut	FRA	
13	Justin Gatlin	USA	
14	Justyn Warner	CAN	

Note: The software commands to produce the output above are given at the end of the chapter.

Take note that the sampling is done with replacement, which means that the same athlete may be chosen more than once in a sample. If this happens, we can just choose to disregard the duplicated athlete.

Based on the MS Excel output, we obtain the sample consisting of Justyn Warner (14), Usain Bolt (23), Richard Thompson (17), and Ben Youssef Meite (4).

An alternative to simple random sampling is taking a *systematic random sample*. It retains the simplicity and the ease of construction of a simple random sample.

Definition 3

In a **systematic random sampling**, a random starting point is selected, and then every k th member of the population is selected.

Here, the value of k is chosen by the researcher. In general, to calculate k , we divide the population size by the desired sample size. For example, when taking a systematic sample of size 4 for the population of size 24 in example 1, $k = \frac{24}{4} = 6$.

To select the first member of the population included in the systematic random sample, a simple random sample is selected from the first k members of the population. Then every k th member of the population after that will be part of the sample.

In the case of example 1, we first select an athlete at random among those numbered from 1 to 6. Assume that this is athlete number 3 (Asafa Powell). Then the other athletes included in the systematic random sample would be Dwain Chambers (9), Kemar Hyman (15), and Suwaibo Sanneh (21).

Before doing systematic random sampling, it is important to note the physical order of the population. If the physical order of the population is related to the characteristic being studied, systematic random sampling cannot be used. For example, if we wish to estimate the average height of students in a class, it cannot be done by selecting a systematic random sample of the students numbered by increasing height. In this case, the resulting sample would not be unbiased.

There are cases where a sample obtained by either simple or systematic random sampling might not cover a portion of our desired population. For example, when selecting a sample of patients who have checked in last week in a given hospital, we might wish to have both male and female patients in our sample. Taking a simple or systematic random sample does not guarantee that both genders are represented. In this case, taking a *stratified random sample* would be more appropriate.

Definition 4

Stratified random sampling is a selection of a simple random sample from each of a given number of subpopulations or **strata**.

The strata are based on the members' shared attributes or characteristics. For example, in an economic study, the population may be grouped by sex (male, female), economic class (A, B, C, D), or geographical location (Luzon, Visayas, Mindanao). In this case, the strata are the gender, the economic class, or the geographical location of the experimental units, respectively. The samples from each stratum are then pooled together to form the stratified random sample.

When doing stratified random sampling, it is customary to have the ratio of the elements from each of the strata in the sample to be the same as that of the population. For example, suppose a company would like to take a stratified random sample of size 10 of their employees by employment status (full-time or part-time). If the company has 40 full-time employees and 10 part-time employees (with a ratio of 4 : 1), then the sample must contain 8 full-time employees and 2 part-time employees (which also has a ratio of 4 : 1).

Another type of random sampling method is cluster sampling. It is usually employed to reduce the cost of sampling population elements spread out over a large geographic area.

Definition 5

Cluster sampling is a selection of clusters from the available clusters in the population. Each member of the selected clusters is then included in the sample.

Unlike stratified random sampling where each stratum consists of members of the population with similar characteristics, in cluster sampling, we wish for each cluster to be as varied as possible. For example, in an economic study where the desired population consists of the residents of Manila, we might define the clusters as the 897 barangays in Manila. A cluster sample could consist of *all* the residents of three of these barangays selected at random.

Note: Although stratified random sampling and cluster sampling both involve dividing the population into subgroups, these sampling methods are completely different. In stratified random sampling, we choose a sample from each subgroup (stratum). In contrast, in cluster sampling, we choose a sample of subgroups (clusters), and include in the sample all members of the selected subgroups.

Example 2

Identify the sampling method illustrated in each of the following:

1. Divide Quezon City into barangays and take a simple random sample from each barangay.
2. Divide Quezon City into city blocks, choose a simple random sample of 10 city blocks, and interview everyone who lives there.
3. Choose an entry at random from the phone book, and select every 50th number thereafter.

Answers:

1. This sample includes representatives from each barangay. Thus, this is an illustration of *stratified random sampling*, with the person's barangay of residence as the stratum.
2. This is an example of *cluster sampling*, where the clusters are the city blocks.
3. This illustrates 1-in-50 *systematic random sampling*.

Let's Practice

I. Write the letter that corresponds to your answer. Write X if your answer is not among the given choices.

- _____ 1. Which is the totality of items, things, or people under consideration?
- | | |
|--------------|---------------|
| a. sample | c. parameter |
| b. statistic | d. population |
- _____ 2. What is any measurable characteristic of a population?
- | | |
|--------------|---------------|
| a. sample | c. parameter |
| b. statistic | d. population |
- _____ 3. What is any measurable characteristic of a sample?
- | | |
|--------------|---------------|
| a. sample | c. parameter |
| b. statistic | d. population |

- _____ 4. Catherine Wong, the director of an industrial company, is concerned by a deteriorating sales trend. Specifically, the average number of customers is stable at 1,500, but they are purchasing less each year. She orders her staff to search for the cause of the downward trend by selecting a focus group (sample) of 40 industrial customers. One question asks the focus group to rate “Merchandise is delivered on time” on a scale of 1 to 5, with 1 meaning “never” and 5 meaning “always.” What measurable characteristic can be gathered on the response of the 40 customers?
- a. parameter
 - b. population
 - c. sample
 - d. statistic
- _____ 5. Which is most likely a population rather than a sample?
- a. Every fifth person who passes a certain intersection
 - b. Respondents to an online survey
 - c. The registered voters in Valenzuela City
 - d. The first 20 shoppers in a department store
- _____ 6. Which is *not* one of the reasons why we need to sample?
- a. It is usually cheaper to study a sample rather than the population.
 - b. It is usually impractical to study all the individuals in a population.
 - c. A sample usually provides more information about the parameter compared to a population.
 - d. A sample is less time-consuming to study than the entire population.
- _____ 7. What is the main purpose of stratified random sampling?
- a. It is to make sure that every member of the population has an equal chance of being selected.
 - b. It is to make sure that the cost of taking the sample is reduced.
 - c. It is to make sure that the sample proportionately represents individuals from different categories.
 - d. It is to make sure that sampling can be done even without a sampling frame.
- _____ 8. You wish to examine the effect of socioeconomic status on diet. To do this, you decide to randomly sample from all employees or patients at a local hospital for the study. However, it is known that more individuals in this group are from a higher, rather than lower, socioeconomic status. What would be the most appropriate sampling method to be used?
- a. cluster sampling
 - b. simple random sampling
 - c. stratified random sampling
 - d. systematic random sampling

- _____ 9. Mr. Cruz wanted to find out if there is a significant difference on the students' performance in mathematics and a foreign language course. To do this, he decided to obtain a list of all the students in each of the eight sections of their school's batch. He then randomly selected two sections, and collected the mathematics and foreign language grades of each student in these two sections. What sampling method is being used by Mr. Cruz?
- cluster sampling
 - simple random sampling
 - stratified random sampling
 - systematic random sampling
- _____ 10. Mrs. Nimfa is deciding which of her 40 students to call for recitation for tomorrow's class. She decides to put all her students' names in a bag, shakes it well, and then draws five names at random. What sampling method is used by Mrs. Nimfa?
- cluster sampling
 - simple random sampling
 - stratified random sampling
 - systematic random sampling

II. Analyze and solve each problem.

1. A consumer group wishes to estimate the average time a person waits in line at the branches of the McBee fastfood chain in Makati City. To do this, they observe the waiting time in some of the chain's branches. After some preliminary research, they were able to obtain the following list of McBee branches in Makati:

No.	Location
01	Makati Cinema Square
02	Guadalupe Commercial Center
03	Glorietta 3
04	Greenbelt 1
05	Jupiter Sreet
06	Glorietta 1
07	Ayala Avenue
08	Madrigal Building
09	Paseo Center
10	SGV Building
11	Reposo

No.	Location
12	Landmark
13	PRC Sta. Ana
14	Jaka Building
15	Power Plant Mall
16	SM Makati
17	Valero
18	People Support
19	Kingswood
20	Evangelista
21	SM Cyberzone 2

- a. Lester puts the numbers 01 to 21 in a bag. Without looking, he draws the numbers 11, 19, and 03. Which branches will be part of his sample?
 - b. To select a random sample of three branches, he decides to use the last two digits of the numbers of the 11th column in *Appendix A*. If he begins on the seventh row, which branches will be included in his sample?
 - c. Use MS Excel to select a simple random sample of three branches. You will need to input the assigned numbers (01 to 30).
 - d. If Lester decides to use systematic random sampling to get his sample of size 3, and he draws the number 06 as his first McBee branch, which other branches will be included in his sample?
2. The following is a list of hospitals in Pasig City. You are interested in estimating the average number of full-time and part-time nurses employed in the hospitals.

No.	Name of Hospital
1	Alfonso Specialist Hospital
2	Glen Eagles Healthcare
3	Health Solutions Corporation
4	Javillonar Clinic and Hospital
5	John F. Cotton Hospital
6	Mary Immaculate Hospital
7	Medicomm Pacific, Inc.
8	Medcor Pasig Hospital and Medical Center
9	Metro Psych Facility
10	Mission Hospital
11	Mother Regina Hospital
12	Pasig City General Hospital
13	Pasig Doctors Medical Center
14	Pasig Medical and Maternity Hospital Foundation, Inc.
15	Rizal Medical Center
16	Sabater General Hospital
17	Salve Regina General Hospital
18	Saint Therese Hospital
19	The Medical City
20	Wellness Pro, Inc.

- a. A sample of five hospitals is to be randomly selected. The random numbers are 16, 08, 47, 61, 15, 07, and 04. Which hospitals will be included in the sample?
 - b. Use the first two digits of the third column in *Appendix A* to develop a sample of five hospitals.
 - c. A sample is to consist of every fourth hospital. If the number 02 is selected as starting point, which hospitals will be included in the sample?
3. Teresa, a senior high school teacher, wishes to study the learning habits of her students. The following is a list of her students, along with their current strand:

No.	Student	Strand
1	Abad, Teresa	ABM
2	Almeda, Clarissa	HUMSS
3	Baluyot, Kristel	HUMSS
4	Barcelona, Alfie	STEM
5	Camus, Gabby	HUMSS
6	Canteras, Christopher	ABM
7	Choi, Selenia	ABM
8	Chua, Andrew	ABM
9	Coo, Benedict	HUMSS
10	Del Rosario, Joseph	ABM
11	Dimacali, Anne	HUMSS
12	Galvez, Kristina	ABM
13	Garcia, Francesca	ABM
14	Gunigundo, Abel	HUMSS
15	Hagad, Moira	HUMSS

No.	Student	Strand
16	Lasam, Isabel	HUMSS
17	Manzano, Lawrence	HUMSS
18	Marte, Eleonor	STEM
19	Medina, Ralph	STEM
20	Nera, Adrian	ABM
21	Palacio, Sharleen	HUMSS
22	Quimpo, Melissa	ABM
23	Santos, Maria	STEM
24	Santos, Lorenzo	HUMSS
25	Sarmenta, Enrique	ABM
26	Sevilla, Louella	ABM
27	Tan, Johann	STEM
28	Veloso, Luis	STEM
29	Villar, Doris	HUMSS
30	Zamora, Antonio	ABM

- a. Teresa wishes to select a random sample of five students to interview. The random numbers are 19, 91, 70, 18, 97, 26, 92, 06, 27. Which students will be selected?
- b. Use MS Excel to select a sample of five students among Teresa's students.
- c. Explain how you would obtain a stratified sample of students based on the strand of the student.

Lesson 2

The Sampling Distribution of the Sample Mean

Learning Outcomes

- At the end of this lesson, you should be able to
 - identify the sampling distribution of the sample mean; and
 - find the mean and variance of the sampling distribution of the sample mean.

Introduction

In the previous lesson, you have learned several ways to select a random sample from a population. In general, our objective in taking such a sample is to use the corresponding sample statistic to estimate the population parameter. As we are only using a subset of a population to say something about the entire population, an important question to ask is, "Can we really do this?" That is, can a sample statistic truly represent the parameter's value?

In the succeeding lessons, we will see that the answer to this question is in the affirmative. However, under certain conditions, we will be able to say much about the *sampling error* or the difference between the statistic and the true value of the parameter.

Definition 1

The difference between the value of a sample statistic and the corresponding population parameter is called the **sampling error**.

Ideally, we would wish for our sampling error to be zero. However, due to the random nature of sampling, this will generally not be the case. For example, when taking a sample of size 5 to estimate the mean age of employees in a company, it is clear that the sample mean will fluctuate, depending on which employees are chosen. Similarly, a survey firm which takes a sample of 200 citizens from a city to determine what percentage agrees with a new city ordinance will generally not be able to obtain the exact proportion based from the sample.

While we cannot ensure that the sampling error of a sample is zero, we can say something about the typical amount of sampling error obtained. To do this, we can examine the resulting values of the statistic. In particular, since the value of a statistic is

dependent on the random sample taken from the population, we can think of a sample statistic as a *random variable*. This means that statistics such as sample means or sample proportions must also have probability distributions.

Definition 2

The probability distribution of a statistic is called a **sampling distribution**. The standard deviation of the sampling distribution is called the **standard error** of the statistic.

The sampling distribution of a statistic is the probability distribution when all possible samples of size n are drawn from the population. Like any other probability distribution, we can then compute the mean and standard deviation of this distribution.

While any sample statistic has its own sampling distribution, the most useful sampling distribution is that of the sample mean, which we shall study in the next example. It turns out that the mean and the standard deviation of the sampling distribution is closely dependent on the sample size drawn from the population, as well as the mean and variance of the parent population.

Example 1

Suppose that we take a sample of size $n = 2$, with replacement, from the discrete uniform population with values 0, 1, 2, 3. That is, each of these four numbers has a probability $p = \frac{1}{4}$ of being selected. Then there are a total of $4^2 = 16$ possible samples. Computing the mean \bar{x} of each sample, we obtain the following table:

No.	Sample	\bar{x}	No.	Sample	\bar{x}
1	0, 0	$\frac{0+0}{2}=0$	5	1, 0	$\frac{1+0}{2}=0.5$
2	0, 1	$\frac{0+1}{2}=0.5$	6	1, 1	$\frac{1+1}{2}=1$
3	0, 2	$\frac{0+2}{2}=1$	7	1, 2	$\frac{1+2}{2}=1.5$
4	0, 3	$\frac{0+3}{2}=1.5$	8	1, 3	$\frac{1+3}{2}=2$

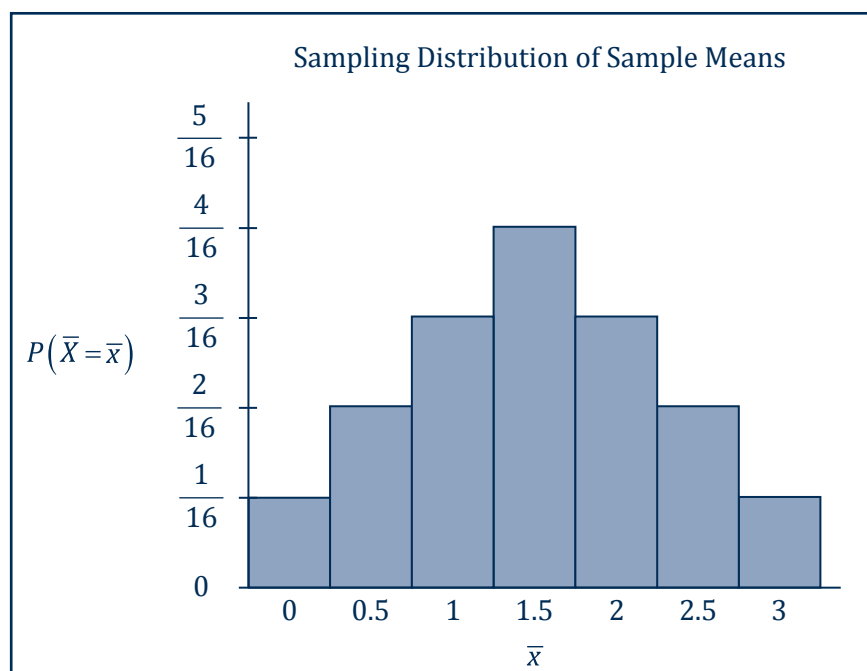
9	2, 0	$\frac{2+0}{2}=1$
10	2, 1	$\frac{2+1}{2}=1.5$
11	2, 2	$\frac{2+2}{2}=2$
12	2, 3	$\frac{2+3}{2}=2.5$

13	3, 0	$\frac{3+0}{2}=1.5$
14	3, 1	$\frac{3+1}{2}=2$
15	3, 2	$\frac{3+2}{2}=2.5$
16	3, 3	$\frac{3+3}{2}=3$

Notice that depending on the sample chosen, the resulting sample mean \bar{x} varies from 0 to 3. Furthermore, these values of \bar{x} do not occur equally often. Since \bar{X} is a random variable, it has a probability distribution. Using the fact that each sample has an equal probability $\frac{1}{16}$ of being chosen, we obtain the following sampling distribution for \bar{X} :

\bar{x}	0	0.5	1	1.5	2	2.5	3
$P(\bar{X} = \bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

The following graph summarizes the results above.



Example 2

Consider again the discrete uniform population with values 0, 1, 2, 3. In this case, the value obtained when taking a sample of size 1 is a random variable X which has a probability of $\frac{1}{4}$ to assume each of the values 0, 1, 2, 3. That is, the probability mass function of X is given by the following table:

x	0	1	2	3
$P(X=x)$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$

Then, recalling the definitions of the mean and variance of a random variable from chapter 2, lesson 4, we can compute the mean and the variance of X as follows:

$$\begin{aligned}
 \mu &= \sum x \cdot p(x) \\
 &= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{4}\right) \\
 &= \frac{3}{2}
 \end{aligned}$$

$$\begin{aligned}
 \sigma^2 &= \sum (x - \mu)^2 \cdot p(x) \\
 &= \left(0 - \frac{3}{2}\right)^2 \left(\frac{1}{4}\right) + \left(1 - \frac{3}{2}\right)^2 \left(\frac{1}{4}\right) + \left(2 - \frac{3}{2}\right)^2 \left(\frac{1}{4}\right) + \left(3 - \frac{3}{2}\right)^2 \left(\frac{1}{4}\right) \\
 &= \frac{5}{4}
 \end{aligned}$$

Example 3

In example 1, we obtained the sampling distribution of the sample mean \bar{X} for samples of size 2 selected from this population which is as follows:

\bar{x}	0	0.5	1	1.5	2	2.5	3
$P(\bar{X} = \bar{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{3}{16}$	$\frac{4}{16}$	$\frac{3}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Taking the mean and variance of the sample mean \bar{x} , we have

$$\begin{aligned}\mu_{\bar{x}} &= \sum \bar{x} \cdot p(\bar{x}) \\ &= 0\left(\frac{1}{16}\right) + 0.5\left(\frac{2}{16}\right) + 1\left(\frac{3}{16}\right) + 1.5\left(\frac{4}{16}\right) + 2\left(\frac{3}{16}\right) + 2.5\left(\frac{2}{16}\right) + 3\left(\frac{1}{16}\right) \\ &= \frac{3}{2}\end{aligned}$$

$$\begin{aligned}\sigma_{\bar{x}}^2 &= \sum (\bar{x} - \mu_{\bar{x}})^2 \cdot p(\bar{x}) \\ &= \left(0 - \frac{3}{2}\right)^2 \left(\frac{1}{16}\right) + \left(0.5 - \frac{3}{2}\right)^2 \left(\frac{2}{16}\right) + \left(1 - \frac{3}{2}\right)^2 \left(\frac{3}{16}\right) + \left(1.5 - \frac{3}{2}\right)^2 \left(\frac{4}{16}\right) \\ &\quad + \left(2 - \frac{3}{2}\right)^2 \left(\frac{3}{16}\right) + \left(2.5 - \frac{3}{2}\right)^2 \left(\frac{2}{16}\right) + \left(3 - \frac{3}{2}\right)^2 \left(\frac{1}{16}\right) \\ &= \frac{5}{8}\end{aligned}$$

Notice that the mean of \bar{X} is the same as the mean of X . On the other hand, the variance of \bar{X} is smaller than that of X , with $\sigma_{\bar{x}}^2 = \frac{5}{8} = \frac{\frac{5}{4}}{2} = \frac{\sigma^2}{2} = \frac{\sigma^2}{n}$, where $n = 2$ is the sample size.

The result in the previous example can be generalized to the case where we are taking samples of size n , with replacement, from any population with finite mean and finite variance.

Generalization: When taking samples (with replacement) of size n from any population with finite mean μ and finite variance σ^2 ,

1. the mean of the sampling distribution of the sample mean is equal to the population mean, that is, $\mu_{\bar{x}} = \mu$; and
2. the variance of the sampling distribution of the sample mean is smaller than the population distribution, and is given as follows:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}.$$

Remark: Since the standard error is the standard deviation of the sampling distribution, this means that the *standard error* of the sample mean for a sample of size n is $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

Example 4

A sample of size 9 is drawn from a population having a mean $\mu = 8$ and standard deviation $\sigma = 5$. Suppose that the resulting sample is 11, 5, 7, 11, 10, 13, 13, 11, 9.

1. What is the sampling error based on this sample?
2. Determine the mean and standard deviation of the sampling distribution of sample means for samples having the same size.

Solution:

1. Using the provided data, the sample mean can be computed as follows:

$$\bar{x} = \frac{11 + 5 + 7 + 11 + 10 + 13 + 13 + 11 + 9}{9} = 10$$

The *sampling error* is the difference between this value and the corresponding population mean, that is, $10 - 8 = 2$.

2. Our population has mean $\mu = 8$ and standard deviation $\sigma = 5$. Using the previous generalization, the sampling distribution of the sample means of samples with size 9 has a mean $\mu_{\bar{x}} = \mu = 8$ and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{9}} = \frac{5}{3}$.

Example 5

A sample is drawn from a population with $\mu = 78.3$ and $\sigma = 5.6$. How is the *variance* of the sample mean affected when the sample size

1. increases from 36 to 64?
2. decreases from 225 to 144?

Solution:

Since the population has standard deviation $\sigma = 5.6$, its variance is $\sigma^2 = 5.6^2 = 31.36$.

1. The variance of the sample mean *decreases* from $\frac{31.36}{36} \approx 0.87$ to $\frac{31.36}{64} = 0.49$.
2. The variance of the sample mean *increases* from $\frac{31.36}{225} \approx 0.14$ to $\frac{31.36}{144} \approx 0.22$.

From example 5, we see that an increase in the sample size results in a decrease in the variance (or equivalently, the standard deviation) of the sample means and vice versa. This is natural, as we would expect a more accurate estimate of the population parameter when taking a larger sample size.

Remark: Suppose that, as before, we take a sample from a population with mean μ and standard deviation σ . If the sample is selected *without* replacement, the resulting standard error of the mean becomes

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

where N is the population size and n is the sample size. The additional factor $\sqrt{\frac{N-n}{N-1}}$ is called the *finite population correction factor*.

Points to Remember

1. The probability distribution of a statistic is called a *sampling distribution*. The standard deviation of the sampling distribution is called the *standard error* of the statistic.
2. When taking samples of size n from any population with finite mean μ and finite variance σ^2 ,
 - the mean of the sampling distribution of the sample mean \bar{x} is equal to the population mean; that is, $\mu_{\bar{x}} = \mu$; and
 - the variance of the sampling distribution of the sample mean \bar{x} is smaller than the population distribution, which is given by $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$.

Let's Practice

I. Write the letter that corresponds to your answer. Write X if your answer is not among the choices.

- _____ 1. Which of the following refers to the standard deviation of a sampling distribution?
- | | |
|-----------------------|----------------------|
| a. sampling deviation | c. standard error |
| b. sampling error | d. standard variance |

- _____ 2. The weekly allowance of all the students in a certain university has an average of ₱1,000 and a standard deviation of ₱10. A random sample of 50 students was asked their weekly allowance. What is the standard error of the sampling distribution of the mean weekly allowance of students in this university?
- a. 10
 - b. $\frac{10}{\sqrt{50}}$
 - c. $\frac{100}{\sqrt{50}}$
 - d. $\frac{1,000}{\sqrt{50}}$
- _____ 3. For the scenario in item 2, what is the mean of the sampling distribution of the mean weekly allowance of the students?
- a. ₱10
 - b. ₱50
 - c. ₱100
 - d. ₱1,000
- _____ 4. Suppose that the sampling error is zero. Which of the following must be true?
- a. There is an error in the computation of the sample mean.
 - b. The sample statistic and the population parameter are proportional.
 - c. The sample statistic and the population parameter are the same.
 - d. None of the above
- _____ 5. How many possible samples of size 2 are there from the population consisting of the values 10, 20, 40, and 70 if the sample is to be taken *without* replacement? Assume the order of selection is irrelevant.
- a. 4
 - b. 6
 - c. 8
 - d. 12

II. Analyze and solve each problem.

1. A population consists of the values 0, 1, and 2.
 - a. List all the possible samples of size 2 when drawing with replacement and compute the mean of each sample.
 - b. Construct the sampling distribution of the sample means.
 - c. Compute the mean and standard deviation of the sample means. Then compare these with the mean and standard deviation of the population.
2. A population consists of the values 2, 4, 6, and 8.
 - a. List all the possible samples of size 2 when drawing with replacement and compute the mean of each sample.
 - b. Construct the sampling distribution of the mean.
 - c. Compute the mean and standard deviation of the sample means. Then compare these with the mean and standard deviation of the population.
3. A normal population has a mean of 100 and variance of 36. How large must the random sample be if we want the standard error of the mean to be 2?
4. A sample is drawn from a population with $\mu = 1,250$ and $\sigma = 160$. How is the standard error of the mean affected when the sample size
 - a. increases from 64 to 196?
 - b. decreases from 324 to 36?
5. Consider the population consisting of the numbers 10, 20, 40, and 70.
 - a. List all the possible samples of size 2 when drawing *without replacement* and compute the mean of each sample. Assume that the order of selection is *not* important.
 - b. Construct the sampling distribution of the sample means.
 - c. Compute the mean and standard deviation of the sample means. Then compare these with the mean and standard deviation of the population.

Lesson 3

The Central Limit Theorem

Learning Outcomes

- At the end of this lesson, you should be able to
 - illustrate the central limit theorem;
 - define the sampling distribution of the sample mean using the central limit theorem;
 - define the sampling distribution of the sample mean for normal populations when the variance is either known or unknown; and
 - solve problems involving the sampling distribution of the sample mean.

Introduction

In the previous lesson, you have learned that the sampling distribution of the sample mean has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. However, nothing yet has been mentioned about the shape of the distribution.

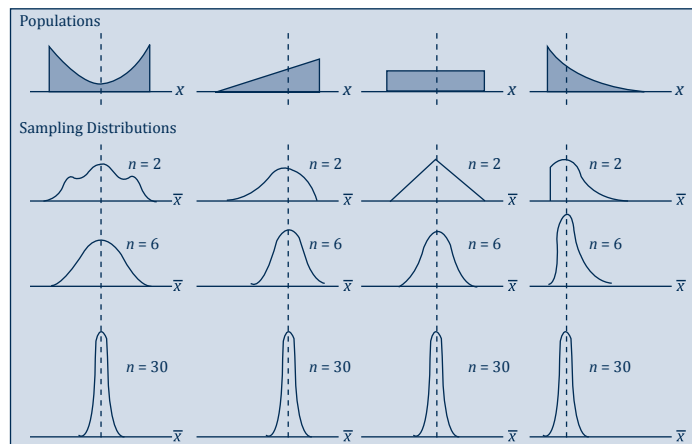
It turns out that as the sample size n becomes large, the sampling distribution of the sample means tends to become bell-shaped and to approximate the normal probability distribution. In fact, this is true for the sampling distribution of the sample mean of *any* population with finite mean and finite standard deviation, regardless of its distribution.

Theorem: The Central Limit Theorem

If random samples of size n are drawn from any population with a finite mean μ and standard deviation σ , then when n is large, the sampling distribution of the sample mean \bar{X} is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

The given figure on the right shows the sampling distributions of the sample mean of different populations as the sample size increases.

Note: For convenience, we shall often refer to the central limit theorem as the CLT.



The central limit theorem on several populations¹

¹ This figure is based from page 227 of the book Basic Statistics for Business and Economics by D.A. Lind, W.G. Marchal, and S.A. Wathen (McGraw-Hill, 2006).

When is the CLT applicable?

- If the sampled population is normal, then the CLT gives more than just an approximation.
In this case, the sampling distribution is normal.
- If the sampled population is almost symmetric, the sampling distribution becomes approximately normal for relatively small values of n .
- If the sampled population is skewed, the sampling distribution becomes approximately normal only for large values of n . Usually, this is when $n \geq 30$.

In example 1 of the previous lesson, recall that the population from which we drew our sample is uniform, and therefore symmetric. Notice that even if the sample size is just 2, the resulting sampling distribution was already approximately normal.

The next example shows how the CLT can be used to compute probabilities involving the mean or the sum of values in a sample.

Example 1

A sample of size 64 is taken from a population with $\mu = 10$ and $\sigma = 25$. Find the probability that

1. the mean of the sample is less than 10.2; and
2. the sum of these 64 values is between 800 and 1,200.

Solution:

Although the distribution of the population is unknown, the CLT still applies since the sample size 64 is greater than 30. In this case, the distribution of the sample means is normal with mean $\mu_{\bar{x}} = 10$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{64}}$.

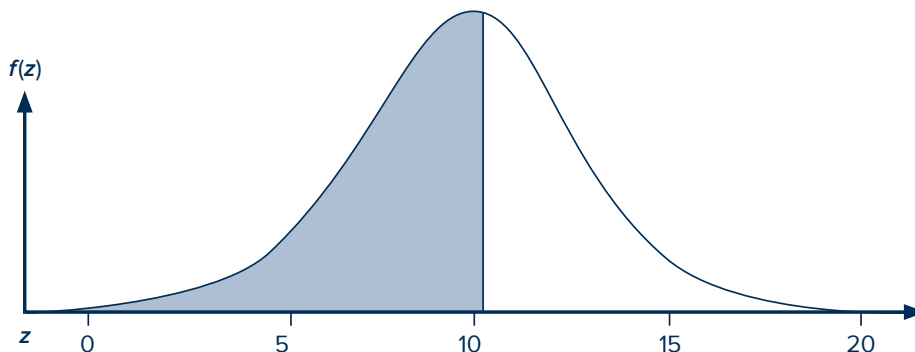
1. The corresponding z-value is

$$z = \frac{10.2 - 10}{\frac{25}{\sqrt{64}}} = 0.06$$

and so

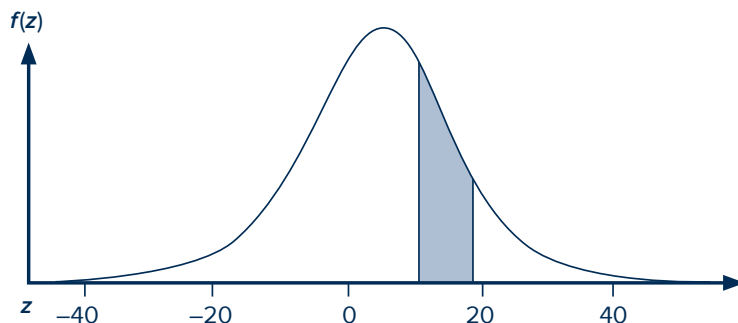
$$\begin{aligned} P(\bar{X} < 10.2) &= P(Z < 0.06) \\ &= 0.5239. \end{aligned}$$

The left-tailed probability we are looking for is shown in the figure below.



2. We first need to convert the problem to one that involves the sample mean. Sums of 800 and 1,200 for 64 values are equivalent to means of $\frac{800}{64} = 12.5$ and $\frac{1,200}{64} = 18.75$.

The probability we are looking for is shown in the figure below



This means that we need to find the probability that the sample mean is between 12.5 and 18.75. These have corresponding z-values as follows:

$$z_1 = \frac{12.5 - 10}{\frac{25}{\sqrt{64}}} = 0.8 \quad \text{and} \quad z_2 = \frac{18.75 - 10}{\frac{25}{\sqrt{64}}} = 2.8.$$

Thus,

$$\begin{aligned} P(12.5 < \bar{X} < 18.75) &= P(0.8 < Z < 2.8) \\ &= P(Z < 2.8) - P(Z < 0.8) \\ &\approx 0.9974 - 0.7881 \\ &= 0.2093. \end{aligned}$$

If the standard deviation of the population is unknown, but the sample size is large ($n \geq 30$), then we can use the sample standard deviation s as replacement for the population standard deviation σ . In this case, the sampling distribution of the mean for samples of size n will still be approximately normally distributed, but with a mean of μ and a standard deviation of $\frac{s}{\sqrt{n}}$.

Example 2

A population of unknown shape has a mean of 75. A sample of 40 is selected from this population, which turns out having a standard deviation of 5. Calculate the probability that the sample mean is

1. greater than 74.
2. between 76 and 77.

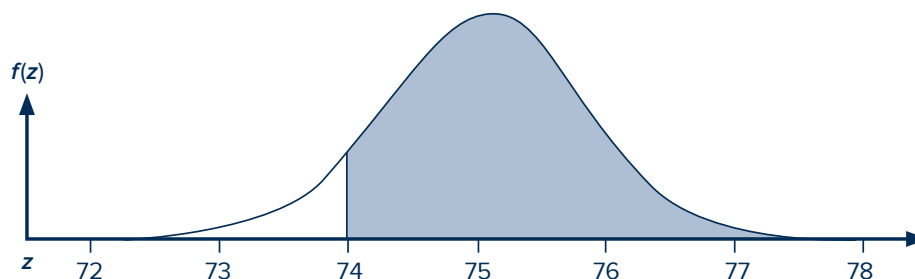
Solution:

Although the population standard deviation is unknown, the sample size is large enough ($n \geq 30$) for us to use the sample standard deviation as a substitute for the population standard deviation. Thus, the distribution of the sample means will be approximately normal with a mean of 75 and standard deviation $\frac{s}{\sqrt{n}} = \frac{5}{\sqrt{40}}$.

1. The z-value is

$$z = \frac{74 - 75}{\frac{5}{\sqrt{40}}} \approx -1.26.$$

The right-tailed probability we are looking for is shown in the figure below.



Thus,

$$\begin{aligned} P(\bar{X} > 74) &= P(Z > -1.26) \\ &= 1 - \Phi(-1.26) \\ &= 1 - 0.1038 \\ &= 0.8962. \end{aligned}$$

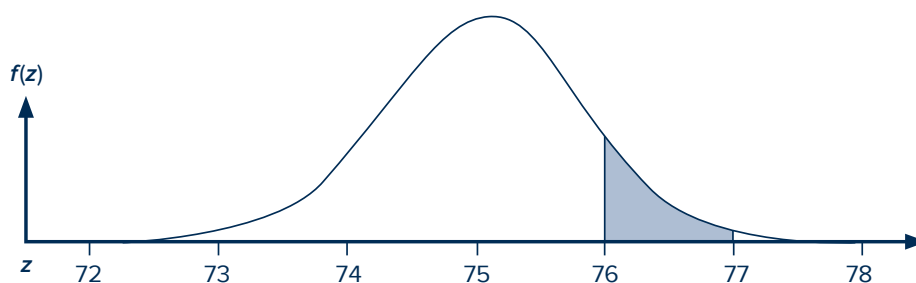
2. The corresponding z-values for 76 and 77 are

$$z_1 = \frac{76-75}{\frac{5}{\sqrt{40}}} \approx 1.26 \quad \text{and} \quad z_2 = \frac{77-75}{\frac{5}{\sqrt{40}}} \approx 2.53.$$

Therefore,

$$\begin{aligned} P(76 < \bar{X} < 77) &= P(1.26 < Z < 2.53) \\ &= \Phi(2.53) - \Phi(1.26) \\ &= 0.9943 - 0.8962 \\ &= 0.0981. \end{aligned}$$

The probability we are looking for is shown in the figure below.



Points to Remember

1. Suppose that we draw samples of size n from any population with mean μ and standard deviation σ . The CLT assures us that when the sample size n is large, then the distribution of the sample means is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
2. For cases where the population is normally distributed, the central limit theorem applies regardless of the sample size.
3. When only the sample standard deviation s is known and the sample size is large ($n \geq 30$), the CLT still applies, but with standard deviation $\frac{s}{\sqrt{n}}$.

The central limit theorem and the concepts in the previous lessons can be used to solve problems involving the sampling distribution of the mean. The next examples illustrate some of these applications.

Example 3

A certain machine makes electric resistors having a mean resistance of 40 ohms (Ω).

1. A sample of 36 resistors is taken and their combined resistance is found to be 1,422 Ω . What is the sampling error of the mean based on this sample?
2. What is the probability that the sample of 36 resistors has combined resistance of more than 1,458 Ω ? Assume that the standard deviation of the sample is 2 Ω .



Solution:

1. In this case, $\bar{x} = \frac{1,422}{36} = 39.5$. Hence, the sampling error is $39.5 - 40 = -0.5$.
2. Since the sample size is 36, which is greater than 30, the CLT applies. Converting the problem to one involving sample means, we can see that we would like to find the probability that \bar{x} is greater than $\frac{1,458}{36} = 40.5$.

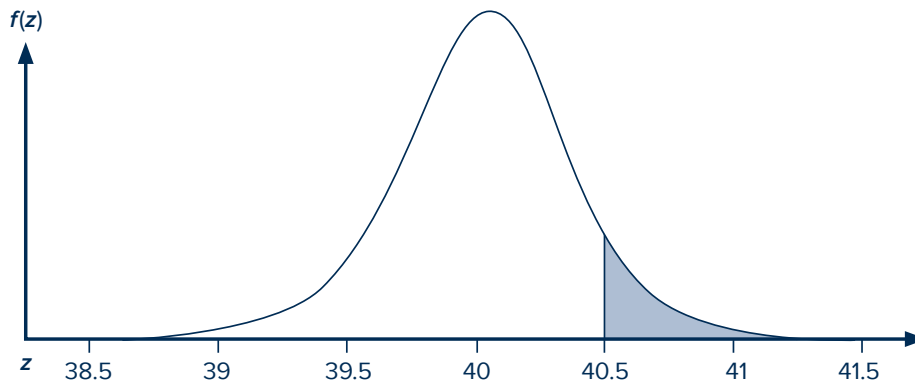
The mean and standard error of the sampling distribution are given by $\mu_{\bar{x}} = 40$ and $\frac{s}{\sqrt{n}} = \frac{2}{\sqrt{36}} = \frac{1}{3}$, respectively. This means that the z-value is given by

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{40.5 - 40}{\frac{2}{\sqrt{36}}} = 1.5.$$

Hence,

$$\begin{aligned} P(\bar{X} > 40.5) &= P(Z > 1.5) \\ &= 1 - \Phi(1.5) \\ &\approx 1 - 0.9332 \\ &= 0.0668. \end{aligned}$$

The right-tailed probability we are looking for is shown in the figure below.



Example 4

The average life span of a bread-making machine is 7 years, with a standard deviation of 1 year. Assuming that the life span of these machines follow a normal distribution, find

1. the probability that the mean life span of a random sample of 16 such machines falls between 6.6 and 7.2 years; and
2. the value of \bar{X} to the right where 15% of the means computed from random samples of size 16 would fall.



