CONTENTS

CHAPTER NO.	CHAPTER NAME	PAGE NO.
1	Introduction to Statistics	1
2	Charts and Graphs	22
3	Descriptive Statistics	50
4	Probability	97
5	Probability Distributions	150
6	Sampling and Sampling Distributions	229
7	Correlation and Simple Regression Analysis	268
8	Multiple Regression Analysis	303
9	Time-Series Forecasting	332
10	Decision Analysis	383
11	Case Studies	418
12	Appendix A (Online)	447

DECISION SCIENCE

CURRICULUM

Decision Science

Introduction to Statistics: Statistics in Business, Basic Statistical Concepts, Variables and Data, Data Measurement.

Data Visualization: Frequency Distributions, Quantitative Data Graphs, Qualitative Data Graphs, Charts and Graphs for Two Variables.

Descriptive Statistics: Measures of Central Tendency: Ungrouped Data, Measures of Variability: Ungrouped Data, Measures of Central Tendency and Variability: Grouped Data, Measures of Shape.

Basics of Probability: Introduction to Probability, Methods of Assigning Probabilities, Structure of Probability, Marginal probability, Union, Joint Probabilities, Addition Laws, Multiplication Laws, Conditional Probability, Bayes' Rule.

Probability Distributions: Discrete Versus Continuous Distribution, Binomial Distribution, Poisson Distribution, Hypergeometric Distribution, The Uniform Distribution, Normal Distribution, using Normal Curve to Approximate Binomial Distribution, Exponential Distribution.

Sampling Techniques and Sampling Distributions: Introduction to sampling, Reasons for Sampling, Random Versus Non-random Sampling, Sampling Error, Non-sampling Errors, Sampling Distribution of "sample mean" and "Sample proportion".

Correlation and Simple Regression: Correlation analysis, Introduction to Simple Regression Analysis, Equation of the Regression Line, Coefficient of Determination.

Multiple Regression Analysis: Multiple Regression Model with Two Independent Variables, Determining the Multiple Regression Equation, Coefficient of Multiple Determination, Adjusted R^2, Interpreting Multiple Regression Computer Output.

Forecasting Techniques: Introduction to Forecasting, Time-Series Components, Measurement of Forecasting Error, Smoothing Techniques, Trend Analysis, Autocorrelation and Autoregression.

Decision Analysis: Decision making under certainty, Decision making under uncertainty, Decision making under risk.

1

CHAPTER

INTRODUCTION TO STATISTICS

CONTENTS

1.1 Introduction
1.2 Basic Statistical Concepts
Self Assessment Questions
Activity
1.3 Data Measurement
1.3.1 Nominal Level
1.3.2 Ordinal Level
1.3.3 Interval Level
1.3.4 Ratio Level
1.3.5 Comparison of the Four Levels of Data
Self Assessment Questions
Activity
1.4 Summary
1.5 Descriptive Questions
1.6 Solutions for Descriptive Questions
1.7 Answers and Hints

LEARNING OBJECTIVES

The primary objective of Chapter 1 is to introduce you to the world of statistics, thereby enabling you to:

- List quantitative and graphical examples of statistics within a business context.
- Define important statistical terms, including population, sample, and parameter, as they relate to descriptive and inferential statistics.
- Explain the difference between variables, measurement, and data.
- Compare the four different levels of data: nominal, ordinal, interval, and ratio.

1.1 INTRODUCTION

Every minute of the working day, decisions are made by businesses around the world that determine whether companies will be profitable and growing or whether they will stagnate and die. Most of these decisions are made with the assistance of information gathered about the marketplace, the economic and financial environment, the workforce, the competition, and other factors. Such information usually comes in the form of data or is accompanied by data. Business statistics provides the tool through which such data are collected, analyzed, summarized, and presented to facilitate the decision-making process, and business statistics plays an important role in the ongoing saga of decision making within the dynamic world of business.

Virtually every area of business uses statistics in decision making. Here are some recent examples:

- According to a national survey of independent business owners conducted by the Institute for Local Self-Reliance in partnership with the Advocates for Independent Business coalition, when asked "Which two public policy changes would most help your business?" (retailers only), 40% said "Pass the Marketplace Fairness Act" and 38% said "Cap Credit Card Swipe Fees".
- A survey of 1465 workers by Hotjobs reports that 55% of workers believe that the quality of their work is perceived the same when they work remotely as when they are physically in the office.
- A survey of 477 executives by the Association of Executive Search Consultants determined that 48% of men and 67% of women say they are more likely to negotiate for less business travel compared with five years ago.
- A global Family Business Survey of 2,378 respondents sponsored by PwC reported that 65% of family businesses reported growth in the last twelve months and 49% of respondents are apprehensive about their ability to recruit skilled staff in the next twelve months.
- A Deloitte Retail "Green" survey of 1080 adults revealed that 54% agreed that plastic, non-compostable shopping bags should be banned.
- A study of consumer electronics spending by a 2,500 member on-line panel of the NPD group showed that consumers expect to spend $555, on average, per person on new consumer electronics devices this year.

You can see from these few examples that there is a wide variety of uses and applications of statistics in business. Note that in most of these examples, business researchers have conducted a study and provided us rich and interesting information.

In this text we will examine several types of graphs for depicting data as we study ways to arrange or structure data into forms that are both meaningful and useful to decision makers. We will learn about techniques for sampling from a population that allow studies of the business world to be conducted more inexpensively and in a more timely manner. We will explore various ways to forecast future values and examine techniques for predicting trends. These and many other exciting statistics and statistical techniques await us on this journey through business statistics. Let us begin.

1.2 BASIC STATISTICAL CONCEPTS

Business statistics, like many areas of study, has its own language. It is important to begin our study with an introduction of some basic concepts in order to understand and communicate about the subject. We begin with a discussion of the word *statistics*. The word *statistics* has many different meanings in our culture. *Webster's Third New International Dictionary* gives a comprehensive definition of **statistics** as *a science dealing with the collection, analysis, interpretation, and presentation of numerical data*. Viewed from this perspective, statistics includes all the topics presented in this text. **Figure 1.1** graphically displays the key elements of statistics.

The study of statistics can be organized in a variety of ways. One of the main ways is to subdivide statistics into two branches: descriptive statistics and inferential statistics. To understand the difference between descriptive and inferential statistics, definitions of *population* and *sample* are helpful. *Webster's Third New International Dictionary* defines **population** as *a collection of persons, objects, or items of interest*. The population can be a widely defined category, such as "all automobiles," or it can be narrowly defined, such as "all Ford Mustang cars produced from 2014 to 2016." A population can

Figure 1.1: The Key Elements of Statistics

be a group of people, such as "all workers presently employed by Microsoft," or it can be a set of objects, such as "all dishwashers produced on February 3, 2016, by the General Electric Company at the Louisville plant." The researcher defines the population to be whatever he or she is studying. When researchers *gather data from the whole population for a given measurement of interest*, they call it a **census**. Most people are familiar with the "Census of India". Every 10 years, the government attempts to measure all persons living in this country.

A **sample** is *a portion of the whole* and, if properly taken, is representative of the whole. For various reasons, researchers often prefer to work with a sample of the population instead of the entire population. For example, in conducting quality-control experiments to determine the average life of lightbulbs, a lightbulb manufacturer might randomly sample only 75 lightbulbs during a production run. Because of time and money limitations, a human resources manager might take a random sample of 40 employees instead of using a census to measure company morale.

If a business analyst is *using data gathered on a group to describe or reach conclusions about that same group*, the statistics are called **descriptive statistics**. For example, if an instructor produces statistics to summarize a class's examination effort and uses those statistics to reach conclusions about that class only, the statistics are descriptive.

Many of the statistical data generated by businesses are descriptive. They might include number of employees on vacation during June, average salary at the Mumbai office, corporate sales for 2016, average managerial satisfaction score on a company-wide census of employee attitudes, and average return on investment for the Tata Motors for the years 1996 through 2016.

Another type of statistics is called **inferential statistics**. If a researcher *gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken*, the statistics are inferential statistics. The data gathered from the sample are used to infer something about a larger group. Inferential statistics are sometimes referred to as *inductive statistics*. The use and importance of inferential statistics continue to grow.

One application of inferential statistics is in pharmaceutical research. Some new drugs are expensive to produce, and therefore tests must be limited to small samples of patients. Utilizing inferential statistics, researchers can design experiments with small randomly selected samples of patients and attempt to reach conclusions and make inferences about the population.

Market researchers use inferential statistics to study the impact of advertising on various market segments. Suppose a soft drink company creates an advertisement depicting a dispensing machine that talks to the buyer, and market researchers want to measure the impact of the new advertisement on various age groups. The researcher could stratify the population into age categories ranging from young to old, randomly sample each stratum, and use inferential statistics to determine the effectiveness of the advertisement for the various age groups in the population. The advantage of using inferential statistics is that they enable the researcher to study effectively a wide range of phenomena without having to conduct

A *descriptive measure of the population* is called a **parameter**. Parameters are usually denoted by Greek letters. Examples of parameters are population mean (μ), population variance (σ^2), and population standard deviation (σ). A *descriptive measure of a sample* is called a **statistic**. Statistics are usually denoted by Roman letters. Examples of statistics are sample mean ($\bar{x}$), sample variance (s^2), and sample standard deviation (s).

Differentiation between the terms *parameter* and *statistic* is important only in the use of inferential statistics. A business researcher often wants to estimate the value of a parameter or conduct tests about the parameter. However, the calculation of parameters is usually either impossible or infeasible because of the amount of time and money required to take a census. In such cases, the business researcher can take a random sample of the population, calculate a statistic on the sample, and infer by estimation the value of the parameter. The basis for inferential statistics, then, is the ability to make decisions about parameters without having to complete a census of the population.

For example, a manufacturer of washing machines would probably want to determine the average number of loads that a new machine can wash before it needs repairs. The parameter is the population mean or average number of washes per machine before repair. A company researcher takes a sample of machines, computes the number of washes before repair for each machine, averages the numbers, and estimates the population value or parameter by using the statistic, which in this case is the sample average. **Figure 1.2** demonstrates the inferential process.

Inferences about parameters are made under uncertainty. Unless parameters are computed directly from the population, the statistician never knows with certainty whether the estimates or inferences made from samples are true. In an effort to estimate the level of confidence in the result of the process, statisticians use probability statements. For this and other reasons, part of this text is devoted to probability (Chapter 4).

NOTES

Business statistics is about measuring phenomena in the business world and organizing, analyzing, and presenting the resulting numerical information in such a way such that better, more informed business decisions can be made. Most business statistics studies contain variables, measurements, and data.

In business statistics, a **variable** is *a characteristic of any entity being studied that is capable of taking on different values*. Some examples of variables in business might include return on investment, advertising dollars, labor productivity, stock price, historical cost, total sales, market share, age of worker, earnings per share, miles driven to work, time spent in store shopping, and many, many others. In business statistics studies, most variables produce a measurement that can be used for analysis. A **measurement** is *when a standard process is used to assign numbers to particular attributes or characteristics of a variable.* Many measurements are obvious, such as time spent in a store shopping by a customer, age of the worker, or the number of miles driven to work. However, some measurements, such as labor productivity, customer satisfaction, and return on investment, have to be defined by the business researcher or by experts within the field. Once such measurements are recorded and stored, they can be denoted as "data." It can be said that **data** are *recorded measurements*. The processes of measuring and data gathering are basic to all that we do in business statistics. It is data that are analyzed by a business statistician in order to learn more about the variables being studied. Sometimes, sets of data are organized into databases as a way to store data or as a means for more conveniently analyzing data or comparing variables. Valid data are the lifeblood of business statistics, and it is important that the business researcher give thoughtful attention to the creation of meaningful, valid data before embarking on analysis and reaching conclusions.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. _____ is a collection of persons, objects, or items of interest.
2. When researchers gather data from the whole population for a given measurement of interest, then it's called a _____.
3. Inferential statistics are sometimes referred to as_____.

State whether the following statements are true/false:

4. Data interpretation is a key element of statistics.
5. Researchers often prefer to work with the population instead of the entire sample.

ACTIVITY

Study the "International Labor Database" containing the civilian unemployment rates in percent from seven countries presented yearly over a 40-year period. The countries are the United States, Canada, Australia, Japan, France, Germany, and Italy. Prepare a comparative report based on your study.

NOTES

1.3 DATA MEASUREMENT

Millions of numerical data are gathered in businesses every day, representing myriad items. For example, numbers represent costs of items produced, geographical locations of retail outlets, weights of shipments, and rankings of subordinates at yearly reviews. All such data should not be analyzed the same way statistically because the entities represented by the numbers are different. For this reason, the business researcher needs to know the *level of data measurement* represented by the numbers being analyzed.

The disparate use of numbers can be illustrated by the numbers 40 and 80, which could represent the weights of two objects being shipped, the ratings received on a consumer test by two different products, or football jersey numbers of a fullback and a wide receiver. Although 80 pounds is twice as much as 40 pounds, the wide receiver is probably not twice as big as the fullback! Averaging the two weights seems reasonable, but averaging the football jersey numbers makes no sense. The appropriateness of the data analysis depends on the level of measurement of the data gathered. The phenomenon represented by the numbers determines the level of data measurement. Four common levels of data measurement follow.

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Nominal is the lowest level of data measurement followed by ordinal, interval, and ratio. Ratio is the highest level of data measurement, as shown in **Figure 1.3**.

Figure 1.3: Hierarchy of Levels of Data

1.3.1 NOMINAL LEVEL

The *lowest level of data measurement* is the **nominal level**. Numbers representing nominal-level data (the word *level* often is omitted) can be *used only to classify or categorize*. Employee identification numbers are an example of nominal data. The numbers are used only to differentiate employees and not to make a value statement about them. Many demographic questions in surveys result in data that are nominal because the questions are used for classification only. The following is an example of such a question that would result in nominal data:

Which of the following employment classifications best describes your area of work?

1. Educator
2. Construction worker
3. Manufacturing worker
4. Lawyer

5. Doctor
6. Other

Suppose that, for computing purposes, an educator is assigned a 1, a construction worker is assigned a 2, a manufacturing worker is assigned a 3, and so on. These numbers should be used only to classify respondents. The number 1 does not denote the top classification. It is used only to differentiate an educator (1) from a lawyer (4).

Some other types of variables that often produce nominal-level data are sex, religion, ethnicity, geographic location, and place of birth. Social Security numbers, telephone numbers, employee ID numbers, and ZIP code numbers are further examples of nominal data. Statistical techniques that are appropriate for analyzing nominal data are limited. However, some of the more widely used statistics, such as the chi-square statistic, can be applied to nominal data, often producing useful information.

1.3.2 ORDINAL LEVEL

Ordinal-level data measurement is higher than the nominal level. In addition to the nominal-level capabilities, ordinal-level measurement can be used to rank or order people or objects. For example, using ordinal data, a supervisor can evaluate three employees by ranking their productivity with the numbers 1 through 3. The supervisor could identify one employee as the most productive, one as the least productive, and one as somewhere between by using ordinal data. However, the supervisor could not use ordinal data to establish that the intervals between the employees ranked 1 and 2 and between the employees ranked 2 and 3 are equal; that is, she could not say that the differences in the amount of productivity between workers ranked 1, 2, and 3 are necessarily the same. With ordinal data, the distances or spacing represented by consecutive numbers are not always equal.

Some questionnaire Likert-type scales are considered by many researchers to be ordinal in level. The following is an example of one such scale:

This computer tutorial is				
___ not helpful	___ somewhat helpful	___ moderately helpful	___ very helpful	___ extremely helpful
1	2	3	4	5

When this survey question is coded for the computer, only the numbers 1 through 5 will remain, not the adjectives. Virtually everyone would agree that a 5 is higher than a 4 on this scale and that ranking responses is possible. However, most respondents would not consider the differences between not helpful, somewhat helpful, moderately helpful, very helpful, and extremely helpful to be equal.

Mutual funds as investments are sometimes rated in terms of risk by using measures of default risk, currency risk, and interest rate risk. These three measures are applied to investments by rating them as having high, medium, and low risk. Suppose high risk is assigned a 3, medium risk a 2, and low risk a 1. If a fund is awarded a 3 rather than a 2, it carries more risk, and so on. However, the differences in risk between categories 1, 2, and 3 are not necessarily equal. Thus, these measurements of risk are only ordinal-level measurements. Another example of the use of ordinal numbers in business is the ranking of the top 50 most admired companies in *Fortune* magazine. The numbers ranking the companies are only ordinal in measurement. Certain statistical techniques are specifically suited to ordinal data, but many other techniques are not appropriate for use on ordinal data. For example, it does not make sense to say that the average of "moderately helpful" and "very helpful" is "moderately helpful and a half."

Because nominal and ordinal data are often derived from imprecise measurements such as demographic questions, the categorization of people or objects, or the ranking of items, *nominal and ordinal data* are **nonmetric data** and are sometimes referred to as *qualitative data.*

1.3.3 INTERVAL LEVEL

Interval-level data measurement is the *next to the highest level of data in which the distances between consecutive numbers have meaning and the data are always numerical.* The distances represented by the differences between consecutive numbers are equal; that is, interval data have equal intervals. An example of interval measurement is Fahrenheit temperature. With Fahrenheit temperature numbers, the temperatures can be ranked, and the amounts of heat between consecutive readings, such as 20°, 21°, and 22°, are the same.

In addition, with interval-level data, the zero point is a matter of convention or convenience and not a natural or fixed zero point. Zero is just another point on the scale and does not mean the absence of the phenomenon. For example, zero degrees Fahrenheit is not the lowest possible temperature. Some other examples of interval-level data are the percentage change in employment, the percentage return on a stock, and the dollar change in stock price.

1.3.4 RATIO LEVEL

Ratio-level data measurement is *the highest level of data measurement.* Ratio data *have the same properties as interval data,* but ratio data have an *absolute zero,* and *the ratio of two numbers is meaningful.* The notion of absolute zero means that zero is fixed, and *the zero value in the data represents the absence of the characteristic being studied.* The value of zero cannot be arbitrarily assigned because it represents a fixed point. This definition enables the statistician to create *ratios* with the data.

Examples of ratio data are height, weight, time, volume, and Kelvin temperature. With ratio data, a researcher can state that 180 pounds of weight is twice as much as 90 pounds or, in other words, make a ratio of 180:90. Many of the data measured by valves or gauges in industry are ratio data.

Other examples in the business world that are ratio level in measurement are production cycle time, work measurement time, passenger miles, number of trucks sold, complaints per 10,000 fliers, and number of employees.

Because interval- and ratio-level data are usually gathered by precise instruments often used in production and engineering processes, in national standardized testing, or in standardized accounting procedures, they are called **metric data** and are sometimes referred to as *quantitative* data.

1.3.5 COMPARISON OF THE FOUR LEVELS OF DATA

Figure 1.4 shows the relationships of the usage potential among the four levels of data measurement. The concentric squares denote that each higher level of data can be analyzed by any of the techniques used on lower levels of data but, in addition, can be used in other statistical techniques. Therefore, ratio data can be analyzed by any statistical technique applicable to the other three levels of data plus some others.

Nominal data are the most limited data in terms of the types of statistical analysis that can be used with them. Ordinal data allow the researcher to perform any analysis that can be done with nominal data and some additional analyses. With ratio data, a statistician can make ratio comparisons and appropriately do any analysis that can be performed on nominal, ordinal, or interval data. Some statistical techniques require ratio data and cannot be used to analyze other levels of data.

Statistical techniques can be separated into two categories: parametric statistics and nonparametric statistics. **Parametric statistics** require that data be interval or ratio. If the data are nominal or ordinal, **nonparametric statistics** must be used. Nonparametric statistics can also be used to analyze interval or ratio data.

Figure 1.5 contains a summary of metric data and nonmetric data.

Figure 1.4: Usage Potential of Various Levels of Data

Figure 1.5: Metric vs. Nonmetric Data

DEMONSTRATION PROBLEM 1.1

Because of increased competition for patients among providers and the need to determine how providers can better serve their clientele, hospital administrators sometimes administer a quality satisfaction survey to their patients after the patient is released. The following types of questions are sometimes asked on such a survey. These questions will result in what level of data measurement?

1. How long ago were you released from the hospital?
2. Which type of unit were you in for most of your stay?

   __Coronary care

   __Maternity care

   __Medical unit

   __Pediatric/children's unit

   __Surgical unit

3. In choosing a hospital, how important was the hospital's location?

   (circle one)

Very Important	Somewhat Important	Not Very Important	Not at All Important

4. What was your body temperature when you were admitted to the hospital?
5. Rate the skill of your doctor:

   __Excellent __Very Good __Good __Fair __Poor

Solution: Question 1 is a time measurement with an absolute zero and is therefore ratio-level measurement. A person who has been out of the hospital for two weeks has been out twice as long as someone who has been out of the hospital for one week.

Question 2 yields nominal data because the patient is asked only to categorize the type of unit he or she was in. This question does not require a hierarchy or ranking of the type of unit. Questions 3 and 5 are likely to result in ordinal-level data. Suppose a number is assigned the descriptors in these two questions. For question 3, "very important" might be assigned a 4, "somewhat important" a 3, "not very important" a 2, and "not at all important" a 1. Certainly, the higher the number, the more important is the hospital's location. Thus, these responses can be ranked by selection. However, the increases in importance from 1 to 2 to 3 to 4 are not necessarily equal. This same logic applies to the numeric values assigned in question 5. In question 4, body temperature, if measured on a Fahrenheit or Celsius scale, is interval in measurement.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

6. _____ is the lowest level of data measurement.
7. Fahrenheit scale is example of _____.
8. _____ is the highest level of data measurement.

State whether the following statements are true/false:

9. Ordinal-level data measurement is higher than the nominal level.
10. Ordinal data are nonmetric data and are sometimes referred to as quantitative data.
11. Interval-level data many not be always numerical.

s

ACTIVITY

From the "Ministry of Statistics and programme Implementation", Government of India website download the "Energy Statistics" report (use http://www.mospi.gov.in/sites/default/files/publication_reports/Energy_Statistics_2018.pdf to download the report). In the report, identify one variable each for the four common levels of data measurement.

1.4 SUMMARY

- Statistics is an important decision-making tool in business and is used in virtually every area of business. In this course, the word *statistics* is defined as the science of gathering, analyzing, interpreting, and presenting numerical data.
- The study of statistics can be subdivided into two main areas: *descriptive statistics* and *inferential statistics*. Descriptive statistics result from gathering data from a body, group, or population and reaching conclusions only about that group. Inferential statistics are generated by gathering sample data from a group, body, or population and reaching conclusions about the larger group from which the sample was drawn.
- Most business statistics studies contain variables, measurements and data. A *variable* is a characteristic of any entity being studied that is capable of taking on different values. Examples of variables might include monthly household food spending, time between arrivals at a restaurant and patient satisfaction rating. A *measurement* is when a standard process is used to assign numbers to particular attributes or characteristics of a variable. Measurements on monthly household food spending might be taken in dollars, time between arrivals might be measured in minutes and patient satisfaction might be measured using a 5-point scale. *Data* are recorded measurements. It is data that are analyzed by business statisticians in order to learn more about the variables being studied.
- The appropriate type of statistical analysis depends on the level of data measurement, which can be (1) *nominal*, (2) *ordinal*, (3) *interval*, or (4) *ratio*. Nominal is the lowest level, representing classification only of such data as geographic location, sex, or Social Security number. The next level is ordinal, which provides rank ordering measurements in which the intervals between consecutive numbers do not necessarily represent equal distances. Interval is the next to highest level of data measurement in which the distances represented by consecutive numbers are equal. The highest level of data measurement is ratio, which has all the qualities of interval measurement, but ratio data contain an absolute zero and ratios between numbers are meaningful. Interval and ratio data sometimes are called *metric* or *quantitative* data. Nominal and ordinal data sometimes are called *nonmetric* or *qualitative* data.
- Two major types of inferential statistics are (1) *parametric statistics* and (2) *nonparametric statistics*. Use of parametric statistics requires interval or ratio data and certain assumptions about the distribution of the data. The techniques presented in this text are largely parametric. If data are only nominal or ordinal in level

KEY WORDS

1. **Census:** The process of gathering data from the whole population for a given measurement of interest is a Census.
2. **Variable:** Variable is a characteristic of any entity being studied that is capable of taking on different values.
3. **Population:** A population is a collection of persons, objects, or items of interest.
4. **Parameter:** A descriptive measure of the population is called a parameter.
5. **Sample:** A sample is a portion of the whole and, if properly taken, is representative of the whole.
6. **Statistic:** A descriptive measure of a sample is called a statistic.
7. **Measurement:** A measurement is when a standard process is used to assign numbers to particular attributes or characteristics of a variable.
8. **Data:** Data are recorded measurements.
9. **Descriptive statistics:** If a business analyst uses data gathered on a group to describe or reach conclusions about that same group, the statistics are called descriptive statistics.
10. **Inferential statistics:** If a researcher gathers data from a sample and uses the statistics generated to reach conclusions about the population from which the sample was taken, the statistics are inferential statistics.
11. **Nominal-level data:** The lowest level of data measurement is the nominal level. Numbers representing nominal-level data can be used only to classify or categorize.
12. **Ordinal-level data:** This measurement is higher than the nominal level. In addition to the nominal-level capabilities, ordinal-level measurement can be used to rank or order people or objects.
13. **Interval-level data:** Is a level of data in which the distances between consecutive numbers have meaning and the data are always numerical.
14. **Ratio-level data:** This measurement is the highest level of data measurement. Ratio data have the same properties as interval data, but ratio data have an absolute zero, and the ratio of two numbers is meaningful.

1.5 DESCRIPTIVE QUESTIONS

1.1. Give a specific example of data that might be gathered from each of the following business disciplines: accounting, finance, human resources, marketing, information systems, production, and management. An example in the marketing area might be "number of sales per month by each salesperson."

1.2. State examples of data that can be gathered for decision making purposes from each of the following industries: manufacturing, insurance, travel, retailing, communications, computing, agriculture, banking, and healthcare. An example in the travel industry might be the cost of business travel per day in various European cities.

1.3. Give an example of *descriptive* statistics in the recorded music industry. Give an example of how *inferential* statistics could be used in the recorded music industry. Compare the two examples. What makes them different?

1.4. Suppose you are an operations manager for a plant that manufactures batteries. Give an example of how you could use *descriptive* statistics to make better managerial decisions. Give an example of how you could use *inferential* statistics to make better managerial decisions.

1.5. There are many types of information that might help the manager of a large department store run the business more efficiently and better understand how to improve sales. Think about this in such areas as sales, customers, human resources, inventory, suppliers, etc., and list five variables that might produce information that could aid the manager in his or her job. Write a sentence or two describing each variable, and briefly discuss some numerical observations that might be generated for each variable.

1.6. Suppose you are the owner of a medium-sized restaurant in a small city. What are some variables associated with different aspects of the business that might be helpful to you in making business decisions about the restaurant? Name four of these variables, and for each variable, briefly describe a numerical observation that might be the result of measuring the variable.

1.7. Classify each of the following as nominal, ordinal, interval, or ratio data.

(a) The time required to produce each tire on an assembly line

(b) The number of quarts of milk a family drinks in a month

(c) The ranking of four machines in your plant after they have been designated as excellent, good, satisfactory, and poor

(d) The telephone area code of clients in the United States

(e) The age of each of your employees

(f) The dollar sales at the local pizza shop each month

(g) An employee's identification number

(h) The response time of an emergency unit

1.8. The Rathburn Manufacturing Company makes electric wiring, which it sells to contractors in the construction industry. Approximately 900 electric contractors purchase wire from Rathburn annually. Rathburn's director of marketing wants to determine electric contractors' satisfaction with Rathburn's wire. He developed a questionnaire that yields a satisfaction score between 10 and 50 for participant responses. A random sample of 35 of the 900 contractors is asked to complete a satisfaction survey. The satisfaction scores for the 35 participants are averaged to produce a mean satisfaction score.

(a) What is the population for this study?

(b) What is the sample for this study?

(c) What is the statistic for this study?

(d) What would be a parameter for this study?

1.6 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

1.1. Examples of data in functional areas:

accounting – cost of goods, salary expense, depreciation, utility costs, taxes, equipment inventory, etc.

finance – World bank bond rates, number of failed savings and loans, measured risk of common stocks, stock dividends, foreign exchange rate, liquidity rates for a single-family, etc.

human resources – salaries, size of engineering staff, years experience, age of employees, years of education, etc.

marketing – number of units sold, dollar sales volume, forecast sales, size of sales force, market share, measurement of consumer motivation, measurement of consumer frustration, measurement of brand preference, attitude measurement, measurement of consumer risk, etc.

information systems – CPU time, size of memory, number of work stations, storage capacity, percent of professionals who are connected to a computer network, dollar assets of company computing, number of "hits" on the Internet, time spent on the Internet per day, percentage of people who use the Internet, retail dollars spent in e-commerce, etc.

production – number of production runs per day, weight of a product; assembly time, number of defects per run, temperature in the plant, amount of inventory, turnaround time, etc.

management – measurement of union participation, measurement of employer support, measurement of tendency to control, number of subordinates reporting to a manager, measurement of leadership style, etc.

1.2. Examples of data in business industries:

manufacturing – size of punched hole, number of rejects, amount of inventory, amount of production, number of production workers, etc.

insurance – number of claims per month, average amount of life insurance per family head, life expectancy, cost of repairs for major auto collision, average medical costs incurred for a single female over 45 years of age, etc.

travel – cost of airfare, number of miles traveled for ground transported vacations, number of nights away from home, size of traveling party, amount spent per day on besides lodging, etc.

retailing – inventory turnover ratio, sales volume, size of sales force, number of competitors within 2 miles of retail outlet, area of store, number of sales people, etc.

communications – cost per minute, number of phones per office, miles of cable per customer headquarters, minutes per day of long distance usage, number of operators, time between calls, etc.

computing – age of company hardware, cost of software, number of CAD/CAM stations, age of computer operators, measure to evaluate competing software packages, size of data base, etc.

agriculture – number of farms per county, farm income, number of acres of corn per farm, wholesale price of a gallon of milk, number of livestock, grain storage capacity, etc.

banking – size of deposit, number of failed banks, amount loaned to foreign banks, number of tellers per drive-in facility, average amount of withdrawal from automatic teller machine, federal reserve discount rate, etc.

healthcare – number of patients per physician per day, average cost of hospital stay, average daily census of hospital, time spent waiting to see a physician, patient satisfaction, number of blood tests done per week.

1.3. **Descriptive statistics** in recorded music industry –

1. RCA total sales of compact discs this week, number of artists under contract to a company at a given time.
2. total dollars spent on advertising last month to promote an album.
3. number of units produced in a day.
4. number of retail outlets selling the company's products.

Inferential statistics in recorded music industry –

1. Measure the amount spent per month on recorded music for a few consumers then use that figure to infer the amount for the population.
2. Determination of market share for rap music by randomly selecting a sample of 500 purchasers of recorded music.
3. Determination of top ten single records by sampling the number of requests at a few radio stations.
4. Estimation of the average length of a single recording by taking a sample of records and measuring them.

The difference between descriptive and inferential statistics lies mainly in the usage of the data. These descriptive examples all gather data from every item in the population about which the description is being made. For example, RCA measures the sales on **all** its compact discs for a week and reports the total.

In each of the inferential statistics examples, a **sample** of the population is taken and the population value is estimated or inferred from the sample. For example, it may be practically impossible to determine the proportion of buyers who prefer rap music. However, a random sample of buyers can be contacted and interviewed for music preference. The results can be inferred to population market share.

1.4. **Descriptive statistics** in manufacturing batteries to make better decisions –

1. total number of worker hours per plant per week – help management understand labor costs, work allocation, productivity, etc.
2. company sales volume of batteries in a year – help management decide if the product is profitable, how much to advertise in coming year, compare to costs to determine profitability.
3. total amount of sulfuric acid purchased per month for use in battery production – can be used by management to study wasted inventory, scrap, etc.

Inferential statistics in manufacturing batteries to make decisions –

1. Take a sample of batteries and test them to determine the average shelf life – use the sample average to reach conclusions about all batteries of this type. Management can then make labeling and advertising claims. They can compare these figures to the shelf-life of competing batteries.
2. Take a sample of battery consumers and determine how many batteries they purchase per year. Infer to the entire population – management can use this information to estimate market potential and penetration.
3. Interview a random sample of production workers to determine attitude towards company management – management can use this survey results to ascertain employee morale and to direct efforts towards creating a more positive working environment which, hopefully, results in greater productivity.

1.5. 1. **Size of sale ($) per customer in men's formal wear.** Either by taking a sample or using a census, management could compute the average sale in men's formal wear of a weekly period and compare the number to the same average taken a year ago or a month ago to determine if more is being sold per customer. Other variables might include number of sales per hour, number of people entering the department per day, number of dress shirts sold per day, etc.

2. **Number of employees working per day.** This variable could indicate the day of the week (certain days have more or less sales), sales activity (how sales are doing overall), or even health of associates. Other variables might include percent of employees absent due to illness, average number of hours worked per week per employee, number of open positions, etc.

3. **Inventory turnover rate.** How fast are items in the store selling? Other variables migh include reorder rate, percent of storage space utilized, number of stockouts per week, etc.

4. **Number of customers that enter the store per hour.** This figure will vary by day, time of day, and season. Compare figures on this variable from period to period can give some indication of sales trends which can help drive human resource planning, etc. Other variables might include amount of time spent per customer in the store per visit, distance that customers travel to shop in the store, number of referrals that customers make to other people annually, etc.

5. **Percentage of people paying with cash.** Percentage of people using credit cards. These can be used to expedite pay systems, investigate employee theft, calculate surcharges associated with credit cards, etc. Other variables might include average time per checkout, average wait time in pay line, etc.

1.6. 1. **Size of bill or tab.** This variable is the total amount in dollars spent by a patron per visit to the restaurant. The bill or tab could be for an individual or a group and would include both food and beverages if they are all included in the bill. Of course, the measurement would be in dollars. This information could be very useful for the manager or owner to know the average size of a bill both in projecting out total revenues over a period or as a baseline before a marketing effort to increase sales.

2. **Percentage of capacity filled.** This variable could be measured at various intervals, times, and days of the week. The measurement would be calculated by taking the number of patrons in the restaurant at any one time divided by the total number of seats in the restaurant (capacity). From this, management could make staffing decisions for various times and days of the week. In addition, management could make decisions about when to expand, how much to advertise, and/or when to run specials.

3. **Length of stay.** The measurement is how many minutes people are actually in the restaurant from the time they are assigned a table until they are leaving? From this, management could determine customer turnover rates which have capacity implications. That is, how many times in a day is an average table "turned over". If people stay longer, do they spend more?

4. **Number of arrivals per 5-minute intervals.** The measurement is how many customers arrive at the front door to be greeted by the maître 'd in any given five-minute period. This figure will likely vary by day of the week, season of the year, and time of day. Management can use this information for staffing decisions and planning.

1.7. (a) ratio

(b) ratio

(c) ordinal

(d) nominal

(e) ratio

(f) ratio

(g) nominal

(h) ratio

1.8. (a) The population for this study is the 900 electric contractors who purchased Rathburn wire.

(b) The sample is the randomly chosen group of thirty-five contractors.

(c) The statistic is the average satisfaction score for the sample of thirty-five contractors.

(d) The parameter is the average satisfaction score for all 900 electric contractors in the population.

1.7 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
1.2 Basic Statistical Concepts	1.	Population
	2.	Census
	3.	Inductive statistics
	4.	True
	5.	False
1.3 Data Measurement	6.	Nominal level
	7.	Interval-level data
	8.	Ratio-level data
	9.	True
	10.	False
	11.	False

2

CHAPTER

CHARTS AND GRAPHS

CONTENTS

2.1 Introduction
2.2 Frequency Distributions
2.2.1 Class Midpoint
2.2.2 Relative Frequency
2.2.3 Cumulative Frequency
Self Assessment Questions
Activity
2.3 Quantitative Data Graphs
2.3.1 Histograms
2.3.2 Using Histograms to Get an Initial Overview of the Data
2.3.3 Frequency Polygons
2.3.4 Ogives
Self Assessment Questions
2.4 Qualitative Data Graphs
2.4.1 Pie Charts
2.4.2 Bar Graphs
Self Assessment Questions
Activity
2.5 Charts and Graphs for Two Variables
2.5.1 Cross Tabulation
2.5.2 Scatter Plot
Self Assessment Questions
Activity
2.6 Summary
2.7 Descriptive Questions
2.8 Solutions for Descriptive Questions
2.9 Answers and Hints

INTRODUCTORY CASELET

N O

DATA VISUALIZATION USING VIZARD

Infruid Labs is a Hyderabad based Business Analytics and Data Visualization Solutions company. The company offers analytics software that can handle big data to provide actionable business intelligence to customers by helping them in finding patterns in large datasets. According to Gartner IT Glossary "*Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.*"

The company was founded by Mahesh Yellai a management graduate from Indian School of Business, Hyderabad and a B. Tech. from Indian Institute of Technology – Madras to help organizations to make data-driven decisions. The company is helping its clients in different business areas such as product strategy, human resource management, new product development, operations management, competitor analysis, financial analysis, etc. The company was awarded best Enterprise Software Startup by HYSEA in 2016 and also got awarded with one of 50 Emerging Startups award by NASSCOM in 2017.

Vizard is Infruid's flagship analytics product. As an instant search-driven analytics platform, Vizard help users to derive deep insights from their data. Vizard helped its clients to democratize analytics across the Organization, so that everyone in the organization, who needs to make decisions, has easy access to instant insights. Some of the best run businesses are using Vizard across Sales, Marketing, Finance, HR and Operations departments.

Vizard's consume-grade user experience hides all the complexity of Big Data processing behind a simple search box, so users can type any question about their Business and slice and dice their data even at big data scale to get visual insights instantaneously.

Source: https://infruid.com/vizard/

Few examples where Infruid's analytics platform helped its clients

- A leading industrial equipment manufacturing company used Vizard to help its Sales team to optimize selling price and maximize revenue. The optimal selling price, for each of thousands of products, was determined in Vizard using historical data. Vizard helped the company to increase its revenue by facilitating data-driven pricing decisions.
- A leading Steel manufacturer used Vizard to derive insights from the inventory data and to interact with Enterprise Resource Planning (ERP) software. Vizard has helped the company to plan its inventory

stock movement across the various distributed warehouses and reduce the Inventory Holding Costs.

- A leading renewable energy company used Vizard to run its Network Operations Centre to analyze Internet of Things (IoT) data from its solar power plants and remotely monitor geographically distributed power plants. Vizard also helped the company to lower its operational expenditure through preventive maintenance.

WAY AHEAD

The company has planned to release enhanced data discovery capabilities to support augmented data discovery. Augmented data discovery feature uses Artificial Intelligence (AI) to auto detect patterns of interest from data and bubble them up to users' attention.

Source: Adapted from Srishti Deoras, "A Lot Of Enterprises Do Not Have A Clear Separation Of Ownership Of Data And Insights, Says Mahesh Yellai Of Infruid", *Analytics India*, https://analyticsindiamag.com/lot-enterprises-not-clear-separation-ownership-data-insights-says-mahesh-yellai-infruid/; Gartner IT Glossary, Big Data, https://www.gartner.com/it-glossary/big-data

LEARNING OBJECTIVES

The overall objective of Chapter 2 is for you to master several techniques for summarizing and depicting data, thereby enabling you to:

- Create a frequency distribution from a set of data.
- Create and evaluate different types of quantitative data graphs, including histograms, frequency polygons, and ogives in order to evaluate the data being graphed.
- Create and evaluate different types of qualitative data graphs, including pie charts, bar graphs, and Pareto charts, in order to analyze the data being graphed.
- Create a cross-tabulation table and analyze basic trends in two-variable scatter plots of numerical data.

2.1 INTRODUCTION

In Chapters 2 and 3 many techniques are presented for reformatting or reducing data so that the data are more manageable and can assist decision makers more effectively. Some of the most effective mechanisms for presenting data in a form meaningful to decision makers are graphical depictions. This chapter focuses on graphical tools for summarizing and presenting data. Through graphs and charts, the decision maker can often get an overall picture of the data and reach some useful conclusions merely by studying the chart or graph. Key characteristics of graphs often suggest appropriate choices among potential numerical methods (discussed in later chapters) for analyzing data. Visual representations of data are often much more effective communication tools than tables of numbers in business meetings.

A first step in exploring and analyzing data is to reduce important and sometimes expensive data to a graphic picture that is clear, concise and consistent with the message of the original data. Converting data to graphics can be creative and artful. In this chapter, guidelines are provided for selecting appropriate graphical representations for data sets. Charts and graphs discussed in Chapter 2 include histograms, frequency polygons, ogives, bar charts, pie charts, and Pareto charts for one-variable data and both cross-tabulation tables and scatter plots for two-variable numerical data.

2.2 FREQUENCY DISTRIBUTIONS

Raw data, or data that have not been summarized in any way, are sometimes referred to as **ungrouped data**. As an example, **Table 2.1** contains 60 years of raw data of the unemployment rates for Canada. *Data that have been organized into a frequency distribution* are called **grouped data. Table 2.2** presents a frequency distribution for the data displayed in Table 2.1.

The distinction between ungrouped and grouped data is important because the calculation of statistics differs between the two types of data. Several of the charts and graphs presented in this chapter are constructed from grouped data.

TABLE 2.1: 60 YEARS OF CANADIAN UNEMPLOYMENT RATES (UNGROUPED DATA)

2.3	7.0	6.3	11.3	9.6
2.8	7.1	5.6	10.6	9.1
3.6	5.9	5.4	9.7	8.3
2.4	5.5	7.1	8.8	7.6
2.9	4.7	7.1	7.8	6.8
3.0	3.9	8.0	7.5	7.2
4.6	3.6	8.4	8.1	7.7
4.4	4.1	7.5	10.3	7.6
3.4	4.8	7.5	11.2	7.2
4.6	4.7	7.6	11.4	6.8
6.9	5.9	11.0	10.4	6.3
6.0	6.4	12.0	9.5	6.0

TABLE 2.2: FREQUENCY DISTRIBUTION OF 60 YEARS OF UNEMPLOYMENT DATA FOR CANADA (GROUPED DATA)

Class Interval	Frequency
1–under 3	4
3–under 5	12
5–under 7	13
7–under 9	19
9–under 11	7
11–under 13	5

One particularly useful tool for grouping data is the **frequency distribution**, which is *a summary of data presented in the form of class intervals and frequencies*. How is a frequency distribution constructed from raw data? That is, how are frequency distributions like the one displayed in Table 2.2 constructed from raw data like those presented in Table 2.1? Frequency distributions are relatively easy to construct. Although some guidelines and rules of thumb help in their construction, frequency distributions vary in final shape and design, even when the original raw data are identical. In a sense, frequency distributions are constructed according to individual business researchers' taste.

When constructing a frequency distribution, the business researcher should first determine the range of the raw data. The **range** often is defined as *the difference between the largest and smallest numbers*. The range for the data in Table 2.1 is 9.7 (12.0–2.3).

The second step in constructing a frequency distribution is to determine how many classes it will contain. One rule of thumb is to select between *5 and 15 classes*. If the frequency distribution contains too few classes, the data summary may be too general to be useful. Too many classes may result in a frequency distribution that does not aggregate the data enough to be helpful. The final number of classes is arbitrary. The business researcher arrives at a number by examining the range and determining a number of classes that will span the range adequately and also be meaningful to the user. The data in Table 2.1 were grouped into six classes for Table 2.2.

After selecting the number of classes, the business researcher must determine the width of the class interval. An approximation of the class width can be calculated by dividing the range by the number of classes. For the data in Table 2.1, this approximation would be 9.7/6 = 1.62. Normally, the number is rounded up to the next whole number, which in this case is 2. The frequency distribution must start at a value equal to or lower than the lowest number of the ungrouped data and end at a value equal to or higher than the highest number. The lowest unemployment rate is 2.3 and the highest is 12.0, so the business researcher starts the frequency distribution at 1 and ends it at 13. Table 2.2 contains the completed frequency distribution for the data in Table 2.1. Class endpoints are selected so that no value of the data can fit into more than one class. The class interval expression "under" in the distribution of Table 2.2 avoids such a problem.

2.2.1 CLASS MIDPOINT

The *midpoint of each class interval* is called the **class midpoint** and is sometimes referred to as the **class mark**. It is *the value halfway across the class interval* and can be calculated as *the average of the two class endpoints*. For example, in the distribution of Table 2.2, the midpoint of the class interval 3–under 5 is 4, or (3 + 5)/2.

The class midpoint is important, because it becomes the representative value for each class in most group statistics calculations. The third column in **Table 2.3** contains the class midpoints for all classes of the data from Table 2.2.

2.2.2 RELATIVE FREQUENCY

Relative frequency is *the proportion of the total frequency that is in any given class interval in a frequency distribution*. Relative frequency is the individual class frequency divided by the total frequency. For example, from Table 2.3, the relative frequency for the class interval 5–under 7 is 13/60 = .2167.

TABLE 2.3: CLASS MIDPOINTS, RELATIVE FREQUENCIES, AND CUMULATIVE FREQUENCIES FOR UNEMPLOYMENT DATA

Interval	Frequency	Class Midpoint	Relative Frequency	Cumulative Frequency
1–under 3	4	2	.0667	4
3–under 5	12	4	.2000	16
5–under 7	13	6	.2167	29
7–under 9	19	8	.3167	48
9–under 11	7	10	.1167	55
11–under 13	5	12	.0833	60
Total	60			

Consideration of the relative frequency is preparatory to the study of prob ability in Chapter 4. Indeed, if values were selected randomly from the dat in Table 2.1, the probability of drawing a number that is "5–under 7" woul be .2167, the relative frequency for that class interval. The fourth column Table 2.3 lists the relative frequencies for the frequency distribution of Table 2.

