Topics	Q. No.	Answers
	38.	True
	39.	False
5.10. Exponential Distribution	40.	right
	41.	continuous
	42.	zero
	43.	False
	44.	True

6

CHAPTER

SAMPLING AND SAMPLING DISTRIBUTIONS

CONTENTS

NOTES

LEARNING OBJECTIVES

The two main objectives for Chapter 7 are to give you an appreciation for the proper application of sampling techniques and an understanding of the sampling distributions of two statistics, thereby enabling you to:

- Contrast sampling to census and differentiate among different methods of sampling, which include simple, stratified, systematic, and cluster random sampling; and convenience, judgment, quota, and snowball nonrandom sampling, by assessing the advantages associated with each
- Describe the distribution of a sample's mean using the central limit theorem, correcting for a finite population if necessary
- Describe the distribution of a sample's proportion using the z formula for sample proportions

6.1 INTRODUCTION

This chapter explores the process of sampling and the sampling distributions of some statistics. How do we obtain the data used in statistical analysis? Why do researchers often take a sample rather than conduct a census? What are the differences between random and nonrandom sampling? This chapter addresses these and other questions about sampling.

In addition to sampling theory, the distributions of two statistics: the sample mean and the sample proportion are presented. Knowledge of the uses of the sample mean and sample proportion is important in the study of statistics and is basic to much of statistical analysis.

6.2 SAMPLING

Sampling is widely used in business as a means of gathering useful information about a population. Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process. In the Introductory Caselet, it is mentioned that the Toro Company created a research and development unit in the 1950s to test new product concepts and conduct agronomic research. In studying the effectiveness of its equipment in irrigating, mowing, mulching, snow removing and other tasks in the lawn and garden industry, this R & D unit could sample customer preferences for brand preference, machine features, and equipment usages. Their agronomic researchers might sample plots of land in studying soil management issues, crop production, grass growth rates, fertilizer effectiveness, and other land-use concerns. Often, a sample provides a reasonable means for gathering such useful decision-making information that might be otherwise unattainable and unaffordable.

6.2.1 REASONS FOR SAMPLING

Taking a sample instead of conducting a census offers several advantages.

1. The sample can save money.
2. The sample can save time.

NOTES

3. For given resources, the sample can broaden the scope of the study.
4. Because the research process is sometimes destructive, the sample can save product.
5. If accessing the population is impossible, the sample is the only option.

For a given number of questions from a survey or a given set of measurements obtained in a study, taking a sample versus a census can result in a savings of both money and time. For example, suppose an eight-minute telephone interview is conducted as part of a survey. Conducting such interviews with a sample of 100 customers is substantially less expensive and time-consuming than taking a census of 100,000 customers. If obtaining the outcomes of a study is a matter of urgency, sampling can produce results more quickly. With the volatility of the marketplace and the constant barrage of new competition and new ideas, sampling has a strong advantage over a census in terms of research turnaround time.

If resources allocated to a research project are fixed, more detailed information can be gathered by taking a sample than by conducting a census. With resources concentrated on fewer individuals or items, a study can be broadened in scope to allow more specialized questions and deeper investigation. As an example, one organization budgeted $80,000 to survey the opinions of its customers and opted to take a census instead of a sample by sending a mail survey to the entire population. The researchers mass-mailed thousands of copies of a brief 20-question survey in which each question could be answered with a Yes or No response. One of the questions was, "Are you satisfied with the service that you received at the XYZ store?" For the same amount of money, the company could have taken a random sample from the population, held interactive one-on-one sessions with highly trained interviewers, and gathered detailed information about customer opinions and attitudes towards products, service, layout, availability, etc.

Some research processes are destructive to the product or item being studied. For example, if light bulbs are being tested to determine how long they burn or batteries are being discharged to determine how long they last, the light bulbs and/or the batteries being analyzed for longevity are ruined in the testing process. By using a sample in destructive testing, only a portion of the population is ruined.

Sometimes a population is virtually impossible to access for research. For example, some people refuse to answer sensitive questions, some telephone numbers are unlisted, and some executives are virtually impossible to access. In such cases, sampling is the only option.

6.2.2 REASONS FOR TAKING A CENSUS

Sometimes it is preferable to conduct a census of the entire population rather than taking a sample. There are at least three reasons why a business researcher may opt to take a census rather than a sample, provided there is adequate time and money available to conduct such a census: (1) to eliminate the possibility that by chance a randomly selected sample may not be representative of the population, (2) for the safety of the consumer, and (3) to benchmark data for future studies.

Even when proper sampling techniques are implemented in a study, there is the possibility a sample could be selected by chance that does not represent

the population. For example, if the population of interest is all truck owners in the state of Colorado, a random sample of truck owners could yield mostly ranchers when, in fact, many of the truck owners in Colorado are urban dwellers. If the researcher or study sponsor cannot tolerate such a possibility, then taking a census may be the only option.

Sometimes a census is taken to protect the safety of the consumer. For example, there are some products, such as airplanes or heart defibrillators, in which the performance of such is so critical to the consumer that 100% of the products are tested, and sampling is not a reasonable option. In addition, companies often want to establish performance baselines in areas like cost, time, and quality by taking a census at least one time. On the basis of such a baseline, future managers can compare their results with past baselines to determine how well their processes are performing.

6.2.3 FRAME

Every research study has a target population that consists of the individuals, institutions, or entities that are the object of investigation. When a sample is drawn from a population, it is actually selected from a list, map, directory, or some other source that represents the population. This list, map, or directory is called the frame. Thus, a **frame** is a list, map, directory, or some other source used in the sampling process to represent the population. Because the sample is drawn from the frame, the frame is sometimes referred to as the working population. Examples of frames can include phone directories, trade association lists, company human resource records, or even lists sold by list brokers. Ideally, a one-to-one correspondence exists between the frame units and the target population units. That is, the frame and the target population are congruent. In reality, the frame and the target population are often different, as shown in **Figure 6.1**. In such cases, a frame can be *overregistered* in that it contains units that are not in the target population; and it can be *underregistered* because it does not contain some of the units in the target population.

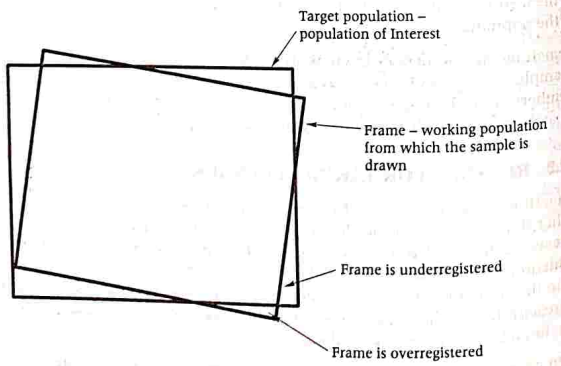


Figure 6.1: The Frame and the Target Population

Suppose the target population is all families living in Detroit. A feasible frame might be the residential pages of the Detroit telephone books. How might the frame differ from the target population? A growing number of families have no "land-line" phone. Other families have unlisted numbers. Still other families might have moved and/or changed numbers since the directory was printed. Some families even have multiple listings under different names.