Solution:

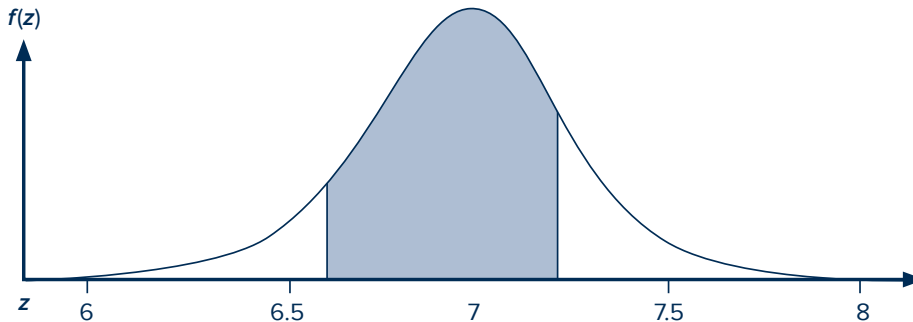
1. Although the sample size of 16 is less than 30, the CLT still applies since the population is normal. Thus, the sampling distribution of the mean life span is also normal with mean $\mu_{\bar{x}} = 7$ and standard deviation $\sigma_{\bar{x}} = \frac{1}{\sqrt{16}} = \frac{1}{4} = 0.25$. The corresponding z-values are

$$z_1 = \frac{6.6 - 7}{0.25} = -1.6 \quad \text{and} \quad z_2 = \frac{7.2 - 7}{0.25} = 0.8.$$

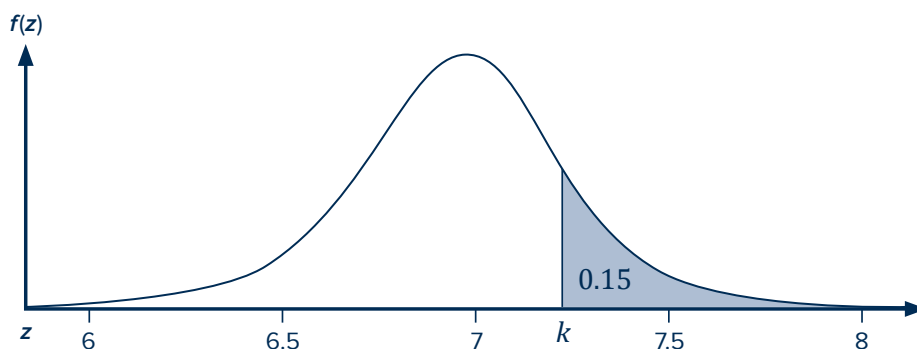
Hence

$$\begin{aligned} P(6.6 < \bar{X} < 7.2) &= P(-1.6 < Z < 0.8) \\ &= \Phi(0.8) - \Phi(-1.6) \\ &\approx 0.7881 - 0.0548 \\ &= 0.7333. \end{aligned}$$

The probability we are looking for is shown in the figure below.



2. Let k be the required value. Since 15% of the sample means lie to the right of k , this means that 85% of them lie to the left of k .



Notice that in the standard normal distribution, the z -value that leaves an area of 0.85 to the left is approximately 1.04. Thus,

$$\begin{aligned}\bar{x} &= z\sigma + \mu \\ &= (1.04)(0.25) + 7 \\ &\approx 7.26.\end{aligned}$$

Let's Practice

Solve each problem.

- A normal population has a mean of 70 and a standard deviation of 12. If a random sample of size 16 is taken from this population, compute the probability that the sample mean is
 - greater than 72.5.
 - less than 67.
 - between 67.5 and 72.
- A population of unknown shape has a mean of 85. You select a sample of 40, and this sample has a standard deviation of 5. Compute the probability that the sample mean is
 - less than 84.
 - between 84 and 86.
 - between 86 and 87.
 - greater than 87.
- If 20 random numbers are sampled from a normal distribution with mean $\mu = 0.5$ and $\sigma = 0.3$, what is the probability that
 - the sum of these numbers is at most 8?
 - the average of these numbers is between 0.575 and 0.625?
- The amount of time a bank teller spends on a customer is a random variable with mean $\mu = 3.2$ minutes and standard deviation of $\sigma = 1.6$ minutes. For a random sample of 36 customers, what is the probability that their mean time at the teller's window is
 - less than 2.7 minutes?
 - at least 3.5 minutes?
 - between 3.2 and 3.6 minutes?

5. The Mr. Tidy Company wants to ensure that their laundry powder refill actually contains 100 grams (g) of detergent. Based on historical summaries from the filling process, the mean amount per refill is 100 g with a standard deviation of 2 g. What is the likelihood that 40 such packs will actually contain an average of 99.8 g of soap or less?
6. Random samples of size 50 are drawn from a population with mean $\mu = 15$ and standard deviation $\sigma = 2$. What value of \bar{X} will 10% of the sample means of such random samples fall below?
7. Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 5 minutes. In a given week, a bus transported passengers 40 times.
 - a. What is the probability that the average transport time was less than 30 minutes?
 - b. If the average transport time exceeds k minutes 20% of the time, what is the value of k ?
8. The random variable X , representing the number of cherries in a cherry puff, has the following probability distribution:

k	4	5	6	7
$P(X = k)$	0.2	0.4	0.3	0.1

- a. Find the mean μ and the variance σ^2 of X .
- b. Find the mean $\mu_{\bar{x}}$ and the variance $\sigma_{\bar{x}}^2$ of the mean for random samples of 30 cherry puffs.
- c. Find the probability that the average number of cherries in a random sample of 30 cherry puffs will be less than 5.5.

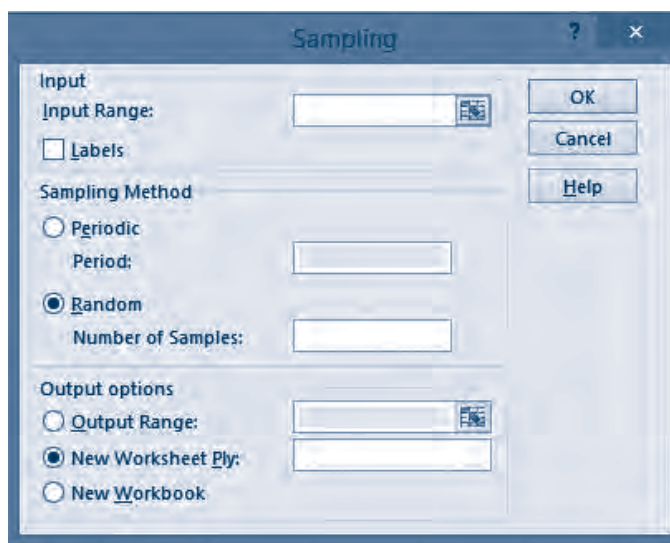
Software Tutorial in MS Excel

To select a simple random sample in MS Excel similar to the one in example 1 of lesson 1, do the following steps:

1. Go to the *Data* ribbon and click the *Data Analysis* button. This displays the list of available Analysis Tools, as shown below.




2. From the list of Analysis Tools, select *Sampling*. This brings up the following screen.

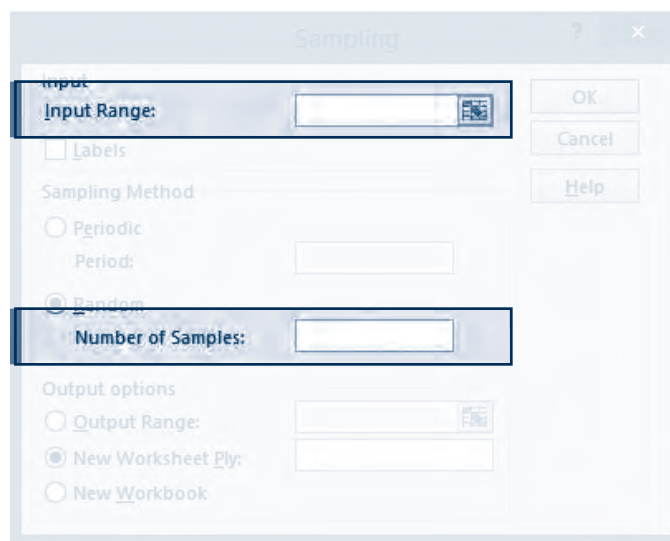


The screenshot shows the 'Sampling' dialog box with the following fields and options:

- Input**
 - Input Range:** A text box with a selection icon to its right.
 - ☐ **Labels**
- Sampling Method**
 - ☐ **Periodic**
 - Period:** A text box.
 - ☒ **Random**
 - Number of Samples:** A text box.
- Output options**
 - ☐ **Output Range:** A text box with a selection icon to its right.
 - ☒ **New Worksheet Ply:** A text box.
 - ☐ **New Workbook**

Buttons on the right: OK, Cancel, Help.

3. Click the  button next to Input Range. This brings you back to your current worksheet, allowing you to select the cells corresponding to the population. Also, input the number of samples desired.



This screenshot is identical to the previous one, but with two blue rectangular highlights:

- One highlight is around the **Input Range:** text box and its selection icon.
- Another highlight is around the **Number of Samples:** text box.

4. Click OK. By default, this places your sample on a separate worksheet. You may just copy the resulting sample to the desired cells.

Chapter Review

- A **population** is the totality of items, things, or people under consideration. A **sample** is a subset of the population.
- Any measurable characteristic of a population is called a **parameter**. Any measurable characteristic of a sample is called a **statistic**.
- **Simple random sampling** is a selection of a subset of a population where each element has an equal chance of being selected.
- A **sampling frame** is a complete list of all the members of the population from which a sample is taken.
- In a **systematic random sampling**, a random starting point is selected, and then every k th member of the population is selected.
- **Stratified random sampling** is a selection of a simple random sample from each of a given number of subpopulations or **strata**.
- **Cluster sampling** is a selection of clusters from the available clusters in the population. Each member of the selected clusters is then included in the sample.
- The difference between the value of a sample statistic and the corresponding population parameter is called the **sampling error**.
- The probability distribution of a statistic is called a **sampling distribution**. The standard deviation of the sampling distribution is called the **standard error** of the statistic.
- When taking samples (with replacement) of size n from any population with finite mean μ , and finite variance σ^2 ,
 1. the mean of the sampling distribution of the sample mean is equal to the population mean, that is, $\mu_{\bar{x}} = \mu$; and
 2. the variance of the sampling distribution of the sample mean is smaller than the population distribution, and is given as follows:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

- If random samples of size n are drawn from any population with a finite mean μ and standard deviation σ , then when n is large, the sampling distribution of the sample mean \bar{X} is approximately normally distributed with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.
- If the standard deviation of the population is unknown, but the sample size is large ($n \geq 30$), then we can use the sample standard deviation s as replacement for the population standard deviation σ . In this case, the sampling distribution of the mean for samples of size n will still be approximately normally distributed, but with a mean of μ and a standard deviation of $\frac{s}{\sqrt{n}}$.

Chapter Performance Tasks

1. The Philippine Stock Exchange Index

Imagine that you are a quantitative analyst commissioned by a certain investor to develop a model to help him decide which stocks to buy and which stocks to sell. Start gathering data and obtain the prices of the 30 stocks that compose the Philippine Stock Exchange Index (PSEi) at <http://www.pse.com.ph/stockMarket/marketInfo-marketActivity.html?tab=1>. Include in your analysis the computation of the mean and standard deviation of the prices of a sample of five of these stocks. To select the sample, use both MS Excel and a random number table generator. Then, make an analysis report based on your findings. Compare the results you obtained from two samples to that which includes all the 30 stocks. Make sure that your calculations are accurate, detailing the steps that you made to arrive at your conclusion and findings. Identify in your report which stocks you will recommend the investor to buy and which stocks you will recommend him to sell.



2. Student Heights

Imagine that you are a school biostatistician. Part of your job is to collect, dissect, and summarize student data, as well as release health information and assessments based on your findings. You are tasked by your school clinic supervisor to generate grade 11 students' profiles by collecting information which includes their heights. Collect the heights (in cm) of one to two classes of grade 11 students in your school. These values will be treated as two populations: one for males and one for females. Specifically, your tasks are as follows:



- Use MS Excel (or any other software) to produce histograms for the heights of male and female students. Based on your results, would you consider the heights to be approximately normally distributed?
- Compute the mean and standard deviation of these two populations.

Assuming that the heights of male and female students are both normally distributed, you are also required to prepare and submit to the clinic supervisor a written report highlighting the answers to the following questions:

- a. If a sample of 9 male students is randomly selected, what is the probability that the mean height is (i) more than 158 cm? (ii) between 148 and 163 cm?
- b. If a sample of 9 female students is randomly selected, what is the probability that the mean height is (i) more than 158 cm? (ii) between 148 and 163 cm?
- c. Is there a significant difference between the probabilities you computed in (a) and (b)? Is this expected?

Make sure that your report is organized, detailed, neat, and accurate.

Chapter Exercises

1. The following table lists the population census data (2015) of provinces from Luzon. Also indicated is the region from which each province is part of.

No.	Province	Population	Region
01	Abra	241,160	CAR
02	Albay	1,314,826	V
03	Apayao	119,184	CAR
04	Aurora	214,336	III
05	Bataan	760,650	III
06	Batanes	17,246	II
07	Batangas	2,694,335	IV-A
08	Benguet	791,590	CAR
09	Bulacan	3,292,071	III
10	Cagayan	1,199,320	II
11	Camarines Norte	583,313	V
12	Camarines Sur	1,952,544	V
13	Catanduanes	260,964	V
14	Cavite	3,678,301	IV-A
15	Ifugao	202,802	CAR
16	Ilocos Norte	593,081	I
17	Ilocos Sur	689,668	I
18	Isabela	1,593,566	II
19	Kalinga	212,680	CAR

No.	Province	Population	Region
20	La Union	786,653	I
21	Laguna	3,035,081	IV-A
22	Marinduque	234,521	IV-B
23	Masbate	892,393	V
24	Mountain Province	154,590	CAR
25	Nueva Ecija	2,151,461	III
26	Nueva Vizcaya	452,287	II
27	Occidental Mindoro	487,414	IV-B
28	Oriental Mindoro	844,059	IV-B
29	Palawan	1,104,585	IV-B
30	Pampanga	2,609,744	III
31	Pangasinan	2,956,726	I
32	Quezon	2,122,830	IV-A
33	Quirino	188,991	II
34	Rizal	2,884,227	IV-A
35	Romblon	292,781	IV-B
36	Sorsogon	792,949	V
37	Tarlac	1,366,027	III
38	Zambales	823,888	III

Source: <https://www.psa.gov.ph/content/highlights-philippine-population-2015- census-population>, retrieved 22 August 2016.

- a. You wish to select a sample of seven from this list. Using the last two digits of the numbers in the 14th column of *Appendix A*, determine which provinces are included in the sample.
- b. You wish to use a systematic sample of every fifth item, and the digit 01 is selected as the starting point. Which provinces are included?
- c. A sample of one province from each region is to be chosen. Describe carefully how you would perform the sampling process in detail. That is, show the random numbers you have selected and the corresponding provinces which are included in your sample.

2. In the law firm Tuano and Associates, there are six associates. Listed below is the number of cases each associate actually tried in court last month:

Associate	Number of Cases
Flores	1
Mendoza	0
Ramos	3
Villegas	3
Mallari	6
Antonio	3

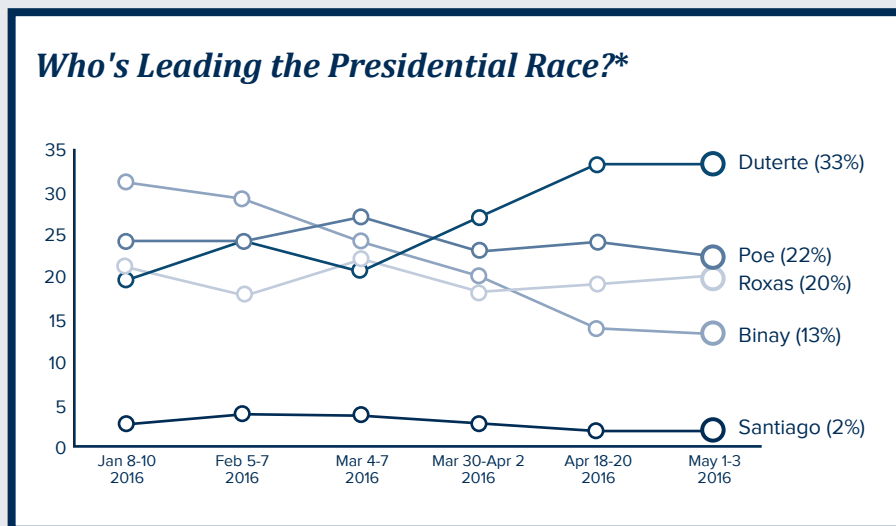
- List all possible samples of size 3, and compute the mean of each sample. Assume that the sample is selected with replacement.
 - Compare the mean of the distribution of sample means to the population mean.
 - Compare the dispersion (standard deviation) in the population with that of the sample means.
3. An employment agency received a client's request for two experienced clerical staff. Based on the client's requirements, the agency has five qualified persons available for these positions. List down the 10 possible combinations of persons A, B, C, D, and E that could be sent to the job. If each of these samples has an equal probability of $\frac{1}{10}$ of being selected, show that the probability is $\frac{2}{5}$ that an individual will be included in the sample.
4. Most wireless mice require one or two AA batteries. Assume that with regular use, these batteries have a mean life of 8.0 weeks. The distribution of the battery lives closely follows a normal distribution with standard deviation of 1.2 weeks. As part of their testing program, the TechOne peripheral company tests samples of 25 of these batteries.
- What can you say about the shape of the distribution of the sample mean?
 - What proportion of the samples will have a mean life of more than 8.3 weeks?
 - What proportion of the samples will have a mean life of between 8.1 and 8.5 weeks?
5. Suppose that the mean age that a man in the Philippines marries is 28 years.² For a random sample of 60 married men, what is the likelihood that the mean age they married is greater than 28.3 years? Assume that the standard deviation of the sample is 2.5 years.

² Based on 2011 figures of the Philippine Statistics Authority (PSA) for registered marriages, 28 is in fact the *median* age for a groom in the country. For this exercise, we assume that the mean is the same as this value.

6. For the scores on an achievement test given to a certain population of students, the expected value is 500 and the standard deviation is 100. Let \bar{X} be the mean of the scores of a random sample of 35 students from the population. Estimate the probability that \bar{X} is between 460 and 500.
7. A sample is drawn from a population with mean $\mu = 78.3$ and standard deviation $\sigma = 5.6$. How is the variance of the sample mean affected when the sample size is
 - a. increased from 64 to 196?
 - b. decreased from 324 to 36?
8. If the standard deviation of the mean for the sampling distribution of random samples of size 36 from a large population is 2, how large must the sample size be if the standard error is to be reduced to 1.2?
9. Suppose that samples of size 25 are selected at random from a normal population with mean 100 and standard deviation 10. What is the probability that the sample mean falls in the interval from $\mu_{\bar{x}} - 1.0\sigma_{\bar{x}}$ to $\mu_{\bar{x}} + 1.8\sigma_{\bar{x}}$?
10. A parking lot is planned for a new apartment with 250 units. For each apartment unit, it is assumed that the number of cars is 0, 1, or 2, with probabilities 0.1, 0.7, and 0.2, respectively. In order to be approximately 95% certain that there is room for all cars, how many spaces should the parking lot have?

Chapter 5

Estimation of Parameters



*Based on the BusinessWorld-SWS May 1-3, 2016 Pre-Election Survey

Popularity surveys are an important part of any election campaign season. Whether or not one wishes to believe the sometimes conflicting results of different survey agencies, these surveys allow the candidates to know their current status and support from the electorate and to modify their campaign strategy accordingly.

These surveys often take samples of 400 or more from the target population, and use the preferences of these 400 people to develop a *point estimate* of the true proportion that support a particular candidate. As we cannot expect this sample proportion to estimate the population proportion exactly, it is necessary to include a *margin of error* for the estimate. This yields an interval of values, called a *confidence interval*, which has a high probability of containing the true proportion. Such point and interval estimates, as well as the appropriate choice of sample size, are the focus of this chapter.

Lesson 1

Point Estimators

Learning Outcomes

- At the end of this lesson, you should be able to
 - illustrate and distinguish between point and interval estimations;
 - identify the point estimator and compute point estimates for the population mean; and
 - identify the point estimator and compute point estimates for the population proportion.

Introduction

There are many practical scenarios where one may wish to estimate the value of a population parameter, especially when it is impractical or even impossible to examine the entire population. For example, a national tourism office might be interested in estimating the mean amount spent by each of the tourists visiting a country. Large television networks would also be interested in knowing the ratings of their shows or what proportion of the viewing public actually watch their shows.

In each of the two cases, the only recourse one can make is to take a sample of the population. To estimate a population parameter, one can then use the information in the sample to compute an *estimator*. Since the value of an estimator is computed based on the sample data, it can be seen as a sample statistic.

Definition 1

A rule, usually expressed as a formula, that tells us how to compute an estimate based on a sample is called an **estimator**.

In this lesson, we will focus on point estimation. In *point estimation*, a single number is calculated based on information in the sample. In this case, the formula is called a *point estimator*, while the computed value is called a *point estimate*.

Choosing a Point Estimator

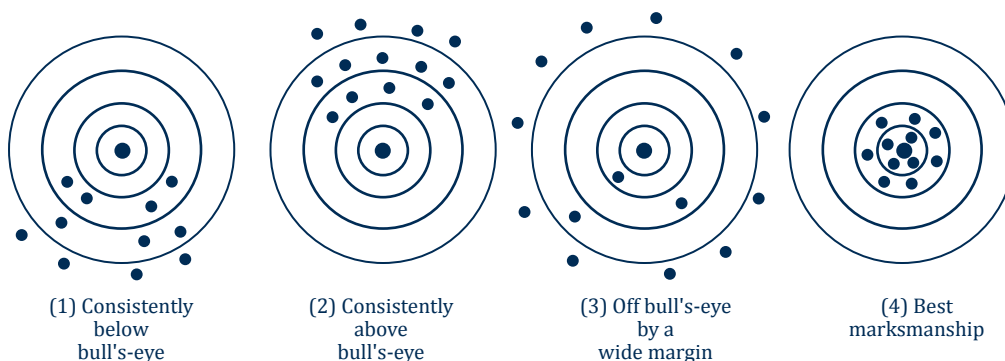
For a given population parameter, there are many possible point estimators available. For example, when estimating the population mean, the obvious choice for a point estimator would be the sample mean. However, one could also use either the sample median, the average of smallest and the biggest values in the sample, or the value obtained in the first sample as your point estimators, among others. So how do we determine which point estimator is the best choice?

Due to the random nature of sampling, the value that will be obtained from the estimator will clearly vary. In fact, it will be futile to expect to get the true value when one obtains an estimate based on a single sample. However, an ideal estimator should be *consistent*. When we take a larger sample, we would intuitively expect our estimates for the population parameter to improve. This is also true for consistent estimators—the probability of obtaining values which are far from the true value of the parameter decreases as the sample size increases.

Aside from consistency, another desirable property of an estimator is *unbiasedness*. We have encountered this term in the previous chapter in the context of a sample being representative of the population. However, for estimators, unbiasedness has a different meaning. It implies that the estimator has no tendency to overestimate or underestimate the true parameter value. This means that while we may obtain estimates which are higher or lower compared to the parameter it estimates, in the long run, the average of the resulting estimates is the same as the true parameter value.

Finally, another important consideration is the *efficiency* of the estimator. This refers to how large the variance of the estimator is. Under the assumption that the estimator is unbiased, an estimator with a smaller variance is preferred. This means that, on the average, the point estimates we will obtain would be located in a small interval only. Consequently, we will not expect estimates to widely vary in value.

The unbiasedness (accuracy) and efficiency (precision) of an estimator are best described by the results of a shooter in a marksmanship competition. Assuming he shoots randomly at the target, the figure below shows four possible results of his shots. The ideal case is, of course, the fourth one, where the shots are consistently close to the target.



An illustration of the concepts of accuracy and precision of an estimate. Ideally, we prefer estimates which consistently hit close to the target, as shown in the fourth picture.

Point Estimates for the Population Mean and the Population Proportion

How do we check these properties given a point estimator? We will need to go back to the corresponding *sampling distribution* of the estimator.

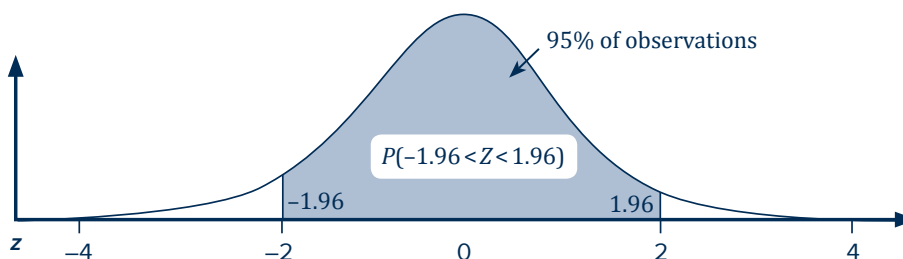
In a real-life sampling situation, all we may have is a point estimate derived from a sample of size n . The first thing we would like to check would be how far it is from the population value.

Definition 2

The distance between an estimate and the parameter being estimated is called the **error of estimate**.

For most of this chapter, we shall assume that our sample sizes are large such that any estimator would have a sampling distribution which is normally distributed by the *central limit theorem* (see chapter 4, lesson 3). Since the samples now have a normal distribution, recall from the *empirical rule* (see chapter 3, lesson 3) that 95% of the observations will lie within two (or, more exactly, 1.96) standard deviations of the mean of the distribution. This means that 95% of our estimates will not differ from the true parameter value by more than 1.96 times the standard error of its sampling distribution.

The figure below shows the corresponding area covered by 95% of the observations.



This quantity, called the **95% margin of error**, provides a practical upper bound for the error of estimation. Let us now look at the typical point estimates and their corresponding 95% margin of errors when estimating the population mean and the population proportion.

A natural choice of point estimate for the population mean μ is the sample mean \bar{X} . Recall from chapter 4, lesson 3 that if the sample size n is large, then the sampling distribution for \bar{X} is approximately normally distributed with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$. This gives rise to the following result:

Point Estimate for the Population Mean

To estimate the population mean μ , we use the point estimator \bar{X} with an estimated standard error of $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and a 95% margin of error of $1.96 \times SE_{\bar{x}}$ when $n \geq 30$.

Example 1

Suppose we take a sample of size $n = 30$ from a population with variance $\sigma^2 = 0.2$. Then the 95% margin of error is

$$1.96 \times SE_{\bar{x}} = 1.96 \times \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{\sqrt{0.2}}{\sqrt{30}} \approx 0.16.$$

This means that by using \bar{X} to estimate the population mean μ , we can be confident that the error of estimate will not exceed 0.16.

If the sample size n is large, we can use the sample standard deviation s in place of the population standard deviation σ in the preceding formula. This is illustrated in the next example.

Example 2

Luisa, a researcher, is concerned about the amount of time children spend on smartphones. In an attempt to determine the average time grade 3 pupils spend per day on smartphones, she asked the parents of a sample of 50 grade 3 pupils from a local elementary school the amount of time their children spend on smartphones. She obtained a sample mean of 35 minutes with a standard deviation of 10 minutes. In this case, the point estimate for the mean time a grade 3 pupil spends daily using a smartphone is $\bar{X} = 35$ minutes, while the 95% margin of error is

$$\frac{s}{\sqrt{n}} = \frac{10}{\sqrt{50}} \approx 1.41 \text{ minutes.}$$

When estimating the population proportion p , the most common choice of point estimator is the *sample proportion*. If there are n observations in our sample and x of these can be classified as “successes,” then the sample proportion of successes is given by $\hat{p} = \frac{x}{n}$. In this case, we have the following result:

Point Estimate for the Population Proportion

To estimate the population proportion p , we use the point estimator \hat{p} (read as “p-hat”) with an estimated standard error of $SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$, where $\hat{q} = 1 - \hat{p}$, and a 95% margin of error of $1.96 \times SE_{\hat{p}}$ when $n\hat{p} > 5$ and $n\hat{q} > 5$.

Example 3

Pulse Asia reported the perceived urgency of selected national issues based on a nationwide survey of 1,200 adults from 25 September–01 October, 2016. Based on the results of the survey, only 31% of the Filipinos surveyed cited fighting criminality as one of the three most urgent issues from a list given.

In this case, our point estimate of the true proportion of Filipinos who find fighting criminality as one of the most urgent issues is $\hat{p} = 0.31$. The corresponding 95% margin of error is approximately

$$1.96 \times \sqrt{\frac{\hat{p}\hat{q}}{n}} = 1.96 \times \sqrt{\frac{0.31(1-0.31)}{1200}} \approx 0.0262.$$

In this case, the estimate can be considered to be no more than 2.6% away from the true proportion.

Let's Practice

I. Write the letter that corresponds to your answer. Write X if your answer is not among the given choices.

- _____ 1. What is the formula that calculates a single number as an estimate based from a sample?
- | | |
|--------------------|---------------------|
| a. point estimate | c. random sample |
| b. point estimator | d. sample statistic |
- _____ 2. What is an estimator that does not have a tendency to underestimate or overestimate the population parameter?
- | | |
|---------------|-------------|
| a. consistent | c. precise |
| b. efficient | d. unbiased |
- _____ 3. What term refers to the variance of an estimator?
- | | |
|----------------|-----------------|
| a. accuracy | c. efficiency |
| b. consistency | d. unbiasedness |
- _____ 4. What do you call the difference between the estimate and the parameter being estimated?
- | | |
|----------------------|-------------------|
| a. error of estimate | c. point estimate |
| b. margin of error | d. standard error |

- _____ 5. What does “consistency” in an estimator mean?
- It means the expected value of the estimator is equal to the population parameter.
 - It means the resulting estimates get better as the sample size becomes larger.
 - It means the resulting estimates have a small error of estimate.
 - It means the resulting estimates have a small variance.

II. Compute the relevant quantities.

- Find and interpret the 95% margin of error for the associated estimates based on the following information:
 - $n = 76, \sigma^2 = 49$
 - $n = 30, s^2 = 144$
 - $n = 150, \hat{p} = 0.2$
 - $n = 100, \hat{p} = 0.68$
- A research firm conducted a survey to determine the mean amount (in pesos) heavy smokers spend on cigarettes in a week. A sample of 36 heavy smokers revealed that $\bar{x} = 1,000$ and $s = 200$.
 - What is the point estimate for the population mean? Explain what it indicates.
 - What would be a 95% margin of error for your estimate in item (a)?
- Karla, the manager of a mall, wants to estimate the mean amount spent per shopping visit by customers. A sample of 10 customers reveals the following amount spent (rounded to the nearest 50 pesos): ₱2,400, ₱2,100, ₱2,600, ₱1,350, ₱2,050, ₱3,200, ₱1,850, ₱1,950, ₱1,150, ₱3,150. From past experience, Karla knows that the amounts are approximately normally distributed with a standard deviation of 650. Give a point estimate for the true mean amount spent per shopping visit by customers. What would be a 95% margin of error for your estimate?
- In a December 2016 survey by Pulse Asia, 74% of the 1,200 adults surveyed disagreed with the view that martial law is needed to resolve the various problems currently facing the country. Use this information to give a point estimate and a 95% margin of error for the true proportion of all Filipinos who believe martial law is not needed to solve national problems.
- A campus organization conducted a survey on the preferred student council president among a school's senior high school students. The results show that 115 prefer candidate A, 73 prefer candidate B, 52 prefer candidate C, and the remaining 10 are undecided. Give point estimates for the proportion of the school's senior high school students who prefer candidate A and the proportion who prefer candidate B. What are the 95% margins of error for these estimates?

Lesson 2

Interval Estimation for a Mean

Learning Outcomes

- At the end of this lesson, you should be able to
 - compute the confidence interval estimate based on the appropriate form of the estimator for the population mean; and
 - solve problems involving confidence interval estimation of the population mean.

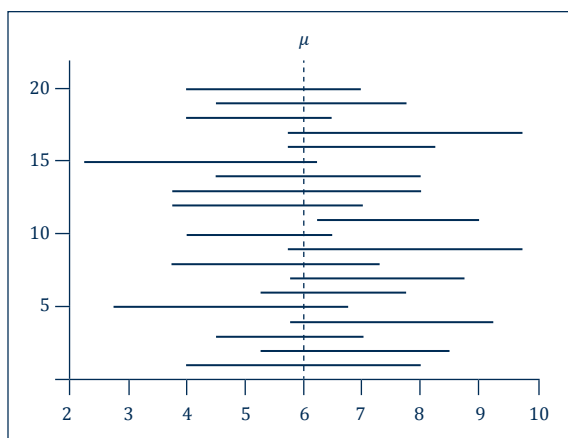
Introduction

In the previous lesson, you have learned how to construct a *point estimator* of a population parameter. However, a much better method for estimating a parameter would be to incorporate a “margin of error” to construct an interval that contains the true parameter value. This method is called an *interval estimation*.

In an interval estimation, two numbers are calculated based on sample data, forming an interval where the parameter’s value is expected to lie. In this case, the formula is called an *interval estimator*, while the range of values obtained is called an *interval estimate* or a *confidence interval*.

Definition 1

The **confidence coefficient**, denoted by $1 - \alpha$, is the probability that a confidence interval will contain the estimated parameter.



Above is a 95% confidence interval μ . Here, the true value of the parameter μ is 6. Among the 20 confidence intervals, 95% of 20, or 19, contain the value 6.

It is important to note that the confidence coefficient is *not* the probability that the parameter is in a particular constructed interval. Instead, it refers to our confidence in the process done to construct the interval.

Once a confidence interval estimate is constructed, it either contains or does not contain the true value of the parameter. Thus, the probability that a specific confidence interval contains the value of the parameter is either 0 or 1. However, if you construct M confidence intervals using the formula, then, in the long run, we could expect that an average of $(1 - \alpha)$ of these M would contain the value of the parameter.

The following is the confidence interval for the case when the population standard deviation is known. It is based on an interval around the point estimate for the mean, which is the sample mean.

**Confidence Interval for the Population Mean
(large sample or normal population, σ is known)**

A $(1 - \alpha) \times 100\%$ confidence interval for μ is given by

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right)$$

where \bar{x} = sample mean;

$z_{\frac{\alpha}{2}}$ = z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

σ = population standard deviation; and

n = sample size.

In the formula above, the number $\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ is known as the *lower limit* and $\bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ is the *upper limit* of the confidence interval.

If the population standard deviation is not known but the sample size is large, we can use the sample standard deviation as an estimate for the population standard deviation.

**Confidence Interval for the Population Mean
(large sample, σ is unknown)**

A $(1 - \alpha) \times 100\%$ confidence interval for μ is given by

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right)$$

where \bar{x} = sample mean;

$z_{\frac{\alpha}{2}}$ = z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

s = sample standard deviation; and

n = sample size.

Notice that this is the same as the formula for the confidence interval for the mean where σ is known, except that the population standard deviation σ is replaced with its corresponding statistic, s .

Example 1

Find and interpret a 95% confidence interval for the population mean given that $n = 36$, $\bar{x} = 13.1$, and $\sigma = 3.42$.

Solution:

For a 95% confidence interval, $1 - \alpha = 0.95$, so $\alpha = 0.05$. Using the z-table in *Appendix B*, we have $z_{\frac{\alpha}{2}} = z_{\frac{0.05}{2}} = z_{0.025} = 1.96$. Substituting these values into the formula above for the confidence interval for a population mean, we have

$$\begin{aligned}\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right) &= \left(13.1 - 1.96 \cdot \frac{3.42}{\sqrt{36}}, 13.1 + 1.96 \cdot \frac{3.42}{\sqrt{36}}\right) \\ &= (13.1 - 1.12, 13.1 + 1.12) \\ &= (11.98, 14.22).\end{aligned}$$

Thus, we can be 95% confident that the interval (11.98, 14.22) contains the true value of the population mean.

Example 2

A random sample of 10 chocolate energy bars of a certain brand has, on the average, 230 calories with known population standard deviation of 15 calories. Construct and interpret a 99% confidence interval for the true mean calorie content of this brand of energy bar. Assume that the distribution of calories is approximately normal.

Solution:

From the given information, we see that $n = 10$, $\bar{x} = 230$, and $\sigma = 15$.

Since we require a 99% confidence interval, we have $1 - \alpha = 0.99$. Thus, $\alpha = 0.01$. Using a z-table, we have $z_{\frac{\alpha}{2}} = z_{\frac{0.01}{2}} = z_{0.005} = 2.575$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049

Since 0.005 is not on the table and is exactly in the middle of 0.0049 and 0.0051, we take the average of the *z*-values corresponding to 0.0049 and 0.0051.

This means that a 99% confidence interval for the mean calorie content of this energy bar is

$$\begin{aligned} \left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) &= \left(230 - 2.575 \cdot \frac{15}{\sqrt{10}}, 230 + 2.575 \cdot \frac{15}{\sqrt{10}} \right) \\ &= (230 - 12.21, 230 + 12.21) \\ &= (217.79, 242.21). \end{aligned}$$

We can therefore be 99% confident that the true mean calorie content of this brand of energy bar is between 217.79 and 242.21 calories.

Example 3

Calvin owns a water refilling station in his neighborhood. To assess the efficiency of his station's operation, he decided to do a study of the water consumption of his customers. He selected 45 households at random whose number of liters (L) of water consumed during the past six months was recorded. The average consumption was found to be 134.6 L with a standard deviation of 21.1 L. What is a 95% confidence interval for the mean water consumption during the past six months among his company's customers?

Solution:

The given mean and standard deviation values are based on the sample of 45 households, so we have $n = 45$, $\bar{x} = 134.6$, and $s = 21.1$. Although the population standard deviation is unknown, we can use the sample standard deviation as its substitute since the sample size is large. Since we wish to find a 95% confidence interval, $\alpha = 0.05$, so the corresponding *z*-value is $z_{\frac{0.05}{2}} = z_{0.025} = 1.96$. Therefore, a 95% confidence interval for the mean water consumption is

$$\begin{aligned} \left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right) &= \left(134.6 - 1.96 \cdot \frac{21.1}{\sqrt{45}}, 134.6 + 1.96 \cdot \frac{21.1}{\sqrt{45}} \right) \\ &= (134.6 - 6.16, 134.6 + 6.16) \\ &= (128.44, 140.76). \end{aligned}$$

The Width of a Confidence Interval and the Margin of Error

Definition 2

The **width** of the confidence interval is the difference of the upper and lower limits.

For the case of a large-sample interval estimate for a population mean, recall that the confidence interval is given by

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right).$$

Thus, its width is

$$\left(\bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) - \left(\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right) = 2z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

Taking half of this quantity, we obtain $z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, which represents the *maximum allowable deviation* or *margin of error* of a confidence interval around the sample mean. Notice that the z -value used is the one that leaves an area of $\frac{\alpha}{2}$ to its right. More importantly, the area under the standard normal distribution between $-z_{\frac{\alpha}{2}}$ and $z_{\frac{\alpha}{2}}$ is $1 - \alpha$.

When $\alpha = 0.05$, this gives $z_{\frac{0.05}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 1.96 \times \frac{\sigma}{\sqrt{n}}$, which is the familiar 95% margin of error formula which we have seen in lesson 1 of this chapter.

Notice that the formula for the maximum allowable error (and therefore, the width of a confidence interval) depends on three quantities:

1. *The critical value $z_{\frac{\alpha}{2}}$.* As the required degree of confidence in the interval increases, $z_{\frac{\alpha}{2}}$ also increases. This means that the width of the interval *also increases*.
2. *The standard deviation σ .* The larger the variability of the population being studied, the wider the margin of error will be. This means that the resulting confidence interval also becomes wider.
3. *The sample size n .* Notice that n is in the denominator of the formula. This means that the larger the sample size, the *narrower* the resulting confidence interval will be.

Example 4

Consider again the given information in example 1: $n = 36$, $\bar{x} = 13.1$, and $\sigma = 3.42$. Compare the widths of a 95% and a 99% confidence interval for μ .

