Ronald L. Moy
Li-Shya Chen
Lie Jane Kao

Study Guide for Statistics for Business and Financial Economics

A Supplement to the Textbook by Cheng-Few Lee, John C. Lee and Alice C. Lee

Springer

Study Guide for Statistics for Business and Financial Economics

Ronald L. Moy • Li-Shya Chen • Lie Jane Kao

Study Guide for Statistics for Business and Financial Economics

A Supplement to the Textbook by Cheng-Few Lee, John C. Lee and Alice C. Lee

Springer

Ronald L. Moy
Tobin College of Business
St. John's University
Staten Island
New York
USA

Li-Shya Chen
National Chengchi University
Taipei
Taiwan

Lie Jane Kao
Department Banking & Finance
KNU Research Center of Corp Finance
Taoyuan
Taiwan

Statistics for Business and Financial Economics, 978-1-4614-5896-8
Here's the book on Springer.com: http://www.springer.com/statistics/business%2C+economics+%26+finance/book/978-1-4614-5896-8

ISBN 978-3-319-11996-0 ISBN 978-3-319-11997-7 (eBook)
DOI 10.1007/978-3-319-11997-7

Library of Congress Control Number: 2014955348

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 1
Introduction

Chapter Intuition

Statistics is an approach that allows us to systematically organize, present, and analyze data. There are two basic uses of statistics: to compare or describe data and to make inferences about the population based on the data. When statistics is used to *describe* or to *summarize*, we call it *descriptive statistics*. Examples of *descriptive statistics* include average points per game for a basketball player, average yards rushing for a football player, grade point average (GPA), and the average salary a graduating senior receives. Descriptive statistics also enables us to make comparisons. For example, if you received a score of 80 on your statistics midterm, it would be difficult for you to know how well you did without knowing how other students in the class performed. By looking at the class average, you can compare your score to the performance of the rest of the class.

When we use statistics to make inference about a population, we are using an approach known as *inferential statistics*, because it allows us to infer facts about a population that is not yet known. For example, polls that predict the outcome of elections use inferential statistics. Inferential statistics is an approach in making educated guesses about what an entire population (for example, all voters) would do based on a smaller sample (for example, a selection of 1000 voters). Inferential statistics is especially important when it is costly or time consuming to survey the entire population.

Chapter Review

1. *Statistics* is a course of study devoted to the collection, organization, and analysis of data.
2. The entire group we are interested in studying is called the *population*. A *sample* is a subset of the population.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_1

3. *Descriptive statistics* allows us to summarize the data from a sample. For example, when we want to compare the heights of the players of the Boston Celtics to that of the Los Angeles Lakers, the sample means of the heights of the two teams are used. The sample mean can be used to measure performance such as the return of AT&T's stock, or the yards Andrian Peterson gains every time he carries the ball, or the SAT score for students entering Stanford University. Other examples of *Descriptive statistics* including the sample variance, the median, the 99th percentile, … , etc. The first part of the text is devoted to descriptive statistics.
4. *Inferential statistics* allows us to draw conclusions about an entire population using only a subset of the population, namely, a sample. For example, drawing a conclusion about the outcome of a presidential election by looking at a sample of voters. The benefit of using inferential statistics is that it enables us to draw conclusions about the population taking into consideration the uncertainty associated with the sample drawn. The second half of the book is devoted to inferential statistics.

Example Problems

Example 1 Descriptive Versus Inferential Statistics

Determine if the following statement is a descriptive or an inferential statistic:

a. The average earnings per share for AT&T over the last 5 years.
b. The number of people who will vote for the Democratic candidate for senator in the upcoming election in California using a sample of 200 potential voters.
c. The number of people who would favor a constitutional amendment requiring Congress to balance the budget, based on a survey of registered voters.
d. The average number of yards a rookie running back is expected to gain, based on a sample of rookie running backs.

Solution

a. Because averages are used to summarize or describe past data, they are descriptive statistics. So, the average EPS is a descriptive statistic.
b. When we use a subset of data to infer the behavior of an entire population, we are using inferential statistics.
c. Inferential statistic.
d. Inferential statistic.

Example 2 Using Descriptive Statistics

Briefly explain how the agent for Clayton Kershaw of the Los Angeles Dodgers could have used statistics to negotiate a contract similar to that of the Detroit Tigers' Justin Verlander in 2013.

Solution For Kershaw's agent to negotiate a contract similar to the one Justin Verlander signed in early 2013, he could have shown that the pitching performance of the two players was similar. To do this, he could have used statistics such as the earned run average of each pitcher (the average number of runs given up per 9 innings pitched), the winning percentages of each player, average strikeouts per year, and so on.

Example 3 Using Descriptive Statistics

Suppose you are graduating this semester. What statistics might be useful in helping you negotiate a fair salary?

Solution A good starting point for your negotiations would be the average starting salary for other students in your major. However, other statistics could be useful too. First, if you have a very good GPA, you could argue that you deserve an above-average starting salary simply because you are well above average. Another useful statistic would be the starting salaries for other fields of study in which you have taken courses. For example, you are graduating with a major in economics and a minor in computer science. If the starting salary for economics majors is $ 35,000 and the starting salary for computer science majors is $ 42,000, you may be able to argue that your computer science background entitles you to be paid more in line with computer science majors rather than being paid as an economics major.

Example 4 Population Versus Sample

Identify each of the following data sets as either a population or a sample:

a. The GPAs of all students at a college
b. The GPAs of a randomly selected group of students on a college campus
c. The ages of the nine Supreme Court Justices of the USA on January 1, 1842
d. The gender of every second customer who enters a movie theater
e. The lengths of Atlantic croakers caught on a fishing trip to the beach

Solution
a. Population
b. Sample
c. Population
d. Sample
e. Sample

Example 4 Population Versus Sample

Give an example of a population and two different characteristics that may be of interest.

Solution All currently registered students at a particular college form a population. Two population characteristics of interest could be the average GPA and the proportion of students over 23 years.

Supplementary Exercises

Multiple Choice

1. Statistics is a
   a. Method for organizing and analyzing data
   b. Method for describing data
   c. Method for making educated guesses based on limited data
   d. Method for summarizing data
   e. All of the above
2. Descriptive statistics can be used to
   a. Compare different sets of data
   b. Compare one observation to a set of data
   c. Make inferences about unknown events
   d. Make guesses about a population using a sample
   e. Both a and b
3. The average points per game for a basketball player represents
   a. A descriptive statistic
   b. An inferential statistic
   c. A deductive statistic
   d. An inductive statistic
   e. A sample

4. Using the Nielsen television ratings to estimate the number of television viewers represents
   a. Descriptive statistics
   b. Inferential statistics
   c. Deductive statistics
   d. Inductive statistics
   e. A population

5. Using the Gallup election poll to predict the outcome of an election represents
   a. Descriptive statistics
   b. Inferential statistics
   c. Deductive statistics
   d. Inductive statistics
   e. A population

6. Your GPA is
   a. Descriptive statistics
   b. Inferential statistics
   c. Deductive statistics
   d. Inductive statistics
   e. A sample

7. When you are interested in comparing data you would use
   a. Descriptive statistics
   b. Inferential statistics
   c. Deductive statistics
   d. Inductive statistics
   e. A census

8. Business people use
   a. Descriptive statistics
   b. Inferential statistics
   c. Deductive statistics
   d. Inductive statistics
   e. All of the above

9. If you were interested in predicting the outcome of the presidential election, you would use
   a. Descriptive statistics
   b. Inferential statistics
   c. Deductive statistics
   d. Inductive statistics
   e. A census

10. If a business is interested in predicting which flavor mouthwash will be favored by all consumers based on a sample of 500 people they should use
    a. Descriptive statistics
    b. Inferential statistics

c. Deductive statistics
d. Inductive statistics
e. A census

True/False (If false, explain why)

1. When we use the Nielsen television ratings to estimate the number of television viewers, we are using descriptive statistics.
2. The average salary for a graduating senior majoring in finance is a descriptive statistic.
3. Descriptive statistics can be useful in contract negotiations.
4. Inferential statistics can be used to compare two sets of data.
5. Descriptive statistics is used to make educated guesses about unknown events.
6. Statistics can be useful in organizing and presenting data.
7. Deductive statistics draws general conclusions based on specific information.
8. Inductive statistics draws specific conclusions based on general information.
9. Descriptive statistics is frequently used in court cases to make comparisons.
10. The true average salary a graduating senior receives, based on a sample of 100 graduating seniors, is a descriptive statistic.
11. The procedure for selecting sample elements from a population is called *sampling*.
12. *Depending on the method of sampling,* a sample can have more observations than the population.
13. A measurable characteristic of a population is called a *parameter*; while a measurable characteristic of a sample is called a *statistic*.
14. The mean of a sample is a statistic, but the standard deviation is not a statistic.
15. In a population with 10 objects, if the simple random sampling method is used to sample 3 objects, 1000 possible samples can be generated.

Questions and Problems

1. List three descriptive statistics commonly encountered in college.
2. List four descriptive statistics in baseball.
3. List four descriptive statistics commonly used in business.
4. Name two well-known sources of inferential statistics.
5. If you were interested in knowing what percentage of the 12 students in your art class will be attending the Picasso exhibit, would it be better to use a sample or a census? Why?
6. If you were interested in knowing what percentage of the 2000 students in your school will be attending the Picasso exhibit, would it be better to use a sample or a census? Why?

7. The owner of a factory regularly requests a graphical summary of all employees' salaries. The graphical summary of salaries is an example of descriptive statistics or inferential statistics?
8. A manager asked 50 employees in a company about their ages. On the basis of this information, the manager states that the average age of all the employees in the company is 39 years. The statement of the manager is an example of descriptive statistics or inferential statistics?
9. Refer to the table below, in which the prices of gasoline, crude oil, and gas oil from Aug. 1, 2013 to Aug. 23, 2013 are given. Answer the following questions.
   a. How many observations are in the data set?
   b. How many variables are in the data set?
10. Refer to the table below, which of the variables are qualitative and which are quantitative variables?

	Gasoline	Crude oil-brent	Gas oil
2013/8/1	291.38	110.05	125.1
2013/8/2	287.13	108.98	126.69
2013/8/5	281.88	108.98	126.3
2013/8/6	278.63	107.96	125.04
2013/8/7	273.63	107.68	124.32
2013/8/8	271.88	106.63	124.32
2013/8/9	279.38	107.47	124.32
2013/8/12	279.88	108.26	124.32
2013/8/13	284.13	109.69	124.32
2013/8/14	287.38	109.72	125.33
2013/8/15	289.13	111.05	126.39
2013/8/16	286.63	111.58	125.97
2013/8/19	283.88	111.8	126.25
2013/8/20	286.88	111.21	124.97
2013/8/21	288.38	111.23	124.91
2013/8/22	291.13	111	125.26
2013/8/23	294.13	112.15	125.01

Answers to Supplementary Exercises

Multiple Choice

1. e
2. e
3. a
4. b
5. b

6. a
7. a
8. e
9. b
10. b

True/False

1. False. Using the Nielsen television ratings to predict the number of television viewers represents inferential statistics.
2. True
3. True
4. False. Descriptive statistics is used to compare two sets of data.
5. False. Inferential statistics is used to make educated guesses.
6. True
7. False. Inductive statistics draws general conclusions based on specific information.
8. False. Deductive statistics draws specific conclusions based on general information.
9. True
10. False. Inferential statistics.
11. True
12. True
13. True
14. False
15. False

Questions and Problems

1. There are many descriptive statistics you may encounter in college including: GPA, mean SAT score of all freshmen, average high school class rank, average starting salary for graduating seniors, percentage of female students, percentage of minorities, and average GMAT score.
2. Again, there are many descriptive statistics including: batting average, slugging percentage, earned run average, average strikeouts per nine innings pitched, and stolen base percentage.
3. In the business world, descriptive statistics might include: average advertising dollars, average sales per month, average revenues per month, average salary of employees, and average bonus per employee.
4. Probably the two most widely recognized sources of inferential statistics are the Nielsen television ratings, and the Gallup polls on voting preferences.

5. Because the size of the population is small, it is probably easier to use a census and simply ask all 12 people whether they are going to the exhibit or not.
6. Because the size of the population is relatively large, it may be better to use a sample of students and use inferential statistics to determine the percentage of all students who will be attending the exhibit.
7. Yes.
8. Yes.
9. (a) 23 observations (b) 3 variables.
10. All the 3 variables are quantitative.

Chapter 2
Data Collection And Presentation

Chapter Intuition

Once the data have been collected, they need to be organized and summarized so they can be used to make inferences about the population. Descriptive statistics is for that purpose. For example, if there are a lot of students in a class, it will be difficult for the teacher to draw any conclusions about the class's performance by simply looking at the students' scores. However, by organizing the data into descriptive statistics such as tables or graphs, the teacher can easily determine the performance of the class. This chapter discusses how to collect and present data in a systematic way so that it can be easily analyzed.

Chapter Review

1. Data can come from either a ***primary source***, which means it is collected specifically for the study, or from a ***secondary source***, which means that the data was originally collected for some other purpose.
2. Before collecting the data, the researcher must decide whether to collect a ***sample*** or a ***census***. In a census, all members of the population are surveyed. When the population is small, a census is possible. For example, to survey the opinion about a movie, a family of five would find little difficulty in surveying all five members. However, when the population is large, it may be infeasible to survey every member. In this case, it may be more desirable to collect a subset of the population, that is, a sample, from the population. For example, if we are interested in the views of all citizens in the USA concerning a national health care policy, it would be better to examine the views of asubset of citizens and make inferences about the entire population based on this sample.
3. There are two types of errors that are associated with primary and secondary data. ***Random error*** is the difference between the value obtained by taking a

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_2

random sample and the value obtained by taking a census. ***Systematic error*** results when there are problems in measurement.

4. One way to organize data is to place them into groups, or classes, and then to present the data using tables.
5. Charts and graphs are another way for presenting data.
   a. A ***pie chart*** shows how the "whole," the pie, is divided into different pieces, the pieces of the pie.
   b. A ***bar chart*** can also be used to display data. A company might use bar graph to see how its advertising dollars are spent on television, radio, and newspapers.
   c. A ***line chart*** can be used to show the relationship between two different variables or how one variable changes over time. For example, a company might be interested in seeing how its advertising expenditures have changed over time.
   d. A ***time series graph*** is a line graph in which the variable on the X-axis represents time, such as the year or month of the data. The line graph is also a time series graph.

Example 1 Presenting Data in Graphs

An economics professor gives the following grades to her class:

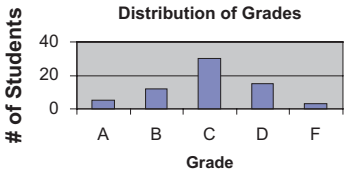
Number of students	Grade
5	A
12	B
30	C
15	D
3	F

a. Use a bar chart to show the distribution of grades.
b. Use a pie chart to show the distribution of grades.
c. Which of these graphs do you think is best for presenting the distribution of grades? Why?

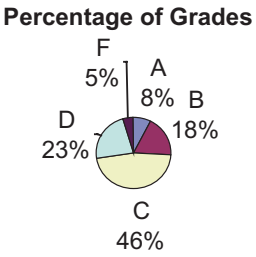
Example Problems

Solution:

a.



b.



c. Both graphs do a good job of presenting the data. However, the two graphs present different aspects of the data. The bar chart gives us information to compare the grades. For example, we can see from the bar chart that most students received a grade C and fewer students received higher and lower grades. On the other hand, the pie chart shows us the allocation of the grades among the 65 students. Pie charts are best used when we are interested in seeing how the whole is divided among smaller subgroups.

Example 2 Primary Versus Secondary Sources of Data

Which of the following are from primary sources and which are from secondary sources?

a. The Dow Jones Industrial Average taken from the *Wall Street Journal*.
b. Responses to a survey on how chief financial officers will respond to a change in the accounting rules.
c. Johnson & Johnson's earnings as given in its annual report.

Solution:

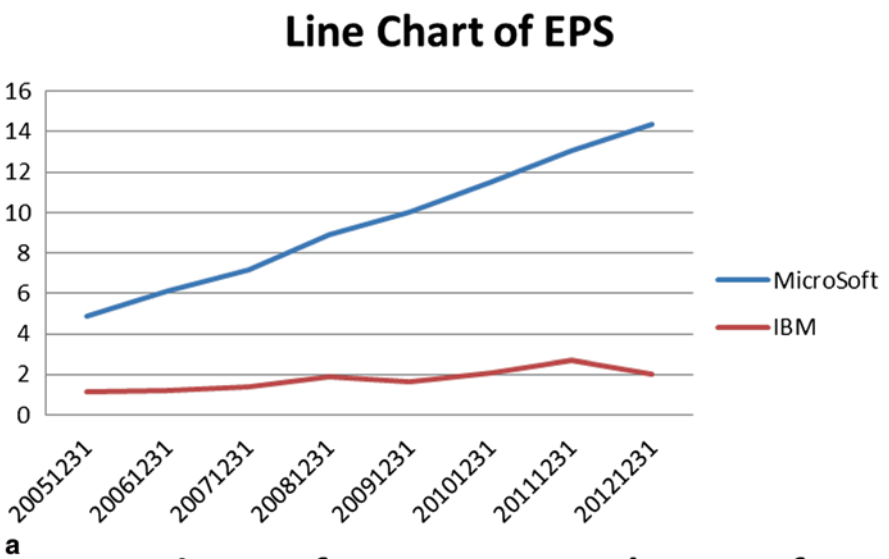
a. Secondary source
b. Primary source
c. Secondary source

Example 3 Charts for Financial Data

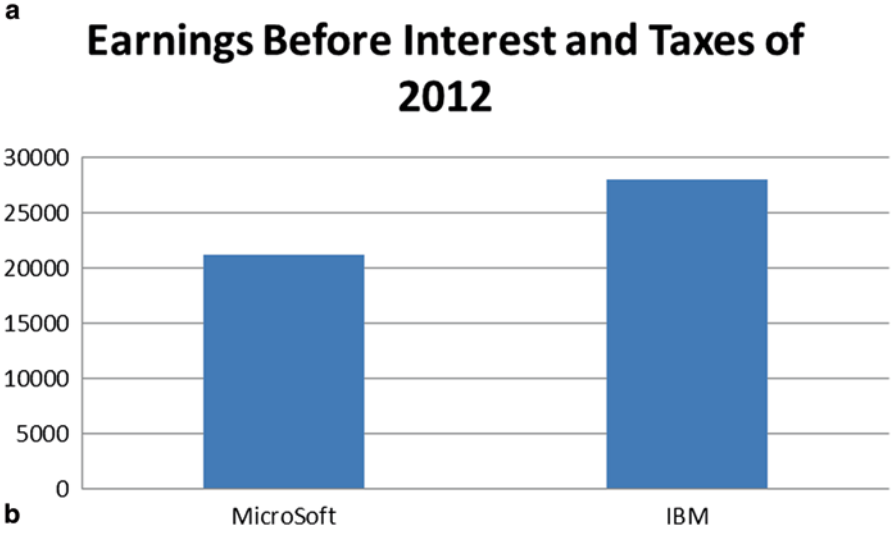
Refer to the data below, in which data on EBIT and EPS, during the period 2005–2012 are given. (a) Draw the line chart of EPS for the Microsoft and IBM during the period 2005–2012. (b) Draw the bar chart for the EBIT of Microsoft and IBM during the year 2012.

	MicroSoft		IBM	
	Earnings before interest and taxes	EPS	Earnings before interest and taxes	EPS
20051231	11,714	4.87	16,642	1.12
20061231	12,614	6.11	17,385	1.2
20071231	14,360	7.18	18,524	1.42
20081231	16,750	8.93	23,992	1.87
20091231	17,719	10.01	20,693	1.62
20101231	18,705	11.52	24,157	2.1
20111231	21,578	13.06	27,161	2.69
20121231	21,173	14.37	27,956	2

Solution:



a



b

Supplementary Exercises

Multiple Choice

1. Data from primary source

   a. are collected for other purposes than the current study.
   b. can be obtained from the newspaper.
   c. are collected specifically for the current study.
   d. are less reliable than data from a secondary source.
   e. indicate correlation.

2. Data from secondary source

   a. are collected for other purposes than the current study.
   b. can be obtained from the newspaper.
   c. are collected specifically for the current study.
   d. are more reliable than data from a primary source.
   e. both a and b.

3. A census

   a. consists of information from all members of the population.
   b. consists of information from a subset of the population.
   c. is a secondary source of data.
   d. is less reliable than a secondary source of data.
   e. is less reliable than a primary source of data.

4. A sample

   a. consists of information from all members of the population.
   b. consists of information from a subset of the population.
   c. is a secondary source of data.
   d. is less reliable than a secondary source of data.
   e. is less reliable than a primary source of data.

5. If you are interested in how a football coach divides the practice session into drills, weight training, scrimmaging and game plans, it would be best to use a

   a. bar graph.
   b. line graph.
   c. pie chart.
   d. times series graph
   e. component-part line chart.

6. If you are interested in how the earnings of a company have fluctuated over time, it would be best to use a

   a. bar graph.
   b. time series graph.
   c. pie chart.

d. component-part line chart.
e. histogram.

7. If you were interested in comparing the average earnings and interest expense for IBM with the average earnings and interest expense for Apple Computers, it would be best to use a

a. bar graph.
b. line graph.
c. pie chart.
d. time series graph.
e. histogram.

8. In measuring the height of students, a systematic error could occur if

a. the ruler used to measure students' height is one inch too long.
b. we measure the height of the wrong students.
c. we ask all students to remove their shoes.
d. both male and female students are measured.
e. we forget to record one student's height.

9. A random error could occur if

a. the ruler used to measure students' height is one inch too long.
b. we measure the height of the wrong students.
c. we forget to have students remove their shoes.
d. both male and female students are measured.
e. we forget to record one student's height.

10. It is best to use a census when conducting a survey if

a. the population is large.
b. the population is small.
c. we have a limited amount of time to conduct the survey.
d. we would like to keep the costs of the survey low.
e. the population is spread over a large geographic region.

11. It is best to use a sample when conducting a survey if

a. the population is large.
b. the population is small.
c. we have a limited amount of time to conduct the survey.
d. we would like to keep the costs of the survey low.
e. all of the above except b.

12. Graphs can be useful for

a. summarizing large amounts of data.
b. showing trends in data.
c. adding visual appeal to business reports.
d. making comparisons.
e. all of the above.

In 13–15, determine if the data is a primary or secondary source.

13. The Educational Testing Service tests 8th graders nationwide to determine average mathematics scores on a standardized test.
14. The captain of a softball team keeps track of batting averages in order to determine batting order.
15. To determine which marketing techniques might be more effective, an insurance executive looks at the composite profile of customers in his region, regularly compiled by his marketing department.

True/False (If False, Explain Why)

1. A golfer interested in knowing the percentage of total shots he hits with each different club should use a line graph.
2. Primary data are data collected specifically for the study.
3. Stock price data collected from the Wall Street Journal are primary data.
4. When the Nielsen television ratings collect data on television viewers to estimate the number of television viewers, they are using secondary data.
5. Random error is the difference between the value taken from a random sample and the value obtained from taking a census.
6. Systematic error results from problems in measurement.
7. Line charts are good for showing how the whole is divided into several parts.
8. Time series graphs show how data fluctuate over time.
9. The Gallup election poll uses primary data.
10. A component-part line chart can be useful for showing how gross national product is divided between consumption, investment, government spending, and net exports.
11. Time series graphs can be useful for examining trends in a company's financial ratios.
12. Financial ratios are often used in accounting and finance because they allow us to compare companies of different sizes.
13. Systematic errors result from the sample differing significantly from the entire population.
14. Random errors can be completely eliminated by appropriate random sampling procedure.
15. Systematic errors can be completely eliminated by appropriate random sampling procedure.
16. A bar chart is used to illustrate the trend of a variable versus time.

Questions and Problems

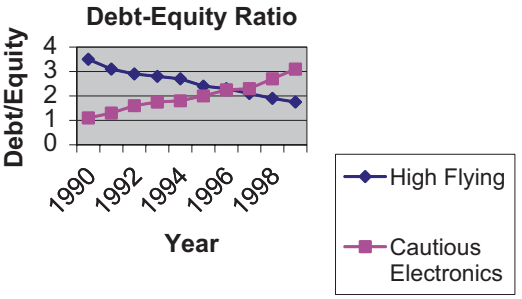
1. Suppose this month you spend $ 250 on rent, $ 125 on food, and $ 75 on entertainment. Use a pie chart to show how your money was spent.

2. Suppose the average earnings per share for ABC Company is $ 3 and the average earnings per share for XYZ Corporation is $ 6. Use a bar chart to compare the EPS for the two companies.
3. Below is the advertising budget for the Shady Lamp Shade Company from *1995 to 1999*.

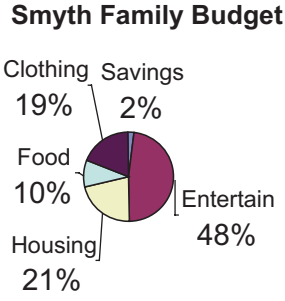
Year	TV	Radio	Newspaper	Total
1995	100	25	18	143
1996	105	31	27	163
1997	115	40	33	188
1998	123	42	34	199
1999	151	47	39	237

Use a component-part line chart to show how the advertising budget has been divided between television, radio, and newspaper advertising over this 5-year period.

4. Briefly explain why financial ratios are preferred to absolute numbers in financial analysis.
5. Below is a time series graph of the debt-equity ratios for the Cautious Electronics Company and the High Flying Electric Company. What conclusions can you draw by examining this graph?



6. Below is a pie graph of the Smyth family's household budget. As their financial planner, what advice might you offer?



7. Refer to the data in Example 3, in which data on EBIT and EPS, during the period 2005–2012 are given. (a) Draw the line chart of EBIT for the Microsoft and IBM during the period 2005–2012. (b) Draw the bar chart for the EPS of Microsoft and IBM during the year 2012.

Answers to Supplementary Exercises

Multiple Choice

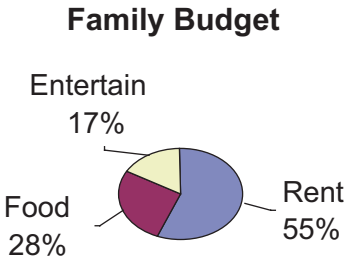
1. c	6. b	11. e
2. e	7. a	12. e
3. a	8. a	13. e
4. b	9. b	14. a
5. c	10. b	15. b

True/False

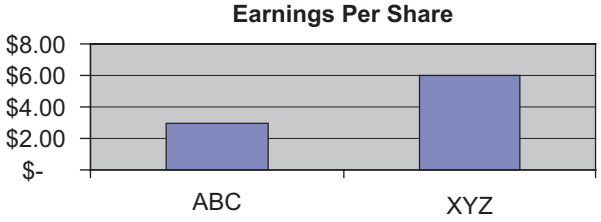
1. False. A pie graph is best for showing how the "whole" is divided into parts.
2. True
3. False. Secondary data.
4. False. Primary data.
5. True
6. True
7. False. Pie chart.
8. True
9. True
10. True
11. True
12. True
13. True
14. False
15. False
16. False

Questions and Problems

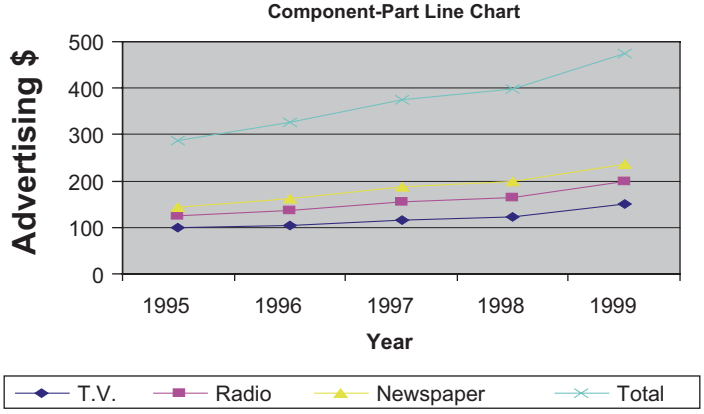
1.



2.



3.

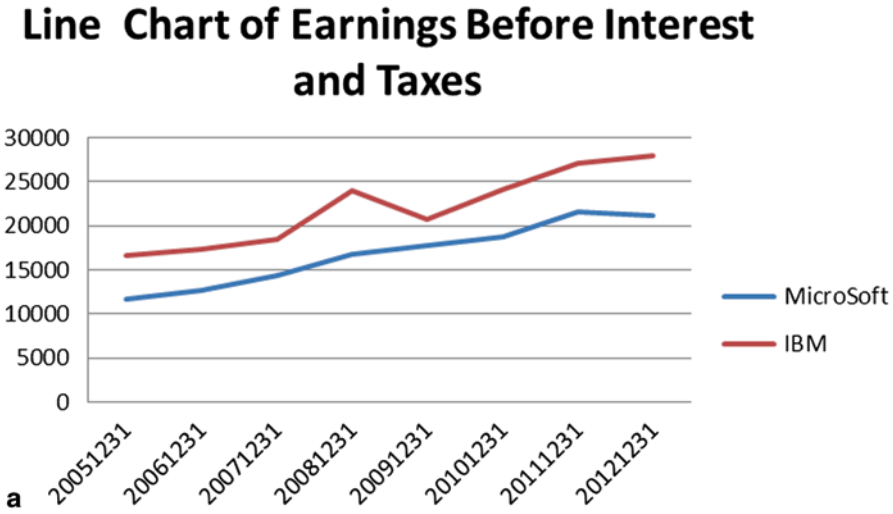


4. Ratios are generally preferred to absolute numbers in financial analysis because they allow us to compare companies of different sizes.
5. From the graph, we can see that the two companies' debt-equity ratios have been moving in different directions over the last few years. High Flying has been

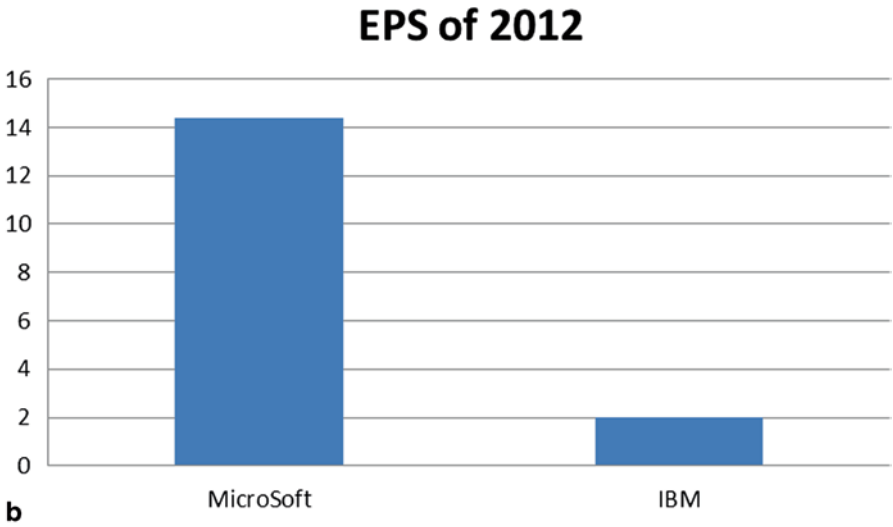
reducing the amount of borrowing (debt) over recent years, whereas Cautious has been increasing its debt. Although the average debt-equity ratios for the two companies may be similar, it is clear that in terms of borrowing they are moving in opposite directions. Time series graphs are important because they allow us to compare trends in the financial ratios of different companies.

6. Looking at the Smyth family's budget, we can see that they spend a very high proportion of their budget on clothing and entertainment and put a very small amount into savings. As their financial planner, you might advise them to reduce their entertainment and clothing expenses and try to increase their savings.

7.



a



b

Chapter 3
Frequency Distributions and Data Analyses

Chapter Intuition

Suppose 400 students are taking a freshman economics class. The midterm grades of the 400 students would be very difficult to analyze due to the large amount of data. However, by separating the grades into groups such as 90–100, 80–89, etc., it becomes much easier to evaluate the performance of the class. By setting up the classifications and observing the number of events falling into each category, we are able to construct a ***frequency table.*** Frequency tables can be useful for a teacher who wants to know how many students scored 90 percent or above in an exam or for a car dealer who wants to keep a tally of the number of cars sold each month.

Chapter Review

1. Once data have been collected, they need to be organized for analysis. For a large amount of data, it would be more useful to ***group the data***. For example, if there are 400 students, the teacher might count the numbers of students who scored within the intervals 96–100, 91–95, 86–90, …, etc. using a ***tally table*** for the midterm exam. This would make it easier to see how the students performed. The number of scores in each *interval is called* the ***frequency***, and the table that shows these frequencies is called a ***frequency table***.
2. *Sometimes we are not only interested in the* ***frequency*** *in each group, but also the* ***cumulative frequency*** *of all the groups which came before. From our previous example, the teacher might be interested in the number of students who scored higher than 90, which is the sum of the frequencies for the first two intervals, i.e., 96–100 and 91–95.*
3. *Sometimes it is more interesting to look at the number of occurrences as a proportion of all the data. When we do this, we are looking at the* ***relative frequency***

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_3

*and the **cumulative relative frequency**. Like cumulative frequency, cumulative relative frequency keeps track of the total proportion of occurrences in the current group and all previous groups.*

4. Graphs can be a very effective way to present the data. A bar chart is used to present the ***frequency table*** of qualitative data, while a ***histogram*** can be used to present the frequency, cumulative frequency, relative frequency and relative cumulative frequency of quantitative data.
5. *A **stem-and-leaf display** is less of a graph and more of a table. In a stem-and-leaf display, the first digits of an observation are presented in one column (the tree trunk) in ascending order. The remaining digits are presented to the right of the tree trunk, i.e., the leaves of the tree.*
6. The ***Lorenz curve*** is a cumulative frequency curve that shows the distribution of a society's income. The ***Gini coefficient*** is used in conjunction with the Lorenz curve to express income distribution. If the Gini coefficient is equal to 0, we have perfect equality of income; that is, 20% of the population has 20% of the society's income, etc. If the Gini coefficient equals 1, then we have absolute inequality of income, that is, one family or person receives all of the society's income.

Example Problems

Example 1 Constructing Tally Tables

Suppose the midterm exam scores from Ms. Hannah's economics course are: 98, 95, 80, 72, 65, 90, 71, 55, 83, 88, 77, 79, 66, 45, 62. Construct a tally table of midterm exams using the following intervals: 90–100, 80–89, 70–79, 60–69, 50–59, and less than 50.

Solution:

90–100	///
80–89	///
70–79	////
60–69	///
50–59	/
< 50	/

Example 2 Frequency and Cumulative Frequency

Use the tally table you constructed in Example 1 to construct a frequency and cumulative frequency table.

Solution:

Frequency Table

Score	Frequency	Cumulative frequency
90–100	3	3
80–89	3	6
70–79	4	10
60–69	3	13
50–59	1	14
< 50	1	15

Example 3 Relative and Cumulative Relative Frequency

Use your results from Example 2 to construct a relative frequency and cumulative relative frequency table.

Solution:

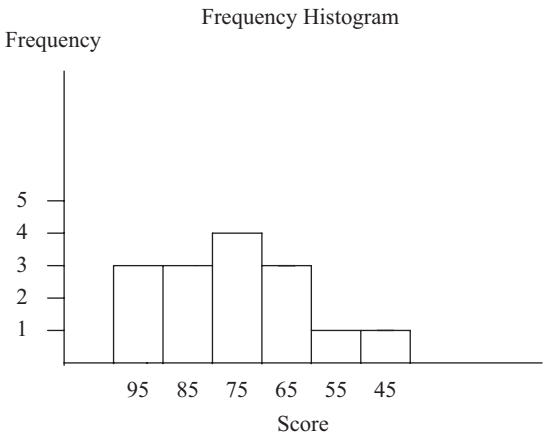
Relative Frequency Table

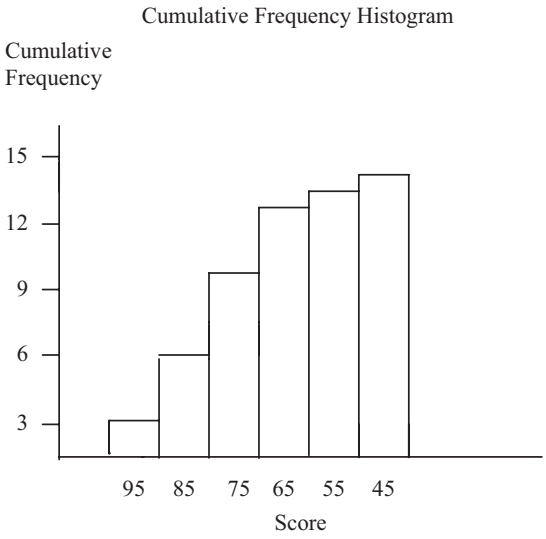
Score	Frequency	Relative frequency	Cumulative relative frequency
90–100	3	0.200	0.200
80–89	3	0.200	0.400
70–79	4	0.267	0.667
60–69	3	0.200	0.867
50–59	1	0.067	0.933
< 50	1	0.067	1.000
Total	15		

Example 4 Frequency and Cumulative Frequency Histograms

Use your results from Example 2 to construct a frequency and cumulative frequency histogram.

Solution:

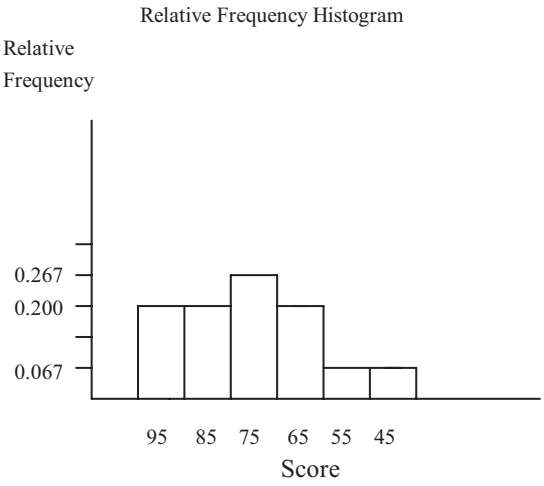


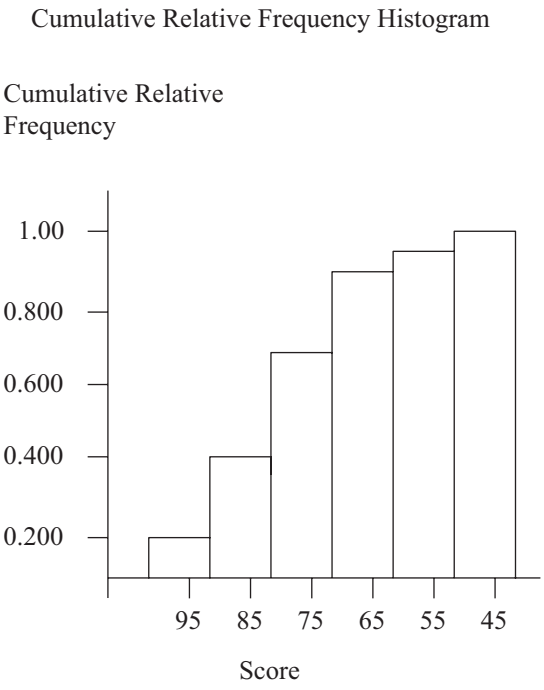


Example 5 Relative and Cumulative Relative Frequency Histograms

Use your results from Exercise 3 to construct a relative and cumulative relative frequency histogram.

Solution:


Relative Frequency Histogram
Relative
Frequency
0.267
0.200
0.067
95 85 75 65 55 45
Score


Cumulative Relative Frequency Histogram
Cumulative Relative
Frequency
1.00
0.800
0.600
0.400
0.200
95 85 75 65 55 45
Score

Example 6 Stem-and-Leaf

Suppose 30 students in a statistics class took a test and made the following scores:

86 80 25 77 73 76 100 90 69 93 90
83 70 73 73 70 90 83 71 95 74
40 58 68 69 100 78 87 97 92

Solution:

2	5
3	
4	0
5	8
6	9 8 9
7	7 3 6 0 3 3 0 1 8 4
8	6 0 3 3 7
9	0 3 0 0 5 7 2
10	0 0

Supplementary Exercises

Multiple Choice

1. A tally table is

   a. used to organize raw data.
   b. the next step after constructing a frequency table.
   c. another name for a frequency histogram.
   d. similar to a stem-and-leaf display.
   e. an example of the Gini coefficient.

2. Data reported in a frequency table with class intervals is

   a. grouped data.
   b. ungrouped data.
   c. raw data.
   d. suitable primarily for academic use.
   e. always reported in a stem-and-leaf display.

3. The number of observations in a particular class interval is known as the

   a. frequency.
   b. relative frequency.

c. cumulative frequency.
d. cumulative relative frequency.
e. fixed frequency.

4. The proportion of observations in a particular class interval is known as the

a. frequency.
a. relative frequency.
b. cumulative frequency.
c. cumulative relative frequency.
d. fixed frequency.

5. The number of observations in the current class interval plus all previous class intervals is known as the

a. frequency.
b. relative frequency.
c. cumulative frequency.
d. cumulative relative frequency.
e. fixed frequency.

6. The proportion of observations in the current class interval as well as all previous class intervals is known as the

a. frequency.
b. relative frequency.
c. cumulative frequency.
d. cumulative relative frequency.
e. fixed frequency.

7. A stem-and-leaf display can be used for summarizing

a. very large amounts of data.
b. small amounts of data.
c. any amount of data.
d. grouped data.
e. tally table results.

8. A frequency polygon is obtained by

a. constructing a frequency histogram.
b. constructing a cumulative frequency histogram.
c. linking the midpoints from a frequency histogram.
d. using a line graph.
e. constructing a stem-and-leaf display.

9. The Lorenz curve is

a. a specific type of histogram.
b. a specific pie chart.
c. a curve showing a society's distribution of income.

d. a stem-and-leaf display.
e. a frequency polygon.

10. When the Gini coefficient is equal to 1 there is

a. absolute equality of income.
b. absolute inequality of income.
c. a 100% income tax.
d. no income in the society.
e. economic growth.

11. When the Gini coefficient is equal to 0 there is

a. absolute equality of income.
b. absolute inequality of income.
c. a 100% income tax.
d. no income in the society.
e. economic growth.

12. Histograms are similar to bar graphs except

a. neighboring bars do not touch each other.
b. the area inside any bar is proportional to the number of observations in the corresponding class.
c. the midpoints of the bars are connected.
d. they are relative to the Gini coefficient.
e. they are derived from frequency polygons.

A local convenience store owner records how many customers enter the store each day over a 25-day period. The results are as follows: 20, 21, 13, 21, 45, 24, 10, 34, 25, 32, 42, 23, 21, 11, 41, 16, 39, 22, 21, 23, 25, 12, 31, 24, 31.

Prepare a stem-and-leaf plot as follows. And answer the problems #13–15.

Stem	Leaf
1	(i)
2	
3	(ii)
4	(iii)

13. What is the content of line (i)?

a. 0,1,2,3,6
b. 1,1,3,6,6
c. 0,1,1,3,6,6
d. 0,1,2,3,7

14. What is the content of line (ii)?

    a. 0,1,2,3,6
    b. 1,1,2,4,9
    c. 1,3,3,4,6
    d. 0,1,2,3,7

15. What is the content of line (iii)?

    a. 0,1,2,3
    b. 1,2,5
    c. 3,4,6
    d. 0,1,2

True/False (If false, explain why)

1. Raw data are data that have been grouped into classes.
2. When the Lorenz curve is a straight line, there is perfect equality of income.
3. It is usually best to limit the amount of data to 100 observations when using a stem-and-leaf display.
4. Grouping data makes handling a large dataset less manageable.
5. Information may be lost when raw data are grouped.
6. Frequencies, cumulative frequencies, relative frequencies, and cumulative relative frequencies can all be graphed using histograms.
7. The area inside each bar of a histogram is proportional to the number of observations in that class.
8. A tally table is usually the first step in creating a frequency table.
9. When grouping data, the number of classes is unimportant.
10. Relative frequencies can be computed by adding together the frequencies of the current and all previous classes.
11. Cumulative frequencies measure the proportion of observations in a particular class.
12. A frequency polygon is obtained by linking the midpoints of the class intervals in a frequency histogram.
13. Cumulative frequencies can be useful for a teacher who would like to know the number of students receiving a grade of C or better.
14. A frequency table is not a descriptive statistic.
15. A frequency table can be used for quantitative as well as qualitative data.

Questions and Problems

1. Suppose the grades on a midterm exam in economics are: 95, 88, 92, 71, 64, 32, 89, 85, 90, 99, 72, 73, and 61. Construct a stem-and-leaf display for the data.
2. Use the data in Problem 1 to construct a tally table, with classes 91–100, 81–90, 71–80, 61–70, and 60 and below.
3. Use the tally table from Problem 2 to construct a frequency and cumulative frequency table.
4. Use the tally table from Problem 2 to construct a relative frequency and cumulative relative frequency table.
5. Draw a frequency and cumulative frequency histogram using the results from Problem 3.
6. Draw a relative and cumulative relative frequency histogram using the results from Problem 3.
7. Sixty adults with gum disease were asked the number of times per week they flossed before their diagnoses. The (incomplete) results are shown below

# Flossing per week	Frequency	Relative frequency	Cumulative relative frEq.
0	27	0.4500	0.4500
1	18		
3			0.9333
6	3	0.0500	
7	1	0.0167	

   a. What is the frequency of adults flossing 3 times per week?
   b. What is the cumulative relative frequency of adults flossing 1 time per week?
   c. What is the cumulative relative frequency of adults flossing 6 times per week?

Answers to Supplementary Exercises

Multiple Choice

1. a	6. d	11. b
2. a	7. b	12. b
3. a	8. c	13. a
4. b	9. c	14. b
5. c	10. a	15. b

True/False

1. False. Raw data have not been grouped.
2. True
3. True
4. False. Grouping data makes a large dataset more manageable.
5. True
6. True
7. True
8. True
9. False. Using too many classes defeats the purpose, whereas using too few classes limits the amount of information.
10. False. Relative frequencies can be computed by dividing the frequency in one class by the total number of observations in all classes.
11. False. Cumulative frequencies are the sum of observations in a particular class plus all preceding classes.
12. True
13. True
14. False
15. True

Questions and Problems

1.

Stem-and-Leaf Display

9	0 2 5 9
8	5 8 9
7	1 2 3
6	1 4
5	
4	
3	2

2. Tally Table

Score	
91–100	///
81–90	////
71–80	///
61–70	//
<60	/

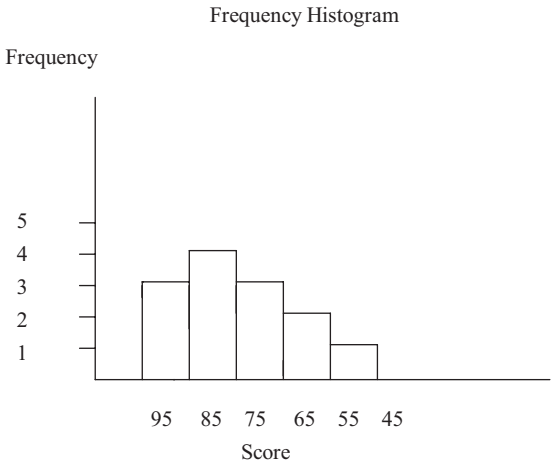
3. Frequency and Cumulative Frequency Table

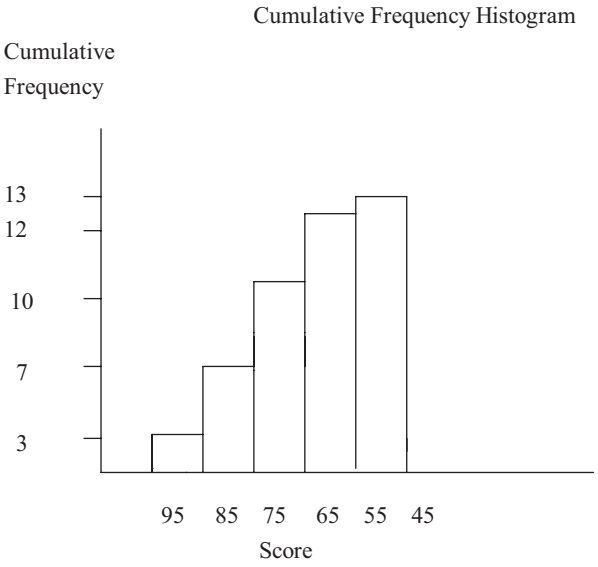
Score	Relative frequency	Cumulative frequency
91–100	3	3
81–90	4	7
71–80	3	10
61–70	2	12
<60	1	13

4. Relative and Cumulative Relative Frequency

Score	Relative frequency	Cumulative relative frequency
91–100	0.231	0.231
81–90	0.308	0.538
71–80	0.231	0.769
61–70	0.154	0.923
<60	0.077	1.000

5.

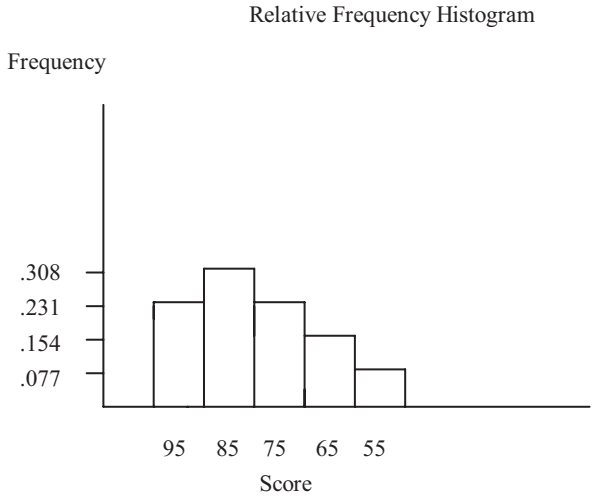


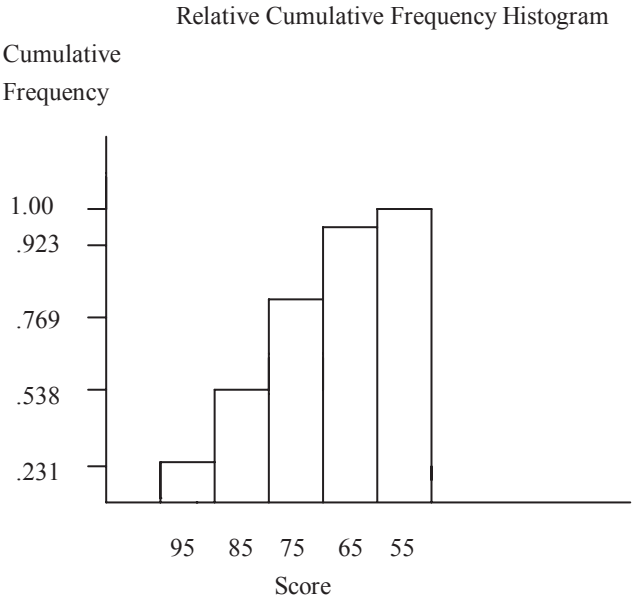


6.

(a) 11
(b) 0.7500
(c) 0.9833

7.




Relative Cumulative Frequency Histogram
Cumulative
Frequency
1.00
.923
.769
.538
.231
95
85
75
65
55
Score

Chapter 4
Numerical Summary Measures

Chapter Intuition

This chapter deals with ways to summarize a large amount of data by using one or two numbers that describe all the data. These summary measures can be used to compare different sets of data or to compare a single observation with the entire set of data. For example, suppose an instructor is teaching two courses in statistics. If you were simply to look at all the scores from each class, it would be difficult to say anything about the performance of each class. However, by computing the ***mean*** or average of each class, a simple comparison can be made. Similarly, a single student's grade could be compared with the class average to see how he or she is doing.

Once we have computed the average *of a set of data*, we are interested in knowing whether this number is meaningful. Measures of dispersion such as the ***variance or standard deviation*** allow us to determine if our measure of central tendency is meaningful. For example, suppose we have two classes each consisting of three students. In the first class, the midterm scores are 0, 50, and 100, and in the second class the scores are 49, 50, and 51. In both classes the mean will be 50; however, in the first class the large dispersion of scores makes the mean not very meaningful, whereas in the second class, the small dispersion of scores makes the mean very meaningful. The standard deviation or variance measures the deviation of each observation from the mean. If the data are very far away from the mean, the mean will not be a good measure of the center of the data, and therefore will not be very meaningful.

The shape of the distribution can be very interesting in business and economics, for example, when we are looking at the distribution of stock returns, thus we are often interested in the shape of the distribution. The ***skewness coefficient*** allows us to determine if the distribution is symmetric or skewed. In a symmetric distribution, the three measures of central tendency, ***mean***, ***median***, and ***mode*** are all the same. This means that observations are as likely to be above the mean as below. If the distribution is skewed, an individual observation is more likely to be below the mean (positive skewness) or above the mean (negative skewness).

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_4

Finally, kurtosis provides another measure of the shape of the distribution. Kurtosis looks at the height and sharpness of the peak, sometimes referred to as "peakedness." Higher values indicate a higher, sharper peak and fatter tails in the distribution. Lower values indicate a lower, less distinct peak with narrower tails in the distribution. This can be valuable because it allows us to see if more of the variability is due to a few extreme differences from the mean (high kurtosis) or a lot of modest differences from the mean (low kurtosis). For a normal distribution, the kurtosis will be 3. It is common practice to use an adjusted version of kurtosis, **excess kurtosis**, which looks at the kurtosis of a distribution relative to the normal distribution. A distribution that has kurtosis less than 3 or negative excess kurtosis is referred to as platykurtic, whereas a distribution that has a kurtosis greater than 3 or positive kurtosis is referred to as leptokurtic.

Chapter Review

A series of data can be described using several descriptive statistics.

1. The ***mean***, ***median***, and ***mode*** are measures of the central tendency of the data. In the simplest terms, measures of central tendency give us an idea of where the average of the data is. The mean is the simple average of the data: the sum of the data divided by the number of observations. The median is the number in the middle. The mode is the most frequently repeating number. The median would be preferred to the other two when the data represent extreme values. For example, the mean salary of baseball players would be greatly influenced by the very large salaries of superstars such as Mark McGuire and Greg Maddox. The median would be a more meaningful figure. We would use the mode when we are interested in looking at the popularity of something. For example, a clothing store might be interested in the shoe size most often purchased. In this case, the mode would provide the store with information about the size of shoes that are in the greatest demand.
2. The ***variance***, ***standard deviation***, ***coefficient of variation***, and ***mean absolute deviation*** are measures of the dispersion of the data or how the data are spread out around their average. The variance measures the average squared distance of the observations from the mean. The standard deviation is just the square root of the variance. The mean absolute deviation measures the average distance of the observations from the mean. If we are interested in using descriptive statistics to compare data with different units, like the risk for two different stocks, we may use the coefficient of variation, which is unit free.
3. The shape of a distribution is described by its ***skewness*** or ***coefficient of skewness*** and by its ***kurtosis***. Skewness measures whether or not the data are symmetrical. Kurtosis measures the peakedness of the distribution.

Useful Formulas

Measures of central tendency	Mean absolute deviation:
Sample arithmetic mean:	$\mathrm{MAD} = \dfrac{\sum_{i=1}^{N} \lvert x_i - \mu \rvert}{N}$
$\bar{x} = \dfrac{\sum_{i=1}^{N} x_i}{N}$	Coefficient of variation:
Sample geometric mean:	$\mathrm{CV} = \dfrac{s}{\bar{x}}$
$\bar{x}_g = (x_1 \times x_2 \times x_3 \times \ldots \times x_N)^{1/N}$	Population standard deviation:
Grouped mean:	$\sigma = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$
$\bar{x} = \dfrac{\sum_{i=1}^{k} f_i m_i}{\sum_{i=1}^{k} f_i}$	Sample standard deviation:
Median:	$s = \sqrt{\dfrac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}}$
$m = L + \dfrac{(N/2 - F)}{f}(U - L)$	Population variance for frequency distribution:
Measures of dispersion	$\sigma^2 = \dfrac{\sum_{i=1}^{k} f_i (m_i - \mu)^2}{N}$
Population variance:	Sample variance for frequency distribution:
$\sigma^2 = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \dfrac{\sum_{i=1}^{k} f_i (m_i - \bar{x})^2}{N-1}$
Sample variance:	
$s^2 = \dfrac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N-1}$	

Measures of Skewness

Skewness:

$$\mu_3 = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^3}{N}}$$

Coefficient of skewness:

$$CS = \frac{\mu_3}{\sigma^3}$$

Measures of Kurtosis

The fourth moment:

$$\mu_4 = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^4}{N}}$$

Coefficient of Kurtosis:

$$k = \frac{\mu_4}{\sigma^4} - 3$$

Example Problems

Example 1 Computing the Average for Ungrouped Data

Suppose you are given the following sales of 12 car dealers for last month. Compute the mean, median, and mode.

38, 41, 67, 63, 32, 50, 58, 74, 28, 69, 43, 63

Solution:

$$\text{mean} = (38 + 41 + 67 + 63 + 32 + 50 + 58 + 74 + 28 + 69 + 43 + 63)/12 = 52.167$$

The median is the number in the middle. When the number of observations is an even number, the median is the mean of the middle two numbers. To find the median, we first rank the data from the smallest to the largest. The middle two numbers in our example are 50 (5th observation) and 58 (6th observation). Their mean is $(50+58)/2 = 54$.

The mode is the number that occurs most frequently. Since each number occurs only once in this problem, there is no mode. Note that the mean and median are equal to each other in this example merely by coincidence.

Example 2 Computing Measures of Dispersion

Using the information from **Example 1**, let's compute the population variance, standard deviation, mean absolute deviation, and coefficient of variation.

Solution:

$$\begin{aligned} s^2 = {} & [(38-52.17)^2 + (41-52.17)^2 + (67-52.17)^2 + (63-52.17)^2 + (32-52.17)^2 \\ & + (50-52.17)^2 + (58-52.17)^2 + (74-52.17)^2 + (28-52.17)^2 + (28-52.17)^2 \\ & + (28-52.17)^2 + (69-52.17)^2 + (43-52.17)^2 + (63-52.17)^2]/(12-1) = 241.24 \end{aligned}$$

$$s = \sqrt{\sigma^2} = \sqrt{114} = 15.53$$

$$\begin{aligned} \text{MAD} = {} & [|38-52.17| + |41-52.17| + |67-52.17| + |63-52.17| \\ & + |32-52.17| + |50-52.17| + |58-52.17| + |74-52.17| \\ & + |28-52.17| + |69-52.17| + |43-52.17| + |63-52.17|]/12 = 13.5 \end{aligned}$$

$$\text{CV} = 15.53/52.17 = .2977$$

Example 3 Computing the Mean for Grouped Data

Suppose we use the data from **Example 1** and group the data from 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, and 66–70.

Solution:

First, group the data and find the midpoints and frequency for each group.

Classes	Midpoints	Frequency	
	m_i	f_i	$f_i m_i$
36-40	38	3	114
41-45	43	2	86
46-50	48	1	48
51-55	53	0	0
56-60	58	1	58
61-65	63	2	126
66-75	68	3	*204*
			636

$$\text{Mean} = [(3\times 38)+(2\times 43)+(1\times 48)+(0\times 53)+(1\times 58) +(2\times 63)+(3\times 68)]/12 = 53$$

$$\text{Median} = 53$$

Example 4 Computing Measures of Dispersion for Grouped Data

Compute the variance, standard deviation, and coefficient of variation for the grouped data of Example 3.

Solution:

$$s^2 = [3(38-53)^2 + 2(43-53)^2 + 1(48-53)^2 + 0(53-53)^2 + (58-53)^2 + 2(63-53)^2 +0(53-53)^2 + 1(58-53)^2 + 2(63-53)^2 + 3(68-53)^2]/(12-1) = 163.63$$

$$s = \sqrt{\sigma^2} = 12.79$$

$$\text{CV} = 10.36/53 = .2414$$

Example 5 Measuring the Shape of the Data—Skewness

Use the data from Example 1 to compute the coefficient of skewness. How are the data skewed?

Solution: First we need to compute μ_3.

$$\begin{aligned}\mu_3 = \; & [(38-52.17)^3+(41-52.17)^3+(67-52.17)^3+(63-52.17)^3+(32-52.17)^3 \\ & +(50-52.17)^3+(58-52.17)^3+(74-52.17)^3+(28-52.17)^3+(28-52.17)^3 \\ & +(28-52.17)^3+(69-52.17)^3+(43-52.17)^3+(63-52.17)^3]/12 = -512.43\end{aligned}$$

$$CS = \mu_3 / \sigma^3 = -512.43 / 3746.97 = -0.1367$$

Example 6 Measuring the Shape of the Data—Kurtosis

Use the data from Example 1 to compute the coefficient of skewness. How are the data skewed?

This result, being a negative number, means the data are slightly skewed to the left.
Solution: First we need to compute μ_4.

$$\begin{aligned}\mu_4 = \; & [(38-52.17)^4+(41-52.17)^4+(67-52.17)^4+(63-52.17)^4+(32-52.17)^4 \\ & +(50-52.17)^4+(58-52.17)^4+(74-52.17)^4+(28-52.17)^4+(28-52.17)^4 \\ & +(28-52.17)^4+(69-52.17)^4+(43-52.17)^4+(63-52.17)^4]/12 = 79503.98\end{aligned}$$

$$k = \mu_4 / \sigma^4 - 3 = (79503.98 / 58197) - 3 = 1.366 - 3 = -1.6336$$

This result, being a negative number, means the data are not fat-tailed.

Supplementary Exercises

Multiple Choice

1. If the data is positively skewed, which of the following statements are true?
   a. The mean is greater than the mode.
   b. The mean is less than the mode.
   c. The mean and mode are the same.
   d. The mean, median, and mode are the same.
   e. The mode will be 0.

2. When comparing the geometric mean and arithmetic mean

   a. the geometric mean and the arithmetic mean will always give the same numeric value.
   b. the arithmetic mean will always be less than or equal to the geometric mean.
   c. the arithmetic mean will always be greater than or equal to the geometric mean.
   d. the mode will always equal the geometric mean.
   e. the mode will always equal the arithmetic mean.

3. The larger the variance is for a set of data, the

   a. more meaningful the mean is.
   b. less meaningful the mean is.
   c. variance plays no part in determining whether the mean is meaningful.
   d. smaller the standard deviation.
   e. smaller the coefficient of variation.

4. Kurtosis measures the

   a. center of the data.
   b. dispersion of the data.
   c. peakedness of the data.
   d. symmetry of the data.
   e. mode of the data.

5. The mode represents

   a. the most frequently repeating score.
   b. middle score.
   c. a geometric average.
   d. an arithmetic average.
   e. a combination of the geometric and arithmetic averages.

6. Which of the following statements is true?

   a. The mean, median, and mode will always be the same for a data set.
   b. The mean, median, and mode will never be the same for a set of data.
   c. The mean will always be greater than the median, but smaller than the mode.
   d. The mean can never be negative.
   e. The variance can never be negative.

7. Which of the following statements is true?

   a. The variance will always be larger than the mean absolute deviation.
   b. The variance will always be smaller than the mean absolute deviation.
   c. The variance and mean absolute deviation will be the same.
   d. The standard deviation can never be negative.
   e. The standard deviation can be negative.

8. Grouping data can

   a. enrich the meaning of the mean and standard deviation.
   b. simplify the computational process for the mean and standard deviation.
   c. change the interpretation of the mean and standard deviation after they are computed.
   d. complicate the computational process for the mean and standard deviation.
   e. not be used for large amounts of data.

9. The range represents the

   a. difference between the highest and lowest value.
   b. middle number.
   c. most frequently repeating number.
   d. highest number.
   e. lowest number.

10. A *z*-score will always have

    a. a mean of 1 and a standard deviation of 0.
    b. a mean of 1 and a standard deviation of 1.
    c. a mean of 0 and a standard deviation of 1.
    d. a mean of 0 and a standard deviation of 0.
    e. a mean of 0 and a standard deviation that is greater than 1.

11. When the variance for a set of data equals 0,

    a. half the numbers will lie above the mean and half below.
    b. the mean will always be 0.
    c. all numbers in the data set will be the same.
    d. the data will be widely dispersed around the mean.
    e. the standard deviation will be 1.

12. Suppose a teacher records the following scores for a test: 87 42 55 37 99 98 47. The median is

    a. 37
    b. 98
    c. 55
    d. 87
    e. 99

13. Suppose the weights of the linemen of the San Francisco 49ers are: 272, 291, 285, 272, 280, and 245. The mode is

    a. 272
    b. 291
    c. 285
    d. 280
    e. 245

14. Suppose a teacher finds that the midterm scores in her accounting class have a mean of 82 and a variance of 9. The standard deviation of the midterms is

    a. 82
    b. 9
    c. 2
    d. 9
    e. 9^2

True/False (If False, Explain Why)

1. Grouping data changes the interpretation of the mean.
2. The standard deviation is a more popular measure of dispersion than the variance because it is in the same units as the mean.
3. The coefficient of variation should not be used to compare the dispersion of different sets of data that are measured in different units.
4. Kurtosis measures the symmetry of the distribution.
5. Positive skewness of a stock's returns is desirable.
6. If the mean, median, and mode all equal 10, the data is positively skewed.
7. The larger the variance, the more meaningful the mean.
8. Skewness measures the peakedness of the distribution.
9. The variance can never be negative.
10. The standard deviation can be negative.
11. The coefficient of variation can never be negative.
12. The coefficient of variation is always positive.

Questions and Problems

1. You are given the following grades from a midterm exam in English: 95, 34, 83, 92, 94, 88, 99.

   Find the median and range of the scores.

2. You are given the following two groups of numbers
   1,000 2,000 3,000
   and 1,000,000 2,000,000 3,000,000
   Is it true that the second group of numbers has a higher dispersion because it has a higher standard deviation?

3. Compute the population mean and variance for the following numbers: 92, 94, 86, 42, 38, 99.

4. Use the data from Exercise 3 to compute the skewness coefficient. Are the data skewed?
5. When examining the desirability of a business venture, we sometimes use the variance of the profits to measure the risk of the project. Briefly explain why the variance may not be a good measure of business risk.
6. Find the median, mode, and standard deviation for the following observations:

   32, 54, 88, 27, 99, 13, 72, 88, 90, 100, 21, 225

7. Compute the skewness coefficient for the data in Exercise 6.
8. Fill in the missing values in the table and find the mean for the frequency distribution.