6.2.4 RANDOM VERSUS NONRANDOM SAMPLING

The two main types of sampling are random and nonrandom. In **random sampling** *every unit of the population has the same probability of being selected into the sample.* Random sampling implies that chance enters into the process of selection. For example, most Americans would like to believe that winners of nationwide magazine sweepstakes or numbers selected as state lottery winners are selected by some random draw of numbers, hence, random sampling.

In **nonrandom sampling** *not every unit of the population has the same probability of being selected into the sample.* Members of nonrandom samples are not selected by chance. For example, they might be selected because they are at the right place at the right time or because they know the people conducting the research.

Sometimes random sampling is called *probability sampling* and nonrandom sampling is called *nonprobability sampling.* Because every unit of the population is not equally likely to be selected, assigning a probability of occurrence in nonrandom sampling is impossible. The statistical methods presented and discussed in this text are based on the assumption that the data come from random samples. *Nonrandom sampling methods are not appropriate techniques for gathering data to be analyzed by most of the statistical methods presented in this text.* However, several nonrandom sampling techniques are described in this section, primarily to alert you to their characteristics and limitations.

6.2.5 RANDOM SAMPLING TECHNIQUES

The four basic random sampling techniques are simple random sampling, stratified random sampling, systematic random sampling, and cluster (or area) random sampling. Each technique offers advantages and disadvantages. Some techniques are simpler to use, some are less costly, and others show greater potential for reducing sampling error.

Simple Random Sampling

The most elementary random sampling technique is **simple random sampling**. Simple random sampling can be viewed as the basis for other random sampling techniques. With simple random sampling, each unit of the frame is numbered from 1 to N (where N is the size of the population). Next, a table of random numbers or a random number generator is used to select n items into the sample. A random number generator is usually a computer program that allows computer-calculated output to yield random numbers. **Table 6.1** contains a brief table of random numbers. Table A.1 in Appendix A contains a full table of random numbers. These numbers are random in all directions. The spaces in the table are there only for ease of reading the values. For each number, any of the 10 digits (0–9) is equally likely, so getting the same digit twice or more in a row is possible.

TABLE 6.1: A BRIEF TABLE OF RANDOM NUMBERS

91567	42595	27958	30134	04024	86385	29880	99730
46503	18584	18845	49618	02304	51038	20655	58727
34914	63974	88720	82765	34476	17032	87589	40836
57491	16703	23167	49323	45021	33132	12544	41035
30405	83946	23792	14422	15059	45799	22716	19792
09983	74353	68668	30429	70735	25499	16631	35006
85900	07119	97336	71048	08178	77233	13916	47564

As an example, from the population frame of companies listed in **Table 6.2**, we will use simple random sampling to select a sample of six companies. First, we number every member of the population. We select as many digits for each unit sampled as there are in the largest number in the population. For example, if a population has 2,000 members, we select four-digit numbers. Because the population in Table 6.2 contains 30 members, only two digits need be selected for each number. The population is numbered from 01 to 30, as shown in **Table 6.3**.

TABLE 6.2: A POPULATION FRAME OF 30 COMPANIES

Alaska Airlines	DuPont	Lubrizol
Alcoa	ExxonMobil	Mattel
Ashland	General Dynamics	Merck
Bank of America	General Electric	Microsoft
Boeing	General Mills	Occidental Petroleum
Chevron	Halliburton	JCPenney
Citigroup	IBM	Procter & Gamble
Clorox	Kellogg's	Ryder
Delta Air Lines	Kroger	Sears
Disney	Lowe's	Time Warner

TABLE 6.3: NUMBERED POPULATION OF 30 COMPANIES

01 Alaska Airlines	11 DuPont	21 Lubrizol
02 Alcoa	12 ExxonMobil	22 Mattel
03 Ashland	13 General Dynamics	23 Merck
04 Bank of America	14 General Electric	24 Microsoft
05 Boeing	15 General Mills	25 Occidental Petroleum
06 Chevron	16 Halliburton	26 JCPenney
07 Citigroup	17 IBM	27 Procter & Gamble
08 Clorox	18 Kellogg	28 Ryder
09 Delta Air Lines	19 Kroger	29 Sears
10 Disney	20 Lowe's	30 Time Warner

The object is to sample six companies, so six different two-digit numbers must be selected from the table of random numbers. Because this population contains only 30 companies, all numbers greater than 30 (31–99) must be ignored. If, for example, the number 67 is selected, we discard the number and continue the process until a value between 01 and 30 is obtained. If the same number occurs more than once, we proceed to another number. For ease of understanding, we start with the first pair of digits in Table 6.1 and proceed across the first row until $n = 6$ different values between 01 and 30 are selected. If additional numbers are needed, we proceed across the second row, and so on. Often a researcher will start at some randomly selected location in the table and proceed in a predetermined direction to select numbers.

In the first row of digits in Table 6.1, the first number is 91. This number is out of range so it is cast out. The next two digits are 56. Next is 74, followed by 25, which is the first usable number. From Table 6.3, we see that 25 is the number associated with Occidental Petroleum, so Occidental Petroleum is the first company selected into the sample. The next number is 95, unusable, followed by 27, which is usable. Twenty-seven is the number for Procter & Gamble, so this company is selected. Continuing the process, we pass over the numbers 95 and 83. The next usable number is 01, which is the value for Alaska Airlines. Thirty-four is next, followed by 04 and 02, both of which are usable. These numbers are associated with Bank of America and Alcoa, respectively. Continuing along the first row, the next usable number is 29, which is associated with Sears. Because this selection is the sixth, the sample is complete. The following companies constitute the final sample.

Alaska Airlines

Alcoa

Bank of America

Occidental Petroleum

Procter & Gamble

Sears

Simple random sampling is easier to perform on small than on large populations. The process of numbering all the members of the population and selecting items is cumbersome for large populations.

Stratified Random Sampling

A second type of random sampling is **stratified random sampling**, in which the population is divided into nonoverlapping subpopulations called strata. The researcher then extracts a random sample from each of the subpopulations (strata). The main reason for using stratified random sampling is that it has the potential for reducing sampling error. Sampling error occurs when, by chance, the sample does not represent the population. With stratified random sampling, the potential to match the sample closely to the population is greater than it is with simple random sampling because portions of the total sample are taken from different population subgroups. However, stratified random sampling is generally more costly than simple random sampling because each unit of the population must be assigned to a stratum before the random selection process begins.