2.2.3 CUMULATIVE FREQUENCY

The **cumulative frequency** is *a running total of frequencies through the classe of a frequency distribution*. The cumulative frequency for each class interva is the frequency for that class interval added to the preceding cumulativ total. In Table 2.3 the cumulative frequency for the first class is the same the class frequency: 4. The cumulative frequency for the second class inte val is the frequency of that interval (12) plus the frequency of the first interva (4), which yields a new cumulative frequency of 16. This process continue through the last interval, at which point the cumulative total equals the su of the frequencies (60). The concept of cumulative frequency is used in man areas, including sales cumulated over a fiscal year, sports scores during contest (cumulated points), years of service, points earned in a course, an costs of doing business over a period of time. Table 2.3 gives cumulati frequencies for the data in Table 2.2.

DEMONSTRATION PROBLEM 2.1

The following data from the Federal Home Loan Mortgage Corporation are the average monthly 30-year fixed rate mortgage interest rates for a recent 40-month period.

5.06	4.89	4.75	4.11	3.81
4.95	4.74	4.95	4.07	3.69
4.88	4.56	4.83	3.99	3.55
4.88	4.43	4.84	3.95	3.60
5.05	4.35	4.64	3.92	3.50
4.99	4.24	4.51	3.89	3.38
4.97	4.30	4.54	3.95	3.35
5.10	4.71	4.27	3.91	3.34

Construct a frequency distribution for these data. Calculate and display th class midpoints, relative frequencies, and cumulative frequencies for this fre quency distribution.

Solution: How many classes should this frequency distribution contain The range of the data is 1.76 (5.10 – 3.34). If 8 classes are used, each clas width is approximately:

$$\text{Class width} = \frac{\text{Range}}{\text{Number of Classes}} = \frac{1.76}{8} = 0.22$$

If a class width of .25 is used, a frequency distribution can be constructe with endpoints that are more uniform looking and allow for presentatio of the information in categories more familiar to mortgage interest ra users.

The first endpoint must be 3.34 or lower to include the smallest value; the last endpoint must be 5.10 or higher to include the largest value. In this case, the frequency distribution begins at 3.25 and ends at 5.25. The resulting frequency distribution, class midpoints, relative frequencies, and cumulative frequencies are listed in the following table:

Interval	Frequency	Class Midpoint	Relative Frequency	Cumulative Frequency
3.25–under 3.50	3	3.375	.075	3
3.50–under 3.75	4	3.625	.100	7
3.75–under 4.00	7	3.875	.175	14
4.00–under 4.25	3	4.125	.075	17
4.25–under 4.50	4	4.375	.100	21
4.50–under 4.75	6	4.625	.150	27
4.75–under 5.00	10	4.875	.250	37
5.00–under 5.25	3	5.125	.075	40
Total	40			

The frequencies and relative frequencies of these data reveal the mortgage interest rate classes that are likely to occur during this period of time. Overall, the mortgage rates are distributed relatively evenly, with the 4.75–under 5.00 class interval containing the greatest frequency (10), followed by the 3.75–under 4.00 class interval (7), and the 4.50–under 4.75 interval (6).

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. _____ is the individual class frequency divided by the total frequency.
2. Class midpoint is also sometimes referred as the _____.
3. The cumulative frequency for each class interval is the _____ for that class interval added to the preceding cumulative total.

ACTIVITY

List three specific uses of cumulative frequencies in business.

2.3 QUANTITATIVE DATA GRAPHS

One of the most effective mechanisms for presenting data in a form meaningful to decision makers is graphical depiction. Through graphs and charts, the decision maker can often get an overall picture of the data and reach some useful conclusions merely by studying the chart or graph. Converting data to graphics can be creative and artful. Often the most difficult step in this process is to reduce important and sometimes expensive data to a graphic picture that is both clear and concise and yet consistent with the message of

the original data. One of the most important uses of graphical depiction in statistics is to help the researcher determine the shape of a distribution. Data graphs can generally be classified as quantitative or qualitative. **Quantitative data graphs** are plotted along a numerical scale, and **qualitative graphs** are plotted using non-numerical categories. In this section, we will examine three types of quantitative data graphs: (1) histogram, (2) frequency polygon, and (3) ogive.

2.3.1 HISTOGRAMS

One of the more widely used types of graphs for quantitative data is the **histogram**. A histogram is a series of contiguous rectangles that represent the frequency of data in given class intervals. If the class intervals used along the horizontal axis are equal, then the heights of the rectangles represent the frequency of values in a given class interval. If the class intervals are unequal, then the areas of the rectangles can be used for relative comparisons of class frequencies. Construction of a histogram involves labelling the x-axis (abscissa) with the class endpoints and the y-axis (ordinate) with the frequencies, drawing a horizontal line segment from class endpoint to class endpoint at each frequency value, and connecting each line segment vertically from the frequency value to the x-axis to form a series of rectangles. **Figure 2.1** is a histogram of the frequency distribution in Table 2.2.

A histogram is a useful tool for differentiating the frequencies of class intervals. A quick glance at a histogram reveals which class intervals produce the highest frequency totals. Figure 2.1 clearly shows that the class interval 7–under 9 yields by far the highest frequency count (19). Examination of the histogram reveals where large increases or decreases occur between classes, such as from the 1–under 3 class to the 3–under 5 class, an increase of 8, and from the 7–under 9 class to the 9–under 11 class, a decrease of 12.

Figure 2.1: Histogram of Canadian Unemployment Data

Note that the scales used along the x- and y-axes for the histogram in Figure 2.1 are almost identical. However, because ranges of meaningful numbers for the two variables being graphed often differ considerably, the graph may have different scales on the two axes.

Figure 2.2 shows what the histogram of unemployment rates would look like if the scale on the y-axis were more compressed than that on the x-axis. Notice that with the compressed graph, Figure 2.2, there appears to be less difference between the lengths of the rectangles than those in Figure 2.1 implying that the differences in frequencies for the compressed graph are not as great as they are in Figure 2.1. It is important that the user of the graph clearly understands the scales used for the axes of a histogram. Otherwise, a graph's creator can "lie with statistics" by stretching or compressing a graph to make a point.*

Figure 2.2: Histogram of Canadian Unemployment Data (y-axis compressed)

**It should be pointed out that the software package Excel uses the term histogram to refer to a frequency distribution. However, by checking Chart Output in the Excel histogram dialog box, a graphical histogram is also created.*

2.3.2 USING HISTOGRAMS TO GET AN INITIAL OVERVIEW OF THE DATA

Because of the widespread availability of computers and statistical software packages to business researchers and decision makers, the histogram continues to grow in importance in yielding information about the shape of the distribution of a large database, the variability of the data, the central location of the data, and outlier data. Although most of these concepts are presented in Chapter 3, the notion of histogram as an initial tool to access these data characteristics is presented here.

A business researcher measured the volume of stocks traded on Wall Street three times a month for nine years resulting in a database of 324 observations. Suppose a financial decision maker wants to use these data to reach some conclusions about the stock market.

Figure 2.3 shows a produced histogram of these data. What can we learn from this histogram? Virtually all stock market volumes fall between zero and 1 billion shares. The distribution takes on a shape that is high on the left end and tapered to the right. In Chapter 3 we will learn that the shape of this distribution is skewed toward the right end. In statistics, it is often useful to determine whether data are approximately normally distributed (bell shaped curve) as shown in **Figure 2.4**.

Figure 2.3: Histogram of Stock Volumes

Normal Distribution

Figure 2.4: Normal Distribution

We can see by examining the histogram in Figure 2.3 that the stock market volume data are not normally distributed. Although the centre of the histogram is located near 500 million shares, a large portion of stock volume observations falls in the lower end of the data somewhere between 100 million and 400 million shares. In addition, the histogram shows some outliers in the upper end of the distribution. Outliers are data points that appear outside of the main body of observations and may represent phenomena that differ from those represented by other data points. By observing the histogram, we notice a few data observations near 1 billion. One could conclude that on a few stock market days an unusually large volume of shares are traded. These and other insights can be gleaned by examining the histogram and show that histograms play an important role in the initial analysis of data.

2.3.3 FREQUENCY POLYGONS

A **frequency polygon**, like the histogram, is a graphical display of class frequencies. However, instead of using rectangles like a histogram, in a frequency polygon each class frequency is plotted as a dot at the class midpoint, and the dots are connected by a series of line segments. Construction of a frequency polygon begins by scaling class midpoints along the horizontal axis and the frequency scale along the vertical axis. A dot is plotted for the associated frequency value at each class midpoint. Connecting these midpoint dots completes the graph. **Figure 2.5** shows a frequency polygon of the distribution data from Table 2.2 produced by using the software package Excel. The information gleaned from frequency polygons and histograms is similar. As with the histogram, changing the scales of the axes can compress or stretch a frequency polygon, which affects the user's impression of what the graph represents.

Figure 2.5: Frequency Polygon of the Unemployment Data

2.3.4 OGIVES

An **ogive** (o-jive) is a *cumulative frequency polygon*. Construction begins by labeling the x-axis with the class endpoints and the y-axis with the frequencies. However, the use of cumulative frequency values requires that the scale along the y-axis be great enough to include the frequency total. A dot of zero frequency is plotted at the beginning of the first class, and construction proceeds by marking a dot at the *end* of each class interval for the cumulative value. Connecting the dots then completes the ogive. **Figure 2.6** presents an ogive produced by using Excel for the data in Table 2.2.

Ogives are most useful when the decision maker wants to see running totals. For example, if a comptroller is interested in controlling costs, an ogive could depict cumulative costs over a fiscal year.

Figure 2.6: Ogive of the Unemployment Data

Steep slopes in an ogive can be used to identify sharp increases in frequencies. In Figure 2.6, a particularly steep slope occurs in the 7–under 9 class signifying a large jump in class frequency totals.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

4. Through _____ and _____, the decision maker can often get an overall picture of the data and reach some useful conclusions.
5. Data graphs can generally be classified as _____ or _____.
6. _____ are data points that appear outside of the main body of observations and may represent phenomena that differ from those represented by other data points.
7. A(n) _____ is a cumulative frequency polygon.

2.4 QUALITATIVE DATA GRAPHS

In contrast to quantitative data graphs that are plotted along a numerical scale, qualitative graphs are plotted using non-numerical categories. In this section, we will examine three types of qualitative data graphs: (1) pie charts, (2) bar charts, and (3) Pareto charts.

2.4.1 PIE CHARTS

A **pie chart** is *a circular depiction of data where the area of the whole pie represents 100% of the data and slices of the pie represent a percentage breakdown of the sublevels.* Pie charts show the relative magnitudes of the parts to the whole. They are widely used in business, particularly to depict such things as budget categories, market share, and time/resource allocations. However, the use of pie charts is minimized in the sciences and technology because pie charts can lead to less accurate judgments than are possible with other types of graphs. Generally, it is more difficult for the viewer to interpret the relative size of angles in a pie chart than to judge the length of rectangles in a bar chart.

Construction of the pie chart begins by determining the proportion of the sub-unit to the whole. **Table 2.4** contains the refining capacity (1,000 barrels per day) of the top five petroleum refining companies in the United States in a recent year.

To construct a pie chart from these data, first convert the raw capacity figures to proportions by dividing each capacity figure by the total capacity figure (15,134). This proportion is analogous to the relative frequency computed for frequency distributions. Because a circle contains 360°, each proportion is then multiplied by 360 to obtain the number of degrees to represent each company in the pie chart. For example, Exxon Mobil's capacity of 5,589 (1,000 barrels) represents a .3693 proportion of the total capacity for these five companies. (5,589/15,134 = 0.3693). Multiplying this value by 360° results in an angle of 132.95°. The pie chart is then constructed by determining each of the other angles and using a compass to lay out the slices. The pie chart in **Figure 2.7**, depicts the data from **Table 2.4**.

TABLE 2.4: TOP FIVE U.S. PETROLEUM REFINING COMPANIES IN CAPACITY

Company	Capacity (1,000 barrels per day)	Proportion	Degrees
Exxon Mobil	5,589	.3693	132.95
Valero Energy	2,777	.1835	66.06
Chevron	2,540	.1678	60.41
ConocoPhillips	2,514	.1661	59.79
Marathon Oil	1,714	.1133	40.79
Totals	15,134	1.0000	360.00

2.4.2 BAR GRAPHS

Another widely used qualitative data graphing technique is the **bar graph** or **bar chart**. A bar graph or chart contains two or more categories along one axis and a series of bars, one for each category, along the other axis. Typically, the length of the bar represents the magnitude of the measure (amount, frequency, money, percentage, etc.) for each category. The bar graph is qualitative because the categories are non-numerical, and it may be either horizontal

Figure 2.7: Pie Chart of the Petroleum Refining Capacity of the Top Five U.S. Companies

or vertical. In Excel, horizontal bar graphs are referred to as **bar charts**, and vertical bar graphs are referred to as **column charts**. A bar graph generally is constructed from the same type of data that is used to produce a pie chart. However, an advantage of using a bar graph over a pie chart for a given set of data is that for categories that are close in value, it is considered easier to see the difference in the bars of bar graph than discriminating between pie slices.

As an example, consider the data in **Table 2.5** regarding how much the average college student spends on back-to-college shopping. Constructing a bar graph from these data, the categories are Electronics, Clothing and Accessories, Dorm Furnishings, School Supplies, and Misc. Bars for each of these categories are made using the dollar figures given in the table. The resulting bar graph produced by Excel is shown in **Figure 2.8.**

TABLE 2.5: HOW MUCH IS SPENT ON BACK-TO-COLLEGE SHOPPING BY THE AVERAGE STUDENT

Category	Amount Spent ($ US)
Electronics	$211.89
Clothing and Accessories	134.40
Dorm Furnishings	90.90
School Supplies	68.47
Misc.	93.72

Figure 2.8: Bar Graph of Back-to-College Spending

DEMONSTRATION PROBLEM 2.2

According to the National Retail Federation and Center for Retailing Education at the University of Florida, the four main sources of inventory shrinkage are employee theft, shoplifting, administrative error, and vendor fraud. The estimated annual dollar amount in shrinkage ($ millions) associated with each of these sources follows:

Employee theft	$17,918.6
Shoplifting	15,191.9
Administrative error	7,617.6
Vendor fraud	2,553.6
Total	$43,281.7

Construct a pie chart and a bar chart to depict these data.

Solution: To produce a pie chart, convert each raw dollar amount to a proportion by dividing each individual amount by the total.

Employee theft	17,918.6/43,281.7 = .414
Shoplifting	15,191.9/43,281.7 = .351
Administrative error	7,617.6/43,281.7 = .176
Vendor fraud	2,553.6/43,281.7 = .059
Total	1.000

Convert each proportion to degrees by multiplying each proportion by 360°.

Employee theft	.414 . 360° = 149.0°
Shoplifting	.351 . 360° = 126.4°
Administrative error	.176 . 360° = 63.4°
Vendor fraud	.059 . 360° = 21.2°
Total	360.0°

Using the raw data above, we can produce the following bar chart.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

8. _____ is widely used in business, particularly to depict things such as budget categories, market share, and time/resource allocations.
9. Bar Graphs are also called _____.
10. In a bar graph the length of the bar represents the _____.
11. A graphical technique for displaying problem causes is _____.

SELF ASSESSMENT QUESTIONS

State whether the following statements are true/false:

12. Pie charts show the relative magnitudes of the parts to the whole.
13. A bar graph can contain maximum two categories along any one axis.

ACTIVITY

Construct an ogive for the variable Type in the financial database. The 100 companies in this database are each categorized into one of seven types of companies. Construct a pie chart of these types and discuss the output. For example, which type is most prevalent in the database and which is the least?

2.5 CHARTS AND GRAPHS FOR TWO VARIABLES

It is very common in business statistics to want to analyze two variables simultaneously in an effort to gain insight into a possible relationship between them. For example, business researchers might be interested in the relationship between years of experience and amount of productivity in a manufacturing facility or in the relationship between a person's technology usage and their age. Business statistics has many techniques for exploring such relationships. Two of the more elementary tools for observing the relationships between two variables are cross tabulation and scatter plot.

2.5.1 CROSS TABULATION

Cross tabulation is *a process for producing a two-dimensional table that displays the frequency counts for two variables simultaneously.* As an example, suppose a job satisfaction survey of a randomly selected sample of 177 bankers is taken in the banking industry. The bankers are asked how satisfied they are with their job using a 1 to 5 scale where 1 denotes very dissatisfied, 2 denotes dissatisfied, 3 denotes neither satisfied nor dissatisfied, 4 denotes satisfied, and 5 denotes very satisfied. In addition, each banker is asked to report his/her age by using one of the three categories: under 30 years, 30 to 50 years, and over 50 years. **Table 2.6** displays how some of the data might look as they are gathered.

TABLE 2.6: BANKER DATA OBSERVATIONS BY JOB SATISFACTION AND AGE

Banker	Level of Job Satisfaction	Age
1	4	53
2	3	37
3	1	24
4	2	28
5	4	46
6	5	62
7	3	41
8	3	32
9	4	29
.		
.		
.		
177	3	51

Note that for each banker the level of their job satisfaction and their age are recorded. By tallying the frequency of responses for each combination of categories between the two variables, the data are cross tabulated according to the two variables. For instance, in this example, there is a tally of how many bankers rated their level of satisfaction as 1 and were under 30 years of age, there is a tally of how many bankers rated their level of satisfaction as 2 and were under 30 years of age, and so on until frequency tallies are determined for each possible combination of the two variables. **Table 2.7** shows the completed cross-tabulation table for the banker survey. A cross-tabulation table is sometimes referred to as a **contingency table**, and Excel refers to such a table as a **Pivot Table.**

TABLE 2.7: CROSS-TABULATION TABLE OF BANKER DATA

		Age Category			
		Under 30	30–50	Over 50	Total
Level of Job Satisfaction	1	7	3	0	10
	2	19	14	3	36
	3	28	17	12	57
	4	11	22	16	49
	5	2	9	14	25
	Total	67	65	45	177

2.5.2 SCATTER PLOT

A **scatter plot** is a *two-dimensional graph plot of pairs of points from two numerical variables*. The scatter plot is a graphical tool that is often used to examine possible relationships between two variables.

Observe the data in **Table 2.8**. Displayed are the values of new residential and new non-residential buildings in the United States for various years over a 35-year period. Do these two numerical variables exhibit any relationship? It might seem logical when new construction booms that it would boom in both residential building and in non-residential building at the same time. However, the scatter plot of these data displayed in **Figure 2.9** shows somewhat mixed results.

TABLE 2.8: VALUE OF NEW CONSTRUCTION OVER A 35-YEAR PERIOD

Residential	Nonresidential	Residential	Nonresidential
169635	96497	252745	169173
155113	115372	228943	167896
149410	96407	197526	135389
175822	129275	232134	120921
162706	140569	249757	122222
134605	145054	274956	127593
195028	131289	251937	139711
231396	155261	281229	153866
234955	178925	280748	166754
266481	163740	297886	177639
267063	160363	315757	175048
263385	164191		

Source: U.S. Census Bureau, *Current Construction Reports* (in millions of constant dollars).

Figure 2.9: Scatter Plot of New Residential and New Nonresidential Construction

The apparent tendency is that more new residential building construction occurs when more new non-residential building construction is also taking place, and less new residential building construction when new non-residential building construction is also at lower levels. The scatter plot also shows that in some years more new residential building and less new non-residential building happened at the same time, and vice versa.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

14. Cross tabulation is a process for producing a two-dimensional table that displays the _____ for two variables simultaneously.
15. A cross-tabulation table is also referred as a _____.

State whether the following statements are true/false:

16. It is uncommon in business statistics to analyze two variables simultaneously.
17. Scatter plots are graphical tools that are often used to examine possible relationships between two or more variables.

ACTIVITY

From the National Stock Exchange (NSE), India website, download the historical data of NIFTY 50 and NIFTY 100 for last one month period. (use https://www.nseindia.com/products/content/equities/indices/historical_index_data.htm to download your data). Plot a scatter plot of the daily close of NIFTY 50 versus NIFTY 100 for the one month period. What does the graph tell you about the relationship of the NIFTY 50 Index to the NIFTY 100 Index? Explain what you observe in the graph and share any conclusions that you reach.

2.6 SUMMARY

- The two types of data are grouped and ungrouped. Grouped data are data organized into a frequency distribution. Differentiating between grouped and ungrouped data is important, because statistical operations on the two types are computed differently.
- Constructing a frequency distribution involves several steps. The first step is to determine the range of the data, which is the difference between the largest value and the smallest value. Next, the number of classes is determined, which is an arbitrary choice of the researcher. However, too few classes over-aggregate the data into meaningless categories, and too many classes do not summarize the data enough to be useful. The third step in constructing the frequency distribution is to determine the width of the class interval.
- Dividing the range of values by the number of classes yields the approximate width of the class interval. The class midpoint is the midpoint of a class interval. It is the average of the class endpoints and represents the halfway point of the class interval.
- Relative frequency is a value computed by dividing an individual frequency by the sum of the frequencies. Relative frequency represents the proportion of total values that is in a given class interval.
- The cumulative frequency is a running total frequency tally that starts with the first frequency value and adds each ensuing frequency to the total.
- Two types of graphical depictions are quantitative data graphs and qualitative data graphs.
- A histogram is a vertical chart in which a line segment connects class endpoints at the value of the frequency. Two vertical lines connect this line segment down to the x-axis, forming a rectangle.
- Ogives are cumulative frequency polygons. Points on an ogive are plotted at the class endpoints.
- A pie chart is a circular depiction of data. The amount of each category is represented as a slice of the pie proportionate to the total. The researcher is cautioned in using pie charts because it is sometimes difficult to differentiate the relative sizes of the slices.
- The bar chart or bar graph uses bars to represent the frequencies of various qualitative categories. The bar chart can be displayed horizontally or vertically.
- Cross tabulation is a process for producing a two-dimensional table that displays the frequency counts for two variables simultaneously.
- The scatter plot is a two-dimensional plot of pairs of points from two numerical variables. It is used to graphically determine whether any apparent relationship exists between the two variables.

KEY WORDS

1. **Ungrouped data:** Raw data, or data that have not been summarized in any way.
2. **Grouped data:** Data that have been organized into a frequency distribution are called grouped data.
3. **Frequency distribution:** A frequency distribution is a summary of data presented in the form of class intervals and frequencies.
4. **Range:** It the difference between the largest and smallest numbers in the data set.
5. **Histogram:** A histogram is a series of contiguous rectangles that represent the frequency of data in given class intervals.
6. **Frequency polygon:** It is a graphical display of class frequencies where each class frequency is plotted as a dot at the class midpoint, and the dots are connected by a series of line segments.
7. **Ogive:** An ogive is a cumulative frequency polygon.
8. **Cross tabulation:** It is a process of producing a two-dimensional table that displays the frequency counts for two variables simultaneously.
9. **Bar graph:** A graph or chart with two or more categories along one axis and a series of bars, one for each category, along the other axis.

KEY WORDS

10. **Column chart:** A vertical bar graph is referred as column chart
11. **Pie chart:** A pie chart is a circular depiction of data where the area of the whole pie represents 100% of the data and slices of the pie represent a percentage breakdown of the sublevels.

2.7 DESCRIPTIVE QUESTIONS

2.1. A northwestern distribution company surveyed 53 of its midlevel ma agers. The survey obtained the ages of these managers, which later we organized into the frequency distribution shown. Determine the cla midpoint, relative frequency, and cumulative frequency for these data

Class Interval	Frequency
20–under 25	8
25–under 30	6
30–under 35	5
35–under 40	12
40–under 45	15
45–under 50	7

2.2. In a medium-sized southern city, 86 houses are for sale, each havi about 2000 square feet of floor space. The asking prices vary. The fr quency distribution shown contains the price categories for the houses. Construct a histogram, a frequency polygon, and an ogive fro these data.

Asking Price	Frequency
$ 80,000–under $100,000	21
100,000–under 120,000	27
120,000–under 140,000	18
140,000–under 160,000	11
160,000–under 180,000	6
180,000–under 200,000	3

2.3 The following figures for U.S. imports of agricultural products an manufactured goods were taken from selected years over a 30-yea period (in $ billions). The source of the data is the U.S. Internationa Trade Administration. Construct a scatter plot for these data an determine whether any relationship is apparent between the U.S. impor of agricultural products and the U.S. imports of manufactured good during this time period.

Agricultural Products	Manufactured Goods
5.8	27.3
9.5	54.0
17.4	133.0
19.5	257.5
22.3	388.8
29.3	629.7

2.4. Shown here is a list of the industries with the largest total toxic releases in a recent year according to the EPA. Construct both a bar chart and a pie chart to depict this information and comment on the advantages of each type of chart in depicting these data.

Industry	Total Releases (billions pounds)
Metal Mining	1.970
Electric Utilities	0.539
Chemicals	0.505
Primary Metals	0.346
Paper	0.175
Hazardous Waste Mgmt.	0.146
Food/Beverages	0.134
All others	0.306

2.8 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

2.1.

Class Interval	Frequency	Class Midpoint	Relative Frequency	Cumulative Frequency
20 – 25	8	22.5	8/53 = .1509	8
25 – 30	6	27.5	.1132	14
30 – 35	5	32.5	.0943	19
35 – 40	12	37.5	.2264	31
40 – 45	15	42.5	.2830	46
45 – 50	7	47.5	.1321	53
TOTAL	53		.9999	

2.2.

Asking Price	Frequency	Cumulative Frequency
$ 80,000 – under $ 100,000	21	21
$ 100,000 – under $ 120,000	27	48
$ 120,000 – under $ 140,000	18	66
$ 140,000 – under $ 160,000	11	77
$ 160,000 – under $ 180,000	6	83
$ 180,000 – under $ 200,000	3	86
	86	

Histogram:

Frequency Polygon:

Ogive:

2.3.

It can be observed that as the U.S. import of agricultural products increased, the U.S. import of manufactured goods also increased. As a matter of fact, a non-linear association may exist between the two variables.

2.4.

Industry	Total Release (billions lbs.)	Proportion	Degrees
Metal Mining	1.970	.478	172
Electric Utilities	0.539	.131	47
Chemicals	0.505	.123	44
Primary Metals	0.346	.084	30
Paper	0.175	.042	15
Hazardous Waste Mgmt.	0.146	.035	13
Food/Beverages	0.134	.033	12
All Others	0.306	.074	27
TOTAL	4.121	1.000	360

Pie Chart:

Bar Graph:

Of the several comparisons of pie charts and bar charts made in problems associated with this chapter, the differences between the two in this problem are the least clear. For example, in this bar chart the three categories from paper through hazardous waste mgmt. are close enough to be problematic in discipering differences. The same goes for electric utilities and chemicals and for primary metals and all others. As for the pie chart, were the categories not labeled, it would be difficult to determine the relative differences in these same categories in the pie chart. Perhaps any advantage that the pie chart has is that pie charts can be pleasing to view. All in all, especially in this problem, the choice between the two is mostly a matter of personal (or professional) preference.

2.9 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
2.2 Frequency Distributions	1.	Relative frequency
	2.	Class mark
	3.	Frequency
2.3 Quantitative Data Graphs	4.	Graphs, charts
	5.	Quantitative, qualitative
	6.	Outliers
	7.	Ogives
2.4 Qualitative Data Graphs	8.	Pie charts
	9.	Bar charts
	10.	Magnitude of the measure
	11.	Pareto analysis
	12.	True
	13.	False
2.5 Charts and Graphs for two Variables	14.	Frequency
	15.	Contingency table
	16.	False
	17.	False

3

CHAPTER

DESCRIPTIVE STATISTICS

CONTENTS

3.1 Introduction
3.2 Measures of Central Tendency: Ungrouped Data
3.2.1 Mode
3.2.2 Median
3.2.3 Mean
Self Assessment Questions
Activity
3.3 Measures of Variability: Ungrouped Data
3.3.1 Range
3.3.2 Interquartile Range
3.3.3 Mean Absolute Deviation, Variance, and Standard Deviation
3.3.4 Mean Absolute Deviation
3.3.5 Variance
3.3.6 Standard Deviation
3.3.7 Meaning of Standard Deviation
3.3.8 z Scores
3.3.9 Coefficient of Variation
Self Assessment Questions
Activity
3.4 Measures of Central Tendency and Variability: Grouped Data
3.4.1 Measures of Central Tendency
3.4.2 Measures of Variability
Self Assessment Questions
Activity
3.5 Measures of Shape
3.5.1 Skewness
3.5.2 Kurtosis
3.5.3 Box-and-Whisker Plots and Five-Number Summary
Self Assessment Questions
Activity
3.6 Summary
3.7 Descriptive Questions
3.8 Solutions for Descriptive Questions
3.9 Answers and Hints

INTRODUCTORY CASELET

LAUNDRY STATISTICS

Unlike USA and Europe Indian household use washing machine at their home and consumption of water increases in this type of uses. Domestic use of washing machine in a typical Indian family can consume 30–50 liters of water where as a Semi-Automatic Machine needs up to 100 liters per cycle. Considering that there are over 1.3 billion households in the India and every week more than 30 kgs load laundry washed in a family of four.

The average Indian spends 74 minutes per wash doing laundry. Washing laundry accounts for between 15% and 40% of the overall water consumption of a typical Indian home. Not only domestic washing of laundry but traditional clothes' washers in the India used between 50 and 60 gallons of water per 30 kgs cloths. Historically, the average washing machine in Europe uses about 4 gallons of water. However, while the average wash cycle of an American wash is about 35 minutes, the average wash cycle in Europe is about 90 minutes and in India average wash cycle is 74 minutes. Like Americans, Indians have traditionally preferred top loading washing machines because they are less expensive; they have a longer life expectancy and the user does not have to bend over to load and unload. However, the percentage of Indian using front loaders is increasing but on the other hand, Europeans prefer front loaders because they use less water, they are stackable, and they use less space. Around 90% of Europeans still prefer front loaders.

Average Americans and Indians uses same style of washing machine for their laundry. But the frequency of using washing machine is more in the India. There are two reasons for that

1. India doesn't have laundry services like US and European countries.
2. The median salary is less in India comparison to US that is why Indians don't want to spend more money for outsourcing it.

Source: Adapted from the following sources:http://www.consumer-voice.org/consumer-durables/washing-machine-buyng-guide/

LEARNING OBJECTIVES

The focus of Chapter 3 is the use of statistical techniques to describe data, thereby enabling you to:

- Apply various measures of central tendency—including the mean, median, and mode—to a set of ungrouped data.
- Apply various measures of variability—including the range, interquartile range, mean absolute deviation, variance, and standard deviation (using the empirical rule and Chebyshev's theorem)—to a set of ungrouped data.
- Compute the mean, median, mode, standard deviation, and variance of grouped data.
- Describe a data distribution statistically and graphically using skewness, kurtosis, and box-and-whisker plots.

3.1 INTRODUCTION

Chapter 2 presented graphical techniques for organizing and displaying data. Even though such graphical techniques allow the researcher to make some general observations about the shape and spread of the data, a more complete understanding of the data can be attained by summarizing the data using statistics. This chapter presents such statistical measures, including measures of central tendency, measures of variability, and measures of shape. The computation of these measures is different for ungrouped and grouped data. Hence we present some measures for ungrouped data and some measures for grouped data.

3.2 MEASURES OF CENTRAL TENDENCY: UNGROUPED DATA

One type of measure that is used to describe a set of data is the **measure of central tendency**. Measures of central tendency *yield information about the center, or middle part, of a group of numbers.* **Table 3.1** displays offer price for the 20 largest U.S. initial public offerings in a recent year according to Securities Data. For these data, measures of central tendency can yield such information as the average offer price, the middle offer price, and the most frequently occurring offer price. Measures of central tendency do not focus on the span of the data set or how far values are from the middle numbers. The measures of central tendency presented here for ungrouped data are the mode, the median, the mean, percentiles, and quartiles.

3.2.1 MODE

The **mode** is *the most frequently occurring value in a set of data.* For the data in Table 3.1 the mode is $19.00 because the offer price that recurred the most

TABLE 3.1: OFFER PRICES FOR THE 20 LARGEST U.S. INITIAL PUBLIC OFFERINGS IN A RECENT YEAR

$14.25	$19.00	$11.00	$28.00
24.00	23.00	43.25	19.00
27.00	25.00	15.00	7.00
34.22	15.50	15.00	22.00
19.00	19.00	27.00	21.00

times (four) was $19.00. Organizing the data into an ordered array (an ordering of the numbers from smallest to largest) helps to locate the mode. The following is an ordered array of the values from Table 3.1.

7.00	11.00	14.25	15.00	15.00	15.50	19.00	19.00	19.00	19.00
21.00	22.00	23.00	24.00	25.00	27.00	27.00	28.00	34.22	43.25

This grouping makes it easier to see that 19.00 is the most frequently occurring number.

In the case of a tie for the most frequently occurring value, two modes are listed. Then the data are said to be **bimodal.** If a set of data is not exactly bimodal but contains two values that are more dominant than others, some researchers take the liberty of referring to the data set as bimodal even without an exact tie for the mode. Data sets with more than two modes are referred to as **multimodal.**

In the world of business, the concept of mode is often used in determining sizes. As an example, manufacturers who produce cheap rubber flip-flops that are sold for as little as $1.00 around the world might only produce them in one size in order to save on machine setup costs. In determining the one size to produce, the manufacturer would most likely produce flip-flops in the modal size. The mode is an appropriate measure of central tendency for nominal-level data.

3.2.2 MEDIAN

The **median** is *the middle value in an ordered array of numbers.* For an array with an odd number of terms, the median is the middle number. For an array with an even number of terms, the median is the average of the two middle numbers. The following steps are used to determine the median.

Step 1 Arrange the observations in an ordered data array.

Step 2 For an odd number of terms, find the middle term of the ordered array. It is the median.

Step 3 For an even number of terms, find the average of the middle two terms. This average is the median.

Suppose a business researcher wants to determine the median for the following numbers.

15 11 14 3 21 17 22 16 19 16 5 7 19 8 9 20 4

The researcher arranges the numbers in an ordered array.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

Because the array contains 17 terms (an odd number of terms), the median is the middle number, or 15.

If the number 22 is eliminated from the list, the array would contain only 16 terms.

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

Now, for an even number of terms, the statistician determines the median by averaging the two middle values, 14 and 15. The resulting median value is 14.5.

Another way to locate the median is by finding the $(n + 1)/2$ term in an ordered array. For example, if a data set contains 77 terms, the median is the 39th term. That is,

$$\frac{n+1}{2} = \frac{77+1}{2} = \frac{78}{2} = 39\text{th term}$$

This formula is helpful when a large number of terms must be manipulated.

Consider the offer price data in Table 3.1. Because this data set contains 20 values, or $n = 20$, the median for these data is located at the $(20 + 1)/2$ term, or the 10.5th term. This result indicates that the median is located halfway between the 10th and 11th terms or the average of 19.00 and 21.00. Thus, the median offer price for the largest 20 U.S. initial public offerings is $20.00.

The median is unaffected by the magnitude of extreme values. This characteristic is an advantage, because large and small values do not inordinately influence the median. For this reason, the median is often the best measure of location to use in the analysis of variables such as house costs, income, and age. Suppose, for example, that a real estate broker wants to determine the median selling price of 10 houses listed at the following prices.

$67,000	$105,000	$148,000	$5,250,000
$91,000	$116,000	$167,000	
$95,000	$122,000	$189,000	

The median is the average of the two middle terms, $116,000 and $122,000, or $119,000. This price is a reasonable representation of the prices of the 10 houses. Note that the house priced at $5,250,000 did not enter into the analysis other than to count as one of the 10 houses. If the price of the tenth house were $200,000, the results would be the same. However, if all the house prices were averaged, the resulting average price of the original 10 houses would be $635,000, higher than 9 of the 10 individual prices.

A disadvantage of the median is that not all the information from the numbers is used. For example, information about the specific asking price of the most expensive house does not really enter into the computation of the median. The level of data measurement must be at least ordinal for a median to be meaningful.

3.2.3 MEAN

The **arithmetic mean** is *the average of a group of numbers* and is computed by summing all numbers and dividing by the number of numbers. Because the arithmetic mean is so widely used, most statisticians refer to it simply as the *mean*.

The population mean is represented by the Greek letter mu (μ). The sample mean is represented by $\bar{x}$. The formulas for computing the population mean and the sample mean are given in the boxes that follow.

Population Mean

$$\mu = \frac{\Sigma x_i}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

Sample Mean

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

The capital Greek letter sigma (Σ) is commonly used in mathematics to represent a summation of all the numbers in a grouping. Also, N is the number of terms in the population, and n is the number of terms in the sample. The algorithm for computing a mean is to sum all the numbers in the population or sample and divide by the number of terms. It is inappropriate to use the mean to analyze data that are not at least interval level in measurement.

Suppose a company has five departments with 24, 13, 19, 26, and 11 workers each. The *population mean* number of workers in each department is 18.6 workers. The computations follow.

$$\begin{array}{r} 24 \\ 13 \\ 19 \\ 26 \\ \underline{11} \\ \Sigma x_i = 93 \end{array}$$

and

$$\mu = \frac{\Sigma x_i}{N} = \frac{93}{5} = 18.6$$

The calculation of a sample mean uses the same algorithm as for a population mean and will produce the same answer if computed on the same data. However, it is inappropriate to compute a sample mean for a population or a population mean for a sample. Because both populations and samples are important in statistics, a separate symbol is necessary for the population mean and for the sample mean.

DEMONSTRATION PROBLEM 3.1

Shown below is a list of the top thirteen shopping centers in the United Kingdom by retail size in 1,000 square meters (m^2). Calculate the mode, median, and mean for this data.

Shopping Centre	Size (1000 m^2)
MetroCentre	190.0
Trafford Centre	180.9
Westfield Stratford City	175.0
Bluewater	155.7
Liverpool One	154.0
Westfield London	149.5
Intu Merry Hill	140.8
Manchester Arndale	139.4
Meadowhall	139.4
Lakeside	133.8
St. David's	130.1
Bullring	127.1
Eldon Square	125.4

https://en.wikipedia.org/wiki/List_of_UK_shopping_centres

Solution:

Mode: The mode for these data is 139.4. Two of the shopping centres, Manchester Arndale and Meadowhall, are 139.4 (1000 m^2) in size.

Median: With 13 different shopping centres in this group, $N = 13$. The median is located at the $(13 + 1)/2 = 7$th position. Because the data are already ordered, the 7th term is 140.8 (1000 m^2), which is the median. Thus, the median shopping centre of this group is Intu Merry Hill.

Mean: The total number (Σx) of square meters for this group of thirteen shopping centres is 1,941.1 (1000 m^2). The mean is calculated as:

$$\mu = \frac{\Sigma x}{N} = \frac{1{,}941.1}{13} = 149.3$$

The mean is affected by each and every value, which is an advantage. The mean uses all the data, and each data item influences the mean. It is also a disadvantage because extremely large or small values can cause the mean to be pulled toward the extreme value. Recall the preceding discussion of the 10 house prices. If the mean is computed for the 10 houses, the mean price is higher than the prices of 9 of the houses because the $5,250,000 house is included in the calculation. The total price of the 10 houses is $6,350,000, and the mean price is $635,000.

The mean is the most commonly used measure of central tendency because it uses each data item in its computation, it is a familiar measure, and it has mathematical properties that make it attractive to use in inferential statistics analysis.

PERCENTILES

Percentiles are *measures of central tendency that divide a group of data into 100 parts*. There are 99 percentiles because it takes 99 dividers to separate a group of data into 100 parts. The nth percentile is the value such that at least n percent of the data are below that value and at most $(100 - n)$ percent are above that value. Specifically, the 87th percentile is a value such that at least 87% of the data are below the value and no more than 13% are above the value. Percentiles are "stair-step" values, as shown in **Figure 3.1**, because there is no percentile between the 87th percentile and the 88th percentile. If a plant operator takes a safety examination and 87.6% of the safety exam scores are below that person's score, he or she still scores at only the 87th percentile, even though more than 87% of the scores are lower.

Percentiles are widely used in reporting test results. Almost all college or university students have taken the SAT, ACT, GRE, or GMAT examination. In most cases, the results for these examinations are reported in percentile form and also as raw scores. Shown next is a summary of the steps used in determining the location of a percentile.

STEPS IN DETERMINING THE LOCATION OF A PERCENTILE

1. Organize the numbers into an ascending-order array.
2. Calculate the percentile location (i) by:

$$i = \frac{P}{100}(N)$$

where

P = the percentile of interest

i = percentile location

N = number in the data set

3. Determine the location by either (a) or (b).

   a. If i is a whole number, the Pth percentile is the average of the value at the ith location and the value at the $(i + 1)^{st}$ location.

   b. If i is not a whole number, the Pth percentile value is located at the whole number part of $i + 1$.

For example, suppose you want to determine the 80th percentile of 1240 numbers, P is 80 and N is 1240. First, order the numbers from lowest to highest. Next, calculate the location of the 80th percentile.

$$i = \frac{80}{100}(1240) = 992$$

Because $i = 992$ is a whole number, follow the directions in step 3(a). The 80th percentile is the average of the 992nd number and the 993rd number.

$$P_{80} = \frac{(992\text{nd number} + 993\text{rd number})}{2}$$

Figure 3.1: Stair-Step Percentiles

DEMONSTRATION PROBLEM 3.2

Determine the 30th percentile of the following eight numbers: 14, 12, 19, 23, 5, 13, 28, 17.

Solution: For these eight numbers, we want to find the value of the 30th percentile, so $N = 8$ and $P = 30$.

First, organize the data into an ascending-order array.

5 12 13 14 17 19 23 28

Next, compute the value of i.

$$i = \frac{30}{100}(8) = 2.4$$

Because i is not a whole number, step 3(b) is used. The value of $i + 1$ is 2.4 + 1, or 3.4. The whole-number part of 3.4 is 3. The 30th percentile is located at the third value. The third value is 13, so 13 is the 30th percentile. Note that a percentile may or may not be one of the data values.

QUARTILES

Quartiles are *measures of central tendency that divide a group of data into four subgroups or parts.* The three quartiles are denoted as Q_1, Q_2, and Q_3. The first quartile, Q_1, separates the first, or lowest, one-fourth of the data from the upper three-fourths and is equal to the 25th percentile. The second quartile, Q_2, separates the second quarter of the data from the third quarter. Q_2 is located at the 50th percentile and equals the median of the data. The third quartile, Q_3, divides the first three-quarters of the data from the last quarter and is equal to the value of the 75th percentile. These three quartiles are shown in **Figure 3.2**.

Figure 3.2: Quartiles

Suppose we want to determine the values of Q_1, Q_2, and Q_3 for the following numbers.

106 109 114 116 121 122 125 129

The value of Q_1 is found at the 25th percentile, P_{25}, by:

$$\text{For } N = 8, i\frac{25}{100}(8) = 2$$

Because i is a whole number, P_{25} is found as the average of the second and third numbers.

$$P_{25} = \frac{(109+114)}{2} = 111.5$$

The value of Q_1 is $P_{25} = 111.5$. Notice that one-fourth, or two, of the values (106 and 109) are less than 111.5.

The value of Q_2 is equal to the median. Because the array contains an even number of terms, the median is the average of the two middle terms.

$$Q_2 = \text{median}\frac{(116+121)}{2} = 118.5$$

Notice that exactly half of the terms are less than Q_2 and half are greater than Q_2.

The value of Q_3 is determined by P_{75} as follows.

$$i = \frac{7}{100}(8) = 6$$

Because i is a whole number, P_{75} is the average of the sixth and the seventh numbers.

$$P_{75} = \frac{(122+125)}{2} = 123.5$$

The value of Q_3 is $P_{75} = 123.5$. Notice that three-fourths, or six, of the values are less than 123.5 and two of the values are greater than 123.5.

DEMONSTRATION PROBLEM 3.3

The following shows the top 16 global marketing categories for advertising spending for a recent year according to *Advertising Age*. Spending is given in millions of U.S. dollars. Determine the first, the second, and the third quartiles for these data.

Category	Ad Spending
Automotive	$22,195
Personal Care	19,526
Entertainment & Media	9,538
Food	7,793
Drugs	7,707
Electronics	4,023
Soft Drinks	3,916
Retail	3,576
Cleaners	3,571
Restaurants	3,553
Computers	3,247
Telephone	2,488
Financial	2,433
Beer, Wine & Liquor	2,050
Candy	1,137
Toys	699

Solution: For 16 marketing organizations, $N = 16$. $Q_1 = P_{25}$ is found by

$$i = \frac{25}{100}(16) = 4$$

Because i is a whole number, Q_1 is found to be the average of the fourth and fifth values from the bottom.

$$Q_1 = \frac{2433 + 2488}{2} = 2460.5$$

$Q_2 = P_{50}$ = median; with 16 terms, the median is the average of the eighth and ninth terms.

$$Q_2 = \frac{3571 + 3576}{2} = 3573.5$$

$Q_3 = P_{75}$ is solved by

$$i = \frac{75}{100}(16) = 12$$

Q_3 is found by averaging the 12th and 13th terms.

$$Q_3 = \frac{7707 + 7793}{2} = 7750$$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. Quartile 2 is also called as _____.
2. Data sets with more than two modes are referred as _____.
3. Quartile 3 is equal to the value of the _____ percentile.

State whether the following statements are true/false:

4. The mean is affected by each and every value.
5. A advantage of the median is that all the information from the numbers is used.

ACTIVITY

Ministry of Communications, Department of Telecommunications (DOT), Telecom Regulatory Authority of India (TRAI) has uploaded one dataset on https://data.gov.in/catalogs. This catalog provides Crowdsourced Mobile Data Speeds and other network information measured using TRAI MySpeed App. Download the dataset and use spreadsheet for calculation.

1. What is the average speed for North East states for CELLONE user under 3G technology?
2. What is the median speed of downloading for Vodafone user in Delhi NCR. Calculate only for 3G technology.