Class	Midpoint	Frequency	
	m_i	f_i	$f_i m_i$
1–5	2.5	6	
6–10	7.5	3	
11–15	12.5	7	
16–20	17.5	2	
21–25	22.5	12	

9. Below are the dollar amounts spent on beer each month by 28 students at McBud University: $2, 7, 25, 18, 19, 21, 38, 30, 27, 19, 22, 31, 35, 29, 33, 22, 17, 18, 39, 33, 21, 3, 6, 24, 33, 28, 9, 17.

   a. Group the data, using groups of 0-9, 10-19, 20-29, and 30-40.
   b. Use a stem-and-leaf graph to show the data.
   c. Compute the mean and variance using the grouped data.

10. The scores of an entrance examination of the applicants are shown below.

24	59	74	79	85	98
36	64	74	79	89	99
49	70	75	81	90	99
52	71	76	83	92	100
56	73	77	84	96	100

   a. What is the mean of the scores?
   b. What is the coefficient of variation of the scores?
   c. What is the median of the scores?
   d. What score is the cutoff for the 90th percentile?

Answers to Supplementary Exercises

Multiple Choice

1. b	6. e	11. c
2. c	7. d	12. c
3. b	8. b	13. a
4. c	9. a	14. d
5. a	10. c	

True/False

1. False. Grouping data does not change the interpretation of the mean, it only simplifies the computation.
2. True
3. False. The coefficient of variation should be used to compare different data sets because it is unit free.
4. False. Kurtosis measures the peakedness of the distribution.
5. True
6. False. The data are symmetric.
7. False. A large variance makes the mean less meaningful.
8. False. Skewness measures symmetry.
9. True
10. False. The standard deviation can never be negative.
11. False. If the mean is negative, the coefficient of variation will be negative.
12. False. See answer to 11.

Questions and Problems

1. Median = 92
   Range = 65
2. No. Because the two sets of data are in different units, it is not appropriate to compare their standard deviations. If we were to use the coefficient of variation as our measure of dispersion, the dispersion would be the same for both sets of data.
3. $\sigma^2 = 634.139$
4. CS = –.6426
5. When we use variance as a measure of risk, we consider values above the mean to be as undesirable as values below the mean. Because profits that lie above the mean are desirable, rather than risky, the variance may not be an appropriate measure of business risk.

6. Median = 80
   Mode = 88
   Standard deviation = 54.48

7. 31.69

Classes	Midpoints	Frequency	
	m_i	f_i	$f_i m_i$
1–5	2.5	6	15
6–10	7.5	3	22.5
11–15	12.5	7	87.5
16–20	17.5	2	35
21–25	22.5	12	270
	____	____	
		30	430

Mean = 430 / 30 = 14.3

8. a.

	Class		Frequency	Midpoint	
			f_i	m_i	$f_i m_i$
	0–9		5	4.5	22.5
	10–19		6	14.5	87.0
	20–29		9	24.5	220.5
	30–39		8	34.5	276.0

b. Stem-and-leaf table

0	2 3 6 7 9
10	7 7 8 8 9 9
20	1 1 2 2 4 5 7 8 9
30	0 1 3 3 3 5 8 9

c. Mean = 606 / 28 = 21.64

Variance = 113.26

9. a. 76.13

   b. 0.25
   c. 78
   d. 99

Chapter 5
Probability Concepts and Their Analysis

Chapter Intuition

Probability is an approach for quantifying uncertainty. For example, suppose we have a bag filled with five balls, three white and two green. If you are asked what are the odds of picking a green ball out of the bag? A straightforward answer is dividing the number of favorable outcomes (green balls) by the total possible outcomes (five possible balls to select). Although it is quite obvious, the counting method is the basic building block for probability and statistics.

Using this same counting method, we can consider a more difficult problem. Suppose you are now interested in knowing the probability of selecting a green ball on your second draw given that the first ball you selected was white and it was not placed back in the bag. Again, your intuition would tell you that there are now four balls in the bag and two are white, so the chance of selecting a white ball is 50 %. What you have just done is to compute a conditional probability, which is the probability of the occurrence of event A (choosing a green ball on the second draw) given the occurrence of event B (the first ball drawn was white).

Although these examples may be obvious, more complicated problems will require a technique to count the possible outcomes and the corresponding probability. The rules of counting will depend on whether the events are dependent or independent. There are different ways to compute probability: (1) the empirical method, (2) the historical method, and (3) the subjective method. This chapter relies on methods of counting and on the established rules of probability. Chapter 6 will show you how to use established probability distributions.

Chapter Review

1. Probabilities can take on values between 0 (no chance of the event occurring) to 1 (event occurs with certainty). In ***classical probability***, we look at how many favorable outcomes can occur as a percentage of the total number of possible

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics*,
DOI 10.1007/978-3-319-11997-7_5

outcomes in an experiment. For example, the probability of tossing a coin and receiving a head is 50% since there is a total of only two possible outcomes, a head or a tail.

2. A ***union of two events*** occurs when either Event A *or* B occurs. An ***intersection of two events*** occurs when both Events A *and* B occur. For example, if we choose one card from a deck of cards, we might be interested in the chance of receiving a face card (Event A) and/or the chance of receiving a club (Event B). The union of Events A and B would deal with the chance of drawing a face card *or* a club, while the intersection of Events A and B would deal with the chance of drawing a face card that was also a club.
3. ***Joint probability*** is the probability that two or more events will occur together. For example, the probability that a Republican wins the presidential election and the Republicans gain control of the Senate is a joint probability.
4. ***Marginal probability*** represents the probability of an individual event, regardless of the occurrence of other events. From our previous example, we might be interested only in the probability that a Republican wins the presidential election, regardless of who has control of the Senate.
5. ***Conditional probability*** is the probability of the occurrence of Event B, given that Event A occurs. For example, we might be interested in the probability of Horse A winning a race given a muddy track.
6. ***Bayes' Theorem*** allows us to update probabilities by using new information.
7. The concepts of ***permutations*** and ***combinations*** provide important tools for computing probabilities. A popular example of probability based on combinations is the birthday problem, where we are interested in the probability that at least two people in a given room have the same birthday. As we increase the number of people in the room, the number of possible combinations increases. With only two people in a room, there is only one chance for a match. With three people in a room, there are now three possible matches: Person A with B; A with C; and B with C. With four people in a room, there are six possible matches and so on. With only 50 people in a room, there is a 97% chance that at least two people have the same birthday. This result comes from the large number of combinations of people who could have the same birthday in a room of 50 people.

Useful Formulas

Probability of Event A:

$$P_r(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$$

Union of Events A and B:

$$P_r\left(A \cup B\right) = P_r(A) + P_r(B) - P_r\left(A \cap B\right)$$

Intersection of Events A and B:

$$P_r\left(A \cap B\right) = P_r(A) + P_r(B) - P_r(A \cup B)$$

Probability of complements:

$$P_r(A \cup \bar{A}) = P_r(A) + P_r(\bar{A}) = 1$$

Conditional probability:

$$P_r(A|B) = \frac{P_r(A|B)}{P_r(B)}$$

Bayes' Theorem:

$$P_r(A|B) = \frac{P_r(B|A)P_r(A)}{P_r(B)}$$

Number of permutations of n-things taken r at a time

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

Number of combinations r objects can be selected from n

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

Example Problems

Example 1 Independent and Mutually Exclusive Events
In each situation below, decide whether the two events are independent, mutually exclusive, or neither.

a. The Detroit Lions win the Super Bowl in 2013 and the Seattle Seahawks win the Super Bowl in 2013.
b. The Boston Celtics win the NBA Championship and the Celtics sell more than 1 million tickets in the same season.
c. A person wins both the New York and New Jersey lotteries.
d. You get a head in one flip of a coin and get a tail in another flip of a coin.

Solution: Events are mutually exclusive if, when one event occurs, the other event cannot occur. Events are independent if the chance of one event occurring is not affected by the other event.

a. Since both the Lions and the Seahawks cannot win the Super Bowl in 1999, these events are mutually exclusive.
b. These two events are not mutually exclusive since both could occur. These events are also not independent, since a winning Celtics team will probably influence ticket sales positively.
c. The events are not mutually exclusive since a person could win both lotteries. The events are independent because the probability of winning one lottery does not affect the probability of winning the other.
d. The events are not mutually exclusive. The events are independent since the probability of flipping a tail is not affected by receiving a head in the first toss.

Example 2 Computing Probability with Replacement

A bag contains ten balls, three red, and seven white. If you draw four balls out of the bag one at a time with replacement, what is the probability that you receive a red ball first, a white ball second, and a red ball third?

Solution: The probability of drawing a red ball first is 3 out of 10 or .3; the probability of drawing a white ball is 7 out of 10 or .7, as the three draws are independent, the probability of red–white–red is

$$.30 \times .70 \times .30 = .063$$

Example 3 Computing Probability Without Replacement

Let's reconsider Example 2, except this time the balls are removed and not placed back in the bag. What is the probability that you receive a red ball first, a white ball second, and a red ball third?

Solution: When the balls are not replaced in the bag, we can see that the probability of drawing a white ball second is conditional on drawing a red ball first. The probability of drawing a red ball first is 3 out of 10 or .30; the probability of drawing a white ball second, given that a red ball has already been drawn is 7 out of 9; the probability of drawing a red ball on the third draw, given that a red ball and a white ball have already been drawn, is 2 out of 8. So the probability that this entire sequence occurs is

$$\begin{aligned} P_r(R_1 \cap W_2 \cap R_3) &= P_r(R_1)\, P_r(W_2 \mid R_1)\, P_r(R_3 \mid R_1 \cap W_2) \\ &= (3/10) \times (7/9) \times (2/8) = 0.058 \end{aligned}$$

Example 4 Conditional Probability

$P_r(E1) = 0.5,\ P_r(E2) = 0.4.$
Obtain $P_r(E1 \cup E2),\ P_r(E1 \mid E2),$ and $P_r(E2 \mid E1)$ given that $P_r(E1 \cap E2) = 0.2.$

Solution:

$$P_r(E1 \cup E2) = P_r(E1) + P_r(E2) - P_r(E1 \cap E2) = .5 + .4 - .2 = .7$$

$$P_r(E1 \mid E2) = P_r(E1 \cap E2) / P_r(E2) = .2 / .4 = .50 \text{ or } 50\%$$

$$P_r(E2 \mid E1) = P_r(E1 \cap E2) / P_r(E1) = .2 / .5 = .40 \text{ or } 40\%$$

Example 5 Permutations

Suppose you are at the racetrack and would like to play the exacta, in which you try to select the first and second place horse in correct order. If there are eight horses in the field, what is the probability of winning the bet?

Solution: Once we have selected the two horses to finish first and second, we can compute the number of ways the other six horses can finish, which is 6!. Since there are 8! ways that these eight horses can finish, we know 6!/8! = 1/56 is the answer.

Example 6 Conditional Probability

Consider Example 5 again, except compute the probability of winning using conditional probability.

Solution: The probability that horse A finishes first is 1 out of 8. The probability that horse B finishes second given that horse A finishes first is 1 out of 7. Therefore, the probability that we win our bet is

$$P_r\left(A_1 \cap B_2\right) = P_r\left(A_1\right)\ P_r\left(B_2 \mid A_1\right) = 1/8 \times 1/7 = 1/56$$

Example 7 Classical Probability

If you roll a die twice, what is the probability that the sum of the rolls equals 5?

Solution: Each roll has a 1/6 probability and the two rolls are independent, so the probability of rolling a certain combination is 1/36 (i.e., $1/6 \times 1/6$). In the first case, the possible combinations that will sum to 5 are: 1,4; 2,3; 3,2; and 4,1. So the probability of two rolls summing to five is just the sum of the probabilities that will add to five:

$$P_r(x = 5) = 1/36 + 1/36 + 1/36 + 1/36 = 1/9$$

Example 8 Union and Intersection of Events

Consider the roll of a six-sided die. Given the following Events *A* and *B*, give the union and the intersection for *A* and *B* in each case.

a. A= {2, 4, 6} B= {1, 3, 5}
b. A= {2, 4} B= {1, 2, 4}
c. A= {1, 2, 3} B= {1, 2, 3}
d. A= {2, 3, 4, 5} B= {1, 4, 5, 6}

Solution: The union of A and B represents all elements that are in A or B. The intersection of A and B represents all elements that are in both A and B.

a. $A \cup B = \{1, 2, 3, 4, 5, 6\}$ $A \cap B = \{\ \}$
b. $A \cup B = \{1, 3, 5\}$ $A \cap B = \{2, 4\}$
c. $A \cup B = \{1, 2, 3\}$ $A \cap B = \{1, 2, 3\}$
d. $A \cup B = \{1, 2, 3, 4, 5, 6\}$ $A \cap B = \{4, 5\}$

Example 9 Combinations

Mary Gop, political candidate for governor, has decided to campaign in four of the seven largest cities in the state. How many different combinations of four cities can she stop in?

Solution:
$$_nC_r = n!/[r!(n-r)!]$$
$$= 7!/[4!(7-4)!] = 35$$

Example 10 Marginal and Joint Probability

Suppose 200 students were randomly asked whether they own a stereo and/or a personal computer. The results are given in the following table.

	PC	No	PC
	(A)	($\bar{A}$)	
(B) Stereo	40	60	100
($\bar{B}$) No stereo	20	80	100
	60	140	200

Obtain $P_r(A \mid B)$, $P_r(A \cup B)$, and $P_r(B)$, Also, show that $P_r(A \mid B) = P_r(A \cap B)/P_r(B)$

Solution:

$$P_r(A|B) = \frac{P_r(AIB)}{P_r(B)} = \frac{40/200}{100/200} = .4$$

$$P_r(AIB) = \frac{40}{200} = .2$$

Example 11 Bayes' Theorem

A factory procures headlamps from two suppliers. Supplier A provides 40% of the headlamps and Supplier B provides 60%. Suppose 10% of the headlamps delivered by Supplier A and 20% of the headlamps delivered by Supplier B are defective. If a bad headlamp is found, what is the probability that it comes from Supplier B?

Solution: If we select a headlamp randomly from the factory, there is a 40% chance that it comes from Supplier A and a 60% chance that it comes from Supplier B. Therefore, the prior probabilities are

$$P_r(A) = 40\% P_r(B) = 60\%$$

Once we have selected a defective headlamp, we now have additional information that we can use. We know that with a headlamp from Supplier A, there is a 10% chance

that it is defective, $P_r(\text{bad}|A) = 10\%$. Similarly, we know $P_r(\text{defective}|B) = 20\%$. Using this information we can solve for

$$P_r(B \mid bad) = \frac{P_r(B \cap bad)}{P_r(bad)} = \frac{P_r(B \mid bad)P_r(B)}{P_r(bad \mid B)P_r(B) + P_r(bad \mid A)P_r(A)}$$
$$= \frac{.20 \times .60}{.20 \times .60 + .10 \times .40} = \frac{.12}{.12 + .04} = \frac{.12}{.16} = \frac{3}{4}$$

Example 12 Bayes' Theorem

In Example 11, why is the posterior distribution $P_r(B \mid bad)$ higher than $P_r(B)$?

Solution: Before we know the headlamp is bad, we know the probability that a headlamp comes from Supplier B is 60%. Because Supplier B has a higher failure rate than Supplier A, this increases still further the probability that a headlamp from Supplier B is bad.

Supplementary Exercises

Multiple Choice

1. The union of two events occurs when
   a. Neither Event A nor Event B occurs
   b. Both Event A and Event B occur
   c. Either Event A or Event B occurs
   d. Event A and Event B are mutually exclusive
   e. Event A and Event B are independent
2. The intersection of two events occurs when
   a. Neither Event A nor Event B occurs
   b. Both Event A and Event B occur
   c. Either Event A or Event B occurs
   d. Event A and Event B are mutually exclusive
   e. Event A and Event B are independent
3. If Events A and B are independent events
   a. They occur simultaneously
   b. Event B is not conditional on Event A occurring
   c. They are not influenced by one another
   d. They cannot occur simultaneously
   e. They are mutually exclusive

4. If Events A and B are mutually exclusive

   a. They occur simultaneously
   b. They cannot occur simultaneously
   c. They are not influenced by one another
   d. They are independent
   e. Event B is conditional on Event A occurring

5. Bayes' Theorem

   a. Gives the probability of two events occurring jointly
   b. Gives the marginal probability of x
   c. Is a method for updating probabilities by using new information
   d. Is a method for constructing Venn diagrams
   e. Is a method for computing probabilities for mutually exclusive events

6. The conditional probability of x given y is

   a. The probability that x and y occur jointly
   b. The probability that y occurs if x has already occurred
   c. The probability that x occurs if y has already occurred
   d. The marginal probability of x minus the marginal probability of y
   e. The same as a joint probability

7. The probability of tossing three heads in a row is

   a. 1/2
   b. 1/4
   c. 1/8
   d. 1/6
   e. 1/16

8. The probability of receiving one head and one tail in two flips of a coin is

   a. The same as the probability of tossing two heads in a row
   b. Is less than the probability of tossing two heads in a row
   c. Is greater than the probability of tossing two heads in a row
   d. Is 0
   e. Is 1/2

9. The multiplication rule in probability is only appropriate when the events

   a. Are independent
   b. Are mutually exclusive
   c. Are dependent
   d. Occur jointly
   e. Are conditional upon one another

10. If you roll two dice, the probability that the sum of the two rolls equals 7 is

    a. 1/36
    b. 2/36
    c. 1/6
    d. 3/36
    e. 1/4

11. If you roll two dice, the probability that the sum of the two rolls equals 12 is

    a. 1/36
    b. 2/36
    c. 1/6
    d. 3/36
    e. 1/4

12. Suppose the probability that you will be elected president of the student body is .30, and the probability that you will get into Harvard Law School is .05. If the two events are independent, the probability that you will not only become student body president but also get into Harvard Law School is

    a. .35
    b. .25
    c. .015
    d. .50
    e. .05

13. Suppose a bag is filled with eight balls, three white and five black. If balls are drawn from the bag *without* replacement, the probability that the second ball drawn is black, given that the first ball drawn is black, is

    a. 1/8
    b. 5/8
    c. 4/7
    d. 2/8
    e. 1/5

14. Suppose a bag is filled with eight balls, three white and five black. If balls are drawn from the bag *with* replacement, the probability that the second ball drawn is black is

    a. The same as the probability that the first ball drawn is black
    b. Greater than the probability that the second ball drawn is black
    c. Less than the probability that the second ball drawn is black
    d. Less than the probability that the second ball is white
    e. Less than the probability that the first ball is white

15. Suppose you draw 1 card from a deck of 52 cards and flip a coin once. The probability that you draw a heart *and* toss a tail is

    a. $1/4+1/2$
    b. $1/4\times1/2$
    c. $1/52\times1/2$
    d. $1/52+1/2$
    e. $1/4-1/2$

16. Suppose a die is rolled once, and we do not know the outcome. The sample space S is:

    a. $S=\{1, 2, 3, 4, 5, 6\}$
    b. $S=\{2, 4, 6\}$
    c. $S=\{1, 3, 5\}$
    d. None of the above

17. Define the following events: $A=\{1, 2, 4\}$, $B=\{2, 3, 4\}$, $C=\{3, 4, 6\}$. A single die is rolled. We are told that Events A, B, and C occur. What is the outcome of the roll?

    a. 1
    b. 3
    c. 4
    d. 7
    e. None of the above

18. A single die is rolled. If either $A=\{1, 2, 4\}$ or $B=\{2, 3, 4\}$ occurs, which of the following outcome is not possible?

    a. 1
    b. 3
    c. 4
    d. 7
    we. None of the above

True/false (If false, explain why)

1. $P_r(A)P_r(B)=P_r(A\cap B)$
2. $P_r(A)+P_r(B)=P_r(A\cap B)$
3. $P_r(A)+P_r(B)=P_r(A\cup B)$
4. $P_r(A\mid B)=P_r(A)$
5. $P_r(A\cap B)=P_r(B\cup A)$
6. $P_r(B\cup A)\geq P_r(A)+P_r(B)$
7. $P_r(B\cap A)\leq P_r(A)+P_r(B)$

8. Because there is a 50% probability of tossing a head and a 50% probability of tossing a tail, there is a greater probability of tossing a head on the first toss and a tail on the second toss than there is of tossing two heads in a row.
9. Bayes' Theorem is a method that allows us to update probabilities using new information.
10. The probability that two mutually exclusive events occur simultaneously is 1.
11. The event of an experiment is the set of all possible outcomes of that experiment: (a) true (b) false.
12. The sample space of an experiment is a subset of the event: (a) true (b) false.
13. Let A and B be two events. The event that either A or B occurs is denoted as $A \cup B$: (a) true (b) false.
14. Let A and B be two events. The event that both A and B occur is denoted as $A \cup B$: (a) true (b) false.

Questions and Problems

1. Suppose you toss a coin twice. What is the probability of tossing two heads?
2. Suppose you toss a coin twice. What is the probability of tossing two heads, given that your first toss was a head?
3. Suppose you are on vacation in Europe and would like to visit 11 cities. Unfortunately, you only have time to visit five of these cities. How many different combinations of five cities can you visit?
4. Suppose you draw four cards from a standard 52-card deck. What is the probability that the first card will be a club, the second a diamond, the third a heart and the fourth a spade if the cards are drawn with replacement?
5. How would your answer to Exercise 4 change if the cards were drawn without replacement?
6. Suppose an auto dealership is interested in the types of options people purchase on cars. The records of 125 people who purchased cars were examined to see which ones purchased air conditioning and which ones purchased automatic transmission. The results are given in the following table.

	A/C	No A/C	
	(A)	(A)	
(B) Automatic	63	10	73
(B) Not automatic	25	27	52
	88	37	125

Find $P_r(B \mid A)$, $P_r(A)$, $P_r(B)$, and $P_r(A \cap B)$

7. A senate committee consists of 12 Democrats and 9 Republicans. In how many ways can a subcommittee consisting of eight members, five Democrats, and three Republicans be formed?
8. All students of a university are assigned ID numbers. The ID number consists of the first three letters of a student's last name, followed by four numbers. How many possible different ID numbers are there?

9. Suppose there are four events A, B, C, and D. The following information is given.

$P(A)=.5$	$P(A \cup D)=.72$
$P(B)=.15$	$P(A\|B)=.25$
$P(C)=.20$	$P(A \cap C)=.04$
	$P(A \cap D)=.03$

   a. Compute P(D)
   b. Compute P(A|D)
   c. Compute P(A∩B)
   d. Compute P(A∪B)
   e. Are A and B mutually exclusive? Explain your answer.
   f. Are A and B independent? Explain your answer.

10. Assume you have applied to two different universities A and B. In the past, 30% of students who applied to University A were accepted, while University B accepted 45% of the applicants. Assume events are independent of each other.

   a. What is the probability that you will be accepted in both universities?
   b. What is the probability that you will be accepted to at least one graduate program?
   c. What is the probability that one and only one of the universities will accept you?
   d. What is the probability that neither university will accept you?

11. Suppose 20% of the employees of company ABC have only a high school diploma; 60% have bachelor degrees; and 20% have graduate degrees. Of those with only a high school diploma, 15% hold management positions; whereas, of those having bachelor degrees, 30% hold management positions. Finally, 60% of the employees who have graduate degrees hold management positions.

   a. What percentage of employees hold management positions?
   b. Given that a person holds a management position, what is the probability that she/he has a graduate degree?

Answers to Supplementary Exercises

Multiple Choice

1 c	6. c	11. a	16.d
2 b	7. c	12. c	17. b
3 c	8. c	13. c	18. e
4 b	9. a	14. a	
5 c	10. c	15. b	

True/False

1. False. True if A and B are independent
2. False. True if A and B are mutually exclusive
3. False. True if A and B are mutually exclusive
4. False. True if A is independent of B
5. False. True if $P_r(A) = P_r(B) = 0$
6. True
7. True
8. False. The probabilities are the same
9. True
10. False. The probability that two mutually exclusive events occur simultaneously is 0
11. True
12. False
13. True
14. True

Questions and Problems

1. 1/4
2. 1/2
3. 462
4. 1/256
5. $1/4 \times 13/51 \times 13/50 \times 13/49$
6. $P_r(B \mid A) = .716$

   $P_r(A) = 88/125$

   $P_r(A \cap B) = 63/125$
7. 41,580
8. $C_3^{26} C_4^{10} = 546000$
9. a. 0.25
   b. 0.12
   c. 0.0375
   d. 0.6125
   e. Since, $P(A \cap B) \neq 0$, therefore they are not mutually exclusive.
   f. Since $P(A \cap B) = 0.0375 \neq P(A)P(B)$, thus they are not independent
10. a. $P(A \cap B) = 0.3 \times 0.45 = 0.135$
    b. $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.45 - 0.135 = 0.615 \leq$
    c. $P(A \cap B^c) + P(A^c \cap B) = 0.48$
    d. $1 - P(A \cup B) = 0.385$

11. a. Define H = {high school degree}, B = {bachelor degree}, G = {graduate degree}; M = {management position}

$$P(M) = P(M \mid H) \times P(H) + P(M \mid B) \times P(B) + P(M \mid G) \times P(G)$$
$$= 0.15 \times 0.2 + 0.30 \times 0.6 + 0.6 \times 0.2 = 0.33$$

b. $P(G \mid M) = 0.6 \times 0.2 / 0.33 = 0.3636$

Chapter 6
Discrete Random Variables and Probability Distributions

Chapter Intuition

In Chap. 5, the concept of probability, the chance of uncertain outcomes was introduced systematically. In this chapter, you will see that different ways in which data are generated results in different probability distributions. For example, if we flip a coin 100 times, a series of outcomes heads (H) or tails (T) are generated. If we are interested in the number X of H, the distribution of interest is the binomial distribution.

In addition to the binomial distribution, this chapter also introduces other distributions including Poisson distribution and hypergeometric distribution. General formulas for the expected value and variance of a random variable are given. The covariance and correlation between two random variables are also discussed.

Chapter Review

This chapter introduces several important discrete distributions and shows how these distributions can be used to solve a wide variety of problems in business and economics.

1. A ***random variable*** is a variable that can take on a certain number of numerical outcomes. Random variables can be either ***discrete*** or ***continuous***. A discrete random variable is the one that can take on only a limited number of values. A continuous random variable is the one that can take on any possible value within an interval. This chapter deals with discrete random variables, while Chaps. 7 and 8 deal with continuous random variables.
2. The ***binomial distribution*** allows us to calculate the probability of X successes out of n independent trials when the probability, p, of success in each trial remains the same. Examples of binomial distributions include the number of heads in ten flips of a coin or the number of sixes in 100 rolls of a die.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_6

3. The ***Poisson distribution*** is the limiting distribution of a binomial distribution B(n, p), where the probability p of success in each trial is very small and the number of trials n is large, such that $np = \lambda$. Thus, $\mathrm{E}(X) = \lambda$ and $\mathrm{Var}(X) = \lambda$. A Poisson random variable X is often used to represent the number of occurrences in an interval or within a given time period.
4. The ***hypergeometric distribution*** is used to represent the number of success in a set of random samples sampled *without replacement*.
5. ***Joint distributions*** represent the likelihood of the occurrence of two or more events simultaneously. ***Marginal probabilities*** deal with the chance of event A occurring regardless of the results of event B. ***Conditional probabilities*** measure the chance of event B occurring, given that event A has already occurred.
6. ***Expected values*** deal with the central tendency of the distribution. ***Variance*** and ***standard deviation*** deal with the dispersion of the distribution around its expected value. ***Covariance*** and ***correlation*** measure the degree of association between two random variables.

Useful Formulas–

Binomial distribution $X \sim \mathrm{B}(n, p)$ $P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$ $\mu_x = np$ $\sigma_x^2 = np(1-p)$	Covariance for two discrete random variables X and Y: $Cov(X,Y) \equiv \sigma_{XY}$ $= \sum (x_i - \mu_x)(y_i - \mu_y) P(X = x_i, Y = y_i)$
Poisson distribution: $P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}$ $\mu_x = \lambda$ $\sigma_x = \sqrt{\lambda}$	Correlation coefficient for two discrete random variables: $\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$
Hypergeometric distribution: $P(X = x) = \frac{C_x^h C_{n-x}^{N-h}}{C_n^N}$ $\mu_x = \left[\frac{h}{N}\right]$ $\sigma_x^2 = n\left(\frac{N-n}{N-1}\right)\left(\frac{h}{N}\right)\left(1-\frac{h}{N}\right)$	Conditional probability: $P(X = x_i \mid Y = y_i) = \frac{P(X = x_i, Y = y_i)}{P(Y = y_i)}$

Expected value for a discrete random variable: $\mu_x = E(X) = \sum_{i=1}^{N} x_i P(x_i)$	Marginal probability of X: $P(X = x_i) = \sum_{j=1}^{m} P(X = x_i, Y = y_j)$
Variance for a discrete random variable: $\sigma^2 = E(x_i - \mu_x)^2 \sum (x_i - \mu_x)^2 P(x_i)$	Marginal probability of Y: $P(Y = y_j) = \sum_{i=1}^{n} P(X = x_i, Y = y_j)$

Example Problems

Example 1 Binomial Distribution

The Kansas City Royals and the Washington Nationals enter the World Series. The Royals are considered to have a 70 % chance of winning, and the Nationals have only a 30 % chance of winning. The team that wins at least four out of seven games will win the series. Find the probability that the Nationals will win the championship.

Solution: Let X be the number of games the Nationals wins, then Nationals will win the championship if X=4, 5, 6, or 7, thus

$$P(X \geq 4 \mid n = 7, p = .30)$$

$$= {}_7C_4 .3^4 (1-.3)^3 + {}_7C_5 .3^5 (1-.3)^2 + {}_7C_6 .3^6 (1-.3)^1 + {}_7C_7 .3^7 (1-.3)^0$$

$$= 0.1260$$

Example 2 Cumulative Distribution of a Binomial Distribution

A production process produces 5 % defective parts. A sample of 20 parts from the production is selected. What is the probability that the sample contains exactly two defective parts? What is the probability that the sample contains more than two defective parts?

Solution: Let X be the number of defective parts, then

$$P(X = 2) = {}_{20}C_2 .05^2 (1-.05)^{18} = 0.1887$$

$$P(X > 2) = 1 - P(X \leq 2) = 1 - \ 0.9245 \ = \ 0.0755$$

Example 3 Poisson Distribution

Suppose the average number of customers at a gas station in any 60 min period is equal to 12. What is the probability that there will be exactly four customers in any 30 min period?

Solution: Because the average of 60 min is 12, thus the average of 30 min is 6.

$$P(X = 4) = [e^{-6} 6^4] / 4! = .1338$$

Example 4 Poisson Distribution—Cumulative Probability

Using the information from Example 3, find the probability that there are more than four customers in any 30 min period.

Solution: The probability that there are more than four customers is the complement of four customers or fewer. Therefore,

$$P_r(X > 4) = 1 - P(X \leq 4)$$

$$= 1 - \left[P(X = 0) + P(X = 1) + \ldots + \ P(X = 4)\right] = .806 \text{ or } 80.6\%$$

Example 5 Mean and Variance for the Sum of Rolling Dices

Write out the probability distribution for the sum of rolling two dices. Compute the mean and variance for the distribution.

Solution:

Sum S_i	Frequency	Probability P_i	S_iP_i
2	1	1/36	2/36
3	2	1/18	3/18
4	3	1/12	4/12
5	4	1/9	5/9
6	5	5/36	30/36
7	6	1/6	7/6

Sum S_i	Frequency	Probability P_i	S_iP_i
8	5	5/36	40/36
9	4	1/9	9/9
10	3	1/12	10/12
11	2	1/18	11/18
12	1	1/36	12/36
			252/36

Total 36

$$\text{Mean} = \mu_s = P_i S_i = 252/36 = 7$$

$$\text{Var} = \sigma^2 = \sum P_i (S_i - \mu_s)^2 = 38.69$$

Example 6 Mean and Variance for Binomial Distribution

Suppose the probability of receiving a response when you send out a resume is .23. If you send out 125 resumes, find the mean and variance for the distribution.

Solution: $\mu = np = .23 \times 125 = 28.75$

$$\sigma^2 = np(1-p) = 125 \times .23 \times (1 - .23) = 22.14$$

Example 7 Mean and Variance for Poisson Distribution

Suppose an average of five orders are placed every 30 min at a mail order computer store. Find the mean and variance of the distribution.

Solution: $\mu = \lambda = 5$ and $\sigma^2 = \lambda = 5$

Example 8 Hypergeometric Distribution

Suppose we have a bag with five balls, three red and two white. If you draw three balls out of the bag without replacement, what is the probability that you receive two red balls and a white ball? What is the probability if there is replacement?

Solution: Without replacement, we have the hypergeometric distribution.

$$\frac{C_x^h C_{n-x}^{N-h}}{C_x^N} = \frac{C_2^3 C_{3-2}^{5-3}}{C_3^5} = \frac{\frac{(3)(2)(1)}{(2)(1)(1)} \times \frac{(2)(1)}{(1)(1)}}{\frac{(5)(4)(3)(2)(1)}{(3)(2)(1)(2)(1)}} = .6$$

With replacement, we have the binomial distribution with a 3/5 chance of drawing a red ball. So if we draw three balls, the probability of getting two red ones is

$$P_r(x = 2) = \binom{3}{2}\left(\frac{3}{5}\right)^2\left(1-\frac{3}{5}\right)^{3-2} = .432$$

Example 9 Expected Value of a Bernoulli Distribution

A fire insurance policy will pay you $ 150,000 if your house burns down. If there is a .00025 chance that the house will burn down, what is the expected value of the policy?

Solution: Assume X = the value of the policy, then X will equal 0 or $ 150,000 depending on whether the house burns down or not.

X	P	PX
0	0.99975	0
150,000	0.00025	50

$$E(X) = \sum PX = 0(0.99975) + 150,000(0.00025) = 37.5$$

Example 10 Poisson Approximation to a Binomial Distribution

In the production process, some defective parts are unavoidable. A statistician monitors the process and finds a defect rate of .0001. If 200 parts are randomly selected, what is the probability that there will be more than two defective items?

Solution: This is a binomial distribution question because we are interested in the probability of obtaining more than two defective items from the 200 items examined. However, using the binomial distribution formula will be very complicated. Fortunately, we can approximate the binomial distribution using the Poisson distribution. For $n = 200$ and $p = .001, \lambda = np = .2$.

$$P_r(x > 2) = 1 - P_r(x \le 2) = 1 - P_r(x = 0) - P_r(x = 1) - P_r(x = 2)$$
$$= 1 - \frac{e^{-2}.2^0}{0!} - \frac{e^{-2}.2^1}{1!} - \frac{e^{-2}.2^2}{2!} = .0012$$

Example 11 Expected Value and Standard Deviation of Two Stock Returns

The following table summarizes the returns of stocks A and B under two economic conditions, namely, high growth or recession.

Economy	r_A	r_B	P_r	$r_A \times r_B$
High growth	10%	14%	30%	1.4%
Recession	8%	−4%	70%	−.32%

There is a 40% chance of a recession and a 60% chance of high growth next year. Compute the expected rate of return for stocks A and B. Compute the standard deviation of the returns for stocks A and B.

Solution: $E(r_A) = 10\% \times 30\% + 8\% \times 70\% = 8.8\%$

$$E(r_B) = 14\% \times 30\% - 4\% \times 70\% = 14.0\%$$

$$\text{s.d.}(r_A) = \left[(10\% - 8.8\%)^2 \times 30\% + (8\% - 8.8\%)^2 \times 70\%\right]^{1/2} = .0938$$

$$\text{s.d.}(r_B) = \left[(14\% - 14.0\%)^2 \times 30\% + (-4\% - 14.0\%)^2 \times 70\%\right]^{1/2} = .1506$$

Example 12 Covariance and Correlation of Two Stock Returns

Use the information given in Example 11 to find the covariance and correlation between stocks A and B.

Solution: $\text{Cov}(r_A, r_B) = E\left[r_A - E(r_A)\right]\left[r_B - E(r_B)\right]$

$$= E(r_A r_B) - E(r_A)E(r_B)$$

$$= (1.4\% \times 30\% - .32\% \times 70\%) - 8.8\% \times 14.0\% = -0.01036$$

$$\text{Corr}(r_A r_B) = \text{Cov}(r_A, r_B) / \left[\text{s.d.}(r_A) \times \text{s.d.}(r_B)\right]$$

$$= -0.01036 / [(.0938)(.1506) = -0.733$$

Example 13 Portfolio Risk and Return

In finance, we are interested in reducing the risk (measured by standard deviation) of investments. One way to do this is to diversify our portfolio by holding more than one stock. Using the information from Examples 12 and 13, let us examine the expected value and standard deviation for a portfolio consisting of 50 % in stock A and 50 % in stock B.

Solution: $E(r_p) = E(r_A) \times 50\% + E(r_B) \times 50\%$

$$= 8.8\% \times 50\% + 10.4\% \times 50\% = 9.6\%$$

$$\mathrm{Var}(r_p) = 50\%^2 \mathrm{Var}(r_A) + 50\%^2 \mathrm{Var}(r_B) + 2\ 50\%(1 - 50\%)\ \mathrm{Cov}(r_A r_B)$$

$$= 50\%^2 \left[.000096 + .04084 + 2(.000768)\right] = .011$$

$$\text{s.d.}(r_p) = 1 = .103$$

Supplementary Exercises

Multiple Choice

1. A Bernoulli process
   a. gives the number of successes in *n* trials
   b. consists of one trial with only two possible outcomes
   c. is the expected value of a binomial distribution
   d. is a continuous distribution
   e. can be used for more than two possible outcomes
2. The hypergeometric distribution is
   a. a continuous distribution
   b. generated with a Bernoulli process with replacement
   c. generated with a Bernoulli process without replacement
   d. the same as the binomial distribution
   e. the expected value of the binomial distribution
3. A discrete random variable
   a. can take on an infinite number of values
   b. can take on a finite and countable number of values

c. is another name for a probability mass function
d. is always generated from a Bernoulli process
e. always has a mean of 0 and a variance of 1

4. The cumulative binomial function gives us the probability that we will have

a. more than x successes in n trials
b. at least x successes in n trials
c. exactly x successes in n trials
d. exactly n successes in n trials
e. no successes in n trials

5. The difference between the binomial distribution and the hypergeometric distribution is that

a. the binomial distribution assumes that the probability of success is constant, whereas the hypergeometric distribution assumes that it changes with each experiment
b. the binomial distribution assumes that the probability of success changes, whereas thev hypergeometric distribution assumes that it remains constant
c. the hypergeometric distribution is a continuous distribution, whereas the binomial distribution is a discrete distribution
d. the hypergeometric distribution is a discrete distribution, whereas the binomial distribution is a continuous distribution
e. there is no difference between the two distributions

6. A continuous random variable is one that can take on

a. only a countable number of values
b. an infinite number of values
c. both a countable and an infinite number of values
d. only integer values
e. only values that are negative

7. The Poisson distribution

a. can be approximated by the binomial distribution
b. allows us to compute the probability of x successes in n trials
c. allows us to compute the probability of x successes in a given time interval
d. is a continuous distribution
e. is another name for a Bernoulli process

8. In the Poisson distribution both the mean and variance of the distribution are

a. the number of successes, x, in n trials
b. the number of trials, n
c. the number of successes, x
d. equal to the average amount of success in a given time interval
e. equal to x! times the number of trials, n

9. Suppose the New York Yankees are playing the Los Angeles Dodgers in the World Series (winner is the team that wins four out of seven games). The probability that the Dodgers win three games or less

   a. can be determined by the binomial distribution
   b. is 1—P_r(Dodgers win four games)
   c. is equal to .5
   d. can be determined by using a Bernoulli process
   e. is 1—P_r(Yankees win four games)

10. The cumulative probability, F(y) is

   a. $P_r(x \geq y)$
   b. $P_r(x = y)$
   c. $P_r(x \leq y)$
   d. always equal to 1
   e. always equal to 0

11. The properties of the probability function for discrete random variables are

   a. $P_r(x_i) > 0$ for all I
   b. $P_r(x_i) \geq 0$ for all I
   c. $\sum P_r(x_i) = 1$
   d. Both a and c
   e. Both b and c

12. The expected value of a discrete random variable, x, can be computed as

   a. $\sum x_i P_r(x_i)$
   b. $\sum x P_r(x_i)$
   c. $\sum [x_i - E(x_i)]^2 P_r(x_i)$
   d. $\sum x_i$
   e. $\sum x$

13. The variance of a discrete random variable, x, can be computed as

   a. $\sum x_i P_r(x_i)$
   b. $\sum x P_r(x_i)$
   c. $\sum [x_i - E(x_i)]^2 P_r(x_i)$
   d. $\sum x_i$
   e. $\sum x$

14. The covariance between two random variables, x and y can be computed as
   a. $E[(x - E(x)]\,[y - E(y)]$
   b. $\sum [y_i - E(y_i)]^2 P_r(y_i)$
   c. $\sum [x_i - E(x_i)]^2 P_r(x_i)$
   d. $\sum x_i P_r(y_i)$
   e. $\sum x P_r(y_i)$

15. The coefficient of correlation between x and y will

    a. always be greater than 0
    b. always be greater than 1
    c. be between 0 and 1
    d. be between −1 and 1
    e. always be negative

Use the following information for 16–18. There are five red balls numbered 1, 4, 6, 8, and 9; and five blue balls numbered 5, 7, 11, 13, and 17.

16. If two balls are drawn one by one in random with replacement from the bag, what is the probability that the sum of the numbers is ten?

    a. 0.0833
    b. 0.1666
    c. 0.0500
    d. 0.1000

17. If two balls are drawn one by one in random from the bag without replacement, what is the probability that the sum of the numbers is ten?

    a. 0.0833
    b. 0.0444
    c. 0.0500
    d. 0.1000

18. If two balls are drawn together, what is the probability that the sum of the numbers is ten?

    a. 0.0833
    b. 0.0444
    c. 0.0500
    d. 0.1000

True/False (If False, Explain Why)

1. The binomial distribution and Poisson distribution are completely unrelated to each other.
2. When we use a binomial distribution to study the opinion of a population, we are assuming that the population is large.
3. The expected value of a random variable is the value we expect the random variable to realize.
4. When we are sampling with replacement, we should use the hypergeometric distribution.
5. When the population is very large and sampling is done without replacement, the binomial and hypergeometric distributions will give similar results.

6. The only way to determine the probability of tossing exactly four heads in ten flips of a coin is to use the binomial distribution.
7. If the covariance between x and y is ten and the covariance between c and d is 25, c and d are 2.5 times more strongly correlated than x and y.
8. If the correlation coefficient between x and y is 1, then when variable x goes up we expect variable y to go down.
9. The variance of a binomial distribution with probability of success, p, is equal to np.
10. A call option can be evaluated by using the binomial distribution.
11. The event of an experiment is the set of all possible outcomes of that experiment.

    a. True
    b. False

12. The sample space of an experiment is a subset of the event.

    a. True
    b. False

13. Let A and B be two events. The event that either A or B occurs is denoted $A \cup B$.

    a. True
    b. False

14. Let A and B be two events. The event that both A and B occur is denoted $A \cup B$.

    a. True
    b. False

Questions and Problems

1. Suppose a coin is tossed 15 times. What is the probability of tossing heads three times or less?
2. A bag contains ten numbers from 1 to 10. If you draw four numbers out of the bag with replacement, what is the probability that all four will be even numbers?
3. Redo Exercise 2 assuming that the four numbers are drawn without replacement.
4. Suppose a grocery store averages six customers per 5 min period. Find the probability that the store will have fewer than three customers in any given 5 min period.
5. You are given the following information about a stock:

$$S = \$50, X = \$55, r = .005, n = 6, u = 1.05, d = .90$$

   Compute the value of the call option.

6. Suppose a coin has a 75 % chance of turning up heads and a 25 % chance of turning up tails. Find the mean and variance for the number of heads tossed in 1000 flips of the coin.

7. Suppose you roll a die five times. What is the probability of receiving exactly one roll of six?
8. Suppose you play a game where a coin is tossed. If the toss is a head you receive $ 100; if the toss is a tail, you receive $ 125. Compute the expected value and standard deviation for this game.
9. You are given the following rates of return for two stocks in two different climates. Stock B is the stock for a beach resort; stock P is the stock for an indoor swimming pool. The possible climates and stock returns are given in the following table

Weather	r_B	r_p	P_r	$r_B r_p$
Sunny	25%	−5%	.30	−1.25%
Rainy	−5%	25%	.70	−1.25%

Compute the expected value and standard deviation for each stock.

10. Use the information given in Exercise 9 to compute the covariance and correlation coefficient between stocks B and P.
11. Use the information given in Exercise 9 and your results from Exercise 10 to find the expected value and standard deviation for a portfolio consisting 60% of stock B and 40% of stock P.
12. The average number of calls received by an operator in a 30-min period is 12. What is the probability that between 17:00 and 17:30 the operator will receive exactly eight calls? What is the probability that between 17:00 and 17:30 the operator will receive more than nine calls but fewer than fifteen calls?
13. In a lot of 200 parts, 50 of them are defective. Suppose a sample of ten parts is selected at random, what is the probability that two of them are defective? What is the expected number of defective parts? What is the standard deviation of the number of defective parts?

Answers to Supplementary Exercises

Multiple Choice

1. a	6. b	11. e	16. c
2. d	7. c	12. a	17. a
3. b	8. d	13. c	18. d
4. d	9. b	14. a	
5. a	10. c	15. d	

True/False

1. False. The Poisson distribution is a special case of the binomial distribution when a binomially distributed random variable is generated from a large number of experiments and the probability of a success in each experiment is very small. In this case, it is easier to use the Poisson distribution to approximate the binomial distribution.
2. True. The population has to be large so that the probability of a certain opinion is not changed from one experiment to another. Otherwise, the hypergeometric distribution is more appropriate.
3. False. The expected value of a random variable should be interpreted as the theoretical average of the random variable and should not be interpreted "literally" as what we expect the random variable to realize.
4. False. When we are sampling with replacement, we should use the binomial distribution. The hypergeometric distribution is used when we sample without replacement.
5. True
6. False. We do not need to use the binomial distribution to determine the probability of tossing exactly four heads in ten flips of the coin, we use the binomial distribution because it makes the computation much easier.
7. False. The covariance between two different pairs of variables cannot be used to compare the degree of correlation.
8. False. When the correlation coefficient between x and y is 1, then when x goes up, we know *y* will also go up.
9. False. The variance is equal to $np(1-p)$. The mean is equal to np.
10. True
11. True
12. False
13. True
14. True

Questions and Problems

1. .0176
2. .0625
3. .0238
4. .1512
5. $ 2.08
6. $\mu = 750 \quad \sigma^2 = 187.5$
7. .402
8. $\mu = \$112.50 \quad \sigma^2 = 156.25$

9. $E(r_B) = 4$ Standard Deviation $(r_B) = 27.61$
   $E(r_p) = 16$ Standard Deviation $(r_p) = 13.75$
10. $\text{Cov}(r_B r_p) = -65.25$

    Correlation coefficient $(r_B r_p) = -.172$
11. Variance (portfolio) $= 273.34$ and Standard deviation $= 12.88$
12. a. $P(X = 8) = 0.0655$

    b. $P(9 < X < 15) = P(X \le 14) - P(X \le 9) = 0.53$
13. $\dfrac{C_8^{150} C_2^{50}}{C_{10}^{200}} = 0.2869; \quad \mu = 10 \times \dfrac{50}{200} = 2.5;$

    $\sigma^2 = 10 \times \dfrac{50}{200} \times \dfrac{150}{200} \times \dfrac{200-10}{200-1} = 1.7902;$

    $\sigma = 1.3380$

Chapter 7
Normal and Lognormal Distributions

Chapter Intuition

This chapter deals with the most important probability distribution in statistics, the normal distribution, which is useful for conducting many kinds of analyses. The ***probability density function*** of a ***normal distribution***, a bell-shaped curve, is completely described by two parameters, the **mean μ** and the **variance σ^2**. Let Z be a ***standard normal random variable*** with mean zero and variance one, the probability that Z is less than a constant a is the area to the left of a under the bell-shaped curve. This area can be found from a standard normal distribution table.

The normal distribution is important because it approximates many different types of distributions, including discrete distributions. One important approximation is the ***sampling distribution*** of a sample mean $\bar{X}$ of a set of *N independent and identically distributed* random samples from an arbitrary population. According to the ***central limit theorem***, under certain conditions, the ***sampling distribution*** of the sample mean $\bar{X}$ follows a normal distribution as N is getting larger, regardless of the distribution of the population. This is very important because the sample mean is often used for inference of the population; however, in many situations, the distribution of the population is unknown.

The second type of distribution discussed in this chapter is the ***lognormal distribution***. If a random variable X takes on only positive values, it is said to have a lognormal distribution if $\log(X)$ follows a normal distribution. The lognormal distribution can be useful for describing variables such as stock prices.

Chapter Review

1. The probability density function of a normal distribution is bell-shaped symmetrical around the mean μ, and the shape of the curve is completely described by its variance, σ^2. The smaller the variance, the less spread out the curve is around the mean. In the following two figures, we show how the distribution changes as the

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_7

mean and variance change. In Fig. 7.1, two normal distributions with variances mean1 < mean2 and in Fig. 7.2, three normal distributions with variances var1 < var2 < var3 are illustrated.

2. The standard normal random variable, Z, is a normal distribution with mean of zero and standard deviation one. The ***standard normal distribution table*** gives us the probability that Z is between 0 and a given critical value. Figure 7.3 shows the cumulative distribution of a standard normal distribution.
3. To use the **standard normal distribution table** for any normal variable X, we need to transform X into a standard normal random variable by $Z = (X - \mu)/\sigma$, where μ is the mean and σ is the standard deviation of the normal random variable X.
4. A normal distribution can be used to approximate the binomial and Poisson distributions.
5. ***Lognormal distribution*** $X = e^Y$ is lognormally distributed if Y is normally distributed. The lognormal distribution is especially useful in business and economics such as stock prices in the Black-Scholes option pricing model, in which the value of a European style call/put options are computed. Figure 7.4 shows the lognormal distribution with a standard deviation of one and two different means.

Useful Formulas

1. Standard normal variable:

$$Z = \frac{X - \mu_x}{\sigma_x}$$

2. Normal approximation of a binomial random variable $X \sim B(n, p)$ for $np > 5$ and $n(1-p) > 5$

$$Z = \frac{X - np}{\sqrt{np(1-np)}}$$

3. Normal approximation of Poisson random variable $X \sim$ Poisson (λ)

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

4. Black-Scholes option pricing formula:

$$P_0 = SN(d_1) - Ee^{-rt}N(d_2)$$

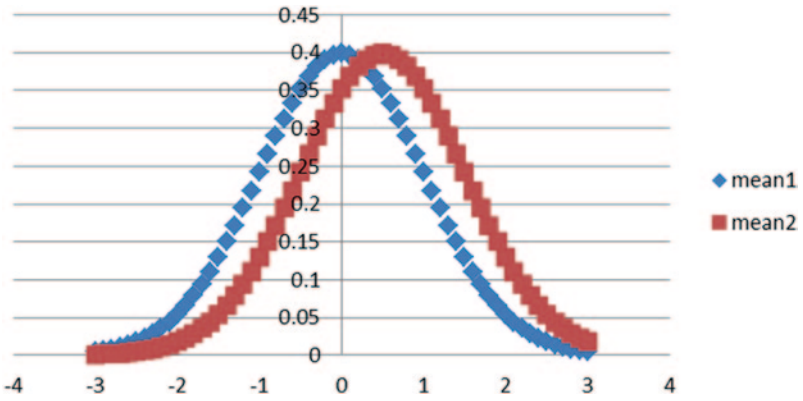


Fig. 7.1 Probability density function of normal distribution with two means

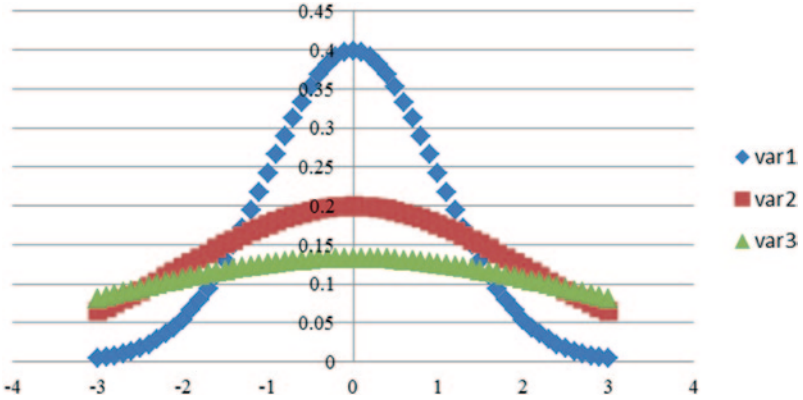


Fig. 7.2 Probability density of normal distribution with three variances

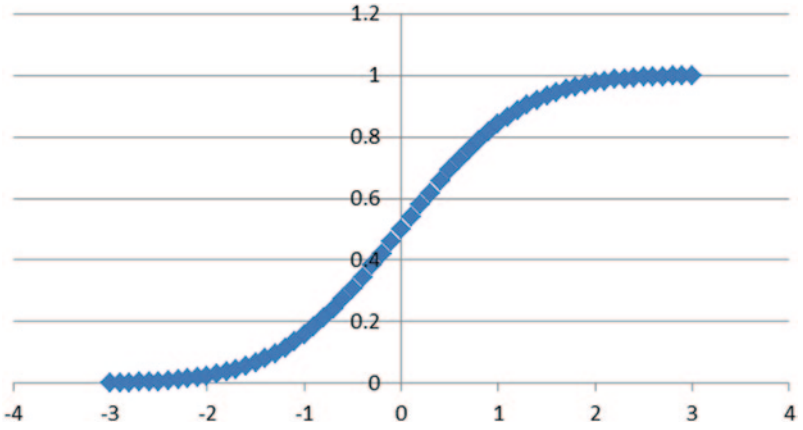


Fig. 7.3 Cumulative distribution of standard normal

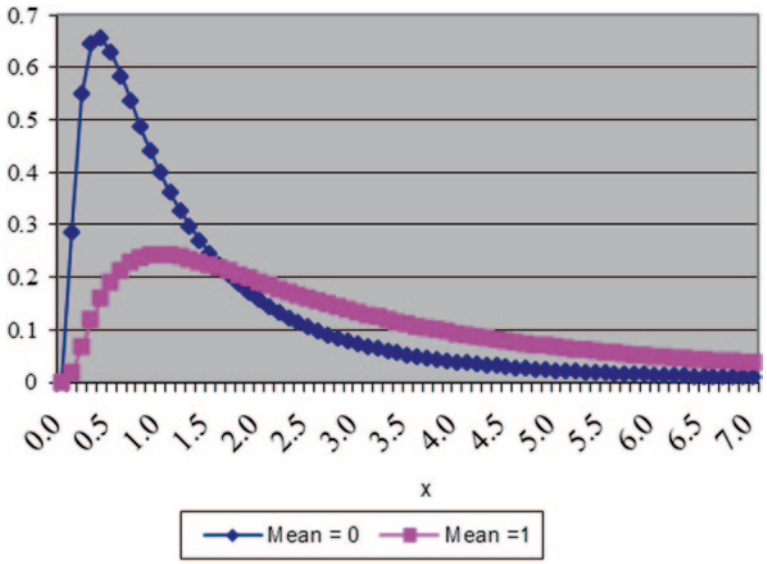


Fig. 7.4 Lognormal distribution different means

$$d_1 = \frac{\ln\left(\frac{P_s}{E}\right)\left(r + \frac{\sigma^2}{2}\right)t}{\sigma\sqrt{t}}$$

$$d_2 = \sigma\sqrt{t}$$

S Current stock price;
E Strike price
σ Volatility
r Risk-free rate

5. Suppose X is a discrete random variable with probability function $f(x)$, then the probability X lies between a and b is

$$P(a \le X \le b) = \sum_{x=a}^{b} f(x)$$

6. Suppose X is a discrete random variable with probability density function $f(x)$, probability that X lies between a and b (continuous random variable) is

$$P(a \le x \le b) = \int_a^b f(x)\,dx$$

7. Lognormal distribution: if $X = e^Y$, where $Y \sim N\left(\mu_y, \sigma_Y^2\right)$. Note the relation of the means is

$$\mu_X = e^{\mu_Y + \sigma_Y^2/2}$$

Example Problems

Example 1 Area Under the Probability Density Function of a Normal Distribution

Calculate the area under the probability density function of standard normal distribution between the following values.

a. $z_1 = 0$ and $z_2 = 1.0$
b. $z_1 = -3.5$ and $z_2 = -1$
c. $z_1 = -1.0$ and $z_2 = 1.0$
d. $z_1 = -3.0$ and $z_2 = 0$

Solution: The areas can be found by looking at the standard normal distribution table. Because the normal distribution is symmetric with mean equal to zero, the area under the curve below zero equals the area under the curve above zero.

a. $P_r(0 < Z < 1.0) = .3413$

b. $P_r(-3.5 < Z < -1) = .1576.$

c. $P_r(-1 < Z < 1) = .6826$

d. $P_r(-3 < Z < 0) = .4987$

Example 2 Area Under the Probability Density Function of a Normal Distribution

Calculate the area under the probability density function of a normal random variable $X \sim N\left(3, \sqrt{2}\right)$ between the following values:

a. $x_1 = 7$ and $x_2 = 8.0$
b. $x_1 = -3.5$ and $x_2 = -1$
c. $x_1 = -1.0$ and $x_2 = 1.0$
d. $x_1 = -4.0$ and $x_2 = 2.0$

Solution: Because these variables are not standard normal, they must be converted to standard normal variables, using the following formula:

$$Z = (X - \mu)/\sigma$$

Once the variables have been standardized, the procedure is the same as for Example 1.

a. $z_1 = [7-3]/\sqrt{2} = 2.83$
$z_2 = [8-3]/\sqrt{2} = 3.54$
$P_r(2.83 < Z < 3.54) = .0021$

b. $z_1 = [-3.5-3]/\sqrt{2} = -4.6$
$z_2 = [-1-3]/\sqrt{2} = -2.83$
$P_r(-4.6 < Z < -2.83) = .023$

c. $z_1 = [-1-3]/\sqrt{2} = -2.83$
$z_2 = [1-3]/\sqrt{2} = 1.41$
$P_r(-2.83 < Z < 1.41) = .9169$

d. $z_1 = [-4-3]/\sqrt{2} = -4.95$
$z_2 = [2-3]/\sqrt{2} = .707$
$P_r(-4.95 < Z < .707) = .7580$

Example 3 Normal Approximation to a Binomial Distribution *B(n, p)*

Using normal approximation to the binomial random variable X with $n = 100$ and $p = .2$.

a. What is the probability X will be greater than 25?
b. What is the probability X will be less than 10?

Solution: Since $np = 20 > 5$ and the variance is $np(1-p) = 16 > 5$, X can be converted to a normal random variable with mean $E(X) = np = 20$, and variance $V(X) = np(1-p) = 16$.

Standardize X to a stand normal random variable Z as follows:

a. $z = [25-(100)(.2)]/[(100)(.2)(1-.2)]^{1/2} = 1.25$
$P_r(Z > 1.25) = .0885$ or 8.85%

b. $z = [10-(100)(.2)]/[(100)(.2)(1-.2)]^{1/2} = -2.5$
$P_r(Z < -2.5) = .0062$ or .62%

Example 4 Normal Approximation to a Poisson Distribution Poisson (λ)

Using the normal approximation to a Poisson random variable $X \sim \text{Poisson}\ (\lambda = 36)$:

a. What is the probability X will be greater than 30?
b. What is the probability X will be between 30 and 48?

Solution: We need to convert X to a standard normal random variable. The mean and variance of a Poisson distribution are both λ,

a. $z = [30 - 36] / \sqrt{36} = -1.000$
   $P(Z > -1) = 0.8413$

b. $z_1 = [30 - 36] / \sqrt{36} = -1.000$
   $z_2 = [48 - 36] / \sqrt{36} = 2.000$
   $P(-1 < Z < 2) = 0.8185.$

Example 5 Probability of a Normal Distribution

A quality control manager has found that the lifespan of light bulbs follows a normal distribution with a mean 320 and a standard deviation 15. Compute the probability that the lifespan of a light bulb is

a. Greater than 340
b. Less than 310
c. Between 340 and 350

Solution: To use the normal distribution table, we must standardize the random variable.

a. $z = [340 - 320] / 15 = 1.33 \rightarrow P(Z > 1.33) = .0918$ or 9.18%

b. $z = [310 - 320] / 15 = -.667 \rightarrow P(Z < -.667) = .2514$ or 25.14%

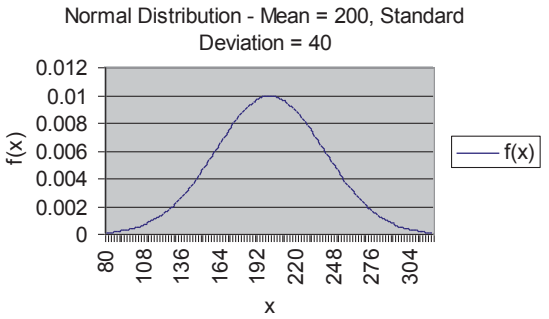
c. $z_1 = [340 - 320] / 15 = 1.33$ and $z_2 = [350 - 320] / 15 = 2$
   $P_r(1.33 < Z < 2) = .064$ or 6.4%

Example 6 Area Under the Normal Distribution

The manager in a local convenience store finds that the gallons of milk sold in a week follows a normal distribution with a mean 200 and a standard deviation 40. If the manager wants to make sure that there is sufficient milk in the store at the beginning of the week so there is only 5% chance that the store will run out of milk, what is the minimum amount of milk she should purchase at the beginning of the week?

Solution: Assume that the amount of milk demanded in a week is X, and that X follows a normal distribution with a mean 200 and a standard deviation 40. We should

find the minimum amount of milk $x_{\min}$ so that the chance is 5 % that X will be greater than $x_{\min}$. The probability density function of X is shown in the following graph:



$P_r(X > x_{\min}) = 5\%$, the probabilities of the standardized X are

$$P\left[(X-200)/40 > (x_{min}-200)/40\right] = 5\% \rightarrow P\left[1.64 > (x_{min}-200)/40\right] = 5\%$$

$$1.64 = (x_{min}-200)/40 \rightarrow x_{min} = 265.6$$

Supplementary Exercises

Multiple Choice

1. The normal distribution is

   a. A discrete distribution
   b. Approximated by the binomial distribution
   c. A continuous distribution completely described by its mean and variance
   d. Approximated by the Poisson distribution
   e. Is positively skewed

2. $P_r(Z < -1.64) =$

   a. 5 %
   b. 95 %
   c. −5 %
   d. −95 %
   e. 90 %

3. The area under the probability density function of a normal distribution between −1 and 3 is

   a. $P_r[X < 3] - P_r[X < -1]$
   b. $P_r[X > 3] - P_r[X < -1]$

c. $P_r[X < 3] + P_r[X < -1]$
d. $P_r[X > 3] + P_r[X < -1]$
e. $P_r[X = 3] + P_r[X = -1]$

4. The probability that a normally distributed random variable X, is greater than 2 is the

   a. Area under the normal distribution to the left of 2
   b. Area under the normal distribution to the right of 2
   c. Height of the normal distribution at 2
   d. Area under the entire normal distribution curve
   e. Area under the entire normal distribution curve divided by 2

5. Suppose we have two normally distributed random variables, X and Y. Variable X has a mean of 100 and a standard deviation of 20, whereas Y has a mean of 100 and a standard deviation of 40. If the two distributions are plotted on the same graph, the

   a. Distribution of Y will be to the right of the distribution of X
   b. Distribution of Y will be to the left of the distribution of X
   c. Two distributions will have the same center; however, the distribution of Y will be flatter
   d. Two distributions will have the same center, but the distribution of Y will be steeper
   e. Distribution of Y will be skewed to the left and the distribution of X will skewed to the right

6. Suppose we have two normally distributed random variables, X and Y. Variable X has a mean of 100 and a standard deviation of 20, whereas Y has a mean of 200 and a standard deviation of 20. If the two distributions are plotted on the same graph, the

   a. Distribution of Y will be to the right of the distribution of X
   b. Distribution of Y will be to the left of the distribution of X
   c. Two distributions will have the same center, but the distribution of Y will be flatter
   d. Two distributions will have the same center, but the distribution of Y will be steeper
   e. Distribution of Y will be skewed to the left and the distribution of X will be skewed to the right

7. A continuous random variable can

   a. Take on only a finite and countable number of values
   b. Take on an infinite number of values
   c. Be approximated by a discrete random variable
   d. Only take on integer values
   e. Never be negative

8. The Black-Scholes model

   a. Is a formula for analyzing stock prices
   b. Explains why the normal distribution is the most important distribution in statistics
   c. Uses the normal distribution to value options
   d. Is used to approximate the binomial distribution
   e. Is used to approximate the Poisson distribution

9. The lognormal distribution

   a. Is valid for nonpositive numbers
   b. Is valid for nonnegative numbers
   c. Can take on any value
   d. Is rarely ever used in finance and accounting
   e. Is a discrete distribution

10. The lognormal distribution is especially useful for describing

    a. Stock prices
    b. Earnings per share
    c. Stock returns
    d. Price/earnings ratios
    e. Earnings growth

11. Approximately 95% of the area under the normal distribution lies

    a. Below the mean
    b. Above the mean
    c. 1 standard deviation away from the mean
    d. 2 standard deviations away from the mean
    e. 3 standard deviations away from the mean

12. The cumulative normal probability at the mean of the distribution is

    a. 1.
    b. 0.
    c. .5
    d. .25
    e. .75

13. A normally distributed random variable can be standardized by

    a. Adding the mean and dividing by the standard deviation
    b. Subtracting the mean and multiplying by the standard deviation
    c. Subtracting the mean and dividing by the standard deviation
    d. Converting to the lognormal distribution
    e. Using the Black-Scholes model

14. If Z is a stand normal random variable, then $P(Z > -1.645) =$

    a. 5%
    b. 95%

c. -5%
d. -95%
e. 90%

15. If Z is a stand normal random variable, then $P(Z < 1.96) =$

a. 2.5%
b. $.97.5\%$
c. -2.5%
d. -97.5%
e. 95%

16. If Z is a stand normal random variable, then $P(Z < -1.96) =$

a. 2.5%
b. 97.5%
c. -2.5%
d. -97.5%
e. 95%

Use the following information for 17 and 18. A test has been devised to measure a student's level of motivation during high school. The motivation scores on this test are approximately normally distributed with a mean of 25 and a standard deviation of 6. The higher the score, the greater the motivation to do well in school.