Strata selection is usually based on available information. Such information may have been gleaned from previous censuses or surveys. Stratification benefits increase as the strata differ more. Internally, a stratum should be

relatively homogeneous; externally, strata should contrast with each other. Stratification is often done by using demographic variables, such as sex, socioeconomic class, geographic region, religion, and ethnicity. For example, if a U.S. presidential election poll is to be conducted by a market research firm, what important variables should be stratified? The sex of the respondent might make a difference because a gender gap in voter preference has been noted in past elections; that is, men and women tended to vote differently in national elections. Geographic region also provides an important variable in national elections because voters are influenced by local cultural values that differ from region to region.

In FM radio markets, age of listener is an important determinant of the type of programming used by a station. **Figure 6.2** contains a stratification by age with three strata, based on the assumption that age makes a difference in programming preference. This stratification implies that listeners 20 to 30 years of age tend to prefer the same type of programming, which is different from that preferred by listeners 30 to 40 and 40 to 50 years of age. Within each age subgroup (stratum), *homogeneity* or alikeness is present; between each pair of subgroups a difference, or *heterogeneity*, is present.

Stratified random sampling can be either proportionate or disproportionate. **Proportionate stratified random sampling** occurs *when the percentage of the sample taken from each stratum is proportionate to the percentage that each stratum is within the whole population.* For example, suppose voters are being surveyed in Boston and the sample is being stratified by religion as Catholic, Protestant, Jewish, and others. If Boston's population is 90% Catholic and if a sample of 1,000 voters is taken, the sample would require inclusion of 900 Catholics to achieve proportionate stratification. Any other number of Catholics would be disproportionate stratification. The sample proportion of other religions would also have to follow population percentages. Or consider the city of El Paso, Texas, where the population is approximately 80% Hispanic. If a researcher is conducting a citywide poll in El Paso and if stratification is by ethnicity, a proportionate stratified random sample should contain 80% Hispanics. Hence, an ethnically proportionate stratified sample of 160 residents from El Paso's 660,000 residents should contain approximately 128 Hispanics. *Whenever the proportions of the strata in the sample are*

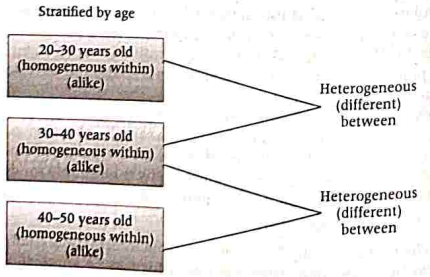


Figure 6.2: Stratified Random Sampling of FM Radio Listeners

different from the proportions of the strata in the population, **disproportionate stratified random sampling** occurs.

6.2.6 SYSTEMATIC SAMPLING

Systematic sampling is a third random sampling technique. Unlike stratified random sampling, systematic sampling is not done in an attempt to reduce sampling error. Rather, systematic sampling is used because of its convenience and relative ease of administration. With **systematic sampling**, *every* k*th item is selected to produce a sample of size* n *from a population of size* N. The value of k, sometimes called the sampling cycle, can be determined by the following formula. If k is not an integer value, the whole-number value should be used.

Determining the value of k (6.1)

$$k = \frac{N}{n}$$

where

n = sample size

N = population size

k = size of interval for selection

As an example of systematic sampling, a business researcher wanted to sample Texas manufacturers as part of a management study. She had enough financial support to sample 1,000 companies. Her frame was the most recent edition of the Texas Manufacturers Register® which listed 26,000 manufacturing companies in alphabetic order. The value of k was 26 (26,000/1,000) and the researcher selected every 26th company in the Texas Manufacturers Register® for her sample.

Did the researcher begin with the first company listed or the 26th or one somewhere between? In selecting every kth value, a simple random number table should be used to select a value between 1 and k inclusive as a starting point. The second element for the sample is the starting point plus k. In the example, $k = 26$, so the researcher would have gone to a table of random numbers to determine a starting point between 1 and 26. Suppose he selected the number 5. He would have started with the 5th company, then selected the 31st (5 + 26), and then the 57th, and so on.

Besides convenience, systematic sampling has other advantages. Because systematic sampling is evenly distributed across the frame, a knowledgeable person can easily determine whether a sampling plan has been followed in a study. However, a problem with systematic sampling can occur if the data are subject to any periodicity, and the sampling interval is in syncopation with it. In such a case, the sampling would be nonrandom. For example, if a list of 150 college students is actually a merged list of five classes with 30 students in each class and if each of the lists of the five classes has been ordered with the names of top students first and bottom students last, then systematic sampling of every 30th student could cause selection of all top students, all bottom students, or all mediocre students; that is, the original list is subject to a cyclical or periodic organization. Systematic sampling methodology is based on the assumption that the source of population elements is random.

6.2.7 CLUSTER (OR AREA) SAMPLING

Cluster (or area) sampling is a fourth type of random sampling. **Cluster (or area) sampling** involves dividing the population into nonoverlapping areas or clusters. However, in contrast to stratified random sampling where strata are homogeneous within, cluster sampling identifies clusters that tend to be internally heterogeneous. In theory, each cluster contains a wide variety of elements, and the cluster is a miniature, or microcosm, of the population. Examples of clusters are towns, companies, homes, colleges, areas of a city, and geographic regions. Often clusters are naturally occurring groups of the population and are already identified, such as states or Standard Metropolitan Statistical Areas.

After randomly selecting clusters from the population, the business researcher either selects all elements of the chosen clusters or randomly selects individual elements into the sample from the clusters. One example of business research that makes use of clustering is test marketing of new products. Often in test marketing, the United States is divided into clusters of test market cities, and individual consumers within the test market cities are surveyed. **Figure 6.3** shows some of the top U.S. cities that are used as clusters to test products.

Sometimes the clusters are too large, and a second set of clusters is taken from each original cluster. This technique is called **two-stage sampling**. For example, a researcher could divide the United States into clusters of cities. She could then divide the cities into clusters of blocks and randomly select individual houses from the block clusters. The first stage is selecting the test cities and the second stage is selecting the blocks.