3.3 MEASURES OF VARIABILITY: UNGROUPED DATA

Measures of central tendency yield information about the center or middle part of a data set. However, business researchers can use another group of analytic tools, **measures of variability**, to *describe the spread or the dispersion of a set of data.* Using measures of variability in conjunction with measures of central tendency makes possible a more complete numerical description of the data.

For example, a company has 25 salespeople in the field, and the median annual sales figure for these people is $1.2 million. Are the salespeople being successful as a group or not? The median provides information about the sales of the person in the middle, but what about the other salespeople? Are all of them selling $1.2 million annually, or do the sales figures vary widely, with one person selling $5 million annually and another selling only $150,000 annually? Measures of variability provide the additional information necessary to answer that question.

Figure 3.3 shows three distributions in which the mean of each distribution is the same ($\mu = 50$) but the variabilities differ. Observation of these distributions shows that a measure of variability is necessary to complement the mean value in describing the data. Methods of computing measures of variability differ for ungrouped data and grouped data. This section focuses on seven measures of variability for ungrouped data: range, interquartile range,

Figure 3.3: Three Distributions with the Same Mean but Different Dispersions

mean absolute deviation, variance, standard deviation, z scores, and coefficient of variation.

3.3.1 RANGE

The **range** is *the difference between the largest value of a data set and the smallest value of a set.* Although it is usually a single numeric value, some business researchers define the range of data as the ordered pair of smallest and largest numbers (smallest, largest). It is a crude measure of variability, describing the distance to the outer bounds of the data set. It reflects those extreme values because it is constructed from them. An advantage of the range is its ease of computation. One important use of the range is in quality assurance, where the range is used to construct control charts. A disadvantage of the range is that, because it is computed with the values that are on the extremes of the data, it is affected by extreme values, and its application as a measure of variability is limited.

The data in Table 3.1 represent the offer prices for the 20 largest U.S. initial public offerings in a recent year. The lowest offer price was \$7.00 and the highest price was \$43.25. The range of the offer prices can be computed as the difference of the highest and lowest values:

$$\text{Range} = \text{Highest} - \text{Lowest} = \$43.25 - \$7.00 = \$36.25$$

3.3.2 INTERQUARTILE RANGE

Another measure of variability is the **interquartile range.** The interquartile range is *the range of values between the first and third quartile.* Essentially, it is the range of the middle 50% of the data and is determined by computing the value of $Q_3 - Q_1$. The interquartile range is especially useful in situations where data users are more interested in values toward the middle and less interested in extremes. In describing a real estate housing market, realtors might use the interquartile range as a measure of housing prices when describing the middle half of the market for buyers who are interested in houses in the midrange. In addition, the interquartile range is used in the construction of box-and-whisker plots.

Interquartile Range

$$Q_3 - Q_1$$

The following data indicate the top 15 trading partners of the United States in exports in 2014 according to the U.S Census Bureau.

Country	Exports (\$ billions)
Canada	312.4
Mexico	240.2
China	123.7
Japan	66.8
United Kingdom	53.8
Germany	49.4
South Korea	44.5
Netherlands	43.1
Brazil	42.4
Hong Kong	40.9
Belgium	34.8
France	31.3
Singapore	30.2
Taiwan	26.7
Switzerland	22.2

What is the interquartile range for these data? The process begins by computing the first and third quartiles as follows.

Solving for $Q_1 = P_{25}$ when $N = 15$:

$$i = \frac{25}{100}(15) = 3.75$$

Because i is not a whole number, P_{25} is found as the fourth term from the bottom.

$$Q_1 = P_{25} = 31.3$$

Solving for $Q_3 = P_{75}$:

$$i = \frac{75}{100}(15) = 11.25$$

Because i is not a whole number, P_{75} is found as the 12th term from the bottom.

$$Q_3 = P_{75} = 66.8$$

The interquartile range is:

$$Q_3 - Q_1 = 66.8 - 31.3 = 35.5$$

The middle 50% of the exports for the top 15 U.S. trading partners spans a range of 35.5 (\$ billions).

3.3.3 MEAN ABSOLUTE DEVIATION, VARIANCE, AND STANDARD DEVIATION

Three other measures of variability are the variance, the standard deviation, and the mean absolute deviation. They are obtained through similar processes and are, therefore, presented together. These measures are not meaningful unless the data are at least interval-level data. The variance and standard deviation are widely used in statistics. Although the standard deviation has some stand-alone potential, the importance of variance and standard deviation lies mainly in their role as tools used in conjunction with other statistical devices.

Suppose a small company started a production line to build computers. During the first five weeks of production, the output is 5, 9, 16, 17, and 18 computers, respectively. Which descriptive statistics could the owner use to measure the early progress of production? In an attempt to summarize these figures, the owner could compute a mean.

x_i
5
9
16
17
18

$$\Sigma x_i = 65 \quad \mu = \frac{\Sigma x_i}{N} = \frac{65}{5} = 13$$

What is the variability in these five weeks of data? One way for the owner to begin to look at the spread of the data is to subtract the mean from each data value. *Subtracting the mean from each value of data* yields the **deviation from the mean** $(x_i - \mu)$. **Table 3.2** shows these deviations for the computer company production. Note that some deviations from the mean are positive and some are negative. **Figure 3.4** shows that geometrically the negative deviations represent values that are below (to the left of) the mean and positive deviations represent values that are above (to the right of) the mean.

An examination of deviations from the mean can reveal information about the variability of data. However, the deviations are used mostly as a tool to compute other measures of variability. Note that in both Table 3.2 and Figure 3.4 these deviations total zero. This phenomenon applies to all cases. For a given set of data, the sum of all deviations from the arithmetic mean is always zero.

Sum of Deviations from the Arithmetic Mean is Always Zero

$$\Sigma(x_i - \mu) = 0$$

This property requires considering alternative ways to obtain measures of variability.

One obvious way to force the sum of deviations to have a nonzero total is to take the absolute value of each deviation around the mean. Utilizing the absolute value of the deviations about the mean makes solving for the mean absolute deviation possible.

TABLE 3.2: DEVIATIONS FROM THE MEAN FOR COMPUTER PRODUCTION

Number (x_i)	Deviations from the mean ($x_i - \mu$)
5	5 − 13 = −8
9	9 − 13 = −4
16	16 − 13 = +3
17	17 − 13 = +4
18	18 − 13 = +5
$\Sigma x_i = 65$	$\Sigma(x_i - \mu) = 0$

Figure 3.4: Geometric Distances from the Mean (from Table 3.2)

3.3.4 MEAN ABSOLUTE DEVIATION

The **mean absolute deviation (MAD)** is *the average of the absolute values of the deviations around the mean for a set of numbers.*

Mean Absolute Deviation

$$\text{MAD} = \frac{\Sigma|x_i - \mu|}{N}$$

Using the data from Table 3.2, the computer company owner can compute a mean absolute deviation by taking the absolute values of the deviations and averaging them, as shown in **Table 3.3**. The mean absolute deviation for the computer production data is 4.8.

Because it is computed by using absolute values, the mean absolute deviation is less useful in statistics than other measures of dispersion. However, in the field of forecasting, it is used occasionally as a measure of error.

3.3.5 VARIANCE

Because absolute values are not conducive to easy manipulation, mathematicians developed an alternative mechanism for overcoming the zero-sum property of deviations from the mean. This approach utilizes the square of the deviations from the mean. The result is the variance, an important measure of variability.

The **variance** is *the average of the squared deviations about the arithmetic mean for a set of numbers.* The population variance is denoted by σ^2.

TABLE 3.3: MAD FOR COMPUTER PRODUCTION DATA

x_i	$x_i - \mu$	$\lvert x_i - \mu \rvert$
5	−8	+8
9	−4	+4
16	+3	+3
17	+4	+4
18	+5	+5
$\Sigma x_i = 65$	$\Sigma(x_i - \mu) = 0$	$\Sigma \lvert x_i - \mu \rvert = 24$

$$\text{MAD} = \frac{\Sigma \lvert x_i - \mu \rvert}{N} = \frac{24}{5} = 4.8$$

TABLE 3.4: COMPUTING A VARIANCE AND A STANDARD DEVIATION FROM THE COMPUTER PRODUCTION DATA

x_i	$x_i - \mu$	$(x_i - \mu)$
5	−8	64
9	−4	16
16	+3	9
17	+4	16
18	+5	25
$\Sigma x_i = 65$	$\Sigma(x_i - \mu) = 0$	$\Sigma(x_i - \mu)^2 = 130$

$$SS_x = \Sigma(x_i - \mu)^2 = 130$$

$$\text{Variance} = \sigma^2 = \frac{SS_x}{N} = \frac{\Sigma(x_i - \mu)^2}{N} = \frac{130}{5} = 26.0$$

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} = \sqrt{\frac{130}{5}} = 5.1$$

Population Variance

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N}$$

Table 3.4 shows the original production numbers for the computer company, the deviations from the mean, and the squared deviations from the mean.

The sum of the squared deviations about the mean of a set of values—called the **sum of squares of** x and sometimes abbreviated as SS_x—is used throughout statistics. For the computer company, this value is 130. Dividing it by the number of data values (5 weeks) yields the variance for computer production.

$$\sigma^2 = \frac{130}{5} = 26.0$$

Because the variance is computed from squared deviations, the final result is expressed in terms of squared units of measurement. Statistics measured in squared units are problematic to interpret. Consider, for example, Mattel Toys attempting to interpret production costs in terms of squared dollars or Troy-Bilt measuring production output variation in terms of squared lawn mowers. Therefore, when used as a descriptive measure, variance can be considered as an intermediate calculation in the process of obtaining the standard deviation.

3.3.6 STANDARD DEVIATION

The standard deviation is a popular measure of variability. It is used both as a separate entity and as a part of other analyses, such as computing confidence intervals and in hypothesis testing (see Chapters 8, 9, and 10).

Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

The **standard deviation** is *the square root of the variance.* The population standard deviation is denoted by σ.

Like the variance, the standard deviation utilizes the sum of the squared deviations about the mean (SS_x). It is computed by averaging these squared deviations (SS_x/N) and taking the square root of that average. One feature of the standard deviation that distinguishes it from a variance is that the standard deviation is expressed in the same units as the raw data, whereas the variance is expressed in those units squared. Table 3.4 shows the standard deviation for the computer production company: $\sqrt{26}$, or 5.1.

What does a standard deviation of 5.1 mean? The meaning of standard deviation is more readily understood from its use, which is explored next. Although the standard deviation and the variance are closely related and can be computed from each other, differentiating between them is important, because both are widely used in statistics.

3.3.7 MEANING OF STANDARD DEVIATION

What is a standard deviation? What does it do, and what does it mean? Insight into the concept of standard deviation can be gleaned by viewing the manner in which it is applied. Two ways of applying the standard deviation are the **empirical rule** and **Chebyshev's theorem.**

EMPIRICAL RULE

The empirical rule is an important rule of thumb that *is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.*

The empirical rule is used only for three numbers of standard deviations: 1σ, 2σ, and 3σ. More detailed analysis of other numbers of σ values is presented in Chapter 6. Also discussed in further detail in Chapter 6 is the normal distribution, a unimodal, symmetrical distribution that is bell (or mound) shaped. The requirement that the data be normally distributed contains some tolerance, and the empirical rule generally applies as long as the data are approximately mound shaped.

Empirical Rule*

Distance from the Mean	Values within Distance
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

*Based on the assumption that the data are approximately normally distributed.

If a set of data is normally distributed, or bell shaped, approximately 68% of the data values are within one standard deviation of the mean, 95% are within two standard deviations, and almost 100% are within three standard deviations.

Suppose a recent report states that for California, the average statewide price of a gallon of regular gasoline is \$4.12. Suppose regular gasoline prices vary across the state with a standard deviation of \$0.08 and are normally distributed. According to the empirical rule, approximately 68% of the prices should fall within $\mu \pm 1\sigma$, or \$4.12 ± 1(\$0.08). Approximately 68% of the prices should be between \$4.04 and \$4.20, as shown in **Figure 3.5A**. Approximately 95% should fall within $\mu \pm 2\sigma$ or \$4.12 ± 2(\$0.08) = \$4.12 ± \$0.16, or between \$3.96 and \$4.28, as shown in **Figure 3.5B**. Nearly all regular gasoline prices (99.7%) should fall between \$3.88 and \$4.36($\mu \pm 3\sigma$).

Note that with 68% of the gasoline prices falling within one standard deviation of the mean, approximately 32% are outside this range. Because the normal distribution is symmetrical, the 32% can be split in half such that 16% lie in each tail of the distribution. Thus, approximately 16% of the gasoline prices should be less than \$4.04 and approximately 16% of the prices should be greater than \$4.20.

Many phenomena are distributed approximately in a bell shape, including most human characteristics such as height and weight; therefore the empirical rule applies in many situations and is widely used.

Figure 3.5: Empirical Rule for One and Two Standard Deviations of Gasoline Prices

DEMONSTRATION PROBLEM 3.4

A company produces a lightweight valve that is specified to weigh 1365 grams. Unfortunately, because of imperfections in the manufacturing process not all of the valves produced weigh exactly 1365 grams. In fact, the weights of the valves produced are normally distributed with a mean weight of 1365 grams and a standard deviation of 294 grams. Within what range of weights would approximately 95% of the valve weights fall? Approximately 16% of the weights would be more than what value? Approximately 0.15% of the weights would be less than what value?

Solution: Because the valve weights are normally distributed, the empirical rule applies. According to the empirical rule, approximately 95% of the weights should fall within $\mu \pm 2\sigma = 1365 \pm 2(294) = 1365 \pm 588$. Thus, approximately 95% should fall between 777 and 1953. Approximately 68% of the weights should fall within $\mu \pm 1\sigma$, and 32% should fall outside this interval. Because the normal distribution is symmetrical, approximately 16% should lie above $\mu + 1\sigma = 1365 + 294 = 1659$. Approximately 99.7% of the weights should fall within $\mu \pm 3\sigma$, and .3% should fall outside this interval. Half of these, or .15%, should lie below $\mu - 3\sigma = 1365 - 3(294) = 1365 - 882 = 483$.

CHEBYSHEV'S THEOREM

The empirical rule applies only when data are known to be approximately normally distributed. What do researchers use when data are not normally distributed or when the shape of the distribution is unknown? Chebyshev's theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or is nonnormal. Even though Chebyshev's theorem can in theory be applied to data that are normally distributed, the empirical rule is more widely known and is preferred whenever appropriate. Chebyshev's theorem is not a rule of thumb, as is the empirical rule, but rather it is presented in formula format and therefore can be more widely applied. Chebyshev's theorem states that *at least $1 - 1/k^2$ values will fall within $\pm k$ standard deviations of the mean regardless of the shape of the distribution.*

Chebyshev's Theorem

Within k standard deviations of the mean, $\mu \pm k\sigma$, lie at least

$$1 - \frac{1}{k^2}$$

proportion of the values.

Assumption: $k > 1$

Specifically, Chebyshev's theorem says that at least 75% of all values are within $\pm 2\sigma$ of the mean regardless of the shape of a distribution because if $k = 2$, then $1 - 1/k^2 = 1 - 1/2^2 = 3/4 = .75$. **Figure 3.6** provides a graphic illustration. In contrast, the empirical rule states that if the data are normally distributed 95% of all values are within $\mu \pm 2\sigma$. According to Chebyshev's theorem, the percentage of values within three standard deviations of the mean is at least 89%, in contrast to 99.7% for the empirical rule. Because a formula is used to compute proportions with Chebyshev's theorem, any value of k greater than 1($k > 1$) can be used. For example, if $k = 2.5$, at least .84 of all values are within $\mu \pm 2.5\sigma$, because $1 - 1/k^2 = 1 - 1/(2.5)^2 = .84$.

DEMONSTRATION PROBLEM 3.5

In the computing industry the average age of professional employees tends to be younger than in many other business professions. Suppose the average age of a professional employed by a particular computer firm is 28 with a standard deviation of 6 years. A histogram of professional employee ages with this firm reveals that the data are not normally distributed but rather are amassed in the 20s and that few workers are over 40. Apply Chebyshev's theorem to determine within what range of ages at least 80% of the workers' ages would fall.

Solution: Because the ages are not normally distributed, it is not appropriate to apply the empirical rule; and therefore Chebyshev's theorem must be applied to answer the question.

Chebyshev's theorem states that at least $1 - 1/k^2$ proportion of the values are within $\mu \pm k\sigma$. Because 80% of the values are within this range, let

$$1 - \frac{1}{k^2} = .80$$

Solving for k yields

$$.20 = \frac{1}{k^2}$$

$$k^2 = 5.000$$

$$k = 2.24$$

Chebyshev's theorem says that at least .80 of the values are within ± 2.24 of the mean.

For $\mu = 28$ and $\sigma = 6$, at least .80, or 80%, of the values are within $28 \pm 2.24(6) = 28 \pm 13.4$ years of age or between 14.6 and 41.4 years old.

POPULATION VERSUS SAMPLE VARIANCE AND STANDARD DEVIATION

The sample variance is denoted by s^2 and the sample standard deviation by s. The main use for sample variances and standard deviations is as estimators of population variances and standard deviations. Because of this, computation of the sample variance and standard deviation differs slightly from computation of the population variance and standard deviation. Both the sample variance and sample standard deviation use $n - 1$ in the denominator instead of n because using n in the denominator of a sample variance results in a statistic that tends to underestimate the population variance. While discussion of the properties of *good estimators* is beyond the scope of this text,

Figure 3.6: Application of Chebyshev's Theorem for Two Standard Deviations

one of the properties of a good estimator is being *unbiased*. Whereas using n in the denominator of the sample variance makes it a *biased* estimator, using $n - 1$ allows it to be an *unbiased* estimator, which is a desirable property in inferential statistics.

Sample Variance

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

Sample Standard Deviation

$$s^2 = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}}$$

Shown here is a sample of six of the largest accounting firms in the United States and the number of partners associated with each firm as reported by the *Public Accounting Report*.

Firm	Number of Partners
Deloitte & Touche	3030
Ernst & Young	2700
PricewaterhouseCoopers	2691
KPMG	1813
RSM McGladrey	644
Grant Thornton	529

The sample variance and sample standard deviation can be computed by:

x_i	$(x_i - \bar{x})^2$
3030	1,274,257.17
2700	638,129.37
2691	623,831.43
1813	7,773.95
644	1,580,476.41
529	1,882,850.51

$\Sigma x = 11{,}407$ $\qquad$ $\Sigma(x_i - \bar{x})^2 = 6{,}007{,}318.84$

$$x = \frac{11{,}407}{6} = 1901.17$$

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1} = \frac{6{,}007{,}318{,}84}{5} = 1{,}201{,}463.77$$

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{1{,}201{,}463.77} = 1096.11$$

The sample variance is 1,201,463.77, and the sample standard deviation is 1096.11

3.3.8 z SCORES

A **z score** represents the number of standard deviations a value (x) is above or below the mean of a set of numbers when the data are normally distributed. Using z scores allows translation of a value's raw distance from the mean into units of standard deviations.

z Score

$$z = \frac{x_i - \mu}{\sigma}$$

For samples,

$$z = \frac{x_i - \bar{x}}{s}$$

If a z score is negative, the raw value (x) is below the mean. If the z score is positive, the raw value (x) is above the mean.

For example, for a data set that is normally distributed with a mean of 50 and a standard deviation of 10, suppose a statistician wants to determine the z score for a value of 70. This value ($x = 70$) is 20 units above the mean, so the z value is

$$z = \frac{70 - 50}{10} = +2.00$$

This z score signifies that the raw score of 70 is two standard deviations above the mean. How is this z score interpreted? The empirical rule states that 95% of all values are within two standard deviations of the mean if the data are approximately normally distributed. **Figure 3.7** shows that because the value of 70 is two standard deviations above the mean ($z = +2.00$), 95% of the values are between 70 and the value ($x = 30$), that is

Figure 3.7: Percentage Breakdown of Scores Two Standard Deviations from the Mean

two standard deviations below the mean, or $z = (30 - 50)/10 = -2.00$. Because 5% of the values are outside the range of two standard deviations from the mean and the normal distribution is symmetrical, $2\frac{1}{2}\%$ $\left(\frac{1}{2} \text{ of the } 5\%\right)$ are below the value of 30 and $2\frac{1}{2}\%$ are above the value of 70. Because a z score is the number of standard deviations an individual data value is from the mean, the empirical rule can be restated in terms of z scores.

Between $z = -1.00$ and $z = +1.00$ are approximately 68% of the values.
Between $z = -2.00$ and $z = +2.00$ are approximately 95% of the values.
Between $z = -3.00$ and $z = +3.00$ are approximately 99.7% of the values.

The topic of z scores is discussed more extensively in Chapter 6.

3.3.9 COEFFICIENT OF VARIATION

The **coefficient of variation** is a statistic that *is the ratio of the standard deviation to the mean expressed in percentage* and is denoted CV.

Coefficient of Variation

$$CV = \frac{\sigma}{\mu}(100)$$

The coefficient of variation essentially is a relative comparison of a standard deviation to its mean. The coefficient of variation can be useful in comparing standard deviations that have been computed from data with different means.

Suppose five weeks of average prices for stock A are 57, 68, 64, 71, and 62. To compute a coefficient of variation for these prices, first determine the mean and standard deviation: $\mu = 64.40$ and $\sigma = 4.84$. The coefficient of variation is:

$$CV_A = \frac{\sigma_A}{\mu_A}(100) = \frac{4.84}{64.40}(100) = .075 = 7.5\%$$

The standard deviation is 7.5% of the mean.

Sometimes financial investors use the coefficient of variation or the standard deviation or both as measures of risk. Imagine a stock with a price that never changes. An investor bears no risk of losing money from the price going down because no variability occurs in the price. Suppose, in contrast, that the price of the stock fluctuates wildly. An investor who buys at a low price and sells for a high price can make a nice profit. However, if the price drops below what the investor buys it for, the stock owner is subject to a potential loss. The greater the variability is, the more the potential for loss. Hence, investors use measures of variability such as standard deviation or coefficient of variation to determine the risk of a stock. What does the coefficient of variation tell us about the risk of a stock that the standard deviation does not?

Suppose the average prices for a second stock, B, over these same five weeks are 12, 17, 8, 15, and 13. The mean for stock B is 13.00 with a standard deviation of 3.03. The coefficient of variation can be computed for stock B as:

$$CV_B = \frac{\sigma_B}{\mu_B}(100) = \frac{3.03}{13}(100) = .233 = 23.3\%$$

The standard deviation for stock B is 23.3% of the mean.

With the standard deviation as the measure of risk, stock A is more risky over this period of time because it has a larger standard deviation. However, the average price of stock A is almost five times as much as that of stock B. Relative to the amount invested in stock A, the standard deviation of $4.84 may not represent as much risk as the standard deviation of $3.03 for stock B, which has an average price of only $13.00. The coefficient of variation reveals the risk of a stock in terms of the size of standard deviation relative to the size of the mean (in percentage). Stock B has a coefficient of variation that is nearly three times as much as the coefficient of variation for stock A. Using coefficient of variation as a measure of risk indicates that stock B is riskier.

The choice of whether to use a coefficient of variation or raw standard deviations to compare multiple standard deviations is a matter of preference. The coefficient of variation also provides an optional method of interpreting the value of a standard deviation.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

6. _____ is the average of the squared deviations about the arithmetic mean for a set of numbers.
7. _____ applies to all distributions regardless of their shape and can be applied to nonnormal distribution.
8. _____ is the measure of dispersion method which applied least for measuring the variability because it is affected by extreme values most.

State whether the following statements are true/false:

9. IQR is the range of the middle 50% of the data.
10. If Stock B has less standard deviation then Stock A then we can say that Stock A is less risky.

ACTIVITY

From bseindia.com download the share price (close) for the month of Jan(2018). You need to download the share price of Bank of Baroda and Punjab National Bank. Apply variance on the both of the dataset and evaluate that which bank is more risky to invest money.

3.4 MEASURES OF CENTRAL TENDENCY AND VARIABILITY: GROUPED DATA

Grouped data do not provide information about individual values. Hence, measures of central tendency and variability for grouped data must be computed differently from those for ungrouped or raw data.

3.4.1 MEASURES OF CENTRAL TENDENCY

Three measures of central tendency are presented here for grouped data: the mean, the median, and the mode.

MEAN

For ungrouped data, the mean is computed by summing the data values and dividing by the number of values. With grouped data, the specific values are unknown. What can be used to represent the data values? The midpoint of each class interval is used to represent all the values in a class interval. This midpoint is weighted by the frequency of values in that class interval. The mean for grouped data is then computed by summing the products of the class midpoint and the class frequency for each class and dividing that sum by the total number of frequencies. The formula for the mean of grouped data follows.

Mean of Grouped Data

$$\mu_{grouped} = \frac{\Sigma f_i M_i}{N} = \frac{\Sigma f_i M_i}{\Sigma f_i}$$

where

f_i = class frequency

N = total frequencies

M_i = class midpoint

Table 3.5 gives the frequency distribution of the unemployment rates of Canada from Table 2.2. To find the mean of these data, we need Σf_i and $\Sigma f_i M_i$. The value of Σf_i can be determined by summing the values in the frequency column. To calculate $\Sigma f_i M_i$, we must first determine the values of M_i or the class midpoints. Next we multiply each of these class midpoints by the frequency in that class interval, f_i, resulting in $f_i M_i$. Summing these values of $f_i M_i$ yields the value of $\Sigma f_i M_i$.

Table 3.6 contains the calculations needed to determine the group mean. The group mean for the unemployment data is 6.93. Remember that because each class interval was represented by its class midpoint rather than by actual values, the group mean is only approximate.

MEDIAN

The median for ungrouped or raw data is the middle value of an ordered array of numbers. For grouped data, solving for the median is considerably more complicated. The calculation of the median for grouped data is done by using the following formula.

TABLE 3.5: FREQUENCY DISTRIBUTION OF 60 YEARS OF UNEMPLOYMENT DATA FOR CANADA (GROUPED DATA)

Class Interval	Frequency	Cumulative Frequency
1–under 3	4	4
3–under 5	12	16
5–under 7	13	29
7–under 9	19	48
9–under 11	7	55
11–under 13	5	60

TABLE 3.6: CALCULATION OF GROUPED MEAN

Class Interval	Frequency (f_i)	Class Midpoint (M_i)	f_iM_i
1–under 3	4	2	8
3–under 5	12	4	48
5–under 7	13	6	78
7–under 9	19	8	152
9–under 11	7	10	70
11–under 13	5	12	60
	$\Sigma f_i = N = 60$		$\Sigma f_iM_i = 416$

$$\mu = \frac{\Sigma f_i M_i}{\Sigma f_i} = \frac{416}{60} = 6.93$$

Median of Grouped Data

$$\text{Median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

where

L = the lower limit of the median class interval

cf_p = a cumulative total of the frequencies up to but not including the frequency of the median class

f_{med} = the frequency of the median class

W = the width of the median class interval

N = total number of frequencies

The first step in calculating a grouped median is to determine the value of $N/2$, which is the location of the median term. Suppose we want to calculate the median for the frequency distribution data in Table 3.5. Since there are 60 values (N), the value of $N/2$ is $60/2 = 30$. The median is the 30th term. The question to ask is: where does the 30th term fall? This can be answered by determining the cumulative frequencies for the data, as shown in Table 3.5.

An examination of these cumulative frequencies reveals that the 30th term falls in the fourth class interval because there are only 29 values in the first three class intervals. Thus, the median value is in the fourth class interval somewhere between 7 and 9. The class interval containing the median value is referred to as the *median class interval.*

Since the 30th value is between 7 and 9, the value of the median must be at least 7. How much more than 7 is the median? The difference between the location of the median value, $N/2 = 30$, and the cumulative frequencies up to but not including the median class interval, $cf_p = 29$, tells how many values into the median class interval lies the value of the median. This is determined by solving for $N/2 - cf_p = 30 - 29 = 1$. The median value is located one value into the median class interval. However, there are 19 values in the median interval (denoted in the formula as f_{med}). The median value is 1/19 of the way through this interval.

$$\frac{\frac{N}{2} - cf_p}{f_{med}} = \frac{30-29}{19} = \frac{1}{19}$$

Thus, the median value is at least 7, the value of L, and is 1/19 of the way across the median interval. How far is it across the median interval? Each class interval is 2 units wide (w). Taking 1/19 of this distance tells us how far the median value is into the class interval.

$$\frac{\frac{N}{2} - cf_p}{f_{med}}(W) = \frac{\frac{60}{2} - 29}{19}(2) = \frac{1}{19}(2) = .105$$

Adding this distance to the lower endpoint of the median class interval yields the value of the median.

$$\text{Median} = 7 + \frac{\frac{60}{2} - 29}{19}(2) = 7 + \frac{1}{19}(2) = 7 + .105 = 7.105$$

The median value of unemployment rates for Canada is 7.105. Keep in mind that like the grouped mean, this median value is merely approximate. The assumption made in these calculations is that the actual values fall uniformly across the median class interval—which may or may not be the case.

MODE

The *mode* for grouped data is *the class midpoint of the modal class. The modal class is the class interval with the greatest frequency.* Using the data from Table 3.6, the 7–under 9 class interval contains the greatest frequency, 19. Thus, the modal class is 7–under 9. The class midpoint of this modal class

is 8. Therefore, the mode for the frequency distribution shown in Table 3.6 is 8. The modal unemployment rate is 8%.

3.4.2 MEASURES OF VARIABILITY

Two measures of variability for grouped data are presented here: the variance and the standard deviation. Again, the standard deviation is the square root of the variance. Both measures have original and computational formulas.

Formulas for Population Variance and Standard Deviation of Grouped Data

Original Formula	Computational Version
$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N}$	$\sigma^2 = \frac{\Sigma f_i M_i^2 - \frac{(\Sigma f_i M_i)^2}{N}}{N}$
$\sigma = \sqrt{\sigma^2}$	

where

f_i = frequency

M_i = class midpoint

$N = \Sigma f_i$, or total frequencies of the population

μ = grouped mean for the population

Formulas for Sample Variance and Standard Deviation of Grouped Data

Original Formula	Computational Version
$s^2 = \frac{\Sigma f_i (M_i - \bar{x})^2}{n-1}$	$s^2 = \frac{\Sigma f_i M_i^2 - \frac{(\Sigma f_i M_i)^2}{n}}{n-1}$
$s = \sqrt{s^2}$	

where

f_i = frequency

M_i = class midpoint

$n = \Sigma f_i$, or total of the frequencies of the sample

$\bar{x}$ = grouped mean for the sample

For example, let us calculate the variance and standard deviation of the Canadian unemployment data grouped as a frequency distribution in Table 3.5. If the data are treated as a population, the computations are as follows.

For the original formula, the computations are given in **Table 3.7**. The method of determining σ^2 and σ by using the computational formula is shown in **Table 3.8**. In either case, the variance of the unemployment data is 7.129 (squared percent), and the standard deviation is 2.67%. As with the computation of the grouped mean, the class midpoint is used

to represent all values in a class interval. This approach may or may not be appropriate, depending on whether the average value in a class is at the midpoint. If this situation does not occur, then the variance and the standard deviation are only approximations. Because grouped statistics are usually computed without knowledge of the actual data, the statistics computed potentially may be only approximations.

TABLE 3.7: CALCULATING GROUPED VARIANCE AND STANDARD DEVIATION WITH THE ORIGINAL FORMULA

Class Interval	f_i	M_i	f_iM_i	$(M_i - \mu)$	$(M_i - \mu)^2$	$f_i(M_i - \mu)^2$
1–under 3	4	2	8	−4.93	24.305	97.220
3–under 5	12	4	48	−2.93	8.585	103.020
5–under 7	13	6	78	−0.93	0.865	11.245
7–under 9	19	8	152	1.07	1.145	21.755
9–under 11	7	10	70	3.07	9.425	65.975
11–under 13	5	12	60	5.07	25.705	128.525
	$\Sigma f_i = N = 60$		$\Sigma f_iM_i = 416$			$\Sigma f_i(M_i - \mu)^2 = 427.740$

$$\mu = \frac{\Sigma f_i M_i}{\Sigma f_i} = \frac{416}{60} = 6.93$$

$$\sigma^2 = \frac{\Sigma f_i (M_i - \mu)^2}{N} = \frac{427.74}{60} = 7.129$$

$$\sigma = \sqrt{7.129} = 2.670$$

TABLE 3.8: CALCULATING GROUPED VARIANCE AND STANDARD DEVIATION WITH THE COMPUTATIONAL FORMULA

Class Interval	f_i	M_i	f_iM_i	$f_iM_i^2$
1–under 3	4	2	8	16
3–under 5	12	4	48	192
5–under 7	13	6	78	468
7–under 9	19	8	152	1216
9–under 11	7	10	70	700
11–under 13	5	12	60	720
	$\Sigma f_i = N = 60$		$\Sigma f_iM_i = 416$	$\Sigma f_iM_i^2 = 3312$

$$\sigma^2 = \frac{\Sigma f_i M_i^2 - \frac{(\Sigma f_i M_i)^2}{N}}{N} = \frac{3312 - \frac{416^2}{60}}{60} = \frac{3312 - 2884.27}{60} = \frac{427.73}{60} = 7.129$$

$$\sigma = \sqrt{7.129} = 2.670$$

DEMONSTRATION PROBLEM 3.6

Compute the mean, median, mode, variance, and standard deviation on the following sample data.

Class Interval	Frequency	Cumulative Frequency
10–under 15	6	6
15–under 20	22	28
20–under 25	35	63
25–under 30	29	92
30–under 35	16	108
35–under 40	8	116
40–under 45	4	120
45–under 50	2	122

Solution: The mean is computed as follows.

Class	f_i	M_i	f_iM_i
10–under 15	6	12.5	75.0
15–under 20	22	17.5	385.0
20–under 25	35	22.5	787.5
25–under 30	29	27.5	797.5
30–under 35	16	32.5	520.0
35–under 40	8	37.5	300.0
40–under 45	4	42.5	170.0
45–under 50	2	47.5	95.0
	$\Sigma f_i = n = 122$		$\Sigma f_iM_i = 3130.0$

$$\bar{x} = \frac{\Sigma f_iM_i}{\Sigma f_i} = \frac{3130}{122} = 25.66$$

The grouped mean is 25.66.

The grouped median is located at the 61st value (122/2). Observing the cumulative frequencies, the 61st value falls in the 20-under 25 class, making it the median class interval; thus, the grouped median is at least 20. Since there are 28 cumulative values before the median class interval, 33 more (61 − 28) are needed to reach the grouped median. However, there are 35 values in the median class. The grouped median is located 33/35 of the way across the class interval which has a width of 5. The grouped median is $20 + \frac{33}{35}(5) = 20 + 4.71 = 24.71$.

The grouped mode can be determined by finding the class midpoint of the class interval with the greatest frequency. The class with the greatest frequency is 20–under 25 with a frequency of 35. The midpoint of this class is 22.5, which is the grouped mode.

The variance and standard deviation can be found as shown next. First, use the original formula.

Class	f_i	M_i	$(M_i - \mu)$	$(M_i - \mu)^2$	$f_i(M_i - \mu)^2$
10–under 15	6	12.5	−13.16	173.19	1039.14
15–under 20	22	17.5	−8.16	66.59	1464.98
20–under 25	35	22.5	−3.16	9.99	349.65
25–under 30	29	27.5	1.84	3.39	98.31
30–under 35	16	32.5	6.84	46.79	748.64
35–under 40	8	37.5	11.84	140.19	1121.52
40–under 45	4	42.5	16.84	283.59	1134.36
45–under 50	2	47.5	21.84	476.99	953.98
	$\Sigma f_i = n = 122$				$\Sigma f_i(M_i - \bar{x})^2 = 6910.58$

$$s^2 = \frac{\sum f_i\left(M_i - \bar{x}\right)^2}{n-1} = \frac{6910.58}{121} = 57.11$$

$$s = \sqrt{57.11} = 7.56$$

Next, use the computational formula.

Class	f_i	M_i	f_iM_i	$f_iM_i^2$
10–under 15	6	12.5	75.0	937.50
15–under 20	22	17.5	385.0	6,737.50
20–under 25	35	22.5	787.5	17,718.75
25–under 30	29	27.5	797.5	21,931.25
30–under 35	16	32.5	520.0	16,900.00
35–under 40	8	37.5	300.0	11,250.00
40–under 45	4	42.5	170.0	7,225.00
45–under 50	2	47.5	95.0	4,512.50
	$\Sigma f_i = n = 122$		$\Sigma f_iM_i = 3,130.0$	$\Sigma f_iM_i^2 = 87,212.50$

$$s^2 = \frac{\sum f_iM_i^2 - \frac{\left(\sum f_iM_i\right)^2}{n}}{n-1} = \frac{87,212.5 - \frac{(3,130)^2}{122}}{121} = \frac{6,910.04}{121} = 57.11$$

$$s = \sqrt{57.11} = 7.56$$

The sample variance is 57.11 and the standard deviation is 7.56.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

11. _____ can be determined by finding the class midpoint of the class interval with the greatest frequency.
12. The standard deviation is the square root of the _____.

State whether the following statements are true/false:

13. We need to calculate cumulative frequency for grouped median.
14. For grouped median, we need to calculate median class and then the mean of the median class.
15. Grouped statistics are usually computed with knowledge of the actual data.

ACTIVITY

From the activity under section 3.2, you need to take the Ministry of Communications, Department of Telecommunications (DOT), Telecom Regulatory Authority of India (TRAI) data. You need to calculate the grouped mean and standard deviation for all companies under 3G technology for North East.

3.5 MEASURES OF SHAPE

Measures of shape are *tools that can be used to describe the shape of a distribution of data.* In this section, we examine two measures of shape, skewness and kurtosis. We also look at box-and-whisker plots.

3.5.1 SKEWNESS

A distribution of data in which the right half is a mirror image of the left half is said to be *symmetrical*. One example of a symmetrical distribution is the normal distribution, or bell curve, shown in Figure 3.8 and presented in more detail in Chapter 6.

Skewness is when *a distribution is asymmetrical or lacks symmetry.* The distribution in **Figure 3.8** has no skewness because it is symmetric. **Figure 3.9** shows a distribution that is skewed left, or negatively skewed, and **Figure 3.10** shows a distribution that is skewed right, or positively skewed.

The skewed portion is the long, thin part of the curve. Many researchers use a skewed distribution to denote that the data are sparse at one end of the distribution and piled up at the other end. Instructors sometimes refer to a grade distribution as skewed, meaning that few students scored at one end of the grading scale, and many students scored at the other end.

Figure 3.8: Symmetrical Distribution

Figure 3.9: Distribution Skewed Left, or Negatively Skewed

Figure 3.10: Distribution Skewed Right, or Positively Skewed

SKEWNESS AND THE RELATIONSHIP OF THE MEAN, MEDIAN, AND MODE

The concept of skewness helps to understand the relationship of the mean, median, and mode. In a unimodal distribution (distribution with a single peak or mode) that is skewed, the mode is the apex (high point) of the curve and the median is the middle value. The mean tends to be located toward the tail of the distribution, because the mean is particularly affected by the extreme values. A bell-shaped or normal distribution with the mean, median, and mode all at the center of the distribution has no skewness. **Figure 3.11** displays the relationship of the mean, median, and mode for different types of skewness.

Figure 3.11: Relationship of Mean, Median, and Mode

3.5.2 KURTOSIS

Kurtosis *describes the amount of peakedness of a distribution.* Distributions that are high and thin are referred to as **leptokurtic** distributions. Distributions that are flat and spread out are referred to as **platykurtic** distributions. Between these two types are distributions that are more "normal" in shape, referred to as **mesokurtic** distributions. These three types of kurtosis are illustrated in **Figure 3.12.**

Figure 3.12: Types of Kurtosis

3.5.3 BOX-AND-WHISKER PLOTS AND FIVE-NUMBER SUMMARY

Another way to describe a distribution of data is by using a box and whisker plot. A **box-and-whisker plot**, sometimes called a *box plot*, is *a diagram that utilizes the upper and lower quartiles along with the median and the two most extreme values to depict a distribution graphically.* The plot is constructed

by using a box to enclose the median. This *box* is extended outward from the median along a continuum to the lower and upper quartiles, enclosing not only the median but also the middle 50% of the data. From the lower and upper quartiles, lines referred to as *whiskers* are extended out from the box toward the outermost data values. The box-and-whisker plot is determined from five specific numbers sometimes referred to as the **five-number summary.**

1. The median (Q_2)
2. The lower quartile (Q_1)
3. The upper quartile (Q_3)
4. The smallest value in the distribution
5. The largest value in the distribution

The box of the plot is determined by locating the median and the lower and upper quartiles on a continuum. A box is drawn around the median with the lower and upper quartiles (Q_1 and Q_3) as the box endpoints. These box endpoints (Q_1 and Q_3) are referred to as the *hinges* of the box.

Next the value of the interquartile range (IQR) is computed by $Q_3 - Q_1$. The interquartile range includes the middle 50% of the data and should equal the length of the box. However, here the interquartile range is used outside of the box also. At a distance of 1.5 . IQR outward from the lower and upper quartiles are what are referred to as *inner fences*. A *whisker*, a line segment, is drawn from the lower hinge of the box outward to the smallest data value. A second whisker is drawn from the upper hinge of the box outward to the largest data value. The inner fences are established as follows.

$$Q_1 - 1.5 \cdot \text{IQR}$$
$$Q_3 + 1.5 \cdot \text{IQR}$$

If data fall beyond the inner fences, then *outer* fences can be constructed:

$$Q_1 - 3.0 \cdot \text{IQR}$$
$$Q_3 + 3.0 \cdot \text{IQR}$$

Figure 3.13 shows the features of a box-and-whisker plot.

Data values outside the mainstream of values in a distribution are viewed as *outliers*. Outliers can be merely the more extreme values of a data set. However, sometimes outliers occur due to measurement or recording errors. Other times they are values so unlike the other values that they should not be considered in the same analysis as the rest of the distribution. Values in the data distribution

Figure 3.13: Box-and-Whisker Plot

that are outside the inner fences but within the outer fences are referred to as *mild outliers*. Values that are outside the outer fences are called *extreme outliers*. Thus, one of the main uses of a box-and-whisker plot is to identify outliers. In some computer-produced box-and-whisker plots, the whiskers are drawn to the largest and smallest data values within the inner fences. An asterisk is then printed for each data value located between the inner and outer fences to indicate a mild outlier. Values outside the outer fences are indicated by a zero on the graph. These values are extreme outliers.

Another use of box-and-whisker plots is to determine whether a distribution is skewed. The location of the median in the box can relate information about the skewness of the middle 50% of the data. If the median is located on the right side of the box, then the middle 50% are skewed to the left. If the median is located on the left side of the box, then the middle 50% are skewed to the right. By examining the length of the whiskers on each side of the box, a business researcher can make a judgment about the skewness of the outer values. If the longest whisker is to the right of the box, then the outer data are skewed to the right and vice versa. We shall use the data given in **Table 3.9** to construct a box-and-whisker plot.

After organizing the data into an ordered array, as shown in **Table 3.10**, it is relatively easy to determine the values of the lower quartile (Q_1), the median, and the upper quartile (Q_3). From these, the value of the interquartile range can be computed.

The hinges of the box are located at the lower and upper quartiles, 69 and 80.5. The median is located within the box at distances of 4 from the lower quartile and 7.5 from the upper quartile. The distribution of the middle 50% of the data is skewed right, because the median is nearer to the lower or left hinge. The inner fence is constructed by

$$Q_1 - 1.5 \cdot \text{IQR} = 69 - 1.5(11.5) = 69 - 17.25 = 51.75$$

and

$$Q_3 + 1.5 \cdot \text{IQR} = 80.5 + 1.5(11.5) = 80.5 + 17.25 = 97.75$$

The whiskers are constructed by drawing a line segment from the lower hinge outward to the smallest data value and a line segment from the upper hinge outward to the largest data value. An examination of the data reveals that no data values in this set of numbers are outside the inner fence.

TABLE 3.9: DATA FOR BOX-AND-WHISKER PLOT

71	87	82	64	72	75	81	69
76	79	65	68	80	73	85	71
70	79	63	62	81	84	77	73
82	74	74	73	84	72	81	65
74	62	64	68	73	82	69	71

TABLE 3.10: DATA IN ORDERED ARRAY WITH QUARTILES AND MEDIAN

62	62	63	64	64	65	65	68	68	69
69	70	71	71	71	72	72	73	73	73
73	74	74	74	75	76	77	79	79	80
81	81	81	82	82	82	84	84	85	87

$Q_1 = 69$
$Q_2 = \text{median} = 73$
$Q_3 = 80.5$
$\text{IQR} = Q_3 - Q_1 = 80.5 - 69 = 11.5$

The whiskers are constructed outward to the lowest value, which is 62, and to the highest value, which is 87.

To construct an outer fence, we calculate $Q_1 - 3 \cdot IQR$ and $Q_3 + 3 \cdot IQR$, as follows,

$$Q_1 - 3 \cdot IQR = 69 - 3(11.5) = 69 - 34.5 = 34.5$$

$$Q_3 + 3 \cdot IQR = 80.5 + 3(11.5) = 80.5 + 34.5 = 115.0$$

Figure 3.14 is the Computer printout for this box-and-whisker plot.

Figure 3.14: Box-and-Whisker Plot

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

16. Values that are outside the outer fences are called_____.
17. A unimodal distribution is a distribution with a _____ peak.
18. Distributions that are flat and spread out are referred to as_____ distributions.

State whether the following statements are true/false:

19. Box and whisker plot is the way to describe a distribution of data.
20. Skewness is when a distribution is symmetrical.

ACTIVITY

https://people.richland.edu/james/spring05/m113/activities/act01-data.html.

From the above mentioned link copy the data and calculate the coefficient of skewness for the time variable. Try to calculate without the help of spreadsheet.