17. What percentage of students taking this test will have scores below 10?

a. 99.38%
b. 0.62%
c. 62.00%
d. 0.38%
e. 38.00%

18. John is told that 35% of the students taking the test have higher motivation scores than he does? What was John's score?

a. 99.01
b. 61.32
c. 25.98
d. 27.34
e. None of the above

19. X is a normally distributed random variable with mean 5 and standard 1.75. Which of the following is standard normal?

a. $(X - 1.75)/5$
b. $(X - 5)/3.0625$
c. $(X - 1.75)/3.0625$
d. $(X - 5)/1.75$
e. $(X - 3.0625)/5$

20. X and Y are independent normal random variables. The mean and variance of X are 2 and 1, respectively. The mean and variance of Y are 3 and 2, respectively. Which of the statements below is true?

    a. X-Y is normal with mean 5 and standard deviation 3
    b. X-Y is normal with mean 5 and standard deviation 3
    c. X-Y is normal with mean −1 and standard deviation −1
    d. X-Y is normal with mean −1 and standard deviation 1
    e. X-Y is normal with mean −1 and standard deviation 1.732

True/false (If false, explain why)

1. The binomial distribution can always be approximated using the normal distribution.
2. The Poisson distribution can be approximated using the normal distribution.
3. The normal distribution can have a negative variance.
4. The normal distribution can be positively or negatively skewed.
5. There are an infinite number of normal distributions.
6. The normal distribution can have a variance equal to 0.
7. The mean, median, and mode are always the same for the normal distribution.
8. The lognormal distribution is better for describing stock prices than the normal distribution, because stock prices can never be negative.
9. The probability density function gives us the probability that the random variable, x, will be less than some given value.
10. The area under the normal distribution curve will depend on the mean and standard deviation.
11. The normal distribution is a discrete distribution.
12. The normal distribution is valid for values of x ranging from $-\infty$ to $+\infty$.
13. The lognormal distribution is valid for values of x ranging from $-\infty$ to $+\infty$.
14. The lognormal distribution is a bimodal distribution.
15. Mean, median, mode of a normal distribution are not necessarily equal.

    a. true
    b. false

16. The probability density curve of a normal distribution is symmetrical

    a. true
    b. false

17. The probability density curve can have more than a single peak
    a. true
    b. false

18. The probability density curve touches the X-axis at $z = -3$ and $z = 3$

    a. true
    b. false

Questions and Problems

1. Compute the z value for $x = 25$, if $\mu = 20$ and $\sigma = 2$.
2. During final exam week, a college professor has an average of 12 students per hour during her office hours. Use the normal approximation to the Poisson to find the probability that she will have fewer than 10 students in any given hour.
3. Suppose you toss a fair coin 50 times. Use the normal approximation to the binomial distribution to find the probability that you will get at least 28 tails.
4. Calculate the area under the normal curve between the following values given that the mean is 125 and the standard deviation is 20.

    a. $x_1 = 90$ and $x_2 = 105$

    b. $x_1 = 125$ and $x_2 = 155$

5. Use the Black-Scholes option pricing formula to compute the value of a call option given the following information.

S	\$ 50 current stock price
E	\$ 45 strike price
r	.045 risk-free interest rate
T	.55 years till option expires
σ	.15 volatility of the stock's return

6. Suppose you work for a company that uses batteries. Every time a truckload of batteries arrives, you examine 400 of the 10,000 batteries delivered. Your company policy says that you should refuse any truckload with more than four bad batteries. Suppose a truck load of batteries arrives that has a total of 200 bad batteries. What is the probability that you accept this shipment? (Hint: What is the probability that you will find 4 bad batteries in the 400 you examine, given that 200 of the 10,000 are bad?)
7. Suppose the number of cups of yogurt sold by a convenience store each week follows a normal distribution with a mean of 600 and a standard deviation of 20. If the store owner wants to make sure that he does not have unsold yogurt in more than 5% of the weeks, how many cups of yogurt should he order?

Answers to Supplementary Exercises

Multiple Choice

1.	c	6.	a	11.	d	16.	a
2.	c	7.	b	12.	c	17.	b
3.	a	8.	c	13.	c	18.	d
4.	b	9.	b	14.	b	19.	d
5.	c	10.	a	15.	d	20.	e

True/False

1. False. The binomial distribution can be approximated by the normal distribution when the sample size is large enough
2. True
3. False. Variance can never be negative for any distribution
4. False. The normal distribution is always symmetric
5. True
6. False. When a distribution has a variance equal to zero, all values from the distribution will be identical
7. True
8. True
9. False. The cumulative distribution function gives us the probability that the random variable, x, will be less than some given value
10. False. Because the normal distribution is a probability distribution, the area under the curve will always be equal to 1, regardless of the mean and standard deviation
11. False. Continuous distribution
12. True
13. False. Valid only for positive numbers
14. False. Unimodal distribution
15. False
16. True
17. False
18. False

Questions and Problems

1. z = 2.5
2. .281
3. .4052
4. a. 1186
   b. 3944
5. $ 6.41
6. We know that 200 out of 10,000 batteries are bad. This means that 2% of the batteriesare bad. If we sample 400 batteries, the probability of returning the shipment is

$$P(X > 4) = P\left(\frac{X}{n} > \frac{4}{400}\right)$$

Using the normal distribution to approximate the binomial distribution, we get

$$P(Z > 4.52) = 0$$

So it is almost impossible for us to return this shipment.

7. $P(X < 600) = 5\%$

$$P\left(Z < \frac{x-600}{20}\right) = .5\%$$

$$\left(\frac{x-600}{20}\right) = -1.645$$

Solving for x gives us x = 567.1 or approximately 567 cups of yogurt.

Chapter 8
Sampling and Sampling Distributions

Chapter Intuition

Often we are interested in the parameters of a population. For example, a manager may be interested in knowing whether the average life of the light bulbs produced yesterday equals or exceeds 300 h. One way to conduct this analysis is to use a census; that is, light up all of the light bulbs produced yesterday and determine the average life. This approach is both impractical and costly. An alternative is to use a sample of 100 light bulbs to estimate the average life of all the light bulbs. Because we are using a sample rather than the whole population, errors can occur if we use the average life of the 100 sample light bulbs to make inferences about the entire population. In our light bulb example, it is possible that the average life of the light bulbs in the population is high (e.g., 350 h); however, the sample average is only 220 h. Using the sample mean of 220, we erroneously conclude that the average life of all the light bulbs is low. Sampling distributions allow us to determine the probability that we will make this type of mistake.

If we use the sample mean to estimate the population mean, we can determine the probability distribution of the estimation error incurred via the sampling distribution of the sample mean. Without the sampling distribution of the sample mean, we will not know the reliability of the estimator, i.e., the sample mean we use. This chapter builds on the basic concepts of probability and distributions discussed in previous chapters. This chapter also forms a building block for inferential statistics, which is discussed in part III of the text.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_8

Chapter Review

1. When analyzing a population, either a ***census*** or a ***sample*** can be used to conduct the analyses. In a census, all members of the ***population*** are examined. In a sample, only a subset of the entire population is drawn and examined. The advantage of a sample over a census is cost reduction as a much smaller set of members in the population are dealt with.
2. Once we have collected a sample, we use the information to draw inferences about a population that is not known with certainty. In order to quantify this uncertainty, we use the probability concepts discussed in Chap. 5 and the probability distribution of Chaps. 6 and 7.
3. Often we are interested in the sampling distribution of the sample mean. Based on the ***central limit theorem***, the sampling distribution of the sample mean can be found. The central limit theorem says that when a set of random samples of size n is drawn from a population with mean μ_x and standard deviation $\sigma_{x,}$ if n is large enough, the distribution of the sample mean $\bar{X}$ will be approximately normal with mean $\mu_{\bar{X}} = \mu_x$ and standard deviation $\sigma_{\bar{X}} = \sigma_x / \sqrt{n}$ regardless of the population distribution.
4. The most basic procedure for selecting a set of random samples is ***simple random sampling*** technique, a procedure in which every member in the population with N members has an equal chance of being selected. If ***simple random sampling*** **without replacement** is used, each subset of n individuals has the same probability $1/C_n^N$ of being chosen for the sample as any other subset of n individuals. If ***simple random sampling*** **with replacement** is used, each subset of n individuals has the same probability $1/N^n$ of being chosen for the sample as any other subset of n individuals. In a set of simple random samples of n individuals X_1 ,…, X_n, the n individuals X_1 ,…, X_n can be considered as **independently and identically distributed** (***i.i.d***.). For a small sample from a large population, sampling without replacement is approximately the same as sampling with replacement.
5. When a set of random samples is used rather than a census, errors may occur. ***Sampling errors*** result from the selection of the samples by chance, i.e., ***random errors***. ***Nonsampling errors*** are due to inaccurate measurement or improper selection of the sample. Nonsampling error is a ***systematic error*** because it affects all the members of the sample in a similar manner. For example, using a ruler that is too short to measure the height of the basketball team. Unlike random errors, systematic error cannot be eliminated by increasing the sample size.

Useful Formulas

1. Sample mean and sample variance:

$$\bar{X} = \frac{1}{n}\sum_{i-1}^{n} X_i \qquad s_X^2 = \frac{1}{n-1}\sum_{i-1}^{n}\left(X_i - \bar{X}\right)^2$$

Standard deviation of the sample mean:

$$\sigma_{\bar{X}} = \sigma_X / \sqrt{n}$$

Mean and variance for a sample proportion:

$$E(\bar{p}) = p \qquad \sigma_{\bar{p}}^2 = \frac{p(1-p)}{n}$$

Example 1 Sample vs. Population

Suppose a city consists of 47,825 people with 13,721 registered democrats, 21,622 registered republicans, and 12,482 independents. If you are interested in who will be nominated by the democrats for the next mayoral election, which group would constitute the population? Give an example of a sample from this population.

Solution: Since we are interested in who will be nominated by the democrats, our population consists of all registered democrats. A set of random samples would consist of a subset of our 21,622 registered democrats, for example, the first 1250 registered democrats alphabetically would be the possible choices of the sample. However, we should be careful when choosing a set of random samples in this manner, because it is not clear that the first 1250 democrats alphabetically will represent the viewpoints of all registered democrats. It might occur if a large family, all with the last name beginning with A, represents a large proportion of the registered democrats.

Example 2 Sampling and Nonsampling Error

Which statement in the following represent sampling error? Nonsampling error?

a. The heights of a set of students are measured with an inaccurate ruler.
b. A set of 57 new houses in the USA shows a mean price of $ 46,200 when the population mean of the prices of new houses is $ 53,100.
c. A restaurant manager asks the first 85 customers about their opinions on the food. By chance, the first 85 customers are all teenagers.

Solution:

a. This represents nonsampling error because the error is due to the inaccurate ruler, not the sample selected.
b. This represents sampling error. It is simply by chance that the sample we selected has a mean above the population mean. We could just as easily have chosen homes that would have given us a mean home price below the actual home price.
c. Once again we have sampling error. It is simply by chance that the first 25 people the owner surveys are teenagers.

Example 3 Distribution of the Sample Mean

The mean life of radial tires produced by the Sure Grip Tire Company is 25,000 miles, with a standard deviation of 4,700 miles. Suppose a set of random sample of 39 tires is drawn.

a. What is the mean of the sample mean life?
b. What is the variance of the sample mean?

Solution:

a. $E(\bar{X}) = 25,000$.

b. $Var(\bar{X}) = Var(\bar{X})/n$

$= 4700^2/39 = 566,410$

Example 4 Probability Associated with the Sample Mean

Suppose, the time a customer spends at a bank is normally distributed with a mean of 19 min and a standard deviation of 5 min. If you take a random sample of five customers, what is the probability that the average time spent will be at least 12 min? What is the mean of the average time spent? What is the standard deviation of the mean waiting time?

Solution:

$$\mu = 16 \text{ minutes.}$$

$$\sigma_{\bar{x}} = \sigma_x/\sqrt{n} = 5/\sqrt{5} = 2.24.$$

$$Z = (\bar{X} - \mu)/\sigma_{\bar{x}} = (12 - 16)/2.24 = -1.78.$$

$$P = (\bar{X} > 12) = P(Z > -1.78) = .9625.$$

Example 5 Sample Proportion

Rah Rah University accepts 60 % of all students who apply for admissions. A random sample of 100 applicants is taken.

a. What is the probability that more than half the applicants sampled are accepted?
b. What is the probability that the sample acceptance rate is between .50 and .75?

Solution:

$$p = 0.6 \qquad \sigma_{\bar{p}} = \sqrt{.6(1-.6)/100} = 0.49.$$

a. $P(\bar{p} > .5) = P(Z > (.5 - .6)/.049) = P(.Z > -2.04) = .9793$.
b. $P(.5 < \bar{p} < .75) = P(-2.04 < Z < 3.06) = .9782$.

Example 6 Sample Proportion

From past history, an auto dealer knows that 8 % of all customers entering the showroom make a purchase. Suppose, 100 people enter the showroom

a. What is the mean of the sample proportion of customers making a purchase?
b. What is the variance of the sample proportion?
c. What is the standard deviation of the sample proportion?
d. What is the probability that the sample proportion is between .05 and .10?

Solution:

a. $\mu = .08.$
b. $\sigma^2 = p(1-p)/n = .08 \times (1 - .08)/100 = .000736$.
c. $\sigma = \sqrt{\hat{\sigma}^2} = \sqrt{.000736} = .0271$.
d. $Z_1 = (.05 - .08)/.0271 = -1.11$
   $Z_2 = (.10 - .08)/.0271 = .738$

$$P(.05 < p < .10) = (-1.11 < Z < .738)$$
$$= P(Z < .738) - P(Z < -1.11) = .7704 - .1335 = .6369 \text{ or } 63.9\%$$

Example 7 Variance of the Sample Mean $\bar{X}$

Find the standard deviation of the sample mean or proportion taken from the following populations.

a. Suppose a random sample of 100 students is taken from the student body of 5,000. The students are asked whether they support the new alcohol policy on campus. Assume 3,000 students support the new policy.
b. A random sample of ten cans of dog food is taken from a case of 24 cans. The mean weight of the dog food is 16 ounces with a standard deviation of .5 ounces.
c. A random sample of ten cans of dog food is taken from the daily production of an assembly line.The mean is 16 ounces and the standard deviation is .5 ounces.

Solution:

$$a.\frac{p(1-p)}{n}=\frac{.6(1-.6)}{100}=.0024.$$

$$b.\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}=\frac{.5}{\sqrt{10}}\sqrt{\frac{24-10}{24-1}}=.123.$$

$$c.\frac{\sigma}{\sqrt{n}}=\frac{.5}{\sqrt{10}}=.158.$$

Example 8 Distribution of the Sample Mean

Use the following information to find the mean and the standard deviation of the sample mean $\bar{X}$ with size n from a population with mean μ and standard deviation of σ.

a. $n=25,\mu=4.73,\sigma=2.2.$
b. $n=25,\mu=6.03,\sigma=3.7.$
c. $n=20,\mu=45,\sigma=7.9.$
d. $n=40,\mu=120,\sigma=23.75.$

Solution:

a. $E(\bar{X})=4.73\ \sigma_{\bar{x}}=2.2/\sqrt{25}=.44$.
b. $E(\bar{X})=6.03\ \sigma_{\bar{x}}=3.7/\sqrt{25}=.74$.
c. $E(\bar{X})=45\ \sigma_{\bar{x}}=7.9/\sqrt{20}=1.77$.
d. $E(\bar{X})=120\ \sigma_{\bar{x}}=23.75/\sqrt{40}=3.76$.

Example 9 Cen tral Limit Theorem

A sample of 200 cans of soda is taken. If a statistician does not know the original distribution of the soda, can he make a statistical inference of the mean based on the sample? If a sample of 20 cans is taken, what kind of assumption is needed before he conducts any statistical inference? If one wants to estimate the mean and the standard deviation of the sample mean, is the above assumption needed?

Solution: Under certain conditions, the statistical inference based on a large sample can be conducted using the central limit theorem. If the sample is small, we need to know the distribution of the population. However, if we are only interested in estimating the mean and the standard deviation, we do not need to make any assumptions about the distribution.

Example 10 Sampling Without Replacement

Suppose, we have a population with five numbers: 1, 6, 8, 10, and 12. If we randomly draw two numbers without replacement, let $\bar{X}$ be the sample mean. What is the probability distribution, the mean and the standard deviation of the sample mean $\bar{X}$?

Solution: The sample space is

(1,6), (1,8), (1,10), (1,12), (6, 8), (6, 10), (6, 12), (8, 10), (8, 12), (10, 12).

The distribution of the sample mean $\bar{X}$ is:

$\bar{X} = a$	$P(\bar{X} = a)$	$aP(\bar{X} = a)$	$[a - E(\bar{X})]^2 P(\bar{X} = a)$
7	1/10	7/10	6.084
9	1/10	9/10	3.364
11	1/10	11/10	1.444
13	1/10	13/10	0.324
14	1/10	14/10	0.064
16	1/10	16/10	0.144
18	2/10	36/10	2.048
20	1/10	20/10	2.704
22	1/10	22/10	5.184
		14.8	21.36

$$E(\bar{X}) = 14.8 \quad Var(\bar{X}) = 21.36$$

Example 11 Distribution of a Sample Proportion

If a town of 300 residents, 30 % approve of the new garbage collection policy. If the local newspaper asks 120 residents about their opinion of the new garbage collection policy, what is the probability that more than 30 people will be in favor of it?

Solution: As the sample is relatively large to the population, so we can use the Z statistics, the distribution of $\bar{p}$ is

$$\bar{p} \sim N\left[p, \left(\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}\right)^2\right]$$

$$\sigma_p^2 = \left(\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}\right)^2 = \left(\sqrt{\frac{0.3(1-0.3)}{120}}\sqrt{\frac{300-120}{300-1}}\right)^2 = 0.0010535$$

$$\sigma_{\bar{p}} = \sqrt{0.0010535} = 0.325$$

$$P\left(\bar{p} > \frac{30}{120}\right) = P\left(\frac{\bar{p} - 0.3}{0.0325} > \frac{0.25 - 0.3}{0.0325}\right) = P(Z > -1.54) = 0.9382.$$

Example 12 Probability

Suppose you play a game where you roll two dice once. You win if you roll a sum of nine or more. What is the probability of winning?

Solution:

Sum	Sample	P_r
2	(1, 1)	1/36
3	(1, 2) (2, 1)	2/36
4	(1, 3) (2, 2) (3, 1)	3/36
5	(1,4) (2, 3) (3, 2) (4, 1)	4/36
6	(1, 5) (2, 4) (3, 3) (4, 2) (5, 1)	5/36
7	(1, 6) (2, 5) (3, 4) (4, 3) (5, 2) (6, 1)	6/36
8	(2, 6) (3, 5) (4, 4) (5, 3) (6, 2)	5/36
9	(3, 6) (4, 5) (5, 4) (6, 3)	4/36

Sum	Sample	P_r
10	(4, 6) (5, 5) (6, 4)	3/36
11	(5, 6) (6, 5)	2/36
12	(6, 6)	1/36

$$P(X \geq 9) = P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$
$$= 4/36 + 3/36 + 2/36 + 1/36 = 10/36.$$

Example 13 Distribution of a Sample Proportion

In a survey by United Airlines of 100 flights between January 2, 2012 and February 15, 2012, 84 of the flights arrived at their destination on time. What is the sample proportion? Please estimate the standard deviation of the sample proportion, and calculate the probability that the sample proportion is less than 80%.

Solution:

$$\bar{P} = 0.84, \sigma\sqrt{\frac{0.84 \times 0.16}{100}} = 0.0367, \qquad (\bar{P} < 0.80) = \mathrm{P}(Z < -1.091) = 0.1376.$$

Example 14 Distribution of a Sample Proportion

On page 370, financial data of 37 companies in the communication and Internet sector are given. In the data set, the debt-to-asset, dividend per share, current ratio, fixed asset turnover, ROA, and P/B ratio in 2011 are given. Please use the sample means and variances of the above six financial variables in 2011 as an estimate of the population means and variances of the year 2012. What is the probability the sample mean of the P/B ratio of the 37 companies is greater than −22.5 but less than 2.5 in 2012? What is the sample proportion of the P/B of the 37 companies that are greater than 1 in 2012? What is the probability the sample proportion of the P/B greater than 1 is greater than 0.5 in 2012?

Solution:

a.
$$\mathrm{P}(-2.5 < \bar{X} < 2.5) = P\left(\frac{-2.5 - 1.712}{1.389} < Z < \frac{2.5 - 1.712}{1.389}\right)$$
$$= \mathrm{P}(-3.03 < \mathrm{Z} < 0.57) = 0.7136.$$

b. $$\overline{p} = 23/37 = 62.16\%.$$

c. $$P(\overline{p} > 0.2) = P\left(Z > \frac{0.5 - 0.6216}{\sqrt{0.2 \times 0.8 / 37}} \right) = 0.9304.$$

Supplementary Exercises

Multiple Choice

1. A census is conducted by

   a. Surveying a subset of the population of interest
   b. Surveying all members of the population of interest
   c. Either surveying all members or a subset of the population of interest
   d. Surveying half the population
   e. Surveying more than half the population

2. A sample is conducted by

   a. Surveying a subset of the population of interest
   b. Surveying all members of the population of interest
   c. Either surveying all members or a subset of the population of interest
   d. Surveying half the population
   e. Surveying more than half the population

3. The advantage of sampling is that it can

   a. Reduce costs
   b. Reduce the time spent on data collection
   c. Increase data manageability
   d. Sometimes be more accurate than a census
   e. All of the above

4. Sampling errors are

   a. Caused by inaccurate measurement
   b. The result of the chance selection of the sampling units
   c. Of no great concern
   d. Larger for a census than for a sample
   e. Always positive for a census

5. Nonsampling errors are

   a. Caused by inaccurate measurement
   b. The result of the chance selection of the sampling units
   c. Of no great concern

d. Larger for a census than for a sample
e. Always positive for a census

6. The central limit theorem says that as the sample size, n, from a given population gets large enough, the sampling distribution of the mean can be approximated by

a. The binomial distribution
b. The normal distribution
c. The Poisson distribution
d. A Bernoulli process
e. Different distributions depending on the given population

7. The finite population multiplier is used to

a. Reduce bias in the sample mean
b. Reduce the bias in the sample variance
c. Make the computation of the variance easier
d. Make the computation of the mean easier
e. Reduce sampling error

8. The mean of a binomial distributed random variable with n experiments and p as the probability of success for each experiment is

a. np
b. $np(1 --p)$
c. $np/\sqrt{p(1-p)}$
d. n
e. p

9. The variance of a binomial distributed random variable with n experiments and p as the probability of success for each experiment is

a. np
b. $np(1 --p)$
c. $np/\sqrt{p(1-p)}$
d. n
e. p

10. If we sample without replacement,

a. It is important to consider the size of the sample relative to the size of the population
b. A larger sample relative to the size of the population is preferred because it will reduce sampling error
c. The sample size is unimportant
d. Use a smaller sample
e. We should take a census

11. If we sample without replacement and the sample is relatively large to the population,

    a. The sample mean will be small
    b. The sample variance will be small
    c. The sample size is unimportant
    d. Use a smaller sample
    e. We should take a census

12. If we sample with replacement,

    a. It is important to consider the size of the sample relative to the size of the population
    b. A larger sample is preferred
    c. The sample variance will not be biased
    d. A smaller sample is preferred
    e. The sample mean will be large

13. If we sample with replacement and sample is large relative to the population,

    a. The sample mean will be biased
    b. The sample variance will be biased
    c. The sample variance must be adjusted
    d. The sample mean will be large
    e. No adjustments need to be made

14. In the following, which is not probability sampling?

    a. Simple random sampling
    b. Stratified sampling
    c. Cluster sampling
    d. Systematic sampling
    e. None of the above

Use the following information for 15 and 16. There are 50 students in a seventh grade class. The teacher wants to know the average hours of television watched each day by the students in his class. Suppose, the He randomly asks ten students how many hours of television they watch each day. The results are 1, 3, 3, 4, 5, 6, 9, 11, 12, and 17 h.

15. What is the sample mean $\bar{X}$ of hours of television watched each day?

    a. 7.100
    b. 2.667
    c. 3.667
    d. 3.5
    e. 3.143

16. What is the sample standard deviation *s*?

    a.2.5
    b.5.021

c.3.556
d.3.067
e.3.678

True/False (If False, Explain Why)

1. The central limit theorem can be used to determine the distribution of the sample mean except distributions for population proportions.
2. To use the central limit theorem, we need to have a large enough sample size.
3. Results from a census will always be more accurate than results from sample.
4. Only continuous distributions can use the central limit theorem.
5. The variance of the sample mean will decrease as the sample size increases.
6. The sample proportions can be found by dividing the number of sample members, x, by the sample size, n.
7. Sampling error cannot be eliminated.
8. Simple random sampling is a procedure in which every element in the population has an equal chance of being selected.
9. In order to compute the mean and the standard deviation of a sample, we must know the distribution of the sample.
10. The standard deviation of the sample mean becomes larger as the sample size increases.
11. Nonsampling errors can be eliminated by taking a census.
12. If we sample with replacement, we do not need to consider the size of the sample relative to the size of the population.
13. Sampling error is a type of systematic error.
14. The lognormal distribution is a bimodal distribution.
15. Convenience sampling is a sampling method based on human choice rather than random selection. Statistical theory cannot be used explain what is happening.

Questions and Problems

1. Suppose you draw a sample of 100 from a population where the mean is 25,225 and the variance is 40,212. Find the mean of the sample mean and the variance of the sample mean.
2. Suppose, 125 residents in Centerville, USA, are asked whether they would approve the construction of a new high school. Of the 125 residents surveyed, 74 are in favor of the new school. Find the sample proportion and variance.
3. Using the information provided in Problem 2, find the probability that the sample proportion is greater than .55.
4. From past history, a stockbroker knows that .05 of all people she "cold calls" will become new clients. Suppose she calls 200 people. What is the mean and the variance of the sample proportions?

5. Using the information from Problem 4, find the probability that the sample proportion is less than .06.
6. Of the 200 people who attended a conference, 40 % are optimistic about the economy, whereas 60 % are pessimistic. If 100 economists were sampled and asked whether they are optimistic or pessimistic about the economy, what kind of distribution is the sample proportion?
7. Of the 4,000 people who attended a conference, 40 % are optimistic about the economy, whereas 60 % are pessimistic. If 100 economists were sampled and asked whether they are optimistic or pessimistic about the economy, what kind of distribution is the sample proportion?
8. In Problem 7, what is the probability that between 50 and 60 of the economists who are sampled are optimistic about the economy.
9. What is the probability of winning a game if the goal of the game is to score 10 or more on two rolls of a die?
10. When sampling without replacement from a population of size N, if the sample size $n(n > 1)$ is larger than 5 % of the total population, i.e., $n > 0.05N$, what is the expected value of the sample mean $\bar{X}$? the standard deviation of the sample mean $\bar{X}$?

Answers to Supplementary Exercises

Multiple Choice

1.	b	6.	b	11	e	16	b
2.	a	7.	b	12	c		
3.	e	8.	a	13	e		
4.	b	9	b	14	d		
5.	a	10.	a	15	a		

True/False

1. False. The sample proportion is a type of mean (average number of successes out of n experiments), and so the central limit theorem can be used.
2. True.
3. False. In some instances, a sample may be more accurate than a census.
4. False. The central limit theorem can also be applied to discrete distributions.
5. True.
6. True.

7. False. Sampling error can be eliminated by taking a census.
8. True.
9. False. To compute the mean and the standard deviation we do not need to know the distribution.
10. False. The standard deviation of the sample mean becomes smaller as the sample size increases.
11. False. Nonsampling error cannot be eliminated by taking a census. Only sampling error can be eliminated by taking a census.
12. True.
13. False. Random error.
14. False.
15. True.

Questions and Problems

1. $E(\bar{x}) = 25,225 \; Var(\bar{x}) = 4,021.2$
2. $\bar{p} = .592 \quad Var = .00193$
3. $P_r(x > .55) = P_r(z > -.9438) = .8264$
4. $E(\bar{p}) = .05 \qquad \bar{\sigma}^2 = .002375$
5. $P_r(\bar{p} < .06) = P_r[(\bar{p} - .05)/.0154 < (.06 - .05)/.154 = .5239$
6. The sampling proportion $\bar{p} = X/n$ is distributed as

$$\bar{p} = \frac{X}{n} \sim N\left[.4, \left(\sqrt{\frac{.4(1-.4)}{100}}\sqrt{\frac{200-100}{200-1}}\right)^2\right]$$

7. The sampling proportion $\bar{p} = X/n$ is distributed as

$$\bar{p} = \frac{X}{n} \sim N\left[.4, \left(\sqrt{\frac{.4(1-.4)}{100}}\right)^2\right]$$

8. $$P\left(.5 < \frac{X}{n} < .6\right) = P\left[\left(\frac{.5-.4}{\sqrt{\frac{.4(1-.4)}{100}}} < Z < \frac{.6-.4}{\sqrt{\frac{.4(1-.4)}{100}}}\right)^2\right]$$
$$= P(2.04 < Z < 4.08) = .0207$$

9. There are 36 possible outcomes in the game, but only six possible rolls that equal or exceed 10: (4, 6), (5, 5), (6, 4), (5, 6), (6, 5), and (6, 6). Therefore, there is a 6/36 chance of winning.
10. Equal to the population meanμ;
$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, where σ is the population standard deviation.

Chapter 9
Other Continuous Distributions and Moments for Distributions

Chapter Intuition

Although the normal distribution, discussed in Chap. 7, can be used to approximate many types of populations, it will not be appropriate and other distributions will be employed when the shape of the population distributions are not symmetrical and have heavy tails.

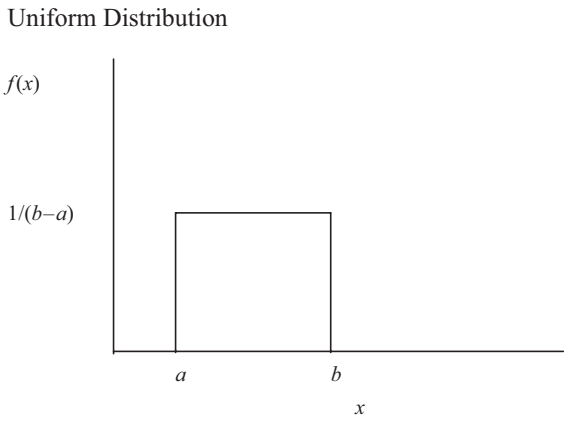
In this chapter, we will learn about two types of continuous distributions. The first type of distributions includes the exponential and uniform distributions. The second type of distributions includes the t, chi-squared X^2, and F distributions. These distributions are usually used to describe the distribution of statistics from a set of random samples and thus are known as sampling distributions.

Chapter Review

1. The simplest of all continuous distributions is the ***uniform distribution***, where the random variable x is assumed to have equal probability of taking on any value over a given interval. For example, if X can only take on values between 5 and 10, and if it is equally likely that X will assume any value between 5 and 10 then X is uniformly distributed. Below is a graph of the uniform distribution.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_9

Uniform Distribution



2. The ***exponential distribution*** is a continuous distribution which is related to the Poisson distribution: The Poisson distribution is used to model the probability distribution of the occurrence times of certain events during a given time interval, while the exponential distribution is used to model the lag time between consecutive events.

When $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two sets of random samples from two independent normal distributions with population means μ_X, μ_Y, and standard deviations σ_X, σ_Y, then we have the following three distributions:

3. ***Student's t distribution*** is used for the statistics

$$\frac{\bar{X} - \mu}{S_X/\sqrt{n}}$$

as the population standard deviation σ_X is unknown and is replaced by the sample standard deviation S_X. Student's t distribution is described by its ***degree of freedom***. As the degree of freedom gets larger, Student's t distribution comes closer to the **standard normal distribution**.

4. When the degree of freedom is small, the t distribution has a heavier tail than the normal distribution.
5. The ***chi-square*** χ^2 ***distribution*** is a continuous distribution that is positively skewed. For a chi-square-distributed random variable W with m degrees of freedom, W can be expressed as the sum of a set of m independent squared standard normal random variables. It is usually used for the statistics

$$\frac{(n-1)S_X^2}{\sigma_X^2}$$

from the samples $X_1, \ldots, X_n$, which is chi-square χ^2 distributed with $(n-1)$ degree of freedom.

6. The ***F distribution*** is a skewed continuous distribution that can be expressed as the ratio of two independent chi-square distributions, adjusted by the degrees of freedom in the numerator and denominator, respectively. It is usually used for the statistics

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

from the two sets of random samples $X_1, \ldots, X_n$, and $Y_1, \ldots, Y_m$, which is F-distributed with $(n-1)$ and $(m-1)$ degrees of freedom.

7. ***Moments*** are measurements used to describe central tendency, dispersion, symmetry, and peakedness of a distribution, respectively.
8. Continuous distributions such as the uniform distribution and the exponential distribution can be used to describe a population. Sampling distributions can be used to test hypotheses about the population.

Useful Formulas

Uniform probability density function: $f(x)=\begin{cases}1/(b-a) & if\ a\le x\le b\\ 0 & otherwise\end{cases}$	$\mu_X=\frac{1}{\lambda},\quad \sigma_X^2=\frac{1}{\lambda^2}\ \frac{(n-1)S_x^2}{\sigma_X^2}=\frac{\sum\left(X_i-\bar{X}\right)^2}{\sigma_X^2}\sim\chi_{n-1}^2$ Exponential probability density function: $f(x)=\lambda e^{\lambda x}, x>0$
Cumulative distribution function: $F(x)=\begin{cases}0 \text{ if } x\le a\\ (x-b)/(b-a) \text{ if } a<x\le b\\ 1 \text{ if } x>b\end{cases}$	Exponential cumulative distribution function: Properties of exponential distribution: $F(x)=1-e^{-\lambda x}$
Mean and variance for uniform distribution: $\mu_x=\frac{a+b}{2},\quad \sigma^2=\frac{b-1}{\sqrt{12}}$	$P_r(X>a)=e^{-\lambda a}$ $P_r(X<b)=1-e^{-\lambda b}$ $P_r(a<X<b)=e^{-\lambda a}-e^{-\lambda b}$
t distribution: $\frac{\bar{X}-\mu}{S_X/\sqrt{n}}\sim t_{n-1}$	Chi-square distribution: F distribution: $\frac{S_X^2/\sigma_x^2}{S_Y^2/\sigma_Y^2}\sim F_{n-1,m-1}$

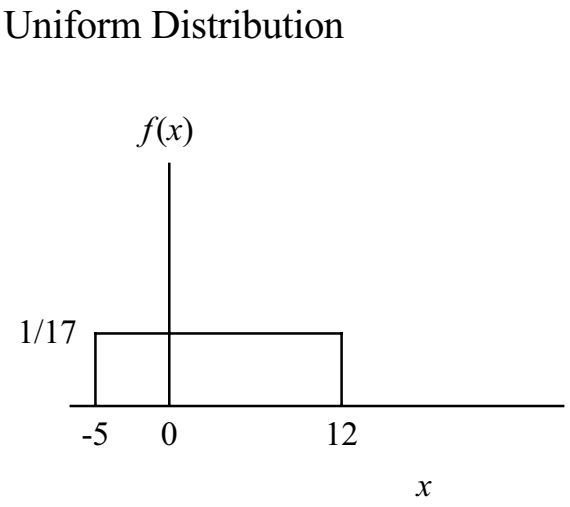
Example Problems

Example 1. Uniform Distribution

Suppose a random variable X can only take on values in the range −5–12 and that the probability that the variable will assume any value within any interval in this range is the same as the probability that X will assume another value in another interval of similar width in the range. What is the distribution of X? Draw the probability density function $f(x)$ for X.

Solution: Since all values of X within the interval have the same probability, X is distributed uniformly.

$$f(x) = \begin{cases} 1/[12-(-5)] & \text{if } -5 \le x \le 12 \\ 0 & \text{otherwise} \end{cases}$$



Example 2 Exponential Distribution

Suppose X has an exponential distribution with mean $E(X)=2.5$. Find the following:

a. $P(X < 5)$
b. $P(X > 7.5)$
c. $P(1 < X < 5)$

Solution:
In order to use the exponential distribution we need to know the value of λ. Because the mean of the exponential distribution is $1/\lambda$, we can substitute $1/E(X)$ for λ.

a. $P(X < 5) = 1 - e^{-(1/2.5)(5)} = .8646$

b. $P(X > 7.5) = e^{-(1/2.5)(7.5)} = .0497$

c. $P(1 < X < 5) = e^{-(1/2.5)(1)} - e^{-(1/2.5)(5)} = .6703 - .1353 = .5349$

Example 3 *t* Distribution

Find the t_α for the following:

a. $\alpha = .05$ and $df = 20$
b. $\alpha = .025$ and $df = 14$
c. $\alpha = .10$ and $df = 8$

Solution: To solve this problem you need to look at the table in the text that provides the values for the *t* distribution.
a. $t_{.05,\,20} = 1.7247$
b. $t_{.025,\,14} = 2.1448$
c. $t_{.10,\,8} = 1.397$

Example 4 Chi-square Distribution

Find the following χ^2 values
a. $\alpha = .025$ and $df = 5$
b. $\alpha = .10$ and $df = 45$
c. $\alpha = .05$ and $df = 10$

Solution: To solve this problem, we use the values for the chi-square distribution listed in the text.

a. $\chi^2_{.025,\,5} = .831$

b. $\chi^2_{.10,\,45} = 33.350$

c. $\chi^2_{.05,\,10} = 3.940$

Example 5 F Distribution

Find the following *F* values when:

a. $\alpha = .05, \nu_1 = 20$ and $\nu_2 = 5$
b. $\alpha = .01, \nu_1 = 5$ and $\nu_2 = 15$
c. $\alpha = .025, \nu_1 = 15$ and $\nu_2 = 5$

Solution: To solve this problem we use the table in the text where the values for the F distribution are given.

a. $F_{.05,\,20,\,10} = 2.77$
b. $F_{.01,\,15,\,5} = 9.72$
c. $F_{.01,\,5,\,15} = 4.56$

Example 6 Uniform Distribution

A professional association gives a standardized test to its new members every year. Suppose the members' grades follow a uniform distribution with 89 points as the maximum and 45 points as the minimum.

a. Find the mean score.
b. Compute the standard deviation of the score.
c. If the passing grade is 72, what percentage of members will pass the course?

Solution:

a. mean of the uniform distribution $= (a+b)/2 = (45+89)/2 = 67$

b. variance of the uniform distribution $= (b-a)/\sqrt{12} = (89-45)/\sqrt{12} = 12.70$

standard deviation $\sqrt{\text{var}} = 3.564$

c. $P(X > c) = (b-c)/(b-a)$ for $a < c < b$

$$P(X > 72) = (89-72)/(89-45) = 0.3864 \text{ or } 38.64\%$$

Example 7 Exponential Distribution

The life of a certain model of car is found to be exponentially distributed with a mean of 75,000 miles. Find the proportion of cars which will have a life greater than 125,000 miles.

Solution: $P(X > 125{,}000) = e^{-(1/75{,}000)(125{,}000)} = .1889$ or 18.89%

Example 8 Chi-square χ^2 Distribution

The stat scores of the students in a university are normally distributed with mean 74 and standard deviation 3. If random sampling a group of 18 students with scores $X_1, \ldots, X_{18}$, find the mean and variance of the sample variance S_X^2.

Solution:

Since $\frac{(n-1)S_X^2}{\sigma_X^2}$ is χ^2 distributed with 17 degree of freedom, thus

$$E\left[\frac{(n-1)S_X^2}{\sigma_X^2}\right] = 17 \text{ and } Var\left[\frac{(n-1)S_X^2}{\sigma_X^2}\right] = 34$$

Thus $E\left[S_X^2\right] = 9$ *and* $Var\left[S_X^2\right] = 9.53$

Example 9 (continued from Example 8) F distribution

The stat scores of the students in another university are also normally distributed with mean 82 and standard deviation 4. The scores of the two universities are independent. If random sampling a group of 25 students with scores Y_1, … , Y_{25}, let S_Y^2 be the sample variance. Find the probability that the sample variance S_Y^2 is larger than S_X^2.

Solution: Since the ratio

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F_{17,24}$$

Thus,

$$P\left(S_X^2 < S_Y^2\right) = P\left(\frac{S_X^2}{S_Y^2} < 1\right) = P\left(\frac{S_X^2}{S_Y^2} \times \frac{16}{9} < \frac{16}{9}\right) = 0.9041$$

Supplementary Exercises

Multiple Choice

1. Sampling distributions
   a. Describe a population.
   b. Describe the distribution of a statistics.
   c. Must always have positive values.
   d. Are not necessarily discrete.
   e. None of the above.

2. Suppose a random variable X follows a uniform distribution in the range 2–7. The probability that X takes on the value of 9 is

   a. 1
   b. 0
   c. 5
   d. 25
   e. Cannot be determined from the information given

3. Suppose a random variable X follows a uniform distribution in the range 2–7. The probability that the variable takes on a value of 9 or less is

   a. 1
   b. 0
   c. 5
   d. 25
   e. Cannot be determined from the information given

4. Suppose a random variable X is normally distributed with mean μ and unknown standard deviation σ. Then the distribution of $(X - \mu)/\sigma$ is

   a. t
   b. Normal
   c. F
   d. χ^2
   e. Binomial

5. The chi-square distribution

   a. Is a continuous distribution described by its degrees of freedom
   b. Can be used to describe the distribution of the sample variance
   c. Is the sum of squared independent standard normal variables
   d. Is positively skewed
   e. All of the above

6. The F distribution is

   a. The ratio of two independent chi-square distributions divided by their respective degrees of freedom
   b. The ratio of two binomial distributions
   c. The ratio of two independent t distributions
   d. The ratio of two independent normal distributions
   e. Always symmetric

7. The exponential distribution is

   a. A discrete distribution
   b. A continuous distribution determined by its parameter λ
   c. The ratio of two independent chi-square distributions
   d. The sum of squared independent normal variables
   e. Described by its degrees of freedom

8. If x follows a uniform distribution in the range 3–11, then the height of the probability density function is

   a. $11+3$
   b. $11-3$
   c. $1/(11+3)$
   d. $1/(11-3)$
   e. $(11-3)^2$

9. The mean for a uniform distribution in the range a–b is
   a. $(b-a)/2$
   b. $(a+b)/2$
   c. $(b-a)/\sqrt{12}$
   d. $(b+a)/\sqrt{12}$
   e. $(b+a)^2$

10. The variance for a uniform distribution in the range a–b is
    a. $(b-a)/2$
    b. $(a+b)/2$
    c. $(b-a)/\sqrt{12}$
    d. $(b+a)/\sqrt{12}$
    e. $(b+a)^2$

11. The probability density function for the exponential distribution, with parameter λ is
    a. $\lambda e^{-\lambda t}$
    b. λ
    c. $\lambda - e^{-\lambda t}$
    d. $1+e^{-\lambda t}$
    e. $e^{-\lambda t}$

12. The cumulative distribution function for the exponential distribution, with parameter λ is

    a. $\lambda e^{-\lambda t}$
    b. λ
    c. $1-e^{-\lambda t}$
    d. $1+e^{-\lambda t}$
    e. $e^{-\lambda t}$

13. If X is uniformly distributed in the range 1–9, then the mean of X is

    a. 1
    b. 9
    c. 5
    d. 2.31
    e. 4.5

14. If X is a uniformly distributed random variable in the range 1–9, then the variance of X is

    a. 1
    b. 9
    c. 5
    d. 2.31
    e. 4.5

15. If X is a uniformly distributed random variable in the range 1–9, then the probability that X takes on a value less than 5 is

    a. 1
    b. 1/2
    c. 1/4
    d. 3/4
    e. 0

16. If X is a uniformly distributed random variable in the range 1–9, then the probability that X takes on a value greater than 3 is

    a. 1
    b. 1/2
    c. 1/4
    d. 3/4
    e. 0

17. If X is an exponentially distributed with $\lambda = 5.7$, then the mean of X is

    a. 1
    b. 1/2
    c. 1/4
    d. 3/4
    e. 0

18. If X is an exponentially distributed random variable with $\lambda = 5.7$, then the variance of X is

    a. 1
    b. 1/2
    c. 1/4
    d. 3/4
    e. 0

19. If X is an exponentially distributed random variable with $\lambda = 5.7$, then the probability that X takes on a value less than 1/2 is

    a. .6321
    b. .3679
    c. .7358
    d. .2642
    e. 0

20. If X is an exponentially distributed random variable with $\lambda = 5.7$, then the probability that X takes on a value greater than 1/2 is

    a. .6321
    b. .3679
    c. .7358
    d. .2642
    e. 0

21. Which of the two chi-square distributions shown below (A, B, or C) has the larger degrees of freedom?

True/False (If False, Explain Why)

1. The *t* distribution is described by its mean, variance, and degrees of freedom.
2. The exponential distribution can be used to examine the probability of waiting time between consecutive occurrences.
3. If x is uniformly distributed in the range 1–12, then the probability that x takes on a value less than 1 is 1.
4. The chi-square distribution is the sum of any squared independent random variables.
5. As the sample size increases, the *t*-distribution gets closer to the normal distribution.
6. The mean and variance for the exponential distribution are both equal to λ.
7. The normal distribution is leptokurtic (coefficient of kurtosis > 3).
8. The chi-square distribution is a positively skewed distribution.
9. The *F* distribution is described by its mean and variance.
10. When X is uniformly distributed in the interval a to b, X is more likely to take on values above the interval $[(a+b)/2, b)$.

Questions and Problems

1. Suppose a random variable, X, follows a uniform distribution in the range 25–37.

    a. Find the height of the probability density function.
    b. Find the probability that x will be less than 30.
    c. Find the probability that x will lie between 29 and 32.

2. Suppose X has an exponential distribution with $\mu_X = 1/2$. Find the following probabilities:

    a. $P_r(X < 1/2)$

    b. $P_r(1/3 < X < 2/3)$

3. Find the t_α for the following:
   a. $\alpha = .10$ and df = 5
   b. $\alpha = .01$ and df = 10
4. Find the χ^2 values when
   a. $\alpha = .05$ and df = 30
   b. $\alpha = .10$ and df = 20
5. Find the *F* values for the following:
   a. $\alpha = .5, v_1 = 10$ and $v_2 = 20$
   b. $\alpha = .5, v_1 = 20$ and $v_2 = 10$
   c. $\alpha = .01, v_1 = 30$ and $v_2 = 20$
6. The Scholastic Aptitude Test (SAT) has a minimum score of 400 and a maximum score of 1600. Assume the scores follow a uniform distribution.
   a. Find the mean SAT score.
   b. Find the variance of the distribution.
   c. Find the percentage of students who will receive above 1200 on the test.
7. The life of Bright White light bulbs is found to be exponentially distributed with a mean of 2500 h. Find the proportion of light bulbs that will have a life greater than 3000 h.
8. A firm is interested in knowing the length of a warranty to issue. From previous experience, it knows that the products' life follows an exponential distribution with mean $1/\lambda = 10$ years. If the firm wants to be sure that no more than 10% of the products will be returned during the warranty period, what should be the length of the warranty?
9. Random variables $Z_1, \ldots, Z_{23}$ are *i.i.d.* standard normal random variables. What is the distribution of the random variable $\sum_{i=1}^{23} Z_i^2$? What is its mean and standard deviation? What is the probability the random variable is larger than 3.25?
10. Suppose random variables $Y_1, \ldots, Y_{15}$ are *i.i.d.* standard normal random variables, and $Y_1, \ldots, Y_{15}$ are independent of the standard normal random variables $Z_1, \ldots, Z_{23}$ in Problem #9. What is the constant *C* so that the random variable is *F*-distributed? What is the degree of freedom of the aforementioned *F* distribution?

$$\frac{S_X^2 / \sigma_X^2}{S_Y^2 / \sigma_Y^2} \sim F_{17,24}$$

Answers to supplementary exercises

Multiple Choice

1. b	6. a	11. a	16. d
2. b	7. b	12. c	17. b
3. a	8. d	13. c	18. b
4. b	9. b	14. d	19. a
5. e	10. c	15. b	20. b

True/False

1. True.
2. True.
3. False. The probability that x takes on a value outside its range is 0.
4. False. The chi-square distribution is the sum of squared independent standard normal random variables.
5. True.
6. False. The mean and standard deviation for the exponential distribution are equal to $1/\lambda$.
7. False. The normal distribution is meso kurtic (CK = 3).
8. True.
9. False. The F distribution is described by its degrees of freedom of its numerator and denominator.
10. False. If X is uniformly distributed, then all values in the interval are equally likely.

Questions and Problems

1. a. $1/(37 - 25)$
   b. $(30 - 25)/(37 - 25) = .417$
   c. $(32 - 25)/(37 - 25) - (29 - 25)/(37 - 25) = .25$
2. a. .632
   b. .250
3. a. 1.476
   b. 2.764
4. a. 44.773
   b. 28.412
5. a. 2.77
   b. 2.35
   c. 2.55

6. a. $(1600 + 400)/2 = 1000$
   b. $(1600 - 400)/\sqrt{12} = 346.41$
   c. $(1600 - 1200)/(1600 - 400) = .333 = 33.3\%$

7. .301

8. Pr(x > a) = .90, where a is the length of the warranty.

$$P_r(x > a) = e^{-.1\,a} = .90$$

Solving for a, we get a = 1.054 years.

Chapter 10
Estimation and Statistical Quality Control

Chapter Intuition

In Chaps. 1–4, we learned about descriptive statistics used to describe data. In Chaps. 5–9, we learned about probability distributions. This Chapter is about the inferences of two populations using descriptive statistics, which concentrates on estimation. In Chap. 11, hypothesis testing will be introduced. Estimation deals with making "educated guesses" about unknown population parameters. For example, advertisers may be interested in knowing what percentage of TV viewers watch a certain television show. Because of the costs involved in surveying every television viewer, the actual percentage of viewers watching the show will never be known. However, using the technique of sampling, discussed in Chap. 8, as well as the concept of estimation, it will be possible to make an "educated guess" about the actual percentage of viewers.

Estimation is a "guess" of an unknown parameter such as the population mean or population proportion. Estimating a parameter is like shooting at a target with a gun. The target is the parameter of interest, the estimator is your gun, and the data are your ammunition. Point estimation is like firing a single shot at the target; you only have one chance to hit the target. On the other hand, interval estimation is like firing a series of shots at a neighborhood around the target, using a machine gun. With interval estimation, we have a greater opportunity of hitting the target because we are firing at a neighborhood around the target.

In order to determine whether an estimator is good or not, we use the three criteria: unbiasedness, efficiency, and consistency. The unbiasedness criterion asks if the expected value of the estimator is the true parameter value. In our gun shooting analogy, an unbiased estimator would be a rifle whose scope is aligned correctly. You may not always hit the target, but your shots should be within a neighborhood around the target: if you shoot several times, the averaged shot should be right at the target. On the other hand, a biased estimator would be like a gun whose scope is incorrectly aligned to the left or the right of the target.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_10

The efficiency criterion refers to the standard error (dispersion) of the estimator. If the estimator is efficient, the standard error will be small. Using our gun and target analogy, efficiency tells us the dispersion of the shots are within a small neighborhood around the target If the shots are far away from each other then the estimator is not efficient.

The final criterion we use is consistency. Consistency implies that the estimator will approach the true parameter value as the sample size increases. This is an important property for an estimator because it means eventually we obtain the true parameter value as long as we have large enough sample size.

Chapter Review

Because we are estimating an unknown population parameter, using an estimator that is stochastic, we can use the sampling and distribution theory discussed in previous chapters to provide a statistical interval for the true population parameter.

1. Calculation of the ***population parameters*** using samples is called ***estimation***. The formulas used to do the calculations are the ***estimators***. The numbers generated by the estimators are called the ***estimates***. A ***point estimate*** is a single number that is obtained from the estimator.
2. Four important criteria for evaluating estimators are:
   a. ***Unbiased.*** The expected value of the estimator is the population parameter.
   b. ***Efficiency***. An estimator is efficient if it has a small variance.
   c. ***Consistency***. An estimator is consistent if the estimator will approach the true parameter value as the sample size increases to infinity.
   d. ***Sufficiency***. A sufficient statistic is an estimator that contains all the information a sample contains about the parameters.
3. ***Interval estimation*** refers to the estimation of the parameter of interest with a ***confidence interval***. In other words, because the true value of the population parameter is unknown, we provide an interval that we believe may contain the true value. A 90% confidence interval implies that if the sampling procedure is repeated 100 times, we can expect the confidence interval to contain the true parameter 90 times. The width of the interval depends on several factors:
   a. The ***confidence level***. For example, if we want to be 99% confident that the confidence interval contains the true parameter value, we need a wider interval than if we only want to be 90% confident.
   b. The variance of the estimator. If the distribution has greater dispersion, the estimator will contain less information concerning the population parameter and the confidence interval will be wider.
   c. The sample size. The greater the size of the sample, the more information we have and the smaller the width of the confidence interval.
4. When the population variance is used, the standard normal distribution is used to compute the confidence intervals for the population mean.

5. When the population variance is *unknown*, the Student's *t* distribution is used to compute the confidence interval for the population mean.
6. Confidence intervals can also be computed for population proportions.
7. The chi-square distribution is used to compute the confidence interval for the population variance.

Useful Formulas

Unbiasedness:

$$E(\hat{\theta}) = \theta$$

Efficiency:

$$\text{Relative Efficiency} = \frac{V(\theta_2)}{V(\theta_1)}$$

Confidence intervals for the population mean:

a. Population variance known:

$$1-\alpha = P_r\left[\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

b. Population variance unknown:

$$1-\alpha = P_r\left[\bar{X} - t_{n-1,\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + t_{n-1,\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right]$$

Confidence interval for a population proportion:

$$1-\alpha = P_r\left[\hat{P}_x - z_{\alpha/2}\sqrt{\frac{\hat{P}_x(1-\hat{P}_x)}{n}} < P < \hat{P}_x + z_{\alpha/2}\sqrt{\frac{\hat{P}_x(1-\hat{P}_x)}{n}}\right]$$

Confidence interval for the population variance:

$$1-\alpha = P_r\left[\frac{(n-1)s_x^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s_x^2}{\chi^2_{n-1,(1-\alpha)/2}}\right]$$

Example Problems

Example 1 Confidence Interval for the Population Mean

A real estate agent in Wisconsin is interested in the mean home price in the state. A random sample of 40 homes shows a mean home price of $ 98,115 and a sample standard deviation of $ 27,100. Construct a 90% confidence interval for the mean home price.

Solution: When we construct a 90% confidence interval, we are interested in finding a range or interval in which the "true" mean home price would fall 90% of the time. The formula for a confidence interval when the population variance is known is

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

For a 90% confidence interval, $\alpha = .10$, and so $z_{\alpha/2} = 1.645$.

$$98{,}115 \pm 1.645(27{,}100/0)$$

So the interval is from $91,066 to $105,164.

Example 2 Confidence Interval for the Population Mean with Known and Unknown Population Standard Deviation

A quality control expert believes that the life of her firm's tires is normally distributed with a standard deviation of 15,225 miles. A random sample of 10 tires gives the following mileage on the life of the tires.

55,000	48,000	73,000	51,000	77,000	52,000	38,000	41,000	68,000	62,000

Construct a 95% confidence interval for the mean life of the tires.

Solution: First, we need to find the sample mean life of the sample. When the sample size is small (less than 30) and the standard deviation of the population is unknown, we can no longer use the normal distribution to construct our confidence intervals. In this case, we need to use the sample standard deviation *s* in place of the population standard deviation σ, and the formula for the confidence interval involves the *t* distribution as

$$\begin{aligned}\bar{x} &= (55,000 + 48,000 + 73,000 + 51,000 + 77,000 + 52,000 \\ &\quad + 38,000 + 41,000 + 68,000 + 62,000)/10 \\ &= 56,500\end{aligned}$$

$$\begin{aligned}\text{where } s^2 &= [(55,000 - 56,500)^2 + (48,000 - 56,500)^2 + (73,000 - 56,500)^2 \\ &\quad + (51,000 - 56,500)^2 + (77,000 - 56,500)^2 + (52,000 - 56,500)^2 \\ &\quad + (38,000 - 56,500)^2 + (41,000 - 56,500)^2 + (68,000 - 56,500)^2 \\ &\quad + (62,000 - 56,500)^2]/9 = 173,611,111\end{aligned}$$

s = 13,176, and the confidence interval is $\bar{X} \pm t_{n-1,\alpha/2} s/\sqrt{n}$

$$56,500 \pm 2.26(13,176/0)$$

So the interval is from 47,083 to 65,916.

Example 3 Confidence Interval for a Population Proportion

A random sample of 500 residents of a city of 7,150 residents shows that 72 % believe that the police commissioner is doing a good job of fighting crime. Construct a 99 % confidence interval for the proportion of all residents that believe that the police commissioner is doing a good job against crime.

Solution: In this example, we want to find the confidence interval for a proportion. The formula for the confidence interval for a proportion is

$$\bar{p} \pm z_{\alpha/2}[\bar{p}(1 - \bar{p})/n]^{1/2}$$

$$.72 \pm 2.81(.72(1 - .72)/500)^{1/2}$$

So the interval is from .6635 to .7764

Example 4 Point Estimation

Construct point estimates from the following situations:

a. A labor union randomly samples 50 of its members and finds that 29 *oppose* the new contract. Estimate the proportion of all workers who *favor* the new contract.

b. A statistics professor randomly samples 75 students in her class and finds that 42 do not know the meaning of a confidence interval. Estimate the proportion of all students in her class who cannot define a confidence interval.

Solution:

a. Since 29 of the 50 workers oppose the contract, 21 must favor the contract. The proportion favoring the contract is 21/50 = .42.
b. A point estimate for the proportion of students who do not know the meaning of a confidence interval is 42/75 = .56.

Example 5 Confidence Intervals for a Population Mean

A random sample of 50 observations from a population yielded the following summary statistics:

$$\sum x = 825 \qquad \sum x^2 = 22,242$$

Construct a 95% confidence interval for the population mean μ.

Solution: To solve this problem we need to compute the sample mean and sample standard deviation. Even though the population standard deviation is not known, we can use the normal distribution because we have a large enough sample.

$$\bar{x} = \sum \frac{x}{n} = \frac{825}{50} = 16.5$$

$$\begin{aligned} s^2 &= [\Sigma x^2 - (\Sigma x)^2/n]/(n-1) \\ &= [22,242 - (825)^2/50]/(50-1) \\ &= 176.11 \end{aligned}$$

$$s = 13.27$$

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

$$16.5 \pm 1.96(13.27/0)$$

So the interval is from 12.82 to 20.18.

Example 6 Confidence Interval for a Population Mean

A random sample of 200 observations from a population yielded the following summary statistics:

$$\sum x_i = 1{,}500 \qquad \sum (x_i - \bar{x})^2 = 31{,}000$$

Construct a 99% confidence interval for the population mean μ.

Solution:

$$\bar{x} = \frac{1{,}500}{200} = 7.5$$

$$s^2 = 31{,}100/199 = 156.28$$

$$s = 12.5$$

$$\bar{x} \pm Z_{\alpha/2}\sigma/\sqrt{n}$$

$$7.5 \pm 2.575(12.5/\sqrt{200})$$

So the interval is from 5.22 to 9.78.

Example 7 Confidence Interval for a Population Variance

Suppose a random sample of 51 boxes of cereal is taken and the sample variance is found to be 4.3 ounces. Construct a 90% confidence interval for the population variance.

Solution: The formula for the confidence interval for the variance is

$$1 - \alpha = P_r\left[\frac{(n-1)s_x^2}{\chi^2_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)s_x^2}{\chi^2_{n-1,(1-\alpha)/2}}\right]$$

$$[(51-1)4.3]/71.42 \text{ to } [(51-1)4.3]/32.57$$

So the interval is from 3.01 to 6.60.

Example 8 Sample Size

The manager of a local soft drink factory wants to construct a 95 % confidence interval to estimate the average amount of soda pumped into the 12-ounce cans. From previous experience, he is confident that the standard deviation of the soda is 0.02. If he wants to control the width of the confidence interval to ±0.005, how many cans of soda should he sample in each experiment?

Solution: The 95 % confidence interval is

$$\bar{x} \pm z_{\alpha/2}\sigma/\sqrt{n}$$

$$z_{\alpha/2}\sigma/\sqrt{n} = 1.96(.02/\sqrt{n}) = .005$$

Solving for n yields n = 61.47. Therefore, he should collect 62 cans for each experiment.

Example 9 Sample Size for Estimation of a Population Proportion

A local newspaper wants to estimate the proportion of voters who favor a tax increase to fund a recycling project. The editor wants to estimate the proportion with a 95 % confidence interval. In addition, she wants the error margin to be within 3 %. How large a sample should she take?

Solution: The error margin is

$$\pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

which should be no more than ±3 %

$$z_{.05/2} = 1.96$$

The proportion, p, that will generate the widest interval (worst case scenario) is 1/2, so,

$$1.96\sqrt{\frac{.5(1-.5)}{n}}=3\%$$

Solving for n yields n = 1076.1. Therefore, she should sample at least 1068 voters.

Supplementary Exercises

Multiple Choice

1. A point estimate is

   a. a single value that is used to estimate the population parameter.
   b. a range of values used to estimate the population parameter.
   c. always unbiased.
   d. always efficient.
   e. sufficient to establish unbiasedness.

2. An interval estimate is

   a. a single value that is used to estimate the population parameter.
   b. a range of values used to estimate the population parameter.
   c. always unbiased.
   d. always efficient.
   e. always a sufficient statistic.

3. The larger the population variance the

   a. narrower the width of the confidence interval, other things being equal.
   b. wider the width of the confidence interval, other things being equal.
   c. larger the mean.
   d. smaller the mean.
   e. larger the point estimate.

4. Other things being equal, the width of a 90% confidence interval will be

   a. wider than a 95% confidence interval.
   b. narrower than a 95% confidence interval.
   c. may be wider or narrower depending on the population variance.
   d. the same width as a 95% confidence interval.
   e. wider than a 99% confidence interval.

5. An unbiased estimator

   a. has the smallest variance among all estimators.
   b. is always the best estimator.

c. has an expected value equal to the true parameter value.
d. always generates the true value of the parameter.
e. is always sufficient.

6. An efficient estimator

a. has a small variance.
b. gets closer to the true parameter value as the sample size increases.
c. has an expected value equal to the true parameter value.
d. always generates the true value of the parameter.
e. is always sufficient.

7. A consistent estimator

a. has the smallest variance among all estimators.
b. gets closer to the true parameter value as the sample size increases.
c. has an expected value equal to the true parameter value.
d. is always sufficient.
e. is always the best estimator.

8. When constructing interval estimates for the population mean we should use the

a. z distribution when the population variance is unknown and the t distribution when the population variance is known.
b. z distribution when the population variance is known and the t distribution when the population variance is unknown and is estimated with a small sample.
c. z distribution whether the population variance is known or not, if the sample size is large enough.
d. chi-squared distribution.
e. F distribution.

9. To construct the confidence interval for the population variance, we should use the

a. normal distribution.
b. t distribution.
c. F distribution.
d. chi-square distribution.
e. binomial distribution.

10. A sufficient statistic

a. is consistent.
b. is unbiased.
c. uses all information a sample contains about the parameter to be estimated.
d. is always efficient.
e. approaches the true parameter value as the sample size decreases.

11. The confidence interval for the population mean when the population variance is known is

    a. $P_r(x \prime\prime z)$

    b. $\bar{X} \pm Z_{\alpha/2}\sigma/\sqrt{n}$

    c. $P_r(x \geq z)$

    d. 1

    e. 0

12. The width of a $(1-\alpha)$ confidence interval for a population mean is

    a. $2\sigma z_{\alpha/2}$

    b. $\sigma\, z_{\alpha/2}$

    c. $2\sigma/\sqrt{n}\; z_{\alpha/2}$

    d. $\sigma/\sqrt{n}\; z_{\alpha/2}$

    e. $2\sigma^2$

13. Suppose a pollster surveys 300 people to see if they favor prayer in school. Of the 300 people surveyed, 127 favor prayer in school. An estimator of the proportion of people who favor prayer in school would be

    a. 127

    b. 300

    c. 127/300

    d. $127/300[1-(127-300)]/300$

    e. 300/127

14. An estimate of the variance of people who favor prayer in school in question 13 would be

    a. 127

    b. 300

    c. 127/300

    d. $127/300[1-(127-300)]/300$

    e. 300/127

15. Other things being equal, width of a confidence interval will be

    a. wider if the population variance is known.
    b. narrower if the population variance is unknown.
    c. wider if the population variance is unknown.
    d. the same width whether the population variance is known or not.
    e. wider if the mean is unknown.

16. Which best describes the lower endpoint of a confidence interval?

    a. Point estimate
    b. Margin of error
    c. Point estimate plus margin of error
    d. Point estimate minus margin of error
    e. None of the above

17. Which value will be at the center of a confidence interval?

    a. Population mean
    b. Population mean standard deviation
    c. Margin error
    d. Sample mean
    e. None of the above

18. What is the relationship between a 95 % confidence interval and a 99 % confidence interval from the same sample?

    a. The 95 % interval will be wider
    b. The 99 % interval will be wider
    c. Both intervals have the same width
    d. Inconclusive

True/False (If False, Explain Why)

1. Other things being equal, the greater the population variance, the narrower the width of the confidence interval.
2. For a given population variance and sample size, the greater the confidence level, the greater the width of the interval.
3. When the population variance is unknown and the sample size is small, we should use the t distribution to construct the confidence interval for the population mean.
4. When the population variance is unknown and the sample size is large, we should use the t distribution to construct the confidence interval.
5. Other things being equal, increasing the sample size increases the width of the confidence interval.
6. To construct a confidence interval for the population variance, we should use the chi-square distribution.