Cluster or area sampling offers several advantages. Two of the foremost advantages are convenience and cost. Clusters are usually convenient to obtain,

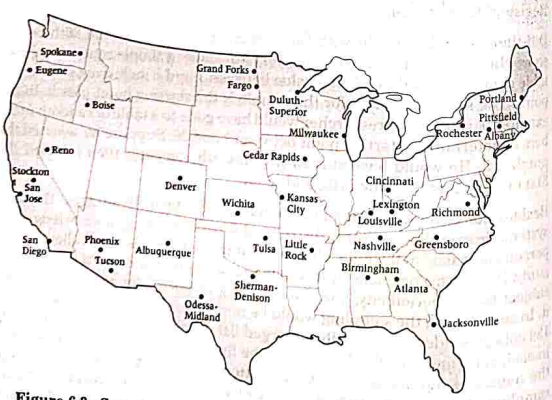


Figure 6.3: Some Top-Rated Test Market Cities in the United States

and the cost of sampling from the entire population is reduced because the scope of the study is reduced to the clusters. The cost per element is usually lower in cluster or area sampling than in stratified sampling because of lower element listing or locating costs. The time and cost of contacting elements of the population can be reduced, especially if travel is involved, because clustering reduces the distance to the sampled elements. In addition, administration of the sample survey can be simplified. Sometimes cluster or area sampling is the only feasible approach because the sampling frames of the individual elements of the population are unavailable, and therefore other random sampling techniques cannot be used.

Cluster or area sampling also has several disadvantages. If the elements of a cluster are similar, cluster sampling may be statistically less efficient than simple random sampling. In an extreme case—when the elements of a cluster are the same—sampling from the cluster may be no better than sampling a single unit from the cluster. Moreover, the costs and problems of statistical analysis are greater with cluster or area sampling than with simple random sampling.

6.2.8 NONRANDOM SAMPLING

Sampling techniques used to select elements from the population by any mechanism that does not involve a random selection process are called **nonrandom sampling techniques**. Because chance is not used to select items from the samples, these techniques are non-probability techniques and are not desirable for use in gathering data to be analyzed by the methods of inferential statistics presented in this text. Sampling error cannot be determined objectively for these sampling techniques. Four nonrandom sampling techniques are presented here: convenenience sampling, judgment sampling, quota sampling, and snowball sampling.

Convenience Sampling

In **convenience sampling**, *elements for the sample are selected for the convenience of the researcher.* The researcher typically chooses elements that are readily available, nearby, or willing to participate. The sample tends to be less variable than the population because in many environments the extreme elements of the population are not readily available. The researcher will select more elements from the middle of the population. For example, a convenience sample of homes for door-to-door interviews might include houses where people are at home, houses with no dogs, houses near the street, first-floor apartments, and houses with friendly people. In contrast, a random sample would require the researcher to gather data only from houses and apartments that have been selected randomly, no matter how inconvenient or unfriendly the location. If a research firm is located in a mall, a convenience sample might be selected by interviewing only shoppers who pass the shop and look friendly.

Judgment Sampling

Judgment sampling occurs when *elements selected for the sample are chosen by the judgment of the researcher.* Researchers often believe they can obtain a representative sample by using sound judgment, which will result in

saving time and money. Sometimes ethical, professional researchers might believe they can select a more representative sample than the random process will provide. They might be right! However, some studies show that random sampling methods outperform judgment sampling in estimating the population mean even when the researcher who is administering the judgment sampling is trying to put together a representative sample. When sampling is done by judgment, calculating the probability that an element is going to be selected into the sample is not possible. The sampling error cannot be determined objectively because probabilities are based on *nonrandom* selection.

Other problems are associated with judgment sampling. The researcher tends to make errors of judgment in one direction. These systematic errors lead to what are called *biases*. The researcher also is unlikely to include extreme elements. Judgment sampling provides no objective method for determining whether one person's judgment is better than another's.

Quota Sampling

A third nonrandom sampling technique is **quota sampling**, which appears to be similar to stratified random sampling. Certain population subclasses, such as age group, gender, or geographic region, are used as strata. However, instead of randomly sampling from each stratum, the researcher uses a nonrandom sampling method to gather data from a stratum until the desired quota of samples is filled. Quotas are described by quota controls, which set the sizes of the samples to be obtained from the subgroups. Generally, a quota is based on the proportions of the subclasses in the population. In this case, the quota concept is similar to that of proportional stratified sampling.

Quotas are often filled by using available, recent, or applicable elements. **Table 6.4** shows how quota sampling might be used to fill quotas of consumers by age.

Note from studying Table 6.4 that the researcher is using strata similar to stratified random sampling. However, the quotas are filled in each case by using convenience sampling, and the result, while appearing to be scientific, is actually nonrandom sampling.

Quota sampling can be useful if no frame is available for the population. For example, suppose a researcher wants to stratify the population into owners of different types of cars but fails to find any lists of Toyota van owners. Through quota sampling, the researcher would proceed by interviewing all car owners and casting out non–Toyota van owners until the quota of Toyota van owners is filled.

TABLE 6.4: USING QUOTA SAMPLING TO FILL QUOTAS OF CONSUMERS BY AGE

Age Category	Quota	How Sample is Obtained
14–18 years old	70	Go to the nearest high school and survey willing students as they leave school until you have surveyed 70 students
25–39 years old	30	Go to junior soccer matches and survey parents in the stands until you have 30 surveys
Over 65 years old	40	Go to the activity center of a retirement community and survey whomever will talk to you

Quota sampling is less expensive than most random sampling techniques because it essentially is a technique of convenience. However, cost may not be meaningful because the quality of nonrandom and random sampling techniques cannot be compared. Another advantage of quota sampling is the speed of data gathering. The researcher does not have to call back or send out a second questionnaire if he does not receive a response; he just moves on to the next element. Also, preparatory work for quota sampling is minimal.

The main problem with quota sampling is that, when all is said and done, it still is only a *nonrandom* sampling technique. Some researchers believe that if the quota is filled by *randomly* selecting elements and discarding those not from a stratum, quota sampling is essentially a version of stratified random sampling. However, most quota sampling is carried out by the researcher going where the quota can be filled quickly. The object is to gain the benefits of stratification without the high field costs of stratification. Ultimately, it remains a nonprobability sampling method.

Snowball Sampling

Another nonrandom sampling technique is **snowball sampling**, in which *survey subjects are selected based on referral from other survey respondents.* The researcher identifies a person who fits the profile of subjects wanted for the study. The researcher then asks this person for the names and locations of others who would also fit the profile of subjects wanted for the study. Through these referrals, survey subjects can be identified cheaply and efficiently, which is particularly useful when survey subjects are difficult to locate. It is the main advantage of snowball sampling; its main disadvantage is that it is nonrandom.

6.2.9 SAMPLING ERROR

Sampling error occurs *when the sample is not representative of the population.* When random sampling techniques are used to select elements for the sample, sampling error occurs by chance. Many times the statistic computed on the sample is not an accurate estimate of the population parameter because the sample was not representative of the population. This result is caused by sampling error. With random samples, sampling error can be computed and analyzed.

6.2.10 NONSAMPLING ERRORS

All errors other than sampling errors are **nonsampling errors**. The many possible nonsampling errors include missing data, recording errors, input processing errors, and analysis errors. Other nonsampling errors result from the measurement instrument, such as errors of unclear definitions, defective questionnaires, and poorly conceived concepts. Improper definition of the frame is a nonsampling error. In many cases, finding a frame that perfectly fits the population is impossible. Insofar as it does not fit, a nonsampling error has been committed.