3.6 SUMMARY

- Statistical descriptive measures include measures of central tendency, measures of variability, and measures of shape.
- Measures of central tendency and measures of variability are computed differently for ungrouped and grouped data. Measures of central tendency are useful in describing data because they communicate information about the more central portions of the data. The most common measures of central tendency are the mode, median, and mean. In addition, percentiles and quartiles are measures of central tendency.
- The mode is the most frequently occurring value in a set of data. Among other things, the mode is used in business for determining sizes.
- The median is the middle term in an ordered array of numbers containing an odd number of terms. For an array with an even number of terms, the median is the average of the two middle terms. A median is unaffected by the magnitude of extreme values. This characteristic makes the median a most useful and appropriate measure of location in reporting such things as income, age, and prices of houses.
- The arithmetic mean is widely used and is usually what researchers are referring to when they use the word *mean*. The arithmetic mean is the average. The population mean and the sample mean are computed in the same way but are denoted by different symbols. The arithmetic mean is affected by every value and can be inordinately influenced by extreme values.
- Percentiles divide a set of data into 100 groups. There are 99 percentiles. Quartiles divide data into four groups. The three quartiles are Q_1, which is the lower quartile; Q_2, which is the middle quartile and equals the median; and Q_3, which is the upper quartile.
- Measures of variability are statistical tools used in combination with measures of central tendency to describe data. Measures of variability provide information about the spread of the data values. These measures include the range, mean absolute deviation, variance, standard deviation, interquartile range, z scores, and coefficient of variation for ungrouped data.
- One of the most elementary measures of variability is the range. It is the difference between the largest and smallest values. Although the range is easy to compute, it has limited usefulness. The interquartile range is the difference between the third and first quartile. It equals the range of the middle 50% of the data.
- The mean absolute deviation (MAD) is computed by averaging the absolute values of the deviations from the mean. The mean absolute deviation has limited usage in statistics but is occasionally used in the field of forecasting as a measure of error.
- Variance is widely used as a tool in statistics but is used little as a stand-alone measure of variability. The variance is the average of the squared deviations about the mean.
- The square root of the variance is the standard deviation. It also is a widely used tool in statistics, but it is used more often than the variance as a stand-alone measure. The standard deviation is best understood by examining its applications in determining where data are in relation to the mean. The empirical rule and Chebyshev's theorem are statements about the proportions of data values that are within various numbers of standard deviations from the mean.

- The empirical rule reveals the percentage of values that are within one, two, or three standard deviations of the mean for a set of data. The empirical rule applies only if the data are in a bell-shaped distribution.
- Chebyshev's theorem also delineates the proportion of values that are within a given number of standard deviations from the mean. However, it applies to any distribution.
- The z score represents the number of standard deviations a value is from the mean for normally distributed data.
- The coefficient of variation is a ratio of a standard deviation to its mean, given as a percentage. It is especially useful in comparing standard deviations or variances that represent data with different means.
- Some measures of central tendency and some measures of variability are presented for grouped data. These measures include mean, median, mode, variance, and standard deviation. Generally, these measures are only approximate for grouped data because the values of the actual raw data are unknown.
- Two measures of shape are skewness and kurtosis. Skewness is the lack of symmetry in a distribution. If a distribution is skewed, it is stretched in one direction or the other. The skewed part of a graph is its long, thin portion.
- Kurtosis is the degree of peakedness of a distribution. A tall, thin distribution is referred to as leptokurtic. A flat distribution is platykurtic, and a distribution with a more normal peakedness is said to be mesokurtic.
- A box-and-whisker plot is a graphical depiction of a distribution. The plot is constructed by using the median, the lower quartile, the upper quartile, the smallest value, and the largest value. It can yield information about skewness and outliers.

KEY WORDS

1. **Distribution:** A distribution is the arrangement of data by the values of one variable in order, from low to high. This arrangement, and its characteristics such as shape and spread, provide information about the underlying sample.
2. **Mean:** The arithmetic mean is simply the average add up all the scores and divide by the number of observations.
3. **Median:** The median is defined as the "middle score." It is the observed score (or an extrapolation of observed scores) that splits the distribution in half so that 50% of the remaining scores are less than the median and 50% of the remaining scores exceed the median.
4. **Mode:** The mode is the most frequent score in the set. Multimodal refers to a distribution with more than one mode; bimodal refers to a distribution with 2 modes.
5. **Skew:** When there are more scores toward one end of the distribution than the other, this results in skew.

KEY WORDS

6. **Range:** The range is defined as the difference between the largest and the smallest score.
7. **Variance:** Variance is the statistical average of the dispersion of scores in a distribution.
8. **Standard deviation:** The standard deviation of a distribution is the average deviation between individual distribution scores and the distribution's mean.
9. **Interquartile range (IQR):** The IQR is the difference between the score delineating the 75th percentile and the 25th percentile, the third and first quartiles, respectively.
10. **Percentile:** The pth percentile is that score such that p percent of the observations score lower or equal to the score.
11. **Population variance:** The population variance is defined as the average squared deviation from the mean and in many statistical procedures is also called a mean square or MS.

3.7 DESCRIPTIVE QUESTIONS

3.1. The 2010 U.S. Census also asked for each person's age. Suppose that a sample of 40 households taken from the census data showed the age of the first person recorded on the census form to be as follows.

42	29	31	38	55	27	28
33	49	70	25	21	38	47
63	22	38	52	50	41	19
22	29	81	52	26	35	38
29	31	48	26	33	42	58
40	32	24	34	25		

Compute P_{10}, P_{80}, Q_1, Q_3, the interquartile range, and the range for these data.

3.2. Shown here are the top twelve biggest oil and gas companies in the world according to *Forbes*. Use these as population data and answer the questions that follow.

Company	Production Volume (million barrels per day)
Saudi Aramco (Saudi Arabia)	12.5
Gazprom (Russia)	9.7
NIOC (Iran)	6.4
ExxonMobil Corp. (USA)	5.3
PetroChina (China)	4.4
BP (UK)	4.1
Royal Dutch/Shell (NL/UK)	3.9

Company	Production Volume (million barrels per day)
Pemex (Mexico)	3.6
Chevron Corp. (USA)	3.5
KPC (Kuwait)	3.2
ADNOC (UAE)	2.9
Sonatrach (Algeria)	2.7

(a) What are the values of the mean and the median? Compare the answers and state which you prefer as a measure of location for these data and why.

(b) What are the values of the range and interquartile range? How do they differ?

(c) What are the values of variance and standard deviation for these data?

(d) What is the z score for ADNOC? What is the z score forExxonMobil? Interpret these z scores.

3.3. The U.S. Department of the Interior releases figures on mineral production. Following are the 15 leading states in nonfuel mineral production in the United States for 2014.

State	Value ($ Billions)
Arizona	8.06
Nevada	7.49
Minnesota	4.71
Texas	4.24
Utah	4.18
California	3.51
Alaska	3.51
Florida	2.99
Missouri	2.48
Michigan	2.41
Wyoming	2.35
Colorado	2.32
New Mexico	1.87
Wisconsin	1.78
Georgia	1.60

(a) Calculate the mean and median.

(b) Calculate the range, interquartile range, mean absolute deviation, sample variance, and sample standard deviation.

(c) Sketch a box-and-whisker plot.

3.4. A research agency administers a demographic survey to 90 telemarketing companies to determine the size of their operations. When asked to report how many employees now work in their telemarketing operation, the companies gave responses ranging from 1 to 100. The agency's analyst organizes the figures into a frequency distribution.

Number of Employees Working in Telemarketing	Number of Companies
0–under 20	32
20–under 40	16
40–under 60	13
60–under 80	10
80–under 100	19

(a) Compute the mean, median, and mode for this distribution.

(b) Compute the sample standard deviation for these data.

3.5. The Polk Company reported that the average age of a car on U.S. roads in a recent year was 7.5 years. Suppose the distribution of ages of cars on U.S. roads is approximately bellshaped. If 99.7% of the ages are between 1 year and 14 years, what is the standard deviation of car age? Suppose the standard deviation is 1.7 years and the mean is 7.5 years. Between what two values would 95% of the car ages fall?

3.6. According to a *Human Resources* report, a worker in the industrial countries spends on average 419 minutes a day on the job. Suppose the standard deviation of time spent on the job is 27 minutes.

(a) If the distribution of time spent on the job is approximately bell shaped, between what two times would 68% of the figures be? 95%? 99.7%?

(b) If the shape of the distribution of times is unknown, approximately what percentage of the times would be between 359 and 479 minutes?

(c) Suppose a worker spent 400 minutes on the job. What would that worker's z score be, and what would it tell the researcher?

3.7. According to the Bureau of Labor Statistics, the average annual salary of a worker in Detroit, Michigan, is $45,140. Suppose the median annual salary for a worker in this group is $41,369 and the mode is $39,500. Is the distribution of salaries for this group skewed? If so, how and why? Which of the measures of central tendency would you use to describe these data? Why?

3.8 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

3.1. P_{10}: $i = \frac{10}{100}(40) = 4$

P_{10} = 4.5th term = 23

P_{80}: $i = \frac{80}{100}(40) = 32$

P_{80} = 32.5th term = 49.5

$Q_1 = P_{25}$: $i = \frac{25}{100}(40) = 10$

P_{25} = 10.5th term = 27.5

$Q_3 = P_{75}$: $i = \frac{75}{100}(40) = 30$

P_{75} = 30.5th term = **47.5**

IQR = $Q_3 - Q_1$ = 47.5 – 27.5 = **20**

Range = 81 – 19 = **62**

3.2. **(a)** $\mu = \frac{\sum x}{N} = \frac{62.2}{12} = 5.183$

The median is located at the $\frac{n+1}{2}$th value = 13/2 = 6.5th value of an ascending array. It is the average of the 6th term and the 7th term:

$$\text{Median} = \frac{3.9 + 4.1}{2} = 4.0$$

(b) Range = 12.5 – 2.7 = **9.8**

$Q_3 = 5.85$ $Q_1 = 3.35$ IQR = $Q_3 - Q_1$ = **2.5**

(c) Variance:

$$\sigma^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N} = \frac{421.92 - \frac{62.2^2}{12}}{12} = 8.293$$

Standard Deviation:

$$\sigma = \sqrt{\sigma^2} = \sqrt{(8.293)} = 2.880$$

(d) ADNOC:

$$z = \frac{x - \mu}{\sigma} = \frac{2.9 - 5.183}{2.880} = -0.79$$

ExxonMobil:

$$z = \frac{x - \mu}{\sigma} = \frac{5.3 - 5.183}{2.880} = 0.04$$

3.3. **(a) Mean:** $\bar{x} = \frac{\sum x}{N} = \frac{53.5}{15} = 3.57$

Median: The median is located at the $\frac{n+1}{2}$th value = (15+1)/2 = 8th value of an ascending array. Thus, the median is **2.99**

(b) Range: 8.06 – 1.60 = **6.46**

Q_1: ¼(15) = 3.75 Located at the 4th term. Q_1 = 2.32

Q_3: ¾(15) = 11.25 Located at the 12th term. Q_3 = 4.24

IQR = $Q_3 - Q_1$ = 4.24 – 2.32 = **1.92**

x	$\lvert x - \bar{x} \rvert$	$(x - \bar{x})^2$
8.06	4.49	20.1601
4.71	1.14	1.2996
4.24	0.67	0.4489
4.18	0.61	0.3721
3.51	0.06	0.0036
2.99	0.58	0.3364
2.48	1.09	1.1881
2.41	1.16	1.3456
2.35	1.22	1.4884
2.32	1.25	1.5625
1.87	1.70	2.8900
1.78	1.79	3.2041
1.60	1.97	3.8809
	$\sum \lvert x - \bar{x} \rvert = 21.71$	$\sum (x - \bar{x})^2 = 53.5503$

$$\text{MAD} = \frac{\sum \lvert x - \bar{x} \rvert}{n} = \frac{21.71}{15} = 1.45$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{53.5503}{14} = 3.8250$$

$$s = \sqrt{s^2} = \sqrt{3.8250} = 1.9558$$

(c) Use Q_1 = 2.32, Q_2 = 2.99, Q_3 = 4.24, IQR = 1.92

Extreme Points: 7.49 and 8.06

Inner Fences: 2.32 – 1.5(1.92) = –0.56

4.24 + 1.5(1.92) = 7.12

Outer Fences: 2.32 – 3.0(1.92) = –3.44

4.24 + 3.0(1.92) = 10.0

The uppermost two points are mild outliers as shown in the boxplot below.

3.4.

Class	f	M	fM	fM^2
0 – 20	32	10	320	3,200
20 – 40	16	30	480	14,400
40 – 60	13	50	650	32,500
60 – 80	10	70	700	49,000
80 – 100	19	90	1,710	153,900
	$\Sigma f = 90$		$\Sigma fM = 3{,}860$	$\Sigma fM^2 = 253{,}000$

(a) Mean: $\bar{x} = \dfrac{\Sigma fM}{n} = \dfrac{\Sigma fM}{\Sigma} = \dfrac{3{,}860}{90} = \mathbf{42.89}$

Median: $20 + \dfrac{45-32}{16}(20) = 20 + \dfrac{13}{16}(20) = 20 + 16.25 = 36.25$

Mode: The Modal Class is 0-20. The midpoint of this class is the mode = **10.**

(b) Sample Standard Deviation:

$$s = \sqrt{\frac{\Sigma fM^2 - \dfrac{(\Sigma fM)^2}{n}}{n-1}} = \sqrt{\frac{253{,}000 - \dfrac{(3860)^2}{90}}{89}}$$

$$= \sqrt{\frac{253{,}000 - 165{,}551.1}{89}}$$

$$= \sqrt{\frac{87{,}448.9}{89}} = \sqrt{982.572} = \mathbf{31.346}$$

3.5. $\mu = 7.5$ Each of the numbers, 1 and 14, are 6.5 units away from the mean.

From the Empirical Rule: 99.7% of the values lie in $\mu \pm 3\sigma$

$3\sigma = 14 - 7.5 = 6.5$ Solving for 3s = 6.5 for $\sigma \cdot \sigma = \mathbf{2.167}$

Suppose that $\mu = 7.5$, $\sigma = 1.7$.

95% lie within $\mu \pm 2\sigma = 7.5 \pm 2(1.7) = 7.5 \pm 3.4$

Between 4.1 and 10.9 lie 95% of the values.

3.6. $\mu = 419$, $\sigma = 27$

(a) 68%: $\mu \pm 1\sigma$ 419 ± 27 **392 to 446**

95%: $\mu \pm 2\sigma$ $419 \pm 2(27)$ **365 to 473**

99.7%: $\mu \pm 3\sigma$ $419 \pm 3(27)$ **338 to 500**

(b) Use Chebyshev's Theorem:

Each of the points, 359 and 479 is a distance of 60 from the mean, $\mu = 419$.

k = (distance from the mean)/σ = 60/27 = 2.22

Proportion = $1 - 1/k^2 = 1 - 1/(2.22)^2 = .797 =$ **79.7%**

(c) Since $x = 400$, $z = \dfrac{400 - 419}{27} = \mathbf{-0.704}$. This worker is in the lower half of workers but within one standard deviation of the mean.

3.7. Mean $45,140

Median $41,369

Mode $39,500

Since these three measures are not equal, the distribution is skewed. The distribution is skewed to the right because the mean is greater than the median. Often, the median is preferred in reporting income data because it yields information about the middle of the data while ignoring extremes.

3.9 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
3.2 Measures of Central Tendency: Ungrouped Data	1.	median
	2.	multimodal
	3.	75 percentile
	4.	True
	5.	False
3.3 Measures of Variability: Ungrouped Data	6.	Variance
	7.	Chebyshev's theorem
	8.	Range
	9.	True
	10.	False
3.4 Measures of Central Tendency and Variability: Grouped Data	11.	Grouped Mode (Mode)
	12.	variance
	13.	True
	14.	True
	15.	False
3.5 Measures of Shape	16.	extreme Outliers
	17.	single
	18.	platykurtic
	19.	True
	20.	False

4

CHAPTER

PROBABILITY

CONTENTS

4.1 Introduction
4.2 Probability
4.2.1 Classical Method of Assigning Probabilities
4.2.2 Relative Frequency of Occurrence
4.2.3 Subjective Probability
Self Assessment Questions
Activity
4.3 Structure of Probability
4.3.1 Experiment
4.3.2 Event
4.3.3 Elementary Events
4.3.4 Sample Space
4.3.5 Unions and Intersections
4.3.6 Mutually Exclusive Events
4.3.7 Independent Events
4.3.8 Collectively Exhaustive Events
4.3.9 Complementary Events
4.3.10 Counting the Possibilities
Self Assessment Questions
Activity
4.4 Marginal, Union, Joint, and Conditional Probabilities
Self Assessment Questions
Activity
4.5 Addition Laws
4.5.1 Joint Probability Tables
4.5.2 Special Law of Addition
Self Assessment Questions
Activity
4.6 Multiplication Laws
4.6.1 General Law of Multiplication
4.6.2 Special Law of Multiplication

LEARNING OBJECTIVES

The main objective of Chapter 4 is to help you understand the basic principles of probability, thereby enabling you to:

- Understand the concepts of probability and differentiate among the three methods of assigning probabilities.
- Define the elements of probability such as experiments, sample spaces, and events.
- Classify mutually exclusive, collectively exhaustive, complementary, and independent events.
- Calculate probabilities using the general law of addition, along with a joint probability table, the complement of a union, and the special law of addition.
- Calculate joint probabilities of both independent and dependent events using the general and special laws of multiplication.
- Understand and calculate the conditional probabilities with various forms of law of conditional probability.
- Calculate conditional probabilities using Bayes' rule.

4.1 INTRODUCTION

In business, most decision making involves uncertainty. For example, an operations manager does not know definitely whether a valve in the plant is going to malfunction or continue to function—or, if it continues, for how long. When should it be replaced? What is the chance that the valve will malfunction within the next week? In the banking industry, what are the new vice president's prospects for successfully turning a department around? The answers to these questions are uncertain.

In the case of a high-rise building, what are the chances that a fire-extinguishing system will work when needed if redundancies are built in? Business people must address these and thousands of similar questions daily. Because most such questions do not have definite answers, the decision making is based on uncertainty. In many of these situations, a probability can be assigned to the likelihood of an outcome. This chapter is about learning how to determine or assign probabilities.

4.2 PROBABILITY

Chapter 1 discussed the difference between descriptive and inferential statistics. Much statistical analysis is inferential, and probability is the basis for inferential statistics. Recall that inferential statistics involves taking a sample from a population, computing a statistic on the sample, and inferring from the statistic the value of the corresponding parameter of the population. The reason for doing so is that the value of the parameter is unknown. Because it is unknown, the analyst conducts the inferential process under uncertainty. However, by applying rules and laws, the analyst can often assign a probability of obtaining the results. **Figure 4.1** depicts this process.

Suppose a quality control analyst for a light bulb manufacturing company randomly selects a sample of 40 light bulbs from a population of brand X bulbs and burns each of them until they burn out. The number of hours of luminance for each bulb is recorded; from the data, the average bulb life is determined to be 1,000 hours. Using techniques presented later in this text, the analyst uses this figure as an estimate of the average number of hours of bulb life for the *population* of brand X bulbs. Because the light bulbs being analyzed are only a sample of the population, the average number of hours of luminance for the 40 bulbs may or may not accurately estimate the average for all bulbs in the population. The results are uncertain. By applying the laws presented in this chapter, the inspector can assign a value of probability to this estimate.

Figure 4.1: Probability in the Process of Inferential Statistics

In addition, probabilities are used directly in certain industries and industry applications. For example, the insurance industry uses probabilities in actuarial tables to determine the likelihood of certain outcomes in order to set specific rates and coverages. The gaming industry uses probability values to establish charges and payoffs. In other industries, such as manufacturing and aerospace, it is important to know the life of a mechanized part and the probability that it will malfunction at any given length of time in order to protect the firm from major breakdowns.

The three general methods of assigning probabilities are (1) the classical method, (2) the relative frequency of occurrence method, and (3) subjective probabilities.

4.2.1 CLASSICAL METHOD OF ASSIGNING PROBABILITIES

When probabilities are assigned based on laws and rules, the method is referred to as the **classical method of assigning probabilities**. This method involves an experiment, which is *a process that produces outcomes*, and an event, *which is an outcome of an experiment.*

When we assign probabilities using the classical method, the probability of an individual event occurring is determined as the ratio of the number of items in a population containing the event (n_e) to the total number of items in the population (N). That is, $P(E) = n_e/N$. For example, if a company has 200 workers and 70 are female, the probability of randomly selecting a female from this company is 70/200 = .35.

Classical Method of Assigning Probabilities

$$P(E) = \frac{n_e}{N}$$

where

N = total possible number of outcomes of an experiment

n_e = the number of outcomes in which the event occurs out of N outcomes

Suppose, in a particular plant, three machines make a given product. Machine A always produces 40% of the total number of this product. Ten percent of the items produced by machine A are defective. If the finished products are well mixed with regard to which machine produced them and if one of these products is randomly selected, the classical method of assigning probabilities tells us that the probability that the part was produced by machine A and is defective is .04. This probability can be determined even before the part is sampled because with the classical method, the probabilities can be determined **a priori**; that is, *they can be determined prior to the experiment.*

Because n_e can never be greater than N (no more than N outcomes in the population could possibly have attribute e), the highest value of any probability is 1. If the probability of an outcome occurring is 1, the event is certain to occur. The smallest possible probability is 0. If none of the outcomes of the N possibilities has the desired characteristic, e, the probability is $0/N = 0$, and the event is certain not to occur.

Range of Possible Probabilities

$$0 \le P(E) \le 1$$

Thus, probabilities are nonnegative proper fractions or nonnegative decimal values greater than or equal to 0 and less than or equal to 1.

Probability values can be converted to percentages by multiplying by 100. Meteorologists often report weather probabilities in percentage form. For example, when they forecast a 60% chance of rain for tomorrow, they are saying that the probability of rain tomorrow is .60.

4.2.2 RELATIVE FREQUENCY OF OCCURRENCE

The **relative frequency of occurrence method** of assigning probabilities is based on cumulated historical data. With this method, *the probability of an event occurring is equal to the number of times the event has occurred in the past divided by the total number of opportunities for the event to have occurred.*

Probability by Relative Frequency of Occurrence

$$\frac{\text{Number of times an event occurred}}{\text{Total number of opportunities for the event to have occurred}}$$

Relative frequency of occurrence is not based on rules or laws but on what has occurred in the past. For example, a company wants to determine the probability that its inspectors are going to reject the next batch of raw materials from a supplier. Data gathered from company record books show that the supplier sent the company 90 batches in the past, and inspectors rejected 10 of them. By the method of relative frequency of occurrence, the probability of the inspectors rejecting the next batch is 10/90, or .11. If the next batch is rejected, the relative frequency of occurrence probability for the subsequent shipment would change to 11/91 = .12.

4.2.3 SUBJECTIVE PROBABILITY

The **subjective method** of *assigning probability is based on the feelings or insights of the person determining the probability.* Subjective probability comes from the person's intuition or reasoning. Although not a scientific approach to probability, the subjective method often is based on the accumulation of knowledge, understanding, and experience stored and processed in the human mind. At times it is merely a guess. At other times, subjective probability can potentially yield accurate probabilities. Subjective probability can be used to capitalize on the background of experienced workers and managers in decision making.

Suppose a director of transportation for an oil company is asked the probability of getting a shipment of oil out of Saudi Arabia to the India within three weeks. A director who has scheduled many such shipments, has a knowledge of Saudi politics, and has an awareness of current climatological and economic conditions may be able to give an accurate probability that the shipment can be made on time.

Subjective probability also can be a potentially useful way of tapping a person's experience, knowledge, and insight and using them to forecast the occurrence of some event. An experienced airline mechanic can usually assign a meaningful probability that a particular plane will have a certain type of mechanical difficulty. Physicians sometimes assign subjective probabilities to the life expectancy of people who have cancer.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. The _____ of assigning probability is based on the feelings or insights of the person determining the probability.
2. _____ is a process that produces outcomes.
3. The _____ of occurrence method of assigning probabilities is based on cumulated historical data.

State whether the following statements are true/false:

4. An event is an outcome of an experiment.
5. The probability of any event can be –1 and 1, where –1 is no probability and 1 means 100% probability.

ACTIVITY

From the below mentioned link, copy the dataset for batting of Indian team. Calculate the probability.

1. What will be the probability that Virat Kohli will make a run on the next ball?
2. What will be the probability that Ravindra Jadeja will take wicket on the next ball?

https://knoema.com/rbkxcsd/2013-best-performance-in-one-day-international-cricket

4.3 STRUCTURE OF PROBABILITY

In the study of probability, developing a language of terms and symbols is helpful. The structure of probability provides a common framework within which the topics of probability can be explored.

4.3.1 EXPERIMENT

As previously stated, an **experiment** is *a process that produces outcomes.* Examples of business-oriented experiments with outcomes that can be statistically analyzed might include the following.

- Interviewing 20 randomly selected consumers and asking them which brand of appliance they prefer
- Sampling every 200th bottle of ketchup from an assembly line and weighing the contents
- Testing new pharmaceutical drugs on samples of cancer patients and measuring the patients' improvement
- Auditing every 10th account to detect any errors

4.3.2 EVENT

Because an **event** is *an outcome of an experiment*, the experiment defines the possibilities of the event. If the experiment is to sample five bottles coming off a production line, an event could be to get one defective and four good bottles. In an experiment to roll a die, one event could be to roll an even number and another event could be to roll a number greater than two. Events are denoted by uppercase letters; italic capital letters (e.g., A and $E_1, E_2, \ldots$) represent the general or abstract case, and roman capital letters (e.g., H and T for heads and tails) denote specific things and people.

4.3.3 ELEMENTARY EVENTS

Events that cannot be decomposed or broken down into other events are called **elementary events**. Elementary events are denoted by lowercase letters (e.g., $e_1, e_2, e_3, \ldots$). Suppose the experiment is to roll a die. The elementary events for this experiment are to roll a 1 or roll a 2 or roll a 3, and so on.

Rolling an even number is an event, but it is not an elementary event because the even number can be broken down further into events 2, 4, and 6.

In the experiment of rolling a die, there are six elementary events {1, 2, 3, 4, 5, 6}. Rolling a pair of dice results in 36 possible elementary events (outcomes). For each of the six elementary events possible on the roll of one die, there are six possible elementary events on the roll of the second die, as depicted in the tree diagram in **Figure 4.2**. **Table 4.1** contains a list of these 36 outcomes.

In the experiment of rolling a pair of dice, other events could include outcomes such as two even numbers, a sum of 10, a sum greater than five, and others. However, none of these events is an elementary event because each can be broken down into several of the elementary events displayed in Table 4.1.

4.3.4 SAMPLE SPACE

A sample space is *a complete roster or listing of all elementary events for an experiment*. Table 4.1 is the sample space for the roll of a pair of dice. The sample space for the roll of a single die is {1, 2, 3, 4, 5, 6}.

Figure 4.2: Possible Outcomes for the Roll of a Pair of Dice

Sample space can aid in finding probabilities. Suppose an experiment is to roll a pair of dice. What is the probability that the dice will sum to 7? An examination of the sample space shown in Table 4.1 reveals that there are six outcomes in which the dice sum to 7—{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)}—in the total possible 36 elementary events in the sample space. Using this information, we can conclude that the probability of rolling a pair of dice that sum to 7 is 6/36, or .1667. However, using the sample space to determine probabilities is unwieldy and cumbersome when the sample space is large. Hence, business statisticians usually use other more effective methods of determining probability.

4.3.5 UNIONS AND INTERSECTIONS

Set notation, the use of braces to group numbers, is used as *a symbolic tool for unions and intersections* in this chapter. The **union** of *X, Y is formed by combining elements from each of the sets* and is denoted $X \cup Y$. An element qualifies for the union of *X, Y* if it is in either *X* or *Y* or in both *X* and *Y*. The union expression $X \cup Y$ can be translated to "*X* or *Y*." For example, if

$$A = \{1, 4, 7, 9\} \quad \text{and} \quad B = \{2, 3, 4, 5, 6\}$$

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

TABLE 4.1: ALL POSSIBLE ELEMENTARY EVENTS IN THE ROLL OF A PAIR OF DICE (SAMPLE SPACE)

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

Note that all the values of A and all the values of B qualify for the union. However, none of the values is listed more than once in the union. In **Figure 4.3**, the shaded region of the Venn diagram denotes the union.

Figure 4.3: A Union

An intersection is denoted $X \cap Y$. To qualify for intersection, an element must be in both *X* and *Y*. The **intersection** *contains the elements common to both sets*. Thus, the intersection symbol, $\cap$, is often read as *and*. The intersection of *X, Y* is referred to as *X* and *Y*. For example, if

$$A = \{1, 4, 7, 9\} \quad \text{and} \quad B = \{2, 3, 4, 5, 6\}$$

$$A \cap B = \{4\}$$

Note that only the value 4 is common to both sets A and B. The intersection is more exclusive than and hence equal to or (usually) smaller than the union. Elements must be characteristic of both *X* and *Y* to qualify. In **Figure 4.4**, the shaded region denotes the intersection.

Figure 4.4: An Intersection

4.3.6 MUTUALLY EXCLUSIVE EVENTS

Two or more events are **mutually exclusive events** if *the occurrence of one event precludes the occurrence of the other event(s)*. This characteristic means that mutually exclusive events cannot occur simultaneously and therefore can have no intersection.

A manufactured part is either defective or okay: The part cannot be both okay and defective at the same time because "okay" and "defective" are mutually exclusive categories. In a sample of the manufactured products, the event of selecting a defective part is mutually exclusive with the event of selecting a nondefective part. Suppose an office building is for sale and two different potential buyers have placed bids on the building. It is not possible for both buyers to purchase the building; therefore, the event of buyer A purchasing the building is mutually exclusive with the event of buyer B purchasing the building. In the toss of a single coin, heads and tails are mutually exclusive events. The person tossing the coin gets either a head or a tail but never both.

The probability of two mutually exclusive events occurring at the same time is zero.

Mutually Exclusive Events X and Y

$$P(X \cap Y) = 0$$

Figure 4.5 shows a Venn diagram of mutually exclusive events.

Figure 4.5: A Venn Diagram of Mutually Exclusive Events X and Y

4.3.7 INDEPENDENT EVENTS

Two or more events are **independent events** if *the occurrence or nonoccurrence of one of the events does not affect the occurrence or nonoccurrence of the other event(s)*. Certain experiments, such as rolling dice, yield independent events; each die is independent of the other. Whether a 6 is rolled on the first die has no influence on whether a 6 is rolled on the second die. Coin tosses always are independent of each other. The event of getting a head on the first toss of a coin is independent of getting a head on the second toss. It is generally believed that certain human characteristics are independent of other events. For example, left-handedness is probably independent of the possession of a credit card. Whether a person wears glasses or not is probably independent of the brand of milk preferred.

Many experiments using random selection can produce either independent or non-independent events, depending on how the experiment is conducted. In these experiments, the outcomes are independent if sampling is done with replacement; that is, after each item is selected and the outcome is determined, the item is restored to the population and the population is shuffled. This way, each draw becomes independent of the previous draw. Suppose an inspector is randomly selecting bolts from a bin that contains 5% defects. If the inspector samples a defective bolt and returns it to the bin, on the second draw there are still 5% defects in the bin regardless of the fact that the first outcome was a defect. If the inspector does not replace the first draw, the second draw is not independent of the first; in this case, fewer than 5% defects remain in the population. Thus, the probability of the second outcome is dependent on the first outcome.

If X and Y are independent, the following symbolic notation is used.

Independent Events X and Y

$$P(X|Y) = P(X) \quad \text{and} \quad P(Y|X) = P(Y)$$

$P(X|Y)$ denotes the probability of X occurring given that Y has occurred. If X and Y are independent, then the probability of X occurring given that Y has occurred is just the probability of X occurring. Knowledge that Y has occurred does not impact the probability of X occurring because X and Y are independent. For example, P (prefers Pepsi | person is right-handed) = P (prefers Pepsi) because a person's handedness is independent of brand preference.

4.3.8 COLLECTIVELY EXHAUSTIVE EVENTS

A list of **collectively exhaustive events** contains *all possible elementary events for an experiment*. Thus, all sample spaces are collectively exhaustive lists. The list of possible outcomes for the tossing of a pair of dice contained in Table 4.1 is a collectively exhaustive list. The sample space for an experiment can be described as a list of events that are mutually exclusive and collectively exhaustive. Sample space events do not overlap or intersect, and the list is complete.

4.3.9 COMPLEMENTARY EVENTS

The **complement** of event X is denoted X', pronounced "not X." All *the elementary events of an experiment not in X comprise its complement.* For example, if in rolling one die, event X is getting an even number, the complement of X is getting an odd number. If event X is getting a 5 on the roll of a die, the complement of X is getting a 1, 2, 3, 4, or 6. The complement of event X contains whatever portion of the sample space that event X does not contain, as the Venn diagram in **Figure 4.6** shows.

Figure 4.6: The Complement of Event X

Using the complement of an event sometimes can be helpful in solving for probabilities because of the following rule.

Probability of the Complement of X

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

Suppose 32% of the employees of a company have a college degree. If an employee is randomly selected from the company, the probability that the person does not have a college degree is 1 – .32 = .68. Suppose 42% of all parts produced in a plant are molded by machine A and 31% are molded by machine B. If a part is randomly selected, the probability that it was molded by neither machine A nor machine B is 1 – .73 = .27. (Assume that a part is only molded on one machine.)

4.3.10 COUNTING THE POSSIBILITIES

In statistics, a collection of techniques and rules for counting the number of outcomes that can occur for a particular experiment can be used. Some of these rules and techniques can delineate the size of the sample space. Presented here are three of these counting methods.

Suppose a customer decides to buy a certain brand of new car. Options for the car include two different engines, five different paint colors, and three interior packages. If each of these options is available with each of the others, how many different cars could the customer choose from? To determine this number, we can use the ***mn* counting rule.**

The *mn* Counting Rule

For an operation that can be done m ways and a second operation that can be done n ways, the two operations then can occur, in order, in mn ways. This rule can be extended to cases with three or more operations.

Using the mn counting rule, we can determine that the automobile customer has (2)(5)(3) = 30 different car combinations of engines, paint colors, and interiors available.

Suppose a scientist wants to set up a research design to study the effects of sex (M, F), marital status (single never married, divorced, married), and economic class (lower, middle, and upper) on the frequency of airline ticket purchases per year. The researcher would set up a design in which 18 different samples are taken to represent all possible groups generated from these customer characteristics.

$$\text{Number of Groups} = (\text{Sex})(\text{Marital Status})(\text{Economic Class})$$
$$= (2)(3)(3) = 18 \text{ Groups}$$

Sampling from a Population with Replacement

In the second counting method, sampling n items from a population of size N *with replacement* would provide

$$(N)^n \text{ possibilities}$$

where

N = population size

n = sample size

For example, each time a die, which has six sides, is rolled, the outcomes are independent (with replacement) of the previous roll. If a die is rolled three times in succession, how many different outcomes can occur? That is, what is the size of the sample space for this experiment? The size of the population, N, is 6, the six sides of the die. We are sampling three dice rolls, $n = 3$. The sample space is

$$(N)^n = (6)^3 = 216$$

Suppose in a lottery six numbers are drawn from the digits 0 through 9, with replacement (digits can be reused). How many different groupings of six numbers can be drawn? N is the population of 10 numbers (0 through 9) and n is the sample size, six numbers.

$$(N)^n = (10)^6 = 1{,}000{,}000$$

That is, a million six-digit numbers are available!

Combinations: Sampling from a Population Without Replacement

The third counting method uses **combinations**, sampling n items from a population of size N *without replacement* provides possibilities.

$${}_NC_n = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

For example, suppose a small law firm has 16 employees and 3 are to be selected randomly to represent the company at the annual meeting of the American Bar Association. How many different combinations of lawyers could be sent to the meeting? This situation does not allow sampling with replacement because three *different* lawyers will be selected to go. This problem is solved by using combinations. $N = 16$ and $n = 3$, so

$${}_NC_n = {}_{16}C_3 = \frac{16!}{3!13!} = 560$$

A total of 560 combinations of three lawyers could be chosen to represent the firm.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

6. Events that cannot be decomposed or broken down into other events are called _____.
7. The _____ of two sets is formed by combining elements from each of the sets.
8. The probability of two mutually exclusive events occurring at the same time is _____.

State whether the following statements are true/false:

9. Mutually exclusive events can occur simultaneously and therefore can have some intersection points.
10. A sample space is a complete roster or listing of all elementary events for an experiment.

ACTIVITY

In this activity you need to do it at your end literally. You toss a Rs 10 coin 3 times and write down the probability of four heads, HHHH.

- What is the probability that you observe exactly one heads?
- Given that you have observed at least one heads, what is the probability that you observe at least two heads?

Now you might have found that numbers of head is not that expected because it should have been come for 2 times. Don't think that coin in fake but now try to toss the coin 20 times and every time writes the number of head. What is the percentage to heads? Does it change?

4.4 MARGINAL, UNION, JOINT, AND CONDITIONAL PROBABILITIES

Four particular types of probability are presented in this chapter. The first type is **marginal probability**. Marginal probability is denoted $P(E)$, where E is some event. A marginal probability is usually *computed by dividing some subtotal by the whole*. An example of marginal probability is the probability that a person owns a Ford car. This probability is computed by dividing the number of Ford owners by the total number of car owners. The probability of a person wearing glasses is also a marginal probability. This probability is computed by dividing the number of people wearing glasses by the total number of people.

A second type of probability is the union of two events. **Union probability** is denoted $P(E_1 \cup E_2)$, where E_1 and E_2 are two events. $P(E_1 \cup E_2)$ is the probability that E_1 will occur or that E_2 will occur or that both E_1 and E_2 will occur. An example of union probability is the probability that a person owns a Ford or a Chevrolet. To qualify for the union, the person only has to have at least one of these cars. Another example is the probability of a person wearing glasses or having red hair. All people wearing glasses are included in the union, along with all redheads and all redheads who wear glasses. In a company, the probability that a person is male or a clerical worker is a union probability. A person qualifies for the union by being male or by being a clerical worker or by being both (a male clerical worker).

A third type of probability is the intersection of two events, or **joint probability**. The joint probability of events E_1 and E_2 occurring is denoted $P(E_1 \cap E_2)$. Sometimes $P(E_1 \cap E_2)$ is read as the probability of E_1 and E_2. To qualify for the intersection, both events must occur. An example of joint probability is the probability of a person owning both a Ford and a Chevrolet. Owning one type of car is not sufficient. A second example of joint probability is the probability that a person is a redhead and wears glasses.

Figure 4.7: Marginal, Union, Joint, and Conditional Probabilities

The fourth type is conditional probability. **Conditional probability** is denoted $P(E_1 | E_2)$. This expression is read: the probability that E_1 will occur given that E_2 is known to have occurred. Conditional probabilities involve knowledge of some prior information. The information that is known or given is written to the right of the vertical line in the probability statement. An example of conditional probability is the probability that a person owns a Chevrolet given that she owns a Ford. This conditional probability is only a measure of the proportion of Ford owners who have a Chevrolet—not the proportion of total car owners who own a Chevrolet. Conditional probabilities are computed by determining the number of items that have an outcome out of some subtotal of the population. In the car owner example, the possibilities are reduced to Ford owners, and then the number of Chevrolet owners out of those Ford owners is determined. Another example of a conditional probability is the probability that a worker in a company is a professional given that he is male. Of the four probability types, only conditional probability does not have the population total as its denominator. Conditional probabilities have a population subtotal in the denominator. **Figure 4.7** summarizes these four types of probability.

Several tools are available for use in solving probability problems. These tools include sample space, tree diagrams, Venn diagrams, the laws of probability, joint probability tables, and insight. Because of the individuality and variety of probability problems, some techniques apply more readily in certain situations than in others. No best method is available for solving all probability problems. In some instances, the joint probability table lays out a problem in a readily solvable manner. In other cases, setting up the joint probability table is more difficult than solving the problem in another way. The probability laws almost always can be used to solve probability problems.

Four laws of probability are presented in this chapter: the addition laws, the multiplication laws, conditional probability, and Bayes' rule. The addition laws and the multiplication laws each have a general law and a special law.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

11. A _____ is usually computed by dividing some subtotal by the whole.
12. The joint probability is the probability of _____ of two events.
13. _____ probabilities involve knowledge of some prior information.

State whether the following statements are true/false:

14. Venn diagram is one of the several tools which are available for use in solving probability problems.
15. In joint probability, we use subtotal of the possible outcomes in denominator.

ACTIVITY

In a study it has been found that at Goa University, 30% of the students are vegetarian. If a student is a vegetarian, there is a 75% chance that he/she eats Goan Bread. If a student is non-vegetarian, there is a 60% chance that he/she eats Goan Bread.

We are interested in some probabilities when we pick a student at random.

(a) The probability of getting a vegetarian who eats Goan Bread.

(b) The probability of getting a vegetarian who does not eat Goan Bread.

(c) The probability of getting a non-vegetarian who eats Goan Bread.

(d) The probability of getting a student who eats Goan Bread

(e) The probability of getting a vegetarian, if we pick a student who eats Goan Bread.

We can represent the probability situation with a tree diagram:

First branching: Vegetarian or not.

ACTIVITY

We know

P [Vegetarian] = 0.30, so P [Non-vegetarian] = 0.70

Second branching: Eats Goan Bread or not.

We know

P [eats Goan Bread | Vegetarian] = 0.75 and P [eats Goan Bread | non-vegetarian] =0.60.

So, our tree is:

The numbers on the branches are the conditional probabilities for the next step. To obtain the probability of an outcome, we multiply along the path leading to that outcome.

Then we can answer the questions given:

(a) P [Vegetarian and eats Goan Bread] = 0.225

Now you need to answer all the above-mentioned questions.

4.5 ADDITION LAWS

The general law of addition is used to find the probability of the union of two events, $P(X \cup Y)$. The expression $P(X \cup Y)$ denotes the probability of X occurring or Y occurring or both X and Y occurring.

General Law of addition

$$P(X \cup Y) = P(X) + P(Y) - P(X \cap Y)$$

where X, Y are events and $(X \cap Y)$ is the intersection of X and Y.

Yankelovich Partners conducted a survey for the American Society of Interior Designers in which workers were asked which changes in office design would increase productivity. Respondents were allowed to answer more than one type of design change. The number one change that 70% of the workers said would increase productivity was reducing noise. In second place was more storage/filing space, selected by 67%. If one of the survey respondents was randomly selected and asked what office design changes would increase worker productivity, what is the probability that this person would select reducing noise *or* more storage/filing space?

Let N represent the event "reducing noise." Let S represent the event "more storage/filing space." The probability of a person responding with N or S can be symbolized statistically as a union probability by using the law of addition.

$$P(N \cup S)$$

To successfully satisfy the search for a person who responds with reducing noise or more storage/filing space, we need only find someone who wants *at least one* of those two events. Because 70% of the surveyed people responded that reducing noise would create more productivity, $P(N) = .70$. In addition, because 67% responded that increased storage space would improve productivity, $P(S) = .67$. Either of these would satisfy the requirement of the union. Thus, the solution to the problem seems to be

$$P(N \cup S) = P(N) + P(S) = .70 + .67 = 1.37$$

However, we already established that probabilities cannot be more than 1. What is the problem here? Notice that all people who responded that *both* reducing noise *and* increasing storage space would improve productivity are included in *each* of the marginal probabilities $P(N)$ and $P(S)$. Certainly a respondent who recommends both of these improvements should be included as favoring at least one. However, because they are included in the $P(N)$ *and* the $P(S)$, the people who recommended both improvements are *double counted*. For that reason, the general law of addition subtracts the intersection probability, $P(N \cap S)$.

In **Figure 4.8**, Venn diagrams illustrate this discussion. Notice that the intersection area of N and S is double shaded in diagram A, indicating that it has been counted twice. In diagram B, the shading is consistent throughout N and S because the intersection area has been subtracted out once. Thus diagram B illustrates the proper application of the general law of addition.

Figure 4.8: Solving for the Union in the Office Productivity Problem

So what is the answer to Yankelovich Partners' union probability question? Suppose 56% of all respondents to the survey had said that *both* noise reduction *and* increased storage/filing space would improve productivity; $P(N \cap S) = .56$. We can use the general law of addition to solve for the probability that a person responds that *either* noise reduction *or* increased storage space would improve productivity as:

$$P(N \cup S) = P(N) + P(S) - P(N \cap S) = .70 + .67 - .56 = .81$$

Hence, 81% of the workers surveyed responded that *either* noise reduction *or* increased storage space would improve productivity.

4.5.1 JOINT PROBABILITY TABLES

In addition to the formulas, another useful tool in solving probability problems is using a joint probability table. A **joint probability table** *displays the intersection (joint) probabilities along with the marginal probabilities of a given problem.* Union probabilities or conditional probabilities are not directly displayed in a joint probability table but can be computed using values from the table. Generally, a joint probability table is constructed as a two-dimensional table with one variable on each side of the table. For example, in the office design problem, noise reduction would be on one side of the table and increased storage space on the other. In this problem, a Yes row and a No row would be created for one variable and a Yes column and a No column would be created for the other variable, as shown in **Table 4.2**.

TABLE 4.2: JOINT PROBABILITY TABLE FOR THE OFFICE DESIGN PROBLEM

		Increase Storage Space	
		Yes	*No*
Noise Reduction	*Yes*		
	No		

Once the joint probability table is created, we can enter the marginal probabilities. $P(N) = .70$ is the marginal probability that a person responds yes to noise reduction. This value is placed in the "margin" in the row of Yes to noise reduction, as shown in **Table 4.3**. If $P(N) = .70$, then .30 of the people surveyed did not think that noise reduction would increase productivity, $P(\text{not } N) = 1 - .70 = .30$. This value, also a marginal probability, goes in the row indicated by No under noise reduction. In the column under Yes for increased storage space, the marginal probability $P(S) = .67$ is recorded. Finally, the marginal probability of No for increased storage space, $P(\text{not } S) = 1 - .67 = .33$, is placed in the No column. In this joint probability table, all four marginal probabilities are given or can be computed simply by using the probability of a complement rule, $P(\text{not } S) = 1 - P(S)$.