Solution:

For a 95% confidence interval, $\alpha = 0.05$, so the width of the interval is

$$2z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 2z_{0.025} \cdot \frac{\sigma}{\sqrt{n}} = 2(1.96) \left(\frac{3.42}{\sqrt{36}} \right) \approx 2.23.$$

On the other hand, the width of a 99% confidence interval is

$$2z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 2z_{0.005} \cdot \frac{\sigma}{\sqrt{n}} = 2(2.575) \left(\frac{3.42}{\sqrt{36}} \right) \approx 2.94.$$

As expected, the 99% confidence interval is wider than the 95% confidence interval. Note that we could also compute for these widths by subtracting the lower limit of each confidence interval from its corresponding upper limit.

For a 95% confidence interval, the width of the interval is

$$13.1 + (1.96) \left(\frac{3.42}{\sqrt{36}} \right) - \left[13.1 - (1.96) \left(\frac{3.42}{\sqrt{36}} \right) \right] = 2(1.96) \left(\frac{3.42}{\sqrt{36}} \right) = 2.23$$

For a 99% confidence interval, the width of the interval is

$$13.1 + (2.575) \left(\frac{3.42}{\sqrt{36}} \right) - \left[13.1 - (2.575) \left(\frac{3.42}{\sqrt{36}} \right) \right] = 2(2.575) \left(\frac{3.42}{\sqrt{36}} \right) = 2.94$$

Let's Practice

- Find and interpret a $(1 - \alpha) \times 100\%$ confidence interval for the population mean μ given the following values:
 - $\alpha = 0.05, n = 64, \bar{x} = 14.1, \sigma^2 = 4.32$
 - $\alpha = 0.01, n = 36, \bar{x} = 7.23, s^2 = 0.3047$
 - $\alpha = 0.10, n = 98, \bar{x} = 66.3, s^2 = 2.48$
- A random sample of n measurements is selected from a population with unknown mean μ and known standard deviation $\sigma = 20$. Calculate the width of the 95% confidence interval for μ for the following values of n :
 - $n = 75$
 - $n = 150$
 - $n = 300$
- Compare the confidence intervals you obtained in item 2. What effect does each of the following actions have on the width of a confidence interval?
 - Doubling the sample size
 - Quadrupling the sample size
- For a fixed value of the sample mean \bar{x} and sample standard deviation s , which one do you expect to be wider: the 90% confidence interval for μ or the 99% confidence interval for μ ? Explain.
- A commonly used IQ test is scaled to have a mean of 100 and a standard deviation of 15. A school counselor was curious about the average IQ of the students in her school and took a random sample of forty students' IQ scores. The average of these scores was 107.9. Find a 95% confidence interval for the mean student IQ in the school.
- A random sample of 100 car owners in Metro Manila shows that a car is driven on the average 13,500 kilometers per year with a standard deviation of 1,900 kilometers. Construct
 - a 90% confidence interval; and
 - a 99% confidence intervalfor the average number of kilometers a car is driven annually in Metro Manila.

7. Many companies are becoming more involved in *flextime*, in which a worker schedules his or her own work hours or compresses workweeks. A company that was contemplating the installation of a flextime schedule estimated that it needed a minimum mean of 7 hours per day for each assembly worker in order to operate effectively. Each of a random sample of 80 of the company's assemblers was asked to submit a tentative flextime schedule. Suppose that the mean number of hours per day was 6.7 hours and the standard deviation was 2.7 hours.
- Construct and interpret a 95% confidence interval for the mean number of hours worked per day for assemblers.
 - Would it be reasonable to believe that the mean number of hours worked is as low as 6.5 hours? Explain.
8. The manager of a grocery store found, on the basis of a random sample of 60 customers taken when the store was crowded, that it took customers who bought 10 or fewer items an average of 13.5 minutes of waiting in the express counter before they were able to check out their groceries and have them bagged. The standard deviation of the sample is 3.4 minutes.
- What can be asserted, with 98% confidence, about the maximum error in the estimate of 13.5 minutes of the true average waiting time it takes a customer to check out their groceries and have them bagged when the store is crowded?
 - Construct a 90% confidence interval for the true average time using the given data above.
9. A student conducted a study and reported that the 95% confidence interval for the mean ranged from 36 to 44. He was sure that the sample mean was 40, the sample standard deviation was 16, and that the sample size was at least 30, but could not remember the exact number. Using the given information, help the student determine that exact number.
10. Prove that, in general, 98% large-sample confidence intervals of the mean are about 19% wider than the corresponding 95% confidence intervals.

Lesson 3

Interval Estimation for a Proportion

Learning Outcomes

- At the end of this lesson, you should be able to
 - compute the confidence interval estimate of the population proportion; and
 - solve problems involving confidence interval estimation of the population proportion.

Introduction

The confidence intervals presented in the previous lesson focused on data on the ratio scale of measurement. That is, we used data which involve quantities such as income, distances, and times. We now wish to consider scenarios such as the following:

- A study by the Bangko Sentral ng Pilipinas (BSP) showed that only 11.5 percent of all Filipino adults maintained bank accounts as of the previous year.
- A Social Weather Stations survey in March 2018 found that 59 percent of Filipinos believed that the death penalty should be restored for those proven in court to have committed heinous crimes.
- A company representative claims that 10 percent of a fastfood chain's sales are made through the drive-thru window.

All of these examples illustrate the nominal scale of measurement. In the nominal scale, an observation is classified into one of two or more mutually exclusive groups. We are then interested in estimating the *proportion*, or the ratio or fraction of the sample or population that has a specific characteristic of interest.

We can construct a confidence interval for a population proportion. As in the case of the population mean, it is an interval around the corresponding point estimate, which is the sample proportion. Under a large sample assumption, the CLT yields the following confidence interval formula:

Confidence Interval for the Population Proportion (large sample)

A $(1 - \alpha) \times 100\%$ confidence interval for p is given by

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

where \hat{p} = estimated proportion based from the sample;

$\hat{q} = 1 - \hat{p}$;

n = sample size; and

$z_{\frac{\alpha}{2}}$ = z-value that leaves an area of $\frac{\alpha}{2}$ to the right.

Example 1

Find a 98% confidence interval for the population proportion if $\hat{p} = 0.4$ and $n = 47$.

Solution:

Since $1 - \alpha = 0.98$, $\alpha = 0.02$, and so $z_{\frac{\alpha}{2}} = z_{0.01} = 2.33$. Also, $\hat{q} = 1 - \hat{p} = 1 - 0.4 = 0.6$. We now substitute all these into the confidence interval formula:

$$\begin{aligned}\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) &= \left(0.4 - 2.33 \sqrt{\frac{(0.4)(0.6)}{47}}, 0.4 + 2.33 \sqrt{\frac{(0.4)(0.6)}{47}} \right) \\ &= (0.4 - 0.1665, 0.675 + 0.1665) \\ &= (0.2335, 0.8415)\end{aligned}$$

Example 2

A random sample of 80 adults was asked whether each favored a new city ordinance. Among these adults, 54 answered in the affirmative. Find and interpret a 95% confidence interval for the proportion of adults favoring the ordinance.

Solution:

The sample size is $n = 80$ while the point estimate for the proportion is $\hat{p} = \frac{54}{80} = 0.675$. Also, $\hat{q} = 1 - \hat{p} = 1 - 0.675 = 0.325$.

Since a 95% confidence interval is required, $\alpha = 0.05$, and so $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$.

Substituting these values into the above formula gives

$$\begin{aligned}\left(\hat{p} - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) &= \left(0.675 - 1.96 \sqrt{\frac{(0.675)(0.325)}{80}}, 0.675 + 1.96 \sqrt{\frac{(0.675)(0.325)}{80}} \right) \\ &= (0.675 - 0.1026, 0.675 + 0.1026) \\ &= (0.5724, 0.7776).\end{aligned}$$

This means that we can be 95% confident that between 57.24% and 77.76% of the population of adults favor the ordinance.

Recall that the *width* of a confidence interval is the difference of the upper limit and its lower limit. For a $(1 - \alpha)100\%$ confidence interval for a population proportion, this is given by

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) - \left(\hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) = 2z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

As in the case of the mean in lesson 2, half of this quantity can also be interpreted as the maximum allowable deviation, or maximum allowable error for the estimate of a population proportion. When $\alpha = 0.05$, this becomes $1.96 \times \sqrt{\frac{\hat{p}\hat{q}}{n}}$, which is the 95% margin of error formula that we encountered in lesson 1.

Points to Remember

1. A $(1 - \alpha) \times 100\%$ confidence interval is an interval constructed from the sample statistic where the value of the parameter is expected to lie. The confidence coefficient $(1 - \alpha)$ refers to the long run percentage of intervals constructed in this manner which contains the true value of the parameter.
2. A $(1 - \alpha) \times 100\%$ large-sample confidence interval for the population mean μ is given by

$$\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right).$$

If σ is unknown, we can replace it by the sample standard deviation s , provided the sample size is large.

3. A $(1 - \alpha) \times 100\%$ large-sample confidence interval for the population proportion p is given by

$$\left(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right).$$

Let's Practice

Analyze and solve each problem.

1. Find and interpret a 95% confidence interval for the population proportion given the following information:
 - a. $\hat{p} = 0.75, n = 200$
 - b. $\hat{p} = 0.27, n = 125$
 - c. $\hat{q} = 0.4, n = 200$
 - d. $\hat{q} = 0.11, n = 125$
2. Suppose that the value of the sample proportion is held fixed at $\hat{p} = 0.2$. Compute the widths of the resulting 98% confidence intervals of the following sample sizes:
 - a. $n = 100$
 - b. $n = 200$
 - c. $n = 400$
3. Compare the widths of the confidence intervals in Item 2. What effect does each of the following have on the widths of the confidence intervals for the population proportion?
 - a. Doubling the sample size
 - b. Quadrupling the sample size
4. In the Pulse Asia study referred in example 3 of lesson 1, 46% of the 1,200 adults surveyed cited improving or increasing the pay of workers as an issue that should be a top concern of the government. Use this information to construct a 90% confidence interval for the proportion of adults who believe that increasing wages is a top concern.
5. The owner of a gasoline station wishes to determine the proportion of its customers who pay using credit card or debit card. He decides to check the payment records of 200 customers and finds that 120 of them paid in cash. Construct a 95% confidence interval for the true proportion of customers who pay by credit or debit card .
6. In a random sample survey, 250 young professionals who work in Makati are asked whether they commuted to the city or brought their own car. If 88 of them commuted to work, construct a 99% confidence interval for the true proportion of those who commute to work.

7. Last year's records of car accidents occurring on a given section of highway were classified according to whether the resulting damage was at least ₱10,000 or less. The number of accidents involving injuries was also taken. The data are as follows:

	Less than ₱10,000	At least ₱10,000
Number of Accidents	42	31
Number Involving Injuries	10	23

- Give a point estimate for the true proportion of accidents involving injuries when the damage was at least ₱10,000 for similar sections of highway and find the margin of error corresponding to a 95% confidence level.
- Find a 95% confidence interval for the proportion in (a).

Lesson 4

The t -distribution

Learning Outcomes

- At the end of this lesson, you should be able to
 - illustrate and construct the t -distribution;
 - identify regions under the t -distribution corresponding to different t -values, as well as percentiles; and
 - solve problems involving a small-sample confidence interval estimate for the population mean.

Introduction

Suppose that the variance of the population from which we select our random sample is unknown. If the sample size n is at least 30, then the sample variance s^2 can be used to estimate σ^2 and the CLT still applies. However, if the sample size is small and the variance is unknown, the sampling distribution for the mean can no longer be approximated by the normal distribution.

It was the English statistician William Sealy Gosset who was one of the first to discover the distribution of these means. In 1908, he published a paper “The Probable Error of a Mean” in the *Biometrika* journal under the pseudonym “Student,” which discusses the details for the sampling distribution for the mean of a small sample. This distribution is now called the *Student’s t -distribution*, or more simply, the *t -distribution*.



William Sealy Gosset
(1876–1937)

The following are the properties of the t -distribution:

1. It is mound-shaped and symmetric about 0.
2. It is more variable than the standard normal distribution (see figure A). In particular, it does not approach the horizontal axis as quickly as z does.
3. Its shape is dependent on the sample size n (see figure B). Due to this, we often speak of having a *family* of t -distributions.
4. As n increases, the variability of the t -distribution decreases.
5. For large values of n , the t -distribution is approximately normal.

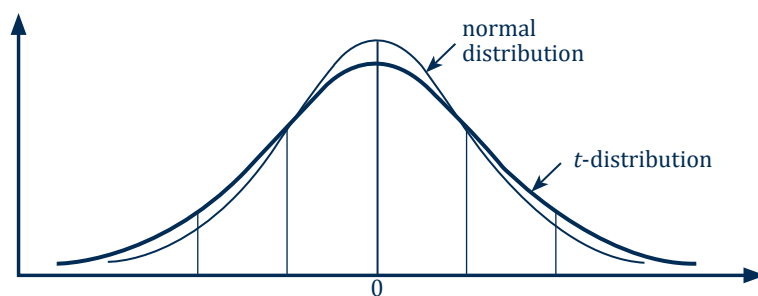


Figure A: Comparison of a standard normal distribution and a t -distribution (with $df = 5$)

When defining a t -distribution, it is necessary to specify the number of *degrees of freedom* (df). While the term originates from statistical theory, we can intuitively understand it by thinking of the number of observations that are free to vary when estimating the mean for a sample of size n .

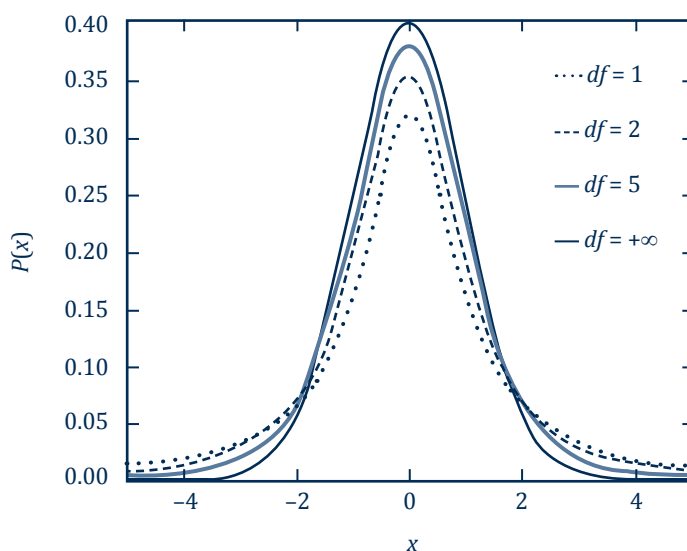


Figure B: Different t -distributions (having different degrees of freedom, df)

For example, suppose that we wish to choose 10 numbers x_1, x_2, \dots, x_{10} which have a mean of 8. This means that the 10 observations must have a sum of 80. Initially, we are free to select any values for the first nine observations. However, once these first nine values are selected, the 10th value is automatically determined: it must be 80 minus the sum of the first nine values chosen. Thus, for this set of 10 observations with a specified mean, we have 9 *degrees of freedom*, which is 1 less than the sample size.

In general, for a sample of size n , the corresponding t -distribution has $n - 1$ degrees of freedom. We shall use the notation $t_{\alpha, n-1}$ to denote the critical value of the t -distribution with $n - 1$ degrees of freedom and which leaves an area of α to its right.

Computing t -values

To compute critical values involving the t -distribution, statistical tables have been constructed which give the t -values for different levels of significance α and different degrees of freedom. A portion of the table in *Appendix C* is given below.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10

A portion of the t -table

To look for the corresponding t -value, you need to look for the corresponding level of significance (α) at the column header and the corresponding number of degrees of freedom (df) in the first column.

For example, suppose we wish to find the t -value that leaves an area of 0.05 to the right and with 6 degrees of freedom, $t_{0.05, 6}$, which has a value of 1.943. To obtain this using the t -table, we look at the column containing $t_{.050}$ and the row corresponding to 6. The entry in the table in this row and in this column is the required t -value, as shown in the figure below.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.314	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.132	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10

$t_{0.050, 6} = 1.943$

Determining $t_{0.050, 6}$ using the t -table

Example 1

For a t -distribution with 14 degrees of freedom, the t -value that leaves an area of 0.025 to the right can be found by looking at the row for $df = 14$ and the column for $t_{.025}$. In this case, we obtain $t_{0.025,14} = 2.145$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	df
1	3.078	6.314	12.706	1
2	1.886	2.920	4.303	2
3	1.638	2.353	3.182	3
4	1.533	2.132	2.776	4
5	1.476	2.015	2.571	5
6	1.440	1.943	2.447	6
7	1.415	1.895	2.365	7
8	1.397	1.860	2.306	8
9	1.383	1.833	2.262	9
10	1.372	1.812	2.228	10
11	1.363	1.796	2.201	11
12	1.356	1.782	2.179	12
13	1.350	1.771	2.160	13
14	1.345	1.761	2.145	14
15	1.341	1.753	2.131	15

Example 2

To find the t -value with 14 degrees of freedom that leaves an area of 0.975 to the right, $t_{0.975,14}$, or equivalently, 0.025 to the left, we can use the symmetry of the t -distribution about the y -axis. That is, the required t -value is the same as the *negative* of the one that leaves an area of 0.025 to the right. Thus, $t_{0.975,14} = -t_{0.025,14} = -2.145$.

Example 3

A random variable T has a t -distribution. If $P(T < 1.708) = 0.95$, how many degrees of freedom does T have?

Solution:

A portion of the table is shown to determine the corresponding df such that $t_{0.05, df} = 1.708$.

df	$t_{.100}$	$t_{.050}$	df	$t_{.100}$	$t_{.050}$
1	3.078	6.314	16	1.337	1.746
2	1.886	2.920	17	1.333	1.740
3	1.638	2.353	18	1.330	1.734
4	1.533	2.132	19	1.320	1.729
5	1.476	2.015	20	1.325	1.725
6	1.440	1.943	21	1.323	1.721
7	1.415	1.895	22	1.321	1.717
8	1.397	1.860	23	1.319	1.714
9	1.383	1.833	24	1.318	1.711
10	1.372	1.812	25	1.316	1.708
11	1.363	1.796			
12	1.356	1.782			
13	1.350	1.771			
14	1.345	1.761			
15	1.341	1.753			

Since $P(T < 1.708) = 0.95$, we have $P(T > 1.708) = 0.05$. If we let df be the required number of degrees of freedom, this means that we need to find df such that $t_{0.05, df} = 1.708$. Looking at the values on the column of $t_{0.05}$ in Appendix C, we see that $df = 25$.

Example 4

Let T be a random variable having a t -distribution with 9 degrees of freedom. Find

1. $P(-1.833 < T < 3.25)$.
2. the 90th percentile of T .

Solution:

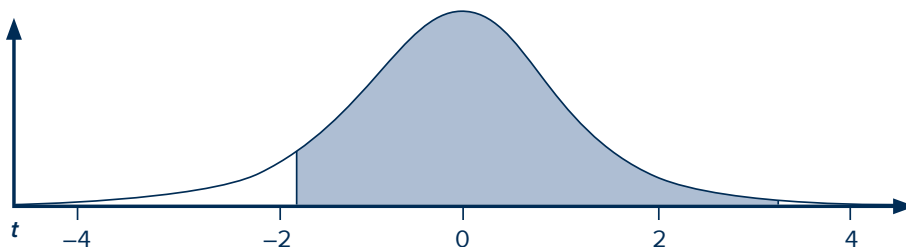
1. Note that $P(-1.833 < T < 3.25) = P(T < 3.25) - P(T < -1.833)$. We compute each of these probabilities individually.

Looking at *Appendix C*, we know that T leaves an area of 0.005 to the right. Thus,

$$P(T < 3.25) = 1 - P(T > 3.25) = 1 - 0.005 = 0.995.$$

Furthermore, the value 1.833 corresponds to $t_{0.05}$, which leaves an area of 0.05 to the right. By the symmetry of the t -distribution, -1.833 leaves an area of 0.05 to the left. That is, $P(T < -1.833) = 0.05$.

The figure below shows the corresponding area for $P(-1.833 < T < 3.25)$.



Combining these two results, we obtain

$$P(-1.833 < T < 3.25) = 0.995 - 0.05 = 0.945.$$

2. The 90th percentile is the value k for which $P(T \leq k) = 0.90$. That is, it is the t -value that leaves an area of 0.90 to its left. This is just $t_{0.10, 9} = 2.821$.

The t -distribution and Confidence Intervals

Assuming the population is approximately normal, we can use the t -distribution to construct a confidence interval for the mean when the sample size is small ($n < 30$) and the population standard deviation is unknown.

Confidence Interval for the Population Mean (small sample, σ is unknown)

A $(1 - \alpha) \times 100\%$ confidence interval for μ is given by

$$\left(\bar{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right)$$

where \bar{x} = estimated mean based from the sample;

$t_{\frac{\alpha}{2}, n-1}$ = t -value with $n - 1$ degrees of freedom that leaves an area of $\frac{\alpha}{2}$ to the right;

s = sample standard deviation; and

n = sample size.

Example 5

A random sample of 12 graduates of a certain secretarial school typed an average of 73.9 words per minute (wpm) with a standard deviation of 8.7 wpm. Assuming that the number of words typed per minute is approximately normally distributed, find a 95% interval for the average number of words typed by all the graduates of this school.

Solution:

From the given, we have $n = 12$, $\bar{x} = 73.9$, and $s = 8.7$. Notice that the sample size is small ($n < 30$) and the population standard deviation is unknown, so the t -distribution must be used.

Since we require a 95% confidence interval, $1 - \alpha = 0.95$, and so $\frac{\alpha}{2} = 0.025$. The number of degrees of freedom is $n - 1 = 12 - 1 = 11$. Thus, $t_{\frac{\alpha}{2}, n-1} = t_{0.025, 11} = 2.201$, and so, a 95% confidence interval for the mean number of words typed by the graduates of the school is

$$\begin{aligned}\left(\bar{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right) &= \left(73.9 - 2.201 \cdot \frac{8.7}{\sqrt{12}}, 73.9 + 2.201 \cdot \frac{8.7}{\sqrt{12}} \right) \\ &= (73.9 - 5.53, 73.9 + 5.53) \\ &= (68.37, 79.43).\end{aligned}$$

Example 6

A machine produces cylindrical metal pieces. A sample of pieces is taken, and the diameters are found to be 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 cm. Find a 99% confidence interval for the mean diameter of pieces from this machine. Assume that the diameters are approximately normally distributed. Round your answers to four decimal places. Based on this interval, is it believable that the mean diameter of the pieces produced by this machine is 1 cm?

Solution:

From the given data, we obtain the sample mean \bar{x} and sample variance s^2 as follows:

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^9 X_i}{9} = \frac{1.01 + 0.97 + 1.03 + \dots + 1.03}{9} \approx 1.0056 \\ s^2 &= \frac{\sum_{i=1}^9 (x_i - \bar{x})^2}{9 - 1} = \frac{(1.01 - 1.0056)^2 + \dots + (1.03 - 1.0056)^2}{8} \approx 0.0006\end{aligned}$$

Thus, the sample standard deviation is $\sqrt{0.0006} \approx 0.0246$. (Alternatively, these can be computed using any calculator with a STAT mode or through software such as MS Excel.) Since the sample size $n = 9$ is small and the population standard deviation is unknown, we use the formula involving the t -distribution above. Here, $\alpha = 0.01$, and the number of degrees of freedom is $n - 1 = 8$. Thus, the critical value is $t_{0.005,8} = 3.355$. A 99% confidence interval is then

$$\begin{aligned} \left(\bar{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right) &= \left(1.0056 - 3.355 \cdot \frac{0.0246}{\sqrt{9}}, 1.0056 + 3.355 \cdot \frac{0.0246}{\sqrt{9}} \right) \\ &= (1.0056 - 0.0275, 1.0056 + 0.0275) \\ &= (0.9781, 1.0331). \end{aligned}$$

Since the value 1 lies within this interval, there is no reason to say that the true mean diameter is not 1 cm.

Points to Remember

1. Assuming approximate normality of the population, the t -distribution models the sampling distribution of the mean when the sample size is small and the population standard deviation is unknown.
2. The t -distribution is mound-shaped and symmetric like the standard normal distribution. However, it approaches the horizontal axis more slowly than the standard normal distribution.
3. As the number of degrees of freedom increases, the t -distribution approaches the standard normal distribution.
4. When the sample size is small and the population standard deviation is unknown, we can construct a $(1 - \alpha) \times 100\%$ confidence interval for μ by using the formula

$$\left(\bar{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{s}{\sqrt{n}} \right).$$

Let's Practice

I. Write True if the statement is correct; otherwise, write False.

- _____ 1. The shape of the normal distribution approaches the t -distribution as its standard deviation becomes larger.
- _____ 2. The t -distribution has a smaller area at its tails than that of the normal distribution.
- _____ 3. When using the t -distribution to construct a confidence interval, we need to assume that the population is normally distributed.
- _____ 4. For a sample of size 64 from a population with unknown shape, it is necessary for the population to be approximately normally distributed to construct a confidence interval.
- _____ 5. For a sample of size 24 from a normal population with unknown variance, the sampling distribution of the mean has a t -distribution with 23 degrees of freedom.

II. Analyze and solve each problem.

- 1. Evaluate the following probabilities:
 - a. $P(T < 2.365)$ when $df = 7$
 - b. $P(T > 1.318)$ when $df = 24$
 - c. $P(-1.356 < T < 2.179)$ when $df = 12$
 - d. $P(T < -2.567)$ when $df = 17$
- 2. Find the following critical values of the t -distribution:
 - a. $t_{0.10, 10}$
 - b. $-t_{0.025, 14}$
 - c. $t_{0.995, 6}$
- 3. Let T be a random variable having a t -distribution with 23 degrees of freedom. Find the value of k such that
 - a. $P(-2.807 < T < k) = 0.095$.
 - b. $P(k < T < 2.069) = 0.965$.
 - c. $P(-k < T < k) = 0.90$.

4. For a random variable with a t -distribution having 19 degrees of freedom, find
 - a. the 95th percentile.
 - b. the 99th percentile.
 - c. the 10th percentile.
5. There is strong evidence supporting the claim that sugar-sweetened soft drinks contribute to the development of diabetes. AJ, curious about why this is the case, examined the sugar content (in grams per serving) of a sample of eight different sugar-sweetened drinks, and obtained an average of 32 grams with a standard deviation of 4 grams. Use this information to construct a 98% confidence interval for the true average sugar content per serving of sweetened drinks. Assume that the sugar contents are normally distributed.
6. A tire manufacturer wishes to investigate the tread life of its tires. A sample of 10 tires driven for 30,000 km revealed a sample mean of 0.12 cm of tread remaining with a standard deviation of 0.035 cm.
 - a. Construct a 99% confidence interval for the population mean amount of tread remaining for the tires of this manufacturer. What assumption did you use?
 - b. Would it be reasonable to conclude that after 30,000 km, the population mean amount of tread remaining is 0.11 cm?
7. DJ Printer Company is planning to introduce a new line of desk jet printers to the market. As part of its advertising campaign, it would like to include the number of pages a user can expect from a print cartridge. A sample of 15 cartridges revealed the following number of pages printed:

2,698	2,395	1,927	2,379	2,099
2,028	2,372	3,006	2,214	2,385
2,474	2,475	2,334	1,995	2,361

Develop a 95% confidence interval for the population mean.

8. A study was conducted to determine the effect of cigarette smoking on the carbon monoxide diffusing capacity of the lung. The carbon monoxide diffusing capacities of a random sample of 20 current smokers are listed below.

104	89	73	123	91
92	62	91	84	76
101	88	71	82	89
103	109	73	107	90

- Construct a 90% confidence interval for the mean carbon monoxide diffusing capacity of current smokers.
 - Based on the interval in (a), would you agree that the mean diffusing capacity could be 100? Explain.
9. A sample of size $n > 30$ is drawn. If the mean and standard deviation of the sample are \bar{x} and s , respectively, approximately what level of confidence is associated with the following intervals?

a. $\left(\bar{x} - 1.51 \cdot \frac{s}{\sqrt{n}}, \bar{x} + 1.51 \cdot \frac{s}{\sqrt{n}} \right)$

b. $\left(\bar{x} - 0.96 \cdot \frac{s}{\sqrt{n}}, \bar{x} + 1.06 \cdot \frac{s}{\sqrt{n}} \right)$

Lesson 5

Finding the Sample Size for Estimating Population Parameters

• Learning Outcomes

- At the end of this lesson, you should be able to
 - compute an appropriate sample size using the length of the interval; and
 - solve problems which involve determining the sample size.

Introduction

One of the most important things that a researcher must consider when conducting a statistical study is the sample size. If the sample size is too small, then there is a high possibility that the estimates may be unreliable. On the other hand, if the sample size is too large, it might be too costly or time-consuming for the researcher. The higher the accuracy we demand for the resulting estimate, the larger the size of the sample we need to obtain.

Below are the three factors which affect the sample size used to estimate a population parameter.

1. *The amount of error (E).* Since sampling is random, we do not expect an exact value of the population parameter when computing an estimate based from the sample. The more accuracy demanded for the resulting estimate, the larger the sample size required.
2. *Degree of confidence ($1 - \alpha$) of the accuracy of the estimate.* This refers to the probability that the amount of error of the estimate will not exceed the amount of error E .
3. *Variability in the population.* The more homogenous the population, the smaller the sample size necessary to achieve the same level of confidence that the error of estimate will not exceed a certain amount. This variability is usually given by the standard deviation of the population and can be based on prior knowledge of the population, or through a pilot study.

The simplest formula that takes into consideration each of the factors mentioned previously is as follows:

Sample Size for Estimating a Population Mean

To be $(1 - \alpha) \times 100\%$ confident that the estimate for the population mean is within E units of the true value, the minimum sample size n necessary is given by

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$$

where $Z_{\frac{\alpha}{2}}$ = z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

σ = estimated standard deviation of the population; and

E = maximum allowed error in the estimate.

The formula above is based on the maximum allowed error of a confidence interval for the population mean as given in lesson 2. To have a maximum deviation E on each side around the point estimate \bar{x} , we need

$$E = Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Solving for n gives

$$\sqrt{n} = \frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E}$$

$$n = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2$$

Example 1

A store manager wishes to be 90% confident that his estimate for the mean monthly family grocery expense is correct within ± 500 pesos. Based on prior information, he believes that the monthly family grocery expense follows a normal distribution, and has arrived at an estimate of ₱9,000 for the standard deviation. Determine the minimum number of families to be taken as sample to meet his criteria.

Solution:

It is given that $\alpha = 0.1$ so $z_{\frac{\alpha}{2}} = z_{0.05} = 1.645$. Also, it is given that $E = 500$ and $\sigma = 9,000$.

Substituting these values into the sample size formula gives

$$n = \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{E} \right)^2 = \left[\frac{(1.645)(9,000)}{500} \right]^2 = 876.75 \approx 877.$$

Therefore, he needs at least 877 families as sample.

The next formula gives the minimum sample size required to estimate a population proportion with a specific accuracy.

Sample Size for Estimating a Population Proportion

To be $(1 - \alpha) \times 100\%$ confident that the estimate for the population proportion is within E units of the true value, the minimum sample size n necessary is

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \hat{p}\hat{q}$$

where $z_{\frac{\alpha}{2}}$ = z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

E = maximum allowed error in the estimate;

\hat{p} = the prior estimate for the proportion; and

$\hat{q} = 1 - \hat{p}$.

The formula above is based on the maximum allowed error of a confidence interval for a population proportion as given in lesson 3. For this quantity to be E , we need

$$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\hat{q}}{n}}.$$

Solving for n , we have

$$\begin{aligned} \sqrt{n} &= \frac{z_{\frac{\alpha}{2}} \sqrt{\hat{p}\hat{q}}}{E} \\ n &= \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \hat{p}\hat{q}. \end{aligned}$$

Note that E must be expressed in decimal. For example, if the maximum allowed error is 3%, then $E = 0.03$. Moreover, the prior estimate \hat{p} may be based on the opinion of the researcher, from past studies, or from a pilot study. In case such an estimate is not available, it is customary to choose \hat{p} for which $\hat{p}\hat{q} = \hat{p}(1 - \hat{p}) = \hat{p} - \hat{p}^2$ is maximized. This occurs when $\hat{p} = 0.5$.

Example 2

A government office wishes to audit 1,256 new appointments to estimate the proportion p who have been incorrectly processed by its payroll department.

1. What must be the sample size for the sample proportion to have a 95% chance of lying within 0.05 of p ?
2. Past audits suggest that p will not be greater than 0.1. Using that information, recalculate the sample size required in item (1).

Solution:

1. Since there is no prior estimate for p , we choose $\hat{p} = 0.5$, and so $\hat{q} = 1 - \hat{p} = 0.5$. From the given information, the maximum error allowed is $E = 0.05$. Since we require 95% confidence from our sample, $\alpha = 0.05$, so $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$. Thus, the required sample size is

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \hat{p}\hat{q} = \left(\frac{1.96}{0.05} \right)^2 (0.5)(0.5) = 384.16 \approx 385.$$

2. As in part (1), $E = 0.05$ and $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$. However, since a prior estimate of the population proportion is known, we can set $\hat{p} = 0.1$. This results in a smaller sample size as follows:

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{E} \right)^2 \hat{p}\hat{q} = \left(\frac{1.96}{0.05} \right)^2 (0.1)(0.9) = 138.3 \approx 139$$

Points to Remember

1. Always round up sample size estimates to the next larger integer, even if the tenths place is less than 0.5. This ensures that the level of confidence will never fall below $(1 - \alpha) \times 100\%$.
2. The sample size formulas discussed do not depend on the size of the population from which the sample is drawn. This means that once a specific maximum error and degree of confidence are given, the same sample size applies regardless of whether the population size is 200 or 2,000,000.

As a final note, let us examine how survey firms such as Social Weather Stations (SWS) or Pulse Asia obtain sample sizes for their surveys. Below is a news article on the 2016 presidential election season reporting on a Pulse Asia survey.

Grace Poe, Duterte top Pulse Asia's March survey

(3rd UPDATE) Poe and Duterte mark similar increases of 2 percentage points from the previous February survey, while Binay drops by 3 percentage points. Roxas improves by only 1 percentage point.

Source: <http://www.rappler.com/nation/politics/elections/2016/125867-grace-poe-rodry-duterte-pulse-asia-march-survey>, retrieved 19 April 2017.

From the article, we can say that the study is based on a nationwide survey with the following specifications:

The Pulse Asia survey was conducted among 2,600 registered voters, with biometrics, and has a $\pm 1.9\%$ error margin at the 95% confidence level.

To obtain the sample size used in the study, we set $E = 1.9\%$ and confidence coefficient to be $1 - \alpha = 0.95$. Then $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$. Choosing $\hat{p} = 0.5$, we obtain the minimum sample size to estimate the proportion of registered voters preferring a specific candidate as

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{E}\right)^2 \hat{p}\hat{q} = \left(\frac{1.96}{0.019}\right)^2 (0.5)(0.5) = 2,660.39 \approx 2,661$$

which the survey firm has approximated as 2,600. This is, theoretically, the appropriate sample size for the specified margin of error and confidence coefficient, regardless of the size of the population of registered voters in the Philippines.

The following table shows the approximate sample size for estimating a proportion for different error margins, assuming a 95% confidence coefficient and no prior knowledge on a population proportion:

Sample size	Margin of Error
200	7.1%
400	5.0%
700	3.8%
1,000	3.2%
1,200	2.9%
1,500	2.6%

Sample size	Margin of Error
2,000	2.2%
3,000	1.8%
4,000	1.6%
5,000	1.4%

Notice that as the sample size increases, the corresponding improvement in the margin of error decreases.

Let's Practice

Solve each problem.

1. A population is estimated to have a standard deviation of 10 units. If the population mean is to be estimated with a minimum error of 2 units, how large must be the sample if we require
 - a. a 95% level of confidence?
 - b. a 99% level of confidence?
2. The estimate of the population proportion is to be within 0.10 and the best estimate of the population proportion is 0.45. How large must the sample be if we require
 - a. a 95% level of confidence?
 - b. a 99% level of confidence?
3. The university registrar wants to estimate the mean grade point average (GPA) of all graduates during the past 10 years. The mean GPA is to be estimated within 0.05 of the population mean. The standard deviation is estimated to be 0.297. Using a 99% level of confidence, how many graduates should be included in the study?
4. A fruit wholesaler packs mangoes into boxes of 20 for shipment. To determine the average weight of these boxes in preparation for shipment, a few of these boxes are weighed. The mean weight is 9.3 kg, with a standard deviation of 0.23 kg. How many boxes must the wholesaler sample to be 95% confident that the sample mean does not differ from the population mean by more than 0.1 kg?

5. In the Pulse Asia study given in example 3 of lesson 1, 31% of the 1,200 adults polled said that fighting criminality is one of the top three urgent national issues. How large must a sample be if a similar study who wishes to be 90% confident that the estimated percentage will be within 2% of the true percentage is to be conducted?
6. Past surveys indicate that 26% of tourists in Macau gambled at the city's casinos. A research organization conducted a new study to update this information.
 - a. The new study used a 90% confidence level. If the estimate must be within 1% of the true population proportion, what must be the sample size?
 - b. What sample size will be necessary if no prior estimate were available?
 - c. The organization said that the sample size determined above is too large. What can be done to reduce the sample size? Explain.

Software Tutorial in MS Excel

Probabilities of the t -distribution and t -values

1. Cumulative left-tail areas in the t -distribution can be obtained using the command “=T.DIST(x,deg_freedom,cumulative)”. Here, “deg_freedom” refers to the number of degrees of freedom of our desired t -distribution. Since we are getting cumulative left-tail probabilities, we set the value of “cumulative” to 1 (or TRUE).

For example, to find the cumulative left-tail area of 2.365 for a random variable T which has a t -distribution with degrees of freedom of 7, we type in “=T.DIST(2.365,7,1)” on the command line to obtain 0.975014.

	A	B	C
1	0.975014	=T.DIST(2.365,7,1)	

2. To find the corresponding t -value in the t -distribution that leaves a cumulative left-tail area of size α , we use the command “=T.INV(α ,deg_freedom)”. As before, “deg_freedom” refers to the number of degrees of freedom of the t -distribution.

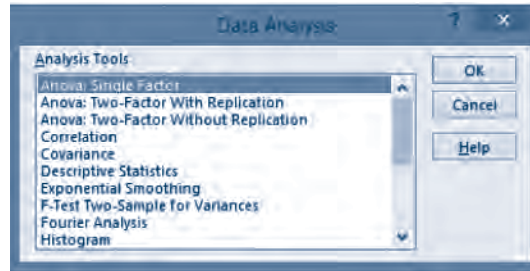
For example, to find the t -value for a t -distribution with degrees of freedom of 23 that leaves an area of 0.01 to the right (and therefore, 0.99 to the left), we type in “=T.INV(0.99,23)” on the command line and obtain the value of 2.499867.

	A	B	C
1	2.499867	=T.INV(0.99,23)	

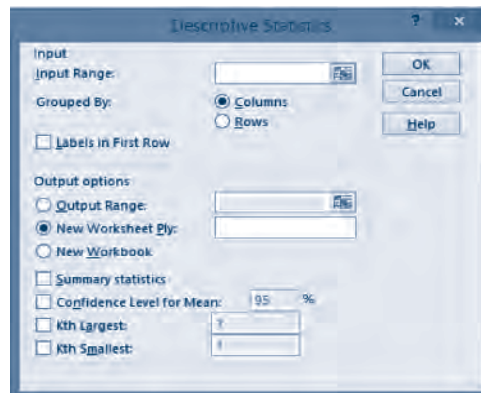
Small-sample confidence interval for the mean

While there is no direct command to compute for a small-sample confidence interval for the mean, the *Descriptive Statistics* option under *Data Analysis* can give the maximum allowable deviation $t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$. This can then be combined with the sample mean to compute a confidence interval. To do this, we have the following steps:

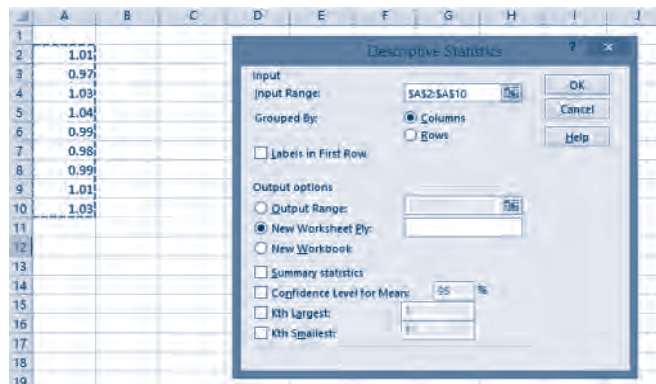
Step 1: From the *Data* ribbon, click the *Data Analysis* button. This displays the list of available *Analysis Tools*, as shown below.