Response errors are also nonsampling errors. They occur when people do not know, will not say, or overstate. Virtually no statistical method is available to measure or control for nonsampling errors. The statistical

techniques presented in this text are based on the assumption that none of these nonsampling errors were committed. The researcher must eliminate these errors through carefully planning and executing the research study.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. A frame can be _____ in that it contains units that are not in the target population.
2. In _____, every unit of the population has the same probability of being selected into the sample.
3. In one of the random sampling techniques, the population is divided in non-overlapping subpopulations called _____.

State whether the following statements are true/false:

4. Taking a sample versus a census can result in wasting of both money and time.
5. When proper sampling techniques are implemented in a study, there will be no possibility of a sample error.

ACTIVITY

1. You need to prepare a flow chart for random sampling and non-random sampling. Unlike traditional flow chart you need to put examples for all the types of sampling. The examples should be from daily life.
2. You need to prepare a report for favourite television serial's (including news serial) TRP.
   (a) What will be the population frame for this research?
   (b) What type of sampling you can use for it?

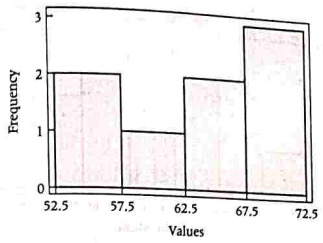
6.3 SAMPLING DISTRIBUTION OF $\bar{x}$

In this section, we explore aspects of the sample mean, $\bar{x}$. The sample mean is one of the more common statistics used in the inferential process. To compute and assign the probability of occurrence of a particular value of a sample mean, the researcher must know the distribution of the sample means. One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.

Suppose a small finite population consists of only $N = 8$ numbers:

54 55 59 63 64 68 69 70

Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.



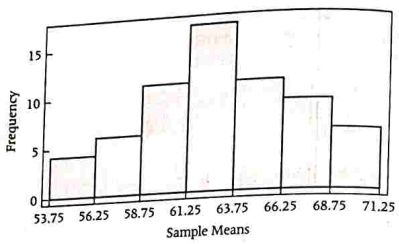
Suppose we take all possible samples of size $n = 2$ from this population with replacement. The result is the following pairs of data.

(54, 54)	(55, 54)	(59, 54)	(63, 54)
(54, 55)	(55, 55)	(59, 55)	(63, 55)
(54, 59)	(55, 59)	(59, 59)	(63, 59)
(54, 63)	(55, 63)	(59, 63)	(63, 63)
(54, 64)	(55, 64)	(59, 64)	(63, 64)
(54, 68)	(55, 68)	(59, 68)	(63, 68)
(54, 69)	(55, 69)	(59, 69)	(63, 69)
(54, 70)	(55, 70)	(59, 70)	(63, 70)
(64, 54)	(68, 54)	(69, 54)	(70, 54)
(64, 55)	(68, 55)	(69, 55)	(70, 55)
(64, 59)	(68, 59)	(69, 59)	(70, 59)
(64, 63)	(68, 63)	(69, 63)	(70, 63)
(64, 64)	(68, 64)	(69, 64)	(70, 64)
(64, 68)	(68, 68)	(69, 68)	(70, 68)
(64, 69)	(68, 69)	(69, 69)	(70, 69)
(64, 70)	(68, 70)	(69, 70)	(70, 70)

The means of each of these samples follow.

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
61	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.



Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population. The sample means appear to "pile up" toward the middle of the distribution and "tail off" toward the extremes.

As another example, consider **Figure 6.4**, which contains an Excel-produced histogram of a Poisson population with a mean (λ) of 1.3 produced from 5,000 data points. Describe the shape of the population. Is it approximately symmetrical or is it skewed to the right or to the left?

Note that this graph is skewed to the right with a mode of 1 with the numbers 0 and 2 also occurring in large numbers.

Now consider **Figure 6.5**, which displays the Excel-produced histogram of sample means computed on 1,000 random samples of size 30 taken from a Poisson population with $\lambda = 1.3$. Is the graph also skewed to the right like the Poisson population from which the samples are drawn? Or is something else happening here?

Observe that even though these 1,000 samples (size $n = 30$) were taken from a skewed Poisson distribution with a λ of 1.3, the distribution of sample means is nearly symmetrical and is approaching the shape of the normal distribution.

Suppose a population is uniformly distributed. If samples are selected randomly from a population with a uniform distribution, how are the sample means distributed? **Figure 6.6** displays the histogram distributions of sample means from five different sample sizes. Each of these histograms represents

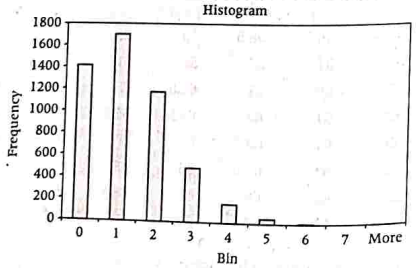


Figure 6.4: An Excel-Produced Histogram of 5,000 Random Numbers Generated from a Poisson Population with $\lambda = 1.3$

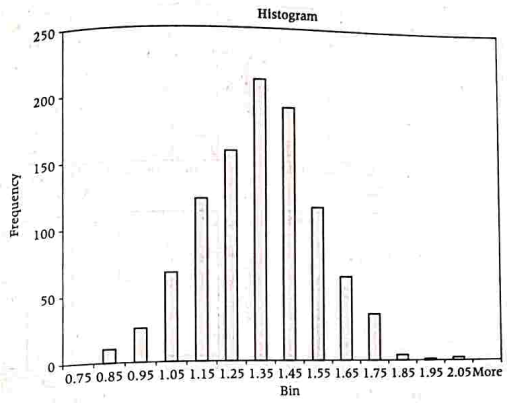


Figure 6.5: An Excel-Produced Histogram of the Sample Means from 1,000 Samples Taken from a Poisson Population with $\lambda = 1.3$

the distribution of sample means from 90 samples generated randomly from a uniform distribution in which $a = 10$ and $b = 30$. Observe the shape of the distributions. Notice that even for small sample sizes, the distributions of sample means for samples taken from the uniformly distributed population begin to "pile up" in the middle. As sample sizes become much larger, the sample mean distributions begin to approach a normal distribution and the variation among the means decreases.

So far, we examined three populations with different distributions. However, the sample means for samples taken from these populations appear to be approximately normally distributed, especially as the sample sizes become larger. What would happen to the distribution of sample means if we studied populations with differently shaped distributions? The answer to that question is given in the **central limit theorem**.