The intersection of noise reduction and increased storage space is given as $P(N \cap S) = .56$. This value is entered into the joint probability table in the cell under Yes Yes, as shown in Table 4.3. The rest of the table can be determined by subtracting the cell values from the marginal probabilities. For example, subtracting .56 from .70 and getting .14 yields the value for the cell under Yes for noise reduction and No for increased storage space. In other words, 14% of all respondents said that noise reduction would improve productivity and increased storage space would not. Filling out the rest of the table results in the probabilities shown in Table 4.3.

TABLE 4.3: JOINT PROBABILITY TABLE FOR THE OFFICE DESIGN PROBLEM

		Increase Storage Space: Yes	Increase Storage Space: No	
Noise Reduction	Yes	.56	.14	.70
	No	.11	.19	.30
		.67	.33	1.00

TABLE 4.4: YES ROW AND YES COLUMN FOR THE JOINT PROBABILITY TABLE OF THE OFFICE DESIGN PROBLEM

		Increase Storage Space: Yes	Increase Storage Space: No	
Noise Reduction	Yes	.56	.14	.70
	No	.11		
		.67		

Now we can solve the union probability, $P(N \cup S)$, in at least two different ways using the joint probability table. The focus is on the Yes row for noise reduction and the Yes column for increase storage space, as displayed in **Table 4.4**. The probability of a person suggesting noise reduction or increased storage space as a solution for improving productivity, $P(N \cup S)$, can be determined from the joint probability table by adding the marginal probabilities of Yes for noise reduction and Yes for increased storage space and then subtracting the Yes Yes cell, following the pattern of the general law of probabilities.

$$P(N \cup S) = .70(\text{from Yes row}) + .67(\text{from Yes column}) - .56(\text{From Yes Yes cell}) = .81$$

Another way to solve for the union probability from the information displayed in the joint probability table is to sum all cells in any of the Yes rows or columns. Observe the following from Table 4.4.

$$P(N \cup S) = .56(\text{from Yes Yes cell}) + .14\ (\text{from Yes on noise reduction and No on increase storage space}) + .11\ (\text{from No on noise reduction and Yes on increase storage space}) = .81$$

DEMONSTRATION PROBLEM 4.1

A client company data reveals that 155 employees worked one of four types of positions. Shown here again is the cross-tabulation table (also called a contingency table) with the frequency counts for each category and for subtotals and totals containing a breakdown of these employees by type of position and by sex. If an employee of the company is selected randomly, what is the probability that the employee is female or a professional worker?

COMPANY HUMAN RESOURCE DATA

		Sex: Male	Sex: Female	
Type of Position	Managerial	8	3	11
	Professional	31	13	44
	Technical	52	17	69
	Clerical	9	22	31
		100	55	155

Solution: Let F denote the event of female and P denote the event of professional worker. The question is

$$P(F \cup P) = ?$$

By the general law of addition,

$$P(F \cup P) = P(F) + P(P) - P(F \cap P)$$

Of the 155 employees, 55 are women. Therefore, $P(F) = 55/155 = .355$. The 155 employees include 44 professionals. Therefore, $P(P) = 44/155 = .284$. Because 13 employees are both female and professional, $P(F \cap P) = 13/155 = .084$. The union probability is solved as

$$P(F \cup P) = .355 + .284 - .084 = .555.$$

Suppose you want to solve this problem using a joint probability table. A joint probability table for this problem can be constructed from the contingency table by dividing all data by 155. The resulting table is:

		Sex: Male	Sex: Female	Total
Type of Position	Managerial	.052	.019	.071
	Professional	.200	.084	.284
	Technical	.335	.110	.445
	Clerical	.058	.142	.200
		.645	.355	1.000

Solving for the probability that the selected worker is female or a professional worker using a joint probability table is done by adding the values in the highlighted cells shown below:

$$.019 + .084 + .110 + .142 + .200 = .555$$

or by using the general law of multiplication from the table,

$$.355 + .284 - .084 = .555$$

		Sex		
		Male	Female	Total
Type of Position	Managerial	.052	.019	.071
	Professional	.200	.084	.284
	Technical	.335	.110	.445
	Clerical	.058	.142	.200
		.645	.355	1.000

DEMONSTRATION PROBLEM 4.2

Shown here are the cross-tabulation table and corresponding joint probability table for the results of a national survey of 200 executives who were asked to identify the geographic locale of their company and their company's industry type. The executives were only allowed to select one locale and one industry type.

CROSS-TABULATION TABLE

		Geographic Location				
		Northeast D	Southeast E	Midwest F	West G	
Industry Type	Finance A	24	10	8	14	56
	Manufacturing B	30	6	22	12	70
	Communications C	28	18	12	16	74
		82	34	42	42	200

By dividing every value of the cross-tabulation table by the total (200), the corresponding joint probability table can be constructed.

JOINT PROBABILITY TABLE

		Geographic Location				
		Northeast D	Southeast E	Midwest F	West G	
Industry Type	Finance A	.12	.05	.04	.07	.28
	Manufacturing B	.15	.03	.11	.06	.35
	Communications C	.14	.09	.06	.08	.37
		.41	.17	.21	.21	1.00

Suppose a respondent is selected randomly from these data.

(a) What is the probability that the respondent is from the Midwest (F)?

(b) What is the probability that the respondent is from the communications industry (C) or from the Northeast (D)?

(c) What is the probability that the respondent is from the Southeast (E) or from the finance industry (A)?

Solution:

(a) $P(\text{Midwest}) = P(F) = .21$

(b) $P(C \cup D) = P(C) + P(D) - P(C \cap D) = .37 + .41 - .14 = .64$

(c) $P(E \cup A) = P(E) + P(A) - P(E \cap A) = .17 + .28 - .05 = .40$

In computing the union by using the general law of addition, the intersection probability is subtracted because it is already included in both marginal probabilities. This adjusted probability leaves a union probability that properly includes both marginal values and the intersection value. If the intersection probability is subtracted out a *second* time, the intersection is removed, leaving the probability of X *or* Y but not *both*.

$$P(X \text{ or } Y \text{ but not both}) = P(X) + P(Y) - P(X \cap Y) - P(X \cap Y)$$
$$= P(X \cup Y) - P(X \cap Y)$$

Figure 4.9 is the Venn diagram for this probability.

Figure 4.9: The X or Y but Not Both Case

Complement of a Union

The probability of the union of two events X and Y represents the probability that the outcome is *either* X *or* it is Y or it is *both* X and Y. The union includes everything except the possibility that it is neither (X or Y). Another way to state it is as *neither* X *nor* Y, which can symbolically be represented as $P(\text{not } X \cap \text{not } Y)$. Because it is the only possible case other than the union of X or Y, it is the **complement of a union**. Stated more formally,

$$P(\text{neither } X \text{ nor } Y) = P(\text{not } X \cap \text{not } Y) = 1 - P(X \cup Y).$$

Examine the Venn diagram in **Figure 4.10**. Note that the complement of the union of X, Y is the shaded area outside the circles. This area represents the neither X nor Y region.

Figure 4.10: The Complement of a Union: The Neither/Nor Region

In the survey about increasing worker productivity by changing the office design discussed earlier, the probability that a randomly selected worker would respond with noise reduction or increased storage space was determined to be

$$P(N \cup S) = P(N) + P(S) - P(N \cap S) = .70 + .67 - .56 = .81$$

The probability that a worker would respond with *neither* noise reduction *nor* increased storage space is calculated as the complement of this union.

$$P(\text{neither } N \text{ nor } S) = P(\text{not } N \cap \text{not } S) = 1 - P(N \cup S) = 1 - .81 = .19$$

Thus, 19% of the workers selected neither noise reduction nor increased storage space as solutions to increasing productivity. In Table 4.3, this *neither/nor* probability is found in the No No cell of the table, .19.

4.5.2 SPECIAL LAW OF ADDITION

If two events are mutually exclusive, the probability of the union of the two events is the probability of the first event plus the probability of the second event. Because mutually exclusive events do not intersect, nothing has to be subtracted.

Special Law of Addition

If X, Y are mutually exclusive,

$$P(X \cup Y) = P(X) + P(Y).$$

The special law of addition is a special case of the general law of addition. In a sense, the general law fits all cases. However, when the events are mutually exclusive, a zero is inserted into the general law formula for the intersection, resulting in the special law formula.

Yankelovich Partners conducted a survey for William M. Mercer in which workers were asked what most hinders their productivity, and they were only allowed to choose from one of the following: Lack of direction, Lack of support, Too much work, Inefficient process, Not enough equipment/supplies, and Low pay/chance to advance.

Lack of direction was cited by the most workers (20%), followed by lack of support (18%), too much work (18%), inefficient process (8%), not enough equipment/supplies (7%), low pay/chance to advance (7%), and a variety of other factors added by respondents. If a worker who responded to this survey is selected and asked which of the given selections most hinders his or her productivity, what is the probability that the worker will respond that it is either too much work or inefficient process?

Let M denote the event "too much work" and I denote the event "inefficient process." The question is:

$$P(M \cup I) = ?$$

Because 18% of the survey respondents said "too much work,"

$$P(M) = .18$$

Because 8% of the survey respondents said "inefficient process,"

$$P(I) = .08$$

Because it was not possible to select more than one answer,

$$P(M \cup I) = .0000$$

Implementing the special law of addition gives

$$P(M \cup I) = P(M) + P(I) = .18 + .08 = .26$$

Twenty-six percent responded either "too much work" or "inefficient process."

DEMONSTRATION PROBLEM 4.3

If a worker is randomly selected from the company described in Demonstration Problem 4.1, what is the probability that the worker is either technical or clerical? What is the probability that the worker is either a professional or a clerical?

Solution: Examine the cross-tabulation table of the company's human resources data shown in Demonstration Problem 4.1. In many cross-tabulation tables like this one, the rows are nonoverlapping or mutually exclusive, as are the columns. In this matrix, a worker can be classified as being in only one type of position and as either male or female but not both. Thus, the categories of type of position are mutually exclusive, as are the categories of sex, and the special law of addition can be applied to the human resource data to determine the union probabilities.

Let T denote technical, C denote clerical, and P denote professional. The probability that a worker is either technical or clerical is

$$P(T \cup C) = P(T) + P(C) = \frac{69}{155} + \frac{31}{155} = \frac{100}{155} = .645$$

The probability that a worker is either professional or clerical is

$$P(P \cup C) = P(P) + P(C) = \frac{44}{155} + \frac{31}{155} = \frac{75}{155} = .484$$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

16. General law of addition is used to find the probability of the _____ of two events.
17. When the events are mutually exclusive into the general law formula for the intersection, we insert _____.

State whether the following statements are true/false:

18. If X and Y are mutually exclusive events, then the probability of the union of the two events will be $P(X) + P(Y)$.
19. In computing the union by using the general law of addition, the intersection probability is multiplied.
20. Compliment of union can be calculated from subtracting the probability of X occurring or Y occurring or both X and Y occurring.

ACTIVITY

Click on the below mentioned link and write down all the information available on a well-shuffled pack of card.

https://www.thoughtco.com/standard-deck-of-cards-3126599

Create the joint probability table for drawing a king. Write all conditions where drawing a king can be a mutually exclusive event.

4.6 MULTIPLICATION LAWS

4.6.1 GENERAL LAW OF MULTIPLICATION

As stated in Section 4.4, the probability of the intersection of two events $(X \cap Y)$ is called the joint probability. The general law of multiplication is used to find the joint probability.

General Law of Multiplication

$$P(X \cap Y) = P(X) \cdot P(Y|X) = P(Y) \cdot P(X|Y)$$

The notation $X \cap Y$ means that *both X and Y* must happen. The general law of multiplication gives the probability that *both* event X and event Y will occur at the same time.

A recent survey conducted for the National Restaurant Association, United States revealed that 47% of adults said they would be likely to patronize a mobile food truck parked near their home. Suppose 70% of adults who would be likely to patronize such a mobile food truck earn less than $40,000 annually. If an adult is randomly selected, what is the probability that the adult would be likely to patronize a mobile food truck parked near their home and the adult earns less than $40,000 annually? This question is a joint (intersection) probability. Let M signify that an adult would be likely to patronize such a mobile food truck and let F signify that the adult earns less than $40,000 annually. In these terms, the probability question is:

$$P(M \cap F) = ?$$

What is given in this problem? The 47% of adults likely to patronize such a mobile food truck is a marginal percentage because it is a percentage of all adults. Thus, $P(M) = .47$. However, the 70% earning less than $40,000 annually is only of the adults likely to patronize a mobile food truck. This is a conditional percentage because the percentage is not of all adults but only of those who are likely to patronize a mobile food truck. Writing this as a conditional probability, $P(F|M) = .70$ With this information, the intersection or joint probability can be solved for as:

$$P(M \cap F) = P(M) \cdot P(F|M) = (47) \cdot (70) = .3290$$

Conceptually, what this means is that if 70% of adults who would patronize a mobile food truck earn less than $40,000 and if 47% of adults would patronize a mobile food truck, then 32.9% of adults would patronize a mobile food truck and earn less than $40,000. We are taking 70% of the 47% and hence

multiplying the conditional by the marginal probabilities. The Venn diagram in **Figure 4.11** shows these relationships and the joint probability.

Figure 4.11: Joint Probability that an Adult Patronizes a Mobile Food Truck and Earns Less than $40,000

TABLE 4.5: JOINT PROBABILITY TABLE OF COMPANY HUMAN RESOURCE DATA

		Sex		
		Male	*Female*	
Type of Position	*Managerial*	.052	.019	.071
	Professional	.200	.084	.284
	Technical	.335	.110	.445
	Clerical	.058	.142	.200
		.645	.355	1.000

Note that there is a second version of the general law of multiplication, which if, applied here, would be: $P(M) \cdot P(F|M)$ This expression could be used if we knew the probability that an adult earned less than $40,000 annually, $P(F)$, a marginal probability, and if we knew the proportion (or percentage) of adults earning less than $40,000 who would patronize a mobile food truck, $P(F|M)$, a conditional probability. Neither of these probabilities is given in this problem; therefore, this version of the general law of multiplication was not used in this problem.

Determining joint probabilities from joint probability tables is easy because every cell of these matrices is a joint probability, hence the name "joint probability table."

For example, suppose the cross-tabulation table of the client company data from Demonstration Problem 4.1 and the Decision Dilemma is converted to a joint probability table by dividing by the total number of employees ($N = 155$), resulting in **Table 4.5**. Each value in the cell of Table 4.5 is an intersection, and the table contains all possible intersections (joint probabilities) for the events of sex and type of position. For example, the probability that a randomly selected worker is male *and* a technical worker, $P(M \cap T)$ is .335. The probability that a randomly selected worker is female *and* a professional worker, $P(F \cap P)$ is .084. Once a joint probability table is constructed for a problem, usually the easiest way to solve for the joint probability is to find the appropriate cell in the

table and select the answer. However, sometimes because of what is given in a problem, using the formula is easier than constructing the table.

DEMONSTRATION PROBLEM 4.4

A company has 140 employees, of which 30 are supervisors. Eighty of the employees are married, and 20% of the married employees are supervisors. If a company employee is randomly selected, what is the probability that the employee is married and is a supervisor?

Solution: Let M denote married and S denote supervisor. The question is:

$$P(M \cap S) = ?$$

First, calculate the marginal probability.

$$P(M) = \frac{80}{140} = .5714$$

Then, note that 20% of the married employees are supervisors, which is the conditional probabilty, $P(S|M) = 20$. Finally, applying the general law of multiplication gives

$$P(M \cap S) = P(M) \cdot P(S|M) = (.5714)(.20) = .1143$$

Hence, 11.43% of the 140 employees are married and are supervisors.

DEMONSTRATION PROBLEM 4.5

From the data obtained from the interviews of 200 executives in Demonstration Problem 4.2, find:

(a) $P(B \cap E)$

(b) $P(G \cap A)$

(c) $P(B \cap C)$

JOINT PROBABILITY TABLE

		Geographic Location				
		Northeast D	*Southeast E*	*Midwest F*	*West G*	
Industry Type	*Finance A*	.12	.05	.04	.07	.28
	Manufacturing B	.15	.03	.11	.06	.35
	Communications C	.14	.09	.06	.08	.37
		.41	.17	.21	.21	1.00

Solution:

(a) From the cell of the joint probability table, $P(B \cap E) = .03$. To solve by the formula, $P(B \cap E) = P(B) \cdot P(E|B)$, first find $P(B)$:

$$P(B) = .35$$

The probability of E occurring given that B has occurred, $P(E|B)$, can be determined from the joint probability table as $P(E|B) = .03/.35$ Therefore,

$$P(B \cap E) = P(B) \cdot P(E|B) = (.35)\left(\frac{.03}{.35}\right) = .03$$

Although the formula works, finding the joint probability in the cell of the joint probability table is faster than using the formula.

An alternative formula is $P(B \cap E) = P(E) \cdot P(B|E)$, but $P(E) = .17$. Then $P(B|E)$ means the probability of B if E is given. There are .17 Es in the probability matrix and .03 Bs in these Es. Hence,

$$P(B|E) = \frac{.03}{.17} \quad \text{and} \quad P(B \cap E) = P(E) \cdot P(B|E) = (.17)\left(\frac{.03}{.17}\right) = .03$$

(b) To obtain $P(G \cap A)$, find the intersecting cell of G and A in the probability matrix, .07, or use one of the following formulas:

$$P(G \cap A) = P(G) \cdot P(A|G) = (.21)\left(\frac{.07}{.21}\right) = .07$$

or

$$P(G \cap A) = P(A) \cdot P(G|A) = (.28)\left(\frac{.07}{.28}\right) = .07$$

(c) The probability $P(B \cap C)$ means that one respondent would have to work both in the manufacturing industry and the communications industry. The survey used to gather data from the 200 executives, however, requested that each respondent specify only one industry type for his or her company. The joint probability table shows no intersection for these two events. Thus B and C are mutually exclusive. None of the respondents is in both manufacturing and communications. Hence,

$$P(B \cap C) = .0$$

4.6.2 SPECIAL LAW OF MULTIPLICATION

If events X and Y are independent, a special law of multiplication can be used to find the intersection of X and Y. This special law utilizes the fact that when two events X, Y are independent, $P(X|Y) = P(X)$ and $P(Y|X) = P(Y)$. Thus, the general law of multiplication, $P(X \cap Y) = P(X) \cdot P(Y|X)$, becomes $P(X \cap Y) = P(X) \cdot P(Y)$ when X and Y are independent.

Special Law of Multiplication

If X, Y are independent, $P(X \cap Y) = P(X) \cdot P(Y)$

A study released by Bruskin-Goldring Research for SEIKO found that 28% of American adults believe that the automated teller has had a most significant impact on everyday life. Another study by David Michaelson & Associates for Dale Carnegie & Associates examined employee views on team spirit in the workplace and discovered that 71% of all employees believe that working as a part of a team lowers stress. Are people's views on automated tellers independent of their views on team spirit in

the workplace? If they are independent, then the probability of a person being randomly selected who believes that the automated teller has had a most significant impact on everyday life *and* that working as part of a team lowers stress is found as follows. Let A denote automated teller and S denote teamwork lowers stress.

$$P(A) = .28$$
$$P(S) = .71$$
$$P(A \cap S) = P(A) \cdot P(S) = (.28)(.71) = .1988$$

Therefore, 19.88% of the population believes that the automated teller has had a most significant impact on everyday life *and* that working as part of a team lowers stress.

DEMONSTRATION PROBLEM 4.6

A manufacturing firm produces pads of bound paper. Three percent of all paper pads produced are improperly bound. An inspector randomly samples two pads of paper, one at a time. Because a large number of pads are being produced during the inspection, the sampling being done, in essence, is with replacement. What is the probability that the two pads selected are both improperly bound?

Solution: Let I denote improperly bound. The problem is to determine

$$P(I_1 \cap I_2) = ?$$

The probability of $I = .03$, or 3% are improperly bound. Because the sampling is done with replacement, the two events are independent. Hence,

$$P(I_1 \cap I_2) = P(I_1) \cdot P(I_2) = (.03)(.03) = .0009$$

Most cross-tabulation tables contain variables that are not independent. If a cross-tabulation table contains independent events, the special law of multiplication can be applied. If not, the special law cannot be used. In Section 4.6 we explore a technique for determining whether events are independent. **Table 4.6** contains data from independent events.

TABLE 4.6: CROSS-TABULATION TABLE OF DATA FROM INDEPENDENT EVENTS

	D	E	
A	8	12	20
B	20	30	50
C	6	9	15
	34	51	85

DEMONSTRATION PROBLEM 4.7

Use the data from Table 4.6 and the special law of multiplication to find $P(B \cap D)$.

Solution:

$$P(B \cap D) = P(B) \cdot P(D) = \frac{50}{85} \cdot \frac{34}{85} = .2353$$

This approach works *only* for cross-tabulation and joint probability tables in which the variable along one side of the table is *independent* of the variable along the other side of the table. Note that the answer obtained by using the formula is the same as the answer obtained by using the cell information from Table 4.6.

$$P(B \cap D) = \frac{20}{85} = .2353$$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

21. If X and Y are _____, then law of multiplication will be mathematical multiplication of their probabilities ($P(x) * P(y)$).
22. If a _____ table contains independent events, the special law of multiplication can be applied.

State whether the following statements are true/false:

23. The notation $X \cap Y$ means that both X and Y must happen.
24. Most cross-tabulation tables contain variables that are independent.
25. General law of multiplication is $P(X \cap Y) = P(X) \,.\, P(Y|X) = P(Y) \,.\, P(X|Y)$.

ACTIVITY

For this activity, you need to consider the activity given in Section 4.4 (in the following graph):

Can you find out some examples of multiplication probability for dependent and independent activity?

4.7 CONDITIONAL PROBABILITY

Conditional probabilities are computed based on the prior knowledge that a business researcher has on one of the two events being studied. If X, Y are two events, the conditional probability of X occurring given that Y is known or has occurred is expressed as $P(X|Y)$ and is given in the *law of conditional probability*.

Law of Conditional Probability

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X) \cdot P(Y|X)}{P(Y)}$$

The conditional probability of $(X|Y)$ is the probability that X will occur given Y. The formula for conditional probability can be derived by dividing both sides of the general law of multiplication by $P(Y)$.

In the study by Yankelovich Partners to determine what changes in office design would improve productivity, 70% of the respondents believed noise reduction would improve productivity and 67% said increased storage space would improve productivity. In addition, suppose 56% of the respondents believed both noise reduction and increased storage space would improve productivity. A worker is selected randomly and asked about changes in office design. This worker believes that noise reduction would improve productivity. What is the probability that this worker believes increased storage space would improve productivity? That is, what is the probability that a randomly selected person believes storage space would improve productivity *given that* he or she believes noise reduction improves productivity? In symbols, the question is

$$P(S|N) = ?$$

Note that the given part of the information is listed to the right of the vertical line in the conditional probability. The formula solution is

$$P(S|N) = \frac{P(S \cap N)}{P(N)}$$

but

$$P(N) = .70 \quad \text{and} \quad P(S \cap N) = .56$$

Therefore

$$P(S|N) = \frac{P(S \cap N)}{P(N)} = \frac{.56}{.70} = .80$$

Eighty percent of workers who believe noise reduction would improve productivity believe increased storage space would improve productivity.

Note in **Figure 4.12** that the area for N in the Venn diagram is completely shaded because it is given that the worker believes noise reduction will improve productivity. Also notice that the intersection of N and S is more heavily shaded. This portion of noise reduction includes increased storage space. It is the only part of increased storage space that is in noise reduction, and because the person is known to favor noise reduction, it is the only area of interest that includes increased storage space.

Figure 4.12: Conditional Probability of Increased Storage Space Given Noise Reduction

Examine the joint probability table in **Table 4.7** for the office design problem. None of the probabilities given in the table are conditional probabilities. To reiterate what has been previously stated, a joint probability table contains only two types of probabilities, marginal and joint. The cell values are all joint probabilities and the subtotals in the margins are marginal probabilities. How are conditional probabilities determined from a joint probability table? The law of conditional probabilities shows that a conditional probability is computed by dividing the joint probability by the marginal probability. Thus, the joint probability table has all the necessary information to solve for a conditional probability.

TABLE 4.7: OFFICE DESIGN PROBLEM JOINT PROBABILITY TABLE

		Increase Storage Space		
		Yes	*No*	
Noise Reduction	*Yes*	[illegible]	.14	.70
	No	[illegible]	.19	.30
		.67	.33	1.00

What is the probability that a randomly selected worker believes noise reduction would not improve productivity given that the worker does believe increased storage space would improve productivity? That is,

$$P(\text{not } N|S) = ?$$

The law of conditional probability states that

$$P(\text{not } N|S) = \frac{P(\text{not } N \cap S)}{P(S)}$$

Notice that because S is given, we are interested only in the column that is shaded in Table 4.7, which is the Yes column for increased storage space. The marginal probability, $P(S)$, is the total of this column and is found in the margin at the bottom of the table as .67. $P(\text{not } N \cap S)$ is found as the intersection of No for noise and Yes for storage. This value is .11. Hence, $P(\text{not } N \cap S)$ is .11. Therefore,

$$P(\text{not } N|S) = \frac{P(\text{not } N \cap S)}{P(S)} = \frac{.11}{.67} = .164$$

The second version of the conditional probability law formula is

$$P(X|Y) = \frac{P(X) \cdot P(Y|X)}{P(Y)}$$

This version is more complex than the first version, $P(X \cap Y)/P(Y)$. However, sometimes the second version must be used because of the information given in the problem—for example, when solving for $P(X|Y)$ but $P(Y|X)$ is given. The second version of the formula is obtained from the first version by substituting the formula for $P(X \cap Y) = P(X) \cdot P(Y|X)$ into the first version. In general, this second version of the law of conditional probabilities is likely to be used for solving $P(X|Y)$ when $P(X \cap Y)$ is unknown but $P(Y|X)$ is known.

DEMONSTRATION PROBLEM 4.8

A survey of web retailers published by *Internet Researcher* found that 55% of respondents had shipment tracking on their Web sites. In addition, 43% had a shipping costs calculator on their Web sites. Suppose that of the respondents who had shipment tracking on their Web sites, 60% had a shipping costs calculator on their Web sites. Based on this information, what is the probability that a randomly selected respondent said that it has shipping tracking on its Web sites given that it has a shipping costs calculator on its Web sites?

Solution: Let C = shipping costs calculator and T = shipment tracking. The question is to determine the probability that a respondent has shipping tracking on its Web sites given that it has a shipping costs calculator on its Web sites. This is a conditional probability that can be written as $P(T|C) = ??$.

In this problem, two marginal probabilities are given:

$$P(T) = .55 \text{ and } P(C) = .43$$

The third probability given in this problem is a conditional probability. The given condition is shipment tracking and the conditional probability is $P(C|T) = .60$.

Note that we are solving for $P(T|C)$ but $P(C|T)$ is given. This is a candidate for solving for a conditional probability using the second form of the conditional probability law formula:

$$P(T|C) = \frac{P(T) \cdot P(C|T)}{P(C)} = \frac{(.55)(.60)}{.43} = .7674$$

Thus, 76.74% of those respondents who have a shipping costs calculator on their Web sites have shipment tracking on their Web sites.

DEMONSTRATION PROBLEM 4.9

The data from the executive interviews given in Demonstration Problem 4.2 are repeated here. Use these data to find:

(a) $P(B|F)$ (b) $P(G|C)$ (c) $P(D|F)$

JOINT PROBABILITY TABLE

		Geographic Location				
		Northeast D	*Southeast E*	*Midwest F*	*West G*	
Industry Type	*Finance A*	.12	.05	.04	.07	.28
	Manufacturing B	.15	.03	.11	.06	.35
	Communications C	.14	.09	.06	.08	.37
		.41	.17	.21	.21	1.00

Solution:

(a) $P(B|F) = \frac{P(B \cap F)}{P(F)} = \frac{.11}{.21} = .524$

Determining conditional probabilities from a joint probability table by using the formula is a relatively painless process. In this case, the joint probability, $P(B \cap F)$, appears in a cell of the table (.11); the marginal probability, $P(F)$, appears in a margin (.21). Bringing these two probabilities together by formula produces the answer, .11/.21 = .524. This answer means that 52.4% of the Midwest executives (the F values) are in manufacturing (the B values).

(b) $P(G|C) = \frac{P(G \cap C)}{P(C)} = \frac{.08}{.37} = .216$

This result means that 21.6% of the responding communications industry executives (C) are from the West (G).

(c) $P(D|F) = \frac{P(D \cap F)}{P(F)} = \frac{.00}{.21} = .00$

Because D and F are mutually exclusive, $P(D \cap F)$ is zero and so is $P(D|F)$. The rationale behind $P(D|F) = 0$ is that, if F is given (the respondent is known to be located in the Midwest), the respondent could not be located in D (the Northeast).

4.7.1 INDEPENDENT EVENTS

Independent Events X, Y

To test to determine if X and Y are independent events, the following must be true.

$$P(X|Y) = P(X) \quad \text{and} \quad P(Y|X) = P(Y)$$

In each equation, it does not matter that X or Y is given because X and Y are *independent*. When X and Y are independent, the conditional probability is solved as a marginal probability.

Sometimes, it is important to test a cross-tabulation table of raw data to determine whether events are independent. If any combination of two events from the different sides of the table fail the test, $P(X|Y) = P(X)$, the table does not contain independent events.

DEMONSTRATION PROBLEM 4.10

Test the cross-tabulation table for the 200 executive responses to determine whether industry type is independent of geographic location.

CROSS-TABULATION TABLE

		Geographic Location				
		Northeast D	*Southeast E*	*Midwest F*	*West G*	
Industry Type	*Finance A*	24	10	8	14	56
	Manufacturing B	30	6	22	12	70
	Communications C	28	18	12	16	74
		82	34	42	42	200

Solution: Select one industry and one geographic location (say, *A*—Finance and *G*—West). Does $P(A|G) = P(A)$?

$$P(A|G) = \frac{14}{42} \text{ and } P(A) = \frac{56}{200}$$

Does 14/42 = 56/200? No, .33 ≠ .28. Industry and geographic location are not independent because at least one exception to the test is present.

DEMONSTRATION PROBLEM 4.11

Determine whether the cross-tabulation table shown as Table 4.6 and repeated here contains independent events.

	D	*E*	
A	8	12	20
B	20	30	50
C	6	9	15
	34	51	85

Solution: Check the first cell in the table to find whether $P(A|D) = P(A)$.

$$P(A|D) = \frac{8}{34} = .2353$$

$$P(A) = \frac{20}{85} = .2353$$

The checking process must continue until all the events are determined to be independent. In this table, all the possibilities check out. Thus, Table 4.6 contains independent events.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

26. The conditional probability of $(X|Y)$ is the probability that X will occur when _____ Y.
27. A joint probability table contains only two types of probabilities, first is _____ and second is joint.
28. When X and Y are _____, the conditional probability is solved as a marginal probability.

State whether the following statements are true/false:

29. The law of conditional probabilities shows that a conditional probability is computed by dividing the joint probability by the marginal probability.
30. If X and Y are independent events, then $P(X|Y) = P(Y|X)$.

ACTIVITY

Using a Venn diagram draw the area for union and intersection in your notebook. Also show the probability of occurrence of *B* if *A* has already occurred.

4.8 REVISION OF PROBABILITIES: BAYES' RULE

An extension to the conditional law of probabilities is Bayes' rule, which was developed by and named for Thomas Bayes (1702–1761). **Bayes' rule** is *a formula that extends the use of the law of conditional probabilities to allow revision of original probabilities with new information.*

Bayes' Rule

$$P(X_i|Y) = \frac{P(X_i)\cdot P(Y|X_i)}{P(X_1)\cdot P(Y|X_1) + P(X_2)\cdot P(Y|X_2) + \cdots + P(X_n)\cdot P(Y|X_n)}$$

Recall that the law of conditional probability for

$$P(X_i|Y)$$

is

$$P(X_i|Y) = \frac{P(X_i)\cdot P(Y|X_i)}{P(Y)}$$

Compare Bayes' rule to this law of conditional probability. The numerators of Bayes' rule and the law of conditional probability are the same—the intersection of X_i and Y shown in the form of the general rule of multiplication. The new feature that Bayes' rule uses is found in the denominator of the rule:

$$P(X_1)\cdot P(Y|X_1)+P(X_2)\cdot P(Y|X_2)+\cdots+P(X_n)\cdot P(Y|X_n)$$

The denominator of Bayes' rule includes a product expression (intersection) for every partition in the sample space, Y, including the event (X_i) itself. The denominator is thus a collectively exhaustive listing of mutually exclusive outcomes of Y. This denominator is sometimes referred to as the "total probability formula." It represents a weighted average of the conditional probabilities, with the weights being the prior probabilities of the corresponding event.

By expressing the law of conditional probabilities in this new way, Bayes' rule enables the statistician to make new and different applications using conditional probabilities. In particular, statisticians use Bayes' rule to "revise" probabilities in light of new information.

As an application of Bayes' rule, suppose the American Printer Company sells printer cartridges under its own brand name that are produced by two suppliers: Alamo and South Jersey. Of the printer cartridges sold by the American Printer Company, Alamo produces 65% and South Jersey produces 35%. All cartridges are made to the same specifications and are shipped to the American Printer Company, who puts their own label on the cartridges. Once the cartridges have the American Printer Company label on them, it is virtually impossible to determine from which supplier they came. If one of these cartridges is randomly selected, what is the probability that the cartridge was supplied by Alamo? Since Alamo produces 65% of the cartridges, the probability that an American Printer Company cartridge comes from Alamo, $P(A)$, is .65 where A denotes Alamo. The probability that the cartridges were produced by South Jersey is $P(SJ) = .35$ where SJ denotes South Jersey. These marginal probabilities are sometimes referred to as *prior* probabilities because they are determined before any new information is obtained. Suppose, however, the printer cartridge is tested and determined to be defective. How can this new information be used to change (or revise) the probability that the cartridge came from a particular supplier? Using Bayes' Rule, and some additional information, we are able to revise these original (marginal) probabilities in light of new information.

If both suppliers produced the same percentage of defects, knowing that the cartridge is defective would not change the marginal probabilities. However, suppose that 8% of Alamo cartridges are defective and 12% of South Jersey cartridges are defective. If it is determined that the randomly selected cartridge is defective, the probabilities that the printer cartridge came from a particular supplier will change because a higher percentage of South Jersey cartridges are defective than those produced by Alamo. Each of these defective probabilities by supplier is a conditional probability. That is, the probability that a cartridge is defective given that is comes from Alamo is .08. This can be written as $P(\text{defective}|\text{Alamo}) = .08$. For South Jersey, the conditional probability is $P(\text{defective}|\text{South Jersey}) = .12$.

Next, we solve for the joint probabilities that the cartridge is from a particular supplier and is defective. If 65% of the cartridges are produced by Alamo and if 8% of Alamo cartridges are defective, then (.65)(.08) = .052 or 5.2% of all cartridges used by the American Printer Company are Alamo cartridges and defective. We have determined the joint probability of a cartridge being from Alamo and defective by multiplying the marginal probability by the conditional probability:

$$P(\text{Alamo}\cap\text{defective}) = P(\text{Alamo})\cdot P(\text{defective}|\text{Alamo}) = (.65)(.08) = .052$$

In a similar manner, we can solve for the joint probability for South Jersey:

$$P(\text{South Jersey}\cap\text{defective}) = P(\text{South Jersey})\cdot P(\text{defective}|\text{South Jersey}) = (.35)(.12) = .042.$$

From these probabilities, we can build the Bayesian table shown in **Table 4.8.**

Because it was determined that the randomly selected cartridge is defective, we need not be concerned about good cartridges and can focus on defective cartridges only, as in Table 4.8. We know that 5.2% of all cartridges are from Alamo and are defective and 4.2% of all cartridges are from South Jersey and are defective. What percentage of all cartridges is defective? If we sum the two joint probabilities from both of the suppliers, .052 + .042 = .094, we find that .094 or 9.4% of all cartridges are defective. That is, P(defective) = .094. Notice that until now, we only knew that probability of each supplier producing a defective cartridge, not the total.

Based on this information, we can now revise the original probabilities that the cartridge came from a particular supplier by following the rest of Bayes' rule. The probabilities are revised by dividing each supplier's joint probability by the total probability that a cartridge is defective (.094). For Alamo:

$$\text{Revised Probability} = \frac{P(\text{Alamo}\cap\text{Defective})}{P(\text{Defective})} = \frac{.052}{.094} = .553$$

TABLE 4.8: BUILDING THE BAYESIAN TABLE: CALCULATING THE JOINT PROBABILITY

Supplier	Prior Probability $P(S_i)$	Conditional Probability $P(d\|S_i)$	Joint Probability $P(d\cap S_i)$
Alamo	.65	.08	.052
South Jersey	.35	.12	.042

Given that the cartridge is defective, the probability that a randomly selected cartridge is from Alamo is .553. Shown in **Table 4.9** are the calculations for both revised probabilities.

Tree diagrams are another common way to solve Bayes' rule problems. **Figure 4.13** shows the solution for the cartridge problem. Note that the tree diagram contains all possibilities, including both defective and acceptable cartridges. When new information is given, only the pertinent branches are selected and used. The joint probability values at the end of the appropriate branches are used to revise and compute the posterior probabilities.

Using the total number of defective cartridges, .052 + .042 = .094, the calculation is as follows.

$$\text{Revised Probability: Alamo} = \frac{.052}{.094} = .553$$

$$\text{Revised Probability: South Jersey} = \frac{.042}{.094} = .447$$

TABLE 4.9: REVISED PROBABILITIES FOR THE PRINTER CARTRIDGE PROBLEM

Supplier	Prior Probability $P(S_i)$	Conditional Probability $P(d\|S_i)$	Joint Probability $P(d \cap S_i)$	Revised or Posterior Probabilities
Alamo	.65	.08	.052	.052/.094 = .553
South Jersey	.35	.12	.042	.042/.094 = .447
		P(defective) = .094		

Figure 4.13: Tree Diagram for Cartridge Problem Probabilities

DEMONSTRATION PROBLEM 4.12

Machines A, B, and C all produce the same two parts, X and Y. Of all the parts produced, machine A produces 60%, machine B produces 30%, and machine C produces 10%. In addition,

40% of the parts made by machine A are part X.

50% of the parts made by machine B are part X.

70% of the parts made by machine C are part X.

A part produced by this company is randomly sampled and is determined to be an X part. With the knowledge that it is an X part, revise the probabilities that the part came from machine A, B, or C.

Solution: The prior probability of the part coming from machine A is .60, because machine A produces 60% of all parts. The prior probability is .30 that the part came from B and .10 that it came from C. These prior probabilities are more pertinent if nothing is known about the part. However, the part is known to be an X part. The conditional probabilities show that different machines produce different proportions of X parts. For example, .40 of the parts made by machine A are X parts, but .50 of the parts made by machine B and .70 of the parts made by machine C are X parts. It makes sense that the probability of the part coming from machine C would increase and that the probability that the part was made on machine A would decrease because the part is an X part.

The following table shows how the prior probabilities, conditional probabilities, joint probabilities, and marginal probability, $P(X)$, can be used to revise the prior probabilities to obtain posterior probabilities.

Event	Prior $P(E_i)$	Conditional $P(X\|E_i)$	Joint $P(X \cap E_i)$	Posterior
A	.60	.40	(.60)(.40) = .24	$\frac{.24}{.46} = .52$
B	.30	.50	.15	$\frac{.15}{.46} = .33$
C	.10	.70	.07	$\frac{.07}{.46} = .15$
			$P(X) = .46$	

After the probabilities are revised, it is apparent that the probability of the part being made at machine A decreased and that the probabilities that the part was made at machines B and C increased. A tree diagram presents another view of this problem.

$$\text{Revised Probabilities: Machine } A\text{: } \frac{.24}{.46} = .52$$

$$\text{Machine } B\text{: } \frac{.15}{.46} = .33$$

$$\text{Machine } C\text{: } \frac{.07}{.46} = .15$$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

31. The Bayes' theorem (or Rule) is an extension to the _____ law of probabilities.
32. Statisticians use Bayes' rule to_____ probabilities in light of new information. State whether the following statements are true/false:
33. The numerators of Bayes' rule and the law of conditional probability are the different.
34. Tree diagrams are another common way to solve Bayes' rule problems.
35. Bayes' theorem is a formula that extends the use of the law of conditional probabilities to allow revision of original probabilities with new information.

ACTIVITY

From the below mentioned link make the notes of Bayes' rule and solve the case "The Flu".

https://www.bayestheorem.net/

4.9 SUMMARY

- The study of probability addresses ways of assigning probabilities, types of probabilities, and laws of probabilities. Probabilities support the notion of inferential statistics. Using sample data to estimate and test hypotheses about population parameters is done with uncertainty. If samples are taken at random, probabilities can be assigned to outcomes of the inferential process.
- Three methods of assigning probabilities are (1) the classical method, (2) the relative frequency of occurrence method, and (3) subjective probabilities. The classical method can assign probabilities a priori, or before the experiment takes place. It relies on the laws and rules of probability. The relative frequency of occurrence method assigns probabilities based on historical data or empirically derived data. Subjective probabilities are based on the feelings, knowledge, and experience of the person determining the probability.
- Certain special types of events necessitate amendments to some of the laws of probability: mutually exclusive events and independent events. Mutually exclusive events are events that cannot occur at the same time, so the probability of their intersection is zero. With independent events, the occurrence of one has no impact or influence on the occurrence of the other. Some experiments, such as those involving coins or dice, naturally produce independent events. Other experiments produce independent events when the experiment is conducted with replacement. If events [illegible] multiplying the marginal probabilities, which is a special case of the law of multiplication.
- Three techniques for counting the possibilities in an experiment are the mn counting rule, the N^n possibilities, and combinations. The mn counting rule is used to determine how many total possible ways an experiment can occur in a series of sequential operations. The N^n formula is applied when sampling is being done with replacement or events are independent. Combinations are used to determine the possibilities when sampling is being done without replacement.
- Four types of probability are marginal probability, conditional probability, joint probability, and union probability. The general law of addition is used to compute the probability of a union. The general law of multiplication is used to compute joint probabilities. The conditional law is used to compute conditional probabilities.
- Bayes' rule is a method that can be used to revise probabilities when new information becomes available; it is a variation of the conditional law. Bayes' rule takes prior probabilities of events occurring and adjusts or revises those probabilities on the basis of information about what subsequently occurs.

KEY WORDS

1. **Event:** A subset of the outcomes of a process.
2. **Independent event:** Events A and B are independent if the outcome of event A has no effect on the outcome of event B.
3. **Mutually exclusive event:** Events A and B are mutually exclusive if the occurrence of event A implies event B can't occur.
4. **Random experiment:** It is an experiment whose outcome cannot be predicted with certainty, before the experiment is run.
5. **Outcomes:** The possible results of an event is called outcomes.
6. **Set:** A set is a well-defined collection of objects.
7. **Subset:** set A is a subset of set B if all of the elements of set A are contained in set B.
8. **Union:** combining the elements of two or more sets.
9. **Intersection:** the intersection of two sets is the set of elements that are in both sets.
10. **Venn diagram:** It is a graphic means of showing intersection and union of sets by representing them as bounded regions.
11. **Probability:** A measure of the likeliness that an event will happen.
12. **Experimental probability:** The estimated probability of an event; obtained by dividing the number of successful trials by the total number of trials.
13. **Conditional probability:** The probability that A will occur given that B has already occurred is the 'conditional probability of A given B' and is denoted by $P(A|B)$.
14. **Tree diagram:** A tree diagram displays all the possible outcomes of an event.

4.10 DESCRIPTIVE QUESTIONS

Calculating the Statistics

4.1. Use the values in the cross-tabulation table to solve the equations given.

		Variable 1	
		D	E
Variable 2	A	10	20
	B	15	5
	C	30	15

(a) $P(E) =$ ____

(b) $P(B \cup D) =$ ____

(c) $P(A \cap E) =$ ____

(d) $P(B|E) =$ ____

(e) $P(A \cup B) =$ ____

(f) $P(B \cap C) =$ ____

(g) $P(D|C) =$ ____

(h) $P(A|B) =$ ____

(i) Are variables 1 and 2 independent? Why or why not?

4.2. The following joint probability table contains a breakdown on the age and gender of U.S. physicians in a recent year, as reported by the American Medical Association.

U.S. PHYSICIANS IN A RECENT YEAR

		Age (years)					
		<35	35–44	45–54	55–64	>65	
Gender	Male	.11	.20	.19	.12	.16	.78
	Female	.07	.08	.04	.02	.01	.22
		.18	.28	.23	.14	.17	1.00

(a) What is the probability that one randomly selected physician is 35–44 years old?

(b) What is the probability that one randomly selected physician is both a woman and 45–54 years old?

(c) What is the probability that one randomly selected physician is a man or is 35–44 years old?

(d) What is the probability that one randomly selected physician is less than 35 years old or 55–64 years old?

(e) What is the probability that one randomly selected physician is a woman if she is 45–54 years old?

(f) What is the probability that a randomly selected physician is neither a woman nor 55–64 years old?

Testing Your Understanding

4.3. Purchasing Survey asked purchasing professionals what sales traits impressed them most in a sales representative. Seventy-eight percent selected "thoroughness." Forty percent responded "knowledge of your own product." The purchasing professionals were allowed to list more than one trait. Suppose 27% of the purchasing professionals listed both "thoroughness" and "knowledge of your own product" as sales traits that impressed them most. A purchasing professional is randomly sampled.

(a) What is the probability that the professional selected "thoroughness" or "knowledge of your own product"?

(b) What is the probability that the professional selected neither "thoroughness" nor "knowledge of your own product"?