Step 2: From the list of *Analysis Tools*, select *Descriptive Statistics* then click *OK*. Then the following screen will appear:



Step 3: Click the  button next to *Input Range*. This brings you back to your current worksheet, allowing you to select the cells corresponding to the population.



Step 4: Tick the *Summary Statistics* and the *Confidence Level for Mean* checkboxes. Specify the desired confidence level on the provided text box then click *OK*. This gives the following output:

	A	B
1	Column1	
2		
3	Mean	1.005556
4	Standard Error	0.008184
5	Median	1.01
6	Mode	1.01
7	Standard Deviation	0.024552
8	Sample Variance	0.000603
9	Kurtosis	-1.42778
10	Skewness	0.011798
11	Range	0.07
12	Minimum	0.97
13	Maximum	1.04
14	Sum	9.05
15	Count	9
16	Confidence Level(95.0%)	0.018872
17		

From the output, the sample mean \bar{x} is 1.0056 (see cells A3 and B3) and the 95% margin of error is 0.0189 (see cells A16 and B16). This means that a 95% confidence interval for μ is

$$(1.0056 - 0.0189, 1.0056 + 0.0189) = (0.9867, 1.0245).$$

Alternatively, one can compute the maximum allowable deviation using the Excel command “=CONFIDENCE.T(alpha,standard_dev,size)”. Here, “alpha” refers to the value of 1 minus the confidence level of the desired interval, “standard_dev” is the standard deviation of the sample, and “size” is the sample size.

For example, suppose we wish to compute the maximum allowable deviation for the data where the given output is based. Since we wish to compute for a 95% confidence interval, $1 - \alpha = 0.95$, so “alpha” = 0.05. Based on the *Summary Statistics* output, we also have “standard_dev” = 0.024552 and “size” = 9. This gives the value $0.018872 \approx 0.0189$ which we have obtained previously.

	A	B	C	D	E
1	0.018872	=CONFIDENCE.T(0.05,0.024552, 9)			

Note: A similar Excel command “=CONFIDENCE.NORM(alpha,standard_dev,size)” can be used to compute for the maximum available deviation when the standard normal distribution applies instead.

Chapter Review

- A rule, usually expressed as a formula, that tells us how to compute an estimate based on a sample is called an **estimator**. In *point estimation*, a single number is calculated based on information in the sample. In this case, the formula is called a *point estimator*, while the computed value is called a *point estimate*.
- The distance between an estimate and the parameter being estimated is called the **error of estimate**. To estimate the population mean μ , we use the point estimator with an estimated standard error of $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and a 95% margin of error of $1.96 \times SE_{\bar{x}}$ when $n \geq 30$.
- To estimate the population proportion p , we use the point estimator with an estimated standard error of $SE_{\hat{p}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$ and a 95% margin of error of when $n\hat{p} > 5$ and $n\hat{q} > 5$.
- A **$(1 - \alpha) \times 100\%$ confidence interval** is an interval constructed from the sample statistic where the value of the parameter is expected to lie.
- The **confidence coefficient**, denoted by $1 - \alpha$, is the probability that a confidence interval will contain the estimated parameter.
- **Confidence Interval for the Population Mean (large sample, σ is known)**
A $(1 - \alpha) \times 100\%$ confidence interval for μ is given by

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

where \bar{x} is the sample mean;

$z_{\alpha/2}$ is the z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

σ is the population standard deviation; and

n is the sample size.

- **Confidence Interval for the Population Mean (large sample, σ is unknown)**

A $(1 - \alpha) \times 100\%$ confidence interval for μ is given by

$$\left(\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$

where \bar{x} is the sample mean;

$z_{\alpha/2}$ is the z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

s is the sample standard deviation; and

n is the sample size.

- The **width** of the confidence interval is the difference between its upper and lower limits.

- **Confidence interval for the population proportion (large sample)**

A $(1 - \alpha) \times 100\%$ confidence interval for p is given by

$$\left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \right)$$

where \hat{p} is the estimated proportion based from the sample;

$$\hat{q} = 1 - \hat{p};$$

$z_{\alpha/2}$ is the z-value that leaves an area of $\frac{\alpha}{2}$ to the right; and

n is the sample size.

- The t -distribution models the sampling distribution of the mean when the sample size is small and the population standard deviation is unknown.

- **Confidence Interval for the Population Mean (small sample, σ is unknown)**

A $(1 - \alpha) \times 100\%$ confidence interval for μ is given by

$$\left(\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right)$$

where \bar{x} is the sample mean;

$t_{\alpha/2, n-1}$ is the t -value with $n - 1$ degrees of freedom that leaves an area of $\frac{\alpha}{2}$ to the right;

s is the sample standard deviation; and

n is the sample size.

- To be confident that the *estimate for the population mean* is within E units of the true value, the minimum sample size n necessary is

$$n = \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2$$

where $z_{\alpha/2}$ is the z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

σ is the estimated standard deviation of the population; and

E is maximum allowed error in the estimate.

- To be confident that the *estimate for the population proportion* is within E units of the true value, the minimum sample size n necessary is

$$n = \left(\frac{z_{\alpha/2}}{E} \right)^2 \hat{p} \hat{q}$$

where $z_{\alpha/2}$ is the z-value that leaves an area of $\frac{\alpha}{2}$ to the right;

E is maximum allowed error in the estimate; and

\hat{p} is the estimated proportion based from the sample; and

$\hat{q} = 1 - \hat{p}$.

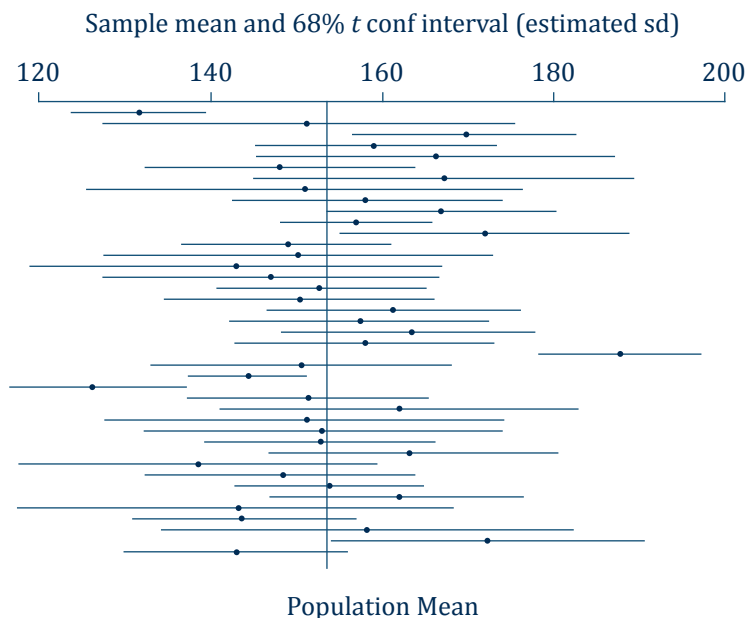
Chapter Performance Tasks

1. Mean Weights and Confidence Intervals

Imagine again that you are a school biostatistician. This time, your school clinic supervisor tasked you to make an analysis of the mean weight of students. Start with one class of grade 11 students and collect their weights (in kg). To select your sample, write each weight measurement on a slip of paper, put the slips into a small bag, mix them up, pick out 5 slips at random, and write down the numbers you picked. Then, return the slips in the bag and repeat the process until you have 20 samples of 5 numbers. Include the following in your analysis:



- For each of the samples of weights obtained, compute the resulting point estimate and 95% margin of error for the mean weight of students in the class.
- For each of the 20 samples, construct the resulting 68% confidence interval for the average weight of all the students in the class.
- Compute the mean and standard deviation of the population of weights.
- Plot all your confidence intervals on a horizontal axis as horizontal segments, stacked vertically. Then mark using a vertical line the true mean population weight that you obtained in part (c). Your diagram should look something like the figure below.



The vertical line represents the true mean weight of the population of students in the class.

Answer the questions below.

- How many of the 20 confidence intervals would you expect to contain the true mean?
- Based from your results, how many actually contain the true mean?

Submit a written report containing the summary of your analysis and findings. Make sure that it is accurate, neat, detailed, and organized.

2. K-Pop

A modern genre of music that has gained popularity over the past few years is Korean pop music, or more popularly known as K-Pop. The French Institut National de L'audiovisuel defined “K-Pop” as a fusion of synthesized music, sharp dance routines, and fashionable, colourful outfits.



Imagine that you are a feature writer of your school's newsletter, and the current theme of the newsletter revolves around genres of music and their impact to the listeners. For your article entry, you want to make sure first that the values and figures you will use are statistically accurate. Conduct a survey among your classmates to determine their interest in K-Pop. In particular, ask them whether

- they listen to K-Pop or watch K-Pop videos or not; and
- for those who listen to K-Pop or watch K-Pop videos, the reasons they listen to K-Pop. (Some typical reasons may include (a) the music itself, (b) the dance moves, (c) the fashion, and (d) the personalities/singers involved, among others. They may choose more than one of these reasons.)

Tabulate the results and summarize the number of students in class who fall in each category.

- Using the survey results, what is your point estimate and margin of error for the proportion of students who listen to K-Pop?
- Construct a 99% confidence interval estimate for the true proportion of students who listen to K-Pop.
- Construct a 90% confidence interval estimate for the true proportion of students who listen to K-Pop. Would you expect it to be shorter or longer than your confidence interval in (b)?
- Develop 99% confidence intervals for the true proportion of students who listen to or watch K-Pop for each of the given reasons in your list.
- If you wanted to be 95% confident of estimating the proportion of students who listen to K-Pop within 0.03 of the true value, what would be an appropriate sample size to draw from the population of students?

Chapter Exercises

Solve each problem.

1. A certain manufacturer produces light bulbs that have a life span that is approximately normally distributed with a standard deviation of 40 hours.
 - a. What would be the best estimate for the mean life span of all light bulbs produced by this manufacturer if a sample of 30 bulbs has an average life span of 735 hours? What would be the 95% margin of error of this estimate?
 - b. What must be the sample size if you wish to be 96% confident that your sample is within 10 hours of the true mean?
2. A survey was administered to 300 basketball fans and it revealed that 63 answered Team G as their favorite team.
 - a. Give a point estimate for the true proportion of fans whose favorite is Team G.
 - b. Compute a 95% confidence interval for the proportion of fans whose favorite is Team G.
3. In a survey of randomly selected households, 962 have computers while 288 do not have computers.
 - a. Find and interpret a 98% confidence interval for the proportion of households with computers.
 - b. What must be the sample size to have a 95% reliability of having a maximum allowable difference of 0.02 from the population proportion?
4. High airline occupancy rates on scheduled flights are essential to the profitability of an airline company. Suppose that a scheduled flight must average at least 60% occupancy in order to be profitable. Suppose further that an examination of the occupancy rate for an airline company's 120 10:00 a.m. flights from Manila to Bacolod showed a mean occupancy per flight of 58%, with a standard deviation of 11%. Use this information to develop a 90% confidence interval for the mean occupancy for this flight. Based on this interval, would you agree that this flight is profitable? Explain.
5. The Greater Manila Area Chamber of Commerce wants to estimate the mean travel time going to work of those who are employed in the main business district. A sample of 15 employees reveals an average travel time of 35 minutes with a standard deviation of 6 minutes. Develop a 98% confidence interval for the population mean.

6. The proportion of public accountants who have changed jobs within the last three years is to be estimated within 3%. The 95% level of confidence is to be used. A study conducted several years ago revealed that the percentage of public accountants changing jobs within the last three years was 21%.
 - a. If there is a need to update this information, how many public accountants must be included in the study?
 - b. How many public accountants should be interviewed if no previous estimates of the population proportion are available?
7. A geneticist is interested in the proportion of African males who have a certain minor blood disorder. In a random sample of 100 African males, 27 are found to be afflicted with this disorder.
 - a. Compute a 99% confidence interval for the proportion of African males who have this blood disorder.
 - b. What can we assert with 99% confidence about the possible size of our error if we estimate the proportion of African males with this blood disorder to be 0.27?
8. An automated teller machine (ATM) was installed in the corporate offices of the Laruan Atbp. Company. The ATM was for the exclusive use of Laruan's 405 employees. After half a year of operation, a sample of 75 employees of Laruan's revealed the following information regarding their use of the ATM in a month:

Number of Times ATM is Used	Frequency
0	15
1	20
2	25
3	10
4	5

- a. Construct a 93% confidence interval for the proportion of employees who do not use the ATM in a month.
 - b. Develop a 99% confidence interval for the mean number of transactions in a month.
9. Verify that for random samples of size 5 from normal populations, 99% confidence intervals of the mean are almost 66% wider than the corresponding 95% intervals.

Chapter 6

Tests of Hypotheses



When your doctor prescribes you any medicine, how does he or she know that it will be helpful for your particular illness? He or she will, of course, depend on his or her experience, medical knowledge and training, and the information provided by *clinical trials* for the said drug. These clinical trials typically involve four phases of testing on different groups of people to judge the drug's safety and side effects. These tests also include determining whether the drug is significantly more effective than either a placebo or an existing medicine or treatment for an illness. Many of the steps in these clinical trials are essentially *tests of hypotheses* on the proportion of subjects who experience relief or get cured due to the new medication. Such tests of hypotheses on proportions, as well as those for means, will be the focus of this chapter.

Lesson 1

The Hypothesis Testing Procedure

Learning Outcomes

- At the end of this lesson, you should be able to
 - illustrate null and alternative hypotheses, level of significance, rejection region, and types of error in hypothesis testing; and
 - identify the parameter to be tested given a real-life problem.

Introduction

In the previous chapter, you have learned how to estimate the value of a population parameter with a certain degree of confidence based on the data obtained from a sample. We will now examine the second common task in inferential statistics: testing hypotheses.

In testing a hypothesis, we are trying to find out the value of a population parameter based on some preconceived idea as to what this value should be. In both estimation and hypothesis testing, the general theme is trying to determine whether or not the data we have obtained are indicative of the general or the future behavior of the phenomenon being studied.

Hypothesis Testing and a Criminal Trial

To understand the idea of hypothesis testing, it is often useful to look at the context of a criminal trial.

In a criminal trial, a jury must decide between two hypotheses. The hypothesis that the jury initially assumes, called the **null hypothesis** or H_0 , is that “the defendant is innocent.” When proven otherwise, this leads to accepting an **alternative hypothesis** or H_a , which is “the defendant is guilty.”



The jury does not know which hypothesis is true. They must make a decision based on the evidence presented in court. If enough evidence is presented showing the guilt of the defendant, the jury decides to reject the null hypothesis and say that the defendant is guilty (beyond reasonable doubt). Otherwise, they acquit the defendant. Note that acquitting the defendant does not necessarily mean that he or she is innocent, but only that the evidence presented is insufficient for the jury to reject the hypothesis that he or she is innocent.

For a more statistical example, consider an automobile company looking for additives to increase gas mileage. As an initial study, they send 30 identical cars on a road trip from Manila to Dau. Without the additive, it is known that these cars average 10 km per liter (km/L), with a standard deviation of 1.4 km/L.



Suppose that it turns out that the 30 cars averaged 11 km/L with the additive. Can we say that the additive was effective, or the improvement just happened by chance?

Questions such as these are common in statistical hypothesis testing. We wish to determine based on sample data whether a *hypothesis* about a population is true or false. A definition which is sufficient for our purposes in this book is given as follows:

Definition 1

A **statistical hypothesis** is either a statement about the value of a population parameter or a statement about the probability distribution that a certain random variable follows.

For example, one may be interested in knowing whether or not the distribution of heights of grade 11 students follows a normal distribution. In this case, we may test a hypothesis about the shape of the population of heights of grade 11 students.

On the other hand, a researcher might be interested in studying the average height of grade 11 students in the Philippines. In this case, the relevant statistical hypothesis will be a statement on the *population mean*.

Testing a Statistical Hypothesis

To test a statistical hypothesis, one can perform the following procedure:

Six-step Procedure for Testing a Statistical Hypothesis

1	State the null and alternative hypotheses.
2	Select a level of significance.
3	Select the test statistic.
4	Formulate the decision rule.
5	Compute the value of the test statistic.
6	Make a decision.

Let us look at the steps in closer detail.

Step 1: State the null and alternative hypotheses.

Definition 2

A **null hypothesis** is a statement about the value of a population parameter formulated with the hope of it being rejected. It is usually denoted by H_0 . If H_0 is rejected, we will be led to accept an **alternative hypothesis**, usually denoted by H_a .

We can think of the null hypothesis as the current value of the population parameter, which we hope to disprove in favor of our alternative hypothesis. When doing a hypothesis test, our objective is to determine whether there is enough evidence for the new value of the population parameter as stated by the alternative hypothesis.

A null hypothesis always involves an equality symbol. (*Note:* Some books allow the null hypothesis to contain the inequality symbols “ \geq ” or “ \leq .” In this book, we shall only use “ $=$ ” in the null hypothesis.) In contrast, the alternative hypothesis contains the inequality symbol “ $>$,” “ $<$,” or “ \neq .”

Example 1

A common problem in many public schools is the large class size. Suppose that the average number of students in a class in a certain city was 65. We wish to know whether the classrooms built over the years have succeeded in reducing this class size.

Since we wish to test whether the *mean* class size has been reduced to less than 65, the alternative hypothesis must contain the symbol $<$. That is:

$$H_0: \mu = 65$$

$$H_a: \mu < 65$$

Example 2

A Pulse Asia survey conducted from 2–8 July 2016 reported that 44% of Filipinos believe that the Constitution should not be amended at the moment. An anti-charter change group may claim that this percentage is too low, and verify this claim by conducting a survey of their own.

In this case, the group wishes to test the hypothesis that the *proportion* of Filipinos opposed to charter change is higher than 0.44. The parameter we are testing is the population proportion p , and the null and alternative hypotheses are the following:

$$H_0: p = 0.44$$

$$H_a: p > 0.44$$

Example 3

Suppose that the turnover rate of a 200-tablet bottle of multivitamins follows the normal distribution with a mean of 6.0 and a standard deviation of 0.50. We would like to know if the mean turnover has changed and is no longer 6.0.

Here, we wish to test a statement on the population mean μ . The following are the null and alternative hypotheses:

$$H_0: \mu = 6.0$$

$$H_a: \mu \neq 6.0$$

Note that the null and alternative hypotheses are statements on a population or a population parameter. They must never be on a sample statistic. For example, the hypotheses $H_0: \bar{X} = 15$ and $H_a: \bar{X} > 15$ are incorrect as they both say something about a *sample* mean.

In general, the alternative hypothesis is more important, as it represents what we are investigating.

Step 2: Select a level of significance.

To test a hypothesis, we take a sample from a population and use the information obtained in the sample to decide whether the hypothesis is likely to be true or false. If the evidence from the sample is inconsistent with the hypothesis, we reject it. Otherwise we accept it. As in a criminal trial, accepting the hypothesis does not mean that it is true; it only means that there is insufficient evidence to reject it.

As we are taking only a sample, we cannot say with 100% certainty whether our null hypothesis is true or false. However, if the sample data have results which are quite far from our null hypothesis, then we can be quite confident that we are not making a mistake by rejecting H_0 . This level of confidence is captured by the *level of significance* of a hypothesis test.

Definition 3

The **level of significance** (α) of a test is the probability of rejecting the null hypothesis when it is true.

The smaller the value of α , the surer we are that we are not making an error if we end up rejecting H_0 . Thus, a smaller α leads to a higher amount of “evidence” needed before we reject H_0 in favor of H_a .

There is no fixed value of α to use in any hypothesis test. While $\alpha = 0.05$ tends to be chosen as a default value, the choice may differ depending on the application. Usually, α is selected 0.05 for consumer research projects, 0.01 for quality assurance, and 0.10 for political polling.

Step 3: Select the test statistic.

Definition 4

Any function of the observed data whose numerical value dictates whether the null hypothesis is accepted or rejected is called a **test statistic**.

As we will see in lessons 2 and 3, the test statistic (and the distribution of the test statistic) depends on the parameter being tested. Intuitively, the value of the test statistic represents the amount of evidence the sample data contains against the null hypothesis.

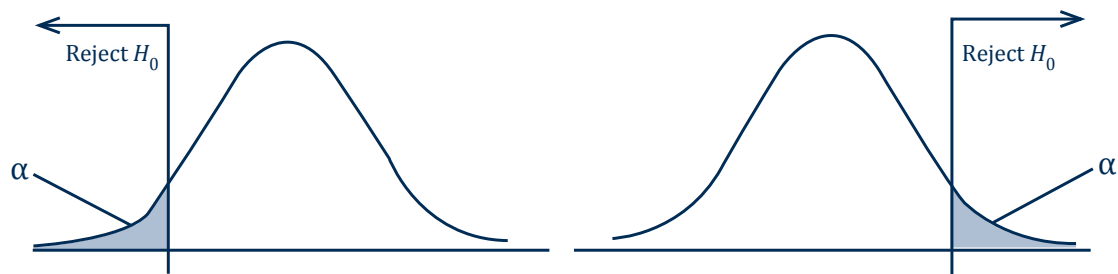
Step 4: Formulate the decision rule.

The *decision rule* indicates the condition(s) where the null hypothesis is rejected. It depends on whether the alternative hypothesis is one-sided (with inequality symbol $>$ or $<$) or two-sided (with inequality symbol \neq).

Definition 5

A test of hypothesis where the alternative hypothesis is one-sided is called a **one-tailed test**. If the alternative hypothesis is two-sided, it is called a **two-tailed test**.

In a one-tailed test, the set of values which lead to the rejection of the null hypothesis, also called the *rejection region*, is in either the upper or the lower tail of the sampling distribution of the statistic.

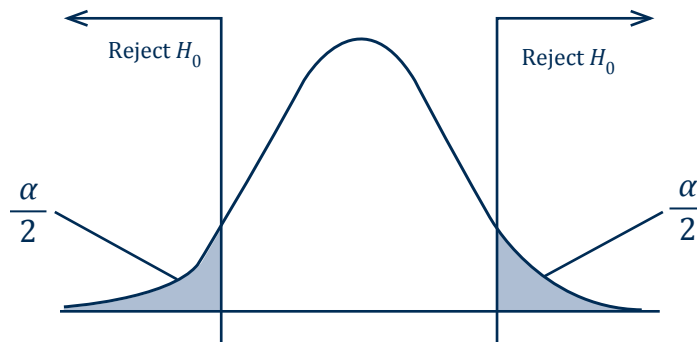


One tailed tests. The first figure shows a left-tailed test, where the entire rejection region is in the lower tail of the sampling distribution. The second figure shows a right-tailed test, where the rejection region lies in the upper tail of the sampling distribution.

The left-tailed case occurs when the alternative hypothesis contains the symbol $<$. In this case, we reject the null hypothesis in favor of the alternative hypothesis if the value of the test statistic based on the sample is *too small* as compared to the value given in the null hypothesis. These consist of the values which are on the left end of the sampling distribution. The problem in example 1 involves a left-tailed test.

On the other hand, the right-tailed case occurs when the alternative hypothesis contains the symbol $>$. Here, we reject the null hypothesis if the value of the test statistic using the sample information is *too large* as compared to the value specified in the null hypothesis. These values are on the right end of the sampling distribution. The problem in example 2 involves a right-tailed test.

In a two-tailed test, the rejection region is divided into two parts and is found in both the upper and lower tails.



A two-tailed test. The area α is divided into two and is distributed among the two ends or “tails” of the distribution.

The two-tailed case occurs when the alternative hypothesis contains the inequality symbol \neq . In this case, we reject the null hypothesis if the values are either too big or too small as compared to the hypothesized value. The scenario in example 3 is an example of a two-tailed test.

The rejection region for a hypothesis test is also called the *critical region*. To specify the critical region for a hypothesis test, it is necessary to indicate at which value of the test statistic the critical region begins.

Definition 6

The set of values of the test statistic that results in the rejection of the null hypothesis is called the **critical region** or the **region of rejection**. The particular point in the critical region that separates the rejection region from the acceptance region is called the **critical value**.

The critical value depends on the level of significance of the test, the alternative hypothesis, and the distribution of the test statistic.

Step 5: Compute the value of the test statistic.

Using the given information in the sample, one must then substitute these numbers to the test statistic and compute its value.

Step 6: Make a decision.

If the value of the test statistic falls within the critical region, we reject H_0 in favor of the alternative hypothesis. Otherwise, we do not reject H_0 . In the latter case, we sometimes say that there is insufficient evidence to reject H_0 .

In the second case, note that we do not say that we are “accepting” H_0 . This is because by not rejecting H_0 , we are not saying that H_0 is true; it only means that there is insufficient evidence to say that it is false. In contrast, when we reject H_0 , we are saying that, based on the sample data, there is only a small probability (specifically, α) of mistakenly rejecting it.

We shall see how these steps are applied in actual statistical tests in the coming lessons.

Let's Practice

I. Fill in the blanks with the correct answer.

1. The _____ is a statement about a population parameter formulated with the hope of it being rejected. If it is rejected, then we are led to accept the _____.
2. The _____ of a test is the probability of rejecting the null hypothesis when it is true.
3. The _____ indicates the condition(s) that will lead to the rejection of the null hypothesis.
4. The set of values of the test statistic for which H_0 will be rejected is called the _____. The particular point that separates this region from the values which do not lead to rejection of H_0 is known as the _____.
5. If the alternative hypothesis for testing a population parameter contains the symbol $>$ or $<$, then the test is said to be _____. Otherwise, the alternative hypothesis contains the symbol \neq , and the test is said to be _____.

II. Specify the null and alternative hypotheses in the following scenarios. In addition, for items 1 to 4, identify the parameter being tested.

1. A store manager of air conditioners tells the higher management that around 90% of its customers are “fully satisfied” with their overall purchase performance. The CEO would like to verify whether this claim is true.
2. A mobile phone manufacturer recently announced their latest smartphone model and claimed that under regular use and a full battery, it can last 2 full days without recharging. Being someone who values long battery life, would you buy this smartphone?

3. Kayla is the founder and head instructor of a review center who claims that their center's innovative teaching methods allow their students to have a higher entrance exam passing rate than the 35% achieved by the majority. You wish to know if this is indeed true.
4. A local police office is concerned about the increased incidence of obesity among its policemen and staff. As such, it has launched a "Get Fit, Eat Right" campaign, which is said to be able to reduce the average weight of the men in police offices within the city to less than 75 kg. The campaign has reached the upper echelon who wish to adapt this campaign in other cities if the claims are true.
5. Julius a junior supervisor, is considering taking up the degree of Master in Business Administration (MBA). He was told that MBA graduates have higher salaries than those with only a bachelor's degree. Help him in making a decision.
6. Camille, a mother of several young children, wishes to decrease the possibility that her children will get sick. She plans to stock up on green barley supplements as she has read that these can help build up resistance against illnesses.

Lesson 2

Tests Involving the Population Mean

Learning Outcomes

At the end of this lesson, you should be able to

- formulate the appropriate null and alternative hypotheses on a population mean;
- identify the appropriate form of the test statistic when the population variance is either assumed to be known or the central limit theorem is to be used;
- identify the appropriate rejection region for a given level of significance when the population variance is either assumed to be known or the central limit theorem is to be used;
- compute the test statistic value and draw the corresponding conclusion for tests involving the population mean based on this value and the rejection region; and
- solve problems involving a test of hypothesis on the population mean.

Introduction

In the previous lesson, you have learned how to construct hypothesis statements based on assumptions or claims made by people concerned. In testing the validity of a hypothesis, it is usually impractical to examine the entire population. Thus, researchers typically examine only a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected.

Large Sample Tests for the Mean

Consider a random sample of n measurements drawn from a population with mean μ and *known* standard deviation σ . To test a hypothesis of the form $H_0: \mu = \mu_0$ against either a one-tailed or two-tailed alternative, we have the following test statistic:

Large Sample Test Statistic for Testing a Population Mean (when σ is known)

When testing the null hypothesis $H_0: \mu = \mu_0$ where the sample size n is large and σ is known, the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

where \bar{x} = sample mean;

n = sample size;

μ_0 = hypothesized value of the population mean; and

σ = population standard deviation.

Here, the test statistic Z has approximately a standard normal distribution when H_0 is true.

Example 1

Consider the following hypotheses:

- $H_0: \mu = 250$
- $H_a: \mu \neq 250$

The sample mean is 253 and the sample size is 50. The population is normally distributed with a standard deviation of 16. Test the hypotheses at 0.05 level of significance.

Solution:

We follow the six-step procedure from the previous lesson.

Step 1: State the null and alternative hypotheses.

In this case, these are already given, and we are testing $H_0: \mu = 250$ against the alternative hypothesis $H_a: \mu \neq 250$.

Step 2: Select a level of significance.

We are asked to perform the hypothesis test at $\alpha = 0.05$. Recall that this represents the probability of rejecting a true null hypothesis.

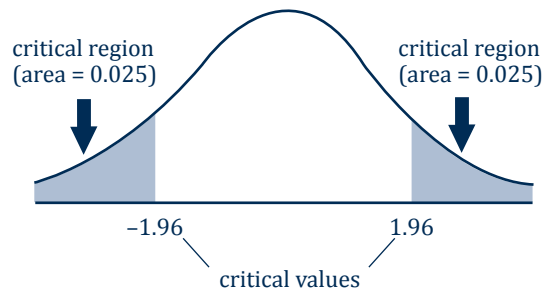
Step 3: Select the test statistic.

Since the population is normally distributed and σ is known, we use

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \text{ as our test statistic.}$$

Step 4: Formulate the decision rule.

We need to determine the critical values of Z for the specified level of significance. Since the alternative hypothesis contains the symbol \neq , we have a two-tailed test, and half of 0.05, or 0.025, is placed in each tail.



Critical region, z , two-tailed test, $\alpha = 0.05$

We therefore need the value of the standard normal distribution which leaves an area of 0.025 to the right. Looking at the z -table in *Appendix B*, we see that this is $z_{0.025} = 1.96$. This means that our decision rule is to reject H_0 if the value of the test statistic $Z > 1.96$ or $Z < -1.96$. Notice that the critical region is divided into two parts because the test is two-tailed.

Step 5: Compute the value of the test statistic.

From the given information, the sample mean is $\bar{x} = 253$, the hypothesized mean is $\mu_0 = 250$, the population standard deviation is $\sigma = 16$, and the sample size is $n = 50$. Substituting these values to the test statistic given in step 3, we have

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{253 - 250}{\frac{16}{\sqrt{50}}} \approx 1.33.$$

Step 6: Make a decision.

Since the computed value of Z based from the sample is neither bigger than 1.96 nor smaller than -1.96 , we do not reject H_0 . Therefore, there is insufficient evidence to say that the mean is different from 250.

Example 2

An oil company claims that their new gasoline formula contains an additive that results in increased fuel efficiency. To test this claim, they collaborate with an automobile company to send 30 identical cars on a road trip from Manila to Dau. The average mileage of these cars turns out to be 10.8 km/L. Without the additive, it is known that these cars' average mileage is 10 km/L, with a standard deviation of 1.4 km/L. At 0.01 level of significance, should we agree with the company's claim?

Solution:

We follow the steps outlined in the previous lesson.

Step 1:

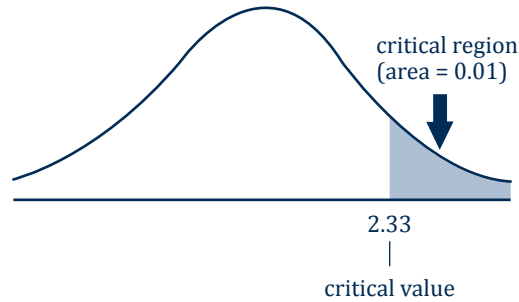
- $H_0: \mu = 10$
- $H_a: \mu > 10$

Step 2: The level of significance is $\alpha = 0.01$.

Step 3: Since the test involves the population mean, the sample size is large, and $\sigma = 1.4$ is known, we use the formula

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

Step 4: The alternative hypothesis is that $\mu > 10$. Therefore, we have a right-tailed test. To determine the critical value, we need to find the value of $z_{0.01}$, which is the z -value that leaves an area of 0.01 to its right. Using the z -table in *Appendix B*, we have $z_{\alpha} = 2.33$. Hence, we shall reject H_0 if the value of the test statistic $Z > z_{0.01} \approx 2.33$.



Critical region, z , right-tailed test, $\alpha = 0.01$

Step 5: Based on the given, the sample mean is $\bar{x} = 10.8$ and the sample size is $n = 30$. Also, the hypothesized value of the mean is $\mu_0 = 10$. Substituting these values to the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{10.8 - 10}{\frac{1.4}{\sqrt{30}}} \approx 3.13.$$

Step 6: Since the computed value of the test statistic is larger than 2.33, we reject H_0 . The average mileage is therefore greater than 10 km/L.

In an actual hypothesis testing scenario, it is not common to know the value of the population standard deviation σ . Fortunately, if the sample size n is large, recall that we can replace σ with the sample standard deviation s without substantially changing the value of the test statistic. In this case, the test statistic becomes as follows, as a result of the CLT:

Large Sample Test Statistic for Testing a Population Mean (when σ is unknown)

When testing the null hypothesis $H_0: \mu = \mu_0$ where the sample size n is large and σ is unknown, the test statistic is

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where \bar{x} = sample mean;

n = sample size;

μ_0 = hypothesized value of the population mean; and

s = sample standard deviation.

Here, the test statistic Z has approximately a standard normal distribution when H_0 is true.

Example 3

A soft drink dispenser is designed to dispense 330 mL of drink per cup. However, there have been recent reports to the management of a fastfood restaurant of both underfilled and overfilled cups. Seeking to investigate this matter, they fill up 50 cups using this machine. If the mean amount of drink in the 50 cups is 325 mL with standard deviation of 20 mL, is there cause for concern for the management? Use a level of significance of 0.01.

Solution:

Step 1:

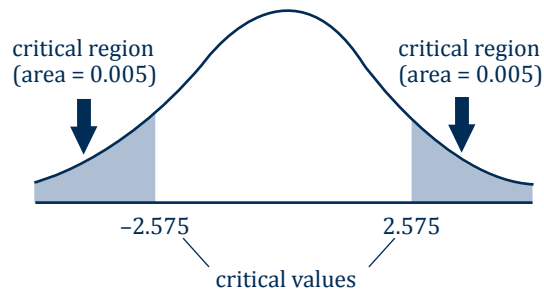
- $H_0: \mu = 330$
- $H_a: \mu \neq 330$

Step 2: The level of significance is $\alpha = 0.01$.

Step 3: Since the test involves the population mean, and the sample size $n = 30$ is large, we use the test statistic

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Step 4: The alternative hypothesis is $\mu \neq 330$. Therefore, we have a two-tailed test, and we divide α equally into two. That is, the critical values leave an area of 0.005 to the right and to the left, and we need $\pm z_{0.005}$.



Critical region, z , two-tailed test, $\alpha = 0.01$

Using the z -table in *Appendix B*, we have $z_{0.005} = 2.575$. Hence, we shall reject H_0 if $Z > 2.575$ or $Z < -2.575$.

Step 5: Based on the given, the sample mean is $\bar{x} = 325$ and the sample size is $n = 50$. The sample standard deviation is $s = 20$. Also, the hypothesized value of the mean is $\mu_0 = 330$. Substituting these values to the test statistic gives

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{325 - 330}{\frac{20}{\sqrt{50}}} \approx -1.77.$$

Step 6: Since the computed value of the test statistic is neither less than -2.575 nor greater than 2.575 , we do not reject H_0 . Therefore, there is insufficient evidence to say that the mean volume of drink dispensed is significantly different from 330 mL.

Small Sample Tests for the Mean

Sometimes, the size of our sample may not be large enough so that the CLT is applicable to the problem. If the standard deviation of the population is also unknown, but the population is approximately normal, then the following test statistic, which has a *Student's t-distribution* instead applies:

Small Sample Test Statistic for Testing a Population Mean (when σ is unknown)

When testing the null hypothesis $H_0: \mu = \mu_0$ where the sample size n is small and σ is unknown, the test statistic is

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

where \bar{x} = sample mean;

μ_0 = hypothesized value of the population mean;

s = sample standard deviation; and

n = sample size.

Here, the test statistic T has a student's t -distribution with $n - 1$ degrees of freedom when H_0 is true.

Example 4

Test at $\alpha = 0.01$ the null hypothesis $H_0: \mu = 1.75$ against the alternative hypothesis $H_a: \mu < 1.75$ if a sample of size $n = 10$ has a mean of 1.722 and a standard deviation of 0.0339. Assume that the population is approximately normal.

Solution:

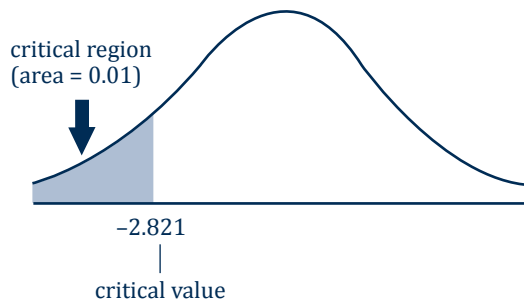
Step 1: The null hypothesis is given as $H_0: \mu = 1.75$ while the alternative hypothesis is $H_a: \mu < 1.75$.

Step 2: Clearly, $\alpha = 0.01$ from the given.

Step 3: Since we are testing a population mean, the sample size $n = 10$ is small, and the population standard deviation is not given (note that only the *sample* standard deviation is given), a student's t -test statistic is appropriate for this problem:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Step 4: The alternative hypothesis contains the symbol $<$. Thus, the hypothesis test is left-tailed. Since the sample size $n = 10$, the t -statistic has $n - 1 = 10 - 1 = 9$ degrees of freedom. Using the t -table in *Appendix C*, we have the critical value $-t_{0.01,9} = -2.821$.



Critical region, t , left-tailed test, $df = 9$, $\alpha = 0.01$

Therefore, our decision rule involves rejecting H_0 if $T < -2.821$.

Step 5: All the relevant quantities in the test statistic are already provided:

$\bar{x} = 1.722$, $\mu_0 = 1.75$, $s = 0.0339$, and $n = 10$. Substituting these values to the formula for the test statistic yields

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{1.722 - 1.75}{\frac{0.0339}{\sqrt{10}}} \approx -2.612$$

which is *not* less than -2.821 .

Step 6: Since the computed value of the test statistic from the sample is outside the critical region $T < -2.821$, we do not reject H_0 . Therefore, there is not enough evidence to say that the population mean is less than 1.75 at $\alpha = 0.01$.

Example 5

According to a report by Quartz in 2013, the Philippines is the number one coffee consumer in Asia. The typical Filipino drinks an average of 0.608 cup of coffee per day, which is equivalent to an average of 4.256 cups per week (using the conversion scale 1 week = 7 days). Suppose that a sample of 12 senior citizens revealed they consumed the following number of cups of coffee during the last week.

3, 3, 4, 4, 5, 7, 7, 5, 5, 3, 4, 4

At the 0.05 significance level, do the sample data suggest that there is a difference between the national average and the sample mean from the senior citizens? Assume that the population is approximately normal.