7. A confidence interval is really just a range of values constructed by sampling that will cover the true value $1-\alpha$ percent of the time.
8. A larger population variance increases the width of the confidence interval because there is more dispersion around the population mean.
9. We can obtain an unbiased estimate of mean earnings per person in a city by randomly sampling people who live on the same street.
10. Consistent estimators reduce the need to increase the sample size.
11. To construct a $1-\alpha$ confidence interval for the population mean with a known population variance, we should use z_{α}.
12. 12. To construct a $1-\alpha$ confidence interval for the population mean with an unknown population variance, we should use t_{α}.
13. When the population variance is unknown, it is inappropriate to use the standard normal distribution to construct a confidence interval.
14. A sufficient estimator utilizes all the information a sample contains about the parameters.

Questions and Problems

1. Suppose a golfer is interested in the average distance he hits a five iron. He hits 100 shots with the five iron and finds the sample mean to be 175 yards. If the population variance for his shots is known to be 64, construct a 95% confidence interval for the population mean.
2. Suppose a golfer is interested in the average distance he hits a five iron. He hits 25 shots with the five iron and finds the sample mean to be 175 yards. Assume that the population variance for his shots is unknown, but the sample variance is estimated to be 64, construct a 95% confidence interval for the population mean.
3. A random sample of 200 residents of Home Town, U.S.A. shows that 65% believe the superintendent of schools is doing a good job. Construct a 95% confidence interval for the proportion of all residents that believe the superintendent is doing a good job.
4. Suppose a college professor randomly samples 125 graduating seniors and finds that 47 have job offers. Estimate the proportion of students who have job offers.
5. A random sample of 30 radial tires is taken and found to have sample standard deviation of 2,352 miles. Construct a 95% confidence interval for the population variance.
6. To obtain an estimate of the proportion of 'full time' university students who have a part time job in excess of 20 h per week, the student union decides to interview a random sample of full time students. They want the length of their 95% confidence interval to be no greater than 0.1. What size sample, should be taken?
7. An industrial designer wants to determine the average amount of time it takes an adult to assemble an "easy to assemble" toy. A sample of 16 attempts yielded an average time of 19.92 min, with a sample standard deviation of 5.73 min. Assuming the assembly times are normally distributed. Provide a 95% confidence interval for the mean assembly time.

Answers to Supplementary Exercises

Multiple Choice

1. a	6. a	11. b	16. d
2. b	7. b	12. c	17. d
3. b	8. d	13. c	18. b
4. b	9. d	14. d	
5. c	10. c	15. c	

True/False

1. False. The confidence interval will be wider.
2. False. The confidence interval will be narrower.
3. True
4. False. When the sample size is large enough, we can also use the z-distribution, although the t-distribution is also correct.
5. False. The confidence interval gets narrower.
6. True
7. True
8. True
9. False. Because people with similar incomes tend to live in the same neighborhood, the mean earnings will be biased.
10. False. The advantage of a consistent estimator appears when the sample size is large.
11. False. We should use $z_{\alpha/2}$.
12. False. We should use $t_{\alpha/2}$.
13. False. If the sample size is large enough, we can use the standard normal distribution even if the population variance is unknown.
14. True

Questions and Problems

1. $\bar{x} \pm z_{\alpha/2} \dfrac{\sigma}{\sqrt{n}}$

$$175 \pm 1.96 \frac{8}{\sqrt{100}}$$

So the interval is from 173.43 to 176.57.

2. Because the sample size is small and the variance is unknown, we must use the t-distribution to construct the confidence interval.

$$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

$$175 \pm 2.064 \frac{8}{\sqrt{25}}$$

So the interval is from 171.70 to 178.30.

3. $\bar{p} \pm z_{\alpha/2} \left[\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \right]$

$$.65 \pm 1.96 \left[\sqrt{\frac{.65(1-.65)}{200}} \right]$$

So the interval is from .584 to .716.

4. $\hat{p}=47/125=.376$

5. $1-\alpha = P_r \left[\frac{(n-1)s_x^2}{\chi^2_{n-1,\,\alpha/2}} < \sigma^2 < \frac{(n-1)s_x^2}{\chi^2_{n-1,(1-\alpha)/2}} \right]$

$$\left(\frac{(30-1)(2352^2}{42.557} \quad \frac{(30-1)(2352^2}{17.708} \right)$$

So the interval is from 3,769,655 to 9,059,477.

6. $1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}} = 0.1 \rightarrow n = 96.$

7. The confidence interval is

$$\bar{x} \pm t_{n-1,\alpha/2} s/\sqrt{n}$$

$$= 19.92 \pm 2.13 \times \frac{5.73}{\sqrt{16}} = (16.87, 22.97)$$

Chapter 11
Hypothesis Testing

Chapter Intuition

The second part of inferential statistics is *hypothesis testing*. In hypothesis testing, unlike estimation, we are not interested in making a guess about the true value of some parameter. Instead, we are interested in deciding if enough evidence exists to overthrow what we believe is true.

This concept can best be understood by the following example. Suppose a scientific study indicates that a cigarette containing more than 20 units of nicotine is extremely dangerous to smokers. Our hypothesis is that the average amount of nicotine per cigarettes is 20 units or more. In estimation, we would collect a sample to "guess" the average nicotine content. In hypothesis testing, we collect a sample of cigarettes to decide whether a cigarette contains more than 20 units of nicotine or not. Because, high dose of nicotine is dangerous, the hypothesis is $\mu \geq 20$. Unless we have sufficient evidence to prove the hypothesis is wrong, we want to believe that the cigarettes have a level of nicotine that is dangerous. The mean level of nicotine for the sample of cigarettes is used to conduct the test. If the averaged level of nicotine is much smaller than 20, we have enough evidence to say that the hypothesis $\mu \geq 20$ is wrong. Otherwise, we will accept the hypothesis.

In a hypothesis test, we consider a pair of hypotheses, the *null hypothesis* (H_0) and the *alternative hypothesis* (H_1). The null hypothesis refers to a general statement or default position unless we have enough evidence to reject it. On the other hand, the alternative hypothesis is the rival hypothesis. In the example above, the null hypothesis is that cigarettes have a dangerous level of nicotine; alternative hypothesis is that cigarettes have a safe level of nicotine.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_11

Chapter Review

1. *Hypothesis testing* is a systematic procedure used in testing the correctness of assumptions made about a population parameter. The ***null*** *hypothesis* is the hypothesis which conforms to the status quo and is what we are trying to disprove. In other words, if we do not have enough evidence to reject the null hypothesis, we are willing to live with it. For example, in our legal system, the null hypothesis is that a person is innocent. If we do not have enough evidence to show the person is guilty "beyond a reasonable doubt," we consider the person as innocent. The alternative hypothesis is the hypothesis we are trying to prove if we have substantial evidence to reject the null hypothesis.
2. In conducting hypothesis testing, it is possible to make two kinds of errors. If we reject the null hypothesis when it is actually true, we have made a *type I error*. If we accept the null hypothesis when it is actually false, we have made a *type II error*.
   There are several ways to set up an alternative hypothesis. If we specify only one value for the population parameter, we are consider a *simple hypothesis*. If we set up a range of values for the population parameter, we are consider a *composite hypothesis*.
3. Steps for hypothesis testing:
   a. Set up the null and alternative hypotheses. The status quo statement should be the null hypothesis. The statement we are trying to prove should be the alternative hypothesis.
   b. Study the parameter of interest and select an appropriate test statistics to conduct the test.
   c. Carefully examine the type I and type II errors, then decide on an α level ((the maximum probability of making a type I error that is tolerable). Check the appropriate probability distribution table to obtain the critical value of the rejection region of the test statistics. How large an α level we choose depends on the costs of making a type I error.
   d. Compute the test statistic chosen in step (b) and compare it with the rejection region in step (c) in order to determine if we accept or reject the null hypothesis.
4. The p-value of a test is the estimated probability of rejecting the null hypothesis (H_0) when that hypothesis is true
5. The *power of a test* refers to the ability of the test to reject the null hypothesis. The more powerful a test, the more likely it is to reject the null hypothesis.
6. *One-tailed tests* are used when the range of values for rejecting the null hypothesis lies entirely on one side of the null hypothesis values. For example, if we were to test the null hypothesis that the average nicotine content is greater than or equal to 20 units, the evidence that leads to the rejection of the null hypothesis should fall entirely on the extreme left-hand side of the null hypothesis value.
   Two-tailed tests are used when the parameter value for the alternative hypothesis can lie on either side of the parameter value in the null hypothesis. For example, if we are interested in whether a coin is fair, then getting a percentage of heads

which is either significantly higher or significantly lower than 50 % will constitute enough evidence to reject the null hypothesis that $p = 50\%$.

7. In deciding which test statistic is appropriate for the test you are performing, the following rules apply:

   a. To conduct a test on the population mean, we use the z-statistic or student's t-statistic depending on the following:

      i. If the population variance is *known* or if the sample size is *large* ($n > 30$), we use the z-statistic.
      ii. If the population variance is *unknown* and the sample is *small* ($n < 30$), we use the t-distribution.

   b. To test that the population variance is equal to some value, use the chi-squared statistic.

   c. To test the equality of two variances, use the F-statistic.

Useful Formulas

Chi-square distribution: $\frac{(n-1)S_x^2}{\sigma_x^2} = \chi^2_{n-1}.$	Testing one mean-unknown variance with small sample: $\frac{\bar{x} - \mu_0}{\frac{s_x}{\sqrt{n}}} = t_{n-1}.$
Testing one mean-known variance: $\frac{\bar{x} - \mu_0}{\frac{\sigma_x}{\sqrt{n}}} = z.$	Testing a proportion: $\frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{(\bar{x}_n - \bar{x}_2) - (\mu_1 - \mu_2)}}} = z$ $\frac{}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = z.$
Testing one mean-unknown variance with large sample: $\frac{\bar{x} - \mu_0}{\frac{s_x}{\sqrt{n}}} = z$	
Testing the difference between two means-small sample: $\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1 - n_2 - 2)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = t.$	
Testing the difference between two means-small sample:	

Example Problems

Example 1 Critical Values for Standard Normal Distribution

Find the critical values for the following standard normal distributions.

a. Two-tailed test for $\alpha = 0.01$
b. One-tailed test for $\alpha = 0.025$
c. Two-tailed test for $\alpha = 0.05$
d. One-tailed test for $\alpha = 0.01$
e. Two-tailed test for $\alpha = 0.10$

Solution: For a two-tailed test, we want the critical value for $z_{\alpha/2}$.

a. $z_{0.01/2} = z_{0.005} = 2.57.$

b. $z_{0.025} = 1.96.$

c. $z_{0.05/2} = z_{0.025} = 1.96.$

d. $z_{0.01} = 2.33.$

e. $z_{0.10/2} = z_{0.05} = 1.645.$

Example 2 One-Tailed Test for the Mean

A college professor randomly selects 25 freshmen to study the mathematical background of the incoming freshman class. The average SAT score of these 25 students is 565 and the standard deviation is estimated to be 40. Using this information, can the professor reject the null hypothesis that the average test score is 550 or less? Use a 5% level of significance.

Solution:

Step 1: Set up the hypotheses.

$$H_0 : \mu \leq 550.$$

$$H_1 : \mu > 55.$$

Step 2: Select a test statistic.

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t \qquad df = n - 1 = 25 - 1 = 24.$$

Step 3: Establish the rejection area.
From the *t*-table, we can determine that $t_{5\%,\,24} = 1.71$.

Step 4: Carry out the test

$$\frac{565 - 550}{40/\sqrt{25}} = 1.875 > 1.71 \text{Reject } H_0.$$

Note: In step 2, we used a t-test because the standard deviation was *unknown* and was estimated with a *small* sample (n = 25). If $n - 1 \geq 30$, we may use the normal distribution to approximate the t-distribution.

Example 3 Testing the Difference of Two Means-Small Sample Case

Suppose the college professor in Example 2 wants to know whether the mathematical aptitude of freshmen has improved over the past year. He randomly draws 25 SAT mathematics scores from last year's freshman class and obtains a sample mean of 560 and a sample standard deviation of 35. Compare the results from last year with the results from this year. Do the data offer enough evidence for rejecting the null hypothesis that there is no improvement? Use a 5% level of significance.

Solution:

Step 1: Set up the hypotheses

$$H_0 : \mu_{new} - \mu_{old} \leq 0.$$

$$H_1 : \mu_{new} - \mu_{old} > 0.$$

Step 2: Select a test statistic.

$$df = 25 + 25 - 2 = 48.$$

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - n_2 - 2)}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = t.$$

Step 3: Establish the rejection area.
From the *t*-table, we can determine that $t_{5\%,\,48} = 1.64$.

Step 4: Carry out the test.

$$\frac{(565-560)-0}{\sqrt{\frac{(25-1)40^2+(25-1)35^2}{(25-25-2)}\left(\frac{1}{25}+\frac{1}{25}\right)}}=0.47<1.64.$$

The evidence is not strong enough to reject the null hypothesis.

v: The *t*-test is used because $n_{new}=n_{old}=25$ and these are small samples.

Example 4 Test of the Difference Between Two Means-Large Sample

In example 3, assume that the college professor obtained a grant to study whether there was an improvement in the mathematics ability of freshmen. Using this grant money, he is able to obtain larger samples. The results are compiled in the following table.

	Last year	This year
Size of sample	125	100
Average score	562	568
Standard deviation	40	42

Can the professor reject the null hypothesis that the average scores improved? Use a 5% level of significance.

Solution:

Step 1: Set up the hypotheses.

$$H_0:\mu_{new}-\mu_{old}\le 0.$$

$$H_1:\mu_{new}-\mu_{old}>0.$$

Step 2: Select a test statistic.

$$\frac{(\bar{x}_{new}-\bar{x}_{old})-(\mu_{new}-\mu_{old})}{\sqrt{\frac{\sigma^2_{new}}{n_{new}}+\frac{\sigma^2_{old}}{n_{old}}}}\sim z.$$

Step 3: Establish the rejection area.

From the standard normal table, we can determine *that* $z_{5\%}=1.64$.

Step 4: Carry out the test.

$$\frac{(568-562)-0}{\sqrt{\frac{40^2}{125}+\frac{42^2}{100}}}=1.087<1.64.$$

The evidence is not strong enough to reject the null hypothesis.

Example 5 Test of a Proportion

In a recent poll, 800 voters were randomly sampled and asked whether they would approve of the building of an incinerator in the state. According to the poll, 450 voters approved and 350 voters disapproved. Is there enough evidence to argue that more than 50% of the voters support the building of the incinerator? Use a 5% level of significance.

Solution:

Step 1: Set up the hypotheses.

$$H_0 : p \le 0.5.$$

$$H_1 : p > 0.5.$$

Step 2: Select a test statistic.

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

Step 3: Establish the rejection region.
From the standard normal table, we can determine that $z_{5\%}=1.64$.

Step 4:

$$\frac{450/800-0.5}{\sqrt{\frac{0.5(1-0.5)}{800}}}=3.33>1.64.$$

So we can reject the null hypothesis.

Example 6 Testing the Equality of Two Proportions

A survey shows that 360 out of 500 females and 600 out of 800 males enjoy a game show. Based on this evidence, would you argue that the ratio of female and male viewers who enjoy the game show differs? Use a 5% level of significance.

Solution:

Step 1: Set up the hypotheses.

$$H_0 : p_f - p_m = 0.$$

$$H_1 : p_f - p_m \neq 0.$$

Step 2: Select a test statistic.

$$\frac{(\bar{p}_f - \bar{p}_m) - (p_f - p_m)}{p_c(1-p_c)\left(\frac{1}{n_f} + \frac{1}{n_m}\right)} = z,$$

where p_c = combined proportion.

Step 3: Establish the v region.

This is a two-tailed test so we use $z_{\alpha/2}$ as our critical value. From the normal distribution table $z_{0.025} = 1.96$. So we will reject the null hypothesis if our statistic $|z| > 1.96$.

Step 4: Carry out the test.

$$p_c = \frac{360 + 600}{500 + 800} = 0.74$$

$$\frac{(360/500 - 600/800) - 0}{0.74(1-0.74)\left(\frac{1}{500} + \frac{1}{800}\right)} = -1.2 > 1.96.$$

There is not enough evidence to reject the null hypothesis.

Example 7 Test of the Variance

The Yum-Yum Company owns a machine that dispenses dog food into 16 ounce cans. The quality of the machine is judged by the variance of the dog food dispensed into the cans. The quality control manager of the company collects 30 cans of dog food and weighs their contents. She estimates the variance to be 1 ounce. From this evidence, will she be able to claim that the machine has a variance higher than 0.5? Use a 5% level of significance.

Solution:

Step 1: Set up the hypotheses.

$$H_0 : \sigma^2 \leq 0.5.$$

$$H_1 : \sigma^2 > 0.5.$$

Step 2: Select a test statistic.

$$\frac{(n-1)s_x^2}{\sigma_x^2} = \chi^2_{n-1}.$$

Step 3: Establish the rejection region.
From the χ^2 table, the critical value for 5% and 29 degrees of freedom is 42.557.

Step 4: Carry out the test.

$$\frac{(30-1)1^2}{0.5} = 58 > 42.557.$$

So we can reject the null hypothesis.

Example 8 Testing the Equality of Two Variances

Suppose the quality control manager in Example 7 is not satisfied with the packaging machine. She is thinking of replacing it with a better one. She tests a new machine that offers a 30-day money-back guarantee. A sample of 30 cans from the new machine has a variance of 0.8. Is there enough evidence to argue that the new machine has a smaller variance than the old machine?

Solution:

Step 1: Set up the hypotheses.

$$H_0 : \sigma_{old} \leq \sigma_{new}.$$

$$H_1 : \sigma_{old} > \sigma_{new}.$$

Step 2: Select a test statistic.

$$\frac{s_{old}^2}{s_{new}^2} = f_{n-1,n-1}.$$

Step 3: Establish the rejection region.

From the F-distribution table, the critical value for $\alpha = 5\%$ and 29 degrees of freedom in the numerator and 29 degrees of freedom in the denominator is approximately 1.85.

Step 4: Carry out the test.

$$\frac{1}{0.8} = 1.25 < 1.85.$$

There is not enough evidence to reject the null hypothesis.

Example 9 the Power of a Test for a Means Test

You are working for a light-bulb manufacturer. Your job is to make sure that the average life of the bulbs produced is greater than 300 h. You randomly select 60 bulbs from the bulbs produced last week and find the sample average is 325 and the sample standard deviation is 32. Using this information, you test the null hypothesis that the mean life is less than 300 h at the 5% level of significance. What is the power of the test when the true mean life is 312?

Solution:

Step 1: Obtain the rejection criteria.

Assume that the critical value for rejecting the null hypothesis of $\mu = 300$ is $\bar{X}_c$. Then

$$\frac{\bar{x}_c - 300}{32/\sqrt{60}} = 1.64.$$

We can solve for $\bar{x}_c = 300 + 32\sqrt{60}(1.64) = 306.78.$

Step 2: Obtain the probability of rejecting H_0 assuming $\mu = 312$.

$$\begin{aligned} &P_r(\bar{x} > 3.06.78 \mid \mu = 312) \\ &= P_r\left(\frac{\bar{x} - 312}{32/\sqrt{60}} > \frac{306.78 - 312}{32/\sqrt{60}}\right) \\ &= P_r(z > 1.26) = 0.8962. \end{aligned}$$

Example 10 the Power of a Test for a Proportion

You are working for a furniture manufacturer. Your job is to make sure that the springs purchased are not stronger than 40 pounds (to keep the springs from poking through the fabric). Each time a truckload of springs arrives, you check 600 springs. You will then test the null hypothesis that at least 2% of the springs are bad using a 5% level of significance. One day a truckload arrives that actually contains only 1.5% bad springs. What is the probability of taking this shipment? [Note: this is a power of a test question because we are interested in the probability of rejecting the null hypothesis.]

Solution:

Step 1: Find the critical value for

$$\frac{\bar{p}_c - 0.02}{\sqrt{\frac{0.02(1-0.02)}{600}}}.$$

Step 2: Obtain the probability of rejecting the null hypothesis.

$$P_r(\bar{p} < 0.0106 \mid p = 0.015.$$

$$P_r\left(\frac{\bar{p} - 0.015}{\sqrt{\frac{0.015(1-0.015}{600}}} < \frac{0.0106 - 0.015}{\sqrt{\frac{0.015(1-0.015}{600}}}\right).$$

$$= P_r(z < -0.88) = 0.19$$

Supplementary Exercises

Multiple Choice

1. A type I error

   a. Results from accepting the null hypothesis when it is actually false
   b. Results from rejecting the null hypothesis when it is true
   c. Is a type of sampling error
   d. Represents the size of the test
   e. Is the difference between taking a sample and a census

2. A type II error

   a. Results from accepting the null hypothesis when it is actually false
   b. Results from rejecting the null hypothesis when it is true
   c. Is a type of sampling error
   d. Represents the size of the test
   e. Is the difference between taking a sample and a census

3. The null hypothesis

   a. Is the hypothesis we are trying to reject
   b. Conforms to the status quo
   c. Usually has the greatest cost when it is incorrectly rejected
   d. Results from a type I error
   e. Results from a type II error

4. A one-tailed test of the population would be appropriate if
   a. $H_1 : \mu \neq 0$
   b. $H_0 : \mu = 0$
   c. $H_1 : \mu > 0$
   d. All of the above hypotheses were true
   e. Would not be appropriate for any of the above hypotheses

5. A two-tailed test of the population would be appropriate if
   a. $H_1 : \mu \neq 0$
   b. $H_1 : \mu < 0$
   c. $H_0 : \mu > 0$
   d. All of the above hypotheses were true
   e. Would not be appropriate for any of the above hypotheses

6. A simple hypothesis is

   a. Another name for the null hypothesis
   b. Another name for the alternative hypothesis
   c. One where we specify only one value for the population parameter
   d. A sampling error
   e. One where we specify a range of values for the population parameter

7. A composite hypothesis is

   a. Another name for the null hypothesis
   b. Another name for the alternative hypothesis
   c. One where we specify only one value for the population parameter, Θ
   d. A sampling error
   e. One where we specify a range of values for the population parameter, Θ

8. The power of a test refers to the test's

   a. Ability to control the type I error
   b. Significance level
   c. Ability to reject the null hypothesis
   d. Ability to control the type II error
   e. Ability to reject the alternative hypothesis

9. For a one-tailed test using the normal distribution, a significance level of 0.10 would have a critical value of

   a. 1.645
   b. 1.96
   c. 1.282
   d. 2.575
   e. 1.382

10. For a two-tailed test using the normal distribution, a significance level of 0.10 would have a critical value of

    a. 1.645
    b. 1.96
    c. 1.282
    d. 2.575
    e. 1.382

11. For a one-tailed test using the normal distribution, a significance level of 0.05 would have a critical value of

    a. 1.645
    b. 1.96
    c. 1.282
    d. 2.575
    e. 1.382

12. For a two-tailed test using the normal distribution, a significance level of 0.05 would have a critical value of

    a. 1.645
    b. 1.96
    c. 1.282
    d. 2.575
    e. 1.382

13. For a one-tailed test using the *t*-distribution, a significance level of 0.05, and 20 degrees of freedom would have a critical value of

    a. 1.725
    b. 1.96
    c. 2.86
    d. 2.528
    e. 1.625

14. For a two-tailed test using the *t*-distribution, a significance level of 0.05, and 20 degrees of freedom would have a critical value of

    a. 1.725
    b. 1.96
    c. 1.625
    d. 2.528
    e. 2.086

15. Given $n_x = n_y = 49$ with sample standard deviations $S_x = 7.2$ and $S_y = 3.5$ from two independent, normal populations. What is the critical value on the *F* distribution for with $\alpha = 0.05$?

    a. 6.39
    b. 1.55
    c. 2.5
    d. 1.62
    e. 3.95

16. Given $n_x = n_y = 49$ with sample standard deviations $S_x = 7.2$ and $S_y = 3.5$ from two independent, normal populations. Which of the following is true for testing the null hypothesis H_0: $\sigma_x = \sigma_y$?

    a. The null hypothesis can be rejected at $\alpha = 0.05$
    b. The null hypothesis cannot be rejected at $\alpha = 0.05$
    c. The null hypothesis can be both rejected at $\alpha = 0.10$ and $\alpha = 0.05$
    d. The null hypothesis cannot be rejected at $\alpha = 0.10$
    e. b and d

True/False (If False, Explain Why)

1. The power of a test is $1 - \alpha$.
2. Other things being equal, a two-tailed test will always have a larger absolute critical value than a one-tailed test.
3. There are greater costs associated with making a type II error than with making a type I error.

4. A type I error is the P_r(reject H_0 | H_0 is true).
5. A type II error is the P_r(accept H_0 | H_0 is false).
6. A test with low power will lead to over rejection of the null hypothesis.
7. A two-tailed test is used when we want to prove that the population mean, μ, is not equal to a specified value of μ_0.
8. A one-tailed test is used when we want to prove that the population mean, μ, is either much larger or much smaller than a specified value of μ_0.
9. A simple hypothesis is one where we specify a range of values for the population parameter.
10. Higher p-values lead to rejection of H_0, whereas lower p-values lead to acceptance of H_0.
11. When conducting a hypothesis test with 25 observations and an *unknown variance*, we should use the z distribution.
12. Other things being equal, the critical value will always be smaller when the variance is unknown.
13. There is a tradeoff between the size and power of a test.
14. Testing for the $H_0 : \sigma^2 \geq 10$ vs. $H_1 : \sigma^2 < 10$, we should use the normal distribution.
15. When comparing the variances of two normal populations, we should use the normal distribution.

Questions and Problems

1. Find the critical value for a sample size of 10 and a known variance of 2, for a two-tailed test, given the following levels of significance.
   a. $\alpha = 0.10$
   b. $\alpha = 0.05$
   c. $\alpha = 0.01$
2. Find the critical value for a sample size of 10 and a known variance of 2, for a one-tailed test, given the following levels of significance.
   a. $\alpha = 0.10$
   b. $\alpha = 0.05$
   c. $\alpha = 0.01$
3. Find the critical value for a sample size of 10 and an *unknown* variance of 2, for a two-tailed test, given the following levels of significance.
   a. $\alpha = 0.10$
   b. $\alpha = 0.05$
   c. $\alpha = 0.05$
   d. $\alpha = 0.05$

4. Find the critical value for a sample size of 10 and an unknown variance of 2, for a one-tailed test, given the following levels of significance.
   a. $\alpha = 0.10$
   b. $\alpha = 0.05$
   c. $\alpha = 0.01$
5. Suppose you are given the following information:

$$x = 100, \sigma^2 = 64, n = 30.$$

   Conduct the following hypothesis test at the 0.05 level of significance:

$$H_0: \mu = 75 \text{ vs. } H_1: \mu > 75.$$

6. A sample of 25 students taking the GMATs at Business School University has a sample mean of 525 and a sample standard deviation of 90. Test the hypothesis that the mean GMAT score of BSU students is equal to 500, against the alternative hypothesis that the mean GMAT score is higher than 500, at the 0.05 level of significance.
7. The Play Like a Pro Golf School claims that the variance of a student's golf score will be less than six strokes per round. A random sample of 30 students who took the course found the variance to be four strokes per round. Assuming a normal distribution, test the golf school's claim at the 5% level of significance.
8. Suppose we are interested in the proportion of people in a small town who favor prayer in school. A random sample of 100 people finds 45 in favor of prayer in school. Test the null hypothesis that p equals 0.50 against the alternative hypothesis that p is less than 0.50 at the 0.10 level of significance.
9. A sample of 100 students in a high school have a sample mean score of 505 on the English portion of the SATs. If the sample standard deviation is 110, test the hypothesis that the high school's mean SAT score is 480 against the alternative hypothesis that the school's mean SAT score is greater than 480, at the 0.10 level of significance.
10. A GMAT review course claims that the variance of test scores of its graduates is less than 100. A random sample of 30 students who took the course is taken and found to have a variance of 95. Assuming a normal distribution, test the review course's claim at the 5% level of significance.
11. A manager claims that the standard deviation in their mean delivery time is less than 2.5 days. A sample of 25 customers is taken. The average delivery time in the sample was four days with a standard deviation of 1.2 days. Suppose the delivery times are normally distributed, at 95% confidence, test the manager's claim.
12. A candidate believes that more than 30% of the citizens will vote for him. A random sample of 250 citizens was taken and 101 of them vote for the candidate.
    a. State the null and alternative hypotheses.
    b. Using the critical value approach, test the hypotheses at the $\alpha = 1\%$ level of significance.

c. Using the *p*-value approach, test the hypotheses at the $\alpha = 1\%$ level of significance.

13. A poll on the preference of two presidential candidates A and B are shown below.

Candidate	Voters surveyed	Voters favoring this candidate
A	500	292
B	400	225

At 99 % confidence, test to determine whether or not there is a significant difference between the preferences for the two candidates.

Answers to Supplementary Exercises

Multiple Choice

1. b	6. c	11. a	16. c
2. a	7. e	12. b	
3. d	8. c	13. a	
4. c	9. c	14. e	
5. a	10. a	15. d	

True/False

1. False. The power of the test is $1 - \beta$.
2. True.
3. False. Because of the way we construct the null hypothesis, a type I error usually has greater costs than a type II error.
4. True.
5. True.
6. False. Low power will lead to an under rejection of the null hypothesis.
7. True.
8. True.
9. False. A simple hypothesis is one where we specify only one value for the parameter.
10. False. Lower p-values lead to rejection of H_0.

11. False. When the variance is unknown and the sample size is small, we should use the t-distribution.
12. False. Other things being equal, the critical value will be larger when the variance is unknown.
13. True.
14. False. Use the chi-square distribution.
15. False. Use the F-distribution.

Questions and Problems

1. a. 1.645
   b. 1.96
   c. 2.575
2. a. 1.282
   b. 1.645
   c. 2.327
3. a. 1.833
   b. 2.262
   c. 3.250
4. a. 1.383
   b. 1.833
   c. 2.821
5. We compute z = 17.17 and the critical value for a one-tailed test with $\alpha = 0.05$ is 1.645, so we reject the null hypothesis.
6. We compute t = 1.389 and the critical value for a one-tailed test with $\alpha = 0.05$ and 24 degrees of freedom is 1.711 so we reject the null hypothesis.
7. We compute $\chi 2 = 19.33$, critical value with $\alpha = 0.05$ and 29 degrees of freedom is 42.557, so we cannot reject the null hypothesis.
8. We compute $z = -1.005$, critical value for a one-tailed test with $\alpha = 0.10$ is -1.28, so we cannot reject the null hypothesis.
9. $z = (505 - 480) / (110/\sqrt{100}) = 2.27$

   The critical value for $\alpha = 0.10$ is 1.28. Therefore, we can reject the null hypothesis that the mean SAT score is equal to 480.

10. H_0: $\sigma^2 \geq 100$ VS. H_1: $\sigma^2 < 100$

    $x^2_{n-1} = [(n-1)s^2]/\sigma^2 = [(30-1)95]/100 = 27.55$

    The critical value for $\alpha = 0.05$ and 29 degrees of freedom is 42.557. Therefore, we are unable to reject the null hypothesis.

11. $H_0 : \mu \le 2.5$ vs. $H_0 : \mu > 2.5$

Test Statistics $\dfrac{\bar{X} - 2.5}{S/\sqrt{25}} = \dfrac{4 - 2.5}{1.2/\sqrt{25}} = 6.25$

As $t_{0.05} = 1.711 \rightarrow$ Reject H_0

12. a. $H_0 : p \le 0.3$ vs. $H_0 : p > 0.3$

b. Test Statistics $\dfrac{\bar{P} - 0.3}{\sqrt{0.3 \times 0.7/250}} = \dfrac{0.404 - 0.3}{0.02898} = 3.5883$

$z_{0.01} = 2.33 \rightarrow$ Reject H_0

c. p-value $= 0.000166 \rightarrow$ Reject H_0

13. $H_0 : p_1 = p_2$ vs. $H_0 : p_1 \ne p_2$

$p_{\text{pooled}} = (292 + 225)/(500 + 400) = 0.5744$

Test Statistics $\dfrac{\bar{P}_1 - \bar{P}_2}{\sqrt{p_{pooled} \times (1 - p_{pooled})/N}} = 1.3045$

$z_{0.01} = 2.58 \rightarrow$ Accept H_0.

Chapter 12
Analysis of Variance and Chi-Square Tests

Chapter Intuition

In Chap. 11, we learned how to test for differences in the means of two groups. When we are interested in comparing the means of more than two groups, we use a procedure known as analysis of variance (ANOVA). For example, suppose a factory operates three shifts. If the manager of the factory wants to test the null hypothesis of equal productivity for all three shifts, ANOVA represents the appropriate technique.

ANOVA uses an *F*-statistic to compare the means of three or more groups. To understand how ANOVA works, consider the numerator of the *F*-statistic, which measures the difference between the average productivity of each shift and the average productivity of all three shifts. If there is similar productivity in all three shifts, then the average productivity of each shift should be close to the overall average productivity and the numerator will be small. When there is different productivity in all three shifts, the numerator should be large.

Two-way ANOVA is an extension of one-way ANOVA that considers two categorical variables on one numerical variable. It is also a method for comparing three or more means while controlling for a second factor. For example, in the previous factory example, we are interested in the productivity of the three shifts. However, we also recognize that the number of years of experience of the employees may have a bearing on the productivity. Therefore, we must account for this factor in order to draw correct conclusions.

The second topic covered in this chapter is goodness of fit tests. Goodness of fit tests are used to test whether the data are generated from a presumed distribution. To conduct this test, we need to determine what the frequency distribution should be if the data are really generated from the presumed distribution. In order to determine if the data do come from the presumed distribution, we compare the expected frequency of the presumed distribution with the observed frequency from the data. If the observed frequencies are similar to the expected frequencies, we can conclude that the data do come from the presumed distribution. Otherwise, we must reject the null hypothesis and conclude that the data may come from some other distribution.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_12

Another use of a goodness of fit test is to test if two variables are independent. To **test the independence** of two variables, we construct a ***contingency table*** and compare the expected frequencies with the observed frequencies.

Chapter Review

1. When we are interested in knowing whether the means of three or more populations equal each other or not, we use ***analysis of variance*** **(ANOVA)**. Testing for the equality of means can be useful for examining a wide variety of problems in business. For example, suppose a factory operates three shifts. If the manager wants to test the null hypothesis of equal productivity in the three shifts, we can use ANOVA to answer this question.
2. **ANOVA** tests the equality of more than two means by comparing the variances between groups with the variances within groups. The variation between groups is measured by the ***sum of squares between treatments*** (SSB). The SSB measures the variation among the population means of the treatment groups. The variation within each group is measured by the ***sum of squares within treatments*** (SSW). The SSW measures the variation within each treatment group. The total variation for the data is measured by the ***total sum of squares*** (TSS).
3. **One-way ANOVA** tests the equality of more than two means, assuming that only one factor determines the differences in the means. From our factory example, the manager assumes that only one factor, the shift, determines productivity, while productivity is the response variable.
4. The results of one-way ANOVA are reliable if the following assumptions are met:

   a. For each population, the response variable is normally distributed.
   b. The variance of the response variable is the same for all populations.
   c. The observations are independent.

5. **Two-way ANOVA** tests the equality of more than two means, assuming that more than one factor determines the differences in the means. For example, when the manager examines the differences in productivity of the three shifts, he may also wish to control for the experience of the workers in each shift. A two-way ANOVA can be run under either one of the followings:

   a. With replication (more than one observation for each combination of the two factors) and thus can be tested whether interaction effect exists.
   b. Without replication (only one observation for each combination of the two factors), for example, one factor is composed of treatments, the other is composed of blocks.

6. To illustrate how one-way ANOVA works, let us return to our factory work example. If we assume that the means of the three shifts are the same, then most of the variation of the data results from the within-group variation (SSW) and there will be very little between-group variation (SSB), because all groups have the same mean. On the other hand, if the means of the three shifts are different,

there will be a large between-group variation and a smaller within-group variation. By taking the ratio of SSW to SSB (with an adjustment for degrees of freedom), we can form an *F*-statistic that allows us to determine if the means are equal.

7. **Goodness of fit tests** are used to determine whether the data are generated from a hypothesized distribution. The test compares the observed frequency (f_o) with the expected frequency (f_e). The expected frequency is obtained by assuming that the null hypothesis is true; that is, it is the frequency we expect to see if the null hypothesis is true. The observed frequency represents the frequency we actually observe. When the observed frequency (our evidence) is far enough away from the null hypothesis, the test statistic will generate a large value and lead to a rejection of the null hypothesis.
8. Another use of a goodness of fit test is to determine if two variables are independent. To **test the independence** of two variables, we construct a ***contingency table*** and compare the expected frequencies with the observed frequencies.

Useful Formulas

One-way ANOVA:	Two-way ANOVA:
Testing the equality of three or more means:	1. **One observation in each cell:**
$F = \frac{MST}{MSW} = \frac{SST/(m-1)}{SSW/(n-m)}$ $SST = \sum_{j}^{m} n_j(\bar{x}_j - \bar{x})^2 = \sum_{j}^{m} \frac{x_{\cdot j}^2}{n_j} - \frac{x_{\cdot\cdot}^2}{n}$ $SSW = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = \sum_{j}^{m}\sum_{i=1}^{n_j} x_{ij}^2 - \sum_{j=1}^{m}\frac{x_{\cdot j}^2}{n_j}$ $TSS = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2 = \sum_{j}^{m}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{x_{\cdot\cdot}^2}{n}$ Also, $SST = TSS - SSW$	**Test treatment effect:** $F_{(J-1),\,(I-1)\,(J-1)} = \frac{SST/(J-1)}{SSE/(I-1)(J-1)}$
Comparing the difference between two population means: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$	**Test for block effect:** $F_{(I-1),(I-1)(J-1)} = \frac{SSB/(I-1)}{SSE/(I-1)(J-1)}$ $SST = \sum_{j=1}^{J} I(\bar{x}_{\cdot j} - \bar{x})^2 = \sum_{j=1}^{J} \frac{x_{\cdot j}^{\,2}}{I} - \frac{x_{\cdot\cdot}^{\,2}}{IJ}$ $SSB = \sum_{i=1}^{I} J(\bar{x}_{i\cdot} - \bar{x})^2 = \sum_{i=1}^{I} \frac{x_{i\cdot}^{\,2}}{J} - \frac{x_{\cdot\cdot}^{\,2}}{IJ}$ $TSS = \sum_{j=1}^{J}\sum_{i=1}^{I}(x_{ij} - \bar{x})^2 = \sum_{j}^{m}\sum_{i=1}^{n_j} x_{ij}^{\,2} - \frac{x_{\cdot\cdot}^{\,2}}{IJ}$ $SSE = TSS - SST - SSB$

One-way ANOVA:	Two-way ANOVA:
	2. **K observations in each cell:** Test treatment (or factor 1) effect: $F = \dfrac{SST/(J-1)}{SSE/IJ(k-1)}$
	Test block (or factor 2) effect: $F = \dfrac{SSB/(I-1)}{SSE/IJ(k-1)}$
	Test interaction effect: $F = \dfrac{SSB/(I-1)(J-1)}{SSE/IJ(k-1)}$
	$SST = IK\sum_{j=1}^{J}(\bar{x}_{.j.} - \bar{x})^2$ $SSB = JK\sum_{i=1}^{I}(\bar{x}_{i..} - \bar{x})^2$ $SSI = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$ $TSS = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(x_{ijk} - \bar{x})^2$ $SSE = TSS - SST - SSB - SSI$
Chi-square test for goodness of fit: $\chi^2_{k-1} = \sum_{i=1}^{k}\dfrac{(f_i^o - f_i^e)^2}{f_i^e}$ f_i^o : observed frequency f_i^e : expected frequency k : number of groups	**Chi-square test of independence:** $\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\dfrac{(f_{ij}^o - f_{ij}^e)^2}{f_{ij}^e}$ f_{ij}^o : observed frequency f_{ij}^e : expected frequency $f_{ij}^e = f_{i.}^o f_{.j}^o / f_{..}^o$

Example Problems

Example 1 One-way ANOVA

The following table shows the sales figures (in thousands) for salespersons with different years of experience. Can you reject the null hypothesis of no difference among various experiences? Do a 5% test.

Salesperson's experience (years)			
	0–2	3–4	5 or more
Sales ($ 1000)	13	18	16
	17	17	17
	16	15	20
	18	14	19
	16	21	18
Average	16	17	18

Solution:

$$\bar{x}_1 = 16, \bar{x}_2 = 17, \bar{x}_3 = 18, \bar{x} = 17.$$

Worksheet for calculating within-treatment sum of squares (SSW):

	$(x_{i1} - \bar{x}_1)^2$	$(x_{i1} - \bar{x}_2)^2$	$(x_{i3} - \bar{x}_3)^2$
	$(13-16)^2$	$(18-17)^2$	$(16-18)^2$
	$(17-16)^2$	$(17-17)^2$	$(17-18)^2$
	$(16-16)^2$	$(15-17)^2$	$(20-18)^2$
	$(18-16)^2$	$(14-17)^2$	$(19-18)^2$
	$(16-16)^2$	$(21-17)^2$	$(18-18)^2$
Total	14	30	10

Hence,

$$SSW = \sum_{i=1}^{5}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2 + \sum_{i=1}^{5}(x_{i3} - \bar{x}_3)^2 = 14 + 30 + 10 = 54$$

$$SST = \sum_{j=1}^{3} n_j(\bar{x}_j - \bar{x})^2 = 5\left[(16-17)^2 + (17-17)^2 + (18-17)^2\right] = 10$$

$$F = \frac{MST}{MSW} = \frac{SST/(m\text{-}1)}{SSW/(n\text{-}m)}$$
$$= \frac{10/(3-1)}{54/(15-3)}$$
$$= 1.11 < F_{2,12;0.05} = 3.89$$

Therefore, we do not reject the null hypothesis.

Example 2 Two-way ANOVA with One Observation in Each Cell

Data for ABC Company's recruitment of salespeople resulted in 15 applicants who were classified into three groups according to their past experiences and also their ranks on an aptitude test. Their yearly sales figures (in thousands) for one year after recruitment are presented in the following table.

Past experience (years)					Row means
		0–2	3–4	5 or more	
Aptitude test	E	8	9	10	9
	D	6	9	12	9
	C	8	8	11	9
	B	14	16	18	16
	A	18	16	20	18
Column means		10.8	11.6	14.2	12.2

Based on $\alpha = 0.05,$ can you argue whether the experience or the rank of the aptitude test makes any difference?

Solution:

$$\text{SST} = \sum_{j=1}^{3} 5(\overline{x}_j - \overline{x})^2$$
$$= 5[(10.8-12.2)^2 + (11.6-12.2)^2 + (14.2-12.2)^2] = 5[6.32] = 31.6$$

$$\text{SSB} = \sum_{i=1}^{5} 3(\overline{x}_j - \overline{x})^2$$
$$= 3[(9-12.2)^2 + (9-12.2)^2 + (9-12.2)^2 + (16-12.2)^2 + (18-12.2)^2]$$
$$= 3[78.8] = 236.4$$

$$\text{TSS} = \sum_{j=1}^{3}\sum_{i=1}^{5}(x_{ij} - \overline{x})^2 = [(8-12.2)^2 + (6-12.2)^2 + \cdots + (20-12.2)^2] = 278.4$$

$$\left(\text{Also, TSS} = \sum_{i=1}^{5}\sum_{j=1}^{3} x_{ij}^{\ 2} - \frac{x_{..}^{\ 2}}{IJ} = 2511 - \frac{183^2}{5\times 3} = 278.4\right)$$

$$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB} = 278.4 - 31.6 - 236.4 = 10.4$$

$$\text{MST} = \text{SST}/(J-1) = 31.6/(3-1) = 15.8$$

$$\text{MSB} = \text{SSB}/(I-1) = 236.4/(5-1) = 59.1$$

$$\text{MSE} = \text{SSE}/(I-1)(J-1) = 10.4/(5-1)\,(3-1) = 1.3$$

Since $\text{MST/MSE} = 15.8/1.3 = 12.16 > F_{2,8;0.05} = 4.46,$

so reject the null hypothesis that past experience does not make any difference.

Since

$$\text{MSB/MSE} = 59.1/1.3 = 45.46 > F_{4,8;0.05} = 3.84,$$

so reject the null hypothesis that the rank of aptitude test does not make any difference.

Example 3 Two-way ANOVA with More Than One Observation in Each Cell

During the recruitment of salespeople, the Lee Corporation had 30 applicants who were classified into three groups according to their past experiences and ranks on an aptitude test. Their yearly sales figures (in thousands) for one year after recruitment are presented in the following table.

Past experience (years)					Row means
		0–2	3–4	5 or more	
Aptitude test	E	8, 9	9, 10	10, 11	9.5
	D	7, 6	9, 10	12, 13	9.5
	C	9, 8	8, 9	11, 12	9.5
	B	14, 15	17, 16	18, 19	16.5
	A	18, 19	16, 17	20, 21	18.5
Column means		11.3	12.1	14.7	12.7

Based on $\alpha = 0.05,$ does the experience or rank of aptitude test or the interaction make any difference?

Solution:

$$I = 5, J = 3, K = 2$$

$$SST = \sum_{j=1}^{3}(5)(2)(\bar{x}_{.j.} - \bar{x})^2$$

$$= 10[(11.3 - 12.7)^2 + (12.1 - 12.7)^2 + (14.7 - 12.7)^2]$$

$$= 10[6.32] = 63.2$$

$$SSB = \sum_{i=1}^{5}(3)(2)(\bar{x}_{i..} - \bar{x})^2$$

$$= 6[(9.5 - 12.7)^2 + (9.5 - 12.7)^2 + (9.5 - 12.7)^2 + (16.5 - 12.7)^2 + (18.5 - 12.7)^2]$$

$$= 6[78.8] = 472.8$$

$$SSI = 2\sum_{j=1}^{3}\sum_{i=1}^{5}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$$

$$= 2[(8.5 - 9.5 - 11.3 + 12.7)^2 + (9.5 - 9.5 - 12.1 + 12.7)^2 + \ldots$$

$$+ (20.5 - 18.5 - 14.7 + 12.7)^2]$$

$$= 2[10.4] = 20.8$$

$$TSS = \sum_{j=1}^{3}\sum_{i=1}^{5}\sum_{k=1}^{2}\left(x_{ijk} - \bar{x}\right)^2$$

$$= [(8 - 12.7)^2 + (9 - 12.7)^2 + \cdots + (21 - 12.9)^2]$$

$$= 564.3$$

$$\left(\text{Also, } TSS = \sum_{i=1}^{5}\sum_{j=1}^{3}\sum_{k=1}^{2}x_{ijk}^{\ 2} - x_{...}^{\ 2}/IJK = 5403 - 381^2/(5 \times 3 \times 2) = 564.3\right)$$

$$SSE = TSS - SST - SSB - SSI = 564.3 - 63.2 - 472.8 - 20.8 = 7.5$$

$$MST = SST/(J-1) = 63.2/(3-1) = 31.6$$

$$MSB = SSB/(I-1) = 472.8/(5-1) = 118.2$$

$$MSI = SSI/(I-1)(J-1) = 20.8/(4)(2) = 2.6$$

$$\text{Since } MST/MSE = 31.6/0.5 = 63.2 > F_{2,15;0.05} = 3.68,$$

so reject the null hypothesis that past experience does not make any difference.

$$\text{Since MSB/MSE} = 118.2/0.5 = 236.4 > F_{4,15;0.05} = 3.06,$$

so reject the null hypothesis that the rank of aptitude test does not make any difference.

$$\text{Since MSI/MSE} = 2.6/0.5 = 5.2 > F_{8,15;0.05} = 2.64,$$

so reject the null hypothesis that the interaction does not exist.

Example 4 Goodness of Fit Test

Four hundred students were randomly sampled and asked who they would vote for in the coming student government election. The results are given below.

Candidate	Smith	Gomez	Blackwell	Friedman
Votes	131	121	99	49

Do the four candidates acquire different levels of support? Use a 5% level of significance.

Solution: H_0: No difference in support.
H_1: Some difference in support.

Test statistic is

$$\chi^2_{k-1} = \sum_{i=1}^{k} (f_i^o - f_i^e)^2 / f_i^e,$$

where k = 4.

f_i^o	f_i^e	$(f_i^o - f_i^e)^2 / f_i^e$
131	100	9.61
121	100	4.41
99	100	0.01
49	100	26.01
	Sum	40.04

Since $40.04 > \chi^2_{3,0.05} = 7.81$, we reject the null hypothesis that there is no difference in the support for the four candidates.

Example 5 Goodness of Fit Test

Using the information in Example 4, assume that 2 days before the election, a debate was held among the four candidates. After the debate, another poll was conducted to study whether the voting pattern changed. Can you argue that the voting pattern changed? Use a 5% level of significance. Below are the results of the new poll.

Candidate	Smith	Gomez	Blackwell	Friedman
Votes	98	102	81	119

Solution: H_0: voting pattern did not change
H1: voting pattern changed

Test statistic is $\chi^2_{k-1} = \sum_{i=1}^{k} (f_i^o - f_i^e)^2 / f_i^e$, where k=4 and f_i^es are votes before the debate.

f_i^o	f_i^e	$(f_i^o - f_i^e)^2 / f_i^e$
98	131	8.31
102	121	2.98
81	99	3.27
119	49	100
Sum		114.56

Since $114.56 > \chi^2_{3,0.05} = 7.81$, we reject the null hypothesis that there has been no change in the voting pattern.

Example 6 Goodness of Fit Tests

A management consultant wants to develop an inventory control system. She collects weekly demand data for a high-tech machine. The results are given below.

Demands in units	0	1	2	3	4	5	6	7	8	9
Frequencies	17	25	40	46	29	28	8	3	4	0

Use a 5% level of significance to determine whether the data follow a Poisson distribution.

Solution:

H_0: weekly demands is Poisson.

H_1: weekly demands is not Poisson.

To conduct this test, she compares the observed frequency with the expected frequency by assuming the data follows the Poisson distribution. Hence, under the null hypothesis, the probability of having x units demanded in a week is:

$P(x) = e^{-\lambda}\lambda^x/x!$, where λ is the expected weekly demands in units.

Since λ is unknown, it is estimated by using the methods for obtaining the group mean in Chap. 4, which is

$$\bar{x} = [(0\times 17) + (1\times 25) + \cdots + (8\times 4) + (9\times 0)]/200 = 600/200 = 3.$$

Hence, the probability of x units demanded is estimated to be

$$P(x) = e^{-\lambda}\lambda^x/x! = e^{-3}3/x!, \textit{ for } x = 0,\ 1,\ 2,\ \ldots,\ 8,$$

where $P(x \geq 9) = 1 - \sum_{i=0}^{8} P(i) = 0.01$ and they are listed in column (2).

Then, $f_i^e = 200 \times P(i)$ is listed in column (3).

Weekly demands	(1) f_i^o	(2) P(x)	(3) f_i^e	(4) $(f_i^o - f_i^e)^2/f_i^e$
0	17	0.05	10	4.9
1	25	0.15	30	0.83
2	40	0.22	44	0.37
3	46	0.22	44	0.09
4	29	0.17	34	0.74
5	28	0.10	20	3.2
6	8	0.05	10	0.4
7	3	0.02	4	0.25
8	4	0.01	2	2
≥ 9	0	0.01	2	2
Sum	200	1.00	200	14.77

Since $\sum (f_i^o - f_i^e)^2/f_i^e = 14.77 < \chi^2_{8,0.05} = 15.5$, we cannot reject the null hypothesis.

Example 7 Goodness of Fit Test for Normal Distribution

The manager of a gas station would like to test if weekly sales volume is normally distributed. The sales volumes in thousand gallons are reported

Volume	f_i^o
40–42	5
42–44	27
44–46	67
46–48	63
48–50	31
50–52	7
Total	200

Do the data support the hypothesis that the data follow a normal distribution? Do a 5% test.

Solution:

The mean and standard deviation should be estimated using the methods for obtaining the group mean and the group standard deviation in Chap. 4.

$$\bar{x} = \sum_{i=1}^{6} f_i^o m_i / 200 = [5\times 41 + 27\times 43 + \cdots + 7\times 51]/200 = 46.09$$

$$s^2 = \sum_{i=1}^{6} f_i^o (m_i - \bar{x})^2 / 199$$

$$= [5\times (41-46.09)^2 + 27\times (43-46.09)^2 + \cdots + 7\times (51-46.09)^2]/199 = 4.776$$

$$s = \sqrt{4.776} = 2.19$$

Assuming the normal distribution is true, we then obtain the probability of each volume interval.

$$p_1 = P(X \le 42) = P\left(\frac{X-46.09}{2.19} \le \frac{42-46.09}{2.19}\right) = P\left(Z \le \frac{42-46.09}{2.19}\right) = 0.031.$$

$$p_2 = P(42 \le X \le 44) = P\left(\frac{42-46.09}{2.19} \le Z \le \frac{44-46.09}{2.19}\right) = 0.139$$

:

:

$$p_6 = P(X \ge 50) = P\left(Z \ge \frac{50-46.09}{2.19}\right) = 0.037$$

Then, the expected frequency of each interval is obtained by

$$f_i^e = np_i = 200p_i, i = 1, \ldots, 6.$$

Volume	f_i^o	p_i	f_i^e	$(f_i^o - f_i^e)^2 / f_i^e$
40–42	5	0.031	6.182	0.23
42–44	27	0.139	27.809	0.02
44–46	67	0.314	62.731	0.29
46–48	63	0.325	64.965	0.06
48–50	31	0.154	30.893	0.00
50–52	7	0.037	7.420	0.02
Total	200	1.00	200	0.62

Hence, the chi-square statistic is $\sum_{i=1}^{6}(f_i^o - f_i^e)^2/f_i^e = 0.62.$

Since we use x and s to estimate μ and σ, so the degrees of freedom of the chi-square statistic is $6-1-2=3$, and $0.62 < \chi^2_{3,0.05} = 7.81$.

So the null hypothesis that the data are normally distributed is not rejected.

Example 8 Test of Independence

Six hundred students from the campus are asked whether they agree with the new tuition policy. Their opinions are summarized in the following table:

	Agree	No	Disagree
Freshman	70	40	60
Sophomore	70	50	60
Junior	60	50	40
Senior	50	30	20

Can you argue that the students' opinions for the new tuition policy are independent of their class year? Use a 5% level of significance.

Solution:

The expected frequencies are computed under the assumption of independence.

Hence, $f_{ij}^e = (f_{i.}^o \times f_{.j}^o)/n, i = 1,2,3,4, j = 1,2,3$

For example, $f_{11}^e = (f_{1.}^o \times f_{.1}^o)/600 = (250)(170)/600 = 70.83$, and the expected frequencies are in the following table:

f_{ij}^e			$(f_{ij}^o - f_{ij}^e)^2/f_{ij}^e$		
70.83	48.17	51.00	0.01	1.38	1.59
75.00	51.00	54.00	0.33	0.02	0.67
62.50	42.50	45.00	0.10	1.32	0.56
41.67	28.33	30.00	1.67	0.10	3.33

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(f_{ij}^o - f_{ij}^e)^2}{f_{ij}^e}$$
$$= \frac{(70-70.83)^2}{70.83} + \cdots + \frac{(20-30)^2}{30} = 11.8$$
$$\chi^2_{6,0.05} = 12.59$$

Since $11.08 < 12.59$, do not reject the opinion is independent of the class year.

Supplementary Exercises

Multiple Choice

1. Which of the following may be the reason that we can not reject the null hypothesis of $\mu_1 = \mu_2 = \mu_3$ in an ANOVA?

   a. Too few groups
   b. Too much variation in each group
   c. The degrees of freedom of the numerator are too low
   d. The degrees of freedom of the denominator are too high
   e. None of the above

2. In a one-way ANOVA, a large TSS relative to SSW would most likely lead to

   a. Acceptance of the null hypothesis of equal means
   b. Rejection of the null hypothesis that the data come from the Poisson distribution
   c. Acceptance of the null hypothesis that the data come from the Poisson distribution
   d. Rejection of the null hypothesis of equal means
   e. None of the above

3. In a one-way ANOVA, if there is a large difference between the means of the populations, then

   a. SST will be large relative to SSW.
   b. SST will be small relative to SSW.
   c. SSW will be large relative to TSS.
   d. SST and SSW will be equal.
   e. TSS will be large relative to SST.

4. ANOVA uses the

   a. t-statistic.
   b. Chi-square statistic.
   c. z-statistic.
   d. F-statistic.
   e. Binomial distribution.

5. ANOVA is used to test the equality of different population

   a. Variances
   b. Means
   c. Modes
   d. Medians
   e. Distributions

6. ANOVA procedure is applied to data obtained from four samples where each sample contains six observations. The degrees of freedom for the critical value of F are

   a. 4 numerator and 20 denominator degrees of freedom
   b. 3 numerator and 20 denominator degrees of freedom
   c. 4 numerator and 6 denominator degrees of freedom
   d. 4 numerator and 19 denominator degrees of freedom
   e. 3 numerator and 19 denominator degrees of freedom

7. The F value with 3 numerator and 20 denominator degrees of freedom at $\alpha = .05$ is

   a. 3.098
   b. 2.895
   c. 3.127
   d. 3.515
   e. 4.534

8. The mean square is the sum of squares divided by

   a. The total number of observations
   b. Its corresponding degrees of freedom
   c. Its corresponding degrees of freedom minus one
   d. The degrees of freedom of within-treatment variations
   e. None of these alternatives is correct

9. In one-way ANOVA, if the null hypothesis is rejected, then one could conclude that

   a. All treatment means are unequal
   b. All treatment means are equal
   c. Some treatment means are not equal
   d. All variances are unequal
   e. All variances are equal

10. Which of the following statements about Scheffe's multiple comparison is true?

    a. It is to be used once the *F*-test in ANOVA has been rejected
    b. It determines each interval simultaneously to yield a $(1-\alpha)$ confidence level
    c. There are many different multiple comparison methods but all yield the same conclusions
    d. a and b
    e. All are true

11. Which of the following can be used in a test of independence among two categorical variables?

    a. SSB and SSW
    b. SSB, SSW, and SSF
    c. TSS and SSB
    d. ANOVA
    e. contingency table

12. The formula for a goodness of fit test is
    a. $\Sigma(f_o - f_e)/f_e$
    b. $\Sigma(f_o - f_e)^2/f_o$
    c. $\Sigma(f_o - f_e)^2/f_e$
    d. $\Sigma(f_o - f_e)/f_o$
    e. $\Sigma(f_o + f_e)/f_o$

13. To test for goodness of fit, which of the following is derived assuming the null hypothesis is true?

    a. The expected frequencies
    b. The observed frequencies
    c. The sum of the expected frequencies
    d. The sum of the observed frequencies
    e. The sum of the squared observed frequencies

14. To test whether the data follow a binomial distribution you should compute the

    a. Expected frequencies assuming the data are not generated from a binomial distribution
    b. Observed frequencies assuming the data are not generated from a binomial distribution

c. Expected frequencies assuming the data are generated from a binomial distribution
d. Observed frequencies assuming the data are generated from a binomial distribution
e. Squared observed frequencies assuming the data are generated from a binomial distribution

15. In a goodness of fit test, we use the

a. t-statistic
b. Chi-square statistic
c. z-statistic
d. F-statistic
e. Binomial distribution

16. In a goodness of fit test, if the expected and observed frequencies are close to one another then most likely

a. The null hypothesis would be rejected
b. The null hypothesis would not be rejected
c. The alternative hypothesis would be accepted
d. The null hypothesis may be accepted or rejected
e. Both a and c

17. In a goodness of fit test with all parameters being known, the degrees of freedom for the chi-square statistic is the number of

a. Groups
b. Groups minus 1
c. Observations
d. Observations minus 1
c. Observations multiplied by that of groups

18. A contingency table can be used to test

a. The equality of population means.
b. both equality of means and variances
c. The equality of population variances.
d. two-way ANOVA problems
e. Statistical independence

19. To conduct chi-square test for independence, the expected cell frequency is the

a. Number of observations
b. Number of groups
c. Product of row totals times column totals
d. Product of row totals times column totals divided by sample size
e. Number of observations divided by sample size

20. To conduct chi-square test for independence, the degrees of freedom are equal to

    a. The number of rows
    b. The number of columns
    c. The number of rows multiplied by the number of columns
    d. The number of rows multiplied by the number of columns minus 2
    e. The number of rows minus 1 multiplied by the number of columns minus 1

True/False (If False, Explain Why)

1. The ANOVA can be used to test the equality of different group means.
2. In one-way ANOVA, the null hypothesis is that the data come from a single population.
3. ANOVA usually uses frequency data.
4. If the treatments do not create different effects on the group means, the numerator in the ANOVA will be large.
5. In one-way ANOVA, if the null hypothesis is rejected, then we would conclude all treatments are different.
6. In one-way ANOVA, one must always check the results from pairwise comparisons no matter whether F-test is rejected or not.
7. Controlling a second factor in an ANOVA will usually reduce the sum of squared error.
8. Controlling a second factor in an ANOVA will change the sum of squares between the groups corresponding to the first factor.
9. ANOVA can be used to test whether data are generated from a normal distribution.
10. ANOVA assumes that the sampled populations are normally distributed but have different variances.
11. Two-way ANOVA allows us to compare the equality of means when two factors are assumed to influence the population means.
12. Two-way ANOVA with replications allows us to test whether interaction effects among the two factors exist.
13. In a test of independence, the null hypothesis is that the data are dependent.
14. The idea of a goodness of fit test is to compare the observed frequencies with expected frequencies, assuming the alternative hypothesis is true.
15. *F*-statistic can be used to test whether data are from a Poisson distribution.
16. To test whether data follow a Poisson distribution, the expected frequencies are calculated according to that Poisson distribution.
17. Goodness of fit tests use frequency data.
18. The sum of expected frequencies and the sum of observed frequencies are the same.
19. When the probability of one variable occurring is not influenced by the outcome of another variable, the two variables are statistically independent.
20. A contingency table uses an *F*-statistic to test for statistical independence.

Questions and Problems

1. To test whether or not there is a difference between three sales promotions, a sample of 12 stores has been randomly assigned to the three plans, and the sales volumes after one week are presented in the following table:

Promotion		
1	2	3
24	21	34
30	23	38
26	23	36
28	25	40

   a. Do the three promotions differ in average sales volume at $\alpha = 0.05$?
   b. Find the 95% confidence interval for $\mu_1 - \mu_2$.
   c. Find the 95% confidence interval for $\mu_1 - \mu_3$.
   d. Find the 95% confidence interval for $\mu_2 - \mu_3$.
   e. Find the 95% interval of the Scheffe's multiple comparison for $\mu_2 - \mu_3$.

2. Continue Question 1. Suppose the sizes of the 12 stores are different, and after being classified according to the sizes, the data set is represented in the following table:

			Promotion	
		1	2	3
Size	**Mini**	24	21	34
	Small	30	23	38
	Medium	26	23	36
	Large	28	25	40

   a. Do the three promotions differ in average sales volume at $\alpha = 0.05$?
   b. Do the sizes of stores differ in average sales volume at $\alpha = 0.05$?

3. One thousand web surfers were asked which searching engine was the most popular in 2010. The results were

Search engine	Google	Yahoo!	Bing	Ask
Frequency	650	140	110	100

Can you reject the null hypothesis that the popularities for Google is 60% and for Yahoo!, Bing, and Ask is 10%, respectively? Use a 5% level of significance.

4. The same 1000 web surfers were asked the same question again in 2013. The results are as follows

Search engine	Google	Yahoo!	Bing	Ask
Frequency	700	100	140	60

Can you argue that the popularity patterns had changed? Use a 5% level of significance.

5. The following data represent gallons of gasoline sold at each gas station.

Gallons	Number of days
< 1400	20
1400 – 1600	30
1600 – 1800	50
1800 – 2000	45
2000 – 2200	35
> 2200	20
Total	200

Do a 5% test to determine if there is enough evidence to show that the data do not come from a normal distribution with a mean of 1800 and a standard deviation of 200.

6. Six hundred economists were asked about their opinions regarding the new budget. Their background and opinions are summarized in the following table:

Background	Opinion on budget			
	Like it	Indifferent	Hate it	
Conservative	100	80	60	240
Liberal	80	60	40	180
Radical	60	50	70	180
	240	190	170	600

Do a 5% test to determine if the economists' opinions depend on their background.

7. Use the data from Question 6 to determine if radical economists' opinions are equally split. Use a 5% test.

8. Use the data from Question 6 to determine if the pattern of reactions to the budget is the same for liberal and radical economists. Use a 5% test.

Answers to Supplementary Exercises

Multiple Choice

1.b	6. b	11. e	16. b
2.d	7. a	12. c	17. b
3.a	8. b	13. a	18. c
4.d	9. c	14. c	19. d
5.b	10. d	15. b	20. e

True/False

1. True
2. True
3. False. ANOVA uses between group and within group variances
4. False. The numerator will be small
5. False. Conclude treatment means are not all the same
6. False. Check the results from pairwise comparisons only when *F*-test is rejected
7. True
8. False. It does not change
9. False. Goodness of fit tests are used to test if data come from a normal distribution
10. False. ANOVA assumes equal population variances
11. True
12. True
13. False. The null hypothesis is that the data are independent
14. False. Compare the observed frequencies with the expected frequencies assuming the null hypothesis is true
15. False. Use chi-square statistic
16. True
17. True
18. True
19. True
20. False. Contingency table uses a chi-square statistic

Questions and Problems

1. $\bar{x}_1 = 27, \bar{x}_2 = 23, \bar{x}_3 = 37, \bar{x} = 29.$

$$\text{(a) } SST = \sum_{j}^{3} n_j(\bar{x}_j - \bar{x})^2 = 4[(27-29)^2 + (23-29)^2 + (37-29)^2] = 4[104] = 416$$

$$TSS = \sum_{j=1}^{3}\sum_{i=1}^{4}(x_{ij} - \bar{x})^2 = \left[(24-27)^2 + (30-27)^2 + \cdots + (40-27)^2\right] = 464$$

$$\left(\text{Also, } TSS = \sum_{j}^{3}\sum_{i=1}^{4} x_{ij}^{\ 2} - \frac{x_{..}^{\ 2}}{n} = 10556 - \frac{348^2}{12} = 464\right)$$

$$\text{Hence, } SSW = TSS - SST = 464 - 416 = 48.$$

$$\left(\begin{aligned}&\text{Also, } SSW = \sum_{j}^{3}\sum_{i=1}^{4} x_{ij}^{\ 2} - \sum_{j=1}^{3}\frac{x_{j}^{\ 2}}{n_j}\\ &\quad = \sum_{j}^{m}\sum_{i=1}^{n_j} x_{ij}^{\ 2} - \sum_{j=1}^{m}\frac{x_{j}^{\ 2}}{n_j} = 10556 - \frac{1}{4}[108^2 + 92^2 + 148^2] = 10556 - 10508 = 48\\ &\text{Or, } SSW = \sum_{i=1}^{4}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{4}(x_{i2} - \bar{x}_2)^2 + \sum_{i=1}^{4}(x_{i3} - \bar{x}_3)^2 = 20 + 8 + 20 = 48.\end{aligned}\right)$$

$$F = \frac{MST}{MSW} = \frac{SST/(m-1)}{SSW/(n-m)} = \frac{416/(3-1)}{48/(12-3)} = \frac{208}{5.333} = 3 > F_{2,9;0.05} = 4.256.$$

Therefore, we reject the null hypothesis and conclude that the three plans are not the same.