(c) If it is known that the professional selected "thoroughness," what is the probability that the professional selected "knowledge of your own product"?

(d) What is the probability that the professional did not select "thoroughness" and did select "knowledge of your own product"?

4.4. The Steelcase Workplace Index studied the types of work-related activities that Americans did while on vacation in the summer. Among other things, 40% read work-related material. Thirty-four percent checked in with the boss. Respondents to the study were allowed to select more than one activity. Suppose that of those who read work-related material, 78% checked in with the boss. One of these survey respondents is selected randomly.

(a) What is the probability that while on vacation this respondent checked in with the boss and read work-related material?

(b) What is the probability that while on vacation this respondent neither read work-related material nor checked in with the boss?

(c) What is the probability that while on vacation this respondent read work-related material given that the respondent checked in with the boss?

(d) What is the probability that while on vacation this respondent did not check in with the boss given that the respondent read work-related material?

(e) What is the probability that while on vacation this respondent did not check in with the boss given that the respondent did not read work-related material?

(f) Construct a joint probability table for this problem.

4.5. A small independent physicians' practice has three doctors. Dr. Sarabia sees 41% of the patients, Dr. Tran sees 32%, and Dr. Jackson sees the rest. Dr. Sarabia requests blood tests on 5% of her patients, Dr. Tran requests blood tests on 8% of his patients, and Dr. Jackson requests

blood tests on 6% of her patients. An auditor randomly selects a patient from the past week and discovers that the patient had a blood test as a result of the physician visit. Knowing this information, what is the probability that the patient saw Dr. Sarabia? For what percentage of all patients at this practice are blood tests requested?

4.6. According to U.S. Census Bureau figures, 34.0% of all Americans are in the 0–24 age bracket, 13.3% are in the 25–34 age bracket, 13.3% are in the 35–44 age bracket, and 39.4% are in the 45 and older age bracket. A study by Jupiter Media Metrix determined that Americans use their leisure time in different ways according to age. For example, of those who are in the 45 and older age bracket, 39% read a book or a magazine more than 10 hours per week. Of those who are in the 0–24 age bracket, only 11% read a book or a magazine more than 10 hours per week. The percentage figures for reading a book or a magazine for more than 10 hours per week are 24% for the 25–34 age bracket and 27% the 35–44 age bracket. Suppose an American is randomly selected and it is determined that he or she reads a book or a magazine more than 10 hours per week. Revise the probabilities that he or she is in any given age category. Using these figures, what is the overall percentage of the U.S. population that reads a book or a magazine more than 10 hours per week?

4.7. A retail study by Deloitte revealed that 54% of adults surveyed believed that plastic, noncompostable shopping bags should be banned. Suppose 41% of adults regularly recycle aluminum cans and believe that plastic, noncompostable shopping bags should be banned. In addition, suppose that 60% of adults who do not believe that plastic, noncompostable shopping bags should be banned do recycle. If an adult is randomly selected.

(a) What is the probability that the adult recycles and does not believe that plastic, noncompostable shopping bags should be banned?

(b) What is the probability that the adult does recycle?

(c) What is the probability that the adult does recycle or does believe that plastic, noncompostable shopping bags should be banned?

(d) What is the probability that the adult does not recycle or does not believe that plastic, noncompostable shopping bags should be banned?

(e) What is the probability that the adult does not believe that plastic, noncompostable shopping bags should be banned given that the adult does recycle?

4.11 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

4.1.

		Variable 1		
		D	E	
Variable 2	A	10	20	30
	B	15	5	20
	C	30	15	45
		55	40	95

(a) $P(E) = 40/95 = .42105$

(b) $P(B \cup D) = P(B) + P(D) - P(B \cap D)$

$= 20/95 + 55/95 - 15/95 = 60/95 = .63158$

(c) $P(A \cap E) = 20/95 = .21053$

(d) $P(B|E) = 5/40 = .1250$

(e) $P(A \cup B) = P(A) + P(B) = 30/95 + 20/95$

$= 50/95 = .52632$

(f) $P(B \cap C) = .0000$ (mutually exclusive)

(g) $P(D|C) = 30/45 = .66667$

(h) $P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{.0000}{20/95} = .0000$ **(A and B are mutually exclusive)**

(i) $P(A) = P(A|D)$??

Does $30/95 = 10/95$??

Since, $.31579 \neq .18182$, Variables 1 and 2 are *not* independent.

4.2.

		Age(years)					
		<35	35–44	45–54	55–64	>65	
Gender	Male	.11	.20	.19	.12	.16	.78
	Female	.07	.08	.04	.02	.01	.22
		.18	.28	.23	.14	.17	1.00

(a) $P(35–44) = .28$

(b) $P(\text{Woman} \cap 45–54) = .04$

(c) $P(\text{Man} \cup 35–44) = P(\text{Man}) + P(35–44) - P(\text{Man} \cap 35–44) = .78 + .28 - .20 = .86$

(d) $P(<35 \cup 55–64) = P(<35) + P(55–64) = .18 + .14 = .32$

(e) $P(\text{Woman}|45–54) = P(\text{Woman} \cap 45–54)/P(45–54) = .04/.23 = .1739$

(f) $P(\text{not } W \cap \text{not } 55–64) = .11 + .20 + .19 + .16 = .66$

4.3. Let T = thoroughness Let K = knowledge

$P(T) = .78 \qquad P(K) = .40 \qquad P(T \cap K) = .27$

(a) $P(T \cup K) = P(T) + P(K) - P(T \cap K) = .78 + .40 - .27 = .91$

(b) $P(NT \cap NK) = 1 - P(T \cup K) = 1 - .91 = .09$

(c) $P(K|T) = P(K \cap T)/P(T) = .27/.78 = .3462$

(d) $P(NT \cap K) = P(NT) - P(NT \cap NK)$

but $\quad P(NT) = 1 - P(T) = .22$

$P(NT \cap K) = .22 - .09 = .13$

The joint probability table:

		K		
		Yes	No	
T	Yes	.27	.51	.78
	No	.13	.09	.22
		.40	.60	1.00

4.4. Let R = read

Let B = checked in the with boss

$P(R) = .40$ $P(B) = .34$ $P(B|R) = .78$

(a) $P(B \cap R) = P(R) \times P(B|R) = (.40)(.78) = \mathbf{.312}$

(b) $P(NR \cap NB) = 1 - P(R \cup B)$

but $P(R \cup B) = P(R) + P(B) - P(R \cap B) =$

$.40 + .34 - .312 = .428$

$P(NR \cap NB) = 1 - .428 = \mathbf{.572}$

(c) $P(R|B) = P(R \cap B)/P(B) = (.312)/(.34) = \mathbf{.9176}$

(d) $P(NB|R) = 1 - P(B|R) = 1 - .78 = \mathbf{.22}$

(e) $P(NB|NR) = P(NB \cap NR)/P(NR)$

but $P(NR) = 1 - P(R) = 1 - .40 = .60$

$P(NB|NR) = .572/.60 = \mathbf{.9533}$

(f) Joint probability table for Problem 4.5:

	B	NB	
R	.312	.088	.40
NR	.028	.572	.60
	.34	.66	1.00

4.5. Let S = Sarabia Let T = Tran Let J = Jackson Let B = Blood test

$P(S) = .41$ $P(T) = .32$ $P(J) = .27$

$P(B|S) = .05$ $P(B|T) = .08$ $P(B|J) = .06$

Event	Prior $P(E_i)$	Conditional $P(B\|E_i)$	Joint $P(B \cap E_i)$	Revised $P(B_i\|NS)$
S	.41	.05	.0205	.3291
T	.32	.08	.0256	.4109
J	.27	.06	.0162	.2600
			$P(B) = .0623$	

4.6.

Event	Prior $P(E_i)$	Conditional $P(RB\|E_i)$	Joint $P(RB \cap E_i)$	Revised $P(E_i\|RB)$
0–24	.340	.11	.03740	.03740/.25889 = .14446
25–34	.133	.24	.03192	.03192/.25889 = .12330
35–44	.133	.27	.03591	.03591/.25889 = .13871
≥ 45	.394	.39	.15366	.15366/.25889 = .59353
			$P(RB) = .25889$	

4.7. Let B = Believe plastic shopping bags should be banned

Let R = Recycle aluminum cans

$P(B) = .54$ $P(R \cap B) = .41$ $P(R|NB) = .60$

(a) $P(R \cap NB) = P(NB \cap R) = P(NB) \cdot P(R|NB$

but $P(NB) = 1 - P(B) = 1 - .54 = .46$ and $P(R|NB) = .60$.

$P(R \cap NB) = P(NB \cap R) = P(NB) \cdot P(R|NB) = (.46)(.60) = \mathbf{.2760}$.

(b) $P(R) = P(R \cap B) + P(R \cap NB) = .41 + .2760 = \mathbf{.6860}$.

(c) $P(R \cup B) = P(R) + P(B) - P(R \cap B) = .6860 + .54 - .41 = \mathbf{.8160}$

(d) $P(NR \cup NB) = P(NR) + P(NB) - P(NR \cap NB)$

But $P(NB) = .46$, $P(NR) = 1 - P(R) = 1 - .6860 = .3140$, and

$P(NR \cap NB) = 1 - P(R \cup B) = 1 - .8160 = .1840$, therefore

$P(NR \cup NB) = .3140 + .46 - .1840 = \mathbf{.5900}$

(e) $P(NB|R) = P(NB \cap R)/P(R) =$

$P(NB \cap R) = P(R \cap NB) = .2760$ and $P(R) = .6860$, therefore

$P(NB|R) = P(NB \cap R)/P(R) = .2760/.6860 = \mathbf{.4023}$.

Initial joint probability table for Problem 4.10:

		B		
		Y	N	
R	Y	.41		
	N			
		.54	.46	1.00

Multiply .60 by the .46 to get the intersection of R and not B = .2760.

Fill in the rest of the cells.

Final joint probability table for Problem 4.10:

		B		
		Y	N	
R	Y	.41	.276	.686
	N	.13	.184	.314
		.54	.46	1.000

4.12 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
4.2 Probability	1.	Subjective Method
	2.	Experiment
	3.	Relative Frequency
	4.	True
	5.	False
4.3 Structure of Probability	6.	Elementary Events
	7.	Union
	8.	Zero
	9.	False
	10.	True
4.4 Marginal, Union, Joint and Conditional Probabilities	11.	Marginal Probability
	12.	Intersection
	13.	Conditional
	14.	True
	15.	False
4.5 Addition Laws	16.	Union
	17.	Zero
	18.	True
	19.	False
	20.	True
4.6 Multiplication Laws	21.	Independent
	22.	cross-tabulation
	23.	True
	24.	False
	25.	True
4.7 Conditional Probability	26.	Given
	27.	Marginal
	28.	Independent
	29.	True
	30.	True
4.8 Revision of Probabilities: Bayes' Rule	31.	Conditional
	32.	Revise
	33.	False
	34.	True
	35.	True

5

CHAPTER

PROBABILITY DISTRIBUTIONS

CONTENTS

Part A Discrete Distributions

5.1 Introduction
5.2 Discrete Versus Continuous Distributions
Self Assessment Questions
Activity
5.3 Describing a Discrete Distribution
5.3.1 Mean, Variance, and Standard Deviation of Discrete Distributions
Self Assessment Questions
Activity
5.4 Binomial Distribution
5.4.1 Solving a Binomial Problem
5.4.2 Using the Binomial Table
5.4.3 Mean and Standard Deviation of a Binomial Distribution
5.4.4 Graphing Binomial Distributions
Self Assessment Questions
Activity
5.5 Poisson Distribution
5.5.1 Working Poisson Problems by Formula
5.5.2 What to Do When the Intervals are Different
5.5.3 Using the Poisson Tables
5.5.4 Mean and Standard Deviation of a Poisson Distribution
5.5.5 Graphing Poisson Distributions
Self Assessment Questions
Activity
5.6 Hypergeometric Distribution
Self Assessment Questions
Activity

Part B Continuous Distributions

5.7 The Uniform Distribution
5.7.1 Solving for the Height and Length of a Uniform Distribution
5.7.2 The Mean and Standard Deviation of a Uniform Distribution
5.7.3 Determining Probabilities in a Uniform Distribution
Self Assessment Questions
Activity
5.8 Normal Distribution
5.8.1 Characteristics of the Normal Distribution
5.8.2 Probability Density Function of the Normal Distribution
5.8.3 Standardized Normal Distribution
5.8.4 Solving for Probabilities Using the Normal Curve
5.8.5 Using Probabilities to Solve for the Mean, the Standard Deviation, or an x Value in a Normal Distribution
Self Assessment Questions
Activity
5.9 Using the Normal Curve to Approximate Binomial Distribution Problems
5.9.1 Correcting for Continuity
Self Assessment Questions
Activity
5.10 Exponential Distribution
5.10.1 Probabilities of the Exponential Distribution
Self Assessment Questions
Activity
5.11 Summary
5.12 Descriptive Questions
5.13 Solutions for Descriptive Questions
5.14 Answers and Hints

LEARNING OBJECTIVES

The overall learning objective of Chapter 5 is to help you understand a category of probability distributions that produces only discrete outcomes, thereby enabling you to:

- Define a random variable in order to differentiate between a discrete distribution and a continuous distribution
- Determine the mean, variance, and standard deviation of a discrete distribution
- Solve problems involving the binomial distribution using the binomial formula and the binomial table
- Solve problems involving the Poisson distribution using the Poisson formula and the Poisson table
- Solve problems involving the hypergeometric distribution using the hypergeometric formula
- Solve for probabilities in a continuous uniform distribution
- Solve for probabilities in a normal distribution using z scores and for the mean, the standard deviation, or a value of x in a normal distribution when given information about the area under the normal curve
- Solve problems from the discrete binomial distribution using the continuous normal distribution and correcting for continuity
- Solve for probabilities in an exponential distribution and contrast the exponential distribution to the discrete Poisson distribution

5.1 INTRODUCTION

Whereas Part A focussed on the characteristics and applications of discrete distribution, Part B concentrates on information about continuous distributions.

In statistical experiments involving chance, outcomes occur randomly. As an example, a battery manufacturer randomly selects three batteries from a large batch of batteries to be tested for quality. Each selected battery is to be rated as good or defective. The batteries are numbered from 1 to 3, a defective battery is designated with a D, and a good battery is designated with a G. All possible outcomes are shown in **Table 5.1**. The expression $D_1 G_2 D_3$ denotes one particular outcome in which the first and third batteries are defective and the second battery is good. In this chapter, we examine the probabilities of events occurring in experiments that produce such discrete distributions. In particular, we will study the binomial distribution, the Poisson distribution, and the hypergeometric distribution.

TABLE 5.1: ALL POSSIBLE OUTCOMES FOR THE BATTERY EXPERIMENT

$G_1 G_2 G_3$
$D_1 G_2 G_3$
$G_1 D_2 G_3$
$G_1 G_2 D_3$

TABLE 5.1: ALL POSSIBLE OUTCOMES FOR THE BATTERY EXPERIMENT

$D_1 D_2 G_3$
$D_1 G_2 D_3$
$G_1 D_2 D_3$
$D_1 D_2 D_3$

PART A DISCRETE DISTRIBUTIONS

5.2 DISCRETE VERSUS CONTINUOUS DISTRIBUTIONS

A **random variable** is *a variable that contains the outcomes of a chance experiment.* For example, suppose an experiment is to measure the arrivals of automobiles at a turnpike tollbooth during a 30-second period. The possible outcomes are: 0 cars, 1 car, 2 cars, . . . , n cars. These numbers (0, 1, 2, . . . , n) are the values of a random variable. Suppose another experiment is to measure the time between the completion of two tasks in a production line. The values will range from 0 seconds to n seconds. These time measurements are the values of another random variable. The two categories of random variables are (1) discrete random variables and (2) continuous random variables.

A random variable is a **discrete random variable** *if the set of all possible values is at most a finite or a countably infinite number of possible values.* In most statistical situations, discrete random variables produce values that are nonnegative whole numbers. For example, if six people are randomly selected from a population and how many of the six are left-handed is to be determined, the random variable produced is discrete. The only possible numbers of left-handed people in the sample of six are 0, 1, 2, 3, 4, 5, and 6. There cannot be 2.75 left-handed people in a group of six people; obtaining nonwhole number values is impossible. Other examples of experiments that yield discrete random variables include the following:

1. Randomly selecting 25 people who consume soft drinks and determining how many people prefer diet soft drinks
2. Determining the number of defective items in a batch of 50 items
3. Counting the number of people who arrive at a store during a five-minute period
4. Sampling 100 registered voters and determining how many voted for the president in the last election

The battery experiment described at the beginning of the chapter produces a distribution that has discrete outcomes. Any one trial of the experiment will contain 0, 1, 2, or 3 defective batteries. It is not possible to get 1.58 defective batteries. It could be said that discrete random variables are usually generated from experiments in which things are "counted" not "measured."

Continuous random variables *take on values at every point over a given interval.* Thus continuous random variables have no gaps or unassumed values. It could be said that continuous random variables are generated from experiments in which things are "measured" not "counted." For example, if a person is assembling a product component, the time it takes to accomplish that feat could be

any value within a reasonable range such as 3 minutes 36.4218 seconds or 5 minutes 17.5169 seconds. A list of measures for which continuous random variables might be generated would include time, height, weight, and volume. Examples of experiments that yield continuous random variables include the following:

1. Sampling the volume of liquid nitrogen in a storage tank
2. Measuring the time between customer arrivals at a retail outlet
3. Measuring the lengths of newly designed automobiles
4. Measuring the weight of grain in a grain elevator at different points of time

Once continuous data are measured and recorded, they become discrete data because the data are rounded off to a discrete number. Thus in actual practice, virtually all business data are discrete. However, for practical reasons, data analysis is facilitated greatly by using continuous distributions on data that were continuous originally.

The outcomes for random variables and their associated probabilities can be organized into distributions. The two types of distributions are **discrete distributions,** *constructed from discrete random variables,* and **continuous distributions,** *based on continuous random variables.*

In this text, three discrete distributions are presented:

1. binomial distribution
2. Poisson distribution
3. hypergeometric distribution

All three of these distributions are presented in this chapter.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. A _____ is a variable that contains the outcomes of a chance experiment.
2. A random variable is a _____ random variable if the set of all possible values is at most a finite or a countably finite number.
3. Continuous random variables are generated from experiments in which things are_____.

State whether the following statements are true/false:

4. Two categories of random variables are discrete random variables and marginal random variables.
5. Discrete random variables are usually generated from experiments in which things are "measured."

ACTIVITY

Create a table with four columns: First column having Serial Number, second column with Random Number from 0 to 1. For this column, ask your friends (at least 30) to choose any random number from 0 to 1.

ACTIVITY

For the third column, ask your friends that "On a scale of 1 to 5, how much do they like a football game". Last column will be the sum of the first and second columns. Can you identify which column values can be considered as random variable, discrete distribution and continuous distributions.

5.3 DESCRIBING A DISCRETE DISTRIBUTION

How can we describe a discrete distribution? One way is to construct a graph of the distribution and study the graph. The histogram is probably the most common graphical way to depict a discrete distribution.

Observe the discrete distribution in **Table 5.2.** An executive is considering out-of-town business travel for a given Friday. She recognizes that at least one crisis could occur on the day that she is gone and she is concerned about that possibility. Table 5.2 shows a discrete distribution that contains the number of crises that could occur during the day that she is gone and the probability that each number will occur. For example, there is a .37 probability that no crisis will occur, a .31 probability of one crisis, and so on. The histogram in **Figure 5.1** depicts the distribution given in Table 5.2. Notice that the x-axis of the histogram contains the possible outcomes of the experiment (number of crises that might occur) and that the y-axis contains the probabilities of these occurring.

It is readily apparent from studying the graph in Figure 5.1 that the most likely number of crises is 0 or 1. In addition, we can see that the distribution is discrete in that no probabilities are shown the values in between the whole number crises.

TABLE 5.2: DISCRETE DISTRIBUTION OF OCCURRENCE OF DAILY CRISES

Number of Crises	Probability
0	.37
1	.31
2	.18
3	.09
4	.04
5	.01

Figure 5.1: Histogram of Discrete Distribution of Crises Data

Prize (x)	Probability $P(x)$
$1,000	.00002
100	.00063
20	.00400
10	.00601
4	.02403
2	.08877
1	.10479
0	.77175

Solution: The mean is computed as follows.

Prize (x)	Probability $P(x)$	$x \cdot P(x)$
$1,000	.00002	.02000
100	.00063	.06300
20	.00400	.08000
10	.00601	.06010
4	.02403	.09612
2	.08877	.17754
1	.10479	.10479
0	.77175	.00000
	$\Sigma[x \cdot P(x)] = .60155$	
	$\mu = E(x) = \Sigma[x \cdot P(x)] = .60155$	

The expected payoff for a $1 ticket in this game is 60.2 cents. If a person plays the game for a long time, he or she could expect to average about 60 cents in winnings. In the long run, the participant will lose about $1.00 − .602 = .398, or about 40 cents a game. Of course, an individual will never win 60 cents in any one game.

Using this mean, μ = .60155, the variance and standard deviation can be computed as follows.

x	$P(x)$	$(x-\mu)^2$	$(x-\mu)^2 \cdot P(x)$
$1,000	.00002	998797.26190	19.97595
100	.00063	9880.05186	6.22443
20	.00400	376.29986	1.50520
10	.00601	88.33086	0.53087
4	.02403	11.54946	0.27753
2	.08877	1.95566	0.17360
1	.10479	0.15876	0.01664
0	.77175	0.36186	0.27927
		$\Sigma[(x-\mu)^2 \cdot P(x)] = 28.98349$	
		$\sigma^2 = \Sigma[(x-\mu^2) \cdot P(x)] = 28.98349$	

$$\sigma = \sqrt{\sigma^2} = \sqrt{\Sigma[(x-\mu)^2 \cdot P(x)]} = \sqrt{28.98349} = 5.38363$$

The variance is 28.98349 (dollars)^2 and the standard deviation is $5.38.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

6. The _____ is the most common graphical way to depict a discrete distribution.
7. If the process is repeated long enough, the average of the outcomes is most likely to approach a _____ value.
8. For descriptive measures for grouped data instead of using the value of the class midpoint, _____ are used.

State whether the following statements are true/false:

9. We can calculate variance using the outcomes (x) and probabilities of outcomes [$P(x)$].
10. x-axis of the histogram of the discrete distribution contains the possible outcomes of the experiment and that the y-axis contains the probabilities of these occurring.

ACTIVITY

You need to do a survey about the dressing style of corporate people. Through this survey, you need to check "What is the probability of different dressing styles of an executive?".

For this survey, you need to give different marks to the executives on their dressing style.

Mark 1: If they are in proper business suit.

Mark 2: If they do not have proper business suit but a corporate blazer.

Mark 3: If they are without blazer but with proper tie, boot and belt.

Mark 4: If they are just wearing formal trousers and shirts.

Mark 5: If they are in casual dress.

Mark given	Total frequency for different mark (X)	Probability (X/30)
1		
2		
3		
4		
5		

Using the table, also calculate mean, variance and standard deviation of the discrete distribution.

5.4 BINOMIAL DISTRIBUTION

Perhaps the most widely known of all discrete distributions is the **binomial distribution**. The binomial distribution has been used for hundreds of years. Besides its historical significance, the binomial distribution is studied by business statistics students because it is the basis for other important statistical

techniques. In studying various distributions, it is important to be able to differentiate among distributions and to know the unique characteristics and assumptions of each. To assist you in this undertaking, the assumptions of a binomial distribution are given here.

Assumptions of the Binomial Distribution

- The experiment involves n identical trials.
- Each trial has only two possible outcomes denoted as success or as failure.
- Each trial is independent of the previous trials.
- The terms p and q remain constant throughout the experiment, where the term p is the probability of getting a success on any one trial and the term $q = (1 - p)$ is the probability of getting a failure on any one trial.

As the word *binomial* indicates, any single trial of a binomial experiment contains only two possible outcomes. These two outcomes are labeled *success* or *failure*. Usually the outcome of interest to the researcher is labeled a success. For example, if a quality control analyst is looking for defective products, he would consider finding a defective product a success even though the company would not consider a defective product a success. If researchers are studying left-handedness, the outcome of getting a left-handed person in a trial of an experiment is a success. The other possible outcome of a trial in a binomial experiment is called a failure. The word *failure* is used only in opposition to success. In the preceding experiments, a failure could be to get an acceptable part (as opposed to a defective part) or to get a right-handed person (as opposed to a left-handed person). In a binomial distribution experiment, any one trial can have only two possible, mutually exclusive outcomes (right-handed/left-handed, defective/good, male/female, etc.).

The binomial distribution is a discrete distribution. In n trials, only x successes are possible, where x is a whole number between 0 and n inclusive. For example, if five parts are randomly selected from a batch of parts, only 0, 1, 2, 3, 4, or 5 defective parts are possible in that sample. In a sample of five parts, getting 2.714 defective parts is not possible, nor is getting eight defective parts possible.

In a binomial experiment, the trials must be independent. This constraint means that either the experiment is by nature one that produces independent trials (such as tossing coins or rolling dice) or the experiment is conducted with replacement. The effect of the independent trial requirement is that p, the probability of getting a success on one trial, remains constant from trial to trial. For example, suppose 5% of all parts in a bin are defective. The probability of drawing a defective part on the first draw is $p = .05$. If the first part drawn is not replaced, the second draw is not independent of the first, and the p value will change for the next draw. The binomial distribution does not allow for p to change from trial to trial within an experiment. However, if the population is large in comparison with the sample size, the effect of sampling without replacement is minimal, and the independence assumption essentially is met, that is, p remains relatively constant.

Generally, if the sample size, n, is less than 5% of the population, the independence assumption is not of great concern. Therefore the acceptable sample size for using the binomial distribution with samples taken *without* replacement is

where

n = sample size

N = population size

For example, suppose 10% of the population of the world is left-handed and that a sample of 20 people is selected randomly from the world's population. If the first person selected is left-handed—and the sampling is conducted without replacement—the value of $p = .10$ is virtually unaffected because the population of the world is so large. In addition, with many experiments the population is continually being replenished even as the sampling is being done. This condition often is the case with quality control sampling of products from large production runs.

There are many binomial distributions, each of which can be described or characterized by two parameters, n and p, where n is the size of the sample ($n > 1$) and p is the probability of getting a success on a single trial. For every different value of n or p, there is a different binomial distribution. With each binomial distribution, probability questions can be asked. Given here are three examples of business scenarios that produce a binomial probability question. In each one, try to determine the value of n and the value of p.

1. Data produced by the U.S. Bureau of Labor Statistics has shown that 34% of all new business establishments fail within the first two years. Suppose 25 business establishments that were newly established two years ago are randomly sampled. What is the probability that exactly three have failed?
2. Why do clients leave their vendors? According to Small Business Trends, 68% of small business clients who leave do so because they feel unappreciated, unimportant, and taken for granted. Suppose 20 business clients who have left a small business are randomly interviewed and asked why they left. What is the probability that 18 or more left because they felt unappreciated, unimportant, and taken for granted?
3. A study conducted by the Employment Benefit Research Institute showed that only 14% of workers feel "very confident" that they will have enough money to live comfortably in retirement. Suppose 16 workers are randomly selected. If the survey is accurate, what is the probability that exactly six feel "very confident" they will have enough money to live comfortably in retirement?

5.4.1 SOLVING A BINOMIAL PROBLEM

A survey of relocation administrators by Runzheimer International revealed several reasons why workers reject relocation offers. Included in the list were family considerations, financial reasons, and others. Four percent of the respondents said they rejected relocation offers because they received too little relocation help. Suppose five workers who just rejected relocation offers are randomly selected and interviewed. Assuming the 4% figure holds for all workers rejecting relocation, what is the probability that exactly one of the workers rejected the offer because of too little relocation help? Let's start by determining the probability that the *first* worker interviewed rejected the offer because of too little relocation help and the next four workers rejected the offer for other reasons.

Let T represent too little relocation help and R represent other reasons. The sequence of interviews for this problem is as follows:

$$T_1, R_2, R_3, R_4, R_5$$

The probability of getting this sequence of workers is calculated by using the special rule of multiplication for independent events (assuming the workers are independently selected from a large population of workers). If 4% of the workers rejecting relocation offers do so for too little relocation help, the probability of one person being randomly selected from workers rejecting relocation offers who does so for that reason is .04, which is the value of p. The other 96% of the workers who reject relocation offers do so for other reasons. Thus the probability of randomly selecting a worker from those who reject relocation offers who does so for other reasons is 1 − .04 = .96, which is the value for q. The probability of obtaining this sequence of five workers who have rejected relocation offers is

$$P(T_1 \cap R_2 \cap R_3 \cap R_4 \cap R_5) = (.04)(.96)(.96)(.96)(.96) = .03397$$

Obviously, in the random selection of workers who rejected relocation offers, the worker who did so because of too little relocation help could have been the second worker or the third or the fourth or the fifth. All the possible sequences of getting one worker who rejected relocation because of too little help and four workers who did so for other reasons follow.

$$T_1, R_2, R_3, R_4, R_5$$
$$R_1, T_2, R_3, R_4, R_5$$
$$R_1, R_2, T_3, R_4, R_5$$
$$R_1, R_2, R_3, T_4, R_5$$
$$R_1, R_2, R_3, R_4, T_5$$

Thus, as you can see, there are five different ways to get exactly one worker who rejected relocation because of too little relocation help.

The probability of each of these sequences occurring is calculated as follows:

$$(.04)(.96)(.96)(.96)(.96) = .03397$$
$$(.96)(.04)(.96)(.96)(.96) = .03397$$
$$(.96)(.96)(.04)(.96)(.96) = .03397$$
$$(.96)(.96)(.96)(.04)(.96) = .03397$$
$$(.96)(.96)(.96)(.96)(.04) = .03397$$

Note that in each case the final probability is the same. Each of the five sequences contains the product of .04 and four .96s. The commutative property of multiplication allows for the reordering of the five individual probabilities in any one sequence. The probabilities in each of the five sequences may be reordered and summarized as $(.04)^1 (.96)^4$. Each sequence contains the same five probabilities, which makes recomputing the probability of each sequence unnecessary. What *is* important is to determine how many different ways the sequences can be formed and multiply that figure by the probability of one sequence occurring. For the five sequences of this problem, the total probability of getting exactly one worker who rejected relocation because of too little relocation help in a random sample of five workers who rejected relocation offers is

$$5(.04)^1(.96)^4 = .16987$$

An easier way to determine the number of sequences than by listing all possibilities is to use *combinations* to calculate them. (The concept of combinations was introduced in Chapter 4.) Five workers are being sampled, so $n = 5$, and the problem is to get one worker who rejected a relocation offer because of too little relocation help, $x = 1$. Hence ${}_nC_x$ will yield the number of possible ways to get x successes in n trials. For this problem, ${}_5C_1$ tells the number of sequences of possibilities.

$${}_5C_1 = \frac{5!}{1!(5-1)!} = 5$$

Weighting the probability of one sequence with the combination yields

$${}_5C_1(.04)^1(.96)^4 = .16987$$

Using combinations simplifies the determination of how many sequences are possible for a given value of x in a binomial distribution.

As another example, suppose 70% of all Americans believe cleaning up the environment is an important issue. What is the probability of randomly sampling four Americans and having exactly two of them say that they believe cleaning up the environment is an important issue? Let E represent the success of getting a person who believes cleaning up the environment is an important issue. For this example, p = .70. Let N represent the failure of not getting a person who believes cleaning up is an important issue (N denotes not important). The probability of getting one of these persons is q = .30.

The various sequences of getting two E's in a sample of four follow.

$$E_1, E_2, N_3, N_4$$
$$E_1, N_2, E_3, N_4$$
$$E_1, N_2, N_3, E_4$$
$$N_1, E_2, E_3, N_4$$
$$N_1, E_2, N_3, E_4$$
$$N_1, N_2, E_3, E_4$$

Two successes in a sample of four can occur six ways. Using combinations, the number of sequences is

$${}_4C_2 = 6 \text{ ways}$$

The probability of selecting any individual sequence is

$$(.70)^2(.30)^2 = .0441$$

Thus the overall probability of getting exactly two people who believe cleaning up the environment is important out of four randomly selected people, when 70% of Americans believe cleaning up the environment is important, is

$${}_4C_2(.70)^2(.30)^2 = .2646$$

Generalizing from these two examples yields the binomial formula, which can be used to solve binomial problems.

Binomial Formula

$$P(x) = {}_nC_x \cdot p^x \cdot q^{n-x} = \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x}$$

where

n = the number of trials (or the number being sampled)

x = the number of successes desired

p = the probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

The binomial formula summarizes the steps presented so far to solve binomial problems. The formula allows the solution of these problems quickly and efficiently.

DEMONSTRATION PROBLEM 5.2

A Gallup survey found that 65% of all financial consumers were very satisfied with their primary financial institution. Suppose that 25 financial consumers are sampled. If the Gallup survey result still holds true today, what is the probability that exactly 19 are very satisfied with their primary financial institution?

Solution: The value of p is .65 (very satisfied), the value of $q = 1 - p = 1 - .65 = .35$ (not very satisfied), $n = 25$, and $x = 19$. The binomial formula yields the final answer.

$${}_{25}C_{19}(.65)^{19}(.35)^6 = (177{,}100)(.00027884)(.00183827) = .0908$$

If 65% of all financial consumers are very satisfied, about 9.08% of the time the researcher would get exactly 19 out of 25 financial consumers who are very satisfied with their financial institution. How many very satisfied consumers would one expect to get in 25 randomly selected financial consumers? If 65% of the financial consumers are very satisfied with their primary financial institution, one would expect to get about 65% of 25 or (.65)(25) = 16.25 very satisfied financial consumers. While in any individual sample of 25 the number of financial consumers who are very satisfied cannot be 16.25, business researchers understand that x values near 16.25 are the most likely occurrences.

DEMONSTRATION PROBLEM 5.3

Recall early in this section that according to Small Business Trends, 68% of small business clients who leave do so because they feel unappreciated, unimportant, and taken for granted. Suppose 20 business clients who have left a small business are randomly interviewed and asked why they left. What is the probability that 18 or more left because they felt unappreciated, unimportant, and taken for granted?

Solution: This problem must be worked as the union of three probabilities: $x = 18$, $x = 19$, and $x = 20$. In each problem, $n = 20$, $p = .68$, and $q = .32$. The binomial formula gives the following result:

$x = 18$		$x = 19$		$x = 20$	
${}_{20}C_{18} \cdot (.68)^{18} \cdot (.32)^2$	+	${}_{20}C_{19} \cdot (.68)^{19} \cdot (.32)^1$	+	${}_{20}C_{20} \cdot (.68)^{20} \cdot (.32)^0$	=
.0188	+	.0042	+	.0004	= .0234

If, indeed, 68% of small business clients leave because they feel unappreciated, unimportant, and taken for granted and 20 business clients who have left are randomly selected, the probability that 18 or more left because they felt unappreciated, unimportant, and taken for granted is .0234. Thus, this problem is the union of three probabilities. Whenever the binomial formula is used to solve for cumulative success (not an exact number), the probability of each x value must be solved and the probabilities summed.

5.4.2 USING THE BINOMIAL TABLE

Anyone who works enough binomial problems will begin to recognize that the probability of getting $x = 5$ successes from a sample size of $n = 18$ when $p = .10$ is the same no matter whether the five successes are left-handed people, defective parts, brand X purchasers, or any other variable. Whether the sample involves people, parts, or products does not matter in terms of the final probabilities. The essence of the problem is the same: $n = 18$, $x = 5$, and $p = .10$. Recognizing this fact, mathematicians constructed a set of binomial tables containing presolved probabilities.

Since two parameters, n and p, describe or characterize a binomial distribution, binomial distributions actually are a family of distributions. Every different value of n and/or every different value of p gives a different binomial distribution, and tables are available for various combinations of n and p values. Because of space limitations, the binomial tables presented in this text are limited. Table A.2 in Appendix A contains binomial tables. Each table is headed by a value of n. Nine values of p are presented in each table of size n. In the column below each value of p is the binomial distribution for that combination of n and p. **Table 5.5** contains a segment of Table A.2 with the binomial probabilities for $n = 20$.

TABLE 5.5: SOLUTION TO THE X < 10 PROBLEM USING THE BINOMIAL TABLE

$n = 20$

	Probability								
x	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	.122	.012	.001	.000	.000	.000	.000	.000	.000
1	.270	.058	.007	.000	.000	.000	.000	.000	.000
2	.285	.137	.028	.003	.000	.000	.000	.000	.000
3	.190	.205	.072	.012	.001	.000	.000	.000	.000
4	.090	.218	.130	.035	.005	.000	.000	.000	.000
5	.032	.175	.179	.075	.015	.001	.000	.000	.000
6	.009	.109	.192	.124	.037	.005	.000	.000	.000
7	.002	.055	.164	.166	.074	.015	.001	.000	.000
8	.000	.022	.114	.180	.120	.035	.004	.000	.000

(Continued)

$n = 20$									
	Probability								
x	.1	.2	.3	.4	.5	.6	.7	.8	.9
9	.000	.007	.065	.160	.160	.071	.012	.000	.000
10	.000	.002	.031	.117	.176	.117	.031	.002	.000
11	.000	.000	.012	.071	.160	.160	.065	.007	.000
12	.000	.000	.004	.035	.120	.180	.114	.022	.000
13	.000	.000	.001	.015	.074	.166	.164	.055	.002
14	.000	.000	.000	.005	.037	.124	.192	.109	.009
15	.000	.000	.000	.001	.015	.075	.179	.175	.032
16	.000	.000	.000	.000	.005	.035	.130	.218	.090
17	.000	.000	.000	.000	.001	.012	.072	.205	.190
18	.000	.000	.000	.000	.000	.003	.028	.137	.285
19	.000	.000	.000	.000	.000	.000	.007	.058	.270
20	.000	.000	.000	.000	.000	.000	.001	.012	.122

DEMONSTRATION PROBLEM 5.4

One study by CNNMoney reported that 60% of workers have less than $25,000 in total savings and investments (excluding the value of their home). If this is true and if a random sample of 20 workers is selected, what is the probability that fewer than 10 have less than $25,000 in total savings and investments?

Solution: In this problem, $n = 20$, $p = .60$, and $x < 10$, making this a cumulative problem. To solve this problem using the binomial probability formula, one would have to solve for $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$, and 9, a total of 10 different applications of the binomial formula, and then sum the individual answers to get the overall probability of $x < 10$. However, using binomial tables, this problem can easily be solved.

Note that Table 5.5 contains a binomial table for $n = 20$ and since the problem under consideration is for $n = 20$, we can use this binomial table to help us solve the problem. At the top of the binomial table, the word "probability" denotes the various values of p. In this problem, $p = .60$, so we peruse the top of the table until we find a column with .6. Down the left side of the binomial table are various values of x. At the intersection of p and each value of x lies the probability answer for that value of x. Looking at Table 5.5, we can see that the probabilities for $x = 0, 1, 2, 3, 4, 5, 6, 7, 8$, and 9 are .000, .000, .000, .000, .000, .001, .005, .015, .035, and .071, respectively. Summing these values results in an overall probability of .127 for $x < 10$. These probabilities are highlighted in Table 5.5.

5.4.3 MEAN AND STANDARD DEVIATION OF A BINOMIAL DISTRIBUTION

A binomial distribution has an expected value or a long-run average, which is denoted by μ. The value of μ is determined by $n \cdot p$. For example, if $n = 10$ and $p = .4$, then $\mu = n \cdot p = (10)(.4) = 4$. The long-run average or expected value means that, if n items are sampled over and over for a long time and if p is the probability of getting a success on one trial, the average number of successes per sample is expected to be $n \cdot p$. If 40% of all graduate business students at a large university are women and if random samples of 10 graduate business students are selected many times, the expectation is that, on average, four of the 10 students would be women.

Mean and Standard Deviation of a Binomial Distribution

$$\mu = n \cdot p$$

$$\sigma = \sqrt{n \cdot p \cdot q}$$

Examining the mean of a binomial distribution gives an intuitive feeling about the likelihood of a given outcome.

In the study previously mentioned where 14% of workers feel "very confident" that they will have enough money to live comfortably in retirement, if 20 workers are randomly selected, the expected number who feel "very confident" that they will have enough money to live comfortably in retirement is the mean of the binomial distribution $n = 20$ and $p = .14$:

$$\mu = n \cdot p = 20(.14) = 2.8$$

In the long run, if 20 workers are randomly selected over and over and if indeed 14% of workers feel "very confident" that they will have enough money to live comfortably in retirement, then the experiment should average 2.8 workers who feel "very confident" that they will have enough money to live comfortable in retirement. The probability values for $n = 20$ and $p = .14$ are quite low because $x = 6, 7, 8, \ldots, 20$ are not near the mean or expected value. If the binomial probability were computed for $x = 3$, near the mean, the probability is .2409, which is relatively high.

The standard deviation of a binomial distribution is denoted as σ and is equal to $\sqrt{n \cdot p \cdot q}$. The standard deviation for the worker confidence problem discussed above is:

$$\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{20(.14)(.86)} = 1.55$$

Chapter 6 shows that some binomial distributions are nearly bell shaped and can be approximated by using the normal curve. The mean and standard deviation of a binomial distribution are the tools used to convert these binomial problems to normal curve problems.

5.4.4 GRAPHING BINOMIAL DISTRIBUTIONS

The graph of a binomial distribution can be constructed by using all the possible x values of a distribution and their associated probabilities. The x values usually are graphed along the x-axis and the probabilities are graphed along the y-axis.

Table 5.6 lists the probabilities for three different binomial distributions: $n = 8$ and $p = .20$, $n = 8$ and $p = .50$, and $n = 8$ and $p = .80$. **Figure 5.2** displays Excel graphs for each of these three binomial distributions. Observe how the shape of the distribution changes as the value of p increases. For $p = .50$, the distribution is symmetrical. For $p = .20$ the distribution is skewed right and for $p = .80$ the distribution is skewed left. This pattern makes sense because the mean of the binomial distribution $n = 8$ and $p = .50$ is 4, which is in the middle of the distribution. The mean of the distribution $n = 8$ and $p = .20$ is 1.6, which results in the highest probabilities being near $x = 2$ and $x = 1$. This graph peaks early and stretches toward the higher values of x. The mean of the distribution $n = 8$ and $p = .80$ is 6.4, which results in the highest probabilities being near $x = 6$ and $x = 7$. Thus the peak of the distribution is nearer to 8 than to 0 and the distribution stretches back toward $x = 0$.

TABLE 5.6: PROBABILITIES FOR THREE BINOMIAL DISTRIBUTIONS WITH $N = 8$

	Probabilities for		
x	$p = .20$	$p = .50$	$p = .80$
0	.1678	.0039	.0000
1	.3355	.0312	.0001
2	.2936	.1094	.0011
3	.1468	.2187	.0092
4	.0459	.2734	.0459
5	.0092	.2187	.1468
6	.0011	.1094	.2936
7	.0001	.0312	.3355
8	.0000	.0039	.1678

Figure 5.2: Excel Graphs of Three Binomial Distributions with $n = 8$

In any binomial distribution the largest x value that can occur is n and the smallest value is zero. Thus the graph of any binomial distribution is constrained by zero and n. If the value of p for the distribution is not .50, this constraint will result in the graph "piling up" at one end and being skewed at the other end.

DEMONSTRATION PROBLEM 5.5

A manufacturing company produces 10,000 plastic mugs per week. This company supplies mugs to another company, which packages the mugs as part of picnic sets. The second company randomly samples 10 mugs sent from the supplier. If two or fewer of the sampled mugs are defective, the second company accepts the lot. What is the probability that the lot will be accepted if the mug manufacturing company actually is producing mugs that are 10% defective? 20% defective? 30% defective? 40% defective?

Solution: In this series of binomial problems, $n = 10$, $x \leq 2$, and p ranges from .10 to .40. From Table A.2—and cumulating the values—we have the following probability of $x \leq 2$ for each value of p and the expected value ($\mu = n \cdot p$):

p	Lot Accepted $P(x \leq 2)$	Expected Number of Defects (μ)
.10	.930	1.0
.20	.677	2.0
.30	.382	3.0
.40	.167	4.0

These values indicate that if the manufacturing company is producing 10% defective mugs, the probability is relatively high (.930) that the lot will be accepted by chance. For higher values of p, the probability of lot acceptance by chance decreases. In addition, as p increases, the expected value moves away from the acceptable values, $x \leq 2$. This move reduces the chances of lot acceptance.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

11. In any binomial distribution, the largest x value that can occur is n and the smallest value is _____.
12. Multiplication of number of success (n) and probability of event (p) is called as _____ in binomial distribution.
13. In a binomial experiment, the trials must be _____.

State whether the following statements are true/false:

14. The probability of failure can be calculated by adding 1 to the probability of success.
15. Each trial has only two possible outcomes denoted as success or as failure is one of the assumptions in binomial distribution.

ACTIVITY

https://www.tes.com/teaching-resource/binomial-distribution-activities-6420044

From the above link, take the below-mentioned tree chart:

A pair of fair dice is thrown. Now you need to throw the first dice. Write down the result, then throw the second dice, and write down the result again. Copy and complete the following tree diagram, which shows the possible outcomes.

Let E be the event that exactly **one** 4 occurs when a pair of dice is thrown.

Reference:

https://www.tes.com/teaching-resource/binomial-distribution-activities-6420044

5.5 POISSON DISTRIBUTION

A second discrete distribution is the **Poisson distribution**, seemingly different from the binomial distribution, but actually derived from the binomial distribution. Unlike the binomial distribution in which samples of size n are taken from a population with a proportion p, the Poisson distribution, named after Simeon-Denis Poisson (1781–1840), a French mathematician, focuses on the number of discrete occurrences over some interval or continuum.