Solution:

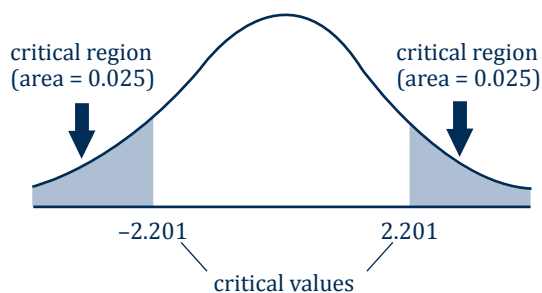
Step 1: We are testing the null hypothesis $H_0: \mu = 4.256$ against the alternative hypothesis $H_a: \mu \neq 4.256$.

Step 2: From the given, the level of significance is $\alpha = 0.05$.

Step 3: Since the sample size $n = 12$ is small, and the population standard deviation is unknown, we use the t -distribution. That is, our test statistic is

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

Step 4: The alternative hypothesis contains the symbol \neq ; thus, we have a two-tailed test. This means that the area of 0.05 will be divided into the two tails:



Critical region, t , two-tailed test, $df = 11$, $\alpha = 0.05$

Since the sample size is $n = 12$, the t -distribution has $n - 1 = 11$ degrees of freedom. Therefore, from the t -table in *Appendix C*, the critical values are $t_{0.025,11} = 2.201$ and $-t_{0.025,11} = -2.201$. We reject H_0 if $T > 2.201$ or $T < -2.201$.

Step 5. We first need to compute the mean \bar{x} and standard deviation s from the given sample. We have

$$\bar{x} = \frac{3+3+4+4+5+7+7+5+5+3+4+4}{12} \approx 4.5; \text{ and}$$

$$s = \sqrt{\frac{(3-4.5)^2 + (3-4.5)^2 + (4-4.5)^2 + \dots + (4-4.5)^2}{12-1}} = 1.3817.$$

Also, $\mu_0 = 4.256$ and $n = 12$ from the given. Plugging these values into our test statistic, we get

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \frac{4.5 - 4.256}{\frac{1.3817}{\sqrt{12}}} \approx 0.61.$$

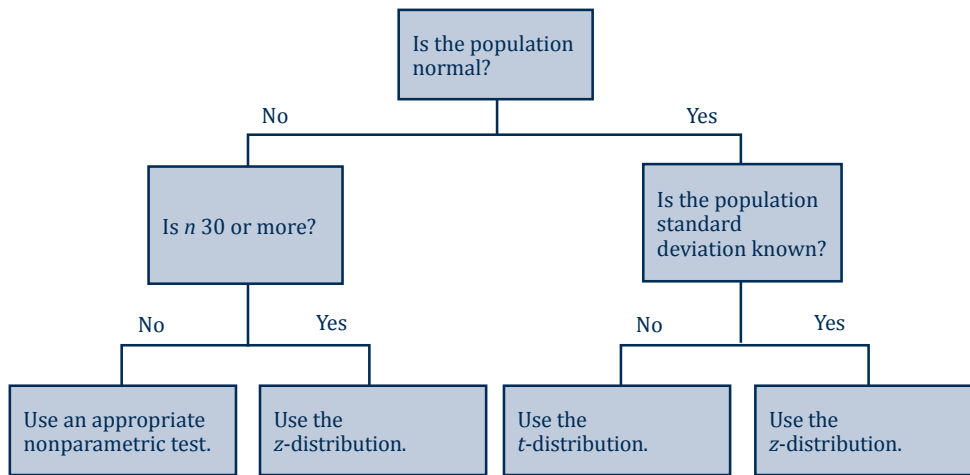
Step 6: Since the value of the test statistic is neither bigger than 2.201 nor smaller than -2.201, we do *not* reject H_0 . Therefore, the average weekly coffee consumption of the senior citizens is not significantly different from the population average of 4.256.

Points to Remember

When performing a test of hypothesis for a single mean, the test statistic depends on three things:

1. whether the population is normally distributed
2. whether the population standard deviation is known
3. whether the sample size is large ($n \geq 30$)

The three cases covered in this book are summarized in the following flowchart:



Flow chart for testing a population mean

Let's Practice

I. Answer the following questions based on the given information:

- a. Is it a one-tailed or two-tailed test?
- b. What is the decision rule?
- c. What is the value of the test statistic?
- d. What is the decision regarding H_0 ?

1. $H_0: \mu = 60$

$H_a: \mu \neq 60$

The sample mean is 59 and the sample size is 36. The population follows a normal distribution with standard deviation 5. Use $\alpha = 0.05$.

2. $H_0: \mu = 8.4$

$H_a: \mu > 8.4$

The sample mean is 9.2, the sample variance is 6.76, and the sample size is 42. Use $\alpha = 0.01$.

3. $H_0: \mu = 10$

$H_a: \mu < 10$

The sample mean is 8, the sample standard deviation is 3, and the sample size is 64. Use $\alpha = 0.025$.

4. $H_0: \mu = 1,375$

$H_a: \mu > 1,375$

The population follows a normal distribution with standard deviation 225, the sample mean is 1,575, and the sample size is 10. Use $\alpha = 0.10$.

5. $H_0: \mu = 0.25$

$H_a: \mu \neq 0.25$

The population is approximately normal, the sample mean is 0.3, the sample standard deviation is 0.1, and the sample size is 25. Use $\alpha = 0.10$.

II. Analyze and solve each problem.

1. a. If the null hypothesis $\mu = 10$ is rejected in favor of the alternative hypothesis $\mu \neq 10$, will it necessarily also be rejected in favor of $\mu > 10$? Why? (Assume that the value of α remains the same.)
 b. If the null hypothesis $\mu = 10$ is rejected in favor of the alternative hypothesis $\mu \neq 10$ at $\alpha = 0.01$, will it necessarily also be rejected when $\alpha = 0.05$? Why?
 c. If the null hypothesis $\mu = 10$ is rejected in favor of the alternative hypothesis $\mu \neq 10$ at $\alpha = 0.05$, will it necessarily also be rejected when $\alpha = 0.01$? Why?
2. As input for a new inflation model, the average cost of a hypothetical basket of basic commodities in the Central Luzon region was predicted by economists to be ₱1,468. The standard deviation of basket prices was assumed to be ₱95, a figure that has held fairly constant over the years. To check this prediction, a sample of 36 baskets representing different parts of the region was checked in late July, and the average cost was ₱1,498. Let $\alpha = 0.05$. Is the difference between the economists' prediction and the sample mean statistically significant?
3. The shipping department manager of an e-commerce company claims that the average order shipped by the firm weighs 10 kg. The general manager thinks this too large and decides to verify this claim by selecting a random sample of 100 orders. What can the manager conclude at 0.10 level of significance if the sample has a mean weight of 9.1 kg with a standard deviation of 4 kg?

4. Bagong Liwanag High School was chosen to participate in the evaluation of the new statistics curriculum. In the recent past, their students were considered “typical,” having earned scores on standardized exams that were consistent with national averages. Two years ago, a cohort of 85 senior high school students of the same school was assigned a special set of statistics classes. According to test results that have just been released, these students got an average rating of 82% on a national standardized statistics exam. Nationwide, senior high school students have an average rating of 79% with a standard deviation of 2.1%. Can it be claimed that the new curriculum had a significant effect? Test at $\alpha = 0.05$ level of significance.
5. Prices of basic commodities are often greatly affected by typhoons. During the aftermath of a recent typhoon, the prices of *bangus* were reported to have been sold at ₱20–₱25 more per kilogram at local wet markets. However, the price was back to its normal average price at ₱100 per kilogram one week after. If a random sample of 10 *bangus* sales transactions from local wet markets have prices (in pesos per kilogram) of ₱107, ₱115, ₱130, ₱95, ₱100, ₱105, ₱98, ₱116, ₱104, and ₱100, assuming that the *bangus* prices are normally distributed, is there sufficient evidence to say that the average price of *bangus* in these markets is greater than ₱100? Use a 0.05 level of significance.
6. It is recognized that cigarette smoking has an adverse effect on lung function. In a study of the effect of cigarette smoking on the carbon monoxide diffusing capacity (DL) of the lung, researchers found that current smokers had DL readings significantly lower than those of either ex-smokers or nonsmokers. The carbon monoxide diffusing capacities for a random sample of 20 current smokers are listed below.

104	89	73	123	91
92	62	91	84	76
101	88	71	82	89
103	198	73	197	90

Do these data indicate that the mean DL reading for current smokers is significantly lower than 100 DL, which is the average reading for nonsmokers? Use $\alpha = 0.01$.

7. A production process is designed to add 50 gallons of water to each of numerous batches of a mixture. The mean number of gallons of water occasionally varies, but the standard deviation is considered to be stable and well established at 2.8 gallons. To verify that the mixtures indeed receive 50 gallons of water, a production engineer observes a random sample of 75 batches of mixture and determines the mean number of gallons of water (\bar{x}) added. For what values of \bar{x} should the production engineer reject the null hypothesis $\mu = 50$ if he uses a two-sided alternative and a level of significance at 0.05?

Lesson 3

p -values in Hypothesis Testing

Learning Outcomes

- At the end of this lesson, you should be able to
 - compute and interpret the p -value obtained when testing a hypothesis on the population mean; and
 - solve problems involving a test of hypothesis on the population mean using a p -value approach.

Introduction

You have learned in the previous two lessons a general procedure in performing a test of hypothesis on a population parameter. This procedure is often called the *critical region* (or *rejection region*) *approach*, as one would need to determine beforehand the critical region, and then, which values of the test statistic would result in the *rejection* of the null hypothesis.

There is, however, another approach in hypothesis testing, which is the main output when one uses statistical software. It involves the *p -value* of the test statistic.

Definition 1

The **p -value** associated with an observed test statistic is the probability of getting a value for that test statistic as extreme or more extreme than what was actually observed (relative to H_0), given that H_0 is true.

Determining the p -value not only allows us to make a decision on H_0 but also gives us an insight into the strength of the relationship. A very small p -value, say 0.0001, means that it is very unlikely that H_0 is true. On the other hand, a large p -value, say 0.3125, means that H_0 is not rejected.

The computation of the p -value depends on the sign in the alternative hypothesis. Suppose that the resulting value from the test statistic T is t . Then,

- if H_a involves the symbol “>,” the p -value is equal to $P(T \geq t)$.
- if H_a involves the symbol “<,” the p -value is equal to $P(T \leq t)$.
- if H_a involves the symbol “ \neq ,” the p -value is equal to $2 \cdot P(T \geq |t|)$.

Note: In the above notation, the variable T is used to represent any test statistic. Depending on what parameter is being tested, it can have any distribution. It does not mean that the test statistic has a Student’s t -distribution.

Example 1

Compute the p -values for the observed values of the test statistics in examples 1 and 2 in the previous lesson.

Solution:

From example 1, we see that the alternative hypothesis is $H_a: \mu \neq 250$, and the value of the test statistic is 1.33. Since the alternative hypothesis is two-tailed and the test statistic is normally distributed,

$$\begin{aligned} p\text{-value} &= 2 \cdot P(Z \geq 1.33) \\ &= 2[1 - \Phi(1.33)] \\ &= 2(1 - 0.9082) \\ &= 0.1836. \end{aligned}$$

On the other hand, the alternative hypothesis in example 2 is $H_a: \mu > 10$ and the value of the standard normal test statistic is 3.13. Since the alternative hypothesis is one-tailed and contains the symbol “>,” there is no need to multiply by 2.

$$\begin{aligned} p\text{-value} &= P(Z \geq 3.13) \\ &= 1 - \Phi(3.13) \\ &= 1 - 0.9991 \\ &= 0.0009 \end{aligned}$$

When using the p -value in hypothesis testing, we simply compare it to our chosen level of significance α . If the p -value of the test is *less than or equal to* α , we reject H_0 . Otherwise, we do not reject H_0 .

In general, we have the following list of steps for testing a hypothesis using the p -value approach.

1	State the null and alternative hypotheses.
2	Select a level of significance α .
3	Select the test statistic.
4	Compute the value of the test statistic.
5	Compute the p -value of the test statistic.
6	If the p -value is smaller than α , reject H_0 . Otherwise, do not reject H_0 .

Here, the determination of the decision rule is replaced by the computation of the p -value.

Example 2

An electricity conservation NGO has published figures on the number of kilowatt-hours (kWh) used annually by various home appliances. It is claimed that a vacuum cleaner uses an average of 45 kWh per year. If a random sample of 39 homes included in a planned study indicates that vacuum cleaners use an average of 41 kWh per year with a standard deviation of 12.9 kWh, does this suggest at the 0.05 level of significance that vacuum cleaners use, on average, less than 45 kWh annually? Use the p -value approach.

Solution:

Step 1: We are testing $H_0: \mu = 45$ against $H_a: \mu < 45$.

Step 2: The level of significance is $\alpha = 0.05$.

Step 3: Although σ is unknown, the sample size $n = 39$ is large enough to apply the CLT. Thus, we use the standard normal test statistic

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Step 4: From the given, $\bar{x} = 41$, $\mu_0 = 45$, $s = 12.9$, and the sample size is $n = 39$. Substituting these values into our test statistic formula yields

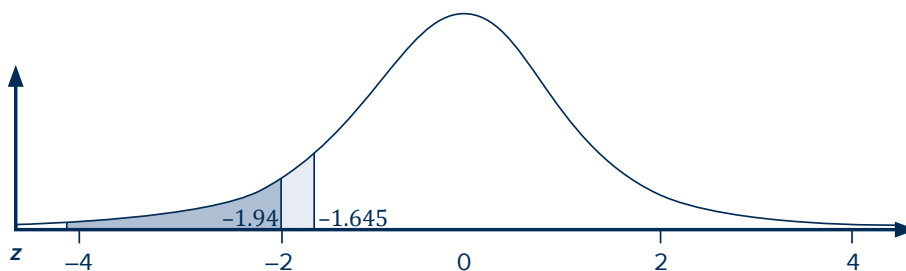
$$Z = \frac{41 - 45}{\frac{12.9}{\sqrt{39}}} \approx -1.94.$$

Step 5: We now compute the p -value associated with this value of the test statistic. Since the alternative hypothesis contains the symbol "<," the computation of the p -value will contain the same inequality symbol. That is,

$$\begin{aligned} p\text{-value} &= P(Z < -1.94) \\ &= \Phi(-1.94) \\ &= 0.0262. \end{aligned}$$

Step 6: Since the p -value 0.0262 is less than 0.05, we reject the null hypothesis $H_0: \mu = 45$ in favor of the alternative hypothesis $H_a: \mu < 45$. Thus, we can say that vacuum cleaners use an average of less than 45 kWh annually.

The figure below shows the area covered by the level of significance $\alpha = 0.05$ and the computed p -value which is 0.0262.



The darker shaded area is the computed p -value, 0.0262.
The sum of the darker and lighter shaded areas is the level of significance $\alpha = 0.05$.

In this example, note that the level of significance of the hypothesis test was at $\alpha = 0.05$. However, we can easily make the relevant decisions for other values of α . For example, it is easy to see that H_0 would still be rejected at $\alpha = 0.10$ but would *not* be rejected at $\alpha = 0.01$ (since $0.0262 > 0.01$). In general, H_0 would be rejected at any level of significance greater than 0.0262. This is the advantage of the p -value approach—there is no need to compute critical values for every value of α ; one simply needs to compare the p -value with the level of significance.

We can also use the p -value approach for tests involving the t -distribution. However, statistical tables involving the t -distribution usually include only the usual values of α , which are 0.01, 0.05, 0.10, and half of these values. As such, one would either resort to approximation (and choosing the nearest value) or use calculators or software such as MS Excel to compute the relevant probability.

Let's Practice

1. Compute the p -values for the hypothesis tests indicated in part I, items 1 to 4 of the exercises in lesson 2 of this chapter. Do they agree with your decision on whether or not to reject H_0 ?
2. Suppose $H_0: \mu = 12$ is to be tested against $H_a: \mu \neq 12$. If $\sigma = 1.06$ and $n = 49$, what p -value would be associated with the sample mean $\bar{x} = 12.16$? Would you reject H_0 ?

3. Recent data on sodium intake suggest that populations around the world are consuming much more sodium than what is physiologically necessary. As such, the World Health Organization (WHO) recommends consuming a maximum of 2 g (or 2,000 mg) of sodium per day. To find out if Filipinos are exceeding this limit, Angelica takes a sample of 100 Filipinos and computes a mean of 2,100 mg with a standard deviation of 1,100 mg.
 - a. Conduct a test of hypothesis at $\alpha = 0.05$ using the p -value approach.
 - b. Without doing any additional computations, would the hypothesis be rejected at 0.10 level of significance? Explain.
4. A local labor union claims that most of their members who are working in a local supermarket chain are being paid less than the minimum daily wage of ₱491. To investigate the matter, the regional wage board took a sample of 81 members of the labor union. Their daily earnings had an average of ₱475 with a standard deviation of ₱90. At $\alpha = 0.05$, is this sufficient evidence to charge the supermarket of violating the minimum wage? Test using both a critical region approach and a p -value approach.
5. Use the t -table in *Appendix C* to approximate the p -value for the t -statistic in the following situations:
 - a. A two-tailed test with $T = 1.92$ and $df = 12$
 - b. A right-tailed test with $T = 3.19$ and $df = 5$
 - c. A left-tailed test with $T = -2.067$ and $df = 27$
6. The daily yield for a local chemical plant has averaged 770 tons for the last several years. The quality control manager would like to know whether this average has changed in recent months. She randomly selects 60 yields from the computer database and computes the average and standard deviation to be 761 tons and 21 tons respectively. Compute the p -value. What is the minimum value of α for which the null hypothesis would be rejected?

Lesson 4

Tests Involving the Population Proportion

Learning Outcomes

- At the end of this lesson, you should be able to
 - formulate the appropriate null and alternative hypotheses on a population proportion;
 - identify the appropriate form of the test statistic when the central limit theorem is to be used;
 - identify the rejection region for a given level of significance when the central limit theorem is to be used;
 - compute the test statistic and draw the corresponding conclusion based from this value and the rejection region; and
 - solve problems involving a test of hypothesis on the population proportion.

Introduction

Recall that a statistical hypothesis is either a statement about the value of a population parameter or a statement about the probability distribution that a certain random variable follows. So far, you have already learned how to test hypotheses concerning the population mean in the previous lessons. In this lesson, you will be dealing with hypothesis tests involving the population proportion.

There are times when we are interested in testing the probability of success p in a binomial experiment. If the sample size is large, the following test statistic can be used to test the hypothesis of the form $p = p_0$ based on the number of successes x in the sample.

Large Sample Test Statistic for a Population Proportion

When testing the null hypothesis $H_0: p = p_0$ where the sample size is large, the test statistic is

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}}$$

where x = the number of successes;

n = the sample size;

p_0 = hypothesized value of the population proportion; and

$q_0 = 1 - p_0$.

Here, the test statistic Z has approximately a standard normal distribution when H_0 is true.

As a rule of thumb, we can consider the sample size to be large enough if both np_0 and $nq_0 = n(1 - p_0)$ are at least 5. In such a case, we can use the CLT to obtain the test statistic given above. Otherwise, we will need to compute the probabilities directly using a binomial distribution. In this book, all tests of proportion will involve sample sizes which are large enough so that the CLT will be applicable.

Example 1

An insurance industry report indicated that 30% of those persons involved in minor traffic accidents this year have been involved in at least one other traffic accident in the last five years. An advisory group decided to investigate this claim, believing it was too large. A sample of 200 traffic accidents this year showed that 56 persons were also involved in another accident within the last five years. Use $\alpha = 0.1$.

Solution:

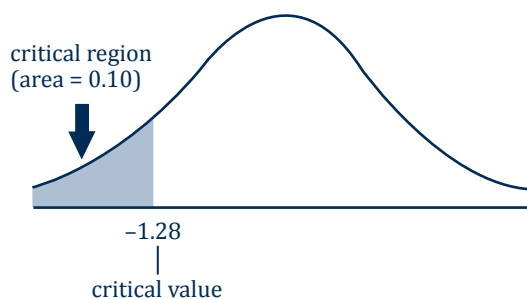
Step 1: Since the advisory group claims that 30% of the persons involved in traffic accidents is too large, the alternative hypothesis must be $p < 0.3$. Thus, the test is left-tailed; that is, we are testing $H_0: p = 0.3$ against $H_a: p < 0.3$.

Step 2: From the given, we have $\alpha = 0.1$.

Step 3: Note that $np_0 = (200)(0.3) = 60 \geq 5$ and $nq_0 = (200)(1 - 0.3) = 140 \geq 5$. Thus, we use the test statistic

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}}.$$

Step 4: Since the test is left-tailed, the critical value is the z-value that gives an area of 0.10 to the left. Therefore, we reject H_0 if $Z < -z_{0.10} \approx -1.28$.



Critical region, z, left-tailed test, $\alpha = 0.10$

Step 5: The number of “successes” x is 56. We also have $p_0 = 0.3$, so $q_0 = 1 - 0.3 = 0.7$. The sample size is $n = 200$. Substituting these values to the test statistic gives

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}} = \frac{56 - 200(0.3)}{\sqrt{200(0.3)(0.7)}} \approx -0.62.$$

Step 6: Since the computed value of the test statistic is -0.62 , which is not less than -1.28 , we fail to reject H_0 . We have insufficient evidence to say that less than 30% of the persons involved in minor traffic accidents this year have been involved in at least one other traffic accident in the last five years.

Example 2

Suppose that in the past, 40% of all adults favored capital punishment. Do we have reason to believe that this proportion has increased, if in a random sample of 150 adults, 80 favored capital punishment? Use a 0.05 level of significance.

Solution:

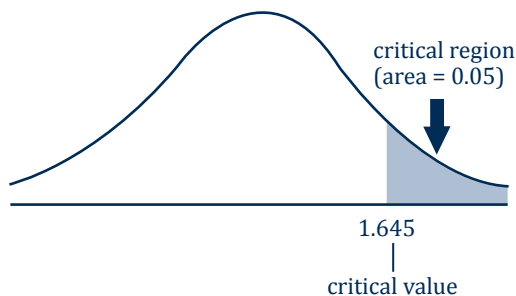
Step 1: We are testing $H_0: p = 0.4$ against $H_a: p > 0.4$.

Step 2: From the given, we have $\alpha = 0.05$.

Step 3: Since the test involves the population proportion, and the sample size is large $np_0 = 150(0.4) = 60 (\geq 5)$ and $nq_0 = 150(0.6) = 90 (\geq 5)$, we use the test statistic

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}}.$$

Step 4: Notice that the alternative hypothesis is that $p > 0.4$. Therefore, we have a one-tailed test, and we reject H_0 if $Z > z_{0.05} \approx 1.645$.



Critical region, z , right-tailed test, $\alpha = 0.05$

Step 5: The number of adults who favor capital punishment based on the sample is $x = 80$. Since the value of the proportion based on the null hypothesis is 0.4, we have $p_0 = 0.4$ and $q_0 = 1 - 0.4 = 0.6$. The sample size is $n = 150$. Substituting these values into the test statistic gives

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}} = \frac{80 - 150(0.4)}{\sqrt{150(0.4)(0.6)}} \approx 3.33.$$

Step 6: Since the computed value of the test statistic is 3.33, which is larger than 1.645, we reject H_0 . Based on this sample, we can be quite confident that the proportion of adults who support capital punishment is now larger than 40%.

Example 3

Verify that the same conclusions are obtained in the previous two examples by computing the p -values.

Solution:

In example 1, the value of the test statistic is -0.62 , and the alternative hypothesis is left-tailed ($H_a: p < 0.3$). Thus, the corresponding p -value is

$$P(Z < -0.62) = \Phi(-0.62) \approx 0.2676.$$

Since the p -value is greater than the level of significance of 0.1, we fail to reject H_0 , which is the same conclusion that we had before.

On the other hand, in example 2, the value of the test statistic is 3.33, and the alternative hypothesis is right-tailed ($H_a: p > 0.4$). Thus, the p -value is

$$P(Z > 3.33) = 1 - \Phi(3.33) \approx 1 - 0.9996 = 0.0004.$$

Since the p -value is significantly less than 0.05, we have extremely strong evidence that H_0 is not true; and we reject H_0 , as before.

Note that the test statistic is also equivalent to

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

where $\hat{p} = \frac{x}{n}$ is the sample proportion. The test statistic is especially useful when the sample proportion is given rather than the number of “successes.” The next example uses this alternative test statistic to perform a hypothesis test for a proportion.

Example 4

In the website of a certain brand of chocolate candies, it was stated that an ideal bag of chocolates contains 24% blue, 20% orange, 16% green, 14% yellow, 13% red, and 13% brown candies. Suppose that we counted the number of blue chocolate candies in 40 chocolate candies sachet packs, and the mean proportion from the sample is 23.04%. At 0.05 level of significance, can we say that the percentage of blue candies in chocolate candy bags these days is no longer 24%?

Solution:

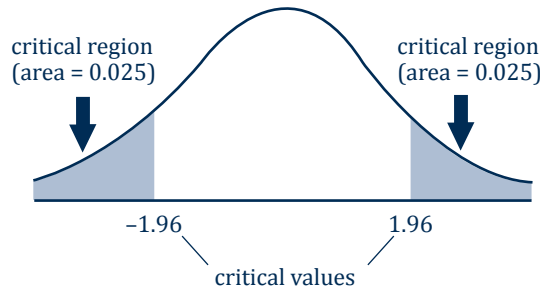
Step 1: We are testing $H_0: p = 0.24$ against $H_a: p \neq 0.24$.

Step 2: From the given, we have $\alpha = 0.05$.

Step 3: Since $np_0 = 40(0.24) = 9.6 \geq 5$ and $nq_0 = 40(1 - 0.24) = 30.4 \geq 5$, we can use the test statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}.$$

Step 4: Since the alternative hypothesis is two-tailed, the 0.05 area is evenly split among the two ends, and we reject H_0 if $Z > z_{0.025} \approx 1.96$ or $Z < -z_{0.025} \approx -1.96$.



Critical region, z, two-tailed test, $\alpha = 0.05$

Step 5: The mean proportion of blue chocolate candies from the sample is given to be $\hat{p} = 23.04\%$ or 0.2304. It was also given that $n = 40$ and $p_0 = 24\%$ or 0.24. Substituting these values into the test statistic gives us

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.2304 - 0.24}{\sqrt{\frac{(0.24)(0.76)}{40}}} \approx -0.14.$$

Step 6: Since the computed value of the test statistic is -0.14 , which is not in the range of $Z > 1.96$ or $Z < -1.96$, we fail to reject H_0 . Based on this sample, we have insufficient evidence to say that the percentage of blue candies in chocolate candy bags is no longer 24%.

Let's Practice

I. Answer the following questions based on the given information:

- Is it a one-tailed or two-tailed test?
- What is the decision rule?
- What is the value of the test statistic?
- What is the decision regarding H_0 ?

1. $H_0: p = 0.35$

$H_a: p > 0.35$

The sample proportion is 0.4 and the sample size is 36. Use $\alpha = 0.1$.

2. $H_0: p = 0.9$

$H_a: p \neq 0.9$

The sample size is 300. Among these, 262 were "successes." Use $\alpha = 0.01$.

3. $H_0: p = 0.5$

$H_a: p < 0.5$

The sample proportion is 0.375 and the sample size is 144. Use $\alpha = 0.05$.

II. Analyze and solve each problem.

- At a certain college, it is estimated that approximately 15% of the students ride bicycles to school. Would you consider this to be a valid estimate if, in a random sample of 90 college students, 19 are found to ride bicycles to class? Use a 0.05 level of significance.
- Boy Diyo was a senatorial candidate who was defeated in the previous election mainly because of poor support from class C, D, and E voters. As such, he has spent the past few years speaking out in favor of issues for the poor and marginalized. A newly released poll for the upcoming election claims to have contacted a random sample of 400 of Boy's current supporters and found that 17.5% were from class C, D, and E voters. In the election where he lost, only 12% of his voters were from Class C, D, and E. Using an $\alpha = 0.05$ level of significance, test the null hypothesis that the proportion of his class C, D, and E supporters has remained the same. Make the alternative hypothesis one-sided.

3. A manufacturer of semiconductor parts produces controllers used in car engine applications. Its customer requires that no more than 5% of these parts be defective at a critical manufacturing step, and that the manufacturer be able to demonstrate this level of quality using $\alpha = 0.05$. The semiconductor manufacturer takes a sample of 240 devices and find that five of them are defective. Would this meet the requirements of the customer?
4. In a survey of randomly selected households, 962 had computers while 288 did not have computers.
 - a. At $\alpha = 0.02$ level of significance, test the claim that computers are not in 80% of households.
 - b. What is the p -value of the test in (a)?
5. A researcher claims that at least 10% of all motorcycle helmets have manufacturing flaws that could potentially cause injury to the wearer. A sample of 200 of these helmets revealed that 25 contained such defects.
 - a. Does this finding support the researcher's claim? Test using the p -value approach and a level of significance of 0.01.
 - b. Explain how the question in (a) could be answered using a confidence interval.
6. Suppose $H_0: p = 0.35$ is to be tested against $H_a: p > 0.35$ at $\alpha = 0.12$ level of significance, where $p = P(\text{success})$. If a sample of size 125 is to be taken, what is the smallest number of successes that would lead to the rejection of the null hypothesis?

Lesson 5

Errors in Hypothesis Testing

Learning Outcomes

- At the end of this lesson, you should be able to
 - illustrate the types of error in hypothesis testing; and
 - calculate the probabilities of committing a type I and type II error.

Introduction

Recall that in a typical hypothesis test, we use the value of the test statistic to determine whether we are to reject the null hypothesis. Since this value is based on a random sample, this may occasionally lead us to make an incorrect conclusion.

There are two typical errors in hypothesis testing, which we shall now formally define.

Definition 1

Rejection of the null hypothesis when it is true is called a **type I error**. The probability of committing a type I error is denoted by α .

Failing to reject the null hypothesis when it is false is called a **type II error**. The probability of committing a type II error is denoted by β .

Example 1

Specify the null and alternative hypotheses in the following scenarios. Then describe a situation where one would commit type I and type II errors in each case.

- Sandra is a young professional who has read about and wishes to try out a weight-loss diet consisting only of organic food. However, she is aware that organic food is more expensive than the usual nonorganic ones.
- A snack food company plans to release a new snack containing peanuts. Initial market studies suggest that this would be a hit among the general public. However, an allergist has raised his concern on this, as he claims that at least 20% of the public is allergic to peanuts.

Solution:

1. The null hypothesis is H_0 : the diet will not result in any change in weight, while the alternative is H_a : the diet will result in weight loss.

Type I error is committed when we reject H_0 when it is true. This means that Sandra would follow an organic diet which in reality, does not result in any weight loss. A possible result of this error is that she lost the opportunity to lose weight with a different weight-loss program.

Type II error is committed when we fail to reject H_0 when it is false. This means that Sandra did not decide to follow the diet even if it actually works. A possible consequence of this would be that she spent to follow another diet which may not actually work.

2. Let p be the true proportion of the general public who are allergic to peanuts. In this case, the null hypothesis is $H_0: p = 0.2$, which we would like to test against the alternative hypothesis $H_a: p > 0.2$.

Type I error is committed when we reject the hypothesis $p = 0.2$ when it is, in fact, true. This leads us to conclude that more than 20% of the population is allergic to peanuts. This may lead the company to defer or even stop the release of the product, and lead to the loss of potential profit for the company.

Type II error is committed if the company believes only a small percentage of the public would be allergic to the snack, even if it is really above 20%. This would probably lead to the commercial release of the snack despite the allergist's concerns. Unless the company specified in the label that the product contains peanuts, this might result in complaints from unaware members of the public of allergic reactions after eating the snack.

Example 2

A manufacturer has developed a new fishing line, which it claims has a mean breaking strength of 15 kg with a standard deviation of 0.5 kg. To test the hypothesis $H_0: \mu = 15$ against $H_a: \mu < 15$, a random sample of 50 fishing lines will be tested. The critical region is defined to be $\bar{X} < 14.9$. Find the probability of committing a type I error when H_0 is true.

Solution:

Note that if H_0 is true, then we have $\mu = 15$. Then by the CLT, \bar{X} is normally distributed with mean 15 and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{50}}$. We then have

$$\begin{aligned} P(\text{Type I error}) &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(\bar{X} < 14.9 | \mu = 15) \\ &= P\left(Z < \frac{14.9 - 15}{\frac{0.5}{\sqrt{50}}}\right) \\ &= \Phi(-1.41) \\ &= 0.0793 \end{aligned}$$

Therefore, there is approximately a 7.93% chance of committing type I error.

The following table summarizes the possible scenarios when testing a statistical hypothesis:

	H_0 is true	H_0 is false
Don't reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Assuming that a given null hypothesis is true, we can compute the value of α once the decision rule is specified. In contrast, to compute β , we will need a specific value of the parameter which is part of the alternative hypothesis.

Example 3

Using the data in the previous example, evaluate β for the alternatives $\mu = 14.7$ and $\mu = 14.8$.

Solution:

From the previous example, recall that the null hypothesis is rejected when $\bar{X} < 14.9$. Since we wish to calculate $\beta = P(\text{Type II error})$, we need the probability that the null hypothesis is **not** rejected, given the specified values of μ . This occurs when $\bar{X} \geq 14.9$.

For $\mu = 14.7$,

$$\begin{aligned}\beta &= P(\text{Fail to reject } H_0 | H_0 \text{ is false}) \\&= P(\bar{X} \geq 14.9 | \mu = 14.7) \\&= P\left(Z \geq \frac{14.9 - 14.7}{0.5/\sqrt{50}}\right) \\&= P(Z \geq 2.83) \\&= 1 - \Phi(2.83) \\&= 1 - 0.9977 \\&= 0.0023.\end{aligned}$$

For $\mu = 14.8$,

$$\begin{aligned}\beta &= P(\text{Fail to reject } H_0 | H_0 \text{ is false}) \\&= P(\bar{X} \geq 14.9 | \mu = 14.8) \\&= P\left(Z \geq \frac{14.9 - 14.8}{0.5/\sqrt{50}}\right) \\&= P(Z \geq 1.41) \\&= 1 - \Phi(1.41) \\&= 1 - 0.9207 \\&= 0.0793.\end{aligned}$$

We can also compute the probabilities of committing a type I or type II error for cases where the CLT does not apply.

Example 4

The proportion of families buying milk from Company A in a certain city is believed to be $p = 0.6$. If a random sample of 12 families shows that at most 3 of them buy milk from Company A, we shall reject the hypothesis that $p = 0.6$, in favor of the alternative $p < 0.6$.

- Find the probability of committing a type I error if the true proportion is $p = 0.6$.
- Find the probability of committing a type II error for the alternative $p = 0.4$.

Solution:

The null hypothesis is $H_0: p = 0.6$, while the alternative hypothesis is $H_a: p < 0.6$.

- a. Let X be the number of families among the 12 who buy milk from Company A. The critical region is defined as $P(X \leq 3)$. Since $np_0 = 12(0.4) = 4.8$, which is less than 5, we cannot apply the CLT to approximate the distribution of \hat{p} . Instead, we use the binomial distribution (see lesson 6 of chapter 2). In this case,

$$\begin{aligned} P(\text{Type I error}) &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(X \leq 3 | p = 0.6) \\ &= \sum_{k=0}^3 \binom{12}{k} (0.6)^k (0.4)^{12-k} \\ &= 0.0153 \end{aligned}$$

- b. We fail to reject H_0 if $X > 3$. Therefore,

$$\begin{aligned} P(\text{Type I error}) &= P(\text{Reject } H_0 | H_0 \text{ is true}) \\ &= P(X > 3 | p = 0.4) \\ &= \sum_{k=4}^{12} \binom{12}{k} (0.4)^k (0.6)^{12-k} \\ &= 1 - \sum_{k=0}^3 \binom{12}{k} (0.4)^k (0.6)^{12-k} \\ &= 1 - 0.2253 \\ &= 0.7747 \end{aligned}$$

Let's Practice

1. Specify the null and alternative hypotheses in the following scenarios. Then describe a situation where one would commit a type I error and a type II error in each case.
 - a. A store manager of air conditioners tells the higher management that around 90% of its customers are “fully satisfied” with their overall purchase performance. The CEO would like to verify whether this claim is true.
 - b. A mobile phone manufacturer recently launched their latest flagship smartphone and claimed that under regular use, the phone can last 2 full days with a fully charged battery. Being someone who values long battery life, would you buy this smartphone?
 - c. Julius a junior supervisor, is considering whether to take up the degree of Master in Business Administration (MBA). He was told that MBA graduates have higher salaries than those with only a bachelor's degree. Help him in making a decision.

- d. Camille, a mother of several young kids, wishes to decrease the possibility that her children will get sick. She plans to stock up on green barley supplements as she has read that these can help build up resistance against illnesses.
- Recall that in a criminal trial, the defendant is assumed innocent unless otherwise proven guilty by the evidence presented during the trial. In this context, which error is more grave, a type I error or a type II error? Explain.
 - Suppose that the hypothesis $H_0: \mu = 45$ is to be tested against $H_a: \mu > 45$ using a sample of size $n = 30$. It was decided that H_0 is to be rejected if the sample mean \bar{X} is greater than or equal to 45.7. If the standard deviation of the sample is 2.4,
 - find α if the null hypothesis is true; and
 - find β for the alternatives $\mu = 45.5$ and $\mu = 45.7$.
 - A researcher is testing the hypothesis $H_0: \mu = 15$ against $H_a: \mu < 15$ with a sample of size 64. She decides that the critical region is to be $\bar{X} \leq 14$. If the sample has a standard deviation 5,
 - find α if the null hypothesis is true; and
 - find β for the alternatives $\mu = 13.5$ and $\mu = 14.5$.
 - A random sample of 400 voters in a certain city is asked if they favor an additional 10% increase in city real-estate taxes to provide badly needed revenues for street repairs. If more than 220 but fewer than 260 favor the tax increase, we shall conclude that 60% of the voters are for it.
 - Find the probability of committing a type I error if 60% of the voters favor the increased tax.
 - What is the probability of committing a type II error using this test procedure if only 50% of the voters are in favor of the additional property tax?
 - In a polygraph test, blood pressure, pulse rate, and other bodily functions are measured. The magnitude of these bodily responses when the subject is asked a relevant question is then used to indicate whether he or she is lying or telling the truth. However, this procedure is not infallible. Suppose seven experienced polygraph examiners were shown the polygraph records of 40 subjects, on the basis of which they had to make a judgment on whether the subject was innocent or guilty. The results are as follows:

Examiner's Decision	Suspect's True Status	
	Innocent	Guilty
"Innocent"	132	13
"Guilty"	8	127

What would be the numerical values of α and β in this context?

7. If $H_0: \mu = 240$ is tested against $H_a: \mu < 240$ at $\alpha = 0.01$ level of significance with a random sample of 25 normally distributed observations, what proportion of the time will the procedure fail to recognize that μ has dropped to 220? Assume that $\sigma = 50$.
8. An urn contains 10 chips. An unknown number of these chips are white while the others are red. We wish to test the following hypotheses
- H_0 : Exactly half of the chips are white.
 - H_a : More than half of the chips are white.

We will draw, without replacement, five chips, and reject H_0 if three or more are white.

- Find α for this decision rule.
- Find β when the urn is (i) 60% white, and (ii) 70% white.

Software Tutorial in MS Excel

Hypothesis tests for single means, where the test statistic is the standard normal distribution, can be done in MS Excel by using the command “=ZTEST(array,x,[sigma]).” Here, “array” refers to the sample data against which the hypothesized mean is to be tested, “x” is the hypothesized mean, and “[sigma]” is an optional argument that represents the population standard deviation. If this optional value is not specified, the function uses the sample standard deviation.

For example, suppose that we wish to test $H_0: \mu = 47.9$ against $H_a: \mu > 47.9$ at the 0.05 level of significance. We have the following sample data, which we assume are drawn from a normal population:

48.2	48.4	47.0	47.3	47.9	48.5	49.0
48.3	48.0	47.9	48.7	48.8	47.4	47.6

Assume that we can use a z-test in this problem. First, copy the sample data into the first column of spreadsheet. Then we type =ZTEST(A1:A14,47.9) in cell B1 to obtain the p -value of 0.141195.