(b) $\bar{x}_1 - \bar{x}_2 =$ 4.95% CI for $\mu_1\ 0\ \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0.025,9}\sqrt{MSW\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= 4 \pm 2.262\ (1.633) = 4 \pm 3.69 = (0.31, 7.69)$$

(c) $\bar{x}_1 - \bar{x}_3 = -10$

95% CI for $\mu_1 - \mu_3$ is

$$(\bar{x}_1 - \bar{x}_3) \pm t_{0.025,9}\sqrt{MSW\left(\frac{1}{n_1} + \frac{1}{n_3}\right)}$$

$$= -10 \pm 2.262\ (1.633) = -10 \pm 3.69 = (-13.69, -6.31)$$

(d) $\bar{x}_2 - \bar{x}_3 = -14.$

95% CI for $\mu_1 - \mu_3$ is

$$(\bar{x}_2 - \bar{x}_3) \pm t_{0.025,9}\sqrt{\text{MSW}\left(\frac{1}{n_2} + \frac{1}{n_3}\right)}$$

$$= -14 \pm 2.262(1.633) = -14 \pm 3.69 = (-17.69, -10.31)$$

(e) Scheffé's 95% simultaneous confidence interval for $\mu_2 - \mu_3$ is

$$(\bar{x}_2 - \bar{x}_3) \pm \sqrt{(3-1)(F_{0.05;;3-1,12-3})\text{MSW}}\sqrt{\frac{1}{n_2} + \frac{1}{n_3}}$$

Since $\sqrt{(3-1)\times 4.256 \times 5.333 \times (4^{-1} + 4^{-1})} = 4.76,$

Scheffes simultaneous CI for $\mu_2 - \mu_3 \; -14 \pm 4.76 = (-18.76, -9.24).$

2. Since 12 observations remain the same, SST and TSS are those computed in Question 1.

$\text{SST} = 416$

$\text{TSS} = 464$

$$\text{SSB} = \sum_{i=1}^{4} 3(\bar{x}_{i.} - \bar{x})^2$$

$$= 3[(26.333 - 29)^2 + (30.333 - 29)^2 + (28.333 - 29)^2 + (31 - 29)^2]$$

$$= 3[13.335] = 40.00$$

$\text{SSE} = \text{TSS} - \text{SST} - \text{SSB} = 464 - 416 - 40 = 8$

$\text{MST} = \text{SST}/(J-1) = 416/(3-1) = 208$

$\text{MSB} = \text{SSB}/(I-1) = 40/(4-1) = 13.333$

$\text{MSE} = \text{SSE}/(I-1)(J-1) = 8/(4-1)(3-1) = 1.333$

(a) Since $\text{MST/MSE} = 208/1.333 = 156.04 > F_{2,6;0.05} = 5.143$,
so reject the null hypothesis that plan does not make any difference.

(b) Since $\text{MSB/MSE} = 13.333/1.333 = 10 > F_{3,6;0.05} = 4.757,$
so reject the null hypothesis that size of store does not make any difference.

3. $H_0: P_1 = 0.6,\ P_2 = 0.1,\ P_3 = 0.1,\ P_4 = 0.1.$

$H_1: P_1 \neq 0.6$ or $P_2 \neq 0.1$ or $P_3 \neq 0.1$ or $P_4 \neq 0.1.$

Test statistic is

$$\chi^2_{k-1} = \sum_{i=1}^{k} (f_i^o - f_i^e)^2 / f_i^e \text{ , where } k = 4.$$

f_i^o	f_i^e	$(f_i^o - f_i^e)^2/f_i^e$
650	600	4.167
140	100	16
110	100	1
100	100	0
Sum		21.167

Since $21.167 > \chi^2_{3,0.05} = 7.81$, we reject the null hypothesis.

4. $H_0: P_1 = 0.65,\ P_2 = 0.15,\ P_3 = 0.11,\ P_4 = 0.1.$
   $H_1: P_1 \neq 0.65 \text{ or } P_2 \neq 0.15 \text{ or } P_3 \neq 0.11 \text{ or } P_4 \neq 0.1.$

f_i^o	f_i^e	$(f_i^o - f_i^e)^2 / f_i^e$
700	650	3.846
100	150	16.667
140	110	8.182
60	100	16.000
	Sum	44.695

Since $44.695 > \chi^2_{3,0.05} = 7.81$, we reject the null hypothesis and conclude the popularities have been changed.

5. To solve this question, we need to compute the expected frequencies.

$$P_r(1400 > \text{gallonssold}) = P_r\left(\frac{1400-1800}{200} > \frac{\text{gallons}-1800}{200}\right)$$
$$= P_r(-2 > z) = 2.28\%$$
$$f_e = 200 \times .028 - 4.56$$

Assuming the normal distribution is true, we then obtain the probability of each interval.

$$p_1 = P(X \leq 1400) = P\left(\frac{X-1800}{200} \leq \frac{1400-1800}{200}\right) = P(Z \leq -2) = 0.0228.$$

$$p_2 = P(1400 \leq X \leq 1600) = P\left(\frac{1400-1800}{200} \leq Z \leq \frac{1600-1800}{200}\right)$$
$$= P(-2 \leq Z \leq -1) = 0.1359$$

$$p_3 = P(1600 \leq X \leq 1800) = P\left(\frac{1600-1800}{200} \leq Z \leq \frac{1800-1800}{200}\right)$$
$$= P(-1 \leq Z \leq 0) = 0.3413$$

$$p_6 = P(X \geq 2200) = P\left(Z \geq \frac{2200-1800}{200}\right) = 0.0228$$

Then, the expected frequency of each interval is obtained by

$$f_i^e = np_i = 200p_i,\ i = 1,\ \ldots,\ 6.$$

Gallons	f_i^o	p_i	f_i^e	$(f_i^o - f_i^e)^2/f_i^e$
< 1400	20	0.0228	4.550	52.462
1400–1600	30	0.1359	27.181	0.292
1600–1800	50	0.3413	68.269	4.889
1800–2000	45	0.3413	68.269	7.931
2000–2200	35	0.1359	27.181	2.249
> 2200	20	0.0228	4.550	52.462
Total	200	1.00	200	120.285

Hence, the chi-square statistic is $\sum_{i=1}^{6}(f_i^o - f_i^e)^2/f_i^e = 120.285$
The degrees of freedom of the chi-square statistic is $6-1=5$, so the critical value is

$$\chi^2_{5,0.05} = 11.07, \text{ and } 120.285 > \chi^2_{5,0.05}\chi = 11.07.$$

Hence, the null hypothesis that the data are normally distributed with mean of 1800 and standard deviation of 200 is rejected.

6. To test for the statistical independence of the economists' background and their opinion, we need to calculate the expected frequencies. The expected frequency for each cell is the product of the column total times the row total divided by the total number of economists, i.e.

$$f_{ij}^e = (f_{i.}^o \times f_{.j}^o)/n, i = 1,2,3, j = 1,2,3$$

For example, $f_{11}^e = (f_{1.}^o \times f_{.1}^o)/600 = (240)(240)/600 = 96$, and the expected frequencies are in the following table:

f_{ij}^e			$(f_{ij}^o - f_{ij}^e)^2/f_{ij}^e$		
96	76	68	0.17	0.21	0.94
72	57	51	0.89	0.16	2.37
72	57	51	2.00	0.86	7.08

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(f_{ij}^o - f_{ij}^e)^2}{f_{ij}^e} = \frac{(100-96)^2}{96} + \cdots + \frac{(70-51)^2}{51} = 14.68$$

Critical value is $\chi^2_{6,0.05} = 9.488$. Since $14.68 > 9.488$, the null hypothesis that the economists' opinions are independent of their background.

7. H_0: uniform vs. H_1: not uniform.
   The expected frequencies are computed under a uniform distribution.
   Thus, $f_i^e = 180/3 = 60$, $i = 1, 2, 3$.

f_i^o	f_i^e	$(f_i^o - f_i^e)^2/f_i^e$
60	60	0
50	60	1.667
70	60	1.667
180	180	$\chi^2 = 3.333$

Since the critical value is $\chi^2_{2,.05} = 5.991$ and $3.333 < 5.991$, so we cannot reject the null hypothesis that the opinions of the radical economists are uniformly distributed.

8. H_0: same pattern vs. H_1: different pattern

f_i^o	f_i^e	$(f_i^o - f_i^e)^2/f_i^e$
80	60	6.67
60	50	2
40	70	12.86
180	180	$\chi^2 = 21.52$

Since the critical value is $\chi^2_{2,.05} = 5.991$, and $21.52 > 5.991$, so reject the null hypothesis and conclude a different pattern.

Chapter 13
Simple Linear Regression and the Correlation Coefficient

Chapter Intuition

We are often interested in measuring the relationship between two variables. In some cases, we know that the relationship is linear, and so we can use linear regression analysis to measure it. For example, elementary economics textbooks state that consumption is a function of income. The equation for the consumption function is

$$\text{CONSUMPTION} = \alpha + \beta\ [\text{INCOME}]$$

The equation for the consumption function represents a straight line, with intercept of α and slope of β. Besides knowing that the consumption function is a straight line, we know that the slope coefficient, β, should be positive, because consumption should rise as income increases.

If we are interested in the actual values of α and β, and hence, the actual equation that describes the relation between consumption and income, then we will collect data and use regression analysis to estimate this relationship.

Simply stated, regression analysis tries to fit a line through a scatter plot of points of the dependent and independent variables. Although there are many ways to fit the line through these points, the most popular approach is the ***least squares method***. The least squares method fits the regression line by minimizing the sum-of-squared error terms. The slope β of the line measures the impact of x (the independent variable) on y (the dependent variable) and represents the expected increment in the response per unit change in x. The intercept term, α, measures the mean value of y when the value of x is zero.

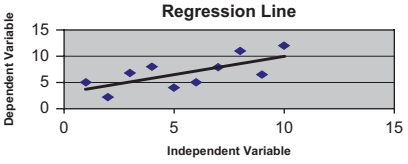
Once we have determined the regression line by estimating α and β by a and b, we are interested in how good a job the line does of explaining the relationship between the two variables. If most of the data points are close to the regression line, then the sum-of-squared error terms will be small and the regression line will explain much of the relationship between the two variables. Likewise, if the data points tend to be far away from the regression line, the sum-of-squared error terms

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_13

will be large and the regression line will not do a good job of explaining the relationship. This relationship can be seen by looking at the regression's coefficient of determination, or R^2. The coefficient of determination measures the amount of variance of the dependent variable that can be explained by the regression.

Chapter Review

1. ***Correlation analysis*** is used to measure the relationship between two different variables. The correlation coefficient, ρ, between two variables will lie between −1, perfect negative correlation and +1, perfect positive correlation. If ρ is positive, the two variables tend to move in the same direction (when one goes up, the other usually goes up). If ρ is negative, the two variables tend to move in opposite directions (when one goes up, the other usually goes down).
2. ***Simple linear regression*** is a systematic procedure for finding a linear relationship between two variables.
3. In simple linear regression, there are only two variables of interest. The ***independent variable*** or ***explanatory variable***, x, is the variable which we assume will explain the **dependent variable**, y.
4. In regression analysis, there are two parameters of interest. The **intercept** coefficient represents the mean value of y if x is zero. The **slope** coefficient, which is usually represented by β, measures the expected increment in the response per unit change in x.
5. A ***scatter diagram*** is used to depict the relationship between the x and y variables. A scatter diagram is just a plot of the x and y variables. Geometrically, regression analysis is just a systematic way of fitting a line through the scatter diagram to best describe the relationship between the x and y variables.
6. One method used for fitting a line through the scatter diagram is known as the ***least squares method***. The least squares method fits the line through the scatter diagram by minimizing the total-squared error terms from the regression line.



7. The assumptions for the linear regression model are:
   a. The values of the independent variable, x, are either fixed numbers or if they are random variables, they are statistically independent of the error term, ε_i.
   b. The error terms, ε_i, are assumed to have a mean of 0 and a constant variance. They are also assumed to be statistically independent of one another.
   c. The error terms, ε_i, are assumed to be normally distributed in order to make valid statistical inference.

8. Once we have estimated the regression line, we would like to know how well this line represents the relationship between x and y. To measure the goodness of fit of the regression line, we can use the **standard error of the residuals** or the **coefficient of determination**. The standard error of the residuals measures the variability of the observed values around the regression line. The coefficient of determination tells us the amount of variation of y that is explained by x.

Useful Formulas

Simple linear regression model: $y = \alpha + \beta x + \varepsilon$	Coefficient of determination: $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
Simple linear regression equation: $E(y) = \alpha + \beta x$	Adjusted coefficient of determination: $\bar{R}^2 = 1 - \frac{SSE/(n-2)}{SST/(n-1)} = 1 - \frac{MSE}{SST/(n-1)}$
Least squares estimates: $b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}$ $a = \bar{y} - b\bar{x}$	Standard error of the residual: $s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$
Sum-of-squares regression: $SSR = \sum(\hat{y}_i - \bar{y})^2 = b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$	Correlation coefficient: $\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$
Sum-of-squares total: $SST = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n$	Sample correlation coefficient: $r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$
Sum-of-squares residual: $SSE = \sum(y_i - \hat{y}_i)^2 = SST - SSR$ $SST = SSR + SSE$	

Example Problems

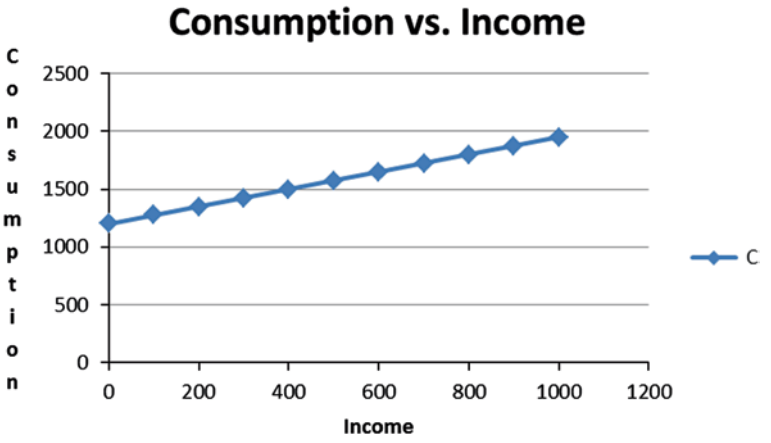
Example 1 Regression Line
Suppose you collect data on household consumption and income in the USA and estimate the following regression equation:

C = 1200 + 0.75 I
where C = consumption
I = income

a. What is the dependent variable? What is the independent variable?
b. Plot the relationship between consumption and income from the above equation.
c. Explain the relationship between consumption and income. How much more will consumption increase if income rises by $ 1?

Solution:

a. Because consumption depends on income, consumption is the dependent variable, while income is the independent variable.
b. Let I = 0, 100, 200, …, 1000, then the plot of C versus I and the regression line is:



c. The above equation for the consumption function says that consumption has two parts. The first part is represented by the intercept term of 1200. This intercept term means that consumption will be 1200 even if there is no income. The second part consists of the .75 I term. This says that for every $1 increase in income (I), consumption will increase by 75 cents.

Example 2 Scatter Diagram

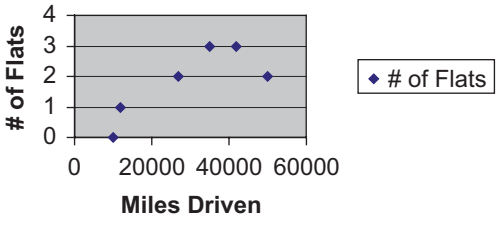
Suppose you are a quality control consultant for the Hug the Road Tire Company. You are interested in the relationship between the number of miles a tire has traveled and the number of flat tires. You collect the following information for six different tires.

Tire	Miles driven	flats
1	10,000	0
2	11,900	1
3	27,000	2
4	35,000	3
5	50,000	2
6	42,000	3

a. Draw a scatter diagram showing the relationship between miles traveled by the tire and the number of flat tires.
b. Is there a direct or an inverse relationship between miles driven and the number of flats?

Solution:

a.



b. There appears to be a direct relationship; that is, as the number of miles increases, the number of flats also increases.

Example 3 Estimating Regression Coefficients

Suppose you collect data for Orange Computer's sales (y) and dollars spent on research and development (x). You compute the following statistics:

S_{xy} = Sample Cov(sales, R&D) = 300
S_y^2 = Sample Var(sales) = 880
S_x^2 = Sample Var(R&D) = 1250
$\overline{y}$ = Sample Mean sales = 1200
$\overline{x}$ = Sample Mean R&D = 895

a. Compute the correlation coefficient between R&D and sales.
b. Calculate the coefficient of determination which would result from a regression of sales on R&D.
c. Calculate the regression parameters a and b.

Solution:

a. $r = \dfrac{S_{xy}}{S_x S_y} = \dfrac{300}{\sqrt{1250}\sqrt{880}} = .286$

b. $R^2 = r^2 = .286^2 = .081$

c. $b = \dfrac{S_{xy}}{S_x{}^2} = \dfrac{300}{1250} = .24$

d. $a = \bar{y} - b\bar{x} = 1200 - 0.24\ (895) = 985.2$

Example 4 Coefficient of Determination

Suppose you estimate a regression and compute sum of squared errors (SSE) and sum of squared residuals (SSR) as

SSE = 53.27
SSR = 202.91
Calculate total sum of squares (SST), R^2 and r using the above information.

Solution:

$$SST = SSE + SSR = 53.27 + 202.91 = 256.18$$

$$R^2 = SSR/SST = 202.91/256.18 = .792$$

$$r = \sqrt{0.792} = .890 \text{ or } r = -\sqrt{0.792} = -0.890$$

Example 5 Estimating Regression Coefficients

You are given the following table

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
20	75	–	–	–	–

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
14	85	–	–	–	–
12	72	–	–	–	–
20	88	–	–	–	–
33	92	–	–	–	–
38	99	–	–	–	–

Fill in the missing values and solve for the regression coefficients a and b.

Compute SSR.
Solution:

	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
	20	75	−2.83	−10.17	28.81	8.03
	14	85	−8.83	−0.17	1.47	78.03
	12	92	−10.83	−13.17	142.64	117.36
	20	88	−2.83	2.83	−8.03	8.03
	33	72	10.17	6.83	69.47	103.36
	38	99	15.17	13.83	209.81	230.03
Sum	137	511	0.00	0.00	444.17	544.83
Mean	22.83	85.16	–	–	–	–

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{444.17}{544.83} = 0.815$$

$$a = \bar{y} - b\bar{x} = 85.16 - .815(22.83) = 66.558$$

$$SSR = b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 0.815(444.17) = 361.999$$

Example 6 Estimating Regression Coefficients

Using the same data as in example 5, fill in the following table and compute the regression coefficients a and b, the correlation coefficient r, and R^2.

x_i	y_i	x_iy_i	x_i^2	y_i^2
20	75			
14	85			
12	72			
20	88			
33	92			
38	99			

Solution:

x_i	y_i	x_iy_i	x_i^2	y_i^2
20	75	1500	400	5625
14	85	1190	196	7225
12	72	864	144	5184
20	88	1760	400	7744
33	92	3036	1089	8464
38	99	3762	1444	9801
137	511	12112	3673	44043

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n} = \frac{12112 - (137)(511)/6}{3673 - (137)(137)/6} = \frac{444.17}{544.83} = 0.815$$

$$a = \bar{y} - b\bar{x} = 85.16 - .815(22.83) = 66.558$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n}{\sqrt{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}\sqrt{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}}$$

$$= \frac{12112 - (137)(511)/6}{\sqrt{544.83}\sqrt{44043 - (511)(5117)/6}} = 0.832$$

$$R^2 = r^2 = 0.832^2 = 0.69$$

Supplementary Exercises

Multiple Choice

1. A scatter diagram is

   a. a line that measures the relationship between the independent and dependent variables.
   b. another name for simple regression.
   c. another name for multiple regression.
   d. a graph of values of the independent and dependent variables.
   e. a line with slope of b and intercept of a.

2. A regression line

   a. is a line measuring the relationship between the independent and dependent variables.
   b. is another name for a scatter diagram.
   c. always has a slope equal to 1.
   d. is a graph of values of the independent and dependent variables.
   e. always has a slope equal to 0.

3. The slope coefficient

   a. is the point where the regression line intersects the *y*-axis.
   b. measures the fit of the regression line.
   c. measures the relationship between the independent and dependent variables.
   d. is always equal to 1.
   e. is another name for the coefficient of determination.

4. The intercept term

   a. is the point where the regression line intersects the y-axis.
   b. measures the fit of the regression line.
   c. measures the relationship between the independent and dependent variables.
   d. is always equal to 1.
   e. is always equal to 0.

5. Which of the following is not an assumption of the linear regression model?

   a. X_i's are either fixed numbers or statistically independent of the random variables ε_i.
   b. The variance of the random variable ε_i is assumed to be constant.
   c. The random variable ε_i is assumed to have a mean of 0.
   d. The variance of the random variable ε_i is assumed to be 0.
   e. The random variables ε_i are assumed to be statistically independent of one another.

6. The covariance between x and y is
   a. $\sum (x_i - \bar{x})^2$
   b. $\sum (y_i - \bar{y})^2$
   c. $\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)$
   d. $r_{xy} s_x s_y$
   e. both c and d.
7. The coefficient of determination
   a. is the point where the regression line intersects the *y*-axis.
   b. measures the fit of the regression line.
   c. measures the relationship between the independent and dependent variables.
   d. is always equal to 1.
   e. is always equal to 0.
8. The standard error of the estimate
   a. is the point where the regression line intersects the y-axis.
   b. measures the fit of the regression line.
   c. measures the relationship between the independent and dependent variables.
   d. is always equal to 1.
   e. is another name for the coefficient of determination.
9. The coefficient of determination measures the
   a. variation of the independent variable.
   b. slope of the regression line.
   c. intercept of the regression line.
   d. total variation of the dependent variable that is explained by the regression.
   e. is always equal to 1.
10. SST is
    a. $\sum (y_i - \bar{y})^2$
    b. $\sum (\hat{y}_i - \bar{y})^2$
    c. $\sum (y_i - \hat{y})^2$
    d. SSR − SSE
    e. SSE − SSR
11. SSR is
    a. $\sum (y_i - \bar{y})^2$
    b. $\sum (\hat{y}_i - \bar{y})^2$
    c. $\sum (y_i - \hat{y})^2$
    d. SSE + SST
    e. SSE − SST

12. SSE is

    a. $\sum(y_i - \bar{y})^2$

    b. $\sum(\hat{y}_i - \bar{y})^2$

    c. $\sum(y_i - \hat{y})^2$

    d. $SSE - SST$

    e. $SSR + SST$

13. If we are interested in measuring the relationship between work experience and earnings using regression analysis,

    a. the independent variable should be earnings.
    b. the independent variable should be work experience.
    c. the dependent variable should be earnings.
    d. the dependent variable should be work experience.
    e. both b and c.

14. If you estimate the following regression, $\hat{y} = .34 + 1.2x$, then the slope coefficient is

    a. x
    b. y
    c. .34
    d. 1.2
    e. 1.2 / .34

15. If you estimate the following regression, $\hat{y} = .34 + 1.2x$, then the intercept term is

    a. x
    b. y
    c. .34
    d. 1.2
    e. 1.2 / .34

16. If you estimate the relationship between earnings (in dollars) and education (in years) as $EARN = 12,201 + 525\ EDUC$, then a person with two additional years of education would be expected to earn an additional

    a. $\$12,201$
    b. $\$525$
    c. $\$24,402$
    d. $\$1,050$
    e. $\$12,201 + \$525$

17. If you estimate the relationship between earnings (in dollars) and education (in years) as $EARN = 12,201 + 525\ EDUC$, then a person with zero years of education would be expected to earn

a. $\$12,201$
b. $\$525$
c. $\$24,402$
d. $\$1,050$
e. $\$12,201 + \$525$

18. If a regression has an R^2 of .80, then the regression line

a. explains 80% of the variation in x.
b. explains 80% of the variation in y.
c. will have a slope of .80.
d. will have an intercept term of .80.
e. will not do a good job of explaining the relationship between x and y.

19. For $n = 10$ observations, the unexplained variation is 4, and the explained variation is 8. What is the adjusted coefficient of determination in this simple regression analysis?

a. 4/12
b. 8/12
c. 5/8
d. 8/13
e. 4/8

20. For $n = 10$ observations, the total variation is 12, and the explained variation is 8. What is the sample standard deviation of the error term for a simple regression?

a. $\sqrt{0.4}$
b. $\sqrt{0.5}$
c. $\sqrt{0.6}$
d. $\sqrt{0.7}$
e. $\sqrt{0.8}$

21. If $r = -1$, then the two variables are

a. perfectly negatively related
b. perfectly positively related
c. not related
d. positively related
e. none of the above

22. Suppose $SST = 4SSE$. Then R^2 is

a. 0.25
b. 0.5
c. 0.75
d. 0.9
e. undetermined

23. Suppose $S_{xy} = 10/(n-1), S_x^2 = 5/(n-1), S_y^2 = 30/(n-1)$. Then, the slope b = ?

    a. 1/6
    b. 0.5
    c. 1/3
    d. 2
    e. None is correct.

24. Suppose $S_{xy} = 10/(n-1), S_x^2 = 5/(n-1), S_y^2 = 30/(n-1)$. Then, the SST = ?

    a. 30
    b. 20
    c. 10
    d. 5
    e. None is correct.

25. Suppose $S_{xy} = 10/(n-1), S_x^2 = 5/(n-1), S_y^2 = 30/(n-1)$. Then, the SSR = ?

    a. 30
    b. 20
    c. 10
    d. 5
    e. None is correct.

True/False (If false, explain why)

1. A scatter diagram measures the relationship between the dependent and independent variables.
2. In regression analysis, the dependent variable is usually placed on the x-axis and the independent variable is usually placed on the y-axis.
3. If we are only interested in measuring the relationship between two variables, correlation analysis can be as useful as regression analysis.
4. The fit of the regression model can be measured by using either the coefficient of determination or the standard error of the estimate.
5. Multiple regression is used when the independent variable is influenced by two or more dependent variables.
6. The slope coefficient in a regression measures the relationship between the dependent and independent variables.
7. The intercept term in a regression is an estimate of the value of the dependent variable when the independent variable is 0.
8. If we are interested in using regression analysis to measure the relationship between high school and college grades, the dependent variable should be high school grades and the independent variable should be college grades.
9. In correlation analysis, it is important to know which is the independent variable and which is the dependent variable.

10. The method of least squares fits the regression line by minimizing the sum of the squared residuals.
11. The coefficient of determination measures the percentage of the independent variable's variance that is explained by the regression.
12. If we are interested in comparing the relationship between two pairs of variables, A and B, and *X* and *Y*, we can use either covariance or correlation.
13. If two variables have a correlation coefficient of −1, they will always move in opposite directions.
14. If an economist reverses the roles of the explanatory variable and the explaned variable, the R^2 will decrease.
15. The R^2 of a simple regression is equal to the square of the correlation coefficient between *x* and *y*.
16. If $b \leq 0$, this implies that *x* has little impact on *y*.
17. If the correlation coefficient equals 1, the data points lie on a straight line.
18. Covariance will always lie between −1 and + 1.
19. The first step in doing a simple linear regression analysis is to construct a scatter diagram.
20. In regression analysis, the variable to be fitted or predicted is usually called as the independent variable.
21. The proportion of variability of the dependent variable (*y*) explained by the independent variable (*x*) is called the correlation coefficient.
22. In a simple linear regression, the correlation coefficient is the square root of the coefficient of determination.
23. If $\hat{y} = 5 - 3x$, and $R^2 = 0.64$. Then, $r = 0.8$.
24. The range of the coefficient of determination is −1–1.
25. Under simple linear regression model, the larger the R^2, the more appropriate the linear relationship is.

Questions and Problems

1. You are given the following information from a regression:

$$SSE = 28.6 \quad SSR = 30.1 \quad n = 50$$

   a. Compute the coefficient of determination.
   b. Compute the standard error of the estimate.

2. You are given the following information about two variables, x and y:

$$\bar{x} = 12, \ \bar{y} = 6, \ S_{xy} = 80, \ S_x^2 = 20, \ S_y^2 = 500$$

   c. Compute the parameters for the slope and intercept of a regression of y on x.

d. Compute r and R^2.
e. If $n = 10$, compute SST, SSE, and s_e.

3. Suppose you are interested in the relationship between income (unit: dollar) and years of schooling. You estimate the following regression:

$$\text{INCOME}_i = 12,000 + 1,250\ \text{SCHOOLING}_i$$

If income is measured in dollars and schooling is measured in years, interpret the results of this regression.

4. Compute the correlation coefficient between x and y, given the following information.

Observation	x	y
1	10	22
2	9	31
3	11	19
4	6	25

5. Use the information given in problem 4 to plot a scatter diagram.
6. Use the information in problem 4 to estimate the regression of y on x.
7. Suppose you estimate the consumption function as

$$\text{CONS} = 1,225 + .82\ \text{INCOME}$$

How much is consumption expected to increase if income rises by $1000? How much will you expect consumption to be if income is $0?

8. Suppose the R^2 from the regression in problem 7 is .75. Briefly explain what does this mean.

Answers to Supplementary Exercises

Multiple Choice

1. d	6. e	11. b	16. d	21. a
2. a	7. b	12. c	17. a	22. c
3. c	8. b	13. e	18. b	23. d
4. a	9. d	14. d	19. c	24. a
5. d	10. a	15. c	20. b	25. b

True/False

1. False. A scatter diagram is a plot of the independent and dependent variables.
2. False. x-independent variable and y-dependent variable.
3. True
4. True
5. False. Multiple regression is used when there are two or more independent variables.
6. True
7. True
8. False. High school grades would be the independent variable and college grades would be the dependent variable.
9. False. In correlation analysis, it doesn't matter which is the dependent variable and which is the independent variable.
10. True
11. False. It measures the percentage of the dependent variable's variance that is explained by the regression.
12. False. We need to use correlation to compare the degree of association between two pairs of variables because correlation is unit-free.
13. True
14. False. R^2 remains the same.
15. True
16. False. It may have significant negative impact.
17. True
18. False. Correlation will lie between -1 and $+1$; covariance can take on any value.
19. True.
20. False. Dependent variable.
21. False. Coefficient of determination.
22. False. The sign of correlation coefficient is the same as that of the slope.
23. False. r = –0.8.
24. False. 0–1.
25. False. The relationship may be nonlinear even R^2 being large.

Questions and Problems

1. a. $R^2 = \dfrac{SSR}{SST} = \dfrac{30.1}{28.6 + 30.1} = .513$

   b. $s_e = \sqrt{\dfrac{SSE}{n-2}} = \sqrt{\dfrac{28.6}{50-2}} = .772$

2. a. $b = \dfrac{S_{xy}}{S_x^2} = \dfrac{80}{20} = 4$

$$a = \bar{y} - b\bar{x} = 6 - 4(12) = -42$$

b. $$r = \frac{S_{xy}}{S_x S_y} = \frac{80}{\sqrt{20 \times 500}} = 0.8$$

$$R^2 = r^2 = 0.64$$

c. $$SST = (n-1)S_y^2 = (9)(500) = 4500$$

$$SSE = (1 - R^2)SST = (0.36)(4500) = 1620$$

$$s_e = \sqrt{MSE} = \sqrt{1620/8} = \sqrt{202.5} = 14.23$$

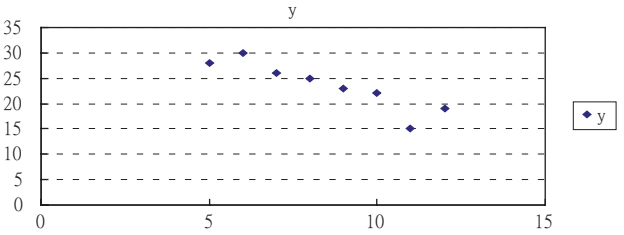
3. The intercept term says that a person with no schooling would be expected to earn \$12,000. The slope coefficient says that for each additional year of schooling, income is expected to rise by \$1250.

4.

	x	y	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
	10	22	−2.25	2.25	2.25
	9	23	−0.25	0.25	0.25
	11	15	−21.25	6.25	72.25
	6	30	−16.25	6.25	42.25
	8	25	−0.75	0.25	2.25
	12	19	−15.75	12.25	20.25
	7	26	−3.75	2.25	6.25
	5	28	−15.75	12.25	20.25
Mean	8.5	23.5			
Sum			−76	42	166

$$r_{x,y} = \frac{-76}{\sqrt{42}\sqrt{166}} = -0.91$$

5.



6.

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{-76}{42} = -1.81$$

$$a = \bar{y} - b\bar{x} = 23.5 - (-1.81)(8.5) = 38.89$$

7. An increase of \$1000 in income is expected to lead to an \$820 increase in consumption. If income is \$0, then consumption would be expected to be \$1225.
8. An R^2 of 0.75 means that 75% of the variation in consumption is explained by the regression of consumption on income.

Chapter 14
Simple Linear Regression and Correlation: Analyses and Applications

Chapter Intuition

In Chap. 13, we learned how regression analysis and correlation can be used to find a relationship between two variables. The estimated regression line provides an estimate of the relationship between the explanatory variables and the explained (or response) variables. However, as we learned in Chap. 10, estimates are nothing more than educated guesses about the true value. Since we are not absolutely certain that our estimates are meaningful, we use statistical theory to test the significance of the parameters we estimated. We can also use statistical theory to provide upper and lower bounds (known as a confidence interval) for the true values of the slope and intercept terms.

When we test the significance of the slope coefficient, we are interested in deciding whether the explanatory variable, x, relates to the y variable. If the slope coefficient is not statistically different from 0, then the explanatory variable is not useful in explaining values of the dependent variable, y.

If there is a significant relationship between y and x, statistical theory can also be used to provide the upper and lower bounds for the mean response and the individual response, respectively.

Chapter Review

1. Once we have estimated the regression coefficients, we are interested in testing the significance of these coefficients. Generally, we test whether the coefficients are significantly different from zero, although in some instances, we may be interested in seeing if the slope coefficient is different from some other number.
2. If the error terms, ε_i, follows a normal distribution, then b will follow a t-distribution with $n - 2$ degrees of freedom. Statistical inference can be done on the slope coefficient, b.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_14

3. Testing the significance of the slope being 0 in a simple regression can be done by using either a *t*-test or an *F*-test.
4. In using regression analysis to ***estimate*** the mean value of the response at a specified value of the explanatory variable, we are sometimes interested in constructing an upper and lower band around the estimate. This band is known as a **confidence band** (or confidence interval). This is done because we are uncertain of the true value of the mean response, and therefore would like to provide some upper and lower bounds on the likely value of the mean response. This confidence interval is not represented by parallel lines around the regression line, but rather by curves that get farther away from the regression line as we get farther away from the mean of x. This is because the farther we are away from the center of our data, the less certain we are about the value of our estimated mean response.
5. In using regression analysis to ***forecast*** the value of y corresponding to a specified value of x, we are also interested in constructing an upper and lower band around the forecast. This band is known as a **prediction band** (or prediction interval). Just like a confidence interval, a prediction interval is not represented by parallel lines around the regression line, but rather by curves that get farther away from the regression line as we get farther away from the mean of x. Moreover, a prediction band is wider than a confidence band for the same value of explanatory variable.
6. One of the assumptions of the simple linear regression model is that the independent variable, x, is measured without error. However, in economics and business, very often the x variable is measured with error. This results in an errors-in-variable problem. The result of measurement error in the x variable is that the estimator of the slope parameter will be biased.
7. The ***market model*** represents one of the most commonly used regressions in finance. It is used to measure the relationship between the returns on a company's stock and the returns on some market index like the DJIA or S&P 500. The estimated slope coefficient is one of the most commonly used measures of a stock's risk.

Useful Formulas

Significance test for $\beta = b_0$: $t_{n-2} = \dfrac{b - b_0}{s_b}$ $s_b = \dfrac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$, $s_e = \sqrt{MSE}$,	*F*-test for the significance of $\beta = 0$: $F_{1,n-2} = \dfrac{SSR/1}{SSE/(n\text{-}2)} = \dfrac{MSR}{MSE}$

Significance test for $\alpha = a_0$: $t_{n-2} = \frac{a - a_0}{s_a}$ $s_a = \sqrt{s_e^2 \frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}}$	Significance test for $\rho = 0$: $t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
Point estimate for $E(Y \mid x_{n+1})$ and y_{n+1} at x_{n+1}: $\hat{y}_{n+1} = a + bx_{n+1}$	Prediction interval for y_{n+1} at x_{n+1}: $\hat{y}_{n+1} \pm t_{\alpha/2,\, n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$
Confidence interval for $E(Y \mid x_{n+1})$: $\hat{y}_{n+1} \pm t_{\alpha/2, n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$	Market model: $R_{jt} = \alpha + \beta R_{mt} + \varepsilon_t$

Example Problems

Example 1 t-Tests for Parameter Estimates

When estimating the relationship between the price of a good and quantity of the good sold (the demand curve), economists sometimes choose to transform the price and quantity data by taking the natural logarithm of both. When this is done, the slope coefficient β can be interpreted as the price elasticity of demand (the sensitivity of quantity demanded to changes in price). Below you are given information on the price and quantity sold of Thirsty Time Cola.

Price ($)	Quantity
2.20	1020
1.85	1521
1.50	1755
1.21	2130
.99	3200
.79	4105

a. Estimate the elasticity of demand for the above data.
b. Use a *t*-test to test the significance of β. Use $\alpha = 0.05$.
c. At $\alpha = 0.05$, test if $\beta < -1$.
d. Construct a 90% confidence interval for the price elasticity.

Solution:

a.

	$p=\ln(P)$	$q=\ln(Q)$	$(p-\bar{p})^2$	$(q-\bar{q})^2$	$(p-\bar{p})(q-\bar{q})$	$\hat{q}$	e	e^2
	.788	6.928	.246	.493	−.348	6.986	−.058	.003
	.615	7.327	.104	.092	−.098	7.211	.116	.014
	.406	7.470	.013	.026	−.018	7.483	−.013	.000
	.191	7.664	.010	.001	−.003	7.762	−.098	.010
	−.010	8.071	.019	.194	−.133	8.022	.048	.002
	−.236	8.320	.279	.476	−.364	8.315	.005	.000
Mean	.292	7.630	.124	.214	−.161			
Sum			.744	1.282	−.965	45.780		.029

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{-.965}{.744} = -1.298$$

$$a = \bar{y} - b\bar{x} = 7.630 - (-1.298)(.292) = 8.009$$

b. $\hat{q} = a + bp = 8.009 - 1.298p$

$e = q - \hat{q}$

$$s_e^2 = \frac{SSE}{n-2} = \frac{\sum e^2}{n-2} = \frac{0.0291}{6-2} = 0.0073$$

$$s_e = \sqrt{0.0073} = 0.0853$$

$$s_b^2 = \frac{s_e^2}{\sum(x-\bar{x})^2} = \frac{0.0073}{0.744} = 0.0098$$

$$s_b = \sqrt{0.0098} = 0.099$$

$$H_0: \beta = 0 \text{ vs. } H_1: \beta \neq 0$$

$$t_{n-2} = \frac{b-0}{s_b} = \frac{-1.298\text{-}0}{0.099} = -13.11$$

$$\text{Critical value}: t_{\alpha/2,n-2} = t_{0.025,4} = 2.776$$

$$\text{Reject } H_0: \beta = 0.$$

c. $H_0 : \beta \geq -1$ vs. $H_1 : \beta - 1.$

$$t_{n-2} = \frac{b - 0}{s_b} = \frac{-1.298 - (-1)}{0.099} = -2.99$$

Critical value: $-t_{\alpha,n-2} = -t_{0.05,4} = -2.132.$

Since $-2.99 < -2.132$, so conclude $\beta < -1$.

d. $b \pm t_{\alpha/2,n-2} s_b = -1.298 \pm 2.132(0.099) = (-1.509, -1.087)$

Example 2 Confidence Interval and Prediction Interval

Suppose you use the data in Example 1 to estimate the demand for Thirsty Time Cola without using a log transformation, and you obtain the following results,

$$\hat{Q}_i = 5149.268 - 2009.907\, P_i$$

$$s_e = 468.943,\ R^2 = .867$$

$$\sum \left(P_i - \overline{P}\right)^2 = 1.426$$

Suppose the company is considering making a price change for Thirsty Time Cola, and would like to know what the quantity will be.

a. If it will charge \$1.75, forecast the demand for the cola.
b. If the price will stay at \$1.75 for several periods, construct a 95% interval for the conditional mean demand quantity.
c. If the price will be \$1.75 for the next period, construct a 95% interval for the demand quantity

Solution:

a. $\hat{Q}_{n+1} = 5149.268 - 2009.90(1.75) = 1631.9$

b.

$$\hat{Q}_{n+1} \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n} + \frac{(P_{n+1} - \overline{P})^2}{\sum_{i=1}^{n} (P_i - \overline{P})^2}}$$

$$= 1631.9 \pm (2.776)(468.943)\sqrt{\frac{1}{6} + \frac{(1.75 - 1.423)^2}{1.426}}$$

$$= 1631.9 \pm 639.9 = (992, 2271.8)$$

c. $\hat{Q}_{n+1} \pm t_{\alpha/2,n-2} s_e \sqrt{1+\frac{1}{n}+\frac{(P_{n+1}-\bar{P})^2}{\sum_{i=1}^{n}(P_i-\bar{P})^2}}$

$= 1631.9 \pm (2.776)(468.943)\sqrt{1+\frac{1}{6}+\frac{(1.75-1.423)^2}{1.426}}$

$= 1631.9 \pm 1450.6 = (181.3, 3082.5)$

Example 3 Confidence Interval for Parameter

Suppose you estimate a regression of y against x and find the following.

$b = 2.8$

$\sigma_b = .6$ standard error of the regression (assumed known)

a. Construct a 95% confidence interval for β.
b. Suppose σ_b is unknown and $s_b = 0.6$, how would this change the way you construct your confidence interval for b? Assume that 25 observations were used to estimate the regression.

Solution:

a. $b \pm z_{\alpha/2}\, \sigma_b$

$2.8 \pm 1.96(.6)$

So the interval is from 1.624 to 3.976.

b. Since σ_b is unknown, we must use an estimate for σ_b, s_b. In this case, we should use the student's t distribution to construct our confidence interval.

$b \pm t_{\alpha/2,n-2}\, s_b$

$2.8 \pm 2.69(.6)$

Example 4 Inference for β

Suppose you estimate the following simple regression:

$\hat{y} = 2300 + 10.12x$

$SSE = 28,225$

$n = 28$

$\sum (x - \bar{x})^2 = 3300$

a. Based on the information provided, test the significance of the slope at $\alpha = 0.05$.
b. Construct a 90% confidence interval for the slope coefficient.

So the interval is from 1.559 to 4.041, which is wider than that obtained in a.

Solution:

$$s_e^2 = \text{SSE}/(n-2) = 28{,}225/(28-2) = 1085.58$$

$$s_b^2 = \frac{s_e^2}{\sum (x-\bar{x})^2} = \frac{1085.58}{3300} = 0.329$$

$$s_b = \sqrt{0.329} = 0.574$$

a. $t_{n-2} = \dfrac{b-0}{s_b} = \dfrac{10.12}{0.574} = 17.63$

Critical value : $t_{0.025,26} = 2.056$

Since $17.63 > 2.056$, reject the null hypothesis that $\beta = 0$ and conclude that β is significantly different from 0.

b. $b \pm t_{\alpha/2,n-2}\, s_b$

$10.12 \pm 1.706(.574) = 10.12 \pm 0.98 = (9.14,\ 11.10)$

So the interval is from 9.14 to 11.10.

Example 5 Test of Significance

Suppose you estimate the following relationship between inches of rainfall and amount of corn in bushels for 100 farmers:

$\text{Bushels}_i = 5100 + 1100\ \text{Rainfall}_i$
SSE = 3221
SSR = 6755

Use an *F* test to test the significance of this regression at the 95% level.

Solution:

$$F_{1,n-2} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{6755/1}{3221/(100-2)} = 205.52$$

$$\text{Critical value} : F_{0.05;1,98} = 3.94$$

Since $205.52 > 3.94$, conclude that the regression is significant.

Supplementary Exercises

Multiple Choice

1. When examining the significance of the regression coefficient, the null hypothesis is usually

   a. $H_0: \beta = 1$
   b. $H_0: \beta = -1$
   c. $H_0: \beta = 0$
   d. $H_0: \beta = 0.5$
   e. $H_0: \beta = -0.5$

2. For a two-tailed test of the significance of β, the alternative hypothesis is usually

   a. $H_1: \beta \neq 1$
   b. $H_1: \beta > 1$
   c. $H_1: \beta \neq 0$
   d. $H_1: \beta > 0$
   e. $H_1: \beta < 0$

3. If we are interested in examining if the slope coefficient is *positive* and significant, the null hypothesis should be

   a. $H_0: \beta \leq 0$
   b. $H_0: \beta \geq 0$
   c. $H_0: \beta = 0$
   d. $H_0: \beta = 1$
   e. $H_0: \beta \geq 1$

4. If we are interested in examining if the slope coefficient is *negative* and significant, the null hypothesis should be

   a. $H_0: \beta \leq 0$
   b. $H_0: \beta \geq 0$
   c. $H_0: \beta = 0$
   d. $H_0: \beta = 1$
   e. $H_0: \beta = -1$

5. When examining the significance of a regression parameter, we usually use the

   a. *Z*-test
   b. chi-square test
   c. *t*-test
   d. binomial distribution
   e. any of the above

6. When examining the significance of a regression, we are

   a. most interested in examining the significance of the intercept term α
   b. most interested in examining the significance of the slope coefficient β
   c. equally interested in the intercept and the slope
   d. most interested in the variance of y
   e. most interested in the variance of x

7. If the slope coefficient β is insignificant, it means that

   a. the independent variable does a good job of explaining the dependent variable
   b. the independent variable does not do a good job of explaining the dependent variable
   c. the dependent variable does a good job of explaining the independent variable
   d. the dependent variable does not do a good job of explaining the independent variable
   e. none of the above

8. If the slope coefficient for a regression is 2.4 and the standard error of the slope coefficient is.8, then the *t*-value used to test H_0: $\beta = 0$ is

   a. $8./2.4$
   b. $2.4/\sqrt{8}$
   c. $(2.4-1)/.8$
   d. $2.4/.8$
   e. $(2.4-1)/\sqrt{.8}$

9. If the slope coefficient for a regression is 2.4 and the standard error of the slope coefficient is.8, then the *t*-value used to test H_0: $\beta = 1$ is

   a. $8./2.4$
   b. $2.4/\sqrt{.8}$
   c. $(2.4-1)/.8$
   d. $2.4/.8$
   e. $(2.4-1)/\sqrt{.8}$

10. Other things being equal, the greater the standard error of the slope coefficient, the

    a. larger the *t*-value for the slope coefficient
    b. smaller the *t*-value for the slope coefficient
    c. larger the intercept term
    d. smaller the intercept term
    e. larger the slope coefficient

11. In simple regression, a significant *t*-value for H_0: $\beta = 0$ implies that

    a. the F-statistic will be significant for the slope coefficient
    b. the F-statistic will not be significant for the regression
    c. the F-statistic could be significant or insignificant
    d. we cannot compare the results from a *t*-test and *F*-test in simple regression
    e. the R^2 will be very small

12. Which of the following statements could be true?

    a. $SSE + SSR > SST$
    b. $R^2 = -.5$
    c. $R^2 = 1.83$
    d. $s_e = -.35$
    e. $t = -2.3$

13. In a regression, it is always true that

    a. $r > 0$
    b. $r < 0$
    c. $t > 0$
    d. $t < 0$
    e. $s_e \geq 0$

14. In a regression, it is always true that

    a. $b \geq 0$
    b. $b \leq 0$
    c. $s_b \geq 0$
    d. $s_b \leq 0$
    e. $r \leq 0$

15. In simple regression, s_a is equal to

    a. $s_e^2 / \sum (x_i - \bar{x})^2$
    b. s_e^2
    c. $s_e^2 / \sum (yi - \bar{y})^2$
    d. $s_e^2 / [a/b]$
    e. $s_e^2 \left[\sum x_i^2 / \left(n \sum (xi - \bar{x})^2 \right) \right]$

16. In simple regression, s_b is equal to

    a. $s_e^2 / \sum (xi - \bar{x})^2$
    b. s_e^2
    c. $s_e^2 / \sum (y_i - \bar{y})^2$
    d. $s_e^2 / [a/b]$
    e. $s_e^2 \left[\sum x_i^2 / \left(n \sum (xi0\bar{x})^2 \right) \right]$

17. For a regression consisting of n observations, an $100(1-\alpha)\%$ confidence interval for the slope coefficient β would be

    a. $b \pm t(\alpha,\ n-2)\ s_b$
    b. $b \pm t(\alpha, n-2) s_b^2$
    c. $b \pm t(\alpha/2,\ n-2) s_b$
    d. $b \pm t(\alpha/2, n-2) s_b^2$
    e. $b \pm t(\alpha/2,\ n-2) s_e$

18. For a regression consisting of n observations, an $100(1-\alpha)\%$ confidence interval for the intercept term α would be

    a. $a \pm t(\alpha,\ n-2) s_a$
    b. $a \pm t(\alpha, n-2) s_a^2$
    c. $a \pm t(\alpha/2,\ n-2) s_a$
    d. $a \pm t(\alpha/2, n-2) s_a^2$
    e. $a \pm t(\alpha/2,\ n-2) s_e$

19. In a data set, if *x* and *y* have a strong negative correlation, then a scatter diagram would fit loosely around

    a. a horizontal line
    b. a vertical line
    c. a line going down to the right
    d. a line going up to the right
    e. All of the above are possible.

20. Which of the following is not the assumption for error terms to make valid statistical inferences?

    a. mean of zero
    b. constant variance
    c. normally distributed
    d. independent to each other
    e. variance of one

True/False (If False, Explain Why)

1. When we are testing the significance of the slope coefficient in a regression, the null hypothesis is usually $\beta = 1$.
2. A negative and significant slope coefficient implies an indirect relationship between the dependent and independent variables.
3. Given a level of significance of $\alpha = .05$, a two-tailed test of the significance of β will have a smaller absolute critical value than a one-tailed test.
4. A *t*-test can be used to test whether the correlation coefficient for two normally distributed random variables is significant.
5. Predictions or forecasts are one of the important uses of regression analysis.
6. The confidence interval for the mean response will get wider the farther we are from the mean of x.
7. The predicted value of y will always be equal to the actual value of y.
8. $t_b = b/s_b$
9. The *t*-statistic for a coefficient can never be negative.
10. In simple regression, a high R^2 would imply a significant slope coefficient.
11. In predicting the value of y using a regression, we are confident about the accuracy of the prediction if the place of prediction is far away from the center of our data.
12. The errors-in-variable problem is due to mismeasurement of the explanatory variable.
13. A negative *t*-statistic for the slope coefficient implies that the explanatory variable is not significant.
14. A *t*-test can be used to test the significance of the covariance between two random variables.
15. The residual in regression analysis is the difference between the fitted value of y and the observed value of y.
16. In simple regression, having R^2 close to 1 would imply y is linearly related to x.
17. In simple regression, an *F*-test can help us to know whether $\beta > 0$.
18. In simple regression, if $R^2 = 0.64$ and $b = 2$, then $r = 0.8$.
19. $s_y^2 = \sum (y_i - \bar{y})^2/(n-1)$. s_y^2 must be greater than s_e^2.
20. In simple regression, if R^2 is close to 0, then the explanatory variable is of no use to account for the variation in the dependent variable.

Questions and Problems

1. You are given the following information:

$$\sum x_i^2 = 92, \sum (x_i - \bar{x})^2 = 3.7, s_e^2 = 1.25, n = 40$$

   a. Compute s_a
   b. Compute s_b.

2. Suppose you estimate the following regression of earnings against years of schooling using 30 observations. (Standard errors are reported in parentheses)

$$\underset{}{\text{EARN}_i} = \underset{(4825)}{11,929} + \underset{(127)}{421}\ \text{SCHOOLING}_i$$

   a. Test the significance of the slope coefficient at the 5% level of significance.
   b. Construct a 95% confidence interval for the slope coefficient.
   c. Test whether the slope is greater than 300 at the 5% level of significance.

3. Suppose the result from a simple regression for 30 paired observations is SSE = 75 and SSR = 81. At $\alpha = 0.05$, test the significance of the regression.
4. Suppose you compute the correlation between two random variables to be.62. If 25 observations were used to estimate the correlation coefficient, use a 5% level of significance to test the significance of the correlation coefficient.
5. You are given the following information:

$$\hat{y} = 50 + 1.5x,\ \text{MSE} = 12,\ \Sigma(x_i - \bar{x})^2 = 9,\ \bar{x} = 10.8,\ n = 18.$$

   a. Suppose $x_{19} = 12$, find the 95% confidence interval for $E(Y_{19} \mid x_{19} = 12)$.
   b. Suppose $x_{19} = 12$, find the 95% prediction interval for Y_{19}.

Answers to Supplementary Exercises

Multiple Choice

1. c	6. b	11. a	16. a
2. c	7. b	12. e	17. c
3. a	8. d	13. e	18. c
4. b	9. c	14. c	19. c
5. c	10. b	15. e	20. e

True/False

1. False. $\beta = 0$.
2. True
3. False. Larger absolute critical value.

4. True
5. True
6. True
7. False. The predicted value is just a guess, and therefore does not have to equal the actual value.
8. False. Unless H_0: $\beta = 0$.
9. False. When the coefficient is negative, the *t*-value will be negative.
10. True
11. False. We become less confident as we move far away from the center of the data.
12. True
13. False. The relation between x and y may be significantly negative.
14. False. A *t*-test can be used to test the significance of the correlation coefficient.
15. False. The residual is the difference between the observed value of y and its fitted value.
16. False. The relation between y and x may be nonlinear.
17. False. *t*-test.
18. True.
19. False. If $SST > (n-1)SSR$, then $SSE = SST - SSR > SST - SST/(n-1) = (n-2)SST/(n-1)$, so $s_e^2 = SST/(n-2) > SST/(n-1) = s_y^2$.
20. False. R^2 close to 0 only means no significant linear relationship among x and y. They may have a nonlinear relationship.

Questions and Problems

1. a. $$s_a^2 = s_e^2 \frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2} = 1.25\left(\frac{92}{(40)(3.7)}\right) = .777$$

$$s_a = \sqrt{.777} = .881$$

b. $$s_b^2 = \frac{s_e^2}{\sum (x_i - \bar{x})^2} = \frac{1.25}{3.7} = .338$$

$$s_b = \sqrt{.338} = .581$$

2. a. $t_b = 421/127 = 3.31$

$t_{.05/2,\ 30-2} = 2.048$, so we reject $H_0: \beta = 0$.

b. $b \pm t_{\alpha/2,\ n-2} s_b$

$421 \pm (2.48)\,127$

So the interval is from 160.90 to 681.10.

3. $$F_{1,n-2} = \frac{SSR/1}{SSE/(n-2)} = \frac{81/1}{75/(30-2)} = 30.24$$

$F_{.05,\ 1,\ 28} = 4.20$, so we reject $H_0: \beta = 0$.

4. $t_{n-2} = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{.62\sqrt{25-2}}{\sqrt{1-.62^2}} = 3.79$

   $t_{.05/2,\,23} = 2.69,$ so reject $H_0: \rho = 0.$

5. $\hat{y}_{n+1} = 50 + (1.5)(12) = 68$

   a. $$\hat{y}_{n+1} \pm t_{\alpha/2,n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

   $$= 68 \pm (2.12)\sqrt{12}\sqrt{\frac{1}{18} + \frac{(12-10.8)^2}{9}}$$

   $$= 68 \pm 3.41 = (64.59,\ 71.41)$$

   b. $$\hat{y}_{n+1} \pm t_{\alpha/2,n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

   $$= 68 \pm (2.12)\sqrt{12}\sqrt{1 + \frac{1}{18} + \frac{(12-10.8)^2}{9}}$$

   $$= 68 \pm 8.10 = (59.9,\ 76.1)$$

Chapter 15
Multiple Linear Regression

Chapter Intuition

In previous chapters, we learned how regression analysis can be used to measure the relationship between two variables. In this chapter, we will learn how we can relate the dependent variable to two or more independent or explanatory variables. For example, suppose a farmer is interested in knowing what factors affect the amount of corn produced. In simple regression, he might assume that corn production depends on either rainfall or the amount of fertilizer. However, in real life, the farmer knows that both rainfall and the amount of fertilizer used determine the amount of corn.

In this case, he could use multiple regression to measure the relationship between a dependent variable (corn production) and two independent variables (rainfall and fertilizer). The slope coefficients for rainfall and fertilizer will measure how each factor affects corn production assuming no change in the other factor. That is, the coefficient for rainfall will measure the effect of changes in rainfall on corn production assuming we do not change the amount of fertilizer. Likewise, the coefficient for fertilizer will measure the effect of changes in the amount of fertilizer assuming no change in the amount of rainfall.

As in simple regression, our estimates of the coefficients are just educated guesses about the actual parameter value. To judge the significance of the coefficients we use statistical theory to test the significance of the coefficients or to construct confidence intervals around the estimates.

Two kinds of significance tests are involved in multiple regression: an F-test for overall significance of all slope coefficient and a t-test for the significance of each individual slope coefficient.

In case that there is a significant relation between the response and the explanatory variables, statistical theory can also be applied to provide the upper and lower bounds for the mean response and the individual response, respectively.

One consideration that is not a concern in simple regression but may be a problem in multiple regression is ***multicollinearity***. Multicollinearity is a phenomenon where two explanatory variables are too highly related. When this problem occurs,

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_15

it is impossible to separate their effects on the dependent variable. For example, serious multicollinearity may occur when we estimate a regression that explains an individual's wages based on experience and age, because as a person's age increases by 1 year, so does experience. In this case, if a person's wages increase by $ 500, it will be impossible to determine whether age or experience was the cause.

Chapter Review

1. ***Multiple linear regression*** is just a natural extension of the simple linear regression that was discussed in the previous chapters. In multiple regression, we assume that the dependent variable, y, can be influenced by more than one factor. For example, the amount of sales a company generates may depend on both the price it charges and the advertising it spends to promote the product.
2. To make valid statistical inferences, the assumptions of the multiple linear regression model are:
   a. The error term is normally distributed with a mean of zero and a constant variance.
   b. The error terms are assumed to be independent of the k independent variables.
   c. Error terms are assumed to be independent of one another.
   d. The independent variables are not highly linearly related to each other. If they are, we say ***multicollinearity*** exists. Multicollinearity is sometimes a problem in multiple regression.
3. The same technique, the ***least squares method*** that was used to derive the simple linear regression model is also used in multiple linear regression, that is, we again minimize the squared error terms in order to find the "best" values for the slope coefficients. In simple regression, we fit a regression line to describe the relationship between x and y. In multiple regression, we fit a ***regression plane*** to describe this relationship.
4. The effect of each explanatory variable on the dependent variable is measured by the ***partial regression coefficient***. A partial regression coefficient measures the effect of the explanatory variable on the dependent variable, assuming that all other explanatory variables are held constant.
5. Because it is quite tedious to estimate the partial regression coefficients manually, we usually use a computer program.
6. When testing the statistical significance of the slope coefficients, there are two ways we can do this. By using an F-test, we can jointly test the significance of all the slope coefficients simultaneously. By using a t-test, we can test the significance of each individual slope coefficient. Recall that the F- and t-tests were equivalent in the simple linear regression model because there was only one slope coefficient. However, when we have more than one independent variable these tests are not equivalent.

7. Given the values of the explanatory variables, we can obtain not only the point estimates of the mean and the individual response, but also their interval estimates.

Useful Formulas

Multiple regression model:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

Regression coefficients with two independent variables:

Let $x_{1i}' = x_{1i} - \bar{x}_1$, $x_{2i}' = x_{2i} - \bar{x}_2$, $y_i' = y_i - \bar{y}$.

$$b_1 = \frac{\Sigma x_{1i}' y_i' \Sigma x_{2i}'^2 - \Sigma x_{2i}' y_i' \Sigma x_{1i}' x_{2i}'}{\Sigma x_{1i}'^2 \Sigma x_{2i}'^2 - (\Sigma x_{1i}' x_{2i}')^2}$$

$$= \frac{\left(\sum_{i=1}^{n} x_{2i}^2 - n\bar{x}_2^2\right)\left(\sum_{i=1}^{n} x_{1i} y_i - n\bar{x}_1\bar{y}\right) - \left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)\left(\sum_{i=1}^{n} x_{2i} y_i - n\bar{x}_2\bar{y}\right)}{\left(\sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2\right)\left(\sum_{i=1}^{n} x_{2i}^2 - n\bar{x}_2^2\right) - \left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)^2}$$

$$b_2 = \frac{\Sigma x_{2i}' y' \text{å}\, x_{1i}'^2 - \Sigma x_{1i}' y_i' \Sigma x_{1i}' x_{2i}'}{\Sigma x_{1i}'^2 \Sigma x_{2i}'^2 - (\Sigma x_{1i}' x_{2i}')^2}$$

$$= \frac{\left(\sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2\right)\left(\sum_{i=1}^{n} x_{1i} y_i - n\bar{x}_2\bar{y}\right) - \left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)\left(\sum_{i=1}^{n} x_{1i} y_i - n\bar{x}_1\bar{y}\right)}{\left(\sum_{i=1}^{n} x_{1i}^2 - n\bar{x}_1^2\right)\left(\sum_{i=1}^{n} x_{2i}^2 - n\bar{x}_2^2\right) - \left(\sum_{i=1}^{n} x_{1i}x_{2i} - n\bar{x}_1\bar{x}_2\right)^2}$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}$$

Sample Variances for coefficients with two independent variables:

$$s_{b1}^2 = \frac{s_e^2 \sum x_{2i}'^2}{\sum x_{1i}'^2 \sum x_{2i}'^2 - \left(\sum x_{1i}' x_{2i}'\right)^2} = \frac{s_e^2}{\left(1 - r_{x1,x2}^2\right)\sum \left(x_{1i} - \bar{x}_2\right)^2}$$

$$s_{b2}^2 = \frac{s_e^2 \sum x_{1i}'^2}{\sum x_{1i}'^2 \sum x_{2i}'^2 - \left(\sum x_{1i}' x_{2i}'\right)^2} = \frac{s_e^2}{\left(1 - r_{x1,x2}^2\right)\sum \left(x_{2i} - \bar{x}_2\right)^2}$$

Coefficient of determination:

$$R^2 = 1 - \frac{SSE}{SST}$$

Adjusted coefficient of determination:

F-ratio for $H_0 : \beta_1 = \ldots = \beta_k = 0$ **vs.** $H_1 : \beta_j \neq 0$, for some j = 1,…k.

$$F_{k,n-k-1} = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{SSR}{SSE}\frac{n-k-1}{k} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\left(\frac{n-k-1}{k}\right)$$

t-statistic for $H_{0j} : \beta_j = 0$ **vs.** $H_{1j} : \beta_j \neq 0$ $t_{n-k-1} = b_j / s_{b_j}$

Point estimate for mean response and individual response:

$$\hat{y}_{n+1} = a + b_1 x_{1,n+1} + \cdots + b_k x_{k,n+1}$$

Example Problems

Example 1 Multiple Regression

Suppose a National Football League (NFL) scout is interested in what physical attributes make for a good running back. He collects data on the weight and seconds in the 40-yard dash of seven running backs and their yards gained for the year. The data are summarized in the following table:

Y (yards)	x_1 (Weight)	x_2 (Seconds in 40-yard dash)
925	210	48
850	185	4.7
1622	225	4.7
1121	215	4.6
658	180	4.9
977	212	4.6
574	195	5.0

Estimate the regression coefficients, b_1 and b_2, and interpret the results.

Solution:

$$\bar{y} = 961, \bar{x}_1 = 203.14, \bar{x}_2 = 4.757$$

$$\sum (x_{1i} - \bar{x}_1)^2 = 1,674.857, \sum (x_{2i} - \bar{x}_2)^2 = .137$$

$$\sum (y_i - \bar{y})(x_{1i} - \bar{x}_1) = 28,417, \sum (y_i - \bar{y})(x_{2i} - \bar{x}_2)$$
$$= -197.9, \sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = -8.457$$

$$b_1 = \frac{(28,417)(.137) - (-197.9)(-8.457)}{(1,674.857)(.137) - (-8.457)^2} = 14.0577$$

$$b_2 = \frac{(-197.9)(1,674.857) - (28,417)(-8.457)}{(1,674.857)(.137) - (-8.457)^2} = -576.1310$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 = 961 - 14.0577\,(203.14) - (-576.131)(4.757) = 846.02$$

The coefficient b_1 indicates that if a running back weighs 1 lb more, he can be expected to gain an additional 14 yards, given that his time in 40-yard dash is unchanged. The coefficient b_2 indicates that if a running back reduces his time in the 40-yard dash for 0.1 s, he can be expected to gain an additional 57.6 yards, given that his weight is unchanged.

Note: For most multiple regression problems, it is much easier to use a computer program to solve for the coefficients.

Example 2 Significance of Regression Estimates

Use the data and your results from Example 1 to test the individual significance of β_1 and β_2 at $\alpha = 0.10$.