The Poisson distribution describes the occurrence of *rare events*. In fact, the Poisson formula has been referred to as the *law of improbable events*. For example, serious accidents at a chemical plant are rare, and the number per month might be described by the Poisson distribution. The Poisson distribution often is used to describe the number of random arrivals per some time interval. If the number of arrivals per interval is too frequent, the time interval can be reduced enough so that a rare number of occurrences is expected. Another example of a Poisson distribution is the number of random customer arrivals per five-minute interval at a small boutique on weekday mornings.

In the field of management science, models used in queuing theory (theory of waiting lines) are usually based on the assumption that the Poisson distribution is the proper distribution to describe random arrival rates over a period of time. In the area of statistical quality control, the Poisson distribution is the basis for the c control chart used to track the number of nonconformances per item or unit.

The Poisson distribution has the following characteristics:

- It is a discrete distribution.
- It describes rare events.
- Each occurrence is independent of the other occurrences.
- It describes discrete occurrences over a continuum or interval.
- The occurrences in each interval can range from zero to infinity.
- The expected number of occurrences must hold constant throughout the experiment.

Examples of Poisson-type situations include the following:

1. Number of telephone calls per minute at a small business
2. Number of hazardous waste sites per county in the United States
3. Number of arrivals at a turnpike tollbooth per minute between 3 A.M. and 4 A.M. in January on the Kansas Turnpike
4. Number of sewing flaws per pair of jeans during production
5. Number of times a tire blows on a commercial airplane per week

Each of these examples represents a rare occurrence of events for some interval. Note that, although time is a more common interval for the Poisson distribution, intervals can range from a county in the United States to a pair of jeans. Some of the intervals in these examples might have zero occurrences. Moreover, the average occurrence per interval for many of these examples is probably in the single digits (1–9).

If a Poisson-distributed phenomenon is studied over a long period of time, a *long-run average* can be determined. This average is denoted **lambda** (λ). Each Poisson problem contains a lambda value from which the probabilities of particular occurrences are determined. Although n and p are required to describe a binomial distribution, a Poisson distribution can be described by λ alone. The Poisson formula is used to compute the probability of occurrences over an interval for a given lambda value.

Poisson Formula

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

where

$x = 0, 1, 2, 3, \ldots$

λ = long-run average

$e = 2.718282$

Here, x is the number of occurrences per interval for which the probability is being computed, λ is the long-run average, and $e = 2.718282$ is the base of natural logarithms.

A word of caution about using the Poisson distribution to study various phenomena is necessary. The λ value must hold constant throughout a Poisson experiment. The researcher must be careful not to apply a given lambda to intervals for which lambda changes. For example, the average number of customers arriving at a Macy's store during a one-minute interval will vary from hour to hour, day to day, and month to month. Different times of the day or week might produce different lambdas. The number of flaws per pair of jeans might vary from Monday to Friday. The researcher should be specific in describing the interval for which λ is being used.

5.5.1 WORKING POISSON PROBLEMS BY FORMULA

Suppose bank customers arrive randomly on weekday afternoons at an average of 3.2 customers every 4 minutes. What is the probability of exactly 5 customers arriving in a 4-minute interval on a weekday afternoon? The lambda for this problem is 3.2 customers per 4 minutes. The value of x is 5 customers per 4 minutes. The probability of 5 customers randomly arriving during a 4-minute interval when the long-run average has been 3.2 customers per 4-minute interval is

$$\frac{(3.2)^5(e^{-3.2})}{5!} = \frac{(335.54)(.0408)}{120} = .1140$$

If a bank averages 3.2 customers every 4 minutes, the probability of 5 customers arriving during any one 4-minute interval is .1140.

5.5.2 WHAT TO DO WHEN THE INTERVALS ARE DIFFERENT

In actuality, the interval of a Poisson problem is used to assure the statistician that the lambda value and the value of x can be compared because they are recorded on the same basis. That is, having the same interval assures us of a "level playing field". Note that in the bank example shown in above, both λ and x were over a 4-minute interval. Suppose, however, the intervals are in the same units (minutes, county, manufactured item), but are not equal (e.g., 2 minutes for lambda but 5 minutes for x)? How do we account for this? Because lambda is a long-run average, we adjust lambda and its interval so that the interval for lambda is the same as the interval for x. If a bank is averaging 3.2 customers in 4 minutes, then it is also averaging 6.4 customers in 8 minutes and 1.6 customers in 2 minutes. **Never change the x interval!** Just because 3 customers arrive in 10 minutes does not mean that 6 customers will arrive in 20 minutes.

As an example, suppose that on Saturday mornings, a specialty clothing store averages 2.4 customer arrivals every 10 minutes. What is the probability that, on a given Saturday morning, 2 customers will arrive at the store in a 6-minute interval? The problem can be summarized as:

$$\lambda = 2.4 \text{ customers}/10 \text{ minutes}$$

$$x = 2 \text{ customers}/6 \text{ minutes}$$

Since the intervals are not the same, they must be reconciled. Remember, do not change the x interval—only the interval for lambda. Thus, we need to change the λ interval from 10 minutes to 6 minutes to match the x interval. One way is to recognize that 6 minutes (x interval) is 0.6 of 10 minutes. If we take 0.6 of both the number of customers and the interval for lambda, we compute the new lambda as:

$$0.6(2.4 \text{ customers}) = 1.44 \text{ customers and } 0.6(10 \text{ minutes}) = 6 \text{ minutes}$$

$$\lambda = 1.44 \text{ customers}/6 \text{ minutes}$$

We can use this new lambda and the old x interval of $x = 2$ customers/6 minutes to solve the problem as:

$$\frac{\lambda^x \cdot e^{-\lambda}}{x!} = \frac{1.44^2 \cdot e^{-1.44}}{2!} = .2456$$

In this case, it was relatively easy to change the interval for lambda into the x interval. Suppose, on the other hand, $\lambda = 2.8$ customers/7 minutes and $x =$ 4 customers/11 minutes. While there are at least a couple of ways to reconcile the two intervals, one sure way is to divide the lambda interval by itself first. As an example, divide this lambda value and the interval by 7, resulting in $\lambda = 0.4$ customers/1 minute. Now multiply this interval by the x interval, which in this case is 11, resulting in a new lambda of $\lambda = 4.4$ customers/11 minutes. This new problem could be worked as:

$$\lambda = 4.4 \text{ customers}/11 \text{ minutes}$$

$$x = 4 \text{ customers}/11 \text{ minutes}$$

5.5.3 USING THE POISSON TABLES

For every new and different value of lambda, there is a different Poisson distribution. However, the Poisson distribution for a particular lambda is the same regardless of the nature of the interval associated with that lambda. Because of this, over time statisticians have recorded the probabilities of various Poisson distribution problems and organized these into what we refer to as the Poisson tables. More recently, computers have been used to solve virtually any Poisson distribution problem, adding to the size and potential of such Poisson tables. In Table A.3 of Appendix A, Poisson distribution tables are given for selected values of lambda. These tables can be used to solve many different Poisson problems and are especially useful when a researcher is working a cumulative problem. **Table 5.7** presents a portion of Table A.3 that contains the probabilities of $x \leq 9$ if lambda is 1.6.

BHARATI BHAWAN

TABLE 5.7: POISSON TABLE FOR $\lambda = 1.6$

x	Probability
0	.2019
1	.3230
2	.2584
3	.1378
4	.0551
5	.0176
6	.0047
7	.0011
8	.0002
9	.0000

DEMONSTRATION PROBLEM 5.6

Suppose that during the noon hour in the holiday season, a UPS store averages 2.3 customers every minute and that arrivals at such a store are Poisson distributed. During such a season and time, what is the probability that more than four customers will arrive in a given minute?

Solution: This is a Poisson probability problem with $\lambda = 2.3$ and $x > 4$. To work this problem using the Poisson probability formula would involve many steps because the formula can only calculate probabilities for one x value at a time. Thus, to solve this problem using the formula, we would have to solve for each x of 5, 6, 7, 8, 9, 10, 11, 12, etc. and sum the answers. However, using the Poisson table for $\lambda = 2.3$, we need merely look up the answers in the table and sum them. Shown below are the Poisson table values for $\lambda = 2.3$ and the solution to this problem.

X	$\lambda = 2.3$
0	.1003
1	.2306
2	.2652
3	.2033
4	.1169
5	.0538
6	.0206
7	.0068
8	.0019
9	.0005
10	.0001
11	.0000
12	.0000

$X > 4$

.0837

If a store is averaging 2.3 customers per minute, the probability that more than four customers would arrive in any one minute is the sum of these table values, or .0837.

5.5.4 MEAN AND STANDARD DEVIATION OF A POISSON DISTRIBUTION

The mean or expected value of a Poisson distribution is λ. It is the long-run average of occurrences for an interval if many random samples are taken. Lambda usually is not a whole number, so most of the time actually observing lambda occurrences in an interval is impossible.

For example, suppose $\lambda = 6.5$/interval for some Poisson-distributed phenomenon. The resulting numbers of x occurrences in 20 different random samples from a Poisson distribution with $\lambda = 6.5$ might be as follows.

6 9 7 4 8 7 6 6 10 6 5 5 8 4 5 8 5 4 9 10

Computing the mean number of occurrences from this group of 20 intervals gives 6.6. In theory, for infinite sampling the long-run average is 6.5. Note from the samples that, when λ is 6.5, several 5s and 6s occur. Rarely would sample occurrences of 1, 2, 3, 11, 12, 13, . . . occur when $\lambda = 6.5$. Understanding the mean of a Poisson distribution gives a feel for the actual occurrences that are likely to happen.

The variance of a Poisson distribution also is λ. The standard deviation is $\sqrt{\lambda}$. Thus, for the Poisson distribution with $\lambda = 6.5$, the variance is 6.5 and the standard deviation is $\sqrt{6.5} = 2.55$.

5.5.5 GRAPHING POISSON DISTRIBUTIONS

The values in Table A.3, Appendix A, can be used to graph a Poisson distribution. The x values are on the x-axis and the probabilities are on the y-axis. **Figure 5.3** is a graph for the distribution of values for $\lambda = 1.6$.

Figure 5.3: Graph of the Poisson Distribution for $\lambda = 1.6$

The graph reveals a Poisson distribution skewed to the right. With a mean of 1.6 and a possible range of x from zero to infinity, the values obviously will "pile up" at 0, 1, and 2. Consider, however, the graph of the Poisson distribution for $\lambda = 6.5$ in **Figure 5.4**. Note that with $\lambda = 6.5$, the probabilities are greatest for the values of 5, 6, and 7. The graph has less skewness, because

the probability of occurrence of values near zero is small, as are the probabil-ities of large values of x.

Figure 5.4: Graph of the Poisson Distribution for $\lambda = 6.5$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

16. Poisson formula has been referred to as the law of _____ events.
17. In Poisson distribution, lambda (λ) is a _____.

State whether the following statements are true/false:

18. Poisson distribution is derived from the binomial distribution.
19. In Poisson distribution, each occurrence is dependent on the other occurrences.

ACTIVITY

You need to calculate Poisson distribution table for values from 0 to 7. For this, consider lambda (Mean) = 1.4, 4.2, 7.4. Also draw a histogram for the different values of lambdas. Explore which value of lambda gives normal distribution curve and why.

5.6 HYPERGEOMETRIC DISTRIBUTION

Another discrete statistical distribution is the hypergeometric distribution. Statisticians often use the **hypergeometric distribution** to complement the types of analyses that can be made by using the binomial distribution. Recall that the binomial distribution applies, in theory, only to experiments in which the trials are done with replacement (independent events). The hypergeometric distribution applies only to experiments in which the trials are done without replacement.

The hypergeometric distribution, like the binomial distribution, consists of two possible outcomes: success and failure. However, the user must know the size of the population and the proportion of successes and failures in the population to apply the hypergeometric distribution. In other words, because the hypergeometric distribution is used when sampling is done without replacement, information about population makeup must be known in order to redetermine the probability of a success in each successive trial as the probability changes.

The hypergeometric distribution has the following characteristics:

- It is a discrete distribution.
- Each outcome consists of either a success or a failure.
- Sampling is done without replacement.
- The population, N, is finite and known.
- The number of successes in the population, A, is known.

Hypergeometric Formula

$$P(x) = \frac{{}_AC_x \cdot {}_{N-A}C_{n-x}}{{}_NC_n}$$

where

N = size of the population

n = sample size

A = number of successes in the population

x = number of successes in the sample; sampling is done *without* replacement

A hypergeometric distribution is characterized or described by three parameters: N, A, and n. Because of the multitude of possible combinations of these three parameters, creating tables for the hypergeometric distribution is practically impossible. Hence, the researcher who selects the hypergeometric distribution for analyzing data must use the hypergeometric formula to calculate each probability. Because this task can be tedious and time-consuming, most researchers use the hypergeometric distribution as a fall-back position when working binomial problems without replacement. Even though the binomial distribution theoretically applies only when sampling is done with replacement and p stays constant, recall that, if the population is large enough in comparison with the sample size, the impact of sampling without replacement on p is minimal. Thus the binomial distribution can be used in some situations when sampling is done without replacement. Because of the tables available, using the binomial distribution instead of the hypergeometric distribution whenever possible is preferable. As a rule of thumb, if the sample size is less than 5% of the population, use of the binomial distribution rather than the hypergeometric distribution is acceptable when sampling is done without replacement. The hypergeometric distribution yields the exact probability, and the binomial distribution yields a good approximation of the probability in these situations.

In summary, the hypergeometric distribution should be used instead of the binomial distribution when the following conditions are present:

1. Sampling is being done without replacement.
2. $n \geq 5\% N$.

Hypergeometric probabilities are calculated under the assumption of equally likely sampling of the remaining elements of the sample space.

As an application of the hypergeometric distribution, consider the following problem. Twenty-four people, of whom eight are women, apply for a job. If five of the applicants are sampled randomly, what is the probability that exactly three of those sampled are women?

This problem contains a small, finite population of 24, or $N = 24$. A sample of five applicants is taken, or $n = 5$. The sampling is being done without replacement, because the five applicants selected for the sample are five different people. The sample size is 21% of the population, which is greater than 5% of the population ($n/N = 5/24 = .21$). The hypergeometric distribution is the appropriate distribution to use. The population breakdown is $A = 8$ women (successes) and $N - A = 24 - 8 = 16$ men. The probability of getting $x = 3$ women in the sample of $n = 5$ is

$$\frac{{}_8C_3 \cdot {}_{16}C_2}{{}_{24}C_5} = \frac{(56)(120)}{42{,}504} = .1581$$

Conceptually, the combination in the denominator of the hypergeometric formula yields all the possible ways of getting n samples from a population, N, including the ones with the desired outcome. In this problem, there are 42,504 ways of selecting five people from 24 people. The numerator of the hypergeometric formula computes all the possible ways of getting x successes from the A successes available and $n - x$ failures from the $N - A$ available failures in the population. There are 56 ways of getting three women from a pool of eight, and there are 120 ways of getting two men from a pool of 16. The combinations of each are multiplied in the numerator because the joint probability of getting x successes and $n - x$ failures is being computed.

DEMONSTRATION PROBLEM 5.7

Suppose 18 major computer companies operate in the United States and 12 are located in California's Silicon Valley. If three computer companies are selected randomly from the entire list, what is the probability that one or more of the selected companies are located in the Silicon Valley?

Solution:

$$N = 18,\ n = 3,\ A = 12,\ \text{and } x \geq 1$$

This problem is actually three problems in one: $x = 1$, $x = 2$, and $x = 3$. Sampling is being done without replacement, and the sample size is 16.6% of the population. Hence this problem is a candidate for the hypergeometric distribution. The solution follows.

$$\begin{array}{ccccc} x = 1 & & x = 2 & & x = 3 \\ \dfrac{{}_{12}C_1 \cdot {}_6C_2}{{}_{18}C_3} & + & \dfrac{{}_{12}C_2 \cdot {}_6C_1}{{}_{18}C_3} & + & \dfrac{{}_{12}C_3 \cdot {}_6C_0}{{}_{18}C_3} = \\ .2206 & + & .4853 & + & .2696 = .9755 \end{array}$$

An alternative solution method using the law of complements would be one minus the probability that none of the companies is located in Silicon Valley, or

$$1 - P(x = 0 | N = 18,\ n = 3,\ A = 12)$$

Thus,

$$1 - \frac{{}_{12}C_0 \cdot {}_6C_3}{{}_{18}C_3} = 1 - .0245 = .9755$$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

20. The hypergeometric distribution applies only to experiments in which the trials are done _____ replacement.
21. In hypergeometric distribution, the population size N is _____ and known.

State whether the following statements are true/false:

22. In hypergeometric distribution each outcome consists of either a success or a failure as in a binomial distribution.
23. When sample size is less than 5% of the population, use of hypergeometric distribution will be better than using binomial distribution.
24. For hypergeometric distribution using Excel, we need only 2 values (sample size and Population size) for getting the output.

ACTIVITY

Use Excel to calculate hypergeometric distribution for the given data.

Sample size = 10

Population size = 30

Number of success in sample = 4

Number of success in Population is 15

Now you have to use sample size from 1 to 10 and try to plot a histogram. Is it normally distributed? Explain the result.

CASELET (PART B)

INDIAN RAILWAY CARGO SERVICES

The first railway on Indian subcontinent ran over a stretch of 21 miles from Bombay to Thane. The idea of a railway to connect Bombay with Thane, Kalyan and with the Thal and Bhore Ghats inclines first occurred to Mr. George Clark, the Chief Engineer of the Bombay Government, during a visit to Bhandup in 1843.

The formal inauguration ceremony was performed on 16th April 1853, when 14 railway carriages carrying about 400 guests left Bori Bunder at 3.30 pm "amidst the loud applause of a vast multitude and to the salute of 21 guns." The first passenger train steamed out of Howrah station destined for Hooghly, a distance of 24 miles, on 15th August, 1854. Thus, the first section of the East Indian Railway was opened to public traffic, inaugurating the beginning of railway transport on the Eastern side of the subcontinent. Today Indian Railways (IR) runs around 11,000 trains every day, of which 5,000 are freight trains.

In 1950–51, IR had 914,000 employees; but in 2016–17, the number has increased to 1,308,000 employees, which is the largest in the world for any organization. IR has Rs. 47,177 billion as total assets out of which Rs. 30,245 billion is at capital-at-charge. IR has 7,349 stations which helps many corporates to move their product to each part of the country. Railways loaded 1,161.66 million tons (mt) freight, up 4.7% from 1,108.79 mt in 2016–17. The years 2017 and 2018 saw the highest-ever loading of coal by railways at 555 mt. Number of passengers carried by railways increased by 0.82% in 2017–18 and 1% in 2016–17 over the previous years. However, there was a drop in passenger numbers by 3% in 2015–16, by 2.5% in 2014–15 and by 3% in 2013–14.

A passenger train has 16 tons of capacity for carrying parcel. Consequently, the IR has an annual parcel-carrying capacity of around 35 mt, of which the current utilization is 5 mt (14%). This generates revenue of about Rs. 500 crore. The cost of haulage and parcel office staff is Rs. 1,800 crore. Thus, the parcel segment is making a loss of Rs. 1,300 crore per annum. Freight transport is the major business for IR and accounts for nearly 70% of its revenues. IR was famous for loss making unit, but it made profit Rs. 4,913 crore in the year 2016–17. But, the real problem is customer satisfaction. Not only passenger, but also companies which transport their materials through railways are not happy with the services. Their major problem is delay which costs them too much, but lower charges and diverse landscape has left no choice for the companies.

It is noted that the average freight-goods train speed in any given year is 23.7 kmph, and it is uniformly distributed in a year with a minimum of 21 kmph and a maximum of 26.4 kmph. Whereas, the terminal dwell time (typically measured in hours, is the time a rail car resides in a rail terminal on the same train) is normally distributed for railcars in India with a mean of 27.5 hours and a standard deviation of 2.6 hours.

CASELET (PART B)

In India, 45% of all freight-goods trains arrive late which is also a major concern in the IR. Because of this, not only IR is having losses, but also many loyal customers are moving to new source of transportation. Due to this, probability of moving to the new style of transportation is increasing day by day.

Sources:

- http://www.indianrailways.gov.in/railwayboard/uploads/directorate/stat_econ/IRSP_2016-17/Annual_Report_Accounts_Eng/Statistical_Summary.pdf
- http://www.business-standard.com/article/economy-policy/indian-railways-posts-highest-ever-passenger-earnings-in-2017-18-118041700288_1.html

PART B: CONTINUOUS DISTRIBUTIONS

Continuous distributions are constructed from continuous random variables in which values are taken on for every point over a given interval and are usually generated from experiments in which things are "measured" as opposed to "counted." With continuous distributions, probabilities of outcomes occurring between particular points are determined by calculating the area under the curve between those points. In addition, the entire area under the whole curve is equal to 1. **Figure 5.5** highlights these points. The many continuous distributions in statistics include the uniform distribution, the normal distribution, the exponential distribution, the t distribution, the chi-square distribution, and the F distribution. This chapter presents the uniform distribution, the normal distribution, and the exponential distribution.

1. Continuous distributions are created from "measured" variables: *volume, distance, temperature, weight,* . . . , E.g. 16.429 cubic feet, 7.3852 km, 81.36 degrees, 2.6641 kg.
2. Area under a continuous distribution curve is 1.
3. Probabilities in a continuous distribution are found by determining the area.

Figure 5.5: Aspects of Continuous Distributions

5.7 THE UNIFORM DISTRIBUTION

Perhaps the simplest of all continuous distributions is the uniform or rectangular distribution. **A uniform distribution** is a continuous distribution in which the same height is obtained over a range of values.

Shown in **Figure 5.6** is a uniform distribution. Note that through the entire distribution from point a to point b, the height of the distribution remains constant and that the distribution takes the shape of a rectangle.

Can you think of phenomena in business that might be uniformly distributed? Consider the demand for breakfast cereal over time. Could it be relatively uniform? What about the demand for some clothing items like dress socks?

5.7.1 SOLVING FOR THE HEIGHT AND LENGTH OF A UNIFORM DISTRIBUTION

Consider again the uniform distribution shown in Figure 5.6. A uniform distribution takes on the shape of a rectangle. The length of the rectangle shown

Figure 5.6: A Uniform Distribution

in Figure 5.6 is $(b - a)$, the distance between the two points defining the ends of the rectangle.

Recall that the area under a continuous distribution curve is by definition equal to 1. But the area of a rectangle is length times width. In the graph displayed in Figure 5.6, the width of the rectangle is actually the height of the curve. Thus, the area of a uniform distribution is the length times the height and is equal to one:

$$\text{Area} = (\text{length})(\text{height}) = 1$$

But the length is $(b - a)$. Thus, Area $= (b - a)(\text{height}) = 1$

Solving this equation yields a formula for the height of a uniform distribution:

$$\text{Height} = \frac{1}{b-a}$$

Figure 5.7 shows a uniform distribution with its length and height.

Figure 5.7: A Uniform Distribution with Length and Height

As an example, suppose a production line is set up to manufacture machine braces in lots of five per minute during a shift. When the lots are weighed, variation among the weights is detected, with lot weights ranging from 41 to 47 grams in a uniform distribution. In this example, $a = 41$, $b = 47$, and the length of the uniform distribution is $b - a = 47 - 41 = 6$.

The height of the distribution is equal to $\frac{1}{b-a} = \frac{1}{47-41} = \frac{1}{6}$. **Figure 5.8** shows this uniform distribution with the length and height.

Figure 5.8: Uniform Distribution of Braces with Length and Height

5.7.2 THE MEAN AND STANDARD DEVIATION OF A UNIFORM DISTRIBUTION

The mean and the standard deviation of a uniform distribution can be calculated using formula 5.1.

Formula 5.1 Mean and Standard Deviation of a Uniform Distribution

$$\mu = \frac{a+b}{2} \qquad \sigma = \frac{b-a}{\sqrt{12}}$$

The mean and the standard deviation can be computed for the machine braces example as

$$\mu = \frac{a+b}{2} = \frac{41+47}{2} = 44$$

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{47-41}{3.4641} = 1.732$$

Figure 5.9 shows the uniform distribution for the braces example with the mean and the standard deviation

Figure 5.9: Mean and Standard Deviation of Braces

5.7.3 DETERMINING PROBABILITIES IN A UNIFORM DISTRIBUTION

Unlike discrete distributions, which have probability functions to yield probability values and in which it is possible to have a probability for a single point, in continuous distributions, probabilities are calculated by determining the area under the curve (or function) over an interval. As an example, note the se cond curve shown in Figure 5.5. Formula 5.2 is used to determine the probabilities of x for a uniform distribution between a and b.

Formula 5.2 Probabilities in a Uniform Distribution

$$P(x) = \frac{x_2 - x_1}{b-a}$$

where: $a \le x_1 \le x_2 \le b$

In a uniform distribution, since a and b define the outer boundaries of the distribution, the area between a and b is equal to one. For values of x between a and b, Formula 5.2 computes the proportion of the total area that lies between two values of x. Consider the machine braces problem with lot weights ranging from 41 to 47 grams (a = 41 and b = 47) in a uniform distribution. What is the probability that a randomly selected brace weighs between 42 and 45 grams (x_1 = 42 and x_2 = 45)? Using Formula 5.2, this probability can be computed as:

$$P(x) = \frac{x_2 - x_1}{b-a} = \frac{45-42}{47-41} = \frac{3}{6} = .5000$$

This solution is displayed in **Figure 5.10.**

Figure 5.10: Solution to the Machine Braces Probability Question

Typically, the values of x_1 and x_2, lie between a and b. What is the probability if $x_1 = a$ and $x_2 = b$? The probability of $x \ge b$ or of $x \le a$ is zero because there is no area above b or below a. If part of the range between x_1 and x_2 lies outside of the area between a and b, that portion will not be considered part of the probability.

As a second example using the machine braces problem, what is the probability that a randomly selected brace weighs more than 43 grams ($x_1 > 43$)?

Because the uniform distribution's upper bound is 47 grams ($b = 47$), there is no area beyond $x = 47$. Thus, x_2 is 47 and using Formula 5.2, the probability can be computed as:

$$P(x) = \frac{x_2 - x_1}{b - a} = \frac{47-43}{47-41} = \frac{4}{6} = .6667$$

This solution is displayed in **Figure 5.11.**

Figure 5.11: Solution to the Second Machine Braces Probability Question

DEMONSTRATION PROBLEM 5.8

Suppose the amount of time it takes to assemble a plastic module ranges from 27 to 39 seconds and that assembly times are uniformly distributed. Describe the distribution. What is the probability that a given assembly will take between 30 and 35 seconds? Fewer than 30 seconds?

Solution:

$$f(x) = \frac{1}{39-27} = \frac{1}{12}$$

$$\mu = \frac{a+b}{2} = \frac{39+27}{2} = 33$$

$$\sigma = \frac{b-a}{\sqrt{12}} = \frac{39-27}{\sqrt{12}} = \frac{12}{\sqrt{12}} = 3.464$$

The height of the distribution is 1/12. The mean time is 33 seconds with a standard deviation of 3.464 seconds.

$$P(30 \le x \le 35) = \frac{35-30}{39-27} = \frac{5}{12} = .4167$$

There is a .4167 probability that it will take between 30 and 35 seconds to assemble the module.

$$P(x < 30) = \frac{30-27}{39-27} = \frac{3}{12} = .2500$$

There is a .2500 probability that it will take less than 30 seconds to assemble the module. Because there is no area less than 27 seconds, $P(x < 30)$ is determined by using only the interval $27 \le x < 30$. In a continuous distribution, there is no area at any one point (only over an interval). Thus the probability $x < 30$ is the same as the probability of $x \le 30$.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

25. A uniform distribution is a _____ distribution in which the same height is obtained over a range of values.
26. The shape of the uniform distribution is _____.
27. The area of the uniform distribution (multiplication of the length and height) is _____.

State whether the following statements are true/false:

28. Excel does not have the capability of directly computing probabilities for the uniform distribution.
29. In continuous distributions, probabilities are calculated by determining the area under the curve over an interval.

ACTIVITY

Assume lower endpoint is 725 and upper endpoint is 775. Now in the place of X, you need to enter different values and then calculate the uniform distribution. Keep the value of X between 750 and 770.

5.8 NORMAL DISTRIBUTION

Probably the most widely known and used of all distributions is the **normal distribution.** It fits many human characteristics, such as height, weight, length, speed, IQ, scholastic achievement, and years of life expectancy, among others. Like their human counterparts, living things in nature, such as trees, animals, insects, and others, have many characteristics that are normally distributed.

Many variables in business and industry also are normally distributed. Some examples of variables that could produce normally distributed measurements

include the annual cost of household insurance, the cost per square foot of renting warehouse space, and managers' satisfaction with support from ownership on a five-point scale. In addition, most items produced or filled by machines are normally distributed.

Because of its many applications, the normal distribution is an extremely important distribution. Besides the many variables mentioned that are normally distributed, the normal distribution and its associated probabilities are an integral part of statistical process control. When large enough sample sizes are taken, many statistics are normally distributed regardless of the shape of the underlying distribution from which they are drawn. **Figure 5.12** is the graphic representation of the normal distribution: the normal curve.

Figure 5.12: The Normal Curve

5.8.1 CHARACTERISTICS OF THE NORMAL DISTRIBUTION

Figure 5.13 displays the characteristics of the normal distribution. Several of these will be useful to us in solving for probabilities under the normal curve. In addition, the normal distribution is a family of curves each defined by a particular combination of mean and standard deviation.

The normal distribution exhibits the following characteristics:

- It is a continuous distribution.
- It is a symmetrical distribution about its mean.
- It is asymptotic to the horizontal axis.
- It is unimodal.
- It is a family of curves.
- Area under the curve is 1.

Figure 5.13: Characteristics of the Normal Distribution

The normal distribution is symmetrical. Each half of the distribution is a mirror image of the other half. Many normal distribution tables contain probability values for only one side of the distribution because probability values for the other side of the distribution are identical because of symmetry.

In theory, the normal distribution is asymptotic to the horizontal axis. That is, it does not touch the x-axis, and it goes forever in each direction. The reality is that most applications of the normal curve are experiments that have finite limits of potential outcomes. For example, even though GMAT scores are analyzed by the normal distribution, the range of scores on the GMAT is from 200 to 800.

The normal curve sometimes is referred to as the *bell-shaped curve*. It is unimodal in that values *mound up* in only one portion of the graph—the center of the curve. The normal distribution actually is a family of curves. Every unique value of the mean and every unique value of the standard deviation result in a different normal curve. In addition, *the total area under any normal distribution is 1*. The area under the curve yields the probabilities, so the total of all probabilities for a normal distribution is 1. Because the distribution is symmetric, the area of the distribution on each side of the mean is 0.5.

5.8.2 PROBABILITY DENSITY FUNCTION OF THE NORMAL DISTRIBUTION

The normal distribution is described or characterized by two parameters: the mean, μ, and the standard deviation, σ. The values of μ and σ produce a normal distribution. The density function of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2[(x-\mu)/(\sigma)]^2}$$

where

μ = mean of x

σ = standard deviation of x

$\pi = 3.14159\ldots$, and

$e = 2.71828\ldots$.

Using integral calculus to determine areas under the normal curve from this function is difficult and time-consuming; therefore, virtually all researchers use table values to analyze normal distribution problems rather than this formula.

5.8.3 STANDARDIZED NORMAL DISTRIBUTION

The normal distribution is described or characterized by two parameters, the mean, μ, and the standard deviation, σ. That is, every unique pair of the values of μ and σ defines a different normal distribution. **Figure 5.14** shows the graphs of normal distributions for the following three pairs of parameters.

1. $\mu = 50$ and $\sigma = 5$
2. $\mu = 80$ and $\sigma = 5$
3. $\mu = 50$ and $\sigma = 10$

Figure 5.14: Normal Curves for Three Different Combinations of Means and Standard Deviations

Note that every change in a parameter (μ or σ) determines a different normal distribution. This characteristic of the normal curve (a family of curves) could make analysis by the normal distribution tedious because volumes of normal curve tables—one for each different combination of μ and σ—would be required. Fortunately, a mechanism was developed by which all normal distributions can be converted into a single distribution: the z distribution. This process yields the **standardized normal distribution** (or curve). The conversion formula for any x value of a given normal distribution follows.

z Formula

$$z = \frac{x - \mu}{\sigma}, \qquad \sigma \neq 0$$

A z **score** is *the number of standard deviations that a value, x, is above or below the mean.* If the value of x is less than the mean, the z score is negative; if the value of x is more than the mean, the z score is positive; and if the value of x equals the mean, the associated z score is zero. This formula allows conversion of the distance of any x value from its mean into standard deviation units. A standard z score table can be used to find probabilities for any normal curve problem that has been converted to z scores. The z **distribution** is *a normal distribution with a mean of 0 and a standard deviation of 1.* Any value of x at the mean of a normal curve is zero standard deviations from the mean. Any value of x that is one standard deviation above the mean has a z value of 1. The empirical rule, introduced in Chapter 3, is based on the normal distribution in which about 68% of all values are within one standard deviation of the mean regardless of the values of μ and σ. In a z distribution, about 68% of the z values are between $z = -1$ and $z = +1$.

The z distribution probability values are given in Table A.5. Because it is so frequently used, the z distribution is also printed inside the cover of this text. For discussion purposes, a list of z distribution values is presented in **Table 5.8.**

Table A.5 gives the total area under the z curve between 0 and any point on the positive z-axis. Since the curve is symmetric, the area under the curve between z and 0 is the same whether z is positive or negative (the sign on the z value designates whether the z score is above or below the mean). The table areas or probabilities are always positive.

TABLE 5.8: z DISTRIBUTION

	SECOND DECIMAL PLACE IN z									
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545

(Continued)

TABLE 5.8: z DISTRIBUTION

1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998									
4.0	.49997									
4.5	.499997									
5.0	.4999997									
6.0	.499999999									

5.8.4 SOLVING FOR PROBABILITIES USING THE NORMAL CURVE

Probabilities for intervals of any particular values of a normal distribution can be determined by using the mean, the standard deviation, the z formula, and the z distribution table. As an example, let's consider information collected and reported by the U.S. Environmental Protection Agency (EPA) on the generation and disposal of waste. According to EPA data, on average there is 4.43 pounds of waste generated per person in the U.S. per day. Suppose waste generated per person per day in the U.S. is normally distributed with a standard deviation of 1.32 pounds. If a U.S. person is randomly selected, what is the probability that the person generates more than 6.00 pounds of waste per day?

We can summarize this problem as:

$$P(x > 6.00 | \mu = 4.43 \text{ and } \sigma = 1.32) = ?$$

Figure 5.15 displays a graphical representation of the problem. Note that the area under the curve for which we are interested in solving is the tail of the distribution.

We begin the process by solving for the z value:

$$z = \frac{x - \mu}{\sigma} = \frac{6.00 - 4.43}{1.32} = 1.19$$

The z of 1.19 indicates that the value of x (6.00) is 1.19 standard deviations above the mean (4.43). Looking this z value up in the z distribution table yields a probability of .3830. Is this the answer? A cursory glance at the top of the z distribution table shows that the probability given in the table is for the area between a z of 0 and the given z value. Thus there is .3830 area (probability) between the mean of the z distribution ($z = 0$) and the z value of interest (1.19). However, we want to solve for the tail of the distribution or the area above $z = 1.19$. The normal curve has an area of 1 and is symmetrical. Thus, the area under the curve in the upper half is .5000. Subtracting .3830 from .5000 results in .1170 for the area of the upper tail of the distribution. The probability that a randomly selected person in the U.S. generates more than 6.00 pounds of waste per day is .1170 or 11.70%. **Figure 5.16** displays the solution to this problem.

Figure 5.15: Graphical Depiction of the Waste-Generation Problem with $x > 6.00$

Figure 5.16: Solution of the Waste-Generation Problem with $x > 6.00$

NOTES

DEMONSTRATION PROBLEM 5.9

Using this same waste-generation example, if a U.S. person is randomly selected, what is the probability that the person generates between 3.60 and 5.00 pounds of waste per day?

We can summarize this problem as:

$$P(3.60 < x < 5.00 | \mu = 4.43 \text{ and } \sigma = 1.32 = ?$$

Figure 5.17 displays a graphical representation of the problem. Note that the area under the curve for which we are solving crosses over the mean of the distribution. Note that there are two x values in this problem (x_1 = 3.60 and x_2 = 5.00). The z formula can handle only one x value at a time. Thus, this problem needs to be worked out as two separate problems and the resulting probabilities added together. We begin the process by solving for each z value:

$$z = \frac{x-\mu}{\sigma} = \frac{3.60-4.43}{1.32} = -0.63$$

$$z = \frac{x-\mu}{\sigma} = \frac{5.00-4.43}{1.32} = 0.43$$

Figure 5.17: Graphical Depiction of the Waste-Generation Problem with 3.60 <x<5.00

Next, we look up each z value in the z distribution table. Since the normal distribution is symmetrical, the probability associated with $z = -0.63$ is the same as the probability associated with $z = 0.63$. Looking up $z = 0.63$ in the table yields a probability of .2357. The probability associated with $z = 0.43$ is .1664. Using these two probability values, we can get the probability that $3.60 < x < 5.00$ by summing the two areas:

$$P(3.60 < x < 5.00 | \mu = 4.43 \text{ and } \sigma = 1.32) = .2357 + .1664 = .4021$$

The probability that a randomly selected person in the U.S. has between 3.60 and 5.00 pounds of waste generation per day is .4021 or 40.21%. **Figure 5.18** displays the solution to this problem.

NOTES

Figure 5.18: Solution of the Waste-Generation Problem with 3.60< x <5.00

DEMONSTRATION PROBLEM 5.10

Using this same waste-generation example, if a U.S. person is randomly selected, what is the probability that the person generates between 5.30 and 6.50 pounds of waste per day?

We can summarize this problem as:

$$P(5.30 < x < 6.50 | \mu = 4.43 \text{ and } \sigma = 1.32) = ?$$

Figure 5.19 displays a graphical representation of the problem. Note that the area under the curve for which we are solving lies completely on one side of the mean of the distribution.

Figure 5.19: Graphical Depiction of the Waste-Generation Problem with 5.30 <x < 6.50

There are two x values in this problem (x_1 = 5.30 and x_2 = 6.50). The z formula can handle only one x value at a time. Thus, this problem needs to be worked out as two separate problems. However, in this problem, the x values are on the same side of the mean. To solve for this probability, we will need to find the area (probability) between x_2 = 6.50 and the mean and the area (probability) between x_1 = 5.30 and the mean, and then subtract the two areas (probabilities). We begin the process by solving for each z value:

$$z = \frac{x-\mu}{\sigma} = \frac{6.50-4.43}{1.32} = 1.57$$

$$z = \frac{x-\mu}{\sigma} = \frac{5.30-4.43}{1.32} = 0.66$$

Next, we look up each z value in the z distribution table. The probability

associated with $z = 1.57$ is .4418. The probability associated with $z = 0.66$ is .2454. Using these two probability values, we can get the probability that $5.30 < x < 6.50$ by subtracting the two areas (probabilities):

$$P(5.30 < x < 6.50 | \mu = 4.43 \text{ and } \sigma = 1.32) = .4418 - .2454 = .1964.$$

The probability that a randomly selected person in the United States has between 5.30 and 6.50 pounds of waste generation per day is .1964 or 19.64%. **Figure 5.20** displays the solution to this problem.

Figure 5.20: Solution of the Waste-Generation Problem with $5.30 < x < 6.50$

5.8.5 USING PROBABILITIES TO SOLVE FOR THE MEAN, THE STANDARD DEVIATION, OR AN X VALUE IN A NORMAL DISTRIBUTION

Note that the z formula contains four variables: μ, σ, z, and x. In solving for probabilities, the researcher has knowledge of μ, σ, x and enters those values into the z formula:

$$z = \frac{x - \mu}{\sigma}$$

Looking up the resulting computed value of z in the z distribution table yields a probability value.

There are times, however, when a business researcher has probability information but wants to solve for one of the mean, the standard deviation, or an x value of a normal distribution. That is, we sometimes want to solve for one of the three variables in the right-hand side of the z formula. Typically, in such problems, we are given two of the three variables along with probability information that will help us solve for z.

As an example, Runzheimer International publishes business travel costs for various cities throughout the world. In particular, they publish per diem totals that represent the average costs for the typical business traveler, including three meals a day in business-class restaurants and single-rate lodging in business-class hotels. Suppose 86.65% of the per diem costs in Buenos Aires, Argentina, are less than \$449. Assuming that per diem costs are normally distributed and that the standard deviation of per diem costs in Buenos Aires is \$36, what is the average per diem cost in Buenos Aires?

What is given in this problem? The standard deviation, σ, is 36. The value of \$449 is the particular value of x that we are working with in this problem, $x = 449$. The mean is unknown. Entering these two values into the z formula results in:

$$z = \frac{449 - \mu}{36}$$

We need only secure the value of z, and μ can be solved for. To do this, we must use the probability information given and the z distribution table to find the value of z.

Study the layout of this problem in **Figure 5.21** in order to determine how to proceed. Think about how probabilities are given in the z distribution table.

Figure 5.21: Layout of Information for Per Diem Problem

How can we use the fact that 86.65% of the per diem rates are less than \$449, as shown in Figure 5.21? Recall that the z distribution table associated with this text always gives probability values for areas between a given x value and the mean. In this problem, if we subtract off the 50% of the area less than the mean, it leaves an area of 36.65% between the mean and x = \$449. Converting this to a proportion, we can say that .3665 of the area lies between the mean and x. Now, we can look up this value as an area (probability) in the z distribution table and back out the associated value of z as shown in **Figure 5.21**.

The resulting z, as shown in Figure 5.22, is $z = 1.11$. In problems such as this (solving for x, μ, or σ), the sign on the z value can matter. In this particular example, x is in the upper half of the distribution and is therefore positive. The standard deviation, σ, is 36, the value of $x = 449$, and $z = 1.11$. Using these values, μ can now be solved for as follows:

$$1.11 = \frac{449 - \mu}{36}$$

$$\mu = 409.04$$

The entries in this table are the probabilities that a standard normal random variable is between 0 and z (the shaded area)

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000		.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398		.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793		.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179		.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554		.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915						.2123	.2157	.2190	.2224
0.6	.2257						.2454	.2486	.2517	.2549
0.7	.2580						.2764	.2794	.2823	.2852
0.8	.2881		.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159		.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413		.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1		.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936

The z value associated with an area of .3665 is z = 1.11

Figure 5.22: Using the z Distribution Table to Locate the z Value Associated with an Area of .3665

DEMONSTRATION PROBLEM 5.11

As a second example, a particular ten-inch (diameter) clay pipe weighs, on average, 44 pounds, and pipe weights are normally distributed in the population. If 74.22% of the pipe weights are more than 40 pounds, what is the value of the standard deviation?

The mean, μ, is 44. The value of x is 40. The standard deviation is unknown. Entering these two values into the z formula results in:

$$z = \frac{40-44}{\sigma}$$

We need only secure the value of z, and σ can be solved for. To do this, we must use the probability information given and the z distribution table to find the value of z.

Study the layout of this problem in **Figure 5.23** in order to determine how to proceed. Think about how probabilities are given in the z distribution table.

Figure 5.23: Layout for the Clay Pipe Problem

How can we use the fact that 74.22% of the weights are more than 40 as shown in Figure 5.23? Recall that the z distribution table associated with this text always gives probability values for areas between a given x value and the mean. In this problem, if we subtract off the 50% of the area greater than the mean, it leaves an area of 24.22% between the mean and $x = 40$. Converting this to a proportion, we can say that .2422 of the area lies between the mean and x. Now, we can look up this value as an area (probability) in the z distribution table and back out the associated value of z, which is 0.65. In this example, since the value of x is to the left of the mean, we must manually assign a negative sign to the z value, resulting in $z = -0.65$.

The mean, μ, is 44, the value of $x = 40$, and $z = -0.65$. Using these values, σ can now be solved for as follows:

$$-0.65 = \frac{40-44}{\sigma}$$

$$\sigma = 6.154$$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

30. The _____ is a normal distribution with a mean of 0 and a standard deviation of 1.
31. Each half of the distribution of normal distribution is a _____ image of the other half.

State whether the following statements are true/false:

32. When large enough sample sizes are taken, many statistics are normally distributed.
33. Normal distribution is a continuous distribution.
34. Area under the curve in normal distribution is zero.

ACTIVITY

We have studied that normal distribution fits many human characteristics. Now write down at least 10 examples where normal distribution can be observed. You need to draw a normal curve and the example you have selected should follow the 68-95-99.7 rule.

5.9 USING THE NORMAL CURVE TO APPROXIMATE BINOMIAL DISTRIBUTION PROBLEMS

For certain types of binomial distribution problems, the normal distribution can be used to approximate the probabilities. As sample sizes become large, binomial distributions approach the normal distribution in shape regardless of the value of p. This phenomenon occurs faster (for smaller values of n) when p is near .50. **Figures 5.25** through **5.26** show three binomial distributions. Note in Figure 6.20 that even though the sample size, n, is only 10, the binomial graph bears a strong resemblance to a normal curve.

The graph in Figure 5.25 ($n = 10$ and $p = .20$) is skewed to the right because of the low p value and the small size. For this distribution, the expected value is only 2 and the probabilities pile up at $x = 0$ and 1. However, when n becomes large enough, as in the binomial distribution ($n = 100$ and $p = .20$) presented in Figure 5.26, the graph is relatively symmetric around the mean ($\mu = n \cdot p = 20$) because enough possible outcome values to the left of $x = 20$ allow the curve to fall back to the x-axis.

Figure 5.24: The Binomial Distribution for $n = 10$ and $p = .50$

Figure 5.25: The Binomial Distribution for $n = 10$ and $p = .20$

Figure 5.26: The Binomial Distribution for $n = 100$ and $p = .20$

For large n values, the binomial distribution is cumbersome to analyze without a computer. Table A.2 goes only to $n = 25$. The normal distribution is a good approximation for binomial distribution problems for large values of n.