	A	B	C	D	E
1	48.2	0.141195	=ZTEST(A1:A14,47.9)		
2	48.4				
3	47				
4	47.3				
5	47.9				
6	48.5				
7	49				
8	48.3				
9	48				
10	47.9				
11	48.7				
12	48.8				
13	47.4				
14	47.6				

Since the p -value is bigger than 0.05, our level of significance, we do not reject H_0 .

Note: The above command can only be used for a right-tailed z-test. For a left-tailed z-test, we can obtain the corresponding p -value by inputting “=1-ZTEST(array,x,[sigma])” instead. For a two-tailed test, we simply multiply the right-tailed p -value; that is, using the command “=2* ZTEST(array,x,[sigma]).”

Chapter Review

- A **statistical hypothesis** is a statement about a population developed for the purpose of testing.
- A **null hypothesis** is a statement about the value of a population parameter formulated with the hope of it being rejected. It is usually denoted by H_0 .
- If H_0 is rejected, we will be led to accept an **alternative hypothesis**, denoted by H_a .
- The **level of significance** of a test, α , is the probability of rejecting the null hypothesis when it is true.
- Any function of the observed data whose numerical value dictates whether the null hypothesis is accepted or rejected is called a **test statistic**.
- A test of hypothesis where the alternative hypothesis is one-sided is called a **one-tailed test**.
- A test of hypothesis where the alternative hypothesis is two-sided is called a **two-tailed test**.
- The set of values of the test statistic that results in the rejection of the null hypothesis is called the **critical region** or the **region of rejection**. The particular point in the critical region that separates the rejection region with the acceptance region is called the **critical value**.
- **Large Sample Test Statistic for Testing a Population Mean, (σ is known)**

When testing the null hypothesis $H_0: \mu = \mu_0$ where the sample size n is large and σ is known, the test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

where \bar{X} = the sample mean;

n = the sample size;

μ_0 = hypothesized value of the population mean; and

σ = the population standard deviation.

Here, the test statistic z has approximately a standard normal distribution when H_0 is true.

- **Large Sample Test Statistic for Testing a Population Mean, (σ is unknown)**

When testing the null hypothesis H_0 where the sample size n is large and σ is unknown, the test statistic is

$$Z = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

where \bar{X} = the sample mean;
 n = the sample size;
 μ_0 = hypothesized value of the population mean; and
 s = the sample standard deviation.

Here, the test statistic z has approximately a standard normal distribution when H_0 is true.

- **Small Sample Test Statistic for Testing a Population Mean (σ is unknown)**

When testing the null hypothesis H_0 where the sample size n is small and σ is unknown, the test statistic is

$$T = \frac{\bar{X} - \mu_0}{s / \sqrt{n}}$$

where \bar{X} = the sample mean;
 μ_0 = hypothesized value of the population mean;
 s = the sample standard deviation; and
 n = the sample size.

Here, the test statistic T has a Student's t -distribution with $n - 1$ degrees of freedom when H_0 is true.

- The **p -value** associated with an observed test statistic is the probability of getting a value for that test statistic as extreme, or more extreme than that was actually observed (relative to H_0), given that H_0 is true.
- If H_a involves a ">", the p -value is equal to $P(T \geq t)$.
 If H_a involves a "<", the p -value is equal to $P(T \leq t)$.
 If H_a involves a " \neq ", the p -value is equal to $2 \cdot P(T \geq |t|)$.

- If the p -value of a test is less than or equal to the level of significance α of a test, we reject H_0 . Otherwise, we do not reject H_0 .
- **Large Sample Test Statistic for a Population Proportion**

When testing the null hypothesis H_0 where the sample size n is large, the test statistic is

$$Z = \frac{x - np_0}{\sqrt{np_0q_0}}$$

where x = the number of successes;

n = the sample size;

p_0 = hypothesized value of the population proportion; and

$q_0 = 1 - p_0$.

Here, the test statistic Z has approximately a standard normal distribution when H_0 is true.

- Rejection of the null hypothesis when it is true is called a **type I error**. The probability of committing a type I error is denoted by α .
- Failing to reject the null hypothesis when it is false is called a **type II error**. The probability of committing a type II error is denoted by β .

	H_0 is true	H_0 is false
Don't reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Chapter Performance Tasks

1. Backpack Weight

Imagine that you are a researcher hired by the school administration to study student safety and care around campus. They have asked you to do a survey on the average weight of backpacks of grade 11 students around the campus, as they have read articles that many students abroad suffer from back, neck, and shoulder pains due to either excessively heavy or improperly used backpacks. Based on your research, you discover that the recommended weight of a backpack should be no more than 10 to 15 percent of the child's body weight. Prepare a presentation to the administration on this issue by doing the following:



- Pick a simple random sample of at least 30 students (from one or several classes, if necessary) who use backpacks, and weigh them as well as their backpacks. Then get the percentage weight of a student's backpack relative to his or her weight. This will be your first sample of size 30.
- For those students in your sample who carry additional bags aside from their backpacks, weigh these bags as well, and add these to the weight of their backpacks. Then, for each student, get the percentage weight of all his or her bags relative to his or her weight. These will comprise your second sample of size 30.
- Use these two samples to test the hypothesis that $H_0: \mu = 15$ vs. $H_a: \mu > 15$. You must have two tests of hypothesis: one for the mean backpack weight and another one for the mean weight of all bags carried. Use the critical region approach, but also include the p -values.
- Prepare a short presentation on your results.

2. Benford's Law

An interesting result on the distribution of the leading digit of many data sets is *Benford's Law*. It states that in many real-life collections of numerical data, the first digit is more likely to be small rather than large. In particular, the digit 1 has about a 30.1% chance of being the first digit, 2 has a 17.6% chance, while 9 has only about a 4.6% chance. In contrast, if each number appears equally often as the first digit, then it would have a $\frac{1}{9} \approx 11.1\%$ chance.



Imagine that you are a statistician who works at the Philippine Statistics Authority. You are to test whether Benford's Law applies to Philippine population data by following the steps below:

- a. Using the detailed census data by the Philippine Statistics Authority in <https://psa.gov.ph/content/highlights-philippine-population-2015-census-population>, choose one of the 18 regions in the country (*Note: For Region IV, you may select between CALABARZON or MIMAROPA region, which have separate files in the given Website*).
- b. Tabulate the first digits of the population of each of the municipalities within the region. Your table should contain three columns: the leading digit, the count, and the percentage.
- c. Use the data you obtained to conduct a test of hypothesis on whether the true proportion of leading digits equal to 1 is $\frac{1}{9}$. Then with the digits 2 and 9, perform a similar test.³
- d. Prepare a written report on your findings.

³ For curious readers, note that a better hypothesis test to use here is a chi-square goodness of fit test, which tests whether the observed frequencies of the digits 1 to 9 conform with the frequencies given by Benford's Law. However, the mechanics of such a test is beyond the scope of this book.

Chapter Exercises

1. Specify the null and alternative hypotheses in the following scenarios. Then describe a situation where one would commit types I and II error in each case.
 - a. The advertising and promotion department of a company has implemented a campus ad campaign for their company's new facial wash product. They have partnered with student organizations from different universities around Metro Manila to increase awareness about their product. The advertising head would like to know whether the campaign has been effective.
 - b. Most buildings nowadays are equipped with fire protection systems. This includes an alarm which alerts everyone in the building as well as the nearest fire station in case of a fire. Relevant to this, the lessor of the BFS Condominium Towers is concerned about whether the building's fire protection system is still functioning properly as the building has already been standing for more than 20 years.
 - c. Various diets have been designed to help people lose weight. Manolo has read about the South Beach diet, which emphasizes eating high-fiber, low-glycemic carbohydrates, unsaturated fats, and lean protein, and wishes to try it for a month to hopefully lose several kilograms in a relatively safe manner.
2. Bukidnon Pineapple, Inc. is concerned that their 16-ounce (oz) can of sliced pineapples is being overfilled. Their quality control department took a random sample of 40 cans and found that the arithmetic mean weight was 16.05 oz, with a sample standard deviation of 0.03 oz. At $\alpha = 0.1$ level of significance, can we conclude that the mean weight is no longer 16 oz?
3. The pH level is a measure of the acidity or alkalinity of water. A reading of 7.0 is neutral, while values in excess of 7.0 indicate alkalinity. Those below 7.0 imply acidity. Suppose that research has indicated that the best chance for a fisherman to catch *tilapia* is when the pH level of the fishing area is in the range 7.5 to 7.9. Based on recent news obtained, you suspect that environmental and human factors are lowering the pH level of your favorite fishing area and wish to determine whether this pH level is now less than 7.5. Suppose a random sample of 30 water specimens from your fishing area gives pH level readings with a mean of 7.3 and a standard deviation of 0.2.
 - a. At $\alpha = 0.05$ level of significance, is this sufficient evidence to conclude that the pH level is now less than 7.5?
 - b. Verify that the same conclusion is obtained using the p -value approach.

4. The Ed4All Foundation is an association composed of students and parents which advocates for affordable education for all Filipinos. Lately, the organization's board was alarmed by the repeated increases in the tuition fees in private high schools, and had received reports that tuition increases for the next school year will average above 10 percent. To verify this, the foundation formed a small group to survey the planned increases of 15 randomly sampled private schools. If the mean and standard deviation of these percentage increases were 10.7 and 1.3, respectively, can they say that the reports are accurate? Test using a 0.10 level of significance.
5. A university library has installed a self-checkout system so that students can scan their own books for checkout. Listed below is the number of students who have used the service during a sample of 16 days during the last semester.

60	54	60	57	59	46	59	45
52	52	56	49	54	58	50	46

Is it reasonable to conclude that the mean number of students using the self-checkout system is more than 50 per day? Use the $\alpha = 0.01$ level of significance.

6. A chemist tested a certain dose of a spray insecticide on 150 insects, in which 102 of them were killed.
 - a. On the basis of this evidence, can we say that the true proportion of insects killed is less than 0.75? Use a 0.05 level of significance.
 - b. Find the p -value. What would be the minimum level of significance with which the null hypothesis would be rejected?
7. A skin care company claims that more than 60% of all target consumers are familiar with the facial wash commercial that they had aired on television and radio during the past month; that is, the company's marketing strategy is effective. A random sample of 400 respondents was recently asked, and only 275 were familiar with the said commercial. At $\alpha = 0.10$ level of significance, test the hypothesis that the company's claim is valid using
 - a. the rejection region approach; and
 - b. the p -value approach.
8. Suppose that the hypothesis $H_0: \mu = 3$ is to be tested against $H_a: \mu \neq 3$ using a sample of size $n = 50$. It was decided not to reject H_0 if the sample mean \bar{X} falls within the interval (2.9, 3.1). If the standard deviation of the sample is 1.25,
 - a. find α if the null hypothesis is true; and
 - b. find β for the alternative hypothesis $\mu > 2.95$.

Chapter 7

Linear Correlation and Simple Linear Regression



An automobile's odometer is an instrument that measures the distance the automobile has traveled. In the business of buying and selling cars, the odometer reading indicates the distance a car has already traveled. For instance, the seller of a second-hand car may present the odometer reading to show that the car is "slightly used." Low odometer reading gives the prospective buyers the impression that the car has not been overly worn out. Hence, the seller of the car with low odometer reading can charge a slightly higher price, as compared to cars of similar models and conditions but with higher odometer readings. Thus, we can conclude that the selling price of a second-hand car may be related to its odometer reading. The lower the odometer reading of a car, the higher its selling price and vice versa. Such relationship between the two variables is a real-life illustration involving *linear correlation* and *simple linear regression*. You will learn about these topics throughout this chapter.

Lesson 1

Linear Correlation

Learning Outcomes

- At the end of this lesson, you should be able to
 - illustrate the nature of bivariate data;
 - construct a scatter plot;
 - describe shape (form), trend (direction), and variation (strength) based on a scatter plot;
 - estimate the strength of correlation between variables based on a scatter plot;
 - calculate the Pearson's sample correlation coefficient; and
 - solve problems involving correlation analysis.

Introduction

In many real-life scientific investigations, the primary objective is to determine if there exists a relationship between two variables. If such relationship can be described mathematically and is sufficiently understood, then it can be used for effectively predicting one variable by the other variable. Below are examples:

1. In business, a retail merchant might want to determine if there exists a relationship between advertising expenses and sales of a product.
2. In human resources, a personnel manager might want to determine if there exists a relationship between an employee's age and his or her number of days of absence from work in a calendar year.
3. In accounting, it may be desired to determine if there exists a relationship between an asset's age and its resale value.
4. In agriculture, it may be desired to determine if there exists a relationship between the height of a tree and the diameter of the trunk of the tree.
5. In pet nutrition, a veterinarian might want to determine if there exists a relationship between the amount of food consumption of a dog and the weight of the dog.

Simple linear correlation is a measure of the degree of the relationship between two variables or the measure of the intensity of their linear dependence. In the late 19th century, English statistician and polymath Sir Francis Galton (1822–1911) created the statistical concept of the correlation coefficient after examining height and forearm measurements. He demonstrated the applications of correlation coefficient in the study of genetics, psychology, and anthropology.

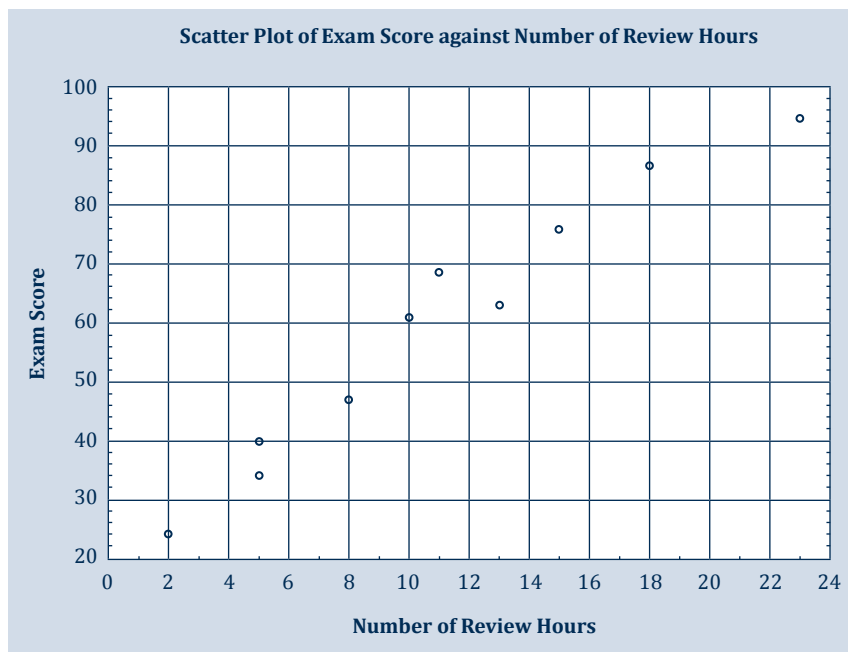
The relationship between two variables may be visualized using a *scatter plot*. Suppose that we have two variables, X and Y , and we measure these from each element in a random sample of n elements. Then we can plot the n ordered pairs (x_i, y_i) , where $i = 1, 2, \dots, n$, on the Cartesian plane, and the resulting two-dimensional graph is the **scatter plot**.

Example 1

Consider the table below which shows the accumulated number of hours in a week that each of the 10 students spends reviewing and their scores in the exam. Data for review times and exam scores, for simplicity in this hypothetical demonstration problem, are rounded off to the nearest whole number. Establish the relationship between the two variables using a scatter plot.

Student	1	2	3	4	5	6	7	8	9	10
Number of Review Hours	5	10	11	15	5	8	13	23	2	18
Score in the Exam	34	61	68	76	40	47	63	94	24	87

If we let X be the number of hours that a student spends reviewing and Y be the score in the exam, then each student corresponds to an ordered pair (x, y) which is represented by a point on the Cartesian plane. The scatter plot for this set of data is shown below.



Examining the scatter plot, we can see that the points closely follow an upward-sloping line. This suggests that there may be a direct linear relationship between the number of hours a student spends reviewing and his or her score in the exam. That is, a student who spends more time reviewing is expected to have a higher score in the exam. It may be concluded that there is a positive correlation between these two variables.

A summary measure that describes the degree and direction of the linear relationship between two quantitative variables is called the **linear correlation coefficient**. These quantitative variables should yield interval to ratio data.

Definition 1

The **population linear correlation coefficient**, denoted by ρ (read as “rho”), is a measure of the extent to which two quantitative variables X and Y tend to move together. In other words, this coefficient measures the degree of association between X and Y . The definition of ρ is given by the ratio of the covariance of X and Y to the product of the standard deviations of X and Y . In symbols,

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The covariance of X and Y is the expected value $E[(X - \mu_X)(Y - \mu_Y)]$, where μ_X and μ_Y are the means of X and Y , respectively. The standard deviation of X is the square root of the variance of X , $\sigma_X = \sqrt{E[(X - \mu_X)^2]}$ and the standard deviation of Y is similarly defined.

A point estimator of ρ is the **Pearson sample product-moment correlation coefficient**, or simply **Pearson's r** , named after Karl Pearson (1857–1936). Pearson is an English statistician who studied various correlation coefficients and other significant statistical concepts. The formula for Pearson's r is given by

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Since $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ and $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$, algebraic manipulation of the above formula yields an alternative formula for Pearson's r as

$$r = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} \sqrt{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}}}.$$

In the formula for Pearson's r , it is customary to denote the numerator

$\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}$ as S_{xy} . In the denominator of Pearson's r , $\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}$ is denoted as

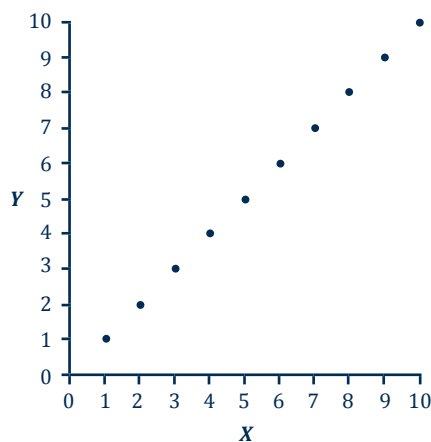
S_{xx} and $\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$ is denoted as S_{yy} . Using these S notations, Pearson's r can be written also

as

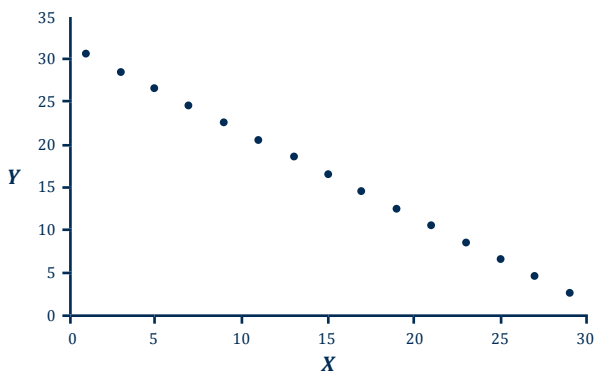
$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}.$$

The value of Pearson's r can range from -1.0 to 1.0 inclusive of endpoints.

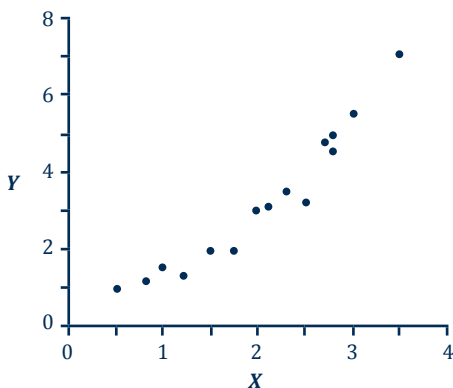
1. If $r = 1.0$, then there is a perfect direct linear relationship between X and Y . Below is an example of a scatter plot for the case when $r = 1.0$. Here, the points lie on a line that is upward sloping to the right.



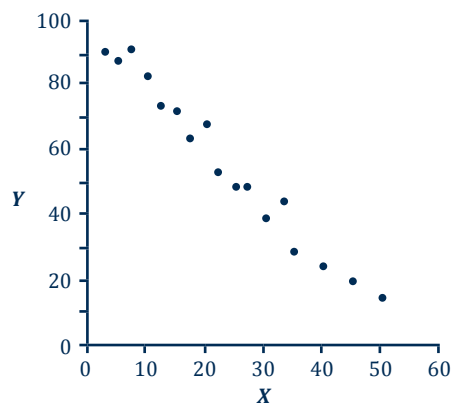
2. If the value of $r = -1.0$, then there is a perfect inverse linear relationship between X and Y . Below is an example of a scatter plot for the case when $r = -1.0$. Here, the points lie on a line that is downward sloping to the right.



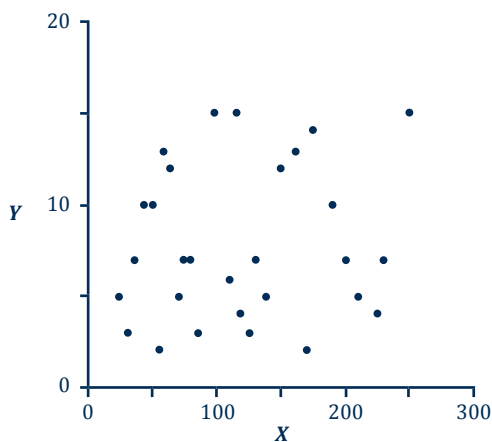
3. If the value of r is close to 1.0, then there is a strong direct linear relationship between X and Y . Below is an example of a scatter plot for the case when r is close to 1.0. Here, the points cluster around a line that is upward sloping to the right. This indicates that as X increases, Y also increases; that is, X and Y go in the same direction.



4. If the value of r is close to -1.0 , then there is a strong inverse linear relationship between X and Y . Below is an example of a scatter plot for the case when r is close to -1.0 . Here, the points cluster around a line that is downward sloping to the right. This indicates that as X increases, the value of Y decreases; that is, X and Y go in opposite directions.



5. If $r = 0$, then there is no linear correlation between X and Y . However, in some cases, X and Y may have a nonlinear relationship such as quadratic or exponential. Below is an example of a scatter plot for the case when $r = 0$ where there is a horizontal band of random points.



Points to Remember

1. The sign of Pearson's r indicates the direction of the linear relationship between X and Y . When $r > 0$, there is a direct relationship between X and Y , where Y is expected to increase as X increases. When $r < 0$, there is an inverse relationship between X and Y , where Y is expected to decrease as X increases.
2. The magnitude of r , its absolute value, indicates the strength or weakness of the linear relationship between X and Y . As a general rule of thumb, we have the following:

Value of $ r $	Description of Linear Relationship Between X and Y
1.0	Perfect
0.90 to under 1.0	Very strong
0.70 to under 0.90	Strong
0.50 to under 0.70	Moderate
0.30 to under 0.50	Weak
Over 0.00 to under 0.30	Very weak
0.0	None

3. On **Correlation** and **Causation**: Correlation or association does not imply causation. A strong linear correlation between two variables does not necessarily mean that there is a cause-and-effect relationship between them. It is possible that these two variables may be correlated to a third variable which is the cause, and the two variables are the effects. As an illustration, suppose that it is found that there is some correlation between the number of cases of leptospirosis and the harvest yield of vegetables and fruits during the rainy season. However, it would be unreasonable to say that leptospirosis cases cause a decrease or increase in the harvest yield of crops, or vice versa. The amount of rainfall or the frequency of typhoons during the rainy season can cause a decrease in harvest yield, as the typhoons could damage the crops, resulting in a drop in supply. In like manner, a high amount of rainfall leads to flood incidence, with the possibility of more people wading in flood water, causing a rise in leptospirosis cases. The amount of rainfall is the third variable which is the cause, and the two variables—cases of leptospirosis and harvest yield of crops—are the effects.

Example 2

Compute the Pearson's r of the given data in example 1. Then interpret the obtained value.

Student	1	2	3	4	5	6	7	8	9	10
Number of Review Hours	5	10	11	15	5	8	13	23	2	18
Scores in the Exam	34	61	68	76	40	47	63	94	24	87

Here, X is the number of hours spent reviewing and Y is the exam score.

Solution:

We construct first the following table:

Student	X	Y	XY	X^2	Y^2
1	5	34	170	25	1,156
2	10	61	610	100	3,721
3	11	68	748	121	4,624
4	15	76	1,140	225	5,776
5	5	40	200	25	1,600
6	8	47	376	64	2,209
7	13	63	819	169	3,969
8	23	94	2,162	529	8,836
9	2	24	48	4	576
10	18	87	1,566	324	7,569
Total	110	594	7,839	1,586	40,036

Therefore, $\sum_{i=1}^{10} x_i = 110$, $\sum_{i=1}^{10} y_i = 594$, $\sum_{i=1}^{10} x_i y_i = 7,839$, $\sum_{i=1}^{10} x_i^2 = 1,586$, and $\sum_{i=1}^{10} y_i^2 = 40,036$.

Substituting these values into the formula of Pearson's r , we have

$$r = \frac{\sum_{i=1}^n x_i y_i - \left(\frac{\sum_{i=1}^n x_i}{n} \right) \left(\frac{\sum_{i=1}^n y_i}{n} \right)}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n}}} \sqrt{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n}}$$

$$\begin{aligned}
 &= \frac{7,839 - \frac{(110)(594)}{10}}{\sqrt{1,586 - \frac{(110)^2}{10}} \sqrt{40,036 - \frac{(594)^2}{10}}} \\
 &= 0.97625
 \end{aligned}$$

A value of $r = 0.97625$ indicates a strong positive linear relationship between the number of hours a student spends reviewing and his or her score in the exam. This means a student who spends more time reviewing is expected to have a higher score in the exam.

Definition 2

The **coefficient of determination**, denoted by r^2 , is the percentage of the total variation in the Y values (the extent to which the Y_i 's are different) that is accounted for or explained by its linear relationship with X .

Points to Remember

1. Since the correlation coefficient r can range from -1 to 1 , the value of the coefficient of determination r^2 can range from 0 to 1 .
2. A high coefficient of determination does not necessarily imply that a cause-and-effect relationship exists between two variables. However, a cause-and-effect relationship between two variables will result in a high coefficient of determination.
3. Suppose that two variables X and Y have a coefficient of determination r^2 and another set of two variables W and Z have a coefficient of determination of $2r^2$. Then it can be concluded that the linear relationship between W and Z is twice as strong as the linear relationship between X and Y .

Example 3

Compute the coefficient of determination of the given data in example 1. Then interpret the obtained value.

Solution:

From example 2, we have computed the correlation coefficient to be 0.97625. Thus, the coefficient of determination $r^2 = (0.97625)^2 = 0.9531$, meaning 95.31% of the total variation in Y , which is the exam scores, is accounted for or explained by its linear relationship with X , which is the number of hours spent reviewing. The remaining 4.69% of the variation in exam scores is explained by other factors affecting it.

Inferences can be drawn regarding the true linear relationship between X and Y . A test of whether a significant linear relationship exists between X and Y may be carried out using a t -test of the population linear correlation coefficient ρ . Then we have the following hypotheses:

$H_0 : \rho = 0$ (There is no linear correlation between X and Y .)

$H_a : \rho \neq 0$ (There exists a linear correlation between X and Y .)

The test statistic for testing the population correlation coefficient is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

This t -test statistic has a Student's t -distribution with $n - 2$ degrees of freedom. At the level of significance α , we reject H_0 when the computed absolute value of the test statistic exceeds the critical number $t_{\alpha/2}$ for a two-tailed test. That is, reject H_0 when $|t| > t_{\alpha/2, n-2}$.

Example 4

Given the data from example 1, is there reason to believe that the number of hours spent reviewing and the exam score have a non-zero correlation? Use a $\alpha = 0.05$ level of significance.

Solution:

We use the six-step rule of testing the hypothesis as follows:

Step 1: $H_0 : \rho = 0$ (There is no linear correlation between the number of hours spent reviewing and the exam score.)

$H_a : \rho \neq 0$ (There exists a linear correlation between the number of hours spent reviewing and the exam score.)

Step 2: $\alpha = 0.05$

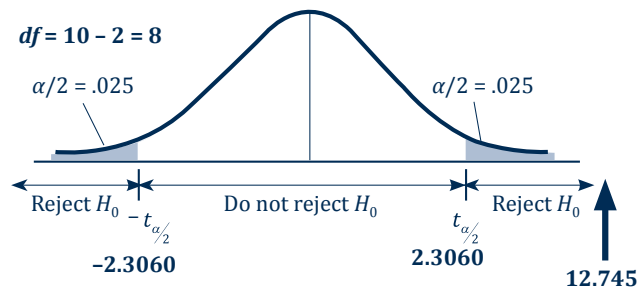
Step 3: The test statistic to be used is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

The critical numbers from the t -distribution with $\alpha/2 = 0.025$ and $n - 2 = 10 - 2 = 8$ degrees of freedom are ± 2.306 .

Step 4: The test statistic value is computed as follows:

$$t = \frac{0.97625\sqrt{10-2}}{\sqrt{1-0.95306}} = 12.745$$

Step 5: We will reject H_0 when $|t| > t_{\alpha/2, n-2}$. Since the computed value of the test statistic, 12.745, is greater than the critical number 2.306, there is sufficient evidence to reject H_0 . This is illustrated in the figure below.



Step 6: At $\alpha = 0.05$, there is sufficient evidence that there exists a significant linear relationship between the number of hours students spend reviewing and their scores in the exam.

Using available statistical software, the p -value of the test is found to be $1.3531820523985 \times 10^{-6}$ or 0.00000135, which is much less than the level of significance of 0.05. Then we can conclude that there is sufficient evidence to reject H_0 .

Example 5

The following data represent the fat contents, in grams, and sodium contents, in milligrams, of two-tablespoon servings of different peanut butter brands. Compute for the Pearson's r , the coefficient of determination of the following set of data. Then test if there is a significant linear relationship between the fat and sodium contents of the peanut butter. Use a 0.05 level of significance.

Brand	A	B	C	D	E	F	G	H	I	J	K	L
Fat	15	16	16	16	16	16	16	12	12	16	16	17
Sodium	100	110	110	65	105	135	150	200	115	150	110	140

Source: <http://www.eatthis.com/peanut-butter-ranked>

Solution:

Let X be the fat content (in grams) and let Y be the sodium content (in milligrams). Then we have the table below.

Brand	X	Y	XY	X^2	Y^2
A	15	100	1,500	225	10,000
B	16	110	1,760	256	12,100
C	16	110	1,760	256	12,100
D	16	65	1,040	256	4,225
E	16	105	1,680	256	11,025
F	16	135	2,160	256	18,225
G	16	150	2,400	256	22,500
H	12	200	2,400	144	40,000
I	12	115	1,380	144	13,225
J	16	150	2,400	256	22,500
K	16	110	1,760	256	12,100
L	17	140	2,380	289	19,600
Total	184	1,490	22,620	2,850	197,600

Thus, $\sum_{i=1}^{12} x_i = 184$, $\sum_{i=1}^{12} y_i = 1,490$, $\sum_{i=1}^{12} x_i y_i = 22,620$, $\sum_{i=1}^{12} x_i^2 = 2,850$, and $\sum_{i=1}^{12} y_i^2 = 197,600$.

Substituting these values into the formula for Pearson's r , we obtain

$$\begin{aligned}
 r &= \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} \sqrt{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}}} \\
 &= \frac{22,620 - \frac{(184)(1,490)}{12}}{\sqrt{2,850 - \frac{(184)^2}{12}} \sqrt{197,600 - \frac{(1,490)^2}{12}}} \\
 &= -0.37727.
 \end{aligned}$$

Then the coefficient of determination is $r^2 = (-0.37727)^2 = 0.14234$. Therefore, only 14.23% of the total variability in sodium content of the peanut butter is accounted for or explained by its linear relationship with its fat content.

Now we employ the six-step rule in testing the hypothesis as follows:

Step 1: $H_0 : \rho = 0$ (There is no linear correlation between the fat and sodium contents of the peanut butter.)

$H_a : \rho \neq 0$ (There exists a linear correlation between the fat and sodium contents of the peanut butter.)

Step 2: $\alpha = 0.05$

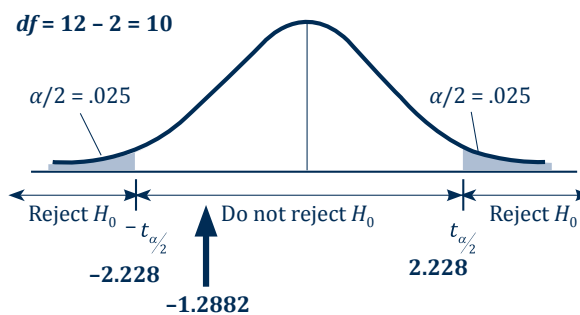
Step 3: The test statistic to be used is $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

The critical numbers from the t -distribution with $\alpha/2 = 0.025$ and $n - 2 = 12 - 2 = 10$ degrees of freedom are ± 2.228 .

Step 4: The test statistic value is computed as follows:

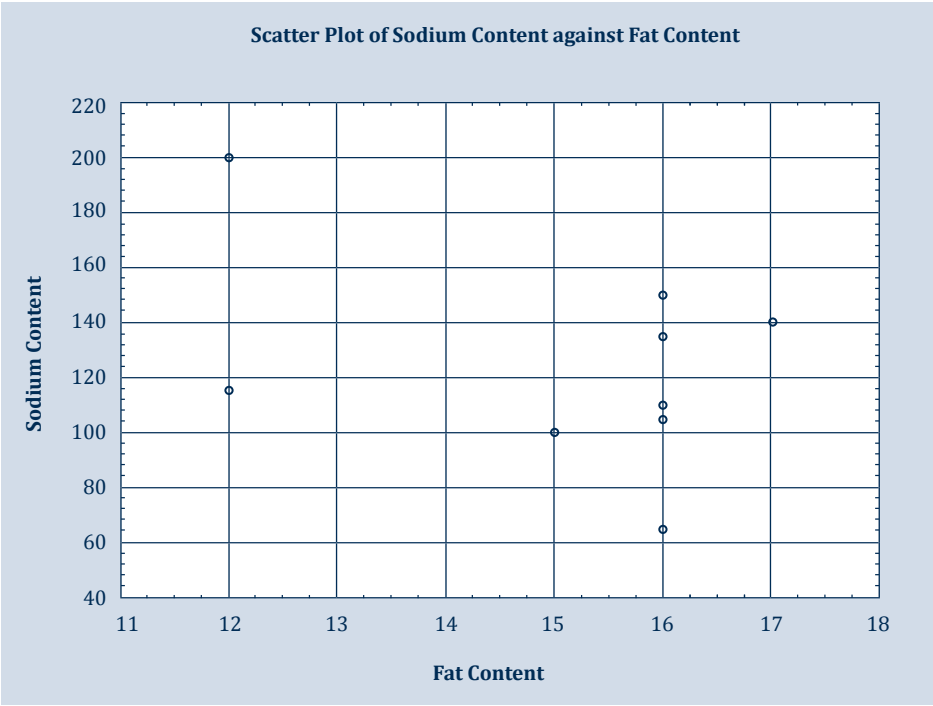
$$t = \frac{-0.37727\sqrt{12-2}}{\sqrt{1-0.14234}} = -1.2882$$

Step 5: We will reject H_0 when $|t| > t_{\alpha/2, n-2}$. Since the computed value -1.2882 is greater than the critical value -2.228 , we have insufficient evidence to reject H_0 . This is illustrated in the figure below.



Step 6: At $\alpha = 0.05$, there is insufficient evidence to conclude that there exists a significant linear relationship between fat and sodium contents in a two-tablespoon serving of different peanut butter brands.

Using available statistical software, the p -value of the test is found to be 0.226669, which is greater than the level of significance of 0.05. Thus, there is insufficient evidence to reject H_0 . The figure below shows the scatter plot of the data which suggests that although there is a slight downward trend in the points, there is a weak linear correlation between the variables.



Example 6

In the case of paired data, an individual or entity is measured twice on the same variable of interest, thus resulting in a pair of observations. Common applications of paired data are cases of “before-and-after” comparative studies. As an illustration, suppose we have data collected on 8 persons that have tried an anti-hunger weight-loss pill for a month. Compute Pearson’s r , the coefficient of determination, and carry out a test if there is a significant linear relationship between the weights before and after taking the pill. Use a 0.05 level of significance.

Person (kg)	Zeta	Yelda	Xerxes	Winnie	Viel	Ursula	Tammy	Sean
Weight before	69	80	57	68	89	65	77	59
Weight after	67	76	55	65	87	66	74	54

Solution:

For the calculations, we let X = weight before the pill program, in kilograms and let Y = weight after the pill program. We have determined the following sums:

$$\sum_{i=1}^8 x_i = 564, \sum_{i=1}^8 y_i = 544, \sum_{i=1}^8 x_i y_i = 39,175, \sum_{i=1}^8 x_i^2 = 40,590, \text{ and } \sum_{i=1}^8 y_i^2 = 37,832.$$

Substituting these values into the formula for Pearson's r , we obtain:

$$r = \frac{39,175 - \frac{(564)(544)}{8}}{\sqrt{40,590 - \frac{(564)^2}{8}} \sqrt{37,832 - \frac{(544)^2}{8}}} = 0.986836$$

The coefficient of determination is $r^2 = 0.973845$. Therefore, about 97.38% of the total variability in the weight after the pill program is accounted for, or explained by its linear relationship with the weight before the pill program.

We employ the standard 6-step procedure, as follows:

Step 1: $H_0: \rho = 0$ (no linear correlation between X and Y)

$H_a: \rho \neq 0$ (there exists a linear correlation between X and Y)

Step 2: $\alpha = 0.05$

Step 3: Test Statistic to be used: $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

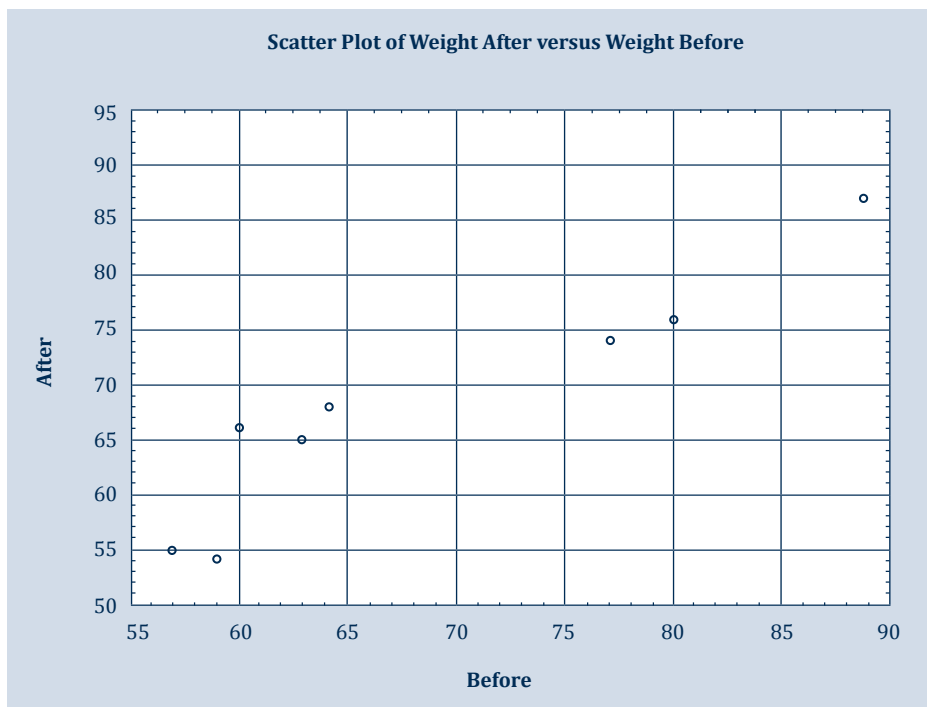
The critical numbers from the t -distribution with $n - 2 = 8 - 2 = 6$ degrees of freedom and $\frac{\alpha}{2} = 0.025$ are ± 2.447 .