Solution: From Example 1, $\hat{y}_i = 846.02 + 14.0577x_{1i} - 576.131x_{2i}$

y_i	x_{1i}	x_{2i}	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
925	210	4.8	1032.718	11603.2
850	185	4.7	738.889	12345.7
1622	225	4.7	1301.197	102915
1121	215	4.6	1218.233	9454.24
658	180	4.9	553.374	10946.6
977	212	4.6	1176.060	39624.8
574	195	5	706.627	17589.8
				204479

Hence, $\text{SSE} = \sum \left(y_i - \hat{y}_i\right)^2 = 204479$

$$s_e^2 = \text{SSE}/(n-k-1) = 204479/4 = 51119.75$$

$$s_{b1}^2 = \frac{s_e^2 \sum \left(x_{2i} - \bar{x}_2\right)^2}{\sum \left(x_{1i} - \bar{x}_1\right)^2 \sum \left(x_{2i} - \bar{x}_2\right)^2 - \left[\sum (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right]^2}$$

$$= \frac{51119.75(0.137)}{(1674.857)(0.137) - (-8.457)^2} = 44.374$$

$$s_{b1} = 6.659 \quad t_{b1} = \frac{b_1 - 0}{s_{b1}} = \frac{14.058 - 0}{6.659} = 2.11$$

$$s_{b2}^2 = \frac{s_e^2 \sum \left(x_{1i} - \bar{x}_1\right)^2}{\sum \left(x_{1i} - \bar{x}_1\right)^2 \sum \left(x_{2i} - \bar{x}_2\right)^2 - \left[\sum \left(x_{1i} - \bar{x}_1\right)\left(x_{2i} - \bar{x}_2\right)\right]^2}$$

$$= \frac{51119.75(1674.857)}{(1674.857)(0.137) - (-8.457)^2} = 542112.3$$

$$s_{b2} = 736.283 \quad t_{b2} = \frac{b_2 - 0}{s_{b2}} = \frac{-576.131 - 0}{736.283} = -0.78$$

Critical value is $t_{.025,4=}2.132$, so neither variable is significant in explaining the number of yards a running back gains. However, β_1 is nearly significant, since the *p*-value is 0.102, which is very close to 0.10. Moreover, $r = 0.84$, which indicates the two explanatory variables are highly correlated.

Solution: A joint test means we are testing H_0: $\beta_1 = \beta_{2=} 0$. The joint significance can be tested by using an *F*-test.

$$F_{k,n-k-1} = [\text{SSR}/k]/[\text{SSE}/(n-k-1)]$$

Example 3 Joint Significance Test

Use the data and your results from Examples 1 and 2 to test the joint significance of β1 and β2 at a 5% level of significance.

From Example 2, $\text{SSE} = 204479$.

$$\text{Also, SST} = \sum \left(y_i - \bar{y}\right)^2 = 717972$$

$$\text{SSR} = 717972 - 204479 = 513493$$

$$F = [513493/2]/[204479/4] = 256746.5/51119.75 = 5.022$$

The critical value is $F_{0.05;2,4=}6.94$ so we are unable to reject the null hypothesis at a 5 % level of significance.
Solution: The regression says that a person is expected to earn $ 18,000 with neither experience nor educational background. For each additional year of experience, we expect his/her income to rise by $ 1,200, given that EDUC is held constant. For each additional year of schooling, we expect income to rise by $ 800, given that EXPER

Example 4 Interpreting Parameter Estimates

Suppose a labor economist is interested in the relationship between experience and education on income. She estimates the following regression.

$$\widehat{\text{INCOME}}_i = 18,000 + 1,200\ \text{EXPER}_i + 800\ \text{EDUC}_i$$

where INCOMEi = income for person i measured in dollars
EXPERi = years of experience for person i
EDUCi = years of education for person i.
Interpret the regression coefficients for EXPER and EDUC.

is held constant.
Solution:

$$SSE = \sum (y_i - \hat{y}_i)^2 = 285$$

Example 5 Coefficient of Determination

Suppose you estimate the following model

$$y_i = \alpha + \beta_1 \times_{1i} + \beta_2 \times_{2i} + \varepsilon_i$$

based on 20 observations and obtain

$$\sum (y_i - \hat{y}_i)^2 = 285, \sum (y_i - \bar{y})^2 = 425$$

Compute the R2. Does the model provide a good fit?

$$R^2 = 1 - SSE/SST = 1 - 285/425 = .3294$$

An R^2 of 0.3294 means that the independent variables x_1 and x_2 explain 32.94% of the variation in the dependent variable y. Whether this is a good fit or not depends on what the dependent variable is. For example, being able to explain 32% of the variation in stock prices may be quite impressive, while being able to explain 32% of the variation in sales may not be very good.

Example 6 Using Computer Programs to Do Multiple Regression Analysis

For the data in Example 1, use the MINITAB program to do the following:

a. Fit the multiple regression.
b. Test the joint significance of β_1 and β_2 at $\alpha = 0.05$.
c. Test the individual significance of β_1 and β_2 at $\alpha = 0.10$.
d. For a running back weighting 200 lbs and finishing a 40-yard dash in 4.8 s, forecast his yards gained. Then, find the 95% interval for his expected yards gained, and the 95% prediction interval for his yards gained.

Solution:

The following is MINITAB output:

```
Regression Analysis: y versus x1, x2
The regression equation is
y = 846 + 14.1 x1 - 576 x2

Predictor     Coef  SE Coef      T      P
Constant       846     4401   0.19  0.857
x1          14.058    6.658   2.11  0.102
x2          -576.1    735.7  -0.78  0.477

S = 226.097   R-Sq = 71.5%   R-Sq(adj) = 57.3%

Analysis of Variance
Source          DF      SS      MS     F      P
Regression       2  513493  256746  5.02  0.081
Residual Error   4  204479   51120
Total            6  717972

Predicted Values for New Observations
New Obs    Fit  SE Fit       95% CI           95% PI
      1  892.1    89.4  (643.8, 1140.4)  (217.1, 1567.2)

Values of Predictors for New Observations
New Obs   x1    x2
      1 200  4.80
```

a. The estimated regression equation is:
$\hat{y}_i = 846 + 14.058x_{1i} - 576.1x_{2i}$

b. $F - \text{test} = 5.02, \text{p} - \text{value} = 0.081$. Do not reject $\beta_1 = \beta_2 = 0$ at $\alpha = 0.05$.a.
$F - \text{test} = 5.02, \text{p} - \text{value} = 0.081$. Do not reject $\beta_1 = \beta_2 = 0$ at $\alpha = 0.05$.

c. t_{b1}a.$= 2.11, \text{p} - \text{value} = 0.1020$. Do not reject $\beta_1 = 0$ at α
$= 0.10,$ when x_2 is in the model.

$t_{b2} = -.78,$ $\text{p} - \text{value} = 0.477$. Do not reject $\beta_2 = 0$ at α
$= 0.10,$ when x_1 is in the model.

d. $\hat{y}_8 = 892.1$ 95% CI for $E(Y_8 \mid 200, 4.8)$ is from 643.8 yards to 1,140.4 yards.

95% PI for Y_8 is from 217.1 yards to $1,567.2$ yards is from
217.1 yards to $1,567.2$ yards.

Supplementary Exercises

Multiple Choice

1. In multiple regression there is
   a. More than one dependent variable but only one independent variable.
   b. More than one independent variable but only one dependent variable.
   c. More than one dependent variable and more than one independent variable.
   d. Only one dependent variable and only one independent variable.
   e. More than two dependent variables and more than one independent variable.

2. When estimating a regression model with two explanatory variables, the geometric interpretation is that we are fitting
   a. Straight line to describe the relationship between the dependent variable and the explanatory variables.
   b. Triangle to describe the relationship between the dependent variable and the explanatory variables.
   c. A plane to describe the relationship between the dependent variable and the explanatory variables.
   d. A circle to describe the relationship between the dependent and explanatory variables.
   e. An ellipse to describe the relationship between the dependent and explanatory variables.

3. In multiple regression, each slope coefficient indicates
   a. The total influence of all the independent variables on the dependent variable
   b. The expected influence of the independent variable on the dependent variable holding other independent variables constant

c. Where the regression plane intersects the *y* axis
d. Both the partial and total influence of the independent variables
e. A point that is equal to the intercept term

4. Multicollinearity occurs when
   a. The error term does not have a zero mean
   b. The error term is not independent of the explanatory variables
   c. Two error terms are correlated with one another
   d. Two independent variables are highly correlated with one another
   e. The variance of the error terms is not constant

5. To test the significance of an individual slope coefficient we use
   a. An *F*-test.
   b. A *t*-test.
   c. A chi-square test
   d. The binomial distribution
   e. The exponential distribution

6. To test the significance of an entire regression we use
   a. An *F*-test.
   b. A *t*-test.
   c. A chi-square test
   d. The binomial distribution
   e. The exponential distribution

7. Other things being equal, as we increase the number of explanatory variables in a regression,
   a. R^2 increases
   b. R^2 decreases
   c. R^2 can increase or decrease
   d. There is no effect on R^2
   e. The probability of multicollinearity decreases

8. 8. The relationship between R^2 and adjusted R^2 is
   a. Adjusted $R^2 = R^2$
   b. Adjusted $R^2 = R^2(n-1)/(n-k-1)$
   c. Adjusted $R^2 = [1-(1-R^2)](n-1)/(n-k-1)$
   d. Adjusted $R^2 = 1-R^2$
   e. Adjusted $R^2 = 1 + R^2$

9. Adjusted R^2 is
   a. A measure of goodness of fit
   b. Adjusted for the use of additional explanatory variables
   c. Measures the percentage of the total variation of the dependent variable explained by the regression
   d. Generally a better measure of goodness of fit than R^2 in multiple regression
   e. All of the above

10. For a regression with n-observations and k-explanatory variables, the relationship between R^2 and F is
    a. $F_{k,\,n-k-1} = [(n-k-1)/k][R^2/(1-R^2)]$
    b. $F_{k,\,n-k-1} = R^2/(1-R^2)$
    c. $F_{k,\,n-k-1} = [n/k][R^2/(1-R^2)]$
    d. $F_{k,\,n-k-1} = [(n-k-1)/k]R^2$
    e. $F_{k,\,n-k-1} = R^2/(1+R^2)$

11. If you estimate an earnings equation for 100 individuals with the number of years of schooling as one explanatory variable and the number of months of school as a second explanatory variable you may
    a. Get estimates that are not BLUE
    b. Suffer from serial correlation
    c. Have unequal variances of the error terms
    d. Suffer from multicollinearity
    e. Both b and d

12. The degrees of freedom for the *t*-statistic used to test the significance of the slope coefficient in a regression consisting of 35 observations and three explanatory variables is
    a. 35
    b. 3
    c. 32
    d. 33
    e. 31

13. The degrees of freedom of the numerator in an *F*-test in a regression consisting of 50 observations and four explanatory variables are
    a. 5
    b. 4
    c. 3
    d. 46
    e. 45

14. The degrees of freedom of the denominator in an *F*-test in a regression consisting of 50 observations and four explanatory variables is
    a. 50
    b. 4
    c. 3
    d. 46
    e. 45

15. One problem that may occur in multiple regression that will not occur in simple regression is
    a. Correlation between error terms
    b. Unequal variances of the error terms
    c. Correlation between the error terms and the explanatory variables

d. Correlation between explanatory variables
e. An error term without a zero mean

16. The coefficient of determination (R^2) of a three-independent variable multiple regression is 0.60. If the number of observations is 20, then the *F*-statistic for this regression is
    a. 1.5
    b. 8.5
    c. 8
    d. 32/9
    e. 16/3

17. In a multiple regression model, the error terms are usually assumed to be
    a. Independent of each other
    b. With equal variance
    c. Independent of each other
    d. a and b
    e. b and c

18. In a multiple regression model, to make valid statistical inferences, which of the following is an unnecessary assumption for the error terms?
    a. with mean 0
    b. with equal variance
    c. independent to each other
    d. normally distributed
    e. with variance 1

19. The estimated regression equation is:
    $\hat{y} = 800 + 15x_1 - 50x_2$.
    What can be concluded?
    a. Y increases 15 units as x_1 increases by 1 unit
    b. Y increases 15 units as x_1 increases by 1 unit holding x_2 constant
    c. Y is expected to increase 15 units as x_1 increases by 1 unit
    d. Y is expected to increase 15 units as x_1 increases by 1 unit holding x_2 constant
    e. None is correct

20. The estimated regression equation is:
    $\hat{y} = 800 + 15x_1 - 50x_2$.
    What can be concluded?
    a. Y is expected to decrease 50 units as x_2 increases by 1 unit
    b. Y is expected to decrease 50 units as x_2 increases by 1 unit holding x_1 constant
    c. Y decreases 50 units as x_2 increases by 1 unit
    d. Y decreases 50 units as x_2 increases by 1 unit holding x_1 constant
    e. None is correct

True/False (If False, Explain Why)

1. When multicollinearity is present, the least squares estimator is still BLUE.
2. It is possible to have insignificant t-values for the slope coefficients and a significant F-value for the regression.
3. If two explanatory variables are perfectly correlated, it will be impossible to use the least squares method.
4. R2 is a better measure of the fit of a regression than adjusted R2 in multiple regression.
5. If the correlation between each pair of explanatory variables is below .5, multicollinearity will never be a problem.
6. A t-statistic can never be negative.
7. A negative t-statistic on one of the slope coefficients implies that this regressor is not important.
8. Simple regression is just a special case of multiple regression.
9. The geometric interpretation of a regression consisting of two explanatory variables is a straight line.
10. Multicollinearity occurs when the explanatory variables are highly correlated with the dependent variable.
11. Adjusted R2 is always larger than R2.
12. A t-test is used to test the joint significance of the slope coefficients in a multiple regression.
13. An *F*-test is used to test the significance of individual slope coefficients.
14. SST = SSR + SSE.
15. SST can never be negative.
16. Adjusted R^2can be negative.
17. If the number of observations is 20, then the number of explanatory variables in a multiple regression model should not exceed 20.
18. After fitting a multiple regression model, we only need to check the results from the individual *t*-tests, in order to know whether any of the regression slopes is significant.
19. To test the significance of each individual slope, multicollinearity should be checked first.
20. If adjusted R^2 is 0.9, then the estimated regression equation will do a good job in forecasting.

Questions and Problems

1. Suppose you are an economist with the Department of Labor and are interested in examining the relationship between earnings and other factors such as age, education level, and work experience.
   a. Which of the above variables should be the dependent variable? Which variables could serve as explanatory variables?

b. Are there any possible problems you might encounter in estimating your regression?

2. 2. Suppose you estimate a regression of stock returns for a company against the earnings per share (EPS) and debt/equity ratio for the company.
   a. Write down the regression model that should be estimated.
   b. Interpret the meaning of the slope coefficients.
3. Suppose you estimate a regression using three explanatory variables that have the following relationship:

$$x_{3i} = 2x_{1i} + x_{2i}$$

   a. Are you likely to encounter any problems in estimating a regression using all three explanatory variables?
   b. Is there any way to solve this problem?
4. Suppose you estimate a regression using 50 observations and four explanatory variables. The R^2 for the regression is.64. Compute the adjusted R^2 for the regression.
5. Use an *F*-test to jointly test the significance of the slope coefficients for the regression given in Problem 4 at the 5% level of significance.
6. You estimate the following regression using 30 observations.

$$\hat{y}_i = 122 + 32.4x_{1i} - .78x_{2i}$$

$$r_{x1,x2}{}^2 = .59$$

$$\sum (x_{1i} - \bar{x}_1)^2 = 23.6$$

$$\sum (x_{2i} - \bar{x}_2)^2 = 12.3$$

$$s^2 = 41.7$$

Test the significance of the slope coefficients at the 5% level.

7. Use the information and your calculations from Problem 6 to construct 95% confidence intervals for the slope coefficients.
8. The sales volume (in 1000 units) (y) of a product over the past 9 months, the unit price (x_1), and the previous month's sales volumes (in 100 units) of its competitor (x_2) are shown below.

Month	y	x_1	x_1
1	28.24	5	15
2	18.63	5.5	17
3	20.44	5.3	16
4	22.24	5.2	17
5	37.61	4.9	12
6	28.3	5	16
7	32.46	4.95	11
8	20.05	5.2	18
9	24.25	5.1	14

Use MINITAB program to answer the following:

a. Write down the estimated regression equation.
b. Test the joint significance of β_1 and β_2 at $\alpha = 0.10$.
c. Test the individual significance of β_1 and β_2 at $\alpha = 0.10$.
d. The price for next month will be set to be 5.2, and sales volume of the 9th month for its competitor is 15,000 units. Forecast its sales volume, determine the 95% interval for its expected sales volume, and the 95% prediction interval for its sales volume for the 10th month.

Answers to Supplementary Exercises

Multiple Choice

1.	b	6.	a	11.	d	16.	c
2.	c	7.	a	12.	e	17.	d
3.	b	8.	c	13.	b	18.	e
4.	d	9.	e	14.	e	19.	d
5.	b	10.	a	15.	d	20.	b

True/False

1. True
2. True
3. True
4. False. Adjusted R^2 is better in multiple regression because it adjusts for the number of explanatory variables
5. False. It's possible for two or more explanatory variables to form a linear combination of another explanatory variable, thus leading to multicollinearity
6. False. The *t*-value will be negative when the coefficient is negative

7. False. Can be significantly negative
8. True
9. False. The geometric interpretation is a plane
10. False. Occurs when the explanatory variables are high correlated with one another
11. False. Adjusted R^2 is always less than or equal to R^2
12. False. An *F*-test is used to test joint significance
13. False. A *t*-test is used to test individual coefficients
14. True
15. True
16. True
17. False. Not exceeding 19, since total degrees of freedom is 19
18. False. It may happen that the *F*-test is significant, but none of the individual t-test is significant due to multicollinearity.
19. True
20. False. Even adjusted R^2 is high, there is no guarantee that the regression model is appropriate for this data set. Moreover, the data point to be forecasted may be far outside the ranges of the explanatory variables observed

Questions and Problems

1. a. Earnings would be the dependent variable, and age, work experience, and education level would be the explanatory variables.

   b. Because a person's age and work experience are likely to be highly correlated, multicollinearity may be a problem.

2. a. $\text{RETURN}_t = \alpha + \beta_1 \text{EPS}_t + \beta_2 \text{DE}_t + \varepsilon_t$

   b. The slope coefficient, $\beta 1$, measures the expected impact of changes in earnings per share on the stock return, assuming that the debt/equity ratio remains constant. The slope coefficient, $\beta 2$, measures the expected impact of changes in the debt/equity ratio on the st$_o$ck return, assuming that the EPS remains constant.

3. a. Because x_3 is a linear combination of x_1 and x_2, multicollinearity will be a problem.

   b. One method for dealing with this problem is to drop one of the explanatory variables from the regression.

4. $$\text{Adjusted } R^2 = 1 - \left(1 - R^2\right)\left[(n-1)/(n-k-1)\right]$$
$$= 1 - (1 - .64)\left[(50-1)/(50-4-1)\right] = .608$$

5. $F_{k,\,n-k-1} = \left[(n-k-1)/k\right]\left[R^2/\left(1-R^2\right)\right]$

$= \left[(50-4-1)/4\right]\left[.64/(1-.64)\right] = 20$

The critical value $F_{.05,\,4,\,45}$ is 2.58, so we can reject the null hypothesis that all slope coefficients are jointly equal to 0.

6. $s_{b1}^2 = \dfrac{s_e^2}{(1-r_{x1,x2}^2)\sum(x_{1i}-\bar{x}_1)^2} = \dfrac{41.7}{(1-.59)(23.6)} = 4.31$

$$s_{b1} = \sqrt{4.31} = 2.08$$

$$t_{b1} = b_1 / s_{b1} = 32.4/2.8 = 15.58$$

$$s_{b2}^2 = \frac{s_e^2}{(1-r_{x1,x2}^2)\sum(x_{2i}-\bar{x}_2)^2} = \frac{41.7}{(1-.59)(12.3)} = 8.27$$

$$s_{b1} = \sqrt{8.27} = 2.88$$

$$t_{b2} = b_2 / s_{b2} = -0.78/2.88 = -0.27$$

Critical value $t_{.05,\,27=}2.052$, so we reject the null hypothesis of $\beta_{1=}0$, but do not reject the null hypothesis of $\beta_{2=}0$.

7. $b_1 \pm s_{b1}t_{\alpha/2,n-k-1} = 32.4 \pm 2.08(2.052)$

95% CI for β_1 is from 28.13 to 36.67.

$$b_2 \pm s_{b2}t_{\alpha/2,n-k-1} = -.78 \pm 2.88(2.052)$$

95% CI for β_2 is from -6.69 to 5.13.

8. MINITAB output:

Regression Analysis: y versus x1, x2

```
The regression equation is
y = 143 - 19.4 x1 - 1.16 x2

Predictor      Coef  SE Coef      T      P
Constant     142.91    27.80   5.14  0.002
x1          -19.414    6.388  -3.04  0.023
x2          -1.1619   0.5183  -2.24  0.066

S = 2.45886   R-Sq = 88.8%   R-Sq(adj) = 85.1%

Analysis of Variance
Source          DF      SS      MS      F      P
Regression       2  288.04  144.02  23.82  0.001
Residual Error   6   36.28    6.05
Total            8  324.31

Predicted Values for New Observations
New Obs     Fit  SE Fit       95% CI            95% PI
      1  24.529   0.962  (22.175, 26.883)  (18.068, 30.990)

Values of Predictors for New Observations
New Obs    x1    x2
      1  5.20  15.0
```

a. The estimated regression equation is:
$\hat{y}_i = 143 - 19.4x_{1i} - 1.16x_{2i}$

b. $F - \text{test} = 23.82, p - \text{value} = 0.001$. Reject $\beta_1 = \beta_2 = 0$ at $\alpha = 0.05$.

c. $t_{b1} = -3.4, p - \text{value} = 0.023$. Reject $\beta_1 = 0$ at $\alpha = 0.10$, when x_2 is in the model.
$t_{b2} = -2.24,\ p - \text{value} = 0.06$. Reject $\beta_2 = 0$ at $\alpha = 0.10,$ when x_1 is in the model.

d. $\hat{y}_{10} = 24.529, \text{i.e.} 24,529$ units

95%CI for $E(Y_{10} \mid 5.2, 15)$ is from 22,175 to 26,883 units.

95% PI for Y_{10} is from 18,068 to 30,990 units.

Chapter 16
Other Topics in Applied Regression Analysis

Chapter Intuition

In the preceding chapters, we learned about linear regression. One of the things which we learned is that for linear regression to be valid, certain rules or assumptions must hold. What problems will occur when the rules are violated and how can we deal with these problems is the focus of this chapter.

Previously, we introduced the concept of ***multicollinearity***. Multicollinearity occurs when explanatory variables are closely related. When this occurs, it will be impossible to separate their impact on the dependent variable.

A second problem that can occur is ***heteroscedasticity***. Heteroscedasticity occurs when the variance of the error terms is not constant. For example, if we were estimating a regression that explains an individual's consumption based on his or her income, it would be reasonable to assume that the error terms will not have a constant variance because wealthier people will have a higher variance in their consumption. Heteroscedasticity, if it occurs, causes the least-squares method to give more weight to high variance observations, and thus causes the model to have a greater variance than it otherwise would.

Autocorrelation occurs when error terms are correlated with previous error terms. This problem will lead to a regression model that is not efficient.

This chapter also tells us how to deal with variables which are qualitative in nature. For example, if an economist believes that one factor influencing a person's earnings is the sex of the individual, she can use a dummy variable to capture this difference.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_16

Chapter Review

1. Previously, you learned that the simple and multiple linear regression models were based on several assumptions. When these assumptions are violated, there may be problems in the estimation of the coefficients or in statistical inference.
2. ***Multicollinearity*** results when two or more independent variables are highly correlated with another, or when one of the independent variables can be written as a linear combination of some of the remaining independent variables.

   a. If the independent variables are perfectly correlated, the least-squares approach cannot be used to estimate the regression coefficients.
   b. If the independent variables are highly but not perfectly correlated, there will be an increase in the standard error of the coefficients and hence the t-values will be extremely small.
   c. Detecting multicollinearity can be difficult. However, three simple methods may enable us to detect multicollinearity. First, we can examine the correlation between each pair of independent variables. If any pair has an absolute correlation of at least 0.8, multicollinearity may be a problem. The second method is to look at the t-values for the regression coefficients and the F-value for the entire regression. If the t-values are insignificant while the F-value is significant, a problem of multicollinearity may exist. The third method is based on k *variance inflationary factor*s (VIFs). If any of the VIFs is at least 10, then multicollinearity may be a problem.

3. ***Heteroscedasticity*** occurs when the variance of the error terms in the regression model is not constant. When heteroscedasticity is present, the parameter estimates will still be unbiased; however, the parameter estimates will be inefficient. Heteroscedasticity is a common problem in cross-sectional regressions.

   a. One way to detect heteroscedasticity is to look at a plot of the residuals against the independent variable or the fitted values of the response variable. If a constant relationship does not appear to hold, heteroscedasticity may be a problem.
   b. A second method for detecting heteroscedasticity is to run a regression using the squared residuals as the dependent variable and the estimated dependent variable as the independent variable, and obtain the R^2. If $nR^2 \geq \chi^2_{1,\alpha}$, then conclude existing heteroscedasticity.
   c. When heteroscedasticity is a problem, we sometimes estimate the model using a method known as weighted least squares.

4. ***Autocorrelation*** occurs when the error terms are correlated with past values of the error term. When autocorrelation is present, the parameter estimates will be unbiased but inefficient. Positive autocorrelation is indicated by the situation that high values for the residuals tend to be followed by high values, and when low values for the residuals tend to be followed by low values.

Negative autocorrelation is indicated by the situation that high values for the residuals tend to be followed by low values, and vice versa.

a. Autocorrelation can be detected by using the ***Durbin–Watson (DW) statistic*** and checking the DW table at the end of the text (Table A9). The problem with using the DW statistic is that the test results may be inconclusive depending on the value of the computed DW statistic. Let d denaote the value of DW statistic. We use the values for the upper DW value d_U and the lower DW value d_L and follow the following rules:

   1. For a one-tailed test of H_0: no autocorrelation vs. H_1: positive autocorrelation. We will reject H_0 if $d < d_L$. We will not reject H_0 if $d > d_U$. The test will be inconclusive if $d_L \le d \le d_U$.
   2. For a one-tailed test of H_0: no autocorrelation vs. H_1: negative autocorrelation. We will reject H_0 if $d > 4 - d_L$. We will not reject H_0 if $d < 4 - d_U$. The test will be inconclusive if $4 - d_U \le d \le 4 - d_L$.
   3. For a two-tailed test of H_0: no autocorrelation vs. H_1: positive or negative autocorrelation. We will reject H_0 if $d < d_L$ or $d > 4 - d_L$. We will not reject H_0 if $d_U < d < 4 - d_U$. The test will be inconclusive if $d_L \le d \le d_U$ or $4 - d_U \le d \le 4 - d_L$.

b. When a lagged-dependent variable is included in the regression model, the DW statistic will not be valid. In this case, it is necessary to use a different statistic known as ***Durbin's H***.

5. ***Specification error*** results when the regression model is incorrectly specified. This can result from the omission of a relevant variable or the inclusion of an irrelevant variable.
6. Sometimes there is not a linear relationship between the x and y variables. In this case, a nonlinear model may be a better choice for the regression. The simplest nonlinear model is the quadratic model in which a squared value of the independent variable is included as an independent variable in order to capture the nonlinear relationship.
7. When we estimate a regression over time, we sometimes believe that the current value of y, y_t depends on past values of y such as y_{t-1} or y_{t-2}. In this case, we can use a lagged-dependent variable as an independent variable in order to capture this effect.
8. In many instances, we are interested in incorporating some qualitative information into our regression, like the sex of the worker or geographic region of the country. When this is the case we can place a binary variable known as a ***dummy variable*** into the regression so we can capture these effects.
9. When two independent variables are assumed to work together in determining the value of the dependent variable as in the case of rainfall and fertilizer on crop production, we can capture this effect by using an interaction variable. An interaction variable can be created by multiplying the two independent variables together and using this new variable as an additional independent variable in the regression.

Useful Formulas

VIF for jth-independent variable:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2},$$

where R_j^2 is the coefficient of determination when the jth-independent variable x_j is regressed against all other independent variables.

Durbin–Watson statistic:

$$\mathrm{DW} = \frac{\sum_{t=2}^{n} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2}.$$

Durbin's H:

$$\mathrm{DH} = \left(1 - \frac{\mathrm{DW}}{2}\right)\sqrt{\frac{n}{1 - n\hat{V}(\hat{\gamma})}},$$

where $\hat{V}(\hat{\gamma})$ is the least-squares estimate of the variance of the coefficient of the lagged variable.

Quadratic regression model:

$$y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Lagged-dependent variable model:

$$y_t = \alpha + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \gamma y_{t-1} + \varepsilon_t$$

Dummy variable model:

$$y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \gamma D_{1i} + \varepsilon_i$$

Interaction variable model:

$$y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 \left(X_{1i} \times X_{2i}\right) + \varepsilon_i.$$

Example Problems

Example 1 Multicollinearity

You are interested in the relationship between y and three possible explanatory variables x_1, x_2, and x_3. You are given the following correlation matrix:

	Y	x_1	x_2	x_3
y	1.00	.55	.66	.89
x_1		1.00	.82	.75
x_2			1.00	.65
x_3				1.00

Does multicollinearity appear to be a problem? If so, between which variables?

Solution:

In the table, we have the correlation between each pair of variables. Because x_1 and x_2 have a correlation of .82, multicollinearity may be a problem.

Example 2 Multicollinearity

Suppose you are interested in the relationship between y and three possible explanatory variables x_1, x_2, and x_3. Let R_i^2 be the coefficient of determination of a regression with x_i being response and the other two variables being explanatory variable, for $i=1,2,3$. Now, $R_1^2 = 0.90$, $R_2^2 = 0.6$, and $R_3^2 = 0.5$. Does multicollinearity appear to be a problem? If it does, what can you do to resolve it?

Solution:

$$\text{VIF}_1 = \frac{1}{1-R_1^2} = 10 \geq 10,\ \text{VIF}_2 = \frac{1}{1-R_2^2} = 2.5,\ \text{VIF}_3 = \frac{1}{1-R_3^2} = 2.$$

Since $\text{VIF}_2 \geq 10$, so multicollinearity appears to be a problem.

Because x_1 is highly correlated to x_2 and x_3, we will use x_2 and x_3 only as independent variables.

Example 3 Heteroscedasticity

A financial analyst is interested in the relationship between dividend per share (DPS) and earnings per share (EPS). He collects information on these two variables to estimate the following regression model:

$$\mathrm{DPS}_i = \alpha + \beta \mathrm{EPS}_i + \varepsilon_i.$$

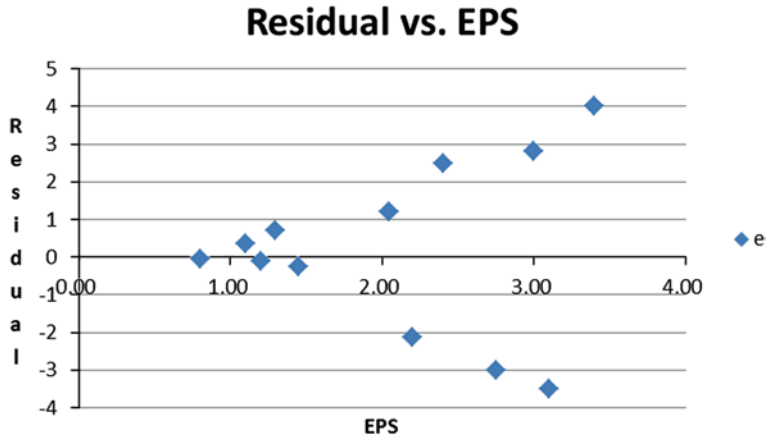
From the estimated regression equation, he computes the error from the regression for each company.

Company	EPS	E	Company	EPS	e
1	$ .80	−.05	7	2.20	−2.12
2	1.10	.35	8	2.40	2.50
3	1.20	−.10	9	2.75	−3.00
4	1.30	.72	10	3.00	2.80
5	1.45	−.25	11	3.10	−3.50
6	2.05	1.21	12	3.40	4.00

Does heteroscedasticity appear to be a problem in this regression?

Solution:

Method I:



From the plot of residuals versus the corresponding EPSs, the variations of residuals become larger as the independent variable increases in size. Hence, we conclude there is a problem of heteroscedasticity.

Method II:

Use MINITAB to fit a regression using e^2 as the dependent variable and EPS as the independent variable. The following is the output.

Regression Analysis: e^2 versus EPS

The regression equation is
e^2 = - 7.15 + 5.81 EPS

Predictor	Coef	SE Coef	T	P	VIF
Constant	-7.148	1.532	-4.67	0.001	
EPS	5.8097	0.6872	8.45	0.000	1.000

S = 2.01169 R-Sq = 87.7% R-Sq(adj) = 86.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	289.28	289.28	71.48	0.000
Residual Error	10	40.47	4.05		
Total	11	329.75			

Hence, $R^2 = 0.877$ and $nR^2 = 10.524$. Since $\chi^2_{1,0.05} = 3.84$, $nR^2 \geq \chi^2_{1,0.05}$, we conclude existing heteroscedasticity.

Example 4 Autocorrelation and DW Statistic

You are given the following error terms from a regression with two independent variables.

Period	E	Period	e
1	21.2	9	−9.6
2	−15.3	10	10.2
3	12.4	11	−13.3
4	−21.0	12	7.4
5	9.4	13	−15.1
6	−4.2	14	12.0
7	−8.4	15	−6.0
8	12.3	16	8.0

Compute the DW *d* statistic. Does autocorrelation appear to be a problem? Use $\alpha = 0.05$.

Solution:

$$d = \sum_{t=2} (e_t - e_{t-1})^2 \Big/ \sum e_t^2$$

$$\sum_{t=2}\left(e_t - e_{t-1}\right)^2 = (-15.3 - 21.2)^2 + (12.4 - (-15.3))^2 + (-21 - 12.4)^2 + \ldots + (8 - (-6))^2 = 8383.40$$

$$\sum_{t=1} e_t^2 = 21.2^2 + (-15.3)^2 + \ldots + 8^2 = 2506.00$$

$$d = 8383.4 / 2506 = 3.35.$$

From the table, we know for $n = 16$, $k = 2$, and $\alpha = 0.05$, $d_L = 0.98$ and $d_U = 1.54$.

Since $d = 3.35 > 4 - d_L = 3.02$, there exists negative autocorrelation.

Example 5 DW Statistic

Suppose you have a sample of 25 observations and two explanatory variables, and you want to test for autocorrelation. What can you say about autocorrelation for each of the following DW statistics at $\alpha = 0.05$?

a. $d = 1.20$
b. $d = 2.00$
c. $d = 3.25$
d. $d = 1.55$
e. $d = 2.55$
f. $d = 1.401$

Solution:

In order to determine whether or not autocorrelation is a problem, we need to examine the table of DW values. Remember, the DW statistic has upper and lower values, and a range where the test is inconclusive. For 25 observations, 2 explanatory variables and a 0.05 significance level,

$$d_L = 1.21,\ d_U = 1.55.$$

a. positive autocorrelation because $d < d_L = 1.21$.
b. no autocorrelation because $1.21 < d < 4 - 1.21$.
c. negative autocorrelation because $d > 4 - 1.21$.
d. no autocorrelation because $1.21 < d < 4 - 1.21$.

e. inconclusive because $4 - 1.55 \leq d \leq 4 - 1.21$.
f. inconclusive because $1.21 \leq d \leq 1.55$.

Example 6 Dummy Variables

Suppose you have been hired by a lawyer who is interested in showing that a company discriminates against women in the wages they pay. You estimate the following regression:

$$\widehat{WAGE}_i = 19{,}000 + \underset{(821)}{2000}\, EXPER_i + \underset{(332)}{1000}\, EDUC_i + \underset{(3400)}{4000}\, SEX_i,$$

where

$WAGE_i$ = wage for person i
$EXPER_i$ = years of experience for person i
$EDUC_i$ = years of education
SEX_i = 1 for female
= 0 for male

Standard errors of the coefficients are reported in the parentheses.

a. Interpret the coefficients for experience and education.
b. Interpret the coefficient for sex. Does discrimination exist?

Solution:

a.
A worker with one more year of experience, his or her wage is expected to increase by $2000 when EDUC and SEX are held constant.
A worker with one more year of education, his or her wage is expected to increase by $1000 when EXPER and SEX are held constant.

b.
The coefficient for sex means that for a female worker, her wage is expected to be higher than a male worker by $4,000, given that they have the same years of experience and education. But the t-ratio is small, the coefficient for sex is not significantly different from 0.
To discuss the problem of discrimination, we are looking for a significant coefficient on our dummy variable for sex. Since t-ratio = 1.18, the coefficient on sex is not significant, we cannot conclude that discrimination exists.

Example 7 Interaction Variables

A biologist is interested in the effect of temperature and humidity on cell growth. She collects the following information from six samples.

Sample	Temperature	Humidity	Cells
1	4°	10%	10,000
2	8	10	11,122
3	12	20	14,025
4	16	20	19,022
5	20	60	26,872
6	24	60	42,308

Estimate the relationship between cells, temperature, and humidity. Use an interaction variable to estimate the interaction effect of temperature and humidity on cell growth. Interpret your results.

Solution:

To capture the interaction effect of temperature and humidity on cell growth, we create a third independent variable by multiplying temperature and humidity. The coefficient on this variable will capture this interaction effect. The regression model is

$$\text{Cell}_i = \alpha + \beta_1 \text{Temp}_i + \beta_2 \text{Humid}_i + \beta_3 \left(\text{Temp}_i \times \text{Humid}_i\right) + \varepsilon_i.$$

Once we have created the interaction variable, we estimate the regression using the multiple regression procedure previously discussed. Because there are three independent variables we used the computer to estimate the regression. The output from MINITAB is:

Regression Analysis: C versus T, H, T*H

```
The regression equation is
C = 15303 + 411 T - 998 H + 53.2 T*H

Predictor     Coef  SE Coef      T      P
Constant     15303     3092   4.95  0.038
T            411.2    250.2   1.64  0.242
H           -997.9    230.0  -4.34  0.049
T*H         53.197    9.729   5.47  0.032

S = 1639.88   R-Sq = 99.3%   R-Sq(adj) = 98.2%

Analysis of Variance
Source          DF         SS         MS      F      P
Regression       3  753099547  251033182  93.35  0.011
Residual Error   2    5378393    2689197
Total            5  758477941
```

Hence, the estimated regression equation is:

$$\hat{C}_i = 15{,}303 + 411.2T_i - 997.9Hi + 53.197\left(T_i \times H_i\right)$$
$$(250.2) \quad (230.0) \quad (9.729)$$

$$R^2 = 0.993$$

$$s_e = 1639.88.$$

Standard errors are reported in parentheses.

Since F-ratio = 93.35 with p-value = 0.011, we reject H_0: $\beta_1 = \beta_2 = \beta_3 = 0$, and conclude this 99.3% of cells variation can be explained by this model.

Also, since t-ratio for H_0: $\beta_3 = 0$ is 5.47 and the corresponding p-value is 0.032, we reject H_0: $\beta_3 = 0$, and conclude existing interaction effect of temperature and humidity.

Supplementary Exercises

Multiple Choice

1. Multicollinearity occurs when

   a. Two or more independent variables are highly correlated
   b. The variance of the error terms is not constant

c. Current and lagged error terms are correlated
d. The independent variable is measured with error
e. We estimate an incorrect version of the true model

2. Multicollinearity gives us

a. Biased parameter estimates.
b. Best linear unbiased estimates (BLUE)
c. Inefficient parameter estimates
d. Problems in statistical inference
e. Two error terms that are correlated with each other

3. To correct the problem of multicollinearity, we can

a. Use a log transformation
b. Omit one or more of the independent variables
c. Use dummy variables
d. Use weighted least squares
e. Use a lagged-dependent variable

4. Heteroscedasticity occurs when

a. Two or more independent variables are highly correlated
b. The variance of the error terms is not constant
c. Current and lagged error terms are correlated
d. The independent variable is measured with error
e. We estimate an incorrect version of the true model

5. Heteroscedasticity gives us

a. Biased parameter estimates
b. Best linear unbiased estimates (BLUE)
c. Efficient parameter estimates
d. Problems in statistical inference
e. A high degree of correlation between the error terms and the dependent variable

6. To correct the problem of heteroscedasticity, we can

a. Omit one or more of the independent variables
b. Use a log transformation
c. Use dummy variables
d. Use weighted least squares
e. First correct for autocorrelation

7. Autocorrelation occurs when

a. Two or more independent variables are highly correlated
b. The variance of the error terms is not constant
c. Current and lagged error terms are correlated
d. The independent variable is measured with error
e. We estimate an incorrect version of the true model

8. Autocorrelation gives us

   a. Biased parameter estimates
   b. Best linear unbiased estimates (BLUE)
   c. Inefficient parameter estimates
   d. A short cut to statistical inference
   e. A high degree of correlation between the error terms and the independent variable

9. Specification error occurs when

   a. The dependent variable is measured with error
   b. We estimate an incorrect version of the true model
   c. Two or more independent variables are highly correlated
   d. The variance of the error terms is not constant
   e. Two or more dependent variables are highly correlated

10. Specification error gives us

    a. Biased parameter estimates
    b. Best linear unbiased estimates (BLUE)
    c. Inefficient parameter estimates
    d. Multicollinearity
    e. Autocorrelation

11. A dummy variable can be used when

    a. The independent variable is quantitative
    b. The independent variable is qualitative
    c. The independent variable is heteroscedastic
    d. There is autocorrelation
    e. Multicollinearity is known not to be a problem

12. If we are interested in using dummy variables to capture the effect of different months on stock returns we should use

    a. Twelve dummy variables, one for each month of the year
    b. One dummy variable
    c. Eleven dummy variables
    d. As many dummy variables as we like
    e. We cannot use dummy variables to capture this effect

13. The use of a dummy variable to measure differences in earnings between males and females assumes that

    a. The slope coefficient for males and females will be the same
    b. The intercept terms for males and females will be the same
    c. Both the slope coefficient and the intercept terms will be the same
    d. Both the slope coefficient and the intercept terms will be different
    e. Multicollinearity is a problem

14. The use of a dummy variable for gender, years of experience, and the interaction between that dummy variable and years of experience to model wages assumes that

    a. The slope coefficient for males and females will be the same
    b. The intercept terms for males and females will be the same
    c. Both the slope coefficient and the intercept terms will be the same
    d. Both the slope coefficients and the intercept terms will be different
    e. Multicollinearity is a problem

15. If we are estimating a *U*-shaped relationship such as the total cost curve in economics, we can best capture this relationship by using

    a. Dummy variables
    b. Simple regression
    c. A lagged-dependent variable
    d. A quadratic regression model
    e. This relationship cannot be modeled using regression analysis

16. If we use a log-log linear model to estimate the demand for ice cream, the slope coefficients would be

    a. Elasticities
    b. Identical to the results we would get from a nonlog model
    c. Always positive
    d. An example of multicollinearity
    e. Impossible to interpret

17. A regression model has two independent variables. The relation between VIF_1 and VIF_2 is

    a. $VIF_1 > VIF_2$
    b. $VIF_1 < VIF_2$
    c. $VIF_1 = VIF_2$
    d. uncertain.
    e. $VIF_1 = VIF_2 = 0.5$

18. For $n=20$, $k=4$, and $\alpha=0.05$, then the lower and upper bounds for DW statistics are

    a. $d_L = 0.69, d_U = 1.97$
    b. $d_L = 0.74, d_U = 1.97$
    c. $d_L = 0.86, d_U = 1.85$
    d. $d_L = 0.90, d_U = 1.83$
    e. $d_L = 0.82, d_U = 1.87$

19. For $n = 20,\ k = 4, \alpha = 0.05,$ and $DW = 1.95$, then conclusion for autocorrelation is

    a. Positive autocorrelation
    b. Negative autocorrelation
    c. No autocorrelation

d. Inconclusive
e. Insufficient information

20. For $n = 20, k = 4, \alpha = 0.05,$ and $DW = 3.45$, then conclusion for autocorrelation is

a. Positive autocorrelation
b. Negative autocorrelation
c. No autocorrelation
d. Inconclusive
e. Insufficient information

True/False (If False, Explain Why)

1. When perfect multicollinearity exists, it will be impossible to use the least-squares method.
2. Multicollinearity may be a problem when an independent variable is highly correlated with the dependent variable.
3. Checking the correlations between pairs of independent variables is sufficient to detect the problem of multicollinearity.
4. Heteroscedasticity leads to biased parameter estimates.
5. Heteroscedasticity can be detected using the DW statistic.
6. The problem of heteroscedasticity can sometimes be detected by examining the correlation between explanatory variables.
7. One way to detect heteroscedasticity is to look at a graph of the residuals against the independent variables or the expected values.
8. One way to mitigate the problem of heteroscedasticity is transforming the dependent variable properly.
9. Autocorrelation leads to biased parameter estimates.
10. The DW statistic can be used to detect autocorrelation when a lagged-dependent variable is included in regression.
11. The DW test will always tell us whether autocorrelation is a problem.
12. If the DW test shows there exists autocorrelation, then we need to modify the regression model to adjust the impact of autocorrelation.
13. Dummy variables can be used when qualitative variables such as sex or race are included in a regression.
14. If a qualitative variable has *m* categories, then *m* dummy variables must be included in the regression model, one for each category.
15. Omitting a relevant explanatory variable in a regression leads to specification error.
16. Including an irrelevant explanatory variable in a regression leads to specification error.
17. Using explanatory variables that are measured with error leads to specification error.

18. Including an irrelevant explanatory variable in a regression is worse than omitting a relevant variable.
19. An interaction variable can be useful when we believe that the effect of an independent variable on the dependent variable will depend on the value of another independent variable.
20. If the effect of humidity on the number of bacterial changes according to the values for temperature, then humidity and the number of bacterial interact.

Questions and Problems

1. At $\alpha = 0.05$, find d_U and d_L for a regression that has k explanatory variables and n observations:

   a. $k = 1, n = 25$

   b. $k = 2, n = 30$

   c. $k = 3, n = 20$

2. Suppose you run a regression using three explanatory variables and 40 observations. If you compute the DW statistic to be 1.25, is autocorrelation a problem? Use $\alpha = 0.05$.
3. Suppose you estimate a regression using one explanatory variable and 50 observations and compute the following:

$$\sum_{t=2}^{50}(e_t - e_{t-1})^2 = 6.27, \sum_{t=1}^{50} e_t^2 = 2.71$$

   a. Compute the DW statistic.
   b. Is autocorrelation a problem? Use $\alpha = 0.05$.

4. Below is the correlation matrix for the variables in a regression with two explanatory variables.

	y	x_1	x_2
y	1.00	0.98	0.75
x_1		1.00	0.85
x_2			1.00

   Does multicollinearity appear to be a problem?

5. You are interested in the relationship between y and three possible explanatory variables x_1, x_2, and x_3. Suppose $R_1^2 = 0.05$, $R_2^2 = 0.96$, $R_3^2 = 0.60$. Does multicollinearity appear to be a problem? If it does, what can you do to resolve it?

6. What possible problems might you encounter in the following regressions?

   c. A regression of IBM's stock returns against its EPS for the last 10 years.
   d. A regression of the EPS for 100 different companies against the size of each company in 1993.
   e. A regression of the consumption of 5,000 people against each person's income in 1993.

7. Suppose you estimate an earnings equation based on experience, education, and the sex of the individual. The equation you estimate is

$$\text{EARN}_i = 12,212 + 722\ \text{EXPER}_i + 927\ \text{EDUC}_i - 2211\ \text{SEX}_i,$$

   where $\text{SEX}_i = 1$ for females and 0 for males.
   Interpret the coefficient on the dummy variable for sex, SEX_i.

8. For the data in Example 7, suppose we are only interested in the relationship between temperature and cell. Do we need to consider other terms transformed from temperature to have a better fit?
9. The relationship between y and x_1, x_2 has been established by a regression analysis. The residual terms are squared, and the MINITAB output for fitting a regression using e^2 as the dependent variable and x_1, x_2 as the independent variable is given below.

Regression Analysis: e^2 versus x1, x2

```
The regression equation is
e^2 = 102 + 1.17 x1 - 0.27 x2

Predictor     Coef  SE Coef      T      P
Constant     101.8    311.6   0.33  0.747
x1           1.169    2.007   0.58  0.566
x2          -0.268    2.355  -0.11  0.911
 S = 198.501   R-Sq = 1.5%   R-Sq(adj) = 0.0%
```

Does heteroscedasticity appear to be a problem in this regression?

Answers to Supplementary Exercises

Multiple Choice

1. a	6. d	11. b	16. a
2. c	7. c	12. c	17. c
3. b	8. c	13. a	18. d

4. b	9. b	14. d	19. c
5. d	10. a	15. d	20. b

True/False

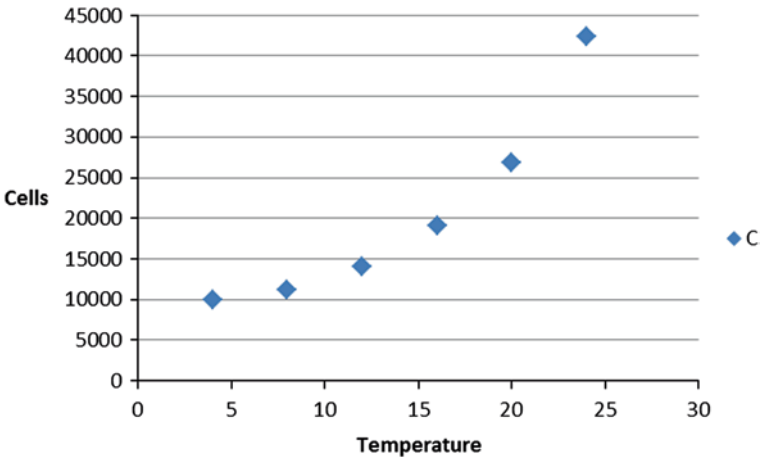
1. True.
2. False. Multicollinearity occurs when independent variables are highly correlated.
3. False. Multicollienarity is caused by part or all of the independent variables are highly related.
4. False. Heteroscedasticity leads to inefficient estimates.
5. False. Autocorrelation is detected using the DW test.
6. False. Heteroscedasticity can be detected by examining a plot of the error terms from the regression.
7. True.
8. True.
9. False. Autocorrelation leads to inefficient estimates.
10. False. Durbin's H should be used when a lagged-dependent variable is used in a regression.
11. False. The DW test has a range where the test is indeterminate.
12. TRUE.
13. True.
14. False. Only need $m-1$ dummy variables.
15. True.
16. False. A t-test may detect the irrelevance of a variable.
17. True.
18. False. It is worse to omit a relevant explanatory variable than to include an irrelevant one.
19. True.
20. False. Humidity and temperature interact.

Questions and Problems

1. a. $d_L = 1.29,\ d_U = 1.45$
   b. $d_L = 1.28,\ d_U = 1.57$
   c. $d_L = 1.00,\ d_U = 1.68$
2. $d_L = 1.34, d_U = 1.66$
   $0 < d < d_L$, so we reject the null hypothesis of no autocorrelation and accept the alternative hypothesis of positive autocorrelation.

3. a. $d = 6.27 / 2.71 = 2.31$
   b. $d_L = 1.50$, $d_U = 1.59$
   Since $1.59 < d < 4 - d_U$, do not reject the null hypothesis of no autocorrelation.
4. The correlation between x_1 and x_2 is 0.85, so multicollinearity may be a problem.
5. $VIF_1 = \frac{1}{1 - R_1^2} = 1.05$, $VIF_2 = \frac{1}{1 - R_2^2} = 25$, $VIF_3 = \frac{1}{1 - R_3^2} = 2.5$

   Since $VIF_2 > 10$, so multicollinearity appears to be a problem.
   Because x_2 is highly correlated to x_1 and x_3, we will use x_1 and x_3 as independent variables.
6. a. In a time series regression like this, autocorrelation may be a problem.
   b. In a cross-sectional regression like this, heteroscedasticity may be a problem because we would expect larger companies to have higher variance in their EPS.
   c. Heteroscedasticity is likely to be a problem because individuals with higher incomes are likely to have a greater variance in their consumption patterns.
7. The coefficient on the dummy variable measures the difference in earnings between males and females, assuming a similar relationship between the experience and education of the individual and his or her earnings. Because the coefficient is −2,211, females are expected to earn $2,211 less than males with similar levels of education and experience.
8. From the scatter diagram of cells versus temperature, there is a nonlinear pattern. Adding a quadratic term for temperature into the simple regression model seems reasonable.



The MINITAB output is:

Regression Analysis: C versus T

```
The regression equation is
C = - 821 + 1527 T

Predictor    Coef  SE Coef      T      P    VIF
Constant     -821     4782  -0.17  0.872
T          1527.1    307.0   4.97  0.008  1.000

S = 5136.90   R-Sq = 86.1%   R-Sq(adj) = 82.6%

Analysis of Variance
Source          DF         SS         MS      F      P
Regression       1  652926877  652926877  24.74  0.008
Residual Error   4  105551064   26387766
Total            5  758477941
```

Regression Analysis: C versus T, T^2

```
The regression equation is
C = 14406 - 1328 T + 102 T^2

Predictor     Coef  SE Coef      T      P
Constant     14406     2570   5.61  0.011
T          -1327.9    420.3  -3.16  0.051
T^2         101.96    14.69   6.94  0.006

S = 1436.49   R-Sq = 99.2%   R-Sq(adj) = 98.6%

Analysis of Variance
Source          DF         SS         MS       F      P
Regression       2  752287402  376143701  182.28  0.001
Residual Error   3    6190538    2063513
Total            5  75847794
```

Since the t-ratio for (temperature)^2 is 6.94 with p-value$=0.006$, the null hypothesis for $\beta_2 = 0$ is reject fit.

Chapter 17
Nonparametric Statistics

Chapter Intuition

So far, you have learned how to conduct statistical analysis assuming that the data is from a specific distribution in order to make inferences for the parameters of the distribution. This is the so-called ***parametric statistics***. However, in many instances, we will not know the distribution from which the data comes. In this case, we need to take a different approach in order to conduct our analysis. This approach is known as ***nonparametric statistics*** because there is no need to assume the data coming from any particular distribution. For example, we learned about ANOVA, which tests whether three or more means equal each other. An important assumption for conducting the test is that the data are drawn from normal distributions with equal variance. In this chapter, we will learn about a test known as the Kruskal–Wallis test, which is used for conducting ANOVA without assuming that the data are drawn from normal distributions. Besides that, five other nonparametric methods applicable for various situations will be introduced. Finally, it is worth mentioning that many of the nonparametric methods discussed in this chapter use the "sign" or "ranking" of data to make comparisons rather than comparing the data directly.

Chapter Review

1. ***Nonparametric tests*** are distribution-free tests. The term "distribution-free" means that the distribution of the test statistic does not depend on the assumption that the data are drawn from a certain distribution.
2. The ***sign test*** can compare the means or central tendencies of two populations using matched pairs. It can also be used to examine the central tendency (median) of a population. The sign test uses the sign of the difference between pairs of numbers coming from different groups of data, or the sign of difference between the sample value and the hypothesized medium. In conducting a sign test, we utilize the test of a proportion. When the sample size is large, the test statistic can be approximated by the standard normal distribution.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_17

3. The ***Wilcoxon matched-pairs signed-rank test*** compares the differences of two related populations. The data used in this method have to be matched pairs and quantitative. This test uses not only the sign of the difference between a pair of numbers but also the quantitative measurements of the difference. The test statistic can be approximated by the standard normal distribution when the sample is large.
4. The ***Mann–Whitney U test*** is used to test whether the two population distributions are identical using two independent samples. The *U* statistic is based on the rank sum of the sample groups. The Mann–Whitney *U* test can be approximated by the standard normal distribution when the sizes of both samples are at least ten.
5. The ***Kruskal–Wallis test*** is used to examine whether three or more independent samples originate from the same distribution. The Kruskal–Wallis test computes the rank sum of each data group. The statistic approximates a chi-square distribution when the sample is large. The degrees of freedom is the number of data groups minus one.
6. ***Spearman's rank correlation*** is used to measure whether the rankings of two variables are correlated. This statistic will generate a number between -1 and 1. A negative Spearman's rank correlation indicates that the rankings of the two variables move in different directions. That is, a higher ranking in one variable implies a lower ranking in the other variable. A *t* test can be used to test the population rank correlation. The degrees of freedom are the sample size minus two.
7. A ***runs test*** is used to study the randomness of data. In conducting the test, we need to count the number of runs, which is defined as consecutive numbers with the same sign. The runs test statistic approximates the standard normal distribution when the sample size is large.

Useful Formulas

Sign test:

Test statistic under normal approximation:

$$Z = \frac{\bar{p} - 0.5}{\sqrt{0.5(0.5)/n}},$$

n is the effective sample size, $\bar{p}$ is the sample proportion of "+,"

Wilcoxon signed-rank test:

$W^+ = \sum_{i=1}^{n} R_i^+$ is the sum of the plus ranks

Test statistic under normal approximation:

$$Z = \frac{W^+ - \mu_W}{\sigma_w}$$

$\mu_w = n(n+1)/4, \sigma_w = \sqrt{n(n+1)(2n+1)/24}$, n = the effective sample size.

Mann–Whitney *U* test:

U statistic:

$$U = n_1 n_2 + \frac{n_1(n_2+1)}{2} - R_1 \quad \text{or} \quad U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Test statistic under normal approximation:

$$Z = \frac{U - \mu_U}{\sigma_U}$$

$$\mu_U = \frac{n_1 n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

Kruskal–Wallis test:

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

$$n = n_1 + \ldots + n_k$$

Spearman's rank correlation:

Spearman's rank correlation:

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Test statistic:

$$t = \frac{r_s}{\sqrt{(1 - r_s^2)/(n-2)}}$$

Runs test:

$$Z = \frac{R - \mu_R}{\sigma_R}$$

$$\mu_R = \frac{2 n_1 n_2}{n} + 1$$

$$\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}}$$

$$n = n_1 + n_2$$

Example Problems

Example 1 Testing the Equality of Proportions

A marketing company is conducting a survey of cola popularity at the mall. Forty people have been asked to taste two different brands of cola, Cola A and Cola B. These people are then asked to score each cola on a scale of 1–4, with 4 being the best flavor. The table below presents how each person scored the two colas.

A	B	A	B	A	B	A	B
1	2	3	4	1	3	4	1
2	1	3	2	1	2	4	2
4	3	1	2	3	2	4	2
1	3	4	1	2	3	2	4
2	4	4	3	4	2	4	1
2	4	2	1	4	2	3	2
2	4	3	4	4	2	2	1
4	3	4	2	1	2	2	3
4	2	3	1	4	2	3	1
4	3	2	1	2	4	4	2

Test whether the two brands of cola are equally popular. Use a 10% level of significance.

Solution:

Let p be the proportion of people who like Cola A better.

The hypotheses of equally popular are:

H_0: equally popular $(p = 1/2)$

H_1: not equally popular $(p \neq 1/2)$

Among the 40 people, 24 like A and 16 like B.

Hence, $\overline{p} = 25/40 = 0.625$

Test statistic is:

$$Z = \frac{\overline{p} - p^*}{\sqrt{\frac{p^*(1-p^*)}{n}}} = \frac{25/40 - 1/2}{\sqrt{\frac{1/2(1-1/2)}{40}}} = 1.58$$

Since $1.58 < z_{0.10/2} = 1.645$, we do not have enough evidence to reject the null hypothesis of A and B being equally popular.

Example 2 Wilcoxon Matched-Pair Signed-Rank Test

The manager of a fast food chain is interested in the effect of coupons on sales. He collected sales data from ten fast food restaurants under his supervision. The sales numbers are for June, when the coupons are in effect, and for July when there are no coupons.

Restaurant	June	July
1	145	132
2	156	160
3	143	131
4	155	132
5	162	160
6	159	159
7	160	165
8	149	152
9	162	153
10	144	140

Do the data support the hypothesis that coupons make a difference? Use a 5 % level of significance.

Solution:

H_0: Sales distributions are identical.

H_1: Sales distributions are not identical.

(1) Restaurant	(2) June	(3) July	d(3) − (2)	| d |	Rank of | d |	(+)	(−)
1	145	132	−13	13	7	–	7
2	156	160	4	4	3.5	3.5	
3	143	131	−12	12	8	–	8

(1) Restaurant	(2) June	(3) July	d(3)−(2)	| d |	Rank of | d |	(+)	(−)
4	155	132	−23	23	9	–	9
5	162	160	−2	2	1	–	1
6	159	159	0	–	–	–	–
7	160	165	5	5	5	5	–
8	149	152	3	3	2	2	–
9	162	153	−9	9	6	–	6
10	144	140	−4	4	3.5	–	3.5
Total	–	–	–	–	–	10.5	34.5

$$w^+ = 10.5, w^- = 34.5.\ \min\,(w^+, w^-) = 10.5.$$

There are nine differences in our data, so n=9.

From Table A.11, the two-tailed value at n=9 and $\alpha = 0.05$ is 6.

Because the value of w+is higher than 6, we reject the null hypothesis of no difference.

Example 3 Testing the Inequality of Two Means

A statistics professor is interested in seeing if there is any difference between the average scores of those students who own the workbook and those who do not. The rankings are summarized below (a higher ranking means a higher grade):

Own			Do not own		
2	4	6	1	3	5
7	8	15	9	10	11
18	17	16	14	13	12
19	20	26	21	22	23
30	28	27	29	25	24

Do a test to determine if students who own the workbook have higher average scores than students who do not own the workbook. Use a 5% level of significance.

Solution: $H_0 : \mu_1 - \mu_2 \le 0$

$$H_1 : \mu_1 - \mu_2 > 0$$

where

μ_1 = the average for students with the workbook.
μ_2 = the average for students without the workbook.

Add the ranks of the data for students who do not own the workbook, and obtain $R_2 = 222$.

$$U = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 15(15) + \frac{15(15+1)}{2} - 222 = 123$$

$$\mu_U = \frac{n_1 n_2}{2} = 112.5$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} 24.11\sigma$$

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{123 - 112.5}{24.11} = 24.11 < 1.645.$$

We do not have enough evidence to reject the null hypothesis that students owning the workbook have a higher average.

Example 4 Testing the Equality of Two Distributions

Two express mail services are competing for a contract with your company. You tested the two services by sending 15 pieces of mail through each service. The delivery times are as follows:

Service A			Service B		
14.3	15.9	17.5	15.6	16.5	17.8
14.4	16.2	18.1	15.8	17	17.9
15.5	16.7	18.2	16	17.1	18.3
15.9	16.8	18.5	16.1	17.2	18.4
15.9	16.9	18.6	16.3	17.3	18.7

Do a 5% test to determine if there are different delivery times between Service A and Service B.

Solution: H_0: Two distributions are identical.

Since the delivery times are from two independent samples, do Mann–Whitney's *U* test.

Pool the data together and rank the data. Then, add the ranks of the data for Service A and obtain $R_1 = 210$.

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 15(15) + \frac{15(15+1)}{2} - 210 = 132$$

$$\mu_U = \frac{n_1 n_2}{2} = 112.5$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = 24.11$$

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{132 - 112.5}{24.11} = 0.81$$

Since $0.81 < z_{0.05/2} = 1.96$, so there is insufficient evidence to reject H_0.

Example 5 Kruskal–Wallis Test

Thirty economists from different backgrounds are asked to estimate the unemployment rate in the next quarter. The unemployment estimates are ranked in the following table.

Academic economist	Private economist	Government economist
1	3	4
10	2	5
9	8	15
6	7	16
11	13	17
12	14	18
20	19	21
22	24	23
28	26	25
29	30	27

Can you argue that the economists working in different professions do not have the same average estimate of the unemployment rate? Use a 5% level of significance.

Solution:

$$R_1 = 148,\ R_2 = 146,\ R_3 = 171$$

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{30(30+1)}\left[\frac{148^2}{10} + \frac{146^2}{10} + \frac{171^2}{10}\right] - 3(30+1) = 0.498 < \chi^2_{2,0.05} = 5.99$$

We do not have enough evidence to reject the null hypothesis.

Example 6 Spearman's Rank-Correlation Test

A consumer organization wants to know if consumers receive their money's worth when purchasing a stereo. They sampled ten stereos and ranked the quality and price of each set. A higher rank indicates a better product and a higher price.

Stereo	(1) Price rank	(2) Quality rank
1	1	1
2	8	3
3	2	2
4	3	4
5	7	5
6	4	6
7	6	7
8	5	8
9	9	9
10	10	10

Use a 5% level of significance to determine if consumers get what they pay for.

Solution: If consumers get what they pay for, the quality and price ranks should have a positive correlation.

$$H_0 : \rho \le 0$$

$$H_1 : \rho > 0$$

Stereo	(1) Price rank	(2) Quality rank	$d_i = (1) - (2)$	d_i^2
1	1	1	0	0
2	8	3	5	25
3	2	2	0	0

Stereo	(1) Price rank	(2) Quality rank	$d_i=(1)-(2)$	d_i^2
4	3	4	−1	1
5	7	5	2	4
6	4	6	−2	4
7	6	7	−1	1
8	5	8	−3	9
9	9	9	0	0
10	10	10	0	0
Total	–	–	–	44

$$\Sigma d_i^2 = 44$$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(44)}{10(10^2-1)} = 0.733$$

$$t = \frac{r_s}{\sqrt{(1-r_s^2)/(n-2)}} = \frac{0.733}{\sqrt{(1-0.733^2)/(10-2)}} = 3.05$$

Since $3.05 > t_{8,0.05} = 1.86$, reject $H_0 : \rho \le 0$ and conclude that there is sufficient evidence for consumers getting what they pay for.

Example 7 Runs Test

In a conference, 40 economists were asked if they think we are out of the recession. The answering pattern, in order, is

y y n n y n y y y n n y y n y y n n n n y y n y y n y n n n y y n n y n y n y n

where y indicates a yes vote and n indicates a no vote.

The person conducting this survey suspects that an economist's answer tends to be affected by the answer just preceding it. Can you verify this? Do a 5% test.