To work a binomial problem by the normal curve requires a translation process. The first part of this process is to convert the two parameters of a binomial distribution, n and p, to the two parameters of the normal distribution, μ and σ. This process utilizes formulas from Chapter 5:

$$\mu = n \cdot p \text{ and } \sigma = \sqrt{n \cdot p \cdot q}$$

After completion of this, a test must be made to determine whether the normal distribution is a good enough approximation of the binomial distribution:

Does the interval $\mu \pm 3\sigma$ lie between 0 and n?

Recall that the empirical rule states that approximately 99.7%, or almost all, of the values of a normal curve are within three standard deviations of the mean. For a normal curve approximation of a binomial distribution problem to be acceptable, all possible x values should be between 0 and n, which are the lower and upper limits, respectively, of a binomial distribution. If $\mu \pm 3\sigma$ is not between 0 and n, do *not* use the normal distribution to work a binomial problem because the approximation is not good enough. Upon demonstration that the normal curve is a good approximation for a binomial problem, the procedure continues. Another rule of thumb for determining when to use the normal curve to approximate a binomial problem is that the approximation is good enough if both $n \cdot p > 5$ and $n \cdot q > 5$.

The process can be illustrated in the solution of the binomial distribution problem.

$$P(x \geq 25 | n = 60 \text{ and } p = .30) = ?$$

Note that this binomial problem contains a relatively large sample size and that none of the binomial tables in Appendix A.2 can be used to solve the problem. This problem is a good candidate for use of the normal distribution.

Translating from a binomial problem to a normal curve problem gives

$$\mu = n \cdot p = (60)(.30) = 18 \text{ and } \sigma = \sqrt{n \cdot p \cdot q} = 3.55$$

The binomial problem becomes a normal curve problem.

$$P(x \geq 25 | \mu = 18 \text{ and } \sigma = 3.55) = ?$$

Next, the test is made to determine whether the normal curve sufficiently fits this binomial distribution to justify the use of the normal curve.

$$\mu \pm 3\sigma = 18 \pm 3(3.55) = 18 \pm 10.65$$

$$7.35 \le \mu \pm 3\sigma \le 28.65$$

This interval is between 0 and 60, so the approximation is sufficient to allow use of the normal curve. **Figure 5.27** is a graph of this binomial distribution. Notice how closely it resembles the normal curve. **Figure 5.28** is the apparent graph of the normal curve version of this problem.

Figure 5.27: Graph of the Binomial Problem: n = 60 and p = .30

Figure 5.28: Graph of Apparent Solution of Binomial Problem Worked by the Normal Curve

5.9.1 CORRECTING FOR CONTINUITY

The translation of a discrete distribution to a continuous distribution is not completely straightforward. A correction of +.50 or –.50 or ±.50, depending on the problem, is required. This correction ensures that most of the binomial problem's information is correctly transferred to the normal curve analysis. This correction is called the **correction for continuity**, which is *made during conversion of a discrete distribution into a continuous distribution.*

Figure 5.29 is a portion of the graph of the binomial distribution, n = 60 and p = .30. Note that with a binomial distribution, all the probabilities are concentrated on the whole numbers. Thus, the answers for $x \ge 25$ are found by summing the probabilities for x = 25, 26, 27, . . . , 60. There are no values between 24 and 25, 25 and 26, . . . , 59, and 60. Yet, the normal distribution is continuous, and values are present all along the x-axis. A correction must be made for this discrepancy for the approximation to be as accurate as possible.

As an analogy, visualize the process of melting iron rods in a furnace. The iron rods are like the probability values on each whole number of a binomial distribution. Note that the binomial graph in Figure 5.29 looks like a series of iron rods in a line. When the rods are placed in a furnace, they melt down and spread out. Each rod melts and moves to fill the area between it and the adjacent rods. The result is a continuous sheet of solid iron (continuous iron) that looks like the normal curve. The melting of the rods is analogous to spreading the binomial distribution to approximate the normal distribution.

Figure 5.29: Graph of a Portion of the Binomial Problem: n = 60 and p = .30

How far does each rod spread toward the others? A good estimate is that each rod goes about halfway toward the adjacent rods. In other words, a rod that was concentrated at x = 25 spreads to cover the area from 24.5 to 25.5; x = 26 becomes continuous from 25.5 to 26.5; and so on. For the problem $P(x \ge 25 | n = 60 \text{ and } p = .30)$, conversion to a continuous normal curve problem yields $P(x \ge 24.5 | \mu = 18 \text{ and } \sigma = 3.55)$. The correction for continuity was –.50 because the problem called for the inclusion of the value of 25 along with all greater values; the binomial value of x = 25 translates to the normal curve value of 24.5 to 25.5. Had the binomial problem been to analyze $P(x > 25)$, the correction would have been +.50, resulting in a normal curve problem of $P(x \ge 25.5)$. The latter case would begin at more than 25 because the value of 25 would not be included.

The decision as to how to correct for continuity depends on the equality sign and the direction of the desired outcomes of the binomial distribution. **Table 5.9** lists some rules of thumb that can help in the application of the correction for continuity.

TABLE 5.9: RULES OF THUMB FOR THE CORRECTION FOR CONTINUITY

Values Being Determined	Corrections
$x >$	+.50
$x \ge$	–.50
$x <$	–.50
$x \le$	+.50
$\le x \le$	–.50 and +.50
$< x <$	+.50 and –.50
$x =$	–.50 and +.50

For the binomial problem $P(x \geq 25 | n = 60 \text{ and } p = .30)$, the normal curve becomes $P(x \geq 24.5 | \mu = 18 \text{ and } \sigma = 3.55)$, as shown in **Figure 5.30**, and

$$z = \frac{x - \mu}{\sigma} = \frac{24.5 - 18}{3.55} = 1.83$$

Figure 5.30: Graph of the Solution to the Binomial Problem Worked by the Normal Curve

The probability (Table 5.9) of this z value is .4664. The answer to this problem lies in the tail of the distribution, so the final answer is obtained by subtracting.

$$\begin{array}{r} .5000 \\ -.4664 \\ \hline .0336 \end{array}$$

Had this problem been worked by using the binomial formula, the solution would have been as shown in **Table 5.10**. The difference between the normal distribution approximation and the actual binomial values is only .0025 (.0361 − .0336).

TABLE 5.10: PROBABILITY VALUES FOR THE BINOMIAL PROBLEM: $n = 60$, $p = 30$, and $x \geq 25$

x Value	Probability
25	.0167
26	.0096
27	.0052
28	.0026
29	.0012
30	.0005
31	.0002
32	.0001
33	.0000
$x \geq 25$	.0361

DEMONSTRATION PROBLEM 5.12

Work the following binomial distribution problem by using the normal distribution.

$$P(x = 12 | n = 25 \text{ and } p = .40) = ?$$

Solution: Find μ and σ.

$$\mu = n \cdot p = (25)(.40) = 10.0$$

$$\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{(25)(.40)(.60)} = 2.45$$

$$\text{test}: \mu \pm 3\sigma = 10.0 \pm 3(2.45) = 2.65 \text{ to } 17.35$$

This range is between 0 and 25, so the approximation is close enough. Correct for continuity next. Because the problem is to determine the probability of x being exactly 12, the correction entails both −.50 and +.50. That is, a binomial probability at $x = 12$ translates to a continuous normal curve area that lies between 11.5 and 12.5. The graph of the problem follows:

Then,

$$z = \frac{x - \mu}{\sigma} = \frac{12.5 - 10}{2.45} = 1.02$$

and

$$z = \frac{x - \mu}{\sigma} = \frac{11.5 - 10}{2.45} = 0.61$$

$z = 1.02$ produces a probability of .3461.

$z = 0.61$ produces a probability of .2291.

The difference in areas yields the following answer:

$$.3461 - .2291 = .1170$$

Had this problem been worked by using the binomial tables, the resulting answer would have been .114. The difference between the normal curve approximation and the value obtained by using binomial tables is only .003.

DEMONSTRATION PROBLEM 5.13

Solve the following binomial distribution problem by using the normal distribution.

$$P(x < 27 | n = 100 \text{ and } p = .37) = ?$$

Solution: Because neither the sample size nor the p value is contained in Table A.2, working this problem by using binomial distribution techniques is

impractical. It is a good candidate for the normal curve. Calculating μ and σ yields

$$\mu = n \cdot p = (100)(.37) = 37.0$$

$$\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{(100)(.37)(.63)} = 4.83$$

Testing to determine the closeness of the approximation gives

$$\mu \pm 3\sigma = 37 \pm 3(4.83) = 37 \pm 14.49$$

The range 22.51 to 51.49 is between 0 and 100. This problem satisfies the conditions of the test. Next, correct for continuity: $x < 27$ as a binomial problem translates to $x \leq 26.5$ as a normal distribution problem. The graph of the problem follows.

Then,

$$z = \frac{x - \mu}{\sigma} = \frac{26.5 - 37}{4.83} = -2.17$$

Table 5.9 shows a probability of .4850. Solving for the tail of the distribution gives

$$.5000 - .4850 = .0150$$

which is the answer.

Had this problem been solved by using the binomial formula, the probabilities would have been the following.

x Value	Probability
26	.0059
25	.0035
24	.0019
23	.0010
22	.0005
21	.0002
20	.0001
$x < 27$	.0131

The answer obtained by using the normal curve approximation (.0150) compares favorably to this exact binomial answer. The difference is only .0019.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

35. One rule of thumb for determining when to use the normal curve to approximate a binomial problem is that the approximation is good enough if both $n \cdot p$ and $n \cdot q$ should be more than_____.
36. The correction for continuity is made during conversion of a discrete distribution into a_____.
37. For large n values, the _____ is cumbersome to analyze.

State whether the following statements are true/false:

38. For a normal curve approximation of a binomial distribution, all possible x values should be between 0 and n, which are the lower and upper limits, respectively.
39. As sample sizes become large, binomial distributions approach the normal distribution in shape but the value of p must be 0.5.

ACTIVITY

From 2011 Census report, draw the histogram and normal curve of your native city with the help of software. Is it normally distributed? Also check if it is following the $\mu \pm 3\sigma$ rule.

5.10 EXPONENTIAL DISTRIBUTION

Another useful continuous distribution is the exponential distribution. It is closely related to the Poisson distribution. Whereas the Poisson distribution is discrete and describes random occurrences over some interval, the **exponential distribution** is *continuous and describes a probability distribution of the times between random occurrences.* The following are the characteristics of the exponential distribution.

- It is a continuous distribution.
- It is a family of distributions.
- It is skewed to the right.
- The x values range from zero to infinity.
- Its apex is always at $x = 0$.
- The curve steadily decreases as x gets larger.

The exponential probability distribution is determined by the following.

Exponential Probability Density Function

$$f(x) = \lambda e^{-\lambda x}$$

where

$$x \geq 0$$

$$\lambda > 0$$

and

$$e = 2.71828\ldots$$

An exponential distribution can be characterized by the one parameter, λ. Each unique value of λ determines a different exponential distribution, resulting in a family of exponential distributions. **Figure 5.31** shows graphs of exponential distributions for four values of λ. The points on the graph are determined by using λ and various values of x in the probability density formula. The mean of an exponential distribution is $\mu = 1/\lambda$, and the standard deviation of an exponential distribution is $\sigma = 1/\lambda$.

Figure 5.31: Graphs of Some Exponential Distributions

5.10.1 PROBABILITIES OF THE EXPONENTIAL DISTRIBUTION

Probabilities are computed for the exponential distribution by determining the area under the curve between two points. Applying calculus to the exponential probability density function produces a formula that can be used to calculate the probabilities of an exponential distribution.

Probabilities of the Right Tail of the Exponential Distribution

$$P(x \geq x_0) = e^{-\lambda x_0}$$

where

$$x_0 \geq 0$$

To use this formula requires finding values of e^{-x}. These values can be computed on most calculators or obtained from Table A.4, which contains the values of e^{-x} for selected values of x. x_0 is the fraction of the interval or the number of intervals between arrivals in the probability question and λ is the average arrival rate.

As an example, arrivals at a bank are Poisson distributed with a λ of 1.2 customers every minute. What is the average time between arrivals and what is the probability that at least 2 minutes will elapse between one arrival and the next arrival? Since the interval for lambda is 1 minute and we want to know the probability that at least 2 minutes transpire between arrivals (twice the lambda interval), x_0 is 2.

Interarrival times of random arrivals are exponentially distributed. The mean of this exponential distribution is $\mu = 1/\lambda = 1/1.2 = .833$ minute (50 seconds). On average, .833 minute, or 50 seconds, will elapse between arrivals at the bank. The probability of an interval of 2 minutes or more between arrivals can be calculated by

$$P(x \geq 2 \mid \lambda = 1.2) = e^{-1.2(2)} = .0907.$$

About 9.07% of the time when the rate of random arrivals is 1.2 per minute, 2 minutes or more will elapse between arrivals, as shown in **Figure 5.32.**

This problem underscores the potential of using the exponential distribution in conjunction with the Poisson distribution to solve problems. In operations research and management science, these two distributions are used together to solve queuing problems (theory of waiting lines). The Poisson distribution can be used to analyze the arrivals to the queue, and the exponential distribution can be used to analyze the interarrival time.

Figure 5.32: Exponential Distribution for $\lambda = 1.2$ and Solution for $x \geq 2$

DEMONSTRATION PROBLEM 5.14

A manufacturing firm has been involved in statistical quality control for several years. As part of the production process, parts are randomly selected and tested. From the records of these tests, it has been established that a defective part occurs in a pattern that is Poisson distributed on the average of 1.38 defects every 20 minutes during production runs. Use this information to determine the probability that less than 15 minutes will elapse between any two defects.

Solution: The value of λ is 1.38 defects per 20-minute interval. The value of μ can be determined by

$$\mu = \frac{1}{\lambda} = \frac{1}{1.38} = .7246$$

On the average, it is .7246 of the interval, or (.7246)(20 minutes) = 14.49 minutes, between defects. The value of x_0 represents the desired number of intervals between arrivals or occurrences for the probability question. In this problem, the probability question involves 15 minutes and the interval is 20 minutes. Thus x_0 is 15/20, or .75 of an interval. The question here is to determine the probability of there being less than 15 minutes between defects. The probability formula always yields the right tail of the distribution—in this case, the probability of there being 15 minutes or more between arrivals. By using

the value of x_0 and the value of λ, the probability of there being 15 minutes or more between defects can be determined.

$$P(x \geq x_0) = P(x \geq .75) = e^{-\lambda x_0} = e^{(-1.38)(.75)} = e^{-1.035} = .3552$$

The probability of .3552 is the probability that at least 15 minutes will elapse between defects. To determine the probability of there being less than 15 minutes between defects, compute $1 - P(x)$. In this case, $1 - .3552 = .6448$. There is a probability of .6448 that less than 15 minutes will elapse between two defects when there is an average of 1.38 defects per 20-minute interval or an average of 14.49 minutes between defects.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

40. Exponential distribution is skewed to the _____.
41. Exponential distribution is closely related to Poisson distribution; only difference is that exponential distribution is _____ and describes a probability distribution of the times between random occurrences.
42. The x values for exponential distribution range from _____ to infinity.

State whether the following statements are true/false:

43. In exponential distribution, the lambda (λ) is equal to μ.
44. The mean and the standard deviation of exponential distribution are same.

ACTIVITY

Suppose the values of lambda are (1.12, 0.8, and 0.5) and values of x are (7, 6, 5, 4, 3, 2, and 1). Use a statistical software tool (e.g., Excel) for calculating the exponential probabilities. Now draw these values and probabilities in the graph (do not use software for graphical presentation).

5.11 SUMMARY

- Probability experiments produce random outcomes. A variable that contains the outcomes of a random experiment is called a random variable. Random variables such that the set of all possible values is at most a finite or countably infinite number of possible values are called discrete random variables. Random variables that take on values at all points over a given interval are called continuous random variables. Discrete distributions are constructed from discrete random variables. Continuous distributions are constructed from continuous random variables. Three discrete distributions are the binomial distribution, Poisson distribution, and hypergeometric distribution.

- The binomial distribution fits experiments when only two mutually exclusive outcomes are possible. In theory, each trial in a binomial experiment must be independent of the other trials. However, if the population size is large enough in relation to the sample size ($n < 5\% N$), the binomial distribution can be used where applicable in cases where the trials are not independent. The probability of getting a desired outcome on any one trial is denoted as p, which is denoted as the probability of getting a success. The binomial formula is used to determine the probability of obtaining x outcomes in n trials. Binomial distribution problems can be solved more rapidly with the use of binomial tables than by formula. Table A.2 of Appendix A contains binomial tables for selected values of n and p.
- The Poisson distribution usually is used to analyze phenomena that produce rare occurrences over some interval. The only information required to generate a Poisson distribution is the long-run average, which is denoted by lambda (λ). The assumptions are that each occurrence is independent of other occurrences and that the value of lambda remains constant throughout the experiment. Poisson probabilities can be determined by either the Poisson formula or the Poisson tables in Table A.3 of Appendix A. The Poisson distribution can be used to approximate binomial distribution problems when n is large ($n > 20$), p is small, and $n \cdot p \leq 7$.
- The hypergeometric distribution is a discrete distribution that is usually used for binomial-type experiments when the population is small and finite and sampling is done without replacement. Because using the hypergeometric distribution is a tedious process, using the binomial distribution whenever possible is generally more advantageous.
- This chapter discussed three different continuous distributions: the uniform distribution, the normal distribution, and the exponential distribution. With continuous distributions, the value of the probability density function does not yield the probability but instead gives the height of the curve at any given point. In fact, with continuous distributions, the probability at any discrete point is .0000. Probabilities are determined over an interval. In each case, the probability is the area under the curve for the interval being considered. In each distribution, the probability or total area under the curve is 1.
- Probably the simplest of these distributions is the uniform distribution, sometimes referred to as the rectangular distribution. The uniform distribution is determined from a probability density function that contains equal values along some interval between the points a and b. Basically, the height of the curve is the same everywhere between these two points. Probabilities are determined by calculating the portion of the rectangle between the two points a and b that is being considered.
- The most widely used of all distributions is the normal distribution. Many phenomena are normally distributed, including characteristics of most machine-produced parts, many measurements of the biological and natural environment, and many human characteristics such as height, weight, IQ, and achievement test scores. The normal curve is continuous,

symmetrical, unimodal, and asymptotic to the axis; actually, it is a family of curves.

- The parameters necessary to describe a normal distribution are the mean and the standard deviation. For convenience, data that are being analyzed by the normal curve should be standardized by using the mean and the standard deviation to compute z scores. A z score is the distance that an x value is from the mean, μ, in units of standard deviations. With the z score of an x value, the probability of that value occurring by chance from a given normal distribution can be determined by using a table of z scores and their associated probabilities.
- The normal distribution can be used to work certain types of binomial distribution problems. Doing so requires converting the n and p values of the binomial distribution to μ and σ of the normal distribution. When worked by using the normal distribution, the binomial distribution solution is only an approximation. If the values of $\mu \pm 3\sigma$ are within a range from 0 to n, the approximation is reasonably accurate. Adjusting for the fact that a discrete distribution problem is being worked by using a continuous distribution requires a correction for continuity. The correction for continuity involves adding or subtracting .50 to the x value being analyzed. This correction usually improves the normal curve approximation.
- Another continuous distribution is the exponential distribution. It complements the discrete Poisson distribution. The exponential distribution is used to compute the probabilities of times between random occurrences. The exponential distribution is a family of distributions described by one parameter, λ. The distribution is skewed to the right and always has its highest value at $x = 0$.

KEY WORDS

1. **Binomial distribution:** As the word binomial indicates, any single trial of a binomial experiment contains only two possible outcomes. These two outcomes are labeled success or failure.
2. **Hypergeometric distribution:** The hypergeometric distribution is a discrete probability distribution that applies only to experiments in which the trials are done without replacement.
3. **Continuous distributions:** The Continuous Distributions take on values at every point over a given interval. Thus continuous random variables have no gaps or unassumed values.
4. **Continuous random variables:** Continuous random variables are generated from experiments in which things are "measured" not "counted."
5. **Mean, or expected value:** The mean or expected value of a discrete distribution is the long-run average of occurrences.
6. **Discrete distributions:** Constructed from discrete random variables.
7. **Poisson distribution:** Focuses on the number of discrete occurrences over some interval or continuum. The average of Poisson distribution is denoted by lamda (λ).

KEY WORDS

8. **Discrete random variables:** A random variable is a **discrete random variable** if the set of all possible values is at most a finite or a countably infinite number of possible values.
9. **Random variable:** A variable that contains the outcomes of a random experiment is called a random variable.
10. **Correction for continuity:** This correction ensures that most of the binomial problem's information is correctly transferred to the normal curve analysis.
11. **Exponential distribution:** The exponential distribution is continuous and describes a probability distribution of the times between random occurrences.
12. **Normal distribution:** Normal distribution is a function that represents the distribution of many random variables as a symmetrical bell-shaped graph curve.
13. **Rectangular distribution:** Rectangular distribution is also known as uniform distribution.
14. **Standardized normal distribution:** A standard normal distribution is a normal distribution with mean 0 and standard deviation 1. Areas under this curve can be found using a standard normal table.
15. **Uniform distribution:** It is sometimes referred to as the rectangular distribution. The uniform distribution is determined from a probability density function that contains equal values along some interval between the points a and b.
16. **z distribution:** The z distribution is a normal distribution with a mean of 0 and a standard deviation of 1.
17. **z score:** A z score is the number of standard deviations that a value, x, is above or below the mean. If the value of x is less than the mean, the z score is negative; if the value of x is more than the mean, the z score is positive.

5.12 DESCRIPTIVE QUESTIONS

5.1. Solve for the probabilities of the following binomial distribution problems by using the binomial formula.

(a) If $n = 11$ and $p = .23$, what is the probability that $x = 4$?

(b) If $n = 6$ and $p = .50$, what is the probability that $x \geq 1$?

(c) If $n = 9$ and $p = .85$, what is the probability that $x > 7$?

(d) If $n = 14$ and $p = .70$, what is the probability that $x \leq 3$?

5.2. In a study by Peter D. Hart Research Associates for the Nasdaq Stock Market, it was determined that 20% of all stock investors are retired peopl(e) In addition, 40% of all U.S. adults invest in mutual funds. Suppose a random sample of 25 stock investors is taken.

(a) What is the probability that exactly seven are retired people?

(b) What is the probability that 10 or more are retired people?

(c) How many retired people would you expect to find in a random sample of 25 stock investors?

(d) Suppose a random sample of 20 U.S. adults is taken. What is the probability that exactly eight adults invested in mutual funds?

(e) Suppose a random sample of 20 U.S. adults is taken. What is the probability that fewer than six adults invested in mutual funds?

(f) Suppose a random sample of 20 U.S. adults is taken. What is the probability that none of the adults invested in mutual funds?

(g) Suppose a random sample of 20 U.S. adults is taken. What is the probability that 12 or more adults invested in mutual funds?

(h) For parts e–g, what exact number of adults would produce the highest probability? How does this compare to the expected number?

5.3. Suppose that, for every lot of 100 computer chips a company produces, an average of 1.4 are defective. Another company buys many lots of these chips at a time, from which one lot is selected randomly and tested for defects. If the tested lot contains more than three defects, the buyer will reject all the lots sent in that batch. What is the probability that the buyer will accept the lots? Assume that the defects per lot are Poisson distributed.

5.4. The National Center for Health Statistics reports that 25% of all Americans between the ages of 65 and 74 have a chronic heart condition. Suppose you live in a state where the environment is conducive to good health and low stress and you believe the conditions in your state promote healthy hearts. To investigate this theory, you conduct a random telephone survey of 20 persons 65 to 74 years of age in your state.

(a) On the basis of the figure from the National Center for Health Statistics, what is the expected number of persons 65 to 74 years of age in your survey who have a chronic heart condition?

(b) Suppose only one person in your survey has a chronic heart condition. What is the probability of getting one or fewer people with a chronic heart condition in a sample of 20 if 25% of the population in this age bracket has this health problem? What do you conclude about your state from the sample data?

5.5. According to a recent survey, the probability that a passenger files a complaint with the Department of Transportation about a particular U.S. airline is .000014. Suppose 100,000 passengers who flew with this particular airline are randomly contacted.

(a) What is the probability that exactly five passengers filed complaints?

(b) What is the probability that none of the passengers filed complaints?

(c) What is the probability that more than six passengers filed omplaints?

5.6. A hair stylist has been in business one year. Sixty percent of his customers are walk-in business. If he randomly samples eight of the people from last week's list of customers, what is the probability that three or fewer were walk-ins? If this outcome actually occurred, what would be some of the explanations for it?

5.7. A survey conducted by the Consumer Reports National Research Center reported, among other things, that women spend an average of 1.2 hours per week shopping online. Assume that hours per week shopping online are Poisson distributed. If this survey result is true for all women and if a woman is randomly selected.

(a) What is the probability that she did not shop at all online over a one-week period?

(b) What is the probability that a woman would shop three or more hours online during a one-week period?

(c) What is the probability that a woman would shop fewer than five hours in a three-week period?

5.8. An office in Albuquerque has 24 workers including management. Eight of the workers commute to work from the west side of the Rio Grande River. Suppose six of the office workers are randomly selected.

(a) What is the probability that all six workers commute from the west side of the Rio Grande?

(b) What is the probability that none of the workers commute from the west side of the Rio Grande?

(c) Which probability from parts (a) and (b) was greatest? Why do you think this is?

(d) What is the probability that half of the workers do not commute from the west side of the Rio Grande?

5.9. Suppose that in the bookkeeping operation of a large corporation the probability of a recording error on any one billing is .005. Suppose the probability of a recording error from one billing to the next is constant, and 1,000 billings are randomly sampled by an auditor.

(a) What is the probability that fewer than four billings contain a recording error?

(b) What is the probability that more than 10 billings contain a billing error?

(c) What is the probability that all 1,000 billings contain no recording errors?

5.10. The U.S. Bureau of Labor Statistics reports that of persons who usually work full-time, the average number of hours worked per week is 43.4. Assume that the number of hours worked per week for those who usually work full-time is normally distributed. Suppose 12% of these workers work more than 48 hours. Based on this percentage, what is the standard deviation of number of hours worked per week for these workers?

5.11. According to an NRF survey conducted by BIGresearch, the average family spends about $237 on electronics (computers, cell phones, etc.) in back-to-college spending per student. Suppose back-to-college family spending on electronics is normally distributed with a standard deviation of $54. If a family of a returning college student is randomly selected, what is the probability that:

(a) They spend less than $150 on back-to-college electronics?

(b) They spend more than $400 on back-to-college electronics?

(c) They spend between $120 and $185 on back-to-college electronics?

5.12. A business convention holds its registration on Wednesday morning from 9:00 A.M. until 12:00 noon. Past history has shown that registrant arrivals follow a Poisson distribution at an average rate of 1.8 every 15 seconds. Fortunately, several facilities are available to register convention members.

(a) What is the average number of seconds between arrivals to the registration area for this conference based on past results?

(b) What is the probability that 25 seconds or more would pass between registration arrivals?

(c) What is the probability that less than five seconds will elapse between arrivals?

(d) Suppose the registration computers went down for a one-minute period. Would this condition pose a problem? What is the probability that at least one minute will elapse between arrivals?

5.13. According to Bureau of Labor Statistics data, the average income for a worker in the metropolitan Boston area is $61,162. Suppose average income for a worker in the metropolitan Boston area is normally distributed with a standard deviation of $4,246. A metropolitan Boston area worker is randomly selected.

(a) What is the probability that the worker's income is more than $70,000?

(b) What is the probability that the worker's income is less than $55,000?

(c) What is the probability that the worker's income is more than $50,000?

(d) What is the probability that the worker's income is between $57,000 and $65,000?

5.14. The Federal Reserve System publishes data on family income based on its Survey of Consumer Finances. When the head of the household has a college degree, the mean before-tax family income is $85,200. Suppose that 60% of the before-tax family incomes when the head of the household has a college degree are between $75,600 and $94,800 and that these incomes are normally distributed. What is the standard deviation of before-tax family incomes when the head of the household has a college degree?

5.15. Coastal businesses along the Gulf of Mexico from Texas to Florida worry about the threat of hurricanes during the season from June through October. Businesses become especially nervous when hurricanes enter the Gulf of Mexico. Suppose the arrival of hurricanes during this season is Poisson distributed, with an average of three hurricanes entering the Gulf of Mexico during the five-month season. If a hurricane has just entered the Gulf of Mexico:

(a) What is the probability that at least one month will pass before the next hurricane enters the Gulf?

(b) What is the probability that another hurricane will enter the Gulf of Mexico in two weeks or less?

(c) What is the average amount of time between hurricanes entering the Gulf of Mexico?

5.16. According to the Air Transport Association of America, the average operating cost of an MD-80 jet airliner is $2,087 per hour. Suppose the operating costs of an MD-80 jet airliner are normally distributed with a standard deviation of $175 per hour. At what operating cost would only 20% of the operating costs be less? At what operating cost would 65% of the operating costs be more? What operating cost would be more than 85% of operating costs?

5.17. Suppose the data below represent the number of sales associates who are working in a department store in any given retail day. Describe the distribution including the mean and standard deviation. Interpret the shape of the distribution and the mean in light of the data being studied. What do the probability statements mean?

CUMULATIVE DISTRIBUTION FUNCTION

Continuous uniform on 11 to 32

x	$P(X <= x)$
28	0.80952
34	1.00000
16	0.23810
21	0.47619

5.18. A manufacturing company produces a metal rod. Use the Excel output shown here to describe the weight of the rod. Interpret the probability values in terms of the manufacturing process.

NORMAL DISTRIBUTION

Mean = 227 mg.
Standard Deviation = 2.3 mg.

x Value	Probability < x Value
220	0.0012
225	0.1923
227	0.5000
231	0.9590
238	1.0000

5.19. A restaurant averages 4.51 customers per 10 minutes during the summer in the late afternoon. Shown here are the statistical software output for this restaurant. Discuss the type of distribution used to analyze the data and the meaning of the probabilities.

EXPONENTIAL DISTRIBUTION

x Value	Probability < x Value
0.1	0.3630
0.2	0.5942
0.5	0.8951
1.0	0.9890
2.4	1.0000

CUMULATIVE DISTRIBUTION FUNCTION

Exponential with mean = 0.221729

x	$P(X <= x)$
0.1	0.363010
0.2	0.594243
0.5	0.895127
1.0	0.989002
2.4	0.999980

5.13 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

5.1.

(a) $P(x = 4 \mid n = 11 \text{ and } p = .23)$

$_{11}C_4(.23)^4(.77)^7 = 330(.0028)(.1605) = .1482$

(b) $P(x \geq 1 \mid n = 6 \text{ and } p = .50)$

$= 1 - P(x < 1) = 1 - P(x = 0)$

$= 1 - [_6C_0(.50)^0(.50)^6] = 1 - [(1)(1)(.0156)] = .9844$

(c) $P(x > 7 \mid n = 9 \text{ and } p = .85) = P(x = 8) + P(x = 9)$

$= {}_9C_8(.85)^8(.15)^1 + {}_9C_9(.85)^9(.15)^0$

$= (9)(.2725)(.15) + (1)(.2316)(1)$

$= .3679 + .2316 = .5995$

(d) $P(x \leq 3 \mid n = 14 \text{ and } p = .70)$

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

$= {}_{14}C_3(.70)^3(.30)^{11} + {}_{14}C_2(.70)^2(.30)^{12} + {}_{14}C_1(.70)^1(.30)^{13} + {}_{14}C_0(.70)^0(.30)^{14}$

$= (364)(.3430)(.00000177) + (91)(.49)(.000000531)$

$= (14)(.70)(.00000016) + (1)(1)(.000000048)$

$= .0002 + .0000 + .0000 + .0000 = .0002$

5.2. $n = 25\ p = .20$ retired

(a) from Table A.2: $P(x = 7) = .1108$

(b) $P(x \geq 10)$: $P(x = 10) + P(x = 11) + \cdots + P(x = 25) = .012 + .004 + .001 = .017$

(c) Expected Value $= \mu = n \cdot p = 25(.20) = 5$

(d) $n = 20\ p = .40$ mutual funds

$P(x = 8) = .1797$

(e) $P(x < 6) = P(x = 0) + P(x = 1) + \cdots + P(x = 5)$

$= .000 + .000 + .003 + .012 + .035 + .075 = .125$

(f) $P(x = 0) = .000$

(g) $P(x \geq 12) = P(x = 12) + P(x = 13) + \cdots + P(x = 20)$

$= .035 + .015 + .005 + .001 = .056$

(h) $x = 8$

Expected Number $= \mu = n \cdot p = 20(.40) = 8$

5.3. $\lambda = 1.4$ defects $\mid \lambda$ lot If $x > 3$, buyer rejects If $x \leq 3$, buyer accepts

$P(x \leq 3 \mid \lambda = 1.4) =$ (from Table A.3)

x	Prob.
0	.2466
1	.3452
2	.2417
3	.1128
	.9463

5.4. **(a)** $n = 20$ and $p = .25$

The expected number $= \mu = n \cdot p = (20)(.25) = 5.00$

(b) $P(x \leq 1 \mid n = 20 \text{ and } p = .25)$

$= P(x = 1) + P(x = 0) = {}_{20}C_1(.25)^1(.75)^{19} + {}_{20}C_0(.25)^0(.75)^{20}$

$= (20)(.25)(.00423) + (1)(1)(.0032) = .0212 + .0032 = .0244$

Since the probability is so low, the population of your state may have a lower percentage of chronic heart conditions than those of other states.

5.5. $n = 100{,}000\ p = .000014$

Worked as a Poisson: $\lambda = n \cdot p = 100{,}000(.000014) = 1.4$

(a) $P(x = 5)$:

from Table A.3 = .0111

(b) $P(x = 0)$:

from Table A.3 = .2466

(c) $P(x > 6)$: (from Table A.3)

x	Prob
7	.0005
8	.0001
	.0006

5.6. $P(x \leq 3) \mid n = 8$ and $p = .60$): From Table A.2:

x	Prob.
0	.001
1	.008
2	.041
3	.124
	.174

17.4% of the time in a sample of eight, three or fewer customers are walk-ins by chance. Other reasons for such a low number of walk-ins might be that she is retaining more old customers than before or perhaps a new competitor is attracting walk-ins away from her.

5.7. $\lambda = 1.2$ hours | week

(a) $P(x = 0 \mid \lambda = 1.2)$ = (from Table A.3) **.3012**

(b) $P(x \geq 3 \mid \lambda = 1.2)$ = (from Table A.3)

x	Prob.
3	.0867
4	.0260
5	.0062
6	.0012
7	.0002
8	.0000
	.1203

(c) $P(x < 5 \mid 3$ weeks)

If $\lambda = 1.2$ for 1 week, the $\lambda = 3.6$ for 3 weeks.

$P(x < 5 \mid \lambda = 3.6)$ = (from Table A.3)

x	Prob.
4	.1912
3	.2125
2	.1771
1	.0984
0	.0273
	.7065

5.8. $N = 24 \quad n = 6 \quad A = 8$

(a) $P(x = 6) = \dfrac{{}_8C_6 \cdot {}_{16}C_0}{{}_{24}C_6} = \dfrac{(28)(1)}{134{,}596} = \mathbf{.0002}$

(b) $P(x = 0) = \dfrac{{}_8C_0 \cdot {}_{16}C_6}{{}_{24}C_6} = \dfrac{(1)(8008)}{134{,}596} = \mathbf{.0595}$

(c) $A = 16$ East Side

$P(x = 3) = \dfrac{{}_{16}C_3 \cdot {}_8C_3}{{}_{24}C_6} = \dfrac{(560)(56)}{134{,}596} = \mathbf{.2330}$

5.9. $p = .005 \quad n = 1{,}000$

$\lambda = n \cdot p = (1{,}000)(.005) = 5$

(a) $P(x < 4) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$

$= .0067 + .0337 + .0842 + .1404 = \mathbf{.265}$

(b) $P(x > 10) = P(x = 11) + P(x = 12) + \cdots$

$= .0082 + .0034 + .0013 + .0005 + .0002 = \mathbf{.0136}$

(c) $P(x = 0) = \mathbf{.0067}$

5.10. $\mu = 43.4$

12% more than 48. $x = 48$

Area between x and μ is $.50 - .12 = .38$

z associated with an area of .3800 is $z = 1.175$

Solving for σ:

$$z = \frac{x - \mu}{\sigma}$$

$$1.175 = \frac{48 - 43.4}{\sigma}$$

$$\sigma = \frac{4.6}{1.175} = 3.915$$

5.11. $\mu = 237 \quad \sigma = 54$

(a) $P(x < 150)$:

$$z = \frac{150 - 237}{54} = -1.61$$

from Table A.5, area for $z = -1.61$ is .4463

$P(x < 80) = .5000 - .4463 = \mathbf{.0537}$

(b) $P(x > 400)$:

$$z = \frac{400 - 237}{54} = 3.02$$

from Table A.5, area for $z = 3.02$ is .4987

$P(x > 400) = .5000 - .4987 = \mathbf{.0013}$

(c) $P(120 < x < 185)$:

$$z = \frac{120 - 237}{54} = -2.17$$

$$z = \frac{185 - 237}{54} = -0.96$$

from Table A.5, area for $z = -2.17$ is .4850

area for $z = -0.96$ is .3315

$P(120 < x < 185) = .4850 - .3315 = \mathbf{.1535}$

5.12. $\lambda = 1.8$ per 15 seconds

(a) $\mu = \frac{1}{\lambda} = \frac{1}{1.8} = = .5556(15 \text{ sec}) = \mathbf{8.33 \text{ sec}}$

(b) For $x_0 > 25$ sec use $x_0 = 25/15 = 1.67$

$P(x_0 > 25 \text{ sec} = e^{-1.6667(1.8)} = \mathbf{.0498}$

(c) $x_0 < 5$ sec $= 1/3$

$P(x_0 < 5 \text{ sec} = 1 - e^{-1/3(1.8)} = 1 - .5488 = \mathbf{.4512}$

(d) $P(x_0 \geq 1 \text{ min.})$:

$x_0 = 1 \text{ min.} = 60/15 = 4$

$P(x_0 \geq 1 \text{ min.}) = e^{-4(1.8)} = \mathbf{.0007}$

5.13. $\mu = 61{,}162$ $\sigma = 4{,}246$

(a) $P(x > 70{,}000)$:

$$z = \frac{x - \mu}{\sigma} = \frac{70{,}000 - 61{,}162}{4{,}246} = \frac{8{,}838}{4{,}246} = 2.08$$

from Table A.5, the area for $z = 2.08$ is .4812

$P(x > 70{,}000) = .5000 - .4812 = \mathbf{.0188}$

(b) $P(x < 55{,}000)$:

$$z = \frac{x - \mu}{\sigma} = \frac{55{,}000 - 61{,}162}{4{,}246} = \frac{-6{,}162}{4{,}246} = -1.45$$

from Table A.5, the area for $z = -1.45$ is .4265

$P(x < 55{,}000) = .5000 - .4265 = \mathbf{.0735}$

(c) $P(x > 50{,}000)$:

$$z = \frac{x - \mu}{\sigma} = \frac{50{,}000 - 61{,}162}{4{,}246} = \frac{-11{,}162}{4{,}246} = -2.63$$

from Table A.5, the area for $z = -2.63$ is .4957

$P(x > 35{,}000) = .5000 + .4957 = \mathbf{.9957}$

(d) $P(57{,}000 < x < 65{,}000)$:

$$z = \frac{x - \mu}{\sigma} = \frac{57{,}000 - 61{,}162}{4{,}246} = \frac{-4{,}162}{4{,}246} = -0.98$$

$$z = \frac{x - \mu}{\sigma} = \frac{65{,}000 - 61{,}162}{4{,}246} = \frac{3{,}838}{4{,}246} = 0.90$$

from Table A.5, the area for $z = -0.98$ is .3365

the area for $z = 0.90$ is .3159

$P(57{,}000 < x < 65{,}000) = .3365 + .3159 = \mathbf{.6524}$

5.14. $\mu = 85{,}200$

60% are between 75,600 and 94,800

$94{,}800 - 85{,}200 = 9{,}600$

$75{,}600 - 85{,}200 = 9{,}600$

The 60% can be split into 30% and 30% because the two x values are equal distance from the mean.

The z value associated with .3000 area is 0.84

$$z = \frac{x - \mu}{\sigma}$$

$$.84 = \frac{94{,}800 - 85{,}200}{\sigma}$$

$$\sigma = \mathbf{11{,}428.57}$$

5.15. $\lambda = 3$ hurricanes | 5 months

(a) $P(x \geq 1 \text{ month} \mid \lambda = 3 \text{ hurricanes per 5 months})$:

$x = 1$ month is 1/5 of the 5 month interval associated with λ.

Thus, $x_0 = 1/5$

$P(x_0 \leq 1) = e^{-\lambda x_0} = e^{-3(1/5)} = e^{-0.6} = \mathbf{.5488}$

(b) $P(x \leq 2 \text{ weeks})$: 2 weeks = 0.5 month = 1/10 of the 5 month interval

$P(x \leq 2 \text{ weeks} \mid \lambda = 0.6 \text{ per month})$

$= 1 - P(x > 2 \text{ weeks} \mid \lambda = 0.6 \text{ per month})$

But $P(x > 2 \text{ weeks} \mid \lambda = 0.6 \text{ per month})$:

Let $x_0 = 0.1$

$P(x > 2 \text{ weeks}) = e^{-\lambda x_0} = e^{-3(1/10)} = e^{-0.3} = .7408$

$P(x \leq 2 \text{ weeks}) = 1 - P(x > 2 \text{ weeks}) = 1 - .7408 = \mathbf{.2592}$

(c) Average time = Expected time = $\mu = 1/\lambda = \mathbf{1.67 \text{ months}}$

5.16. $\mu = 2087$ $\sigma = 175$

If 20% are less, then 30% lie between x and μ.

$z_{.30} = -.84$

$$z = \frac{x-\mu}{\sigma}$$

$$-.84 = \frac{x-2087}{175}$$

$$x = 1940$$

If 65% are more, then 15% lie between x and μ

$z_{.15} = -0.39$

$$z = \frac{x-\mu}{\sigma}$$

$$-.39 = \frac{x-2087}{175}$$

$$x = 2018.75$$

If x is more than 85%, then 35% lie between x and μ.

$z_{.35} = 1.04$

$$z = \frac{x-\mu}{\sigma}$$

$$1.04 = \frac{x-2087}{175}$$

$$x = 2269$$

5.17. This is a uniform distribution with $a = 11$ and $b = 32$.

The mean is $(11 + 32)/2 = 21.5$ and the standard deviation is $(32 - 11)/\sqrt{12} = 6.06$. Almost 81% of the time there are less than or equal to 28 sales associates working. One hundred percent of the time there are less than or equal to 34 sales associates working and never more than 34. About 23.8% of the time there are 16 or fewer sales associates working. There are 21 or fewer sales associates working about 48% of the time.

5.18. The weight of the rods is normally distributed with a mean of 227 mg and a standard deviation of 2.3 mg. The probability that a rod weighs less than or equal to 220 mg is .0012, less than or equal to 225 mg is .1923, less than or equal to 227 is .5000 (since 227 is the mean), less than 231 mg is .9590, and less than or equal to 238 mg is 1.000.

5.19. The exponential distribution has $\lambda = 4.51$ per 10 minutes and $\mu = 1/4.51 = .22173$ of 10 minutes or 2.2173 minutes. The probability that there is less than .1 or 1 minute between arrivals is .3630. The probability that there is less than .2 or 2 minutes between arrivals is .5942. The probability that there is .5 or 5 minutes or more between arrivals is .1049. The probability that there is more than 1 or 10 minutes between arrivals is .0110. It is almost certain that there will be less than 2.4 or 24 minutes between arrivals.

5.14 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
5.2. Discrete Versus Continuous Distributions	1.	random variable
	2.	discrete
	3.	measured
	4.	False
	5.	False
5.3. Describing a Discrete Distribution	6.	histogram
	7.	expected or mean
	8.	discrete experiment's outcomes
	9.	True
	10.	True
5.4. Binomial Distribution	11.	zero
	12.	mean
	13.	independent
	14.	False
	15.	True
5.5. Poisson Distribution	16.	improbable
	17.	long-run average.
	18.	True
	19.	False
5.6. Hypergeometric Distribution	20.	without
	21.	finite
	22.	True
	23.	False
	24.	False
5.7. The Uniform Distribution	25.	continuous
	26.	rectangle
	27.	one
	28.	True
	29.	True
5.8. Normal Distribution	30.	z distribution
	31.	mirror
	32.	True
	33.	True
	34.	False
5.9. Using the Normal Curve to Approximate Binomial Distribution Problems	35.	five
	36.	continuous distribution
	37.	binomial distribution

E S

Topics	Q. No.	Answers
	38.	True
	39.	False
5.10. Exponential Distribution	40.	right
	41.	continuous
	42.	zero
	43.	False
	44.	True

6

CHAPTER

SAMPLING AND SAMPLING DISTRIBUTIONS

CONTENTS

6.1 Introduction
6.2 Sampling
6.2.1 Reasons for Sampling
6.2.2 Reasons for Taking a Census
6.2.3 Frame
6.2.4 Random Versus Nonrandom Sampling
6.2.5 Random Sampling Techniques
6.2.6 Systematic Sampling
6.2.7 Cluster (or Area) Sampling
6.2.8 Nonrandom Sampling
6.2.9 Sampling Error
6.2.10 Nonsampling Errors
Self Assessment Questions
Activity
6.3 Sampling Distribution of $\bar{x}$
6.3.1 Sampling from a Finite Population
Self Assessment Questions
Activity
6.4 Sampling Distribution of $\hat{p}$
Self Assessment Questions
Activity
6.5 Summary
6.6 Descriptive Questions
6.7 Solutions for Descriptive Questions
6.8 Answers and Hints