Central Limit Theorem

If random samples of size n are repeatedly drawn from a population that has a mean of μ and a standard deviation of σ, the sample means, $\bar{x}$, are approximately normally distributed for sufficiently large sample sizes ($n \geq 30$) regardless of the shape of the population distribution. If the population is normally distributed, the sample means are normally distributed for any size sample.

From mathematical expectation,* it can be shown that the mean of the sample means is the population mean

$$\mu_{\bar{x}} = \mu$$

*The derivations are beyond the scope of this text and are not shown.

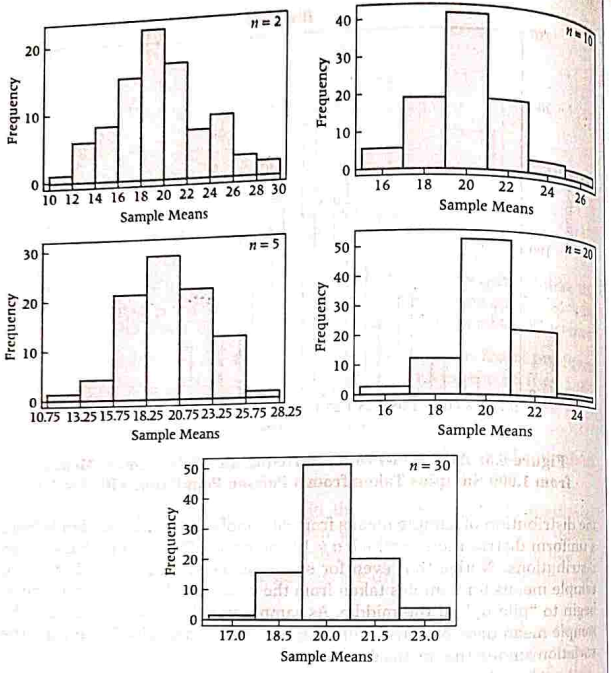


Figure 6.6: Outputs for Sample Means from 90 Samples Ranging in Size from $n = 2$ to $n = 30$ from a Uniformly Distributed Population with $a = 10$ and $b = 30$

and the standard deviation of the sample means (called the standard error of the mean) is the standard deviation of the population divided by the square root of the sample size.

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

The central limit theorem creates the potential for applying the normal distribution to many problems when sample size is sufficiently large. Sample means that have been computed for random samples drawn from normally distributed populations are normally distributed. However, the real advantage of the central limit theorem comes when sample data drawn from populations not normally distributed or from populations of unknown shape also can be analyzed by using the normal distribution because the sample means are normally distributed for sufficiently large sample sizes.* Column 1 of **Figure 6.7** shows four different population distributions. Each

*The actual form of the central limit theorem is a limit function of calculus. As the sample size increases to infinity, the distribution of sample means literally becomes normal in shape.

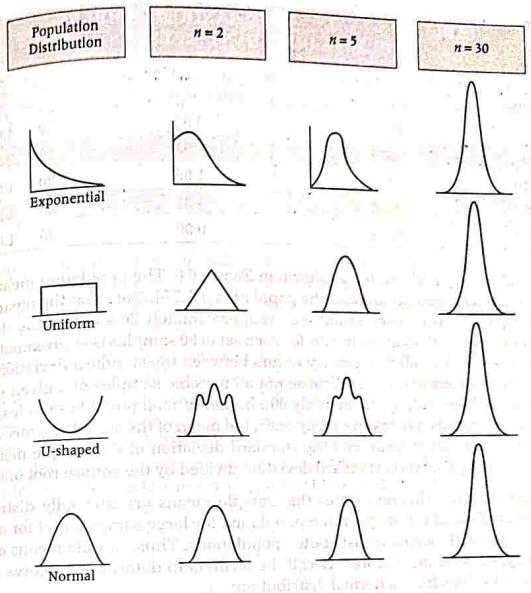


Figure 6.7: Shapes of the Distributions of Sample Means for Three Sample Sizes Drawn from Four Different Population Distributions

succeeding column displays the shape of the distribution of the sample means for a particular sample size. Note in the bottom row for the normally distributed population that the sample means are normally distributed even for $n = 2$. Note also that with the other population distributions, the distribution of the sample means begins to approximate the normal curve as n becomes larger. For all four distributions, the distribution of sample means is approximately normal for $n = 30$.

How large must a sample be for the central limit theorem to apply? The sample size necessary varies according to the shape of the population. However, in this text (as in many others), a sample of *size 30 or larger* will suffice. Recall that if the population is normally distributed, the sample means are normally distributed for sample sizes as small as $n = 1$.

The shapes displayed in Figure 6.7 coincide with the results obtained empirically from the random sampling shown in Figures 6.5 and 6.6. As shown in Figure 6.7, and as indicated in Figure 6.6, as sample size increases, the distribution narrows, or becomes more leptokurtic. This trend makes sense because the standard deviation of the mean is $\sigma/\sqrt{n}$. This value will become smaller as the size of n increases.

In **Table 6.5**, the means and standard deviations of the means are displayed for random samples of various sizes ($n = 2$ through $n = 30$) drawn from the uniform

TABLE 6.5: $\mu_{\bar{x}}$ AND $\sigma_{\bar{x}}$ OF 90 RANDOM SAMPLES FOR FIVE DIFFERENT SIZES*

Sample Size	Mean of Sample Means	Standard Deviation of Sample Means	μ	$\frac{\sigma}{\sqrt{n}}$
$n = 2$	19.92	3.87	20	4.08
$n = 5$	20.17	2.65	20	2.58
$n = 10$	20.04	1.96	20	1.83
$n = 20$	20.20	1.37	20	1.29
$n = 30$	20.25	0.99	20	1.05

distribution of a = 10 and b = 30 shown in Figure 6.6. The population mean is 20, and the standard deviation of the population is 5.774. Note that the mean of the sample means for each sample size is approximately 20 and that the standard deviation of the sample means for each set of 90 samples is approximately equal to $\sigma/\sqrt{n}$. A small discrepancy occurs between the standard deviation of the sample means and $\sigma/\sqrt{n}$, because not all possible samples of a given size were taken from the population (only 90). In theory, if all possible samples for a given sample size are taken exactly once, the mean of the sample means will equal the population mean and the standard deviation of the sample means will equal the population standard deviation divided by the square root of n.