Solution:

If an economist's answer is affected by the previous answer, then the answering pattern should exhibit momentum with fewer changes in the sequence.

H_0: no momentum
H_1: momentum (fewer changes)
Number of runs: R = 24

Also, $n_1(\text{"y"}) = 20, n_2(\text{"n"}) = 20$.

$$\mu_R = \frac{2n_1n_2}{n} + 1 = \frac{2(20)(20)}{40} + 1 = 21$$

$$\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}} = \sqrt{\frac{2(20)(20)(2(20)(20) - 40)}{(40)^2(40-1)}} = 3.12$$

$$Z = \frac{R - \mu_R}{\sigma_R} = \frac{24 - 21}{3.12} = 0.96 > -z_\alpha = -z_{0.05} = -1.645$$

Do not reject H_0, and conclude there is not sufficient evidence that an economist's answer tends to be affected by the answer just preceding it.

Supplementary Exercises

Multiple Choice

1. Which of the following is another name for nonparametric statistics?

   a. Nominal statistics
   b. Distribution-free statistics
   c. Ratio statistics
   d. Scale statistics
   e. *t* statistics

2. In a nonparametric test, it is not necessary to assume that

   a. The data are normally distributed.
   b. The test statistic is not approximately normally distributed.
   c. The data are at least ordinal.
   d. The test statistic is approximately normally distributed.
   e. All of the above.

3. Which of the following statistics is not constructed on the ranks of the data?

   a. The sign test
   b. The Mann–Whitney *U* test
   c. The Wilcoxon signed-rank test
   d. Spearman's rank correlation test
   e. The Kruskal–Wallis test

4. If we are interested in testing whether the median SAT score for Hannahtown High School is 1000, we would use

   a. A runs test
   b. A Mann–Whitney *U* test

c. A Wilcoxon signed-rank test
d. Spearman's rank correlation test
e. Sign test.

5. Which of the following tests should be used to compare the equality of two means if the samples are related?

a. Wilcoxon signed-rank test
b. Mann–Whitney *U* test
c. Kruskal–Wallis test
d. Spearman's rank correlation test
e. Runs test

6. A data containing 21 matched-pairs has been collected, and 20 of them show pairwise difference. The sum of plus (+) ranks is 95. What is the z value for signed-rank test?

a. −0.713
b. −0.187
c. −0.373
d. 0.373
e. 0.713

7. To compare two distributions, which of the following tests is appropriate to use if the samples are independent?

a. Wilcoxon signed-rank test
b. Mann–Whitney *U* test
c. Kruskal–Wallis test
d. Spearman's rank correlation test
e. Runs test

8. In conducting a statistical test, the research assistant messed up the order of the data and the test result was seriously biased. What kind of test was being done?

a. Sign test
b. Mann–Whitney *U* test
c. Spearman's rank correlation test
d. Runs test
e. Kruskal–Wallis test

9. Which of the following statistics is constructed on the ranks of the data?

a. Sign test
b. Runs test
c. F test
d. The Kruskal–Wallis test
e. Pearson's correlation coefficient

10. Which of the following tests is used to compare three or more means from independent samples?

    a. The sign test
    b. Mann–Whitney *U* test
    c. The Wilcoxon signed-rank test
    d. Spearman's rank correlation test
    e. Kruskal–Wallis test

11. Which of the following tests can compare two means of two groups with different sample sizes?

    a. The sign test
    b. The Mann–Whitney *U* test
    c. The Wilcoxon signed-rank test
    d. Spearman's rank correlation test
    e. Kruskal–Wallis test

12. Which of the followings is the nonparametric counterpart of the F test to compare the means of three or more independent populations?

    a. The sign test
    b. The Mann–Whitney *U* test
    c. The Wilcoxon signed-rank test
    d. Spearman's rank correlation tes
    e. Kruskal–Wallis test

13. Which of the following is the nonparametric counterpart of the *t* test to compare the means of two independent populations?

    a. Chi-square goodness of fit test
    b. Chi-square test of independence
    c. Mann–Whitney *U* test
    d. Wilcoxon signed-rank test
    e. Friedman test

14. The approximate distribution for the Kruskal–Wallis *t* statistic is a/an

    a. *t* distribution
    b. Normal distribution
    c. Standard normal distribution
    d. Chi-square distribution
    e. F distribution

15. If you are interested in seeing if a basketball player is a streaky shooter (tends to make many baskets in a row) you would use

    a. A runs test
    b. A Mann–Whitney *U* test
    c. A Wilcoxon signed-rank test

d. Spearman's rank correlation test
e. Kruskal–Wallis test

16. The expected number of runs in n flips of a coin is

a. n
b. $n-1$
c. $2n_1n_2/n$
d. $2n_1n_2/(n+1)$
e. $(2n_1n_2/n)+1$

17. Suppose you ranked four students on the basis of their SAT scores (from 1 to 4, 1 being the highest SAT score). If these students' SAT scores have a rank correlation of -1 with the rank of their high school grades, then the rankings for high school grades would imply that

a. The student with the highest SAT score also had the highest grade.
b. The student with the highest SAT score had the lowest grade.
c. The student with the lowest SAT score also had the lowest grade.
d. The student with the highest SAT score had the middle grade.
e. Cannot be determined by the information given.

18. Suppose you ranked ten students on the basis of their SAT scores (from 1 to 4, 1 being the highest SAT score). If these students' SAT scores have a rank correlation of 0.8 with the rank of their high school grades, then

a. The student with the lowest SAT score had the highest grade.
b. The student with the lowest SAT score also had the lowest grade.
c. The student with the higher SAT score tended to have the lower grade.
d. The student with the higher SAT score tended to have the higher grade.
e. Cannot be determined by the information given.

19. How many runs are there in the following data: $+-+---+-$?

a. 8
b. 7
c. 6
d. 5
e. 4

20. Which of the following can help us to decide whether the sequence of "yes" or "no" responses of walk-in customers are related?

a. A runs test
b. A Mann–Whitney *U* test
c. A Wilcoxon signed-rank test
d. Spearman's rank correlation test
e. A sign test

True/False (If False, Explain Why)

1. One advantage of nonparametric statistics is that some nonparametric tests can be used to analyze nominal or ordinal data values.
2. Nonparametric tests get this name because they do not use test statistics such as the *t* test.
3. Many nonparametric tests use a ranking method to test hypotheses.
4. The Kruskal–Wallis test is used to test the equality of three or more means when the populations are not normal.
5. A rank-sum test can be used to test the hypothesis that changes in stock prices are random.
6. Nonparametric tests assume that the data are normally distributed.
7. In the runs test, the data follow a normal distribution.
8. Spearman's rank correlation measures the correlation between the rankings of two variables.
9. The Wilcoxon signed-rank test is used to compare the difference between two groups of data when the data are paired.
10. The sign test is used to compare the difference between two groups of data when the data are paired.
11. The Spearman's rank correlation test has a maximum value of 1 and a minimum value of 0.
12. If we want to test whether data carry some kind of momentum, we may use the sign test.
13. If we are testing the null hypothesis that data does not carry some kind of momentum, we use the runs test and put the rejection region in the left tail.
14. In conducting the Kruskal–Wallis test, we should make sure that the same amount of data is present in each group.
15. In conducting the Mann–Whitney *U* test, it is necessary to use the rank sum of each group.
16. In conducting the Kruskal–Wallis test, it is necessary to use the rank sum of each group.
17. The Mann–Whitney *U* test can be used to compare the medians of two independent samples with ordinal values.
18. The Wilcoxon's signed rank test is the nonparametric alternative of the *t* test to compare the means of two independent populations.
19. The Mann–Whitney *U* test is the nonparametric alternative to the one-way analysis of variance.
20. When only ordinal values of paired data are available, the Pearson correlation coefficient is recommended to analyze the association between two variables.

Questions and Problems

1. Twenty-five graduates who majored in computer science were asked to give their starting salaries. Using the sign test, can you reject the hypothesis that the median salary is more than $ 31,000? Do a 5 % test.

Salaries (in thousands of dollars)				
35.3	34.3	26.1	28.2	28.9
32.8	37.2	32.8	31.6	32.8
31.2	29.7	26.3	31.3	32.7
32.8	33.8	32.1	32.3	35.4
29.7	31.6	38.6	31.8	31.2

2. A company has designed two sales strategies for a new product, and is interested in their effectiveness. Ten salesmen were selected to implement both strategies. The sales records are shown below.

Salesman	A	B
1	34	36
2	36	40
3	27	25
4	35	32
5	38	30
6	35	32
7	42	40
8	29	28
9	35	32
10	38	33

At $\alpha = 0.1$, can you reject the hypothesis that two strategies are the same? Do a 5 % test.

3. Fifty graduates, 25 males and 25 females, who majored in computer science were asked to give their starting salaries. The results are given below in thousands of dollars.

Males					Females				
25.1	25.2	25.3	25.8	25.9	24.3	24.5	25.7	26.0	26.2
26.3	26.4	26.8	26.9	27.0	26.5	26.7	27.3	27.6	27.9
27.1	27.2	27.4	27.5	27.8	28.1	28.3	28.3	28.4	28.5
27.9	28.1	28.3	28.7	28.8	28.6	28.6	28.7	28.9	28.9
29.3	29.4	29.5	29.7	30.5	29.1	29.2	29.2	29.2	29.2

Use a 5% level of significance to determine if males receive a higher average salary than females.

4. Thirty-five workers are divided into three groups. The first group works in a quiet environment. The second group works while playing classical music. The third group works while playing easy listening music. The productivities of the three groups are given below.

Group 1	1	7	9	12	13	15	16	22	27	21	30	33	35
Group 2	2	4	5	6	11	17	18	19	20	25	28	29	–
Group 3	3	8	10	14	23	24	26	31	32	34	–	–	–

Do a 5% test to determine if the three groups have the same average productivity.

5. Use the data from Problem 4 to compare the first and second groups. Can you argue that the average productivities are different? Do a 5% test.
6. It is widely believed that a student's mathematical ability can help his or her science grade. A mathematics teacher collected the grades of ten students and ranked the grades as follows:

Rank of math grade	1	2	3	4	5	6	7	8	9	10
Rank of science grade	2	4	1	5	9	3	6	10	8	7

Do the above data support the hypothesis that the two grades are positively correlated? Use a 5% level of significance.

7. A baseball player's last 40 at bats are recorded as follows. (H is a hit and N is no hit.)

H N H H H N H H H H N N N N N N N N H H H N N N N N N N N N N H H N N N N H N N N

Is this batter a streaky hitter? Use a 5% level of significance.

Answers to Supplementary Exercises

Multiple Choice

1. b
2. a
3. a
4. e
5. a

6. c
7. b
8. d
9. d
10. e
11. b
12. e
13. c
14. d
15. a
16. e
17. b
18. d
19. c
20. a

True/False

1. True.
2. False. They get their name because the distribution of the data is not known.
3. True.
4. True.
5. False. Use a runs test.
6. False. Data do not necessarily have to be normally distributed.
7. False. Data are dichotomous, so they cannot be normally distributed.
8. True.
9. True.
10. True.
11. False. Minimum value is− 1.
12. False. Use a runs test.
13. True.
14. False. Amount of data can differ.
15. False. If we know the rank sum of one group we can conduct the test.
16. True.
17. True.
18. False. Signed rank test is the nonparametric alternative of the *t* test to compare the means of two related populations
19. False. Kruskal–Wallis test.
20. False. Spearman's rank correlation.

Questions and Problems

1. $H_0 : p \le 1/2$

   $H_1 : p > 1/2$

p = the population proportion of numbers no more than 31.
$\bar{p}$ = the sample proportion of numbers no more than $31 = 19/25$

$$Z = \frac{\bar{p} - 0.5}{\sqrt{0.5(0.5)/n}} = \frac{(19/25) - 0.5}{\sqrt{0.5(0.5)/25}} = 7.6 > z_{0.05/2} = 1.96$$

Reject the null hypothesis.

2.

Salesman	A	B	d = B − A	|d|	Rank of |d|	Signed-rank	Signed +	Signed −
1	34	36	2	2	3	3	3	–
2	36	40	4	4	8	8	8	–
3	27	25	−2	2	3	−3	–	3
4	35	32	−3	3	6	−6	–	6
5	38	30	−8	8	10	−10	–	10
6	35	32	−3	3	6	−6	–	6
7	42	40	−2	2	3	−3	–	3
8	29	28	−1	1	1	−1	–	1
9	35	32	−3	3	6	−6	–	6
10	38	33	−5	5	9	−9	–	9
–	–	–	–	–	–	–	11	44

$$w^+ = 11, w^- = 44. \min(w^+, w^-) = 11, \text{and} n = 10.$$

Method 1. Find the critical value from table A.11.
From Table A.11, the two-tailed value at n = 10 and $\alpha/2 = 0.05$ is 11.
Since the value of w + = 11, we reject the null hypothesis of no difference.
Method 2. Use normal approximation.
Test statistic under normal approximation:

$$\mu_w = n(n+1)/4 = 10(11)/4 = 27.5$$

$$\sigma_w = \sqrt{n(n+1)(2n+1)/24} = \sqrt{10(11)(21)/24} = 9.81$$

$$Z = \frac{W^+ - \mu_W}{\sigma_w} = \frac{11 - 27.5}{9.81} = -1.68 < -z_{0.10/2} = -1.645$$

Reject H_o: No difference.

3. Pool the data together and rank them. The rankings of the males and females are:

Males					Females				
3	4	5	7	8	1	2	6	9	10
11	12	15	16	17	13	14	20	23	25.5
10	19	21	22	24	27.5	30	30	32	33
25.5	27.5	30	36.5	38	34.5	34.5	36.5	39.5	39.5
46	47	48	49	50	41	43.5	43.5	43.5	43.5

Rank sum of males, $R_1 = 599.5$.

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 25(25) + \frac{25(25+1)}{2} - 599.5 = 350.5$$

$$\mu_U = \frac{n_1 n_2}{2} = 3125 \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = 5154$$

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{350.5 - 312.5}{51.54} = 0.74$$

Since $0.74 < z_{0.05} = 1.645$, we cannot reject the null hypothesis.

4. H_0: equally productive.

H_1: not equally productive.

$R_1 = 241$, $R_2 = 184$, $R_3 = 205$

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{3} \frac{R_i^2}{n_i} - 3(n+1)$$

$$= \frac{12}{35(35+1)} \left[\frac{241^2}{13} + \frac{184^2}{12} + \frac{205^2}{10} \right] - 3(35+1)$$

$$= 109.44 - 108 = 1.44 < \chi^2_{2,0.05} = 5.99$$

We do not have enough evidence to reject the null hypothesis.

5. To answer this question, we need to rerank the data in groups 1 and 2. Then, add the ranks of the data for group 1 and obtain $R_1 = 183$.

H_0: equally productive.

H_1: not equally productive.

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 13(12) + \frac{13(13+1)}{2} - 183 = 64$$

$$= \mu U = \frac{n_1 n_2}{2} = 78, \quad \sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = 18.385$$

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{64 - 78}{18.385} = -0.76$$

Since $-0.761 > -z_{0.05/2} = -1.96$, we cannot reject H_0.

6. $H_0 : \rho \le 0$, not positively correlated.

H_1: $\rho > 0$, positively correlated.

x	y	$d_i = x - y$	d_i^2
1	2	−1	1
2	4	−2	4
3	1	2	4
4	5	−1	1
5	9	−4	16
6	3	3	9
7	6	1	1
8	10	−2	4
9	8	1	1
10	7	3	9
			50

$$\Sigma d_i^2 = 50$$

$$r_S = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} = 1 - \frac{6(50)}{10(10^2-1)} = 0.697$$

$$t = \frac{r_s}{\sqrt{(1 - r_s^2)/(n-2)}} = \frac{0.697}{\sqrt{(1 - 0.697^2)/(10-2)}} = 2.75$$

Since $2.75 > t_{8,0.05} = 1.86$, reject $H_0 : \rho \le 0$ and conclude that there is sufficient evidence for positive association.

7. H_0: not a streaky hitter.

H_1: a streaky hitter.

We can test the null hypothesis by examining the number of runs. If the hitter is streaky, there will be fewer runs than we would expect because both the hits and outs will come in bunches.

$$R = 12, n_1\left(\text{for H}\right) = 14,\ n_2 = 26$$

$$\mu_R = \frac{2n_1n_2}{n} + 1 = \frac{2(14)(26)}{40} + 1 = 19.2$$

$$\sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}} = \sqrt{\frac{2(14)(26)(2(14)(26) - 40)}{(40)^2(40-1)}} = 2.83$$

$$Z = \frac{R - \mu_R}{\sigma_R} = \frac{12 - 19.2}{2.83} = -2.54 < -z_\alpha = -z_{0.05} = -1.645$$

We reject the null hypothesis and there is sufficient evidence that the hitter is a streak hitter.

Chapter 18
Time-Series: Analysis, Model, and Forecasting

Chapter Intuition

Previously, you learned how different variables could be related to one another by using regression analysis. In most of the examples that were discussed, you were looking at ***cross-sectional data***. Cross-sectional data are data which occur at the same time, but which are examined across many different individuals or companies. For example, if we were interested in determining factors that influence department store sales in 2003, we could collect information such as sales, prices, and strength of the economy in 2003 for many different stores. As we are determining sales in one time period (2003), across many stores, we are using cross-sectional data.

A different type of analysis that could be conducted is to look at sales over many different months or years for the same store. As we are analyzing the data over different time periods, we are looking at ***time-series data***. While some of the same techniques used in cross-sectional analysis are also applicable to time-series analysis, the use of time-series data may necessitate the use of special techniques not required in cross-sectional analysis. For example, if we are using monthly sales data to find a relationship between the sales of Sears and various economic factors, we should also deal with possible seasonal effects, which occur in retail sales (e.g., higher sales around Christmas). Besides the seasonal component, there may exhibit some gradual movements constituting the trend component over time. This chapter discusses how to deal with some of these problems and presents special techniques that can be used in time-series analysis.

Chapter Review

1. ***Time-series data*** are numbers that are recorded over time. ***Cross-sectional data*** are numbers that are recorded across different companies, industries, or individuals in the same time period.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_18

2. When analyzing a time series, we sometimes find it easier to conduct the analysis if we decompose the time series into several observable components.
   a. The ***trend component*** is the part of the time series that reflects a tendency either to grow or to decrease fairly steadily over time.
   b. The ***seasonal component*** is the intrayear pattern, which constantly repeats itself from year to year. For example, the sales revenues of a toy store would be expected to experience higher sales during the Christmas shopping season.
   c. The ***cyclical component*** consists of long-term oscillatory patterns that are unrelated to seasonal behavior. The US Department of Commerce has specified a number of time series that are used to identify the peaks and valleys in the business cycle. ***Leading indicators*** get their names because they tend to have their turning points prior to turns in economic activity. ***Coincidence indicators*** tend to have their turning points at the same time as turns in economic activity. ***Lagging indicators*** tend to have their turning points following the turning points in economic activity.
   d. The ***irregular component*** of a time series is the random pattern that may show up in either the long run or the short run.
3. One of the simplest approaches to analyze a time series is to use a ***simple moving average***. In an *n*-period simple moving average, we just find the mean of the series for the most recent n periods to forecast the current period. For example, we might look at the mean value of the Dow Jones Industrial Average over the last 30 days.
4. ***Exponential smoothing*** is a method that is commonly used in forecasting in which a weighted average of past values of the variable are used to forecast values of the variable.
5. In a ***time trend regression***, we use a time trend as our independent variable. The time trend can be either linear or nonlinear. In a nonlinear time trend, we use a quadratic model to specify the trend relationship in order to pick up any nonlinear effects.
6. The ***Holt–Winters forecasting model*** consists of both an exponentially smoothed component and a trend component to forecast future values.
7. In an ***autoregressive process***, current values of the dependent variable are assumed to depend on its past values.

Useful Formulas

k-term moving average:

$$Z_t = \frac{1}{k}\sum_{i=0}^{k-1} x_{t-i}, \quad t = k,\ldots,N.$$

k-term weighted moving average:

$$Z_t = \sum_{i=0}^{k-1} w_{t-i} x_{t-i},\ t = k, \ldots, N.$$

where $\sum_{i=0}^{k-1} w_{t-i} = 1$ and $w_{t-i} \geq 0$ for all $i = 1, \ldots, k-1$.

Exponential smoothing:

$$\hat{X}_{t+1} = S_t = \alpha X_t + (1-\alpha) S_{t-1}.$$

Also, $\hat{X}_{t+1} = S_t = S_{t-1} + \alpha(X_t - \alpha S_{t-1})$.

Linear time trend:

$$X_t = \alpha + \beta t + \varepsilon_t.$$

Nonlinear time trend:

$$X_t = \alpha + \beta t + \gamma t^2 + \varepsilon_t,$$

$$X_t = \alpha + \beta t + \gamma t^2 + \delta t^3 + \varepsilon_t.$$

k-th order autoregressive process:

$$y_t = \alpha + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_k y_{t-k} + \varepsilon_t.$$

Example Problems

Example 1 Identifying the Components of a Time Series

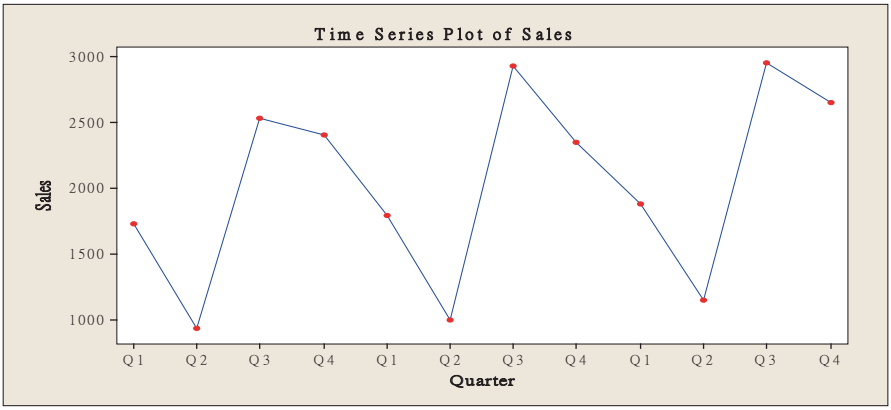
Year	Quarter	Sales
1	1	1,731
	2	935
	3	2,529
	4	2,400
2	1	1,789
	2	1,000
	3	2,931
	4	2,350
3	1	1,875
	2	1,150

Year	Quarter	Sales
	3	2,950
	4	2,650

Graph the data and identify the components of this time series.

Solution:

From the time-series plot, this data exhibits seasonal, trend, and irregular components.



Example 2 Simple Moving Average and Centered Moving Average

Use the data given in Example 1 to find the four-period moving average and the centered moving average.

Solution: The first four-period moving average would be the arithmetic average of the first four observations.

$$\text{Moving average} = (1731 + 935 + 2529 + 2400)/4 = 1898.75.$$

The moving average for the first four observations is centered around observations 1–4 or at observation 2.5. The second moving average is centered around observations 2–5 or at observation 3.5, and so on. In order to find the moving average at observation 3, we need to take the average of our moving averages for observation 2.5 and observation 3.5. This is called a centered moving average. The centered

moving average for the third quarter of the first year is the average of the moving average at quarter 2.5 and the moving average at quarter 3.5.

$$\text{Centered MA for Quarter 3} = (1898.75 + 1913.25)/2 = 1906.$$

The rest of the moving averages and centered moving averages can be computed in a similar manner as follows:

Year	Quarter	Sales	4-Period MA	Centered MA
1	1	1731		
	2	935		
			1898.75	
	3	2529		1906
			1913.25	
	4	2400		1921.38
			1929.5	
2	1	1789		1979.75
			2030	
	2	1000		2023.75
			2017.5	
	3	2931		2028.25
			2039	
	4	2350		2057.75
			2075.5	
3	1	1875		2078.88
			2081.25	
	2	1150		2118.75
			2156.25	
	3	2950		
	4	2650		

Example 3 Seasonal and Irregular Component

Use your results from Example 2 to construct the seasonal and irregular components of the data.

Solution: The seasonal and irregular components are constructed by dividing the sales by the centered moving average and multiplying by 100.

The seasonal and irregular component for the third quarter of the first year would be

$$(2529/1906)100 = 132.69.$$

Year	Quarter	(X_t) Sales	(Z_t) Centered MA	$(X_t/Z_t)100\ S_t \times I_t$
1	1	1731		
	2	935		
	3	2529	1906	132.69
	4	2400	1921.38	124.91
2	1	1789	1979.75	90.36
	2	1000	2023.75	49.41
	3	2931	2028.25	144.51
	4	2350	2057.75	114.20
3	1	1875	2078.88	90.19
	2	1150	2118.75	54.28
	3	2950		
	4	2650		

Example 4 Seasonal Component

Use your results from Example 3 to separate the seasonal and irregular components of the time series.

Solution: Notice in Example 3, we have two seasonal and irregular components for each quarter. To remove the irregular component from the seasonal factor, we take the average of each quarter's values.

$$S_1 = (90.36 + 90.19)/2 = 90.28 \text{ (1st quarter adjustment).}$$

$$S_2 = (49.41 + 54.28)/2 = 51.85 \text{ (2nd quarter adjustment).}$$

$$S_3 = (132.69 + 144.51)/2 = 138.60 \text{ (3rd quarter adjustment).}$$

$$S_4 = (124.91 + 114.20)/2 = 119.56 \text{ (4th quarter adjustment).}$$

Example 5 Seasonally Adjusted Sales

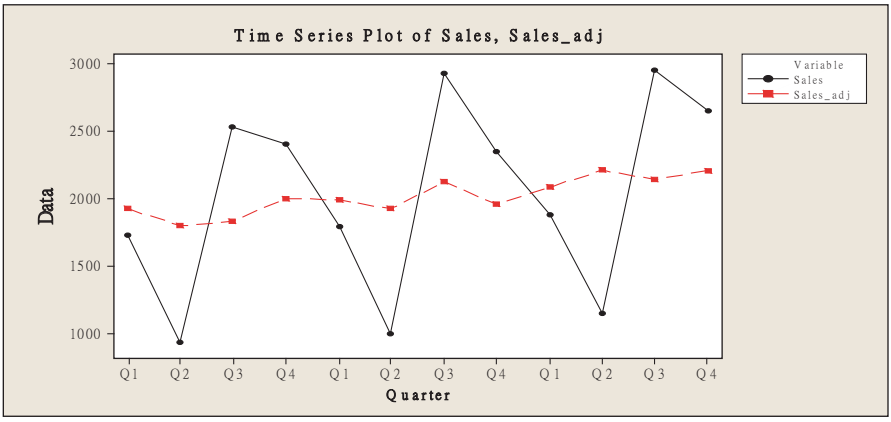
Use your results from Example 4 to seasonally adjust the original sales data. Plot the seasonally adjusted sales and the original sales data. Has the adjustment removed the seasonal effect?

Solution: To seasonally adjust the time series, we divide the original sales data by the seasonal factor and multiply by 100. (Note: We will round the seasonal factors off to 90, 52, 138, and 120). The first year's sales in the first quarter would then be

$$\text{Adjusted Sales} = (1731/90)(100)=1923.33.$$

The rest of the seasonally adjusted sales would be found in a similar manner.

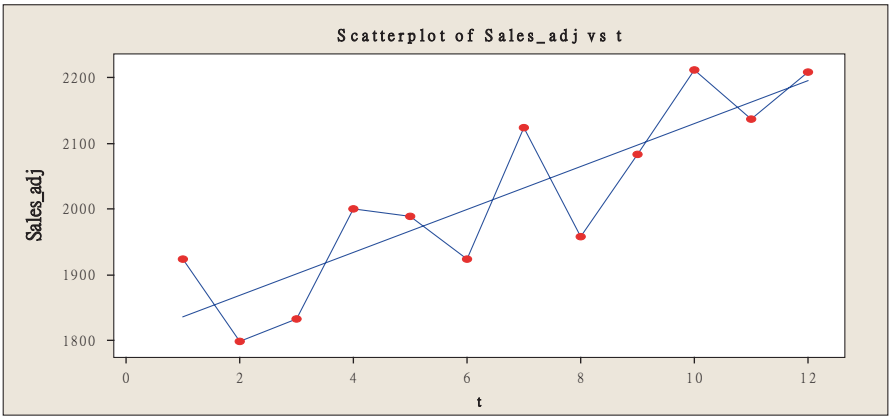
Year	Quarter	(X_t) Sales	(S_t) Seasonal adjustment	(X_t/S_t)100 Adjusted sales
1	1	1731	90	1923.33
	2	935	52	1798.08
	3	2529	138	1832.61
	4	2400	120	2000.00
2	1	1789	90	1987.78
	2	1000	52	1923.08
	3	2931	138	2123.91
	4	2350	120	1958.33
3	1	1875	90	2083.33
	2	1150	52	2211.54
	3	2950	138	2137.68
	4	2650	120	2208.33



Example 6 Linear Trend Regression

Run a regression on the deseasonalized data you computed in Example 5 to estimate the trend line. Use the trend line to forecast sales for the next four quarters.

Solution:



To estimate the trend line, we regress the seasonally adjusted data on time $(t = 1, 2, \ldots, 12)$. The estimated regression is

$$\hat{S}\_adj_t = 1802.64 + 32.77\, t,\ R^2 = 0.735.$$

To forecast adjusted sales in periods 13–16, we substitute the time period into the regression

$$\hat{S}\_adj_{13} = 1802.64 + 32.77(13) = 2228.69.$$

$$\hat{S}\_adj_{14} = 1802.64 + 32.77(14) = 2261.46.$$

$$\hat{S}\_adj_{15} = 1802.64 + 32.77(15) = 2294.24.$$

$$\hat{S}\_adj_{16} = 1802.64 + 32.77(16) = 2327.01.$$

From the above graph, we can see that the trend line forecasts seasonally adjusted sales. To get a forecast of the sales, we need to make an adjustment using the seasonal index we computed. The adjustment is to multiply the seasonal index and divide by 100.

$$\hat{S}_{13} = 2228.69(90)/100 = 2005.82.$$

$$\hat{S}_{14} = 2261.46(52)/100 = 1175.96.$$

$$\hat{S}_{15} = 2294.24(138)/100 = 3166.05.$$

$$\hat{S}_{16} = 2327.01(120)/100 = 2792.41.$$

Example 7 Exponential Smoothing

The following are car sales data at a local dealership during the first 12 weeks of the year.

Week	Sales	Week	Sales
1	15	7	11
2	14	8	10
3	13	9	12
4	12	10	10
5	14	11	12
6	13	12	13

Use the exponential smoothing method to forecast the sales for the first four weeks. Assume $\alpha = 0.2.$

Solution: The formula for exponential smoothing looks at the last period's forecast, plus an adjustment based on our forecast error from the previous period.

$$\hat{X}_{t+1} = S_t = S_{t-1} + \alpha(X_t - \alpha S_{t-1}).$$

$$\hat{X}_1 = X_1 = S_0 = 15.$$

$$\hat{X}_2 = S_1 = S_0 + \alpha(X_1 - S_0) = 15 + .2(15 - 15) = 15.$$

$$\hat{X}_3 = S_2 = 15 + 0.2(14 - 15) = 14.8.$$

$$\hat{X}_4 = S_3 = 14.9 + 0.2(13 - 14.8) = 14.44.$$

Example 8 Autoregressive Model

Use the data given in Example 1 to estimate the parameters of a first-order autoregressive model. Forecast sales one period into the future using this model.

Solution:

To estimate the parameters of a first-order autoregressive model, we regress the sales against last period's sales. The regression we estimate is

$$X_t = \alpha + \beta X_{t-1} + \varepsilon_t.$$

Using any regression package, we can estimate the model as

$$\hat{X}_t = 2,292.81 - .123X_{t-1},\ R^2 = .0142.$$

We can forecast the next quarter's sales by substituting the current period's sales into the estimated regression to get.

$$\hat{X}_{13} = 2,292.81 - .123(2,650) = 1,966.86.$$

Example 9 Exponential Smoothing—Choosing the Smoothing Parameter

Suppose the smoothing parameter α is unknown. Suggest a method for estimating α. Use this method to find the best α using the data from Example 7.

Solution: One method for selecting the best smoothing parameter is to choose the α with the smallest mean squared error (MSE) for the forecasts.

The MSE is computed as $MSE = \sum e_t^2 / n$,

where $n = 11$ is the number of forecasts excluding the 1st period, and $e_t = X_t - \hat{X}_t$, while $\hat{X}_t$ is the forecast of exponential smoothing.

For various values of α, i.e., $\alpha = 0.1, 0.2, \ldots, 0.9$, the corresponding $\hat{X}_t$, e_t^2 are listed in the following table:

0.1		0.20		0.30		0.40		0.50		0.6		0.70		0.80		0.90	
$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2	$\hat{X}_t$	e_t^2
15	0.00	15.00	0.00	15.00	0.00	15.00	0.00	15.00	0.00	15	0.00	15.00	0.00	15.00	0.00	15.00	0.00
15.00	1.00	15.00	1.00	15.00	1.00	15.00	1.00	15.00	1.00	15.00	1.00	15.00	1.00	15.00	1.00	15.00	1.00
14.90	3.61	14.80	3.24	14.70	2.89	14.60	2.56	14.50	2.25	14.40	1.96	14.30	1.69	14.20	1.44	14.10	1.21
14.71	7.34	14.44	5.95	14.19	4.80	13.96	3.84	13.75	3.06	13.56	2.43	13.39	1.93	13.24	1.54	13.11	1.23
14.44	0.19	13.95	0.00	13.53	0.22	13.18	0.68	12.88	1.27	12.62	1.89	12.42	2.51	12.25	3.07	12.11	3.57
14.40	1.95	13.96	0.92	13.67	0.45	13.51	0.26	13.44	0.19	13.45	0.20	13.53	0.28	13.65	0.42	13.81	0.66
14.26	10.60	13.77	7.67	13.47	6.11	13.30	5.31	13.22	4.92	13.18	4.75	13.16	4.65	13.13	4.54	13.08	4.33
13.93	15.45	13.22	10.34	12.73	7.45	12.38	5.67	12.11	4.45	11.87	3.50	11.65	2.71	11.43	2.03	11.21	1.46
13.54	2.36	12.57	0.33	11.91	0.01	11.43	0.33	11.05	0.89	10.75	1.57	10.49	2.27	10.29	2.94	10.12	3.53
13.38	11.45	12.46	6.04	11.94	3.75	11.66	2.75	11.53	2.33	11.50	2.25	11.55	2.40	11.66	2.75	11.81	3.28
13.04	1.09	11.97	0.00	11.36	0.41	10.99	1.01	10.76	1.53	10.60	1.96	10.46	2.36	10.33	2.78	10.18	3.31
12.94	0.00	11.97	1.05	11.55	2.10	11.40	2.57	11.38	2.62	11.44	2.43	11.54	2.13	11.67	1.78	11.82	1.40
MSE =	5.00		3.32		2.65		2.36		2.23		2.178		2.175		2.21		2.27

The MSE for different values of α are presented below.

α	MSE
0.1	5.00
0.2	3.32
0.3	2.65
0.4	2.36
0.5	2.23
0.6	2.178
0.7	2.175
0.8	2.20
0.9	2.27

Since MSE is minimized at $\alpha = 0.7$, the best α is 0.7.

Example 10 Holt–Winters Forecasting Method

Use the Holt–Winters method to forecast the sales in the 10th period for the following data. Assume that the smoothing constants are $\alpha = 0.6$ and $\beta = 0.5$.

Time	Sales
1	37.8
2	38.0
3	38.6
4	40.3
5	41.2
6	43.4
7	43.9
8	44.7
9	45.7
10	46.4

Solution: $S_t = .6X_t + (1-.6)(S_{t-1} + T_{t-1})$

$$T_t = .5(S_t - S_{t-1}) + (1-.5)T_{t-1}$$

To begin the system we use $T_1 = 0$ and $S_1 = X_1 = 37.8$

$$S_2 = \alpha X_2 + (1-\alpha)(S_1 + T_1)$$

$$= .6(38.0) + (1-.6)(37.8+0) = 37.92$$

$$T_2 = \beta(S_2 - S_1) + (1-\beta)T_1$$

$$= .5(37.92 - 37.8) + (1 - .5)0 = .06$$

$$S_3 = \alpha X_3 + (1-\alpha)(S_2 + T_2)$$

$$= .6(38.6) + (1 - .6)(37.92 + .06) = 38.352$$

$$T_3 = \beta(S_3 - S_2) + (1-\beta)T_2$$

$$= .5(38.352 - 37.92) + (1 - .5).06 = .246$$

The rest of the values are presented in the table below.

t	Sales	S_t	T_t
1	37.8	37.8	0
2	38.0	37.920	0.060
3	38.6	38.352	0.246
4	40.3	39.619	0.757
5	41.2	40.870	1.004
6	43.4	42.790	1.462
7	43.9	44.041	1.356
8	44.7	44.979	1.147
9	45.7	45.870	1.019
10	46.4	46.596	0.872

Example 11 Autoregressive Mode

Estimate the first-order autoregressive model for the data given below. Use the model to produce a forecast for the 16th period.

Time	
1	1.5
2	0.48
3	1.06
4	0.78
5	0.95
6	0.89
7	0.92
8	0.88
9	0.84
10	0.83

Time	
11	0.89
12	0.91
13	0.96
14	0.86
15	0.93

Solution:

The first-order autoregressive model is computed by regressing the current period value of *X* against the previous period. The estimated regression is

$$\hat{X}_t = 1.390 - 0.571X_{t-1} \quad R^2 = 0.93$$

To forecast the value in the 16th period we substitute into the estimated regression to get

$$\hat{X}_{16} = 1.390 - 0.571(0.93) = 0.86$$

Supplementary Exercises

Multiple Choice

1. To construct the current 30-day moving average of stock prices, we

   a. Use the first 30 days of stock prices
   b. Use the most recent 30 days of stock prices
   c. Use any 30 days of stock prices
   d. Divide the most recent stock price by 30
   e. Multiply the most recent stock price by 30

2. Time series analysis is used to analyze data

   a. Over different time periods
   b. Across different companies
   c. Across different companies and across different time periods
   d. That are qualitative
   e. Provided they are based on stock prices

3. A record store notices that sales increase on Elvis's birthday. The "Elvis phenomenon" is

   a. A trend component
   b. An irregular component
   c. A seasonal component

d. A cyclical component
e. A random component

4. If the last eight observations of a time series are: a, b, c, d, e, f, g, h, then the current three-period moving average would be
   a. $(a + b + c)/3$
   b. $(a + b + c) \times 3$
   c. $h/3$
   d. $(f + g + h)/3$
   e. $(f + g + h) \times 3$
5. Which of the following would likely be a seasonal component of a time series?
   a. Holidays
   b. Law suits
   c. Recessions
   d. Population growth
   e. Depression
6. Which of the following would likely be a trend component of a time series?
   a. Holidays
   b. Law suits
   c. Recessions
   d. Population growth
   e. Depression
7. Exponential smoothing requires
   a. Past values of the time series
   b. Current values of the time series
   c. Both past and current values of th time series
   d. Estimation of a time trend regression
   e. Seasonal adjustments
8. When using simple exponential smoothing, the use of a larger smoothing coefficient means
   a. More weight is given to past observations
   b. Less weight is given to current observations
   c. More weight is given to current observations
   d. Equal weight is given to past and current observations
   e. It is impossible to determine which observations receive greater weight
9. The Holt–Winters forecasting model
   a. Is a simple exponential smoothing model
   b. Is a time trend model
   c. Is a moving average model

d. Contains both a time trend and exponential smoothing
e. Is a cyclical model

10. An autoregressive model

a. Is a time trend model
b. Is a method for seasonally adjusting data
c. Uses past values of the data to forecast future values
d. Is an exponential smoothing model
e. Is a cyclical model

11. Variables that tend to precede turning points in economic activity are called

a. Leading indicators
b. Coincidence indicators
c. Lagging indicators
d. Moving averages
e. Exponentially smoothed values

12. Variables that tend to match turning points in economic activity are called

a. Leading indicators
b. Coincidence indicators
c. Lagging indicators
d. Moving averages
e. Exponentially smoothed values

13. Variables that tend to follow turning points in economic activity are called

a. Leading indicators
b. Coincidence indicators
c. Lagging indicators
d. Moving averages
e. Exponentially smoothed values

14. Data can be deseasonalized by using

a. The moving average
b. Exponential smoothing
c. An autoregressive process
d. A trend regression
e. Leading indicators

15. If the sales for a company exhibit constant growth over time, the best method for forecasting would be

a. A linear time trend
b. A log-linear time trend
c. A simple moving average
d. A first-order autoregressive model
e. A centered moving average

16. If time series data has been collected on an annual basis only, which component can we ignore?

    a. Trend
    b. Seasonal
    c. Cyclical
    d. Irregular
    e. MSE

17. Which of the followings is correct when using exponential smoothing to forecast?

    a. Data from past periods is of equal importance
    b. Data from past periods is of exponentially increasing importance
    c. Data from past periods is of exponentially decreasing importance
    d. Data from past periods is of linearly decreasing importance
    e. Data from past periods is ignored

18. In using exponential smoothing, the actual value for last month is 10 and its forecast value is 8, what is the forecast value for this month if the smoothing constant is $\alpha = 0.30$?

    a. 10
    b. 9.4
    c. 9
    d. 8
    e. 8.6

19. Which of the followings is equivalent to the "ratios of actuals to moving average" (i.e., x_t / z_t^*)?

    a. The product of seasonal and irregular components
    b. The product of seasonal and cyclical components
    c. The product of trend and irregular components
    d. The product of cyclical and irregular components
    e. The product of trend and cyclical components

20. What is (are) the independent variable(s) in an autoregressive forecasting model?

    a. Time-lagged values of the independent variable
    b. Time-lagged values of the dependent variable
    c. The difference between the current and previous values of the dependent variable
    d. The difference between the current and previous values of the independent variable
    e. Period in time

True/False (If False, Explain Why)

1. A simple moving average could lie consistently above the original data.
2. A simple moving average will exhibit more variability than the original data.
3. A simple moving average can be used to forecast cross-sectional data.
4. Regression cannot be used to forecast a time series.
5. The more stable the time series, the smaller the smoothing parameter should be.
6. If a company's earnings exhibit high earnings in the first and third quarters and lower earnings in second and fourth quarters, the time series is cyclical.
7. The larger the smoothing coefficient in exponential smoothing, the greater the weight given to current observations.
8. Moving averages can be used to deseasonalize the data.
9. For a trend to exist, the data must be growing steadily over time.
10. The high sales that most stores experience around the holiday season are a cyclical effect.
11. Cyclical patterns are long-term oscillatory patterns that are related to seasonal behavior.
12. Leading economic indicators get their name because they are the most important indicators of economic growth.
13. When we are looking at quarterly data, a four-period simple moving average will exhibit a pronounced seasonal effect.
14. A simple time trend regression is the best method for forecasting a variable that shows constant growth over time.
15. The Holt–Winters model uses both smoothing and a trend to forecast a time series.
16. Because cyclical components can confound trend analysis, it is important to deseasonalize the data before using regression analysis to establish a trend model.
17. Forecast error is the difference between the forecast value and the actual value.
18. The seasonal component reflects a regular, multiyear pattern of being above and belowthe trend line.
19. If a time series has no significant trend, cyclical, or seasonal effect, then it is appropriate to use the moving-average method to do forecasting.
20. To choose the smoothing constant (α), a value of α with the largest MSE is usually recommended.

Questions and Problems

1. Briefly explain why we must deseasonalize data before we can use a trend to forecast future values.
2. Briefly explain why we cannot use exponential smoothing when a trend exists.

3. You are given the following quarterly sales information for XYZ Company over the last 3 years.

Year	Quarter	Sales
1	1	3800
	2	3700
	3	3900
	4	4000
2	1	3940
	2	3800
	3	4020
	4	4140
3	1	4100
	2	4080
	3	4300
	4	4440

   a. Find the current four-period moving average.
   b. Find the four-period moving averages, the centered moving averages, and obtain the seasonal index for each quarter.

4. Use your results from Question 3 to seasonally adjust the original sales data.

   a. Plot the seasonally adjusted sales and the original sales data. Has the adjustment removed the seasonal effect?
   b. Run a linear regression on the deseasonalized data you computed in Example 5 to estimate the trend line. Use the trend line to forecast sales for the next four quarters.
   c. Run a quadratic regression on the deseasonalized data you computed in Example 5 to estimate the trend line. Use the trend line to forecast sales for the next four quarters.

5. Below are sales data for XYZ Company.

Time	Sales	Time	Sales
1	44.3	6	50.2
2	47.8	7	51.1
3	47.6	8	50.9
4	48.3	9	52.2
5	49.4	10	53.3

Use the Holt–Winters forecasting model to forecast sales for the 10 periods. Assume that $\alpha = 0.6$ and $\beta = 0.5$.

6. Below are sales data for ABC Company.

Period	Sales	Period	Sales
1	146.5540	16	176.6588
2	159.4821	17	173.3224
3	171.8967	18	170.7909
4	172.9721	19	169.0809
5	176.6022	20	167.7419
6	172.0441	21	172.9483
7	171.3946	22	178.0765
8	169.2607	23	177.8793
9	175.6464	24	173.5440
10	177.4715	25	176.0660
11	176.3656	26	179.8407
12	180.3656	27	173.2309
13	175.0408	28	170.4035
14	177.9652	29	169.2317
15	173.1053		

Use a computer program to estimate the first-order autoregressive model to forecast sales in the 30th period.

7. Below are the sales data for the Smart Computer Company over the last 20 months. Use a time trend regression to forecast sales for the 21st month.

Month	Sales	Month	Sales
1	370.2467	11	370.9667
2	370.8606	12	374.5309
3	371.2098	13	376.6496
4	372.3486	14	376.9085
5	372.7624	15	376.071
6	373.7477	16	378.6685
7	372.0169	17	376.0601
8	373.8907	18	370.9929
9	370.3686	19	379.3183
10	373.3517	20	378.1466

Answers to Supplementary Exercises

Multiple Choice

1.	b	6.	d	11.	a	16.	b
2.	a	7.	c	12.	b	17.	c
3.	c	8.	c	13.	c	18.	e
4.	d	9.	d	14.	a	19.	a
5.	a	10.	c	15.	b	20.	b

True/False

1. True
2. False. The moving average will be less variable.
3. False. A simple moving average is used to forecast time series data.
4. False. Regression analysis can be very useful for forecasting time series data.
5. False. If a time series is stable, we would want to place more weight on the most recent observation, so we should use a larger smoothing coefficient.
6. False. Seasonal effect.
7. True
8. True
9. False. The data could also be falling over time.
10. False. Seasonal effect.
11. False. The cyclical component should be unrelated to seasonal effects.
12. False. Get their name because they precede turning points in economic activity.
13. False. Because each moving average observation incorporates a full year of data, it will not show a seasonal effect.
14. False. A log-linear time trend is best for forecasting a time series that shows constant growth.
15. True
16. False, Seasonal component can confound trend analysis.
17. False. Forecast error = actual value – forecast value.
18. False. Cyclical component.
19. True.
20. False. Smallest MSE.

Questions and Problems

1. Because a simple time trend cannot capture the seasonal effect, we get more accurate forecasts if we remove the seasonal effect from the time series. Once the deseasonalized data have been forecast using a trend, we can adjust the values for the seasonal effect.
2. Exponential smoothing should not be used when a trend exists because the forecasts will always lag the actual values of the time series.
3. a. Moving average for the 13th period

   $= (4,100 + 4,80 + 4,300 + 4,440)/4 = 4,230$

   b. In the following table,

      1. Gives the sales data, x_t
      2. Gives the 4-peiod moving averages (MA), z_t
      3. Gives the centered 4-peiod moving averages (CMA), z_t^*
      4. Gives $100(x_t/z_t^*) = 100(S_t I_t)$
      5. Gives the seasonal index:

Seasonal index for 1st quarter = 100.
Seasonal index for 2nd quarter = 97.
Seasonal index for 3rd quarter = 101.
Seasonal index for 4th quarter = 102.

Year	Quarter	(1) Sales	(2) MA	(3) CMA	(4) $100(S_tI_t)$		(5) Seasonal Index
1	1	3800					
	2	3700					
			3850				
	3	3900		3867.5	100.84	100.73	101
			3885				
	4	4000		3897.5	102.63	102.43	102
			3910				
2	1	3940		3925.0	100.38	99.95	100
			3940				
	2	3800		3957.5	96.02	96.67	97
			3975				
	3	4020		3995.0	100.63		
			4015				
	4	4140		4050.0	102.22		
			4085				
3	1	4100		4120.0	99.51		
			4155				
	2	4080		4192.5	97.32		
			4230				
	3	4300					
	4	4440					

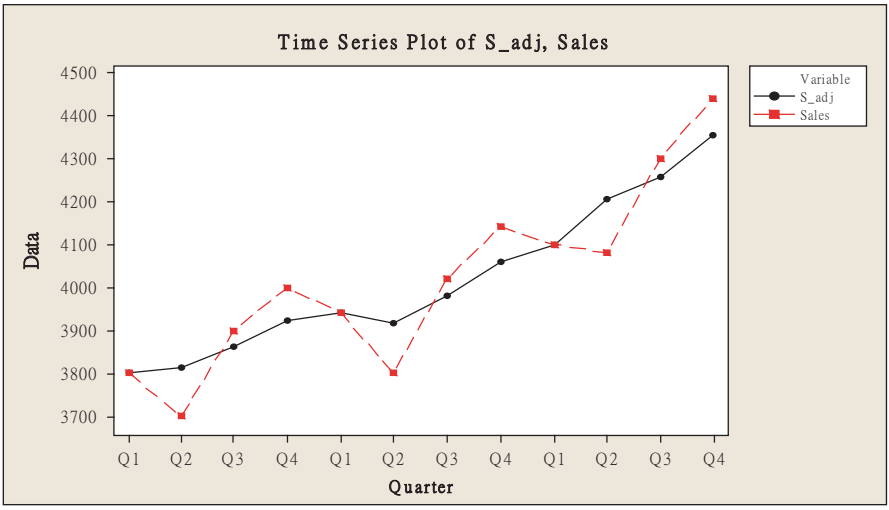
4. a.

$$S\_adj = 100(\text{Sales} / \text{Seasonal Index}).$$

Year	Quarter	Sales	Index	S_adj
1	1	3800	100	3800.00
	2	3700	97	3814.43
	3	3900	101	3861.39
	4	4000	102	3921.57
2	1	3940	100	3940.00
	2	3800	97	3917.53
	3	4020	101	3980.20
	4	4140	102	4058.82
3	1	4100	100	4100.00

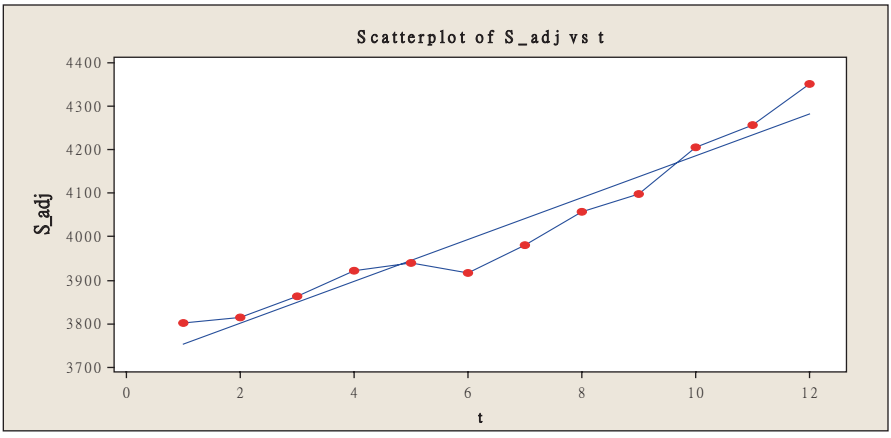
Year	Quarter	Sales	Index	S_adj
	2	4080	97	4206.19
	3	4300	101	4257.43
	4	4440	102	4352.94

Time series plot for adjusted sales and sales:



The adjusted sales has been deseaonalized.

b. To estimate the trend line, we regress the seasonally adjusted data on time ($t = 1, 2, \ldots, 12$). From the scatter plot, a linear trend regression will be fitted.



The estimated regression linear in time is

$$\hat{S}\_adj_t = 3704.04 + 48.23t,\ R^2 = 0.94.$$

To forecast adjusted sales in periods 13–16, we substitute the time period into the regression.

$$\hat{S}\_adj_{13} = 3704.04 + 48.23(13) = 4331.3.$$

$$\hat{S}\_adj_{14} = 3806.48 + 48.23(14) = 4379.26.$$

$$\hat{S}\_adj_{15} = 3806.48 + 48.23(15) = 4427.49.$$

$$\hat{S}\_adj_{16} = 3806.48 + 48.23(16) = 4475.72.$$

To forecast the sales we need to make an adjustment using the seasonal index we computed. The adjustment is to multiply the seasonal index and divide by 100.

$$\hat{S}_{13} = 4331.03(100)/100 = 4331.03.$$

$$\hat{S}_{14} = 4379.26(97)/100 = 4247.88.$$

$$\hat{S}_{15} = 4427.49(101)/100 = 4471.77.$$

$$\hat{S}_{16} = 4475.72(102)/100 = 4565.23.$$

c. The estimated regression quadratic in time is

$$\hat{S}\_adj_t = 3806.48 + 4.326\, t + 3.377\, t^2,\ R^2 = 0.984$$

The forecasts for the adjusted sales in periods 13 to16 are:

$$\hat{S}\_adj_{13} = 3806.48 + 4.326(13) + 3.377(13^2) = 4433.43.$$

$$\hat{S}\_adj_{14} = 3806.48 + 4.326(14) + 3.377\ (14^2) = 4528.94$$

$$\hat{S}\_adj_{15} = 3806.48 + 4.326(15) + 3.377\ (15^2) = 4631.20.$$

$$\hat{S}\_adj_{16} = 3806.48 + 4.326(16) + 3.377\ (16^2) = 4740.21.$$

The forecasts for sales in periods 13–16 are:

$$\hat{S}_{13} = 4433.43\ (100)/100 = 4433.43.$$

$$\hat{S}_{14} = 4528.94\ (97)/100 = 4393.07.$$

$$\hat{S}_{15} = 4631.20\ (101)/100 = 4677.51.$$

$$\hat{S}_{16} = 4740.21(102)/100 = 4835.01.$$

5. $S_t = 0.6X_t + (1-0.6)(S_{t-1} + T_{t-1})$.

$$T_t = 0.5(S_t - S_{t-1}) + (1-0.5)T_{t-1}.$$

To begin the system we use $T_1 = 0$ and $S_1 = X_1 = 44.3$.

t	X_t	S_t	T_t	$\hat{X}_t$
1	44.3	44.300	0.000	44.300
2	47.8	46.400	1.050	47.450
3	47.6	47.540	1.095	48.635
4	48.3	48.434	0.995	49.429
5	49.4	49.411	0.986	50.397
6	50.2	50.279	0.927	51.206
7	51.1	51.142	0.895	52.037
8	50.9	51.355	0.554	51.909
9	52.2	52.084	0.641	52.725
10	53.3	53.070	0.814	53.884

6. To estimate a first-order autoregressive model, we regress sales on sales lagged one period. Using a computer program the estimated model is

$$\hat{X}_t = 96.38 + .447\ X_{t-1}, \qquad R^2 = .466.$$

To forecast sales in the 30th period we substitute in sales in period 29 to get

$$\hat{X}_{30} = 96.38 + 0.447(169.2317) = 172.03.$$

7. Using a computer program we estimate the time trend regression as

$$\hat{X}_t = 370.144 + 0.363t, \qquad R^2 = 0.537.$$

The forecasted sales in period 21 are

$$\hat{X}_{21} = 370.144 + 0.363(21) = 377.77.$$

Chapter 19
Index Numbers and Stock Market Indexes

Chapter Intuition

Throughout the text, we have learned that one of the most important uses of statistics is to summarize data and to make comparisons. Often we are interested in comparing the past and the present. For example, we may be interested in how prices have changed over time. However, simply looking at changes in the price of one good at a time is impractical and not very useful. Although the prices of most items have risen over the past 10 years, the cost of many electronic goods such as computers, VCRs, and televisions have fallen dramatically. In addition, many items commonly used today were not even in existence several years ago. One way to compare price changes in two different time periods is by constructing an index of prices. This chapter explains the different methods for constructing index numbers.

We can also use index numbers to make comparisons in the same time period. For example, many people are interested in how their stocks have done compared to the rest of the stock market. One way to make this comparison is to construct an index of stock prices such as the Dow Jones industrial average (DJIA) and to compare the performance of our stocks to the index.

Chapter Review

1. Simply put, a ***price index*** represents a weighted average cost for purchasing a bundle of goods. By comparing the price index for a bundle of goods in different periods, we can see if it has gotten more or less expensive to purchase these goods.
2. One difficulty with constructing a price index is determining the appropriate bundle of goods. Although food makes an important part of a household's budget, and therefore should be represented in our price index, changes in eating habits mean that we are consuming different foods than previous generations. In addition, there are goods that are an important part of most household budgets

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_19

today, such as DVD players and smart phones that did not even exist 20 years ago.

3. In order to construct an index, we need to determine the ***base year***. The base year is the reference point from which we measure changes in prices, quantities, or values.
4. The simplest price index is a ***price relative***. A price relative is the ratio of the current price of one commodity to its base-year price. Price relatives allow us to look at the price change of a single commodity between the base year and the current year.
5. A ***simple aggregative price index*** is the ratio of the total cost of various commodities in the current year to the total cost in the base year. There are two problems with this type of index: it does not consider the relative importance of each of the commodities, and the units used can affect the index value. For example, if we use the price of one egg, rather than the cost of 1 dozen eggs, when we construct the index, eggs will become less important in the value of the index.
6. ***Weighted aggregative price indexes*** look at both the prices and quantities. The ***Laspeyres price index*** uses base-year quantities. The ***Paasche price*** *index* uses current-year quantities. ***Fisher's ideal price index*** is a geometric average of the Laspeyres and Paasche indexes.
7. If we are interested in how the quantity of goods purchased has changed, we can construct a ***quantity index***. ***Weighted aggregative quantity*** *indexes* also look at both the prices and quantities. The ***Laspeyres quantity index*** uses base-year prices. The ***Paasche quantity*** *index* uses current-year prices. ***Fisher's ideal quantity index*** is a geometric average of the Laspeyres and Paasche indexes.
8. If we are interested in how the total cost of goods purchased has changed compared to the cost of these goods in the base period at the base-year price, we can construct a ***value index***.
9. ***Stock market indexes*** show how the prices of a group of stocks change over time. The three most common types of stock market indexes are: market-value-weighted indexes, price-weighted indexes, and equally weighted indexes.

Useful Formulas

Price Indexes

Price relatives

$$\frac{P_{ti}}{P_{0i}}$$

P_{ti} the price of commodity i at time t,
P_{0i} the price of commodity i at base period.

Simple aggregate price index

$$I_t = \frac{\sum_{i=1}^{n} P_{ti}}{\sum_{i=1}^{n} P_{0i}} \times 100.$$

Simple average of price relatives

$$I_t = \frac{\sum_{i=1}^{n} P_{ti}/P_{0i}}{n} \times 100.$$

Laspeyres index

$$I_t = \frac{\sum_{i=1}^{n} P_{ti}Q_{0i}}{\sum_{i=1}^{n} P_{0i}Q_{0i}} \times 100.$$

Paasche index

$$I_t = \frac{\sum_{i=1}^{n} P_{ti}Q_{ti}}{\sum_{i=1}^{n} P_{0i}Q_{ti}} \times 100.$$

Fisher's ideal price index

$$FI_t = \sqrt{\frac{\sum_{i=1}^{n} P_{ti}Q_{0i}}{\sum_{i=1}^{n} P_{0i}Q_{0i}} \times \frac{\sum_{i=1}^{n} P_{ti}Q_{ti}}{\sum_{i=1}^{n} P_{0i}Q_{ti}}} \times 100.$$

Quantity indexes

Laspeyres

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{0i}}{\sum_{i=1}^{n} Q_{0i}P_{0i}} \times 100.$$

Paasche

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti} P_{ti}}{\sum_{i=1}^{n} Q_{0i} P_{ti}} \times 100.$$

Fisher's quantity index

$$FI_t = \sqrt{\frac{\sum_{i=1}^{n} Q_{ti} P_{0i}}{\sum_{i=1}^{n} Q_{0i} P_{0i}} \times \frac{\sum_{i=1}^{n} Q_{ti} P_{ti}}{\sum_{i=1}^{n} Q_{0i} P_{ti}}} \times 100.$$

Value indexe

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti} P_{ti}}{\sum_{i=1}^{n} Q_{0i} P_{0i}} \times 100.$$

Example Problems

Example 1 Price Relatives

Below are the prices and quantities for three commodities in 2012 and 2013.

Commodity	2012		2013	
	Price ($)	Quantity	Price ($)	Quantity
Eggs	1.00	22 dozen	1.21	20 dozen
Milk	1.95	16 gallon	2.05	18 gallon
Beef	3.22	11 lb.	3.20	12 lb.

Find the price relative for each commodity using 2012 as the base year.

Solution:

$$\text{Price relative} = P_{ti} / P_{0i}$$

$$\text{Eggs} = 1.21 / 1.00 = 1.21$$

$$\text{Milk} = 2.05 / 1.95 = 1.05$$

$$\text{Beef} = 3.20 / 3.22 = 0.99.$$

Example 2 Simple Aggregate Price Index

Using the data from Example 1, compute the simple aggregate price index. Use 2012 as the base year.

Solution:

$$I_t = \frac{\sum_{i=1}^{n} P_{ti}}{\sum_{i=1}^{n} P_{0i}} \times 100 = \frac{1.21 + 2.05 + 3.20}{1.00 + 1.95 + 3.22} \times 100 = 104.7.$$

Example 3 Simple Average of Price Relatives

Use the data and your results from Example 1 to compute the simple average of price relatives.

Solution:

$$I_t = \frac{\sum_{i=1}^{n} P_{ti}/P_{0i}}{n} \times 100 = \frac{1.21 + 1.051 + 0.994}{3} \times 100 = 108.5.$$

Example 4 Laspeyres Price Index

Using the data from Example 1, compute the Laspeyres price index. Use 2012 as the base year.

Solution:

$$I_t = \frac{\sum_{i=1}^{3} P_{ti} Q_{0i}}{\sum_{i=1}^{3} P_{0i} Q_{0i}} \times 100$$

$$= \frac{1.21(22) + 2.05(16) + 3.20(11)}{1.00(22) + 1.95(16) + 3.22(11)} \times 100 = 106.77.$$

Example 5 Paasche Price Index

Using the data from Example 1, compute the Paasche price index. Use 2012 as the base year.

Solution:

$$I_t = \frac{\sum_{i=1}^{3} P_{ti}Q_{ti}}{\sum_{i=1}^{3} P_{0i}Q_{ti}} \times 100$$
$$= \frac{1.21(20) + 2.05(18) + 3.20(12)}{1.00(20) + 1.95(18) + 3.22(12)} \times 100 = 106.14.$$

Example 6 Fisher's Ideal Price Index

Use your results from Examples 4 and 5 to compute Fisher's ideal price index

Solution:

$$FI_t = \sqrt{\frac{\sum_{i=1}^{n} P_{ti}Q_{0i}}{\sum_{i=1}^{n} P_{0i}Q_{0i}} \times \frac{\sum_{i=1}^{n} P_{ti}Q_{ti}}{\sum_{i=1}^{n} P_{0i}Q_{ti}}} \times 100$$
$$= \sqrt{(1.068)(1.061)} \times 100 = 106.4.$$

Example 7 Laspeyres Quantity Index

Below are prices and quantities for three commodities in 2012 and 2013.

Commodity	2012		2013	
	Price ($)	Quantity	Price ($)	Quantity
Shirts	22.00	6	19.00	8
Shoes	71.00	2	78.00	3
Dresses	53.00	8	62.00	7

Compute the Laspeyres quantity index using 2012 as the base year.

Solution:

$$I_t = \frac{\sum_{i=1}^{3} Q_{ti} P_{0i}}{\sum_{i=1}^{3} Q_{0i} P_{0i}} \times 100$$
$$= \frac{8(22) + 3(71) + 7(53)}{6(22) + 2(71) + 8(53)} \times 100 = 108.88.$$

Example 8 Paasche Quantity Index

Use the data in Example 7 to compute the Paasche quantity index. Use 2012 as the base year.

Solution:

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti} P_{ti}}{\sum_{i=1}^{n} Q_{0i} P_{ti}} \times 100$$
$$= \frac{8(19) + 3(78) + 7(62)}{6(19) + 2(78) + 8(62)} \times 100 = 107.05.$$

Example 9 Fisher's Ideal Quantity Index

Use your results from Examples 7 and 8 to compute Fisher's ideal quantity index.

Solution:

$$FI_t = \sqrt{\text{Laspeyres} \times \text{Paasche}}$$
$$= \sqrt{(108.88)(107.05)} = 107.96.$$

Example 10 Value Index

Use the data in Example 7 to compute the value index. Use 2012 as the base year.

Solution:

$$I_t = \frac{\sum_{i=1}^{n} Q_{ti} P_{ti}}{\sum_{i=1}^{n} Q_{0i} P_{0i}} \times 100$$
$$= \frac{8(19) + 3(78) + 7(62)}{6(22) + 2(71) + 8(53)} \times 100 = 117.48.$$

Example 11 Stock Market Index

Suppose the Little Stock Exchange (LYSE) trades only three stocks. Below are the prices for the stocks and the number of shares in 2012 and 2013.

	2012		2013	
Stock	Price ($)	Shares	Price ($)	Shares
Little Auto, Inc.	27	47	32	50
Little Phone Company	18	100	16	105
Little Computer	92	72	90	88

Calculate the market value weighted index and the price weighted index.