Step 4: Computation of the test statistic value: $t = \frac{0.986836\sqrt{8-2}}{\sqrt{1-0.973845}} = 14.94676$

Step 5: We will reject H_0 when $|t| > t_{\alpha/2, n-2}$. Since the computed value of the test statistic, 14.94676, is greater than the critical number 2.447, we have sufficient evidence to reject H_0 .

Step 6: Implication: At the 0.05 level of significance, there is sufficient evidence to say that there exists a significant linear relationship between weights of persons before and after the weight loss pill program.

Using available statistical software, the p -value of the test is found to be 0.00000565, which is much smaller than the level of significance of 0.05. Thus, there is sufficient evidence to reject H_0 . The figure below shows the scatter plot of the data, suggesting that although there is an upward trend in the points, there is strong linear correlation between X and Y .



Let's Practice

I. Write the letter that corresponds to the best answer. Write X if your answer is not among the choices.

- _____ 1. What would happen to the points in a scatter plot as they deteriorate from perfect correlation?
- The points form an upward-sloping line.
 - The points form a downward-sloping line.
 - The points cluster around a horizontal line.
 - The points become more scattered.

- _____ 2. What is the range of the magnitude of the correlation coefficient if there is a very strong correlation between two variables?
- It exceeds 1.0, if the correlation between the two variables is positive.
 - It is either more than 1.0 or less than -1.0 .
 - It is either near 1.0 or near -1.0 .
 - It is less than -1.0 , if the correlation between the two variables is negative.
- _____ 3. Which statement is true?
- If the value of the coefficient of determination is -1.0 , then there is a perfect inverse linear relationship between two variables.
 - It is possible for the value of the coefficient of determination to exceed 1.0.
 - The value of the coefficient of determination cannot be negative.
 - The coefficient of determination is a measure of how well two variables are related as the relationship deteriorates from perfect correlation.
- _____ 4. If the correlation coefficient between X and Y is 0.7, then what is the percentage of variation in Y explained by its linear relationship with X ?
- 0.49%
 - 0.70%
 - 49%
 - 70%
- _____ 5. To perform a t -test whether there is a significant linear relationship between X and Y , what would be the critical values, if there are 15 data points?
Use $\alpha = 0.05$.
- ± 1.761
 - ± 1.753
 - ± 2.160
 - ± 2.145
- _____ 6. Which describes a negative correlation between two variables X and Y ?
- Large values of X are associated with large values of Y .
 - Small values of X are associated with large values of Y .
 - The coefficient of determination is also negative.
 - The points in the scatter plot lie in the third quadrant only.

II. For each of the following sets of data, calculate the correlation coefficient and coefficient of determination then interpret the obtained values. Perform a t -test of the significance of the linear relationship between the variables X and Y , where $\alpha = 0.05$.

1. $n = 5, \sum X = 2.625, \sum Y = 4.8, \sum X^2 = 1.890625, \sum Y^2 = 7.5356, \sum XY = 3.725$
2. $n = 9, \sum X = 219.2, \sum Y = 388.5, \sum X^2 = 5,388.18, \sum Y^2 = 17,055.61, \sum XY = 9,359.75$
3. $n = 10, \sum X = 3.46, \sum Y = 12.06, \sum X^2 = 1.6402, \sum Y^2 = 16.4136, \sum XY = 5.0805$

III. For the following data, construct a scatter plot and calculate the correlation coefficient and coefficient of determination.

1. The table below shows X : the age of a person in years, and Y : the number of one-arm push-ups (workout exercise) that the person does in half a minute.

X	41	65	37	30	59	47	24	38	62	28
Y	9	6	12	15	10	10	19	11	7	16

2. The table shows the sodium content (in mg) and the total carbohydrate content of one serving (42 grams, about half a block of instant noodles with 1 teaspoon of seasoning mix) of a popular brand of instant noodles.

Sodium	800	760	910	860	780	760
Total Carbohydrate	26	27	26	26	25	26

Source: <https://www.nissinfoods.com/Nutrition/Top%20Ramen%20NF%20Ingredients.pdf>

3. The table below shows the food energy and sodium contents of a variety of Subway® sandwiches.

Sandwich	Food Energy (calories)	Sodium Content (milligrams)
Black Forest Ham	290	800
Carved Turkey	330	890
Classic Tuna	480	580
Italian BMT®	410	1,260
Roast Beef	320	660
Oven Roasted Chicken	320	610

Source: <http://www.subway.com/en-us/menunutrition/nutrition>

Lesson 2

Simple Linear Regression

Learning Outcomes

- At the end of this lesson, you should be able to
 - identify the independent and dependent variables in different situations;
 - draw the best-fit line on a scatter plot;
 - calculate the slope and y-intercept of the regression line and interpret the values; and
 - predict the value of the dependent variable given the value of the independent variable.

Introduction

So far, we have discussed linear correlation which is a measure of the degree of association of two quantitative variables X and Y . Now we present another related concept which is *regression analysis*.

Regression analysis deals with constructing a mathematical model to predict one variable by another variable. The basic regression model is the *bivariate linear regression model* which involves only two variables. This model is called **simple linear regression**, whereby one variable is predicted by another variable. In this section, you will learn *simple linear regression analysis*, which examines only the straight-line relationship between the dependent and independent variables.

Points to Remember

- In simple linear regression, the variable denoted by Y refers to dependent, response, or outcome variable; the variable denoted by X refers to independent, explanatory, or predictor variable. Also, the response variable Y is an observable random variable while the predictor variable X is an observable nonrandom variable.
- In simple linear regression analysis, the primary objective is to develop a model to predict the value of Y based on X and to evaluate the impact of X on Y . In other words, simple linear regression enables us to quantify the effect of changes on the predictor variable on the response variable.

Example 1

Identify the dependent variable Y and the independent variable X in each situation:

1. A real-estate agent wants to predict the selling price of a house (in pesos) based on the floor area (in m^2) of the house.
2. A person providing shuttle service wants to predict the resale value of a van (in pesos) based on the age (in years) of the van.
3. It is desired to predict the number of parts of a chemical that dissolve in water based on the water temperature (in $^{\circ}\text{C}$).

Solutions:

1. Y = selling price of a house (in pesos)
 X = floor area of the house (in m^2)
2. Y = resale value of a van (in pesos)
 X = age of the van (in years)
3. Y = parts of a chemical that dissolve in water
 X = water temperature (in $^{\circ}\text{C}$)

Definition 1

The straight-line model is given by

$$Y = \beta_0 + \beta_1 X$$

where Y is the dependent variable;

β_0 is the Y -intercept for the population;

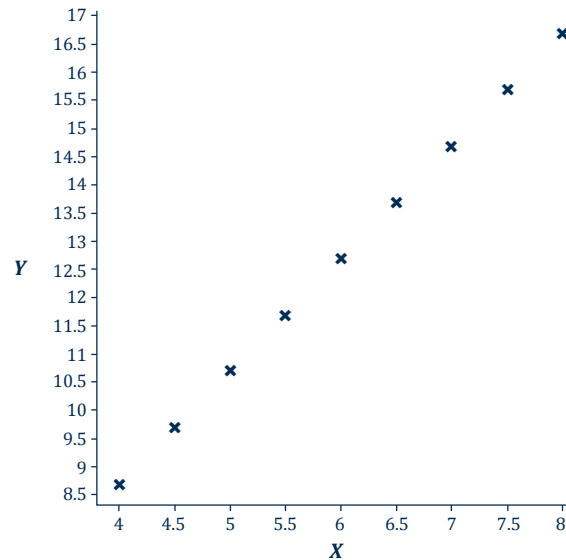
β_1 is the slope for the population; and

X is the independent variable.

The slope of the line is the expected change in Y per unit change in X . The Y -intercept of the line is the mean value of Y when $X = 0$. Such a mathematical model is called a **deterministic model**.

The simple linear regression equation $Y = \beta_0 + \beta_1 X$ is a deterministic model, wherein for a given input value of X , the model returns an exact output for Y . In the scatter plot for a deterministic model, all the points fall exactly on a straight line.

As an illustration, consider the simple linear regression equation $Y = 0.7 + 2X$. If $X = 4$, then the exact predicted value of Y is $Y = 0.7 + 2(4) = 8.7$. Moreover, if $X = 5$, the exact predicted value of Y is $Y = 0.7 + 2(5) = 10.7$. If we obtained more points and plotted them on the Cartesian plane, all the points would fall exactly on a straight line, as shown in the scatter plot below.



However, most of the time, the values of Y are not exactly equal to the values given by the regression equation. For this purpose, a random error term, denoted by ε , is incorporated into the equation. This random error term occurs in the prediction of Y because X does not entirely account for the variation of Y .

Definition 2

The straight-line model is given by

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where Y is the dependent variable;
 β_0 is the Y -intercept for the population;
 β_1 is the slope for the population;
 X is the independent variable; and
 ε is the random error term.

This mathematical model is called a **probabilistic model**.

Definition 3

The simple linear regression equation $Y = \beta_0 + \beta_1 X + \varepsilon$ is a probabilistic model, wherein the portion $\beta_0 + \beta_1 X$ is the deterministic part of the model, and ε is the random error term that allows for the variation in the values of Y for any given value of X . When in all cases ε is zero, all the points fall on a straight line, and the model becomes a deterministic model. β_0 and β_1 are called the **regression beta-coefficients**, or simply **regression coefficients**.

Why does the random error term ε have to be included in the equation? Suppose that you are in your school's computer facility, and you are tasked to develop a model that will predict the number of times a computer breaks down in a year based on its age. It is reasonable to think that the number of times a computer breaks down is related to its age. On the other hand, there are other factors affecting the computer breakdowns that are not accounted for by the age of the computer. For this reason, the regression model to predict computer breakdowns by age of the computer may involve some error. Hence, the random error term ε needs to be included in the equation.

The standard assumptions for the random error term ε in the simple linear regression model are as follows:

1. The ε 's have an expected value of zero.
2. The ε 's are independent.
3. The ε 's are normally distributed.
4. The ε 's have a constant variance σ^2 .

To make it easier for the student to remember these model assumptions, many authors came up with the acronym **L.I.N.E.**, where the "L" stands for the assumption of the linearity between Y and X , thus translating to the ε 's having an expected value of zero; the "I" stands for the assumption of the independence of the ε 's; the "N" stands for the assumption of normality of the ε 's; and the "E" stands for the assumption of an equal variance among the ε 's, referred to as *homoscedasticity*.

The Y -intercept (β_0) and the slope (β_1) are unknown population parameters that can be estimated using sample observations. We denote their respective point estimators as b_0 and b_1 . Therefore, the equation of the best *fitted regression line* contains these point estimators as shown below.

$$\hat{Y} = b_0 + b_1 X$$

where \hat{Y} is the predicted value of the dependent variable;

b_0 is the point estimator of the Y -intercept based on sample data; and

b_1 is the point estimator of the slope based on sample data.

The error term ε is estimated by the regression residual, defined as the difference between the observed value of Y and the predicted value of Y , $e = Y - \hat{Y}$. For a sample of n observations (X_i, Y_i) , $i = 1, 2, \dots, n$, there would be n residual terms e_i , $i = 1, 2, \dots, n$. The process of deriving the point estimators is called the **method of least squares**. This method involves developing the linear regression model by minimizing the sum of the squares of the error values, which is the sum of the squares of the difference between the actual Y and its expected value,

$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2$. Thus the equation $\hat{Y} = b_0 + b_1 X$ is referred to as the **least squares estimated regression equation**. The following formulas for b_0 and b_1 and were obtained using this method, and hence, they are called the **ordinary least squares point estimators** of the Y -intercept and the slope, respectively.

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}$$

$$b_0 = \frac{\sum_{i=1}^n Y_i}{n} - b_1 \left(\frac{\sum_{i=1}^n X_i}{n} \right)$$

Example 2

Find the equation of the fitted regression line of the following set of data from lesson 1 of this chapter. Provide an interpretation of the point estimates and draw the fitted regression line on the scatter plot.

Student	1	2	3	4	5	6	7	8	9	10
Number of Review Hours	5	10	11	15	5	8	13	23	2	18
Scores in the Exam	34	61	68	76	40	47	63	94	24	87

Solution:

From lesson 1, we have obtained the following values for the set of data above.

$$\sum_{i=1}^{10} X_i = 110, \sum_{i=1}^{10} Y_i = 594, \sum_{i=1}^{10} X_i Y_i = 7,839, \sum_{i=1}^{10} X_i^2 = 1,586$$

Substituting these values into the formulas for b_1 and b_0 , we have

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}} = \frac{7,839 - \frac{(110)(594)}{10}}{1,586 - \frac{(110)^2}{10}} = 3.470745$$

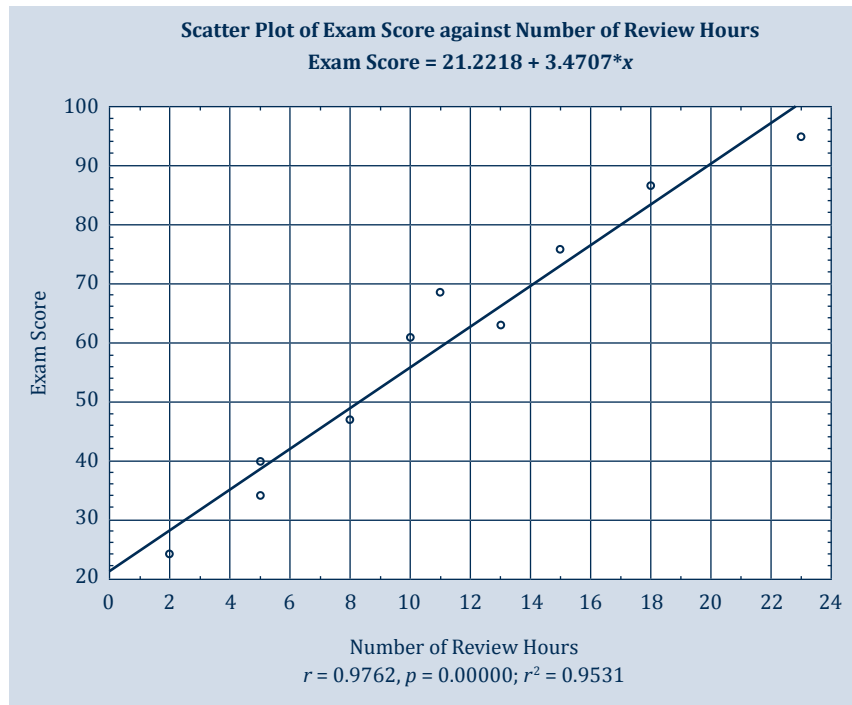
and

$$b_0 = \frac{\sum_{i=1}^n Y_i}{n} - b_1 \left(\frac{\sum_{i=1}^n X_i}{n} \right) = \frac{594}{10} - (3.470745) \left(\frac{110}{10} \right) = 21.221805.$$

We find the point estimate of β_1 to be $b_1 = 3.470745$. This value means that there is an average increase of 3.47074 points in the score in the exam for every one hour increase a student spends reviewing. The point estimate of β_0 is $b_0 = 21.221805$. If $X = 0$ (meaning the student did not review), then the student's expected score in the exam is 21.221805 points.

Now we have the fitted regression line of $\hat{Y} = 21.221805 + 3.470745X$ in the figure below.

Fitted Regression Line for Exam Score and Number of Review Hours



Points to Remember

1. The y -intercept b_0 of the fitted regression line will have a meaningful interpretation if $X = 0$ is within or quite close to the x -data range, as in the following examples:
 - Suppose that X = age of a car in years and Y = the number of times the car had a malfunction. If the fitted regression equation is $\hat{Y} = 0.15 + 3.2X$ based on 15 sample data points with x -values ranging from 1 to 12, then when $X = 0$, the expected number of times that a brand-new car malfunctions is 0.15. It still has a meaningful interpretation since 0 is close to the x -data range.
 - Suppose that X = number of glasses of cow's milk that a 5-year old child drinks in a week and Y = height of the child, in centimeters. If the fitted regression equation is $\hat{Y} = 95.4 + 0.281X$ based on 15 sample data points with x -values ranging from 1 to 10, then when $X = 0$, the expected height of a 5-year old child who doesn't drink cow's milk is 95.4 cm.
2. However, it can happen that the y -intercept b_0 of the fitted regression line will be meaningless, and this happens when 0 is distant from the x -data range or if 0 is not a possible value of X , as in the following situation: X = age of an employee in years, Y = number of days the employee is absent in a calendar year, and the fitted regression equation is $\hat{Y} = 21.5 - 1.38X$ based on 20 sample data points with x -values ranging from 21 to 64. There is no employee whose age is 0.
3. In simple linear regression, one must not extrapolate or predict the response based on an X -value that does not fall within the X -data range, since it is still unproven that the same linear relationship is valid for X -values beyond the X -data range.

Example 3

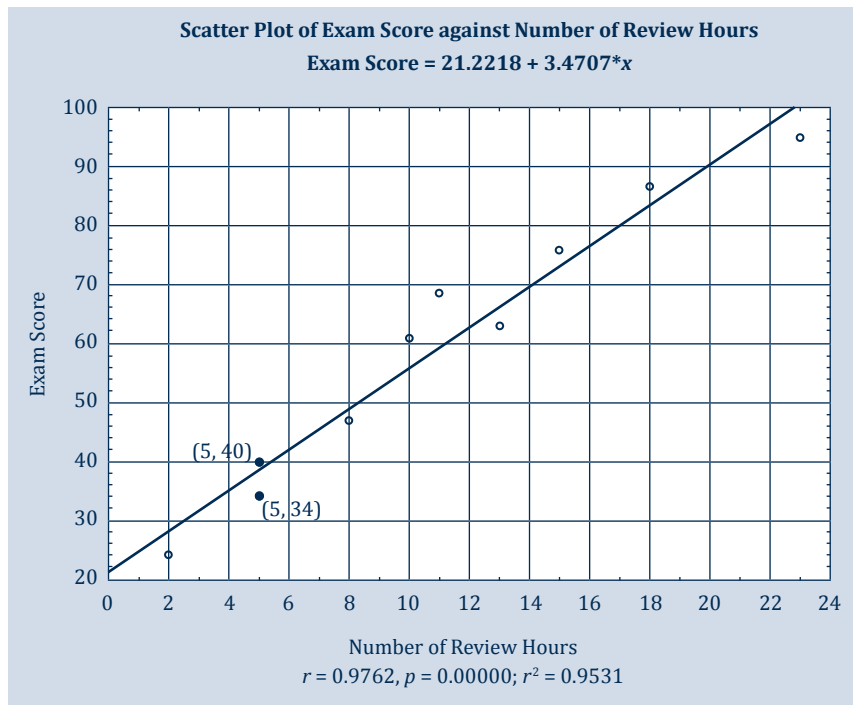
Considering the previous example, predict the exam score of a student who spends 5 hours reviewing.

Solution:

From the given, we have $X = 5$. Substituting this value to $\hat{Y} = 21.221805 + 3.470745X$, we obtain $\hat{Y} = 21.221805 + 3.470745(5) = 38.57553$. Therefore, based on our regression analysis, we predict that a student who spends 5 hours reviewing will get a score, on the average, of 38.57553 points in the exam.

To illustrate the concept of the residual, we look at the data and we see the points (5, 34) and (5, 40). The residual at (5, 34) is $e = 34 - 38.57553 = -4.57553$, whereas the residual at (5, 40) is $e = 40 - 38.57553 = 1.42447$. Graphically, these residuals are the vertical distances between the data points and the regression line. Since (5, 34) is below the regression line, the residual is negative. As for (5, 40) which is above the regression line, the residual is positive.

This is shown in the figure below.



Analysis of the residuals may be performed by plots and statistical tests using computer software. They are useful in checking the model assumptions regarding the errors ε_i .

Let's Practice

I. Write the letter that corresponds to the best answer. Write X if your answer is not among the choices.

- _____ 1. What does the simple linear regression model predict or estimate?
- It predicts or estimates the value of one dependent variable by one independent variable.
 - It predicts or estimates the value of one dependent variable by several independent variables.
 - It predicts or estimates the values of several dependent variables by one independent variable.
 - It predicts or estimates the values of several dependent variables by several independent variables.
- _____ 2. In simple linear regression, which does not describe the variable Y ?
- It is the dependent variable.
 - It is the explanatory variable.
 - It is the response variable.
 - It is the variable to be predicted.
- _____ 3. What does the slope in the fitted regression line $\hat{Y} = b_0 + b_1X$ represent?
- It represents the estimated average change in Y per unit change in X .
 - It represents the predicted value of Y .
 - It represents the predicted value of Y when $X = 0$.
 - It represents the variability around the line of regression.
- _____ 4. In simple linear regression, if the dependent variable is measured in pesos, then what is the unit of the independent variable?
- It may be in any unit.
 - It must be in pesos as well.
 - It must not be in pesos.
 - It must be in terms of a unit of currency other than pesos.

- _____ 5. A regression analysis between sales of a product (in ₱100s) and price in pesos resulted in the equation $\hat{Y} = 400 + 6X$.
Which statement is true about the aforementioned equation?
- An increase of ₱1 in price is associated with an increase of ₱6 in sales.
 - An increase of ₱6 in price is associated with an increase of ₱600 in sales.
 - An increase of ₱1 in price is associated with an increase of ₱600 in sales.
 - An increase of ₱1 in price is associated with a increase of ₱1,000 in sales.
- _____ 6. A regression analysis was applied between sales data (in ₱1,000s) and advertising expenses data (in ₱100s), and the fitted regression line obtained is $\hat{Y} = 18 + 1.2X$. What is the point estimate for the sales if advertising expense is ₱2,000?
- ₱42
 - ₱2,418
 - ₱20,400
 - ₱42,000
- _____ 7. A regression analysis was applied between the age of a computer terminal in terms of years and the number of service maintenance calls of the computer in a year, and the fitted regression line obtained is $\hat{Y} = 0.9 + 2X$. Which statement is *incorrect*?
- The point estimate of the slope is 2.0.
 - The point estimate of the Y -intercept is 0.9.
 - For every one year increase in the age of the computer, the estimated number of service maintenance calls of the computer in a year increases by 2 units.
 - For every one year increase in the age of the computer, the estimated number of service maintenance calls of the computer in a year increases by 0.9 units.
- _____ 8. Which value is minimized through least squares estimation of β_1 and β_0 ?
- $\sum_{i=1}^n (Y_i - \bar{Y})^2$
 - $\sum_{i=1}^n (Y_i - \hat{Y})^2$
 - $\sum_{i=1}^n (\bar{Y} - \hat{Y})^2$
 - $\sum_{i=1}^n (\hat{Y} - Y_i)^2$

- _____ 9. What is true about the slope of the fitted regression line if the correlation coefficient of X and Y is negative?
- The slope can be either negative or positive.
 - The slope is also negative.
 - The slope is positive.
 - The slope is zero.
- _____ 10. What is the magnitude of the coefficient of determination for X and Y if all the points of the scatter plot fall on the fitted regression line?
- 0
 - 1
 - any value between 0 and 1
 - 1.0 if the relationship is positive, or -1.0 if the relationship is negative.

II. For each item, find the least squares estimated regression equation.

1. $n = 10$, $\sum X = 3.46$, $\sum Y = 12.06$, $\sum X^2 = 1.6402$, $\sum Y^2 = 16.4136$, $\sum XY = 5.0805$

Estimate the value of Y when $X = 1.25$.

2. The table below shows X : the age of a person in years, and Y : the number of one-arm push-ups (workout exercise) that the person does in half a minute.

X	41	65	37	30	59	47	24	38	62	28
Y	9	6	12	15	10	10	19	11	7	16

Estimate the value of Y when $X = 36$.

3. Consider the following data on the length of stride (X) of a person and the person's height (Y). Both variables are measured in inches.

X	24	26	16	13	26	18	23	17	12	20	18	16
Y	72	70	59	55	74	64	67	61	53	65	58	57

Estimate the value of Y when $X = 20$.

Lesson 3

Model Adequacy and Inference on the Slope β_1

Learning Outcomes

- At the end of this lesson, you should be able to
 - determine whether a predictor variable contributes significantly to the model; and
 - solve problems involving simple linear regression.

Introduction

In the previous lessons, you learned about the correlation analysis. Under the said topic, we have the Pearson's r and the coefficient of determination, denoted by r^2 , which is defined as the proportion of total variation of the dependent variable Y that is accounted for or explained by its linear relationship with the independent variable X . Recall that the range of values for r^2 is from 0 to 1. If the value of r^2 is 1.0, then X predicts Y perfectly, meaning the total variation in Y is accounted for by X . If the value of r^2 is 0, then none of the variation in Y is accounted for by X , meaning there is no linear regression prediction of Y by X . With this, it is important that an interpretation of r^2 is made to know the amount of variation in the response Y that is reduced on the account of using X as the predictor.

Depending on the use of the simple linear regression model and the context wherein the model was developed, one can have different interpretations on what is considered to be a high or a low value of r^2 . Some researchers in the sciences and engineering who work with precision in systems require r^2 to be as high as 80%. However, some are satisfied to obtain a value of r^2 near 50%.

Points to Remember

1. The coefficient of determination r^2 can only quantify a linear relationship's strength or weakness. It can happen that the value of r^2 is zero, but the variables X and Y have a strong non-linear relationship.
2. It is possible to obtain a high value of r^2 , suggesting a strong linear relationship between X and Y . However, it could happen that upon inspection of the scatter plot, one could deduce that a curvilinear model would even be better than a linear model. Thus, we are cautioned not to misinterpret a high value of r^2 as a gauge that the estimated regression line fits the data well. As an illustration, suppose that r^2 between X and Y is 0.94. This high value of r^2 merely implies that we are better off considering X as the predictor variable than not considering it. However, there could still be some improvement in the model that we have not considered.

Example 1

Consider again the set of data below.

Student	1	2	3	4	5	6	7	8	9	10
Number of Review Hours	5	10	11	15	5	8	13	23	2	18
Scores in the Exam	34	61	68	76	40	47	63	94	24	87

It was computed that its coefficient of determination is 0.95306, which implies that 95.31% of the total variation in exam scores is accounted for or explained by its linear relationship with the number of hours a student spends reviewing. Having this large percentage, we may conclude that the regression line fits well for the data on the time spent reviewing and exam scores. The remaining 4.69% of the variation in exam scores is explained by factors other than the time spent reviewing.

One way to determine how well a linear regression model fits sample data is to carry out a hypothesis test on the slope of the regression model, that is, if the slope is significantly different from zero. A test of whether the population slope of the regression line is different from zero may be carried out using a t -test.

The null hypothesis is $H_0 : \beta_1 = 0$ and the alternative hypothesis is $H_a : \beta_1 \neq 0$. This is a two-tailed test. The test statistic is

$$t = \frac{b_1}{\sqrt{\frac{S_{YY} - b_1^2 S_{XX}}{(n-2)S_{XX}}}}$$

$$\text{where } S_{XX} = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}; \text{ and } S_{YY} = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n}.$$

It is easy to show that this t -test statistic is equivalent to the t -test statistic $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ for a test of the null hypothesis $H_0 : \rho = 0$. We illustrate this in the next example.

Example 2

From the previous example, carry out a t -test of $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ at the 0.05 level of significance.

Solution:

The t -test statistic value is computed as follows:

$$S_{XX} = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i \right)^2}{n} = 1,586 - \frac{(110)^2}{10} = 376$$

$$S_{YY} = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i \right)^2}{n} = 40,036 - \frac{(594)^2}{10} = 4,752.4$$

$$t = \frac{b_1}{\sqrt{\frac{S_{YY} - b_1^2 S_{XX}}{(n-2)S_{XX}}}} = \frac{3.470745}{\sqrt{\frac{4,752.4 - (3.470745)^2 (376)}{(10-2)(376)}}} = 12.745$$

The computed value above is the same as the test statistic value for the test of $H_0 : \rho = 0$ in example 4 of lesson 1. Thus, there is sufficient evidence to reject $H_0 : \beta_1 = 0$ and conclude that the time spent reviewing contributes significant predictability to the exam score.

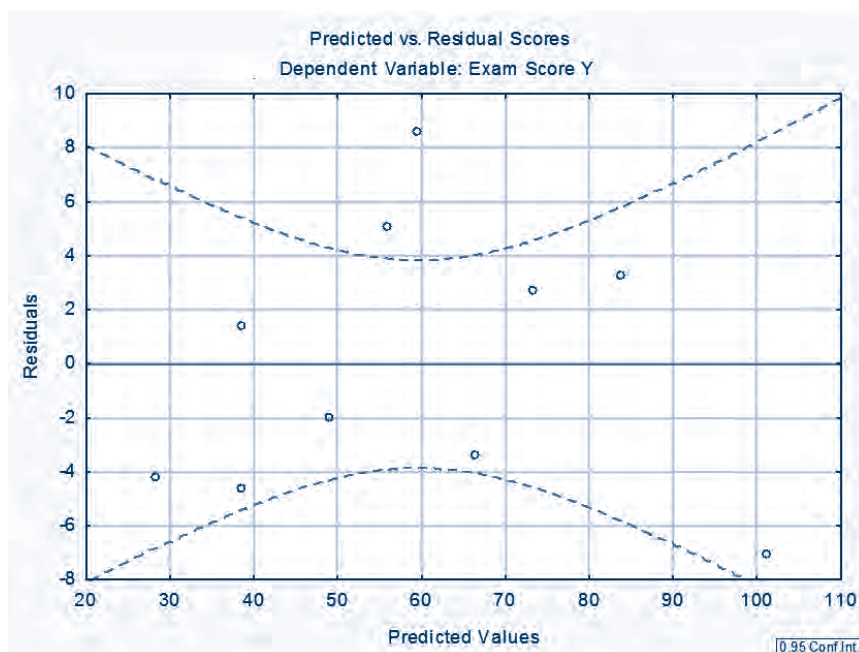
Another way to determine the model adequacy is to create a scatter plot of the residual values versus the fitted values. This plot is called a **residual plot**. It may be used to assess if the model meets the assumption that the errors have an expected value of zero. As an illustration, we use the data in example 1 of lesson 1.

Student	1	2	3	4	5	6	7	8	9	10
Number of Review Hours, X	5	10	11	15	5	8	13	23	2	18
Score in the Exam, Y	34	61	68	76	40	47	63	94	24	87

In lesson 2, we have determined the fitted linear regression equation to be $\hat{Y} = 21.221805 + 3.470745X$. If we substitute the values of the number of review hours X into the equation, we obtain the predicted (fitted) values. The difference between the observed values of Y , the score in the exam, and their corresponding predicted values \hat{Y} are the residuals, $e_i = Y_i - \hat{Y}_i$.

Student	1	2	3	4	5	6	7	8	9	10
Number of Review Hours, X	5	10	11	15	5	8	13	23	2	18
Score in the Exam, Y	34	61	68	76	40	47	63	94	24	87
Predicted Values \hat{Y}	38.576	55.929	59.4	73.283	38.576	48.988	66.342	101.049	28.163	83.695
Residual, e_i	-4.576	5.071	8.6	2.717	1.424	-1.988	-3.341	-7.049	-4.163	3.305

Using statistical software, we create a residual plot with residuals on the vertical axis and predicted values on the horizontal axis. This is shown in the figure below.



A “healthy-looking” residual plot will show no fitted pattern, with the points “randomly bouncing” around the centerline of zero. In our example, although we have only 10 sample data points, we can say there is no pattern in the residual plot, suggesting that the assumption of a linear relationship between the predictor and the response variables is valid. The presence of a pattern such as a curve, a fan-out, or a funnel-in (shaped like a cone or megaphone) in a residual plot may suggest a problem with some aspect of the linear model. Residual analysis such as this is preferred to be carried out with a large number of sample data points. The other model assumptions (independence, normality, and homoscedasticity) for the errors ε may be checked using plots generated from statistical software and by statistical tests using the p -value approach, which are beyond the scope of this text.

Let's Practice

I. Write True if the statement is always true; otherwise, write False.

- _____ 1. The test of $H_0 : \beta_1 = 0$ and the test of $H_0 : \rho = 0$ may or may not yield the same results.
- _____ 2. To reject $H_0 : \beta_1 = 0$ in simple linear regression analysis means that there is a significant linear relationship between the variables X and Y .
- _____ 3. Using a sample of $n = 17$ data points, regression analysis was applied between the aptitude score of a student in mathematics (X) and his or her final grade in calculus (Y). At $\alpha = 0.05$, the critical t -value for testing the significance of the slope for this data is 2.131.
- _____ 4. Smaller value of the coefficient of determination implies that the observations cluster closely around the least squares regression line.
- _____ 5. In simple linear regression analysis, to reject $H_0 : \beta_1 = 0$ implies a cause-and-effect relationship between X and Y .

II. For each item, carry out a t -test to determine if X contributes significant predictability to Y . Use a 0.05 level of significance.

1. $n = 10, \sum X = 3.46, \sum Y = 12.06, \sum X^2 = 1.6402, \sum Y^2 = 16.4136, \sum XY = 5.0805$

2. The table below shows X : the age of a person in years, and Y : the number of one-arm push-ups (workout exercise) that the person does in half a minute.

X	41	65	37	30	59	47	24	38	62	28
Y	9	6	12	15	10	10	19	11	7	16

3. Consider the following data on the length of stride (X) of a person and his or her height (Y). Both variables are measured in inches.

X	24	26	16	13	26	18	23	17	12	20	18	16
Y	72	70	59	55	74	64	67	61	53	65	58	57

Software Tutorial in MS Excel

Simple linear regression analysis and correlation analysis may be done in MS Excel using the Data Analysis tool.

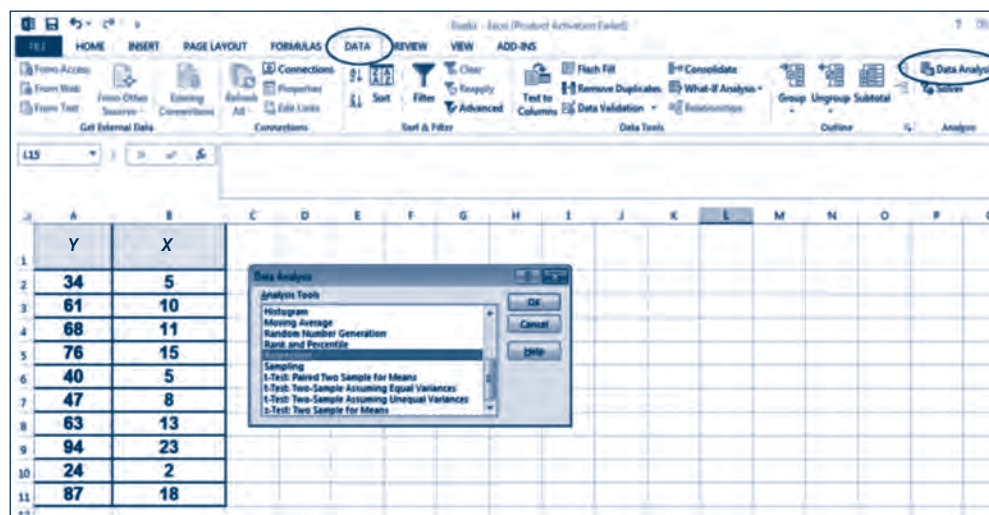
As an illustration, we consider again the following set of data from lesson 1.

Student	1	2	3	4	5	6	7	8	9	10
Number of Hours	5	10	11	15	5	8	13	23	2	18
Scores in the Exam	34	61	68	76	40	47	63	94	24	87

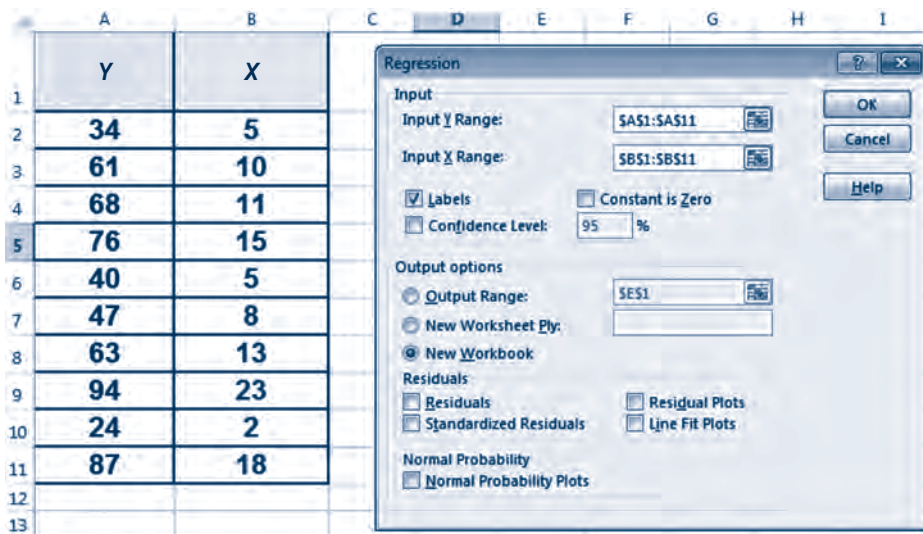
First, input all the observations on an MS Excel spreadsheet as seen in the screenshot below. Let X be the number of hours reviewing and Y be the score in the exam.

	A	B
1	Y	X
2	34	5
3	61	10
4	68	11
5	76	15
6	40	5
7	47	8
8	63	13
9	94	23
10	24	2
11	87	18

Select the *DATA* tab and select the *Data Analysis tool*. When the *Data Analysis* dialog box appears, select *Regression* then click *OK*.



In the *Input Y Range*, drag the mouse on the cells from A1 to A11. In the *Input X Range*, drag the mouse on the cells from B1 to B11. Tick the *Labels* box and the *New Workbook* button then click *OK*.



Then the results will appear as shown below.

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	Regression Statistics						
4	Multiple R	0.976247862					
5	R Square	0.953059887					
6	Adjusted R Square	0.947192373					
7	Standard Error	5.280603558					
8	Observations	10					
9							
10	ANOVA						
11		df	SS	MS	F	Significance F	
12	Regression	1	4529.321809	4529.321809	162.4299275	1.35318E-06	
13	Residual	8	223.0781915	27.88477394			
14	Total	9	4752.4				
15							
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	Intercept	21.22180851	3.429582331	6.187869677	0.000262882	13.31317747	29.13043955
18	Hours Studied X	3.470744681	0.272326335	12.74480002	1.35318E-06	2.842759026	4.098730336

From the figure above, we have the following:

- The Pearson's r is equal to 0.976247862. (See cells A4 and B4.)
- The coefficient of determination r^2 is 0.953059887. (See cells A5 and B5.)
- The point estimate of the slope b_1 is 3.470744681. (See cell B18.)
- The t -test statistic value is 12.7448. (See cell D18.)

Chapter Review

- **Simple linear correlation** is a measure of the degree to which two variables are associated, or a measure of the intensity of the linear dependence of the two variables.
- A summary measure that describes the degree and direction of the linear relationship between two quantitative variables is the **linear correlation coefficient**.
- The **population linear correlation coefficient**, denoted by the Greek letter rho, ρ , measures the dependence or association between X and Y . The definition of ρ is given by the ratio of the covariance of X and Y to the product of the standard deviation of X and the standard deviation of Y .

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

- The covariance of X and Y is the expected value $E[(X - \mu_X)(Y - \mu_Y)]$, where μ_X and μ_Y are the means of X and Y , respectively. The standard deviation of X is the square root of the variance of X , $\sigma_X = \sqrt{E[(X - \mu_X)^2]}$; and the standard deviation of Y is similarly defined.
- A point estimator of ρ , the Pearson sample product moment correlation coefficient or simply **Pearson's r** , is given by

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sqrt{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} \sqrt{\sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}}}$$

- The value of Pearson's r can only range from -1.0 to 1.0 inclusive of endpoints.
- If the value of $r = 1.0$, then there is a perfect direct linear relationship between X and Y . The data points lie on a line that is upward sloping to the right.
- If the value of $r = -1.0$, then there is a perfect inverse linear relationship between X and Y . The data points lie on a line that is downward sloping to the right.
- If the value of r is close to 1.0 , then there is a strong direct linear relationship between X and Y . The data points cluster around a line that is upward sloping to the right, indicating that as X increases, the value of Y also increases, (i.e., X and Y go in the same direction).