The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations. Thus, sample means can be analyzed by using z scores. Recall the formula to determine z scores for individual values from a normal distribution:

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

If sample means are normally distributed, the z score formula applied to sample means would be

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

This result follows the general pattern of z scores: the difference between the statistic and its mean divided by the statistic's standard deviation. In this formula, the mean of the statistic of interest is $\mu_{\bar{x}}$, and *the standard deviation of the statistic of interest* is $\sigma_{\bar{x}}$, sometimes referred to as **the standard error of the mean**. To determine $\mu_{\bar{x}}$, the researcher would randomly draw out all possible samples of the given size from the population, compute the sample means, and average them. This task is virtually impossible to accomplish in any realistic period of time. Fortunately, $\mu_{\bar{x}}$ equals the population mean, μ, which is easier to access. Likewise, to determine directly the value of $\sigma_{\bar{x}}$, the researcher would take all possible samples of a given size from a population, compute the sample means, and determine the standard deviation of sample means. This task also is practically impossible. Fortunately, $\sigma_{\bar{x}}$ can be computed by using the population standard deviation divided by the square root of the sample size.

*Randomly generated from a uniform distribution with a = 10, b = 30.

As sample size increases, the standard deviation of the sample means becomes smaller and smaller because the population standard deviation is being divided by larger and larger values of the square root of n. The ultimate benefit of the central limit theorem is a practical, useful version of the z formula for sample means.

z Formula for Sample Means (6.2)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

When the population is normally distributed and the sample size is 1, this formula for sample means becomes the z formula for individual values that we used in Chapter 6. The reason is that the mean of one value is that value, and when $n = 1$ the value of $\sigma/\sqrt{n} = \sigma$.

As an example of the application of this z formula for sample means (Formula 6.2), suppose the population mean expenditure per customer at a tire store is \$125 and the population standard deviation is \$30. If a random sample of 40 customers is taken, what is the probability that the sample mean expenditure is more than \$133? Because the sample size is greater than 30, the central limit theorem can be used and the sample means are normally distributed allowing us to use Formula 6.2. With $\mu = \$125$, $\sigma = \$30$, and a sample mean, $\bar{x}$, of \$133, z can be computed with Formula 6.2 as:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\$133 - \$125}{\frac{\$30}{\sqrt{40}}} = 1.69$$

From the z distribution (Table A.5), $z = 1.69$ produces a probability of .4545. This is the probability of getting a sample mean between \$125 and \$133. Solving for the tail of the distribution yields

$$.5000 - .4545 = .0455$$

which is the probability of $\bar{x} > \$133$. That is, 4.55% of the time, a random sample of 40 customers from this population would yield a sample mean expenditure of \$133 or more when the population mean is \$125. **Figure 6.8** shows the problem and its solution.

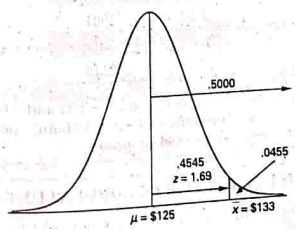
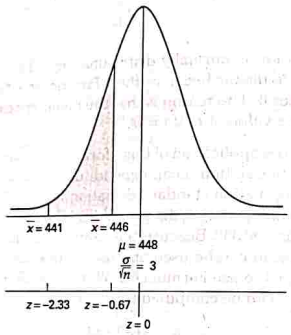


Figure 6.8: Graphical Solution to the Tire Store Problem

DEMONSTRATION PROBLEM 6.1

Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers. What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?

Solution: For this problem, $\mu = 448$, $\sigma = 21$, and $n = 49$. The problem is to determine $P(441 \leq \bar{x} \leq 446)$. The following diagram depicts the problem.



Solve this problem by calculating the z scores and using Table A.5 to determine the probabilities.

$$z = \frac{441 - 448}{\frac{21}{\sqrt{49}}} = \frac{-7}{3} = -2.33$$

and

$$z = \frac{446 - 448}{\frac{21}{\sqrt{49}}} = \frac{-2}{3} = -0.67$$

z Value	Probability
−2.33	.4901
−0.67	−.2486
	.2415

The probability of a value being between $z = -2.33$ and -0.67 is .2415; that is, there is a 24.15% chance of randomly selecting 49 hourly periods for which the sample mean is between 441 and 446 shoppers.

6.3.1 SAMPLING FROM A FINITE POPULATION

The example shown in this section and Demonstration Problem 6.1 was based on the assumption that the population was infinitely or extremely large. In cases of a finite population, *a statistical adjustment can be made to the z formula for sample means.* The adjustment is called the **finite correction factor:** $\sqrt{\frac{N-n}{N-1}}$. It operates on the standard deviation of sample mean, $\sigma_{\bar{x}}$.

Following is the z formula for sample means when samples are drawn from finite populations.

z Formula for Sample Means of a Finite Population (6.3)

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

If a random sample of size 35 were taken from a finite population of only 500, the sample mean would be less likely to deviate from the population mean than would be the case if a sample of size 35 were taken from an infinite population. For a sample of size 35 taken from a finite population of size 500, the finite correction factor is

$$\sqrt{\frac{500-35}{500-1}} = \sqrt{\frac{465}{499}} = .965$$

Thus the standard deviation of the mean—sometimes referred to as the standard error of the mean—is adjusted downward by using .965. As the size of the finite population becomes larger in relation to sample size, the finite correction factor approaches 1. In theory, whenever researchers are working with a finite population, they can use the finite correction factor. A rough rule of thumb for many researchers is that, if the sample size is less than 5% of the finite population size or $n/N < 0.05$, the finite correction factor does not significantly modify the solution. **Table 6.6** contains some illustrative finite correction factors.

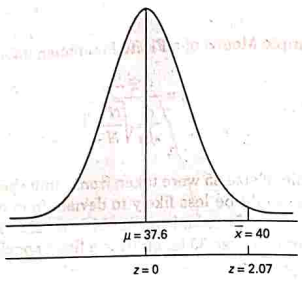
TABLE 6.6: FINITE CORRECTION FACTOR FOR SOME SAMPLE SIZES

Population Size	Sample Size	Value of Correction Factor
2000	30(<5% N)	.993
2000	500	.866
500	30	.971
500	200	.775
200	30	.924
200	75	.793

DEMONSTRATION PROBLEM 6.2

A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years. If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

Solution: The population mean is 37.6, with a population standard deviation of 8.3; that is, $\mu = 37.6$ and $\sigma = 8.3$. The sample size is 45, but it is being drawn from a finite population of 350; that is, $n = 45$ and $N = 350$. The sample mean under consideration is 40, or $\bar{x} = 40$. The following diagram depicts the problem on a normal curve.



Using the z formula with the finite correction factor gives

$$z = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}}\sqrt{\frac{350-45}{350-1}}} = \frac{2.4}{1.157} = 2.07$$

This z value yields a probability (Table A.5) of .4808. Therefore, the probability of getting a sample average age of less than 40 years is .4808 + .5000 = .9808. Had the finite correction factor not been used, the z value would have been 1.94, and the final answer would have been .9738.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

6. The central limit theorem creates the potential for applying the _____ to many problems when sample size is sufficiently large.
7. As sample size increases, the distribution narrows; and if we do analysis of kurtosis, then it will be found as _____.