Solution:

Market-value weighted index
$$= \frac{\sum_{i=1}^{3} Q_{ti} P_{ti}}{\sum_{i=1}^{3} Q_{0i} P_{0i}} \times 100$$
$$= \frac{32(50) + 16(105) + 90(88)}{27(47) + 18(100) + 92(72)} \times 100 = 115.55.$$

Price-weighted index
$$= \frac{\sum p_{ti}/3}{\sum p_{0i}/3} \times 100 = \frac{(32 + 16 + 90)/3}{(27 + 18 + 92)/3} \times 100 = 100.73.$$

Supplementary Exercises

Multiple Choice

1. If small stocks on the New York Stock Exchange (NYSE) rise more than larger stocks, then an equal weighted index of NYSE stocks would

   a. Have a smaller return than the S&P 500 index
   b. Have a smaller return than a value weighted NYSE index
   c. Have a greater return than a value weighted NYSE index
   d. Have the same return as the value weighted NYSE index
   e. Cannot be determined by the information given

2. A value weighted stock price index would be

   a. More sensitive to changes in small stocks than an equal weighted index
   b. Less sensitive to changes in small stocks than an equal weighted index
   c. Equally sensitive to changes in small stocks as an equal weighted index
   d. Equally sensitive to changes in large stocks as an equal weighted index
   e. Less sensitive to changes in large stocks than an equal weighted index

3. Fisher's ideal price index

   a. Is a Laspeyres index
   b. Is a Paasche index
   c. Is a simple average index
   d. Is a geometric average of a Laspeyres index and a Paasche index
   e. Always equals 1

4. The DJIA is

   a. An index of consumer prices
   b. A Laspeyres index
   c. A Paasche index
   d. An index of industrial production
   e. A stock price index

5. The base year

   a. Is the reference year from which changes in the index are measured
   b. Is always the current year
   c. Is always last year
   d. Is the first year the index is created
   e. Is the last year the index is used

6. A price relative is

   a. Another name for the base year
   b. Another name for a Laspeyres index

c. Another name for a Paasche index
d. The ratio of the price in the current year to the price in the base year
e. The ratio of the price in the base year to the price in the current year

7. The GNP deflator is an example of a

a. Laspeyres index
b. Paasche index
c. Simple aggregate price index
d. Base-year price
e. Current-year price

8. The FRB index of industrial production measures

a. Changes in the DJIA
b. Changes in producer prices
c. Changes in consumer prices
d. Changes in the physical value of output
e. Changes in consumer well being

9. Which of the following is *not* one of the leading economic indicators?

a. New orders for durable goods industries
b. Corporate profits after taxes
c. Index of stock prices for 500 common stocks
d. Change in consumer prices
e. Average hourly workweek

10. The consumer price index (CPI) is an example of a

a. Laspeyres index
b. Paasche index
c. Simple aggregate price index.
d. Base-year price
e. Current-year price

11. A large change in the price of commodities that are held in large quantities in the base year but are held in smaller quantities in the current year will lead to

a. A greater change in the Paasche index relative to the Laspeyres index
b. A smaller change in the Paasche index relative to the Laspeyres index
c. An equal change in both the Paasche index and the Laspeyres index
d. A Laspeyres index equal to 0
e. A Paasche index equal to 0

12. A large change in the price of commodities that are held in large quantities in the current year but are held in smaller quantities in the base year will lead to

a. A greater change in the Paasche index relative to the Laspeyres index
b. A smaller change in the Paasche index relative to the Laspeyres index
c. An equal change in the Paasche index relative to the Laspeyres index

d. A Laspeyres index equal to 0
e. A Paasche index equal to 0

13. Which of the following is a market-value-weighted index similar to?

a. A Laspeyres price index
b. A Paasche price index
c. A Laspeyres quantity index
d. A Paasche quantity index
e. A value index

14. Which of the following is a price-weighted index?

a. Consumer price index
b. Dow Jones industrial average
c. S&P 500 index
d. NYSE index
e. Wilshire 5000 equity index

15. Which of the following is correct for index numbers?

a. Index numbers help us to compare means
b. Index numbers help us to compare variances
c. Index numbers help us to compare related items over time
d. Index numbers help us to compare related items over place
e. c and d

16. The FRB index of industrial production is a

a. A Laspeyres price index
b. A Paasche price index
c. A Laspeyres quantity index
d. A Paasche quantity index
e. A value index

17. DJIA is an arithmetic average of the stock prices of

a. All firms listed on NYSE
b. Five-hundred firms
c. All firms listed on OTC
d. Thirty blue-chip firms
e. Five-thousand stocks.

18. Usually, what are the weights for a weighted aggregative price index?

a. Prices of the included items
b. Prices of the items not included
c. Quantities of the included items
d. Quantities of the items not included
e. Equal weights

19. Which of the following is a market-value-weighted index?

    a. Consumer price index
    b. Dow Jones industrial average
    c. S&P 500 index
    d. NYSE index
    e. Both c and d

20. The gross national product (GNP) deflator is a

    a. A Paasche price index
    b. A Laspeyres price index
    c. A Paasche quantity index
    d. A Laspeyres quantity index
    e. A value index

True/False (If False, Explain Why)

1. An index number is a summary measure that allows for a comparison between a group of related items over time.
2. The most recent year is always used as the base year.
3. The CPI measures a market basket of goods.
4. A simple aggregate price index will not be affected by the units used to state prices.
5. A simple aggregate price index does not consider the relative importance of the commodities.
6. The Laspeyres index uses base-year quantities.
7. The GNP deflator is an example of a Laspeyres index.
8. Fisher's ideal price index is a geometric average of the Laspeyres and Paasche indexes.
9. A quantity index measures the change in quantity from a base year to a particular year.
10. One problem with constructing a price index is determining the appropriate market basket.
11. The S&P 500 is an equal weighted stock market index.
12. A stock market index is a statistical measure that allows us to see how the prices of a group of stocks has changed over time.
13. Nominal GDP can be calculated from dividing real GDP by GDP deflator.
14. The NYSE index includes only big firms listed on the NYSE.
15. The DJIA takes number of shares of 30 blue chips into account.
16. The CPI is a Laspeyres price index.
17. The Paasche quantity index for a bunch of foods is 98 for Year 2012. Then, the total cost of those foods is 2% more than what the base year would have cost had they been purchased on 2012.

18. The disadvantage of the Paasche price index is that it tends to five more weight to those items that show a dramatic price increase.
19. The complexity of updating the reference-year quantities for a Laspeyres price index makes it difficult to apply.
20. The advantage of the Paasche price index is that it uses current-year quantities to provide an up-to-data estimate of total expense.

Questions and Problems

Questions 1–10 use the following information:

Commodity	2013		2014	
	Price ($)	Quantity (lbs.)	Price ($)	Quantity (lbs.)
Beef	2.50	50	2.75	40
Chicken	1.50	70	1.25	90
Pork	3.25	35	3.75	35

Assume that 2013 is the base year.

1. Compute the price relatives for each good
2. Construct the simple aggregate price index
3. Construct the simple relative price index
4. Construct the Laspeyres price index
5. Construct the Paasche price index
6. Use your results from Questions 4 and 5 to construct Fisher's ideal price index
7. Construct the Laspeyres quantity index
8. Construct the Paasche quantity index
9. Use your results from Questions 7 and 8 to construct Fisher's ideal quantity index
10. Construct the value index
11. CPI and average personal income of four areas are in the following table:

Area	CPI	Personal income
A	253	51,481
B	226	49,303
C	231	47,758
D	240	46,594

a. Find the purchasing power in each area
b. Express the purchasing power of B, C, and D as a percentage of that of A, respectively

12. The following table contains GDP deflator and GDP from 2009 to 2013:

Year	GDP deflator	GDP (in billions)
2009	100.16	14.56
2010	101.94	15.23
2011	103.78	15.82
2012	105.67	16.42
2013	107.20	17.09

Compute the real GDP for each year.

Answers to Supplementary Exercises

Multiple Choice

1. c	6. d	11. b	16. c
2. b	7. b	12. a	17. d
3. d	8. d	13. e	18. c
4. e	9. d	14. b	19. e
5. a	10. a	15. e	20. a

True/False

1. True
2. False. Any year can be used as the base year, including the current year
3. True
4. False. A simple aggregate price index will be affected by the units used to state the prices
5. True
6. True
7. False. Paasche index
8. True
9. True
10. True
11. False. Market value weighted index
12. True
13. False. real GDP = nominal GDP / (GDP deflator / 100)

14. False. All firms listed on the NYSE are included in NYSE index
15. False. DJIA considers prices of 30 blue-chips only
16. True.
17. False. The total cost of those foods is 2% less than what the base year would have cost had they been purchased on 2012
18. False. Laspeyres price index
19. False. Paasche price index
20. True

Questions and Problems

1. Price relative $= P_{ti} / P_{0i}$
   Beef $= 2.75 / 2.50 = 1.100$
   Chicken $= 1.25 / 1.50 = 0.833$
   Pork $= 3.75 / 3.25 = 1.154.$

2. $$I_t = \frac{\sum p_{ti}}{\sum p_{0i}} \times 100$$
$$= \frac{2.75 + 1.25 + 3.75}{2.50 + 1.50 + 3.25} \times 100 = 106.9.$$

3. $$I_t = \frac{\sum p_{ti} / p_{0i}}{n} \times 100$$
$$= \frac{1.10 + 0.833 + 1.154}{3} \times 100 = 102.9.$$

4. $$I_t = \frac{\sum_{i=1}^{3} P_{ti} Q_{0i}}{\sum_{i=1}^{3} P_{0i} Q_{0i}} \times 100$$
$$= \frac{2.75(50) + 1.25(70) + 3.75(35)}{2.50(50) + 1.50(70) + 3.25(35)} \times 100 = 103.64.$$

5. $$I_t = \frac{\sum_{i=1}^{3} P_{ti} Q_{ti}}{\sum_{i=1}^{3} P_{0i} Q_{ti}} \times 100$$
$$= \frac{2.75(40) + 1.25(90) + 3.75(35)}{2.50(40) + 1.50(90) + 3.25(35)} \times 100 = 101.43$$

6. $$FI_t = \sqrt{\frac{\sum_{i=1}^{n} P_{ti}Q_{0i}}{\sum_{i=1}^{n} P_{0i}Q_{0i}} \times \frac{\sum_{i=1}^{n} P_{ti}Q_{ti}}{\sum_{i=1}^{n} P_{0i}Q_{ti}}} \times 100 = \sqrt{(1.0364)(1.0143)} \times 100 = 102.53.$$

7. $$I_t = \frac{\sum_{i=1}^{3} Q_{ti}P_{0i}}{\sum_{i=1}^{3} Q_{0i}P_{0i}} \times 100$$
$$= \frac{40(2.50) + 90(1.50) + 35(3.25)}{50(2.50) + 70(1.50) + 35(3.25)} \times 100 = 101.45.$$

8. $$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{ti}}{\sum_{i=1}^{n} Q_{0i}P_{ti}} \times 100$$
$$= \frac{40(2.75) + 90(1.25) + 35(3.75)}{50(2.75) + 70(1.25) + 35(3.75)} \times 100 = 99.3.$$

9. $$FI_t = \sqrt{\text{Laspeyres} \times \text{Paasche}}$$
$$= \sqrt{(101.45)(99.3)} = 100.37.$$

10. $$I_t = \frac{\sum_{i=1}^{n} Q_{ti}P_{ti}}{\sum_{i=1}^{n} Q_{0i}P_{0i}} \times 100$$
$$= \frac{40(2.75) + 90(1.25) + 35(3.75)}{50(2.50) + 70(1.50) + 35(3.25)} \times 100 = 102.91.$$

11. a. Purchasing power = personal income/CPI

A: $51481 / 253 = 203.48$
B: $49303 / 226 = 218.15$
C: $47758 / 231 = 206.74$
D: $46594 / 240 = 194.54.$

b.

B: 100(218.15 / 203.48)% = 107.2%
C: 100(206.74 / 203.48)% = 101.6%
D: 100(194.54 / 203.48)% = 95.4%.

12. Real GDP = 100[GDP / (GDP deflator)]

Year	GDP deflator	GDP (in billions)	Real GDP (in billions)
2009	100.16	14.56	14.54
2010	101.94	15.23	14.94
2011	103.78	15.82	15.24
2012	105.67	16.42	15.54
2013	107.2	17.09	15.94

Chapter 20
Sampling Surveys: Methods and Applications

Chapter Intuition

Previously, we learned about the basic methods used in sampling. In many cases, simple random sampling may not represent the best method of selecting members from a population. This chapter continues to discuss sampling by presenting methods that may be useful in designing a sampling survey. For example, when we are conducting a survey over a very large geographic region, it may be very expensive to conduct the survey. One way to deal with this problem might be to conduct a census over several smaller geographic areas. This approach is known as ***cluster sampling***. As another example, suppose we are interested in the earnings of people at a large company such as IBM. If we believe that the earnings of men and women in the company may be substantially different, we may want to ensure that both men and women are fairly represented in our sample. Because simple random sampling may not guarantee that both sexes are fairly represented, we may choose to use a method known as ***stratified sampling***. In stratified sampling, we divide the total population into different strata; in this case there will be two strata: men and women. A random sample of each group is then taken and analyzed. This approach ensures that each group in our population is fairly represented.

Chapter Review

1. From Chap. 8, we know sampling is preferred to a census for four reasons:
   a. Sampling is more economical than a census.
   b. Sampling is a much quicker way of obtaining data than a census.
   c. The large size of the population of interest may make a census infeasible.
   d. A census may be inappropriate for things like quality control.

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_20

2. The easiest approach to obtain a sample is known as ***simple random sampling***. Simple random sampling is a technique in which each member of a population has an equal chance of being selected.
3. ***Stratified random sampling*** is a sampling technique in which the population is subdivided into groups known as ***strata***. This approach may be appropriate when the researcher believes that different members of the population will have different views. For example, a pollster may believe that the views on a civil rights bill may differ by racial and ethnic groups. In this case, the researcher may choose to use a stratified sampling technique in which the different strata are based on a person's racial or ethnic background.
4. When we sample, we incur different types of errors. ***Sampling errors*** occur when the difference between the sample and population parameters are due entirely to the items selected in the sample. ***Nonsampling errors*** are not connected to the type of sampling method used. They occur if the researcher chooses to sample the wrong population or if the responses are biased due to improperly worded questionnaires.
5. ***Cluster sampling*** is used when a researcher is interested in surveying a population which is spread over a large geographic region. In this case, the researcher may choose to divide the population into geographically compact clusters and then sample or take a census of a selected number of the clusters.
6. In ***two-phase sampling***, the researcher conducts a small pilot study in phase I before he or she attempts the large-scale study. The pilot study can be used to evaluate the questionnaire used and the number of nonrespondents.
7. As the purpose of sampling is to make inferences about a population, we sometimes use ratios or regression analysis to improve our estimates.
8. Sometimes a technique known as the ***jackknife method*** is used in conjunction with sampling to remove the bias of an estimator and to produce confidence intervals.
9. An important issue in sampling surveys is ***sample size determination*** to have the desired ***precision***. Formulae to determine sample size for simple random sampling and stratified random sampling are introduced.

Useful Formulas

Simple random sampling:

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample variance: $s^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$

Estimated variance for the sample mean: $\hat{\sigma}^2_{\bar{x}} = \frac{s^2}{n} \times \frac{N-n}{N-1}$

Confidence interval for population mean: $\bar{x} - z_{\alpha/2}\hat{\sigma}_{\bar{x}} < \mu < \bar{x} + z_{\alpha/2}\hat{\sigma}_{\bar{x}}$

Sample size: $n = \dfrac{N\sigma^2}{(N-1)\sigma_{\bar{X}}^2 + \sigma^2}$ if σ^2 is known,

or $n = \dfrac{NS^2}{(N-1)\hat{\sigma}_{\bar{X}}^2 + S^2}$ ifσ^2 is unknown.

Simple random sampling for proportions:

Sample proportion: $\hat{p} = \dfrac{t}{n}$,
where t=number of sample observations with the property

Estimated variance of the sample proportion: $\hat{\sigma}_{\hat{p}}^2 = \dfrac{\hat{p}(1-\hat{p})}{n} \times \dfrac{N-n}{N-1}$

Confidence interval for the population proportion: $\hat{p} - z_{\alpha/2}\hat{\sigma}_{\hat{p}} < \mu < \hat{p} + z_{\alpha/2}\hat{\sigma}_{\hat{p}}$

Sample size: $n = \dfrac{N\hat{p}(1-\hat{p})}{(N-1)\sigma_{\hat{p}}^2 + \hat{p}(1-\hat{p})}$

or $n = \dfrac{Nz_{\alpha/2}^2\hat{p}(1-\hat{p})}{(N-1)d^2 + z_{\alpha/2}^2\hat{p}(1-\hat{p})}$

or $n' = \dfrac{z_{\alpha/2}^2\hat{p}(1-\hat{p})}{d^2}$ when N is large.

Stratified random sampling:

Sample mean for jth stratum: $\bar{x}_j = \dfrac{1}{n_j}\sum_{i=1}^{n_j} X_{ij}$

Estimated population mean: $\bar{x}_{st} = \sum_{j=1}^{H} W_j\bar{x}_j = \dfrac{1}{N}\sum_{j=1}^{H} N_j\bar{x}_j$

Sample variance for jth stratum: $s_j^2 = \dfrac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(X_{ij} - \bar{X}_j\right)^2$

Estimated variance for the sample mean $\bar{x}_j$: $\hat{\sigma}_{\bar{X}_j}^2 = \dfrac{s_j^2}{n_j} \times \dfrac{N_j - n_j}{N_j}$

Estimated variance for $\bar{x}_{st}$: $\hat{\sigma}_{\bar{x}_{st}}^2 = \sum_{j=1}^{H} W_j^2\hat{\sigma}_{\bar{x}_j}^2$

Confidence interval for population mean: $\bar{x}_{st} - z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}} < \mu < \bar{x}_{st} + z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}}$

Sample size of proportional sampling: $n = \dfrac{\sum_{j=1}^{H} N_j s_j^2}{N\sigma_{\bar{x}_{st}}^2 + \dfrac{1}{N}\sum_{j=1}^{H} N_j s_j^2}$

Sample size of optimal allocation: $n = \dfrac{\frac{1}{N}\left(\sum_{j=1}^{H} N_j s_j\right)^2}{N\sigma_{\bar{x}_{st}}^2 + \frac{1}{N}\sum_{j=1}^{H} N_j s_j^2}$

Optimal proportion for jth stratum: $n_j = \dfrac{N_j s_j}{\sum_{i=1}^{H} N_i s_i} \times n$

Cluster sampling when $n_j = N_j$ for all sampled clusters:

Estimated population mean:

$$\hat{\mu} = \frac{1}{\sum_{j=1}^{m} n_j}\left(\sum_{j=1}^{m} n_j \bar{x}_j\right), \text{ if N is unknown,}$$

$$= \frac{1}{\bar{N}m}\left(\sum_{j=1}^{m} n_j \bar{x}_j\right), \text{ if N is known.}$$

Estimated variance for $\hat{\mu}$: $\hat{\sigma}_{\hat{\mu}}^2 = \dfrac{(M-m)}{MmN^2}\dfrac{\sum_{j=1}^{m} n_j^2(\bar{x}_j - \hat{\mu})^2}{m-1}$

Confidence interval: $\hat{\mu} - z_{\alpha/2}\hat{\sigma}_{\mu} < \mu < \mu + z_{\alpha/2}\hat{\sigma}_{\hat{\mu}}$

Ratio method:

Estimated population total of X: $\hat{X}_r = \dfrac{x}{y}Y = \dfrac{\bar{x}}{\bar{y}}Y$

Regression method:

Estimated population mean: $\hat{\mu}_x = \bar{x} + b(\mu_y - \bar{y})$

Example 1 Confidence Interval for the Population Mean

The citizens for fair taxes of Rich City are interested in the average property tax paid by their 2000 residents. A random sample of 25 of these households had a mean property tax of $ 3222 with a standard deviation of $ 811.

a. Find an estimate of the variance of the sample mean.
b. Find a 90 % confidence interval for the population.

Example Problems

Solution:

a. $\hat{\sigma}^2_{\bar{x}} = \frac{s^2}{n} \times \frac{N-n}{N-1} = \frac{811^2}{25} \times \frac{2000-25}{1999} = 25,992.98$

b. $\bar{x} \pm z_{.10/2}\hat{\sigma}_{\bar{x}}$

$$3222 \pm 1.645\sqrt{25,992.98}$$

2956.79 to 2487.21

Example 2 Confidence Interval for the Population Mean

A fitness expert is interested in the mean number of miles marathoners run per week. Of the 100 members of the Crazy Legs Running Club, 40 were randomly sampled and found to run an average of 82.5 miles/week with a standard deviation of 12.1 miles. Find a 95 % confidence interval for the mean number of miles that the runners run each week.

Solution:

$$\hat{\sigma}^2_{\bar{x}} = \frac{s^2}{n} \times \frac{N-n}{N-1} = \frac{12.1^2}{40} \times \frac{100-40}{100-1} = 2.22$$

$$\bar{x} \pm z_{0.05/2}\hat{\sigma}_{\bar{x}} = 82.5 \pm 1.96\sqrt{2.22} = (79.58, 85.42)$$

Example 3 Determining the Sample Size

Suppose the quality control expert knows from past experience that the number of dud bullets in a case of 20,000 has a population standard deviation of 401. If she would like to compute a 95 % confidence interval for the population mean with a standard deviation of 30, how many bullets should she sample?

Solution:

The desired precision is to have $1.96\hat{\sigma}_{\bar{X}} = 30$, so $\hat{\sigma}_{\bar{X}} = \frac{30}{1.96} = 15.31$.

$$n = \frac{N\sigma^2}{(N-1)\hat{\sigma}^2_{\bar{X}} + \sigma^2} = \frac{(2000)(401^2)}{(2000-1)(15.31^2) + 401^2} = 510.999$$

So select a sample of 511.

Example 4 Determining the Sample Size

An auditor would like to estimate the total value of a corporation's accounts receivable. From previous years, the auditor has found the population standard deviation to be \$ 722 for the 1200 accounts receivable. If the auditor would like to have a level of precision of \$ 225, how large a sample should he select?

Solution:

$$n = \frac{1200(722^2)}{(1200-1)(225^2)+722^2} = 10.2$$

So select a sample of 11 (round up when selecting the sample).

Example 5 Confidence Interval for Population Proportion

Suppose the Fly Straight Golf Club Company sends its new golf clubs to 200 professionals. Of the 200 professionals receiving the clubs, 40 are randomly selected and asked if they hit the ball straighter with the new clubs. Twenty-two of the professionals said they did hit the ball straighter. Construct a 95% confidence interval for the population proportion.

Solution:

$$\hat{p} = \frac{t}{n} = \frac{22}{40} = 0.55$$

$$\hat{\sigma}_p^2 = \frac{\hat{p}(1-\hat{p})}{n} \times \frac{N-n}{N-1} = \frac{0.55(1-0.55)}{40} \times \frac{200-40}{200-1} = 0.005$$

$$\hat{p} \pm z_{\alpha/2}\hat{\sigma}_{\hat{p}} = 55 \pm 1.96\sqrt{.005} = (0.411, 0.689)$$

Example 6 Determining the Sample Size for Population Proportion

Use the data from Example 5 to determine the appropriate sample size if we would like to have $d = 0.1$ at 95% confidence.

Solution:

$$\hat{p} = \frac{t}{n} = \frac{22}{40} = 0.55$$

$$n = \frac{Nz^2_{\alpha/2}\hat{p}(1-\hat{p})}{(N-1)d^2 + z^2_{\alpha/2}\hat{p}(1-\hat{p})} = \frac{200(1.96^2)(0.55)(1-0.55)}{(200-1)(1.96^2)(0.55)(1-0.55)} = 64.66$$

So select a sample of 65.

Example 7 Stratified Sampling-Proportional Allocation

Suppose the auditor decides to divide the accounts receivable into stratum. If he would like a level of precision of $ 10, determine the total number of sample observations under a proportional allocation.

Stratum	Population Size	Standard deviation (estimated)
1	300	$ 85
2	375	$ 125
3	275	$ 50
4	250	$ 100

Solution:

$$N = 300 + 375 + 275 + 250 = 1200$$

$$n = \frac{\sum_{j=1}^{H} N_j s_j^2}{N\sigma^2_{\bar{X}} + \frac{1}{N}\sum_{j=1}^{H} N_j s_j^2}$$

$$= \frac{300(85^2) + 375(125^2) + 275(50^2) + 250(100^2)}{1200(10^2) + \frac{1}{1200}[300(85^2) + 375(125^2) + 275(50^2) + 250(100^2)]} = 86.70$$

So select a sample of 87.

Example 8 Stratified Sampling-Optimal Allocation

Use the data from Example 7 to determine the appropriate sample size under an optimal allocation rule.

Solution:

$$n = \frac{\frac{1}{N}\left(\sum_{j=1}^{H} N_j s_j^2\right)^2}{N\sigma_{\bar{x}_{st}}^2 + \frac{1}{N}\sum_{j=1}^{H} N_j s_j^2}$$

$$= \frac{\frac{1}{1200}[300(85) + 375(125) + 275(50) + 250(100)]^2}{1200(10^2) + \frac{1}{1200}[300(85^2) + 375(125^2) + 275(50^2) + 250(100^2)]} = 79.56$$

So select a sample of 80. Notice that selecting the optimal sample reduces the size of the sample over a proportional allocation.

Supplementary Exercises

Multiple Choice

1. In simple random sampling:
   a. all members of the population are examined
   b. the population is divided into several groups and a census of all groups is taken
   c. the population is divided into several groups and a census of some of the groups is taken
   d. all members of the population have an equal chance of being selected
   e. the population mean always equals the sample mean.

2. In cluster sampling:
   a. all members of the population are examined
   b. the population is divided into several groups, and samples are selected from each group

c. the population is divided into several groups and a census of some of the groups is taken
d. all members of the population have an equal chance of being selected
e. the population mean always equals the sample mean.

3. In stratified random sampling:
a. all members of the population are examined
b. the population is divided into several groups, and samples are selected from each group
c. the population is divided into several groups and a sample of some of the groups is taken
d. all members of the population have an equal chance of being selected
e. the population mean always equals the sample mean.

4. Sampling error is
a. unimportant
b. the difference between the sample and population parameters due entirely to the sample
c. due to response error
d. due to nonresponses
e. due to measurement error.

5. Nonsampling error
a. can result from misinformation provided by respondents
b. can result from nonresponses
c. is a type of measurement error
d. can result from choosing an inappropriate population to sample
e. all of the above.

6. Which of the following is *not* a step in survey sampling?
a. determining the relevant information required for the study
b. constructing a population list in terms of the relevant population
c. collecting information on all population members
d. determining the method used to infer population parameters
e. drawing conclusions based on the sample information.

7. If we are surveying a population that is spread over a large geographic region, it is best to use
a. a census
b. simple random sampling
c. stratified random sampling
d. cluster sampling
e. the jackknife method.

8. Suppose we are interested in studying the opinion of voters regarding a national day care policy. If we believe that people in different age groups will have different opinions, it is best to use
a. a census
b. simple random sampling

c. stratified random sampling
d. cluster sampling
e. the jackknife method.

9. An unbiased estimate of the variance for the sample mean from a random sample is
   a. s^2
   b. s^2 / n
   c. $(s^2 / n)[(N-1)/(N-n)]$
   d. $(s^2 / n)[(N-n)/(N-1)]$
   e. $(s^2 / n)[N / n]$.

10. An unbiased estimator for the variance of the sample proportion from a random sample is
   a. $\hat{p}(1-\hat{p})$
   b. $\hat{p}(1-\hat{p})/n$
   c. $[\hat{p}(1-\hat{p})/n][(N-n)/(N-1)]$
   d. $[\hat{p}(1-\hat{p})/n][(N-1)/(N-n)]$
   e. $[\hat{p}(1-\hat{p})/n]\left[N/n\right]$.

11. The unbiased estimator for the population mean in stratified random sampling is
   a. $\bar{x}$
   b. $w_j x_j$
   c. $\Sigma W_j X_j$
   d. $\Sigma W_j \bar{X}_j$
   e. $\Sigma W_j \bar{X}_j / n$

12. Errors such as sample selection bias, response bias, measurement error, and self-selection bias are
   a. sampling errors
   b. standard errors
   c. nonsampling errors
   d. nonstandard errors
   e. either sampling or nonsampling errors.

13. Which of the following statements about sampling errors is correct?
   a. They can be avoided by increasing the sample size.
   b. They can be avoided by using stratified sampling.
   c. They can be avoided by selecting the proper population.
   d. They can be avoided by using proper interviewing techniques.
   e. They cannot be avoided.

14. An unbiased estimate of the population total from a random sample is
   a. $n\bar{x}$
   b. $n\bar{x}_{st}$
   c. $N\bar{x}_{st}$
   d. $N\bar{x}$
   e. $N\bar{x}_{st}\left[(N-n)/(N-1)\right]$.

15. Which of the following is not a reason for sampling being preferred to a census?
    a. Sampling is more economic.
    b. Information needs to be gathered quickly.
    c. Population is very large.
    d. It is sometimes destructive to collect information.
    e. None of the above.

True/False (If False, Explain Why)

1. For a given level of precision in stratified sampling, a proportional allocation will always require a larger sample than an optimal allocation.
2. The jackknife method is a method for removing bias in sampling.
3. The ratio method is used to remove bias in sampling.
4. A pilot study is the first stage in two-stage sampling.
5. Stratified random sampling is used when the population consists of several groups that are expected to have similar means.
6. Cluster sampling can reduce the costs of surveying a population that is spread over a large geographic region.
7. A random number table can be used to select a random sample.
8. A confidence interval can be constructed for the population mean, but not for the population total.
9. The ratio method can be used to forecast the population total.
10. Other things being equal, the smaller the level of precision desired, the smaller the sample size that needs to be collected.
11. A census will always be more accurate than a sample.
12. For a given population size and level of precision, the largest sample will need to be taken at a proportion of .5.
13. Whether the auxiliary variable is correlated to the variable of interest or not, the ratio method can always improve the precision of parameter estimates.
14. The ratio method and the regression method are identical when the regression line passes through the origin.
15. The required sample size for stratified random sampling depends only on the variance and the size of each stratum and the desired level of precision.

Questions and Problems

1. The students for Affordable Education at Bargain College are interested in the average amount of money spent on textbooks by the college's 1500 students. A random sample of 40 of these students had a mean of $ 375 with a standard deviation of $ 50. Construct a 95 % confidence interval for the population mean.

2. Suppose a quality control expert knows from past experience that the number of defective widgets in a case of 50,000 has a population standard deviation of 725. If he would like to compute a 95% confidence interval for the population mean with a standard deviation of 50, how many widgets should he sample?
3. Suppose the Score High SAT Review Course is interested in the proportion of students that improve their scores after taking the course. Five hundred students have completed the course. Of the 500 students completing the course, 50 are randomly selected and asked if they improved their scores. Thirty-five of the students said they improved. Construct a 95% confidence interval for the population proportion.
4. The leader of People for Lower Taxes is interested in the proportion of citizens of Mayberry that favor the governor's tax reform policy. If there are 1000 adults in Mayberry and the leader would like a 95% confidence interval to extend 5% on each side of the mean proportion, how many residents should be surveyed?
5. The results from a stratified sample are:

Stratum (h)	$\bar{x}_h$	S_h	n_h	N_h
1	9	20	30	120
2	8	30	40	200
3	10	10	50	200

   a. Estimate the population mean.
   b. Estimate the variance for $\bar{x}_{st}$.
   c. Find a 95% confidence interval for the population mean.

6. For the data in Problem 5, suppose that we would like a level of precision of 1.2.

   a. Determine the sample size under a proportional allocation.
   b. Decide the sample size under an optimal allocation rule.

7. A sample of five clusters is to be taken from a population with $M = 20$ clusters and $N = 400$ elements in the population. The following table contains N_i and x_i for each sampled cluster.

Cluster	N_i	x_i
1	25	400
2	10	150
3	20	400
4	15	300
5	30	450
Total	100	1700

   a. Estimate the population mean
   b. Find the estimated variance for the estimator of population mean.

c. Develop a 95 % confidence interval for the population mean.
d. Estimate the population total.

Answers to Supplementary Exercises

Multiple Choice

1. d	6. c	11. d
2. c	7. d	12. c
3. b	8. c	13. e
4. b	9. d	14. d
5. e	10. c	15. e

True/False

1. True
2. True
3. False. Used for forecasting.
4. True
5. False. Used when the means of the groups are expected to be different.
6. True
7. True
8. False. Confidence intervals can be constructed for the population mean and the population total.
9. True
10. False. The smaller the precision, the larger the sample size is required.
11. False. As some respondents may provide inaccurate information, a well designed sample survey may be more accurate than a census.
12. True
13. False. They should be correlated.
14. True.
15. False. It also depends on the way to allocate the total sample among the strata.

Questions and Problems

1. $\hat{\sigma}^2_{\bar{x}} = \frac{s^2}{n} \times \frac{N-n}{N-1} = \frac{50^2}{40} \times \frac{1500-40}{1500-1} = 60.874$

$$375 \pm 1.96\sqrt{60.874} = (359.71, 390.29)$$

2. $1.96\hat{\sigma}_{\bar{X}} = 50,\ \hat{\sigma}_{\bar{X}} = \dfrac{50}{1.96} = 25.51,$

$$n = \frac{N\sigma^2}{(N-1)\hat{\sigma}_{\bar{X}}^2 + \sigma^2} = \frac{50,000)(725^2)}{(50,000-1)(25.51^2) + 725^2} = 794.88$$

So select a sample of 795.

3. $\hat{p} = \dfrac{35}{50} = 0.7$

$$\hat{\sigma}_p^2 = \frac{\hat{p}(1-\hat{p})}{n} \times \frac{N-n}{N-1} = \frac{0.7(1-0.7)}{50} \times \frac{500-50}{500-1} = 0.0038$$

$$\hat{p} \pm z_{\alpha/2}\hat{\sigma}_{\hat{p}} = 0.7 \pm 1.96\sqrt{.0038} = (0.58, 0.82)$$

4. $1.96\hat{\sigma}_{\hat{p}} = 0.05,\ \hat{\sigma}_{\hat{p}} = \dfrac{0.05}{1.96} = 0.0255,$

$$n = \frac{Np(1-p)}{(N-1)\sigma_{\hat{p}}^2 + p(1-p)} = \frac{1000(.5)(1-.5)}{(1000-1)(.0255^2) + .5(1-.5)} = 277.9$$

So select a sample of 278.

5. a. $\bar{x}_{st} = \sum_{j=1}^{H} W_j \bar{x}_j \frac{1}{N} \sum_{j=1}^{H} N_j \bar{x}_j = \frac{1}{520}(120 \times 9 + 200 \times 8 + 200 \times 10) = 9.00$

b. $\hat{\sigma}_{\bar{x}_{st}}^2 = \sum_{j=1}^{H} W_j^2 \hat{\sigma}_{\bar{x}_j}^2$

$$\hat{\sigma}_{\bar{x}_j}^2 = \frac{s_j^2}{n_j} \times \frac{N_j - n_j}{N_j}$$

$$\hat{\sigma}_{\bar{x}_1}^2 = \frac{s_1^2}{n_1} \times \frac{N_1 - n_1}{N_1} = \frac{20^2}{30} \times \frac{120-30}{120} = 10,$$

$$\hat{\sigma}_{\bar{x}_2}^2 = \frac{20^2}{30} \times \frac{120-30}{120} = 18, \hat{\sigma}_{\bar{x}_3}^2 = \frac{10^2}{50} \times \frac{200-50}{200} = 1.5$$

$$\hat{\sigma}_{\bar{x}_{st}}^2 = \frac{1}{520^2}(120^2 \times 10 + 200^2 \times 18 + 200^2 \times 1.5) = \frac{924,000}{270,400} = 3.417$$

c. $\bar{x}_{st} \pm z_{\alpha/2}\hat{\sigma}_{\bar{x}_{st}} = 9.00 \pm 1.96\sqrt{3.417} = (5.377, 12.623)$

6. a. $$n = \frac{\sum_{j=1}^{H} N_j s_j^2}{N\sigma_{\bar{X}}^2 + \frac{1}{N}\sum_{j=1}^{H} N_j s_j^2}$$

$$= \frac{120(20^2) + 200(30^2) + 200(10^2)}{520(1.2^2) + \frac{1}{520}[120(20^2) + 200(30^2) + 200(10^2)]} = 202.3$$

So select a sample of 203.

b. $$n = \frac{\frac{1}{N}\left(\sum_{j=1}^{H} N_j s_j^2\right)^2}{N\sigma_{\bar{x}_{st}}^2 + \frac{1}{N}\sum_{j=1}^{H} N_j s_j^2}$$

$$= \frac{\frac{1}{520}[120(20) + 200(30) + 200(10)]^2}{520(1.2^2) + \frac{1}{520}[120(20^2) + 200(30^2) + 200(10^2)]} = 169.69$$

So select a sample of 170.

7. $m = 5, M = 20, N = 400, \bar{N} = M/N = 20$. Also, $n_j = N_j, j = 1,..,5$.

a. $$\hat{\mu} = \frac{1}{\bar{N}m}\left(\sum_{j=1}^{m} n_j \bar{x}_j^2\right) = \frac{1}{20(5)}\left(\sum_{j=1}^{m} n_j \bar{x}_j^2\right) = \frac{1700}{100} = 17$$

b. $$\hat{V}(\hat{\mu}) = \frac{(M-m)}{Mm\bar{N}^2}\frac{\sum_{j=1}^{m} N_j{}^2(\bar{x}_j - \hat{\mu})^2}{m-1} = \frac{(20-5)}{(20)(5)\bar{N}^2}\frac{\sum_{j=1}^{m}\left(X_j - N_j\hat{\mu}\right)^2}{5-1}$$

Since $\sum_{j=1}^{5}(x_j - N_j\hat{\mu})^2 = 10,250$, and $\bar{N} = N/M = 20$

$$\hat{V}(\hat{\mu}) = \frac{(20-5)}{(20)(5)20^2}\frac{10,250}{4} = 0.961$$

c. $\hat{\sigma}_{\hat{\mu}} = \sqrt{0.961} = 0.98$

$\hat{\mu} \pm z_{.05/2}\hat{\sigma}_{\hat{\mu}} = 17 \pm 1.96(0.98).15.08 \le \mu \le 18.92.$

d. $$\hat{X} = \frac{M}{m}\left(\sum_{j=1}^{5} x_j\right) = \frac{20}{5}(1700) = 6800$$

Chapter 21
Statistical Decision Theory: Methods and Applications

Chapter Intuition

Throughout the book, we have learned about the role of statistics in dealing with an uncertain world. The ultimate goal of using statistics is to allow us to make better decisions when the future is uncertain. For example, suppose you are wrongly accused of committing a crime in a different country. The penalties for the crime are a $ 25 fine if you plead guilty, or 5 years in jail if you go to trial and lose. Although most of us would not like to plead guilty to a crime we did not commit, the uncertainty of the trial system in another country and the severe penalty if we should fight and lose, would lead most of us to pay the fine and leave the country as fast as we could. In this case, we have used a decision-making approach known as the ***minimax criterion***. In the minimax criterion, we assume that whatever action we take, the worst possible outcome will prevail. We then try to minimize the worst thing that can happen to us (maximum penalty). In the crime example, by pleading guilty, we have limited the maximum penalty to a $ 25 fine. As another example, suppose an owner of a small business is trying to decide whether to lease a large or small machine. Although leasing a large machine can generate higher profits when future demand is high, it can lead to larger losses when future demand is low. The owner can address this problem in two ways. First, he can use an approach that ignores the probabilities of future states of the world, such as the minimax and ***maximin*** methods. Second, he can incorporate the probability of future states of the world into decision-making. The ***expected monetary value***, ***utility criterion***, and ***Bayesian method*** are approaches that incorporate probabilities into the decision-making. This chapter deals with different strategies for decision-making under uncertainty.

Chapter Review

1. Some of the decisions faced by a business manager are:

   a. What products should be produced?
   b. What investments should be purchased?
   c. What projects should be undertaken by the firm?

R. L. Moy et al., *Study Guide for Statistics for Business and Financial Economics,*
DOI 10.1007/978-3-319-11997-7_21

2. The four key elements in the decision-making process are:
   a. ***Action***: the choices available. For example, should I purchase insurance or not?
   b. ***State of nature***: the uncertain elements of a decision. For example, will I get into an accident or not?
   c. ***Outcomes***: the consequences of each combination of an action and a state of nature. For example, I buy insurance and I have an accident or I buy insurance and I do not have an accident.
   d. ***Probability***: the chance that a state of nature occurs. For example, there is 1 % chance that I will get into an accident.
3. There are two methods of decision-making that do not use probabilities: The ***maximin*** and the ***minimax regret criterion***. The maximin criterion is a very conservative strategy. It assumes that the worst outcome will occur regardless of what action is taken. Using the maximin criterion, we would choose the action that results in the best outcome, given that the worst state will occur. To use the minimax criterion, we have to generate the ***regret matrix***. We then consider that the worst outcome (maximum regret) will occur for whichever action is taken. The decision then is to choose the action that has the smallest maximized regret.
4. Decision-making can also use probabilities. For each action, the payoffs are different depending on the state of nature that occurs. One way to evaluate the actions is to calculate the expected payoffs of each action. The criterion is called the ***expected monetary value criterion*** (EMV).
5. The probability used to calculate the EMV can be improved by incorporating sample information. The original probability distribution before the revision is called the prior distribution. The revised probability is called the posterior distribution.
6. Very often, it is the satisfaction (in economic terms ***utility***) from the monetary value rather than the monetary value itself that determines how people make decisions. Decision making based on satisfaction or utility is called the ***expected utility criterion***. The expected utility is found by averaging the amount of satisfaction a person receives over the different possible states of nature.
7. One simple approach for making decisions is the use of ***decision trees***. The decision tree is similar to the ***probability tree*** that was previously discussed. The advantage of using a decision tree is that all possible outcomes can be easily seen at once.
8. When an investor's utility function is unknown, the ***mean-variance decision criterion*** assuming utility to be risk-aversion can replace the expected utility rule

Useful Formulas

Expected monetary value criterion

$$\mathrm{EMV}(A_1) = \sum_{j=1}^{M} P_j M_{ij}.$$

Expected utility criterion:

$$E[U(A_i)] = \sum_{j=1}^{M} P_j U_{ij}.$$

Bayes' theorem:

$$P_r(E_2 \mid E_1) = \frac{P_r(E_1 \mid E_2) P_r(E_2)}{P_r(E_1)}.$$

Capital market line:

$$E(R_i) = R_f + [E(R_m) - R_f] \frac{\sigma_i}{\sigma_m}.$$

Capital asset pricing model:

$$E(R_i) = R_f + \beta_i [E(R_m) - R_f].$$

Systematic risk:

$$\beta_i = \frac{\mathrm{Cov}(R_i, R_m)}{\mathrm{Var}(R_m)} = \frac{\rho_{im} \sigma_i \sigma_m}{\sigma_m^2} = \frac{\rho_{im} \sigma_i}{\sigma_m}.$$

Sharpe performance measure:

$$SM_i = \frac{\bar{R}_i - R_f}{\sigma_i}.$$

Treynor performance measure:

$$TM_i = \frac{\bar{R}_i - R_f}{\beta_i}.$$

Example Problems

Example 1 Maximin Criterion

A copy center is considering whether it should lease a large machine or a smaller machine for the next year. The net profits for leasing these two machines are reported in the following table.

	Action	
State of nature	Lease a large machine	Lease a small machine
High demand	30	25
Medium demand	10	10
Low demand	−5	5

Use the maximin method to determine the better action.

Solution:

Step 1 Find the minimum payoff for each action. The minimum payoff occurs when we have low demand for both actions.

Step 2 Select the action that has the maximum payoff from step 1. In this case we choose the small machine because the payoff of 5 is greater than the −5 payoff for the larger machine.

Example 2 Minimax Regret Criterion

Use the data given in Example 1 to find the better action based on the minimax regret criterion.

Solution:

Step 1 Obtain the regret matrix, which is the optimal payoff minus the actual payoff for each state of nature.

	Action	
State of nature	Large machine	Small machine
High demand	0	5
Medium demand	0	0
Low demand	10	0

Step 2 Find the maximum regret (highest values) for each action.

Step 3 Choose the action that gives the smaller maximum regret. In this case, we choose to the small machine.

Example 3 Maximin Strategy

A homeowner is considering whether she should insure her house completely, partially, or not at all. The loss matrix is:

	State of nature	
Action	Fire	No fire
Not insure	100,000	0
Partially insure	51,000	1000
Fully insure	2000	2000

What is the best action under a maximin strategy?

Solution: Notice that in this example, the number is the loss, let payoff=− loss Payoff matrix

	State of nature	State of nature
Action	Fire	No fire
Not insure	–100,000	0
Partially insure	–51,000	–1000
Fully insure	–2000	–2000

Step 1 Find the minimum payoff for each action. The minimum payoff for not insured and partially insured would be the fire state of nature. For the fully insure action, both states would have the same payoff.

Step 2 Choose the action that maximizes the minimum payoff. In this case, the choice is to fully insure.

Example 4 Minimax Strategy

Use the information given in Example 3 to find the best action using a minimax criterion.

Solution:

Step 1 Construct the regret matrix.

98,000	0
49,000	1000
0	2000

Step 2 Maximize the regret for each action.

Step 3 Choose the action that minimizes the maximum regret. In this case, the choice is to fully insure.

Example 5 Expected Monetary Values

A company is considering the following three options for an assembly line: (i) do nothing, (ii) upgrade the old assembly line, and (iii) install a new one. The payoff matrix under different demand situations is presented in the following table:

State of nature	Do nothing	Upgrade	Install new	P_r
Low demand	10	5	0	0.3
Moderate demand	10	15	15	0.3
High demand	10	25	30	0.4

Compute the expected monetary value of each action and determine the best action.

Solution:

$$\text{E(do nothing)} = 10(.3) + 10(.3) + 10(.4) = 10.$$
$$\text{E(upgrade)} = 5(.3) + 15(.3) + 25(.4) = 16.$$
$$\text{E(install new)} = 0(.3) + 15(.3) + 30(.4) = 16.5.$$

The best action is to install the new machine because it has the highest EMV.

Example 6 Expected Monetary Values

An ice cream vendor at the local beach knows from past experience that the sales of ice cream on a Sunday depend on the weather. He also knows from past experience that the probability of sunny weather is 0.8 and the probability of rainy weather is 0.2. The net profits from ordering 50 and 100 pounds of ice cream are presented in the following table.

Weather	(A1) 50 pounds	(A2) 100 pounds	Probability
(S1) Sunny	100	150	0.8
(S2) Rainy	80	50	0.2

Compute the expected monetary value for each action and make suggestion.

Solution:

$$\text{EMV}(\text{A1}) = 100(.8) + 80(.2) = 96$$

$$\text{EMV}(\text{A2}) = 150(.8) + 50(.2) = 130$$

Since ordering 100 pounds has larger EMV, it is better than ordering 50 pounds.

Example 7 Bayesian Method

Now suppose that the ice cream vendor of Example 6 uses the weather report on Saturday evening. The weather report on Saturday is for rain on Sunday. The weatherman's track record is summarized below.

$$P(I1 \mid S1) = 0.8 \qquad P(I2 \mid S1) = 0.2$$

$$P(I1 \mid S2) = 0.3 \qquad P(I2 \mid S2) = 0.7,$$

where I1 indicates a sunny report and I2 indicates a rainy report.

a. Use this information to obtain the posterior probability.
b. What is the expected monetary value using the posterior distribution?
c. What is the expected value for the procedure considering weather report?

Solution:

a. $$P(S1 \mid I1) = \frac{P(S1,I1)}{P(I1)} = \frac{P(I1 \mid S1)P(S1)}{P(I1 \mid S1)P(S1) + P(I1 \mid S2)P(S2)} = \frac{.8(.8)}{.8(.8) + .3(.2)} = .914.$$

$$P\ (S2 \mid I1) = 1 - \ P(S1 \mid I1)\ = 0.086.$$

$$P(S1 \mid I2) = \frac{P(S1,I2)}{P(I2)} = \frac{P(I2 \mid S1)P(S1)}{P(I2 \mid S1)P(S1) + P(I2 \mid S2)P(S2)} = \frac{.2(.8)}{.2(.8) + .7(.2)} = 0.533.$$

$$P\ (S2 \mid I2) = 1 - P(S2 \mid I2) = 0.467.$$

b. The expected monetary value given I1 is

$$E(A1 \mid I1) = 100(.914) + 80(.086) = 98.28$$
$$E(A2 \mid I1) = 150(.914) + 50(.086) = 141.4.$$

Given I1, A2 has larger EMV.
The expected monetary value given I2 is

$$E(A1 \mid I2) = 100(.533) + 80(.467) = 90.69$$
$$E(A2 \mid I2) = 150(.533) + 50(.467) = 103.3$$

Given I2, A2 has larger EMV.

c. Since $P(I1) = 0.7$, $P(I2) = 0.3$, the expected value for the procedure considering weather report is $EV = 141.4(0.7) + 103.3(0.3) = 129.88$.

Example 8 Utility Criterion

A farmer harvests 1000 bushels of wheat every year. The total revenue from wheat is 1000P, where P is the price of wheat at the time of the harvest. His utility function is

$$U = (1,000P)^{1/2}.$$

Suppose the farmer knows that there is a 50% probability that the price will be 100 and a 50% probability that the price will be 80. The farmer decides to sell his harvest using the future market and is guaranteed a price of 90.

Is this behavior consistent with his utility function? Is the farmer risk adverse, risk neutral, or a risk taker?

Solution:
The utility of selling in the future market is

$$U(90) = [1,000P]^{1/2} = [1,000(90)]^{1/2} = 300.$$

The expected utility of not selling in the future market is

$$[1,000(100)]^{1/2} \times .50 + [1,000(80)]^{1/2} \times .50 = 299.54.$$

As $300 > 299.54$, his choice is consistent with his utility function. As

$$E(U) < U(90), \text{he is risk averse.}$$

Example 9 Expected Utility

Consider the following three projects.

Project A		Project B		Project C	
Payoff	P_r	Payoff	P_r	Payoff	P_r
5	1/3	4	1/3	5	1/4
10	1/3	10	1/3	10	1/2
15	1/3	16	1/3	15	1/4

a. Compute the expected monetary value of each project.
b. Which is preferred if $U(W) = W - W^2/20$?

Solution:

a. Expected monetary value for project A is

$$\begin{aligned} \text{EMV(A)} &= 5(1/3) + 10(1/3) + 15(1/3) = 10, \\ \text{EMV(B)} &= 4(1/3) + 10(1/3) + 16(1/3) = 10, \\ \text{EMV(C)} &= 5(1/4) + 10(1/2) + 15(1/4) = 10. \end{aligned}$$

b. Use the expected utility to solve this problem.

The expected utility for project A is

$$\text{E[U(W)]} = \sum \text{Pr}_i \, \text{U}_\text{i} = \frac{1}{3}(5 - 5^2/20) + \frac{1}{3}(10 - 10^2/20) + \frac{1}{3}(15 - 15^2/20) = 4.17.$$

The expected utility for project B is

$$\text{E[U(W)]} = \sum \text{Pr}_i \, \text{U}_\text{i} = \frac{1}{3}(4 - 4^2/20) + \frac{1}{3}(10 - 10^2/20) + \frac{1}{3}(16 - 16^2/20) = 3.8.$$

The expected utility for project C is

$$\text{E[U(W)]} = \sum \text{Pr}_i \, \text{U}_\text{i} = \frac{1}{4}(5 - 5^2/20) + \frac{1}{2}(10 - 10^2/20) + \frac{1}{4}(15 - 15^2/20) = 4.375.$$

Choose project C, since it has the largest expected utility.

Even though the expected monetary values are the same for each project, the expected utilities are different.

Supplementary Exercises

Multiple Choices

1. The minimax strategy

   a. Uses probabilities
   b. Uses Bayes' theorem
   c. Maximizes the worst thing that can happen
   d. Minimizes the maximum regret
   e. Is another name for a decision tree

2. The EMV criterion

   a. Maximizes the satisfaction of the decision maker
   b. Uses Bayes' theorem
   c. Maximizes the worst thing that can happen
   d. Minimizes the maximum regret
   e. Looks at the expected value of the payoffs

3. Which of the following is *not* an element of the decision-making process?

   a. Action
   b. State of nature
   c. Outcome
   d. Probability
   e. CAPM

4. The expected utility criterion is computed by

   a. $\sum Pr_i$
   b. $\sum U_i$
   c. $\sum Pr_i U_i$
   d. $\sum Pr_i / U_i$
   e. $\sum U_i / Pr_i$

5. The capital asset pricing model is a decision-making model in finance that shows

   a. The tradeoff between risk and expected return for combinations of a riskless asset and the market portfolio.
   b. The relationship between the expected return on a stock and its total risk.
   c. The expected price of an asset based on its total risk.
   d. The relationship between the expected return on an asset and its nonsystematic risk.
   e. The relationship between the expected price of an asset and its nonsystematic risk.

6. Utility refers to

   a. The amount of satisfaction a person receives
   b. The amount of money a person receives
   c. The expected amount of money a person receives
   d. The minimax strategy
   e. The maximin strategy

7. Which of the following statements is true?

   a. The posterior distribution contains less information than the prior distribution.
   b. The posterior distribution is derived from the prior distribution and the sample information.
   c. The prior information is derived from the posterior distribution and the sample information.
   d. The sample information is derived from the prior distribution.
   e. The sample information is derived from the posterior distribution.

8. In case of having two actions and two states, if $P_r(S_1) = P_r(S_2)$, then

   a. $P_r(S_1 \mid I_1) = P_r(S_1 \mid I_2)$
   b. $P_r(S_1 \mid I_2) = P_r(S_1 \mid I_1)$
   c. $P_r(S_2 \mid I_1) = P_r(S_1 \mid I_2)$
   d. $P_r(S_1 \mid I_2) = P_r(S_2 \mid I_2)$
   e. The posterior distribution equals sample information

9. A risk neutral person

   a. Never takes risk
   b. Takes risk with constant increase of utility
   c. Takes risk with increasing increase of utility
   d. Takes risk with decreasing increase of utility
   e. Takes risk for no good reason

10. A risk lover

    a. Never takes risk
    b. Takes risk with constant increase of utility
    c. Takes risk with increasing increase of utility
    d. Takes risk with decreasing increase of utility
    e. Takes risk for no good reason

11. Which of the followings best describes risk averting behavior?

    a. If two actions have the same EMV, select the one with the smaller standard deviation.
    b. If two actions have the same standard deviation, select the one with the smaller EMV.

c. The utility increases at an increasing rate as payoff increases.
d. The utility increases at a constant rate as payoff increases.
e. None is correct.

12. Which of the followings is risk seeking behavior?

a. If two actions have the same EMV, select the one with the smaller standard deviation.
b. If two actions have the same standard deviation, select the one with the smaller EMV.
c. The utility increases at an increasing rate as payoff increases.
d. The utility increases at a constant rate as payoff increases.
e. None is correct.

13. Which criterion decides the best actions from a table containing the difference between a strategy's payoff and the best strategy's payoff for each state of nature and actions?

a. Maximin criterion
b. Minimax regret criterion
c. Expected monetary value criterion
d. Mean-variance rule
e. Bayes' strategy

14. Which of the following approaches is used to analyze investment opportunities involving a sequence of investment decisions over time?

a. Time series analysis
b. Maximin criterion
c. Minimax regret criterion
d. Decision tree
e. Bayes' strategy

15. Which measure(s) an investor's risk according to the mean-variance rule?

a. The expected return
b. The variance of return
c. The standard deviation of return
d. All of the above are correct
e. b and c

16. Which measure(s) an investor's profitability according to the mean-variance rule?

a. The expected return
b. The variance of return
c. The standard deviation of return
d. Market rate of return
e. Variance of market rate of return

17. In stock investment analysis, what is the systematic risk?

    a. It results from the basic variability of stock prices.
    a. It accounts for the tendency of stock prices to move together with the general market.
    b. It reflects the fluctuations and changes on general market condition.
    c. All of the above are correct.
    d. It is the result of variations peculiar to the firm or industry.

18. The capital market line used to describe the trade-off between expected return and total risk is

    a. $E(R_i) = R_f + [E(R_m) - R_f]$

    b. $E(R_i) = R_f + [E(R_m) - R_f]\sigma_i$

    c. $E(R_i) = R_f + [E(R_m) - R_f]\sigma_i / \sigma_m$

    d. $E(R_i) = R_f + [R_f - E(R_m)]\sigma_i / \sigma_m$

    e. $E(R_i) = R_f + [E(R_m) - R_f]\sigma_m / \sigma_i$

True/False (If False, Explain Why)

1. The EMV criterion for decision making does not use probability.
2. The EMV criterion looks at the expected value of the money received to determine the best action to take.
3. The minimax regret strategy is a conservative approach to decision making.
4. Decision trees can be used in decision making.
5. The capital market line is used to show the trade-off between risk and return for combinations of the market portfolio and the risk free asset.
6. The maximin and minimax approaches to decision making both use probabilities.
7. The utility approach to decision making looks at the expected value of the money received in order to determine the best action to take.
8. A risk averter will never take risk.
9. For a risk averter, the expected utility of wealth is lower than the utility of expected wealth.
10. The maximin strategy is an aggressive strategy.
11. The Bayesian method of decision making uses sample information to revise the prior distribution into posterior distribution.
12. The expected monetary value approach is equivalent to the expected utility approach, when an individual is risk neutral.
13. A risk averter prefers lower risk to higher risk given the same level of expected value.

14. A risk lover prefers higher risk (standard deviation) than lower risk given the same level of expected value.
15. A risk lover prefers to deposit his or her money with a low but guaranteed interest rate, rather than buy a stock expected to have high returns and some chance of losing value.
16. A mutual fund with smaller a Sharpe performance measure is considered to be better than the one with higher a Sharpe performance measure.
17. Assume the expected return on the market is 10%, the risk-free rate is 4%, and beta coefficient for AAA is 1.15, then the expected return on AAA's stock is 10%.
18. The capital asset price model (CAPM) expresses the rate of return for a portfolio by the risk-free rate plus the portfolio's risk premium.

Questions and Problems

1. Below is the payoff matrix for a delivery company trying to decide if it should buy a new delivery truck or repair the old truck.

State	New truck	Repair old	Prob
Few miles driven	45	100	1/3
Average miles driven	70	70	1/3
Many miles driven	125	80	1/3

   a. Use the minimax criterion to find a better option.
   b. Use the maximin regret criterion to find a better option.
   c. Compute the expected monetary value for each action and make suggestion.

2. Suppose the owner of the delivery company in Problem 1 has a utility function of $U = W - W^{1/2}$. Find the expected utility for each action and make suggestion.
3. Bob Jones is in the market for a new car. He knows that he will pay one of two possible prices, $ 10,000 or $ 12,000. Suppose Bob's utility function is $U = P^{-0.1}$. If Bob believes there is a 20% probability of getting the $ 10,000 price and an 80% probability of getting the $ 12,000 price, compute Bob's expected utility.
4. Suppose a company is trying to decide whether to replace an old machine with a new machine. As the new machine is more efficient, it will result in higher profits if sales are high. However, if sales are low the cost of the machine will result in losses to the company. Below is a table of the profits for the two possible actions.

	Action	
State of demand	New machine	Old machine
Low	−50	50
Average	125	100
High	175	60

Use the maximin criterion to decide on the better course of action.

5. Reconsider Problem 4. Use the minimax regret criterion to decide on the better course of action.
6. A hot dog bun vendor at the park discovers that the sales on a Sunday depend on the weather. He also knows from past experience that the probability of sunny weather is 0.9 and the probability of rainy weather is 0.1. The net profits from ordering 100 and 200 buns are presented in the following table.

Weather	(A1) 100 buns	(A2) 200 buns	Probability
(S1) Sunny	150	200	0.9
(S2) Rainy	90	60	0.1

Compute the expected monetary value for each action and make suggestion.

7. Suppose that the vendor of Problem 6 uses the weather report on Saturday evening. The weatherman's track record is summarized below.

$$P(I1 \mid S1) = 0.8 \quad P(I2 \mid S1) = 0.2$$
$$P(I1 \mid S2) = 0.25 \quad P(I2 \mid S2) = 0.75.$$

where I1 indicates a sunny report and I2 indicates a rainy report.

   a. Use this information to obtain the posterior probability.
   b. What is the expected monetary value using the posterior probabilities?
   c. What is the expected value for the procedure considering weather report?

8. The following table provides mean return (%) and standard deviation (%) for 3 mutual funds. Compute the Sharpe investment performance measure for each mutual fund at the risk-free interest being 4 and 5%, respectively. Then, suggest the one to invest.

Fund	Mean return (%)	St. Dev. (%)
A	10	5
B	9	4
C	11	7

9. Reconsider Problem 8. If the beta coefficient for A, B, and C are 1.2, 1.1, and 1.5, respectively. Compute the Trenor investment performance measure for each mutual fund at the risk-free interest being 4 and 5 %, respectively. Then, suggest the one to invest.

Answers to Supplementary Exercises

Multiple Choices

1. d	6. a	11. a	16. a
2. e	7. b	12. c	17. d
3. e	8. e	13. b	18. c
4. c	9. b	14. d	
5. d	10. c	15. e	

True/False

1. False. Use probabilities.
2. True.
3. True.
4. True.
5. True.
6. False. Does not use probabilities.
7. False. Uses the expected value of utility.
8. False. A risk averter will take risk as long as the gain from taking risk outweighs the loss from taking the risk.
9. True.
10. False. In doing maximin, we assume that the worst situation will occur in the first step; then we pick the best out of all worst situations. So, it is a very conservative strategy.
11. True.
12. True.
13. True.
14. False. A risk lover has to be compensated with decreasing expected value for an additional unit of risk taken.
15. False. A risk averter.
16. False. Funds with larger Sharpe performance measures are considered to be better.
17. False. $E(R) = 4\% + 1.15(10\% - 4\%) = 10.9\%$.
18. True.

Questions and Problems

1. (a) Step 1: Find the minimum payoff for each action:
   New truck: 45 at few miles driver
   Repair old: 70 for average miles driven
   Step 2: Select the action with the maximum payoff from step 1
   Hence, repair old

(b) Step 1: Obtain the regret matrix, which is the optimal payoff minus the actual payoff for each state of nature

	Action	
State	New truck	Repair old
Few miles driven	55	0
Average miles driven	0	0
Many miles driven	0	45

Step 2: Find the maximum regret for each action.
Step 3: Choose the action that gives the smaller maximum regret. In this case, we choose to "Repair old".

(c) EMV for the new truck is

$$\text{EMV} = 1/3(45) + 1/3(70) + 1/3(125) = 80.$$

EMV for repairing the old truck is

$$\text{EMV} = 1/3(100) + 1/3(70) + 1/3(80) = 83.33.$$

Repair the old truck using the EMV criterion.

2. For the new truck

$$\text{E(U)} = 1/3[45 - 45^{1/2}] + 1/3[70 - 70^{1/2}] + 1/3[125 - 125^{1/2}] = 71.25.$$

For repairing the old truck

$$\text{E(U)} = 1/3[100 - 100^{1/2}] + 1/3[70 - 70^{1/2}] + 1/3[80 - 80^{1/2}] = 74.23.$$

Select "Repair the old truck" based on the expected utility criterion.

3. The expected utility is

$$E(U) = .20(10,000^{-0.1}) + .80(12,000^{-0.1}) = .3923.$$

4. The minimum profit occurs in the low demand state for both courses of action. The minimax criterion will then tell us to select the action that has the maximum profit from the low demand state. In this case, we select keeping the old machine.
5. To use the maximin regret criterion, we need to obtain the regret matrix, which is found by taking the optimal payoff minus the actual payoff in each state.

	Regret	
State of demand	New machine	Old machine
Low	100	0
Average	0	25
High	0	115

The maximum regrets in the regret matrix are for the new machine in the low-demand state and for the old machine in the high-demand state. The maximin regret criterion tells us to select the action that has the minimum maximum regret. In this case, we should choose the new machine.

6. $EMV(A1) = 150(.9) + 90(.1) = 154.$

$$EMV(A2) = 200(.9) + 60(.1) = 240.$$

Since ordering 200 buns has larger EMV, it is a better action.

7. (a)

$$P(S1 \mid I1) = \frac{P(S1,I1)}{P(I1)} = \frac{P(I1 \mid S1)P(S1)}{P(I1 \mid S1)P(S1) + P(I1 \mid S2)P(S2)} = \frac{.8(.9)}{.8(.9) + .25(.1)} = .966$$

$$P(S2 \mid I1) = 1 - P(S1 \mid I1) = 0.034$$

$$P(S1 \mid I2) = \frac{P(S1,I2)}{P(I2)} = \frac{P(I2 \mid S1)P(S1)}{P(I2 \mid S1)P(S1) + P(I2 \mid S2)P(S2)} = \frac{.2(.9)}{.2(.9) + .75(.1)} = 0.545$$

$$P(S2 \mid I2) = 1 - P(S2 \mid I2) = 0.455$$

(b) The expected monetary value given I1 is

$$E(A1 \mid I1) = 150(.966) + 90(.034) = 147.96$$

$$E(A2 \mid I1) = 200(.966) + 60(.034) = 195.24.$$

Given I1, A2 has larger EMV.

The expected monetary value given I2 is

$$E(A1 \mid I2) = 150(.545) + 90(.455) = 122.7$$
$$E(A2 \mid I2) = 200(.545) + 60(.455) = 136.3.$$

Given I2, A2 has larger EMV.

(c) Since $P(I1) = 0.745,\ P(I2) = 0.255$, the expected value for the procedure considering weather report is

$$EV = 195.24\ (0.745) + 136.3(0.255) = 180.21.$$

8. Since $SM_i = \dfrac{\bar{R}_i - R_f}{\sigma_i}$, the SM for each fund at R_f=4 and 5% are computed and listed in the following table.

Fund	$\bar{R}_i$	σ_i	R_f=4%	R_f=5%
A	10	5	1.2	1
B	9	4	1.25	1
C	11	7	1	0.86

If risk-free interest rate is 4%, then mutual fund B is the best.
If risk-free interest rate is 5%, then both A and B are better than C.

9. $TM_i = \dfrac{\bar{R}_i - R_f}{\beta_i}$, the TM for each fund at R_f=4 and 5% are computed and listed in the following table.

Fund	$\bar{R}_i$	β_i	R_f=4%	R_f=5%
A	10	1.2	0.050	0.042
B	9	1.1	0.045	0.036
C	11	1.5	0.047	0.040

Whether the risk-free interest rate is 4 or 5%, mutual fund A is the best.