- If the value of r is close to -1.0 , then there is a strong inverse linear relationship between X and Y . The data points cluster around a line that is downward sloping to the right, indicating that as X increases, the value of Y decreases, (i.e., X and Y go in opposite directions).
- If the value of r is 0, then there is no linear correlation between X and Y . However, this does not imply that there is an absence of association between the two variables. Two variables X and Y can have zero linear correlation, but may have a non-linear relationship such as quadratic or exponential.
- The **coefficient of determination**, denoted by r^2 , is the percentage of the total variation in the Y values (i.e., the extent to which the Y_i 's are different) that is accounted for or explained by its linear relationship with X .
- In simple linear regression,
 1. the variable designated by Y has three names: dependent variable, response variable, and variable to be predicted.
 2. the variable designated by X has three names: independent variable, explanatory variable, and predictor variable.
 3. the response variable Y is an observable random variable. The predictor variable X is an observable nonrandom variable.
 4. the primary objective in simple linear regression analysis is to develop a model to predict the value of the response variable based on the predictor variable, and to evaluate the impact of the predictor on the response. In other words, simple linear regression enables us to quantify the effect of changes on the predictor variable on the response variable.
- The simple linear regression equation $Y = \beta_0 + \beta_1 X$ is a **deterministic model**, wherein for a given input value of the independent variable X , the model returns an exact output for the dependent variable Y . In the scatter plot for a deterministic model, all the points fall exactly on a straight line.
- The simple linear regression equation $Y = \beta_0 + \beta_1 X + \varepsilon$ is a **probabilistic model**, wherein the portion $\beta_0 + \beta_1 X$ is the deterministic part of the model, and ε is the random error term that allows for the variation in the values of the dependent variable Y for any given value of the independent variable X . β_0 and β_1 are called the regression beta-coefficients, or simply, **regression coefficients**.
- The equation of the best *fitted regression line* is defined as $\hat{Y} = b_0 + b_1 X$, where \hat{Y} is the predicted value of the dependent variable; b_0 is the point estimator of the Y -intercept, based on the sample data; and b_1 is the point estimator of the slope, based on the sample data. The error term in the probabilistic model $Y = \beta_0 + \beta_1 X + \varepsilon$ is estimated by the residual, defined as the difference between the observed value of Y and the predicted value of Y , thus, $e = Y - \hat{Y}$.

- The process of deriving the point estimators, or the *method of least squares*, involves developing the linear regression model by minimizing the sum of the squares of the error values, $\sum_{i=1}^n \varepsilon_i^2$. Thus the equation $\hat{Y} = b_0 + b_1X$ is referred to as the *least squares estimated regression equation*. The following are the formulas for b_0 and b_1 , or the *ordinary least squares point estimators* of the Y -intercept and the slope, respectively.

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - \frac{\left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}$$

$$b_0 = \frac{\sum_{i=1}^n Y_i}{n} - b_1 \left(\frac{\sum_{i=1}^n X_i}{n} \right)$$

- A test of whether the population slope of the regression line is different from zero may be carried out using a t -test. The null hypothesis is $H_0 : \beta_1 = 0$ and the alternative hypothesis is $H_a : \beta_1 \neq 0$ since this is a two-tailed test. The test statistic is

$$t = \frac{b_1}{\sqrt{\frac{S_{YY} - b_1^2 S_{XX}}{(n-2)S_{XX}}}}$$

- The t -test statistic is equivalent to $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ for a test of $H_0 : \rho = 0$.

Chapter Performance Tasks

1. Real-life Measurements

A fitness trainer would like to convince his clients at a gym where he works that inadequate amount of sleep could be related to one's overeating. He needs your help to verify if there is a relationship between a person's waistline measurement and the amount of sleep that the person gets daily. Imagine that you are a teacher of statistics in senior high school, and you are willing to help the fitness trainer.



Collect data from students in your class. Get each student's waistline measurement (in centimeters) and amount of sleep at night (in hours). Based from the data you gathered, make a report, which should include the following:

- a scatter plot, with X = sleep time (horizontal scale) and Y = waistline measurement (vertical scale), using different colors or different plotting symbols to represent the data for males and females. You may use Excel Drawing Tools or make your own scatter plot by hand.
- an interpretation of the scatter plot. Does it look like there is a linear relationship between a person's length of sleep and waistline measurement? Explain.
- a computation of Pearson's r for the data obtained from the males. Also test whether there is a significant linear relationship between X and Y . Use a 0.05 level of significance. Do the same for the data obtained from the females.
- the best fitted regression line for the data obtained from the males. Does it appear that the same straight line could be used to summarize the relationship between length of sleep and waistline measurement for females? Determine the best fitted regression line for the data obtained from the females and compare your answer with the results from the data of the males.

2. For the Coffee Drinkers

Suppose you are a nutritionist studying the possible relationship between the coffee consumption of an adult and his or her systolic blood pressure. Collect data by interviewing 20 adults in your community who drink at least one cup of coffee a day. Ask them the average number of cups of coffee they consumed in a week and their latest blood pressure reading. After collecting the data, perform the following tasks:



- Draw a scatter plot using the data you have collected. Let X be the number of cups of coffee and Y the blood pressure reading. (Note: You may use Excel Drawing Tools to construct the scatter plot.)
- Interpret the scatter plot you have in item (a) on whether or not a relationship exists between the variables X and Y .
- Compute for its Pearson's r and test whether there is a significant linear relationship between these variables. Use a 0.05 level of significance.
- Determine the best fitted regression line for the data you have collected. Explain why you can or cannot use the same straight line to summarize the relationship between the coffee consumption of adults and their blood pressure readings.

Chapter Exercises

I. Identify the dependent and independent variables in the given situations.

1. The dean of a college would like to predict the grade point average (GPA) of a freshman student based on the National Career Assessment Examination (NCAE) rating.
2. An agronomist would like to predict the growth in height of a certain tree based on annual precipitation.
3. A bank manager would like to predict the amount of time it takes to complete the processing of a car loan based on the number of car loan applications processed in a day.
4. You are studying the possibility of predicting water consumption in a city per day based on the high temperature in a day.
5. A medical study involves constructing a model to predict the number of deaths due to lung cancer based on the cigarette smoking rate.
6. In economics, it is desired to predict the demand for a commodity based on the selling price of the commodity.
7. In psychology, a study involves constructing a model to predict a person's reaction time to a given stimulus based on the amount of alcoholic beverage consumption.

II. Write the letter that corresponds to your answer. Write X if your answer is not among the choices.

- _____ 1. Which Pearson's r value indicates a strong linear relationship?
- a. -0.95
 - b. 0.08
 - c. 0.75
 - d. 0.36
- _____ 2. If the correlation coefficient between X and Y is 0.9, then what is the proportion of variation in Y explained by its linear relationship with X ?
- a. 0.81%
 - b. 0.90%
 - c. 81%
 - d. 90%

- _____ 3. If the coefficient of determination between X and Y is 1, which is true about its correlation coefficient?
- It is equal to 1.
 - It must be equal to -1 .
 - It can be any value between -1 and 1.
 - It can be either 1 or -1 .
- _____ 4. A regression analysis between sales (in ₱1,000s) and price (in pesos) resulted in the equation $\hat{Y} = 12,000 - 5X$.
- Which statement is true about the aforementioned equation?
- An increase of ₱1 in price is associated with a decrease of ₱5 in sales.
 - An increase of ₱5 in price is associated with an increase of ₱5,000 in sales.
 - An increase of ₱1 in price is associated with a decrease of ₱7,000 in sales.
 - An increase of ₱1 in price is associated with a decrease of ₱5000 in sales.
- _____ 5. A regression analysis was applied between sales data (in ₱1,000s) and advertising expense data (in ₱100s) and the fitted regression equation $\hat{Y} = 150 + 2.4x$ was obtained. What is the point estimate for the sales if advertising expense is ₱10,000?
- ₱390
 - ₱24,150
 - ₱174,000
 - ₱390,000

III. Analyze and solve each problem.

1. A regression analysis resulted in the following information:

$$\sum_{i=1}^{21} X_i = 443, \sum_{i=1}^{21} Y_i = 368, \sum_{i=1}^{21} X_i Y_i = 8,326, \sum_{i=1}^{21} X_i^2 = 9,545, \sum_{i=1}^{21} Y_i^2 = 8,518,$$

where $n = 21$.

- Compute the Pearson's r and interpret it.
- Compute the coefficient of determination and interpret it.
- Determine the fitted regression line.
- Find the predicted value of Y when $X = 20$.

2. Consider the data below that shows the final grades in English and Math of 10 randomly selected students from a certain public school.

Student	1	2	3	4	5	6	7	8	9	10
Grade in English	84	93	85	89	73	95	84	90	77	84
Grade in Math	83	87	81	79	78	93	83	83	75	74

- Draw a scatter plot, with X = No mathematical logic course and Y = With mathematical logic course
 - Compute Pearson's r and test whether there is a significant linear relationship between X and Y .
3. A certain study aims to construct a simple linear regression model that can predict the farthest distance that a driver can see clearly based on his or her age. Consider the gathered data below of twelve randomly selected drivers for the study.

Driver	1	2	3	4	5	6	7	8	9	10	11	12
Age in years (X)	19	21	24	26	30	33	38	42	48	54	63	66
Distance in meters (Y)	170	140	185	150	165	145	130	125	135	120	120	110

- Draw a scatter plot.
- Compute Pearson's r and interpret the answer.
- Compute the coefficient of determination and interpret the answer.
- Determine the best fitted regression line for the data.
- What is the predicted farthest distance that a 30-year-old driver can see?
- Carry out a test of the significance of the linear relationship between X and Y . Use a 0.05 level of significance.

4. The body mass index (BMI) is defined as the ratio of a person's weight in kilograms to the square of the person's height in meters. A value of a BMI may indicate that a person is underweight, normal, or overweight. Suppose you are studying the possible relationship of a student's BMI and his or her academic performance. Consider the data below of 20 students.

Student	BMI (X)	GPA (Y)
1	22.4	3.59
2	19.6	2.80
3	23.0	2.89
4	20.9	3.17
5	25.2	1.94
6	18.9	3.00
7	26.6	3.20
8	20.5	2.63
9	25.9	2.70
10	21.0	3.86
11	24.1	3.10
12	28.0	3.39
13	19.3	2.00
14	18.4	1.75
15	20.0	2.89
16	17.9	3.43
17	23.4	1.50
18	22.7	3.62
19	22.9	3.20
20	25.4	3.50

- Draw a scatter plot.
- Compute Pearson's r and interpret the answer.
- Compute the coefficient of determination and interpret the answer.
- Determine the best fitted regression line for the data.
- What is the predicted grade point average of a student whose BMI is 20?
- Carry out a test of the significance of the linear relationship between X and Y . Use a 0.05 level of significance.

Appendices

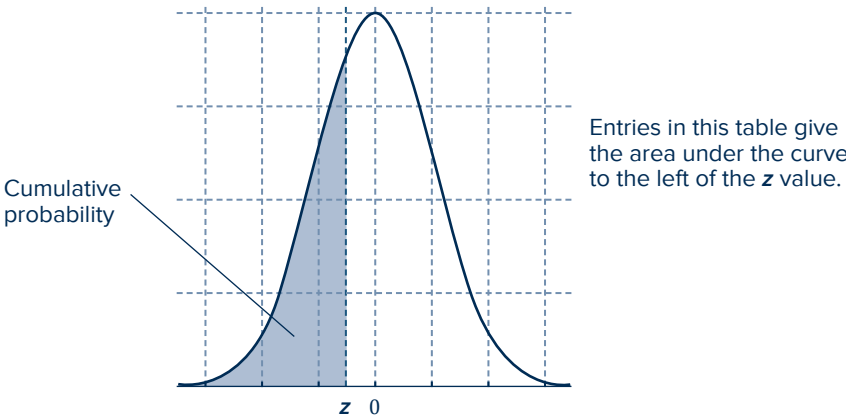
Appendix A: Table of Random Numbers

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	10480	15011	01536	02011	81647	91648	69179	14194	62590	36207	20969	99570	91291	90700
	22368	46573	25595	85393	30995	89198	27982	53402	93965	34095	52666	19174	39615	99505
	24130	48360	22527	97265	76393	64809	15179	24830	49340	32081	30680	19655	63348	58629
	42167	93093	06243	61680	07856	16376	39440	53537	71341	57004	00849	74917	97758	16379
5	37570	39975	81837	16656	06121	91782	60468	81305	49684	60672	14110	06927	01263	54613
	77921	06907	11008	42751	27756	53498	18602	70659	90655	15053	21916	81825	44394	42880
	99562	72905	56420	69994	98872	31016	71194	18738	44013	48840	63213	21069	10634	12952
	96301	91977	05463	07972	18876	20922	94595	56869	69014	60045	18425	84903	42508	32307
	89579	14342	63661	10281	17453	18103	57740	84378	25331	12566	58678	44947	05585	56941
10	85475	36857	53342	53988	53060	59533	38867	62300	08158	17983	16439	11458	18593	64952
	28918	69578	88231	33276	70997	79936	56865	05859	90106	31595	01547	85590	91610	78188
	63553	40961	48235	03427	49626	69445	18663	72695	52180	20847	12234	90511	33703	90322
	09429	93969	52636	92737	88974	33488	36320	17617	30015	08272	84115	27156	30613	74952
	10365	61129	87529	85689	48237	52267	67689	93394	01511	26358	85104	20285	29975	89868
15	07119	97336	71048	08178	77233	13916	47564	81056	97735	85977	29372	74461	28551	90707
	51085	12765	51821	51259	77452	16308	60756	92144	49442	53900	70960	63990	75601	40719
	02368	21382	52404	60268	89368	19885	55322	44819	01188	65255	64835	44919	05944	55157
	01011	54092	33362	94904	31273	04146	18594	29852	71585	85030	51132	01915	92747	64951
	52162	53916	46369	58586	23216	14513	83149	98736	23495	64350	94738	17752	35156	35749
20	07056	97628	33787	09998	42698	06691	76988	13602	51851	46104	88916	19509	25625	58104
	48663	91245	85828	14346	09172	30168	90229	04734	59193	22178	30421	61666	99904	32812
	54164	58492	22421	74103	47070	25306	76468	26384	58151	06646	21524	15227	96909	44592
	32639	32363	05597	24200	13363	38005	94342	28728	35806	06912	17012	64161	18296	22851
	29334	27001	87637	87308	58731	00256	45834	15398	46557	41135	10367	07684	36188	18510
25	02488	33062	28834	07351	19731	92420	60952	61280	50001	67658	32586	86679	50720	94953
	81525	72295	04839	96423	24878	82651	66566	14778	76797	14780	13300	87074	79666	95725
	29676	20591	68086	26432	46901	20849	89768	81536	86645	12659	92259	57102	80428	25280
	00742	57392	39064	66432	84673	40027	32832	61362	98947	96067	64760	64584	96096	98253
	05366	04213	25669	26422	44407	44048	37937	63904	45766	66134	75470	66520	34693	90449
30	91921	26418	64117	94305	26766	25940	39972	22209	71500	64568	91402	42416	07844	69618
	00582	04711	87917	77341	42206	35126	74087	99547	81817	42607	43808	76655	62028	76630
	00725	69884	62797	56170	86324	88072	76222	36086	84637	93161	76038	65855	77919	88006
	69011	65795	95876	55293	18988	27354	26575	08625	40801	59920	29841	80150	12777	48501
	25976	57948	29888	88604	67917	48708	18912	82271	65424	69774	33611	54262	85963	03547
35	09763	83473	73577	12908	30883	18317	28290	35797	05998	41688	34952	37888	38917	88050
	91567	42595	27958	30134	04024	86385	29880	99730	55536	84855	29080	09250	79656	73211
	17955	56349	90999	49127	20044	59931	06115	20542	18059	02008	73708	83517	36103	42791
	46503	18584	18845	49618	02304	51038	20655	58727	28168	15475	56942	53389	20562	87338
	92157	89634	94824	78171	84610	82834	09922	25417	44137	48413	25555	21246	35509	20468
40	14577	62765	35605	81263	39667	47358	56873	56307	61607	49518	89656	20103	77490	18062
	98427	07523	33362	64270	01638	92477	66969	98420	04880	45585	46565	04102	46880	45709
	34914	63976	88720	82765	34476	17032	87589	40836	32427	70002	70663	88863	77775	69348
	70060	28277	39475	46473	23219	53416	94970	25832	69975	94884	19661	72828	00102	66794
	53976	54914	06990	67245	68350	82948	11398	42878	80287	88267	47363	46634	06541	97809
45	76072	29515	40980	07391	58745	25774	22987	80059	39911	96189	41151	14222	60697	59583
	90725	52210	83974	29992	65831	38857	50490	83765	55657	14361	31720	57375	56228	41546
	64364	67412	33339	31926	14883	24413	59744	92351	97473	89286	35931	04110	23726	51900
	08962	00358	31662	25388	61642	34072	81249	35648	56891	69352	48373	45578	78547	81788
	95012	68379	93526	70765	10592	04542	76463	54328	02349	17247	28865	14777	62730	92277
50	15664	10493	20492	38391	91132	21999	59516	81652	27195	48223	46751	22923	32261	85653

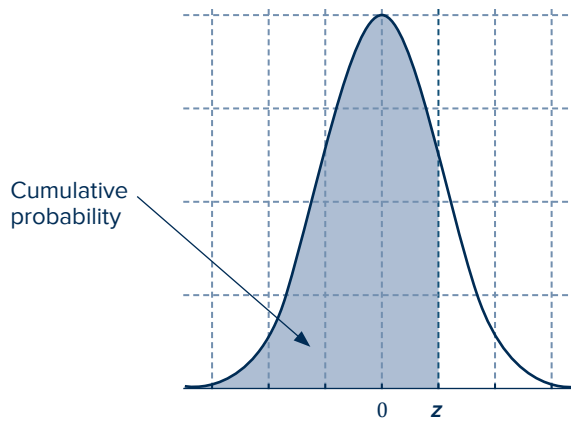
	15	16	17	18	19	20	21	22	23	24	25	26	27	28
51	16408	81899	04153	53381	79401	21438	83035	92350	36693	31238	59649	91754	72772	02338
	18629	81953	05520	91962	04739	13092	97662	24822	94730	06496	35090	04822	86774	98289
	73115	35101	47498	86737	99016	71060	88824	71013	18735	20286	23153	72924	35165	43040
	57491	16703	23167	49323	45021	33132	12544	41035	80780	45393	44812	12515	98931	91202
55	30405	83946	23792	14422	15059	45799	22716	19792	09983	74353	68668	30429	70735	25499
	16631	35006	85900	98275	32388	52390	16815	69298	82732	38480	73817	32523	41961	44437
	96773	20206	42559	78985	05300	22164	24369	54224	35083	19687	11052	91491	60383	19746
	38935	64202	14349	82674	66523	44133	00697	35552	35970	19124	63318	29686	03387	59846
	31624	76384	17403	53363	44167	64486	64758	75366	76554	31601	12614	33072	60332	92325
60	78919	19474	23632	27889	47914	02584	37680	20801	72152	39339	34806	08930	85001	87820
	03931	33309	57047	74211	63445	17361	62825	39908	05607	91284	68833	25570	38818	46920
	74426	33278	43972	10119	89917	15665	52872	73823	73144	88662	88970	74492	51805	93378
	09066	00903	20795	95452	92648	45454	09552	88815	16553	51125	79375	97596	16296	66092
	42238	12426	87025	14267	20979	04508	64535	31355	86064	29472	47689	05974	52468	16834
65	16153	08002	26504	41744	81959	65642	74240	56302	00033	67107	77510	70625	28725	34191
	21457	40742	29820	96783	29400	21840	15035	34537	33310	06116	95240	15957	16572	06004
	21581	57802	02050	89728	17937	37621	47075	42080	97403	48626	68995	43805	33386	21597
	55612	78095	83197	33732	05810	24813	86902	60397	16489	03264	88525	42786	05269	92532
	44657	66999	99324	51281	84463	60563	79312	93454	68876	25471	93911	25650	12682	73572
70	91340	84979	46949	81973	37949	61023	43997	15263	80644	43942	89203	71795	99533	50501
	91227	21199	31935	27022	84067	05462	35216	14486	29891	68607	41867	14951	91696	85065
	50001	38140	66321	19924	72163	09538	12151	06878	91903	18749	34405	56087	82790	70925
	65390	05224	72958	28609	81406	39147	25549	48542	42627	45233	57202	94617	23772	07896
	27504	96131	83944	41575	10573	08619	64482	73923	36152	05184	94142	25299	84387	34925
75	37169	94851	39117	89632	00959	16487	65536	49071	39782	17095	02330	74301	00275	48280
	11508	70225	51111	38351	19444	66499	71945	05422	13442	78675	84081	66938	93654	59894
	37449	30362	06694	54690	04052	53115	62757	95348	78662	11163	81651	50245	34971	52924
	46515	70331	85922	38329	57015	15765	97161	17869	45349	61796	66345	81073	49106	79860
	30986	81223	42416	58353	21532	30502	32305	86482	05174	07901	54339	58861	74818	46942
80	63798	64995	46583	09785	44160	78128	83991	42865	92520	83531	80377	35909	81250	54238
	82486	84846	99254	67632	43218	50076	21361	64816	51202	88124	41870	52689	51275	83556
	21885	32906	92431	09060	64297	51674	64126	62570	26123	05155	59194	52799	28225	85762
	60336	98782	07408	53458	13564	59089	26445	29789	85205	41001	12535	12133	14645	23541
	43937	46891	24010	25560	86355	33941	25786	54990	71899	15475	95434	98227	21824	19585
85	97656	63175	89303	16275	07100	92063	21942	18611	47348	20203	18534	03862	78095	50136
	03299	01221	05418	38982	55758	92237	26759	86367	21216	98442	08303	56613	91511	75928
	79626	06486	03574	17668	07785	76020	79924	25651	88325	88428	85076	72811	22717	50585
	85636	68335	47539	03129	65651	11977	02510	26113	99447	68645	34327	15152	55230	93448
	18039	14367	61337	06177	12143	46609	32989	74014	64708	00533	35398	58408	13261	47908
90	08362	15656	60627	36478	65648	16764	53412	09013	07832	41574	17639	82163	60859	75567
	79556	29068	04142	16268	15387	12856	66227	38358	22478	73373	88732	09443	82558	05250
	92608	82674	27072	32534	17075	27698	98204	63863	11951	34648	88022	56148	34925	57031
	23982	25835	40055	67006	12293	02753	14827	23235	35071	99704	37543	11601	35503	85171
	09915	96306	05908	97901	28395	14186	00821	80703	70426	75647	76310	88717	37890	40129
95	59037	33300	26695	62247	69927	76123	50842	43834	86654	70959	79725	93872	28117	19233
	42488	78077	69882	61657	34136	79180	97526	43092	04098	73571	80799	76536	71255	64239
	46764	86273	63003	93017	31204	36692	40202	35275	57306	55543	53203	18098	47625	88684
	03237	45430	55417	63282	90816	17349	88298	90183	36600	78406	06216	95787	42579	90730
	86591	81482	52667	61582	14972	90053	89534	76036	49199	43716	97548	04379	46370	28672
100	38534	01715	94964	87288	65680	43772	39560	12918	86537	62738	19636	51132	25739	56947

Abridged from Handbook of Tables for Probability and Statistics, Second Edition, edited by William H. Beyer (Cleveland: The Chemical Rubber Company, 1968)

Appendix B: Cumulative Probabilities for the Standard Normal Distribution



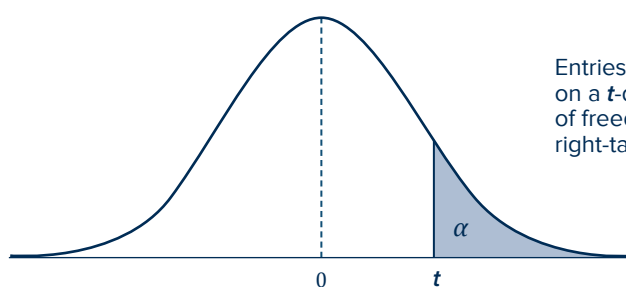
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.409	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641



Entries in this table give the area under the curve to the left of the z value.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Appendix C: Critical Values of the t -distribution



Entries in this table are values on a t -distribution with degrees of freedom (df) for selected right-tail probabilities.

Critical Values of t

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	df
1	3.078	6.134	12.706	31.821	63.657	1
2	1.886	2.920	4.303	6.965	9.925	2
3	1.638	2.353	3.182	4.541	5.841	3
4	1.533	2.162	2.776	3.747	4.604	4
5	1.476	2.015	2.571	3.365	4.032	5
6	1.440	1.943	2.447	3.143	3.707	6
7	1.415	1.895	2.365	2.998	3.499	7
8	1.397	1.860	2.306	2.896	3.355	8
9	1.383	1.833	2.262	2.821	3.250	9
10	1.372	1.812	2.228	2.764	3.169	10
11	1.363	1.796	2.201	2.718	3.106	11
12	1.356	1.782	2.179	2.681	3.055	12
13	1.350	1.771	2.160	2.650	3.012	13
14	1.345	1.761	2.145	2.624	2.977	14
15	1.341	1.753	2.131	2.602	2.947	15
16	1.337	1.746	2.120	2.583	2.921	16
17	1.333	1.740	2.110	2.567	2.898	17
18	1.330	1.734	2.101	2.552	2.878	18
19	1.328	1.729	2.093	2.539	2.861	19
20	1.325	1.725	2.086	2.528	2.845	20
21	1.323	1.721	2.080	2.518	2.831	21
22	1.321	1.717	2.074	2.508	2.819	22
23	1.319	1.714	2.069	2.500	2.807	23
24	1.318	1.711	2.064	2.492	2.797	24
25	1.316	1.708	2.060	2.485	2.787	25
26	1.315	1.706	2.056	2.479	2.779	26
27	1.314	1.073	2.052	2.473	2.771	27
28	1.313	1.701	2.048	2.467	2.763	28
29	1.311	1.699	2.045	2.462	2.756	29
∞	1.282	1.645	1.960	2.326	2.576	∞

Glossary

cluster sampling – selection of clusters from the available clusters in the population

coefficient of determination – percentage of the total variation in the variable Y (i.e. the extent to which the Y_i 's are different) that is accounted for or explained by its linear relationship with the variable X

coefficient of variation – measure of relative dispersion that expresses the standard deviation as a percentage of the mean

confidence coefficient – probability that a confidence interval will contain the estimated parameter

confidence interval – interval constructed from the sample statistic where the value of the parameter is expected to lie

continuous random variable – random variable defined over a continuous sample space

continuous sample space – consists of uncountably infinite number of outcomes

critical region – set of values of the test statistic that results in the rejection of the null hypothesis; also called the *region of rejection*

critical value – particular point in the critical region that separates the rejection region with the acceptance region

discrete random variable – random variable defined over a discrete sample space

discrete sample space – consists of a finite number of elements or has an unending sequence with as many elements as there are counting numbers

empirical rule – specifies that in any normal distribution, approximately 68.26% of the values lie within one standard deviation away from the mean; 95.44% of the values lie within two standard deviations away from the mean; and 99.74% of the values lie within three standard deviations away from the mean

error of estimate – distance between an estimate and the parameter being estimated

estimator – rule or formula that tells us how to compute an estimate based on a sample

experiment – any procedure that can be repeated, theoretically, an infinite number of times and has a well-defined set of possible outcomes

level of significance (of a test) – probability of rejecting the null hypothesis when it is true

linear correlation coefficient – summary measure that describes the degree and direction of the linear relationship between two quantitative variables

null hypothesis – statement about the value of a population parameter formulated with the hope of it being rejected

one-tailed test – test of hypothesis where the alternative hypothesis is one-sided

outlier – observation that is unusually different (higher or lower) from the rest of the observations

parameter – any measurable characteristic of a population

point estimation – calculation of a single number based on information in the sample

population – totality of items, things, or people under consideration

population linear correlation coefficient – measures the dependence or association between two variables

probability – measure of the likelihood of occurrence of an event

probability histogram – graph of the probability mass function of a random variable

random variable – function that assigns a unique real number to each element in a sample space; real-valued function whose domain is the sample space

sample – subset of the population

sample outcome – each possible result of an experiment

sample space – set of all possible outcomes

sampling distribution – probability distribution of a statistic

sampling error – difference of the value of a sample statistic and the corresponding population parameter

sampling frame – complete list of all the members of the population

simple linear correlation – measure of the degree to which two variables are associated, or a measure of the intensity of the linear dependence of the two variables

simple random sampling – selection of a subset of a population where each element has an equal chance of being selected

standard error (of the statistic) – standard deviation of the sampling distribution

standard normal distribution – normal distribution whose mean is 0 and whose standard deviation is 1

statistic – any measurable characteristic of a sample

statistical hypothesis – statement about a population developed for the purpose of testing

stratified random sampling – selection of a simple random sample from each of a given number of subpopulations or *strata*

systematic random sampling – sampling method wherein a random starting point is selected, and then every k th member of the population is selected

t -distribution – models the sampling distribution of the mean when the sample size is small and the population standard deviation is unknown

test statistic – any function of the observed data whose numerical value dictates whether the null hypothesis is accepted or rejected

two-tailed test – test of hypothesis where the alternative hypothesis is two-sided

type I error – rejection of the null hypothesis when it is true

type II error – failing to reject the null hypothesis when it is false

Index

A

alternative hypothesis, 254, 256

B

Bayes's formula, 49

C

central tendency, 83

classical probability approach, 26

coefficient of determination, 311

coefficient of variation, 89

combination, 20

confidence coefficient, 212

confidence interval, 212

 for population mean, 212

 for population proportion, 220

critical region, 260

critical value, 260

D

De Morgan's laws, 6

E

empirical rule, 148

error of estimate, 208

estimator, 206

event, 4

experiment, 2

F

fair game, 94

L

level of significance, 227

linear correlation, 304

 coefficient, 305

linear regression, 321

M

mean, 83

method of least squares, 325

N

nonnegativity property, 68

normal curve, 148

norming property, 68

null hypothesis, 254, 256

O

one-tailed test, 259

outlier, 149

P

parameter, 137, 170

 location, 137

 scale, 137

Pearson's r , see *linear correlation coefficient*

permutation, 17

point estimate, 206

for population mean, 208

for population proportion, 209

point estimator, 206

Poisson experiment, 111

population, 170

probability, 77

density function, 77

histogram, 71

mass function, 68

R

random variable, 62

Bernoulli, 100

binomial, 102

continuous, 76

discrete, 67

geometric, 116

hypergeometric, 107

Poisson, 111

region of rejection, see *critical region*

relative frequency approach, 28

rule on probability, 31

complement, 31

conditional probability, 39

independence, 44

product (general case), 14

product (multiplication principle), 13

sum, 11

total probability, 46

union, 34

union for mutually exclusive events, 34

S

sample outcome, 3

sample space, 3

continuous, 4, 76

discrete, 4, 69

sampling distribution, 182

sampling error, 181

sampling frame, 171

sampling method, 170

cluster sampling, 175

simple random sampling, 171

stratified random sampling, 175

systematic random sampling, 174

set operations, 5

complement, 5

intersection, 5

union, 5

standard deviation, 87

standard error, 182

standard normal distribution, 140

statistic, 170

statistical hypothesis, 255

T

test statistic, 258

two-tailed test, 259

type I error, 287

type II error, 287

Bibliography

Books

Almeda, Josefina, Therese Capistrano, and Genelyn Ma. Ferry Sarte. *Elementary Statistics*. Quezon City: University of the Philippines Press, 2010.

Anderson, David, Dennis Sweeney, and Thomas Williams. *Modern Business Statistics with Microsoft Excel*. 2nd ed. OH: Thomson South-Western, 2006.

Arcilla, Rechel, et al. *Statistical Literacy for Lifelong Learning*. Quezon City: Abiva Publishing House, Inc., 2013.

Black, Ken. *Business Statistics for Contemporary Decision Making*. 4th ed. NJ: John Wiley & Sons Inc., 2004.

Freund, John E., et al. *Elementary Business Statistics: The Modern Approach*. 6th ed. Prentice Hall, 1993.

Gelman, Andrew, and Deborah Nolan. *Teaching Statistics: A Bag of Tricks*. Oxford University Press, 2002.

Ghahramani, Saeed. *Fundamentals of Probability with Stochastic Processes*. 3rd ed. Prentice Hall, 2005.

Grimaldi, Ralph P. *Discrete and Combinatorial Mathematics: An Applied Introduction*. 3rd ed. Addison-Wesley, 1994.

Larsen, Richard, and Morris Marx. *An Introduction to Mathematical Statistics and Its Applications*. 5th ed. Prentice Hall, 2012.

Levine, David M., et al. *Statistics for Managers Using Microsoft Excel*. 3rd ed. Prentice Hall, 2002.

Lind, Douglas, et al. *Basic Statistics for Business & Economics*. 5th ed. NY: McGraw-Hill Education, 2006.

Mendenhall, William, Robert Beaver, and Barbara Beaver. *Introduction to Probability and Statistics*. 14th ed. CA: Cengage Learning, 2012.

Miller, Irwin, and Marylees Miller. *John E. Freund's Mathematical Statistics with Applications*. 7th ed. Prentice Hall, 2004.

Montgomery, Douglas, and George Runger. *Applied Statistics and Probability for Engineers* 4th ed. NJ: John Wiley & Sons Inc., 2007.

Mood, Alexander, Franklin Graybill, and Duane Boes. *Introduction to the Theory of Statistics*. 3rd ed. McGraw-Hill Inc., 1974.

Olofsson, Peter. *Probability, Statistics, and Stochastic Processes*. NJ: John Wiley & Sons Inc., 2005.

Rosen, Kenneth H. *Discrete Mathematics and Its Applications*. 6th ed. Singapore: McGraw-Hill Education, 2007.

Ross, Sheldon. *A First Course in Probability*. 6th ed. Prentice-Hall, 2002.

Ross, Sheldon. *Introduction to Probability Models*. 1st ed. Academic Press, Inc., 1980.

Walpole, Ronald. *Introduction to Statistics*. 3rd ed. New York: Macmillan, 1982.

Walpole, Ronald, et al. *Probability & Statistics for Engineers & Scientists*. 9th ed. Prentice Hall, 2012.

Web sites

- http://highered.mheducation.com/sites/0073521477/student_view0/chapter6/multiple_choice_quiz.html
- http://highered.mheducation.com/sites/0078020522/student_view0/chapter7/multiple_choice_quiz.html
- http://mrscholkarsmathclass.weebly.com/uploads/1/3/2/1/13218566/expected_value_extra_examples.pdf
- <http://philfsis.psa.gov.ph/index.php/id/15/matrix/J30FSTII>
- [http://web.cocc.edu/srule/MTH105/homework/6expectation\(text\).pdf](http://web.cocc.edu/srule/MTH105/homework/6expectation(text).pdf)
- http://wps.prenhall.com/bp_levine_statsexcel_5/65/16644/4260977.cw/content/index.html
- http://wps.prenhall.com/bp_newbold_statbuse_6/53/13700/3507327.cw/-/3507329/index.html
- <http://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/>
- <http://www.eatthis.com/peanut-butter-ranked>
- <http://www.getthehealthystayhealthy.com/articles/how-your-medicines-are-put-test>
- <http://www.ics.uci.edu/~jutts/st13v-06/AnsQuiz6.PDF>
- <http://www.pcsogov.ph/games/lotto-642/>
- <http://www.pcsogov.ph/games/swertres-lotto-3d/>
- http://www.stat.ucla.edu/~nchristo/introeconometrics/introecon_normal_dist.pdf

- <http://www.statisticshowto.com/expected-value/>
- <http://www.subway.com/en-us/menunutrition/nutrition>
- <http://www.subway.com/en-us/menunutrition/nutrition>
- <http://www.themangofactory.com/growing-mangoes/organic-mangoes/growing-organic-mangoes-in-the-philippines/>
- <http://www1.pagasa.dost.gov.ph/index.php/tropical-cyclones/annual-tropical-cyclone-tracks>
- <https://betterexplained.com/articles/understanding-the-monty-hall-problem/>
- <https://idea.library.drexel.edu/islandora/object/idea%3A863>
- https://www.milo.com.ph/sites/milo_philippines2/files/milo_marathon_runners_handbook_2015.pdf
- <https://www.nissinfoods.com/Nutrition/Top%20Ramen%20NF%20Ingredients.pdf>
- <https://www.scientificamerican.com/article/bring-science-home-probability-birthday-paradox/>

(All the above Web sites were last accessed on 19 April 2017.)

About the Authors

Christian Paul O. Chan Shio graduated with a degree of BS Mathematics from Ateneo de Manila University (ADMU), where he also obtained his MS Mathematics degree. He graduated with a PhD in Mathematics from the University of Nice Sophia Antipolis in Nice, France. Currently, he is an assistant professor of the Mathematics Department of ADMU, where he has taught courses in basic and mathematical statistics, and other mathematical fields. He has also served as coach of the school's representatives in inter-university statistics competitions, and as one of the leaders of the country's team to the International Mathematical Olympiad since 2018.

Maria Angeli T. Reyes graduated from De La Salle University (DLSU) with a degree of BS Applied Mathematics as a Faculty Resource Program scholar. She started her teaching career at DLSU as an instructor of the Mathematics Department, and after a year, she pursued graduate studies under the auspices of the Australian Universities International Development Program. She graduated from the University of New South Wales in Sydney, Australia with a degree of Master of Statistics. For many years, she served as lecturer at the Mathematics Department of the Loyola Schools of Ateneo de Manila University and the Ateneo Professional Schools, particularly in the Information Technology Institute and the Graduate School of Business (RCBC Campus), where she taught courses in statistics, mathematics, and quantitative techniques. She also served as lecturer at the De La Salle Graduate School of Business–RCBC Campus, Makati and the Decision Science and Innovation Department of the Ramon V. del Rosario College of Business. She was also as a senior lecturer of the School of Statistics, University of the Philippines Diliman. She was an occasional resource speaker on statistics and probability and has conducted training courses in statistics as well. At present, she is an assistant professorial lecturer of the Mathematics Department of DLSU.

About the Book

Statistics & Probability provides students with the necessary theoretical background along with the practical applications of statistics and probability. This book covers relevant topics in the new K–12 Statistics and Probability curriculum for senior high school, as well as some additional topics for enrichment.

The following are some of the main features of the book:

- *Points to Remember*: Found mostly in chapters with more content and theory, these summarize the main points that the students are expected to know in the section.
- *Chapter Review*: This serves to review the concepts and formulas introduced in the chapter.
- *Chapter Exercises*: These include exercises of varying levels of difficulty, from the routine exercises to strengthen basic proficiencies, to the more challenging items to stimulate the problem solving skills of the faster learners.
- *Software Tutorial in MS Excel*: As most statistical computations involve the use of software, these include instructions on how to do specific tasks or analyses in MS Excel.
- *Chapter Performance Tasks*: These tasks allow the students to integrate everything that they have learned in the chapter in an applied problem. From data collection to data processing and analysis, these challenge the students to perform an actual statistical study.



Excello™ is a continuously expanding aggregation of high-quality learning assessments designed to help students master the Philippine curriculum's key skills and provide teachers with easy-to-use test creation tools.



Published by:
C & E Publishing, Inc.
839 EDSA, South Triangle
Quezon City, Philippines
Tel. No.: (02) 8929-5088
E-mail: info@cebookshop.com
Website: www.cepublishing.com



MEMBER