State whether the following statements are true/false:

8. As sample sizes become larger, variation among the means increases.
9. For all four distributions, the distribution of sample means is approximately normal for $n > 30$.
10. Standard deviation of the sample means is also called the standard error of the mean.

ACTIVITY

You need to take the 2011 Census report for this activity. Calculate the average salary of the state which you belong to. Now calculate the average salary of the states which are neighbor to your state. Is it getting normally distributed? Explain.

6.4 SAMPLING DISTRIBUTION OF $\hat{p}$

Sometimes a business researcher will choose to use a sample proportion, denoted as $\hat{p}$, in the analysis of data rather than a sample mean, $\bar{x}$. If research produces *measurable* data such as weight, distance, time, and income, the sample mean is often the statistic of choice. However, if research results in *countable* items, such as how many people in a sample choose Dr. Pepper as their soft drink or how many people in a sample have a flexible work schedule, the sample proportion is often the statistic of choice. Whereas the mean is computed by averaging a set of values, the **sample proportion** is *computed by dividing the frequency with which a given characteristic occurs in a sample by the number of items in the sample.*

Sample Proportion (6.4)

$$\hat{p} = \frac{x}{n}$$

where

x = number of items in a sample that have the characteristic

n = number of items in the sample

For example, in a sample of 100 factory workers, 30 workers might belong to a union. The value of $\hat{p}$ for this characteristic, union membership, is 30/100 = .30. In a sample of 500 businesses in suburban malls, if 10 are shoe stores, then the sample proportion of shoe stores is 10/500 = .02. The sample proportion is a widely used statistic and is usually computed on questions involving Yes or No answers. For example, do you have at least a high school education? Are you predominantly right-handed? Are you female? Do you belong to the student accounting association?

How does a researcher use the sample proportion in analysis? The central limit theorem applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportions if $n \cdot p > 5$ and $n \cdot q > 5$ (p is the population proportion and $q = 1 - p$). The mean of sample proportions for all samples of size n randomly drawn from a population is p (the population proportion) and the standard deviation of sample proportions is $\sqrt{\frac{p \cdot q}{n}}$, sometimes referred to as the **standard error of the proportion.** Using this information, a z formula (6.5) for sample proportions can be developed.

z Formula for Sample Proportions for $n \cdot p > 5$ and $n \cdot q > 5$ (6.5)

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

$\hat{p}$ = sample proportion

n = sample size

p = population proportion

$q = 1 - p$

Suppose 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that .50 or less use that brand of wire? For this problem,

$$p = .60 \quad \hat{p} = .50 \quad n = 120$$

The z formula yields

$$z = \frac{.50 - .60}{\sqrt{\frac{(.60)(.40)}{120}}} = \frac{-.10}{.0447} = -2.24$$

From Table A.5, the probability corresponding to $z = -2.24$ is .4875. For $z < -2.24$ (the tail of the distribution), the answer is .5000 − .4875 = .0125. **Figure 6.9** shows the problem and solution graphically.

This answer indicates that a researcher would have difficulty (probability of .0125) finding that 50% or less of a sample of 120 contractors use a given brand of wire if indeed the population market share for that wire is .60. If this sample result actually occurs, either it is a rare chance result, the .60 proportion does not hold for this population, or the sampling method may not have been random.

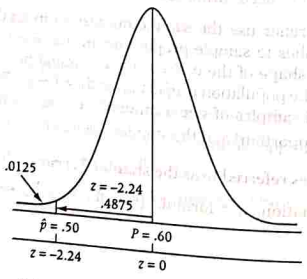


Figure 6.9: Graphical Solution to the Electrical Contractor Example

DEMONSTRATION PROBLEM 6.3

If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more parts are defective?

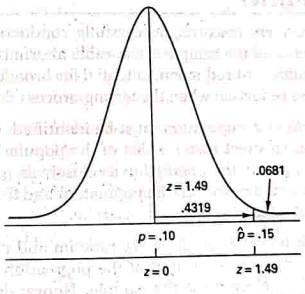
Solution: Here, $p = .10$, $\hat{p} = 12/80 = .15$, and $n = 80$. Entering these values in the z formula yields

$$z = \frac{.15 - .10}{\sqrt{\frac{(.10)(.90)}{80}}} = \frac{.05}{.0335} = 1.49$$

Table A.5 gives a probability of .4319 for a z value of 1.49, which is the area between the sample proportion, .15, and the population proportion, .10. The answer to the question is

$$P(\hat{p} \geq .15) = .5000 - .4319 = .0681.$$

Thus, about 6.81% of the time, 12 or more defective parts would appear in a random sample of 80 parts when the population proportion is .10. If this result actually occurred, the 10% proportion for population defects would be open to question. The diagram shows the problem graphically.



SELF ASSESSMENT QUESTIONS

Fill in the blanks:

11. If research results in _____ items, sample proportion is often the statistic of choice than sample mean.
12. Standard deviation of sample proportions is sometimes referred to as the _____ of the proportion.

State whether the following statements are true/false:

13. The central limit theorem applies to sample proportions in that the normal distribution approximates, if $n \cdot p$ and $n \cdot q$ should be less than 5.

SELF ASSESSMENT QUESTIONS

14. Sample proportion is computed by dividing the number of items in a sample that have the same characteristic by the number of items in the sample.
15. The sample proportion is a widely used statistic and is usually computed on questions involving Yes or No answers.

ACTIVITY

The NIFTY 50 is a diversified 50 stock index accounting for 12 sectors of the economy. It is used for a variety of purposes, such as benchmarking fund portfolios, index-based derivatives and index funds.

Use the data from the website (www.nseindia.com) to calculate a sample proportion of all the industries. Also calculate the sample proportion and 95% confidence interval for the proportion of industry. Which interval is wider? Will this always be the case?

6.5 SUMMARY

- For much business research, successfully conducting a census is virtually impossible and the sample is a feasible alternative. Other reasons for sampling include cost reduction, potential for broadening the scope of the study, and loss reduction when the testing process destroys the product.
- To take a sample, a population must be identified. Often the researcher cannot obtain an exact roster or list of the population and so must find some way to identify the population as closely as possible. The final list or directory used to represent the population and from which the sample is drawn is called the frame.
- The two main types of sampling are random and nonrandom. Random sampling occurs when each unit of the population has the same probability of being selected for the sample. Nonrandom sampling is any sampling that is not random. The four main types of random sampling discussed are simple random sampling, stratified sampling, systematic sampling, and cluster or area, sampling.
- In simple random sampling, every unit of the population is numbered. A table of random numbers or a random number generator is used to select n units from the population for the sample.
- Stratified random sampling uses the researcher's prior knowledge of the population to stratify the population into subgroups. Each subgroup is internally homogeneous but different from the others. Stratified random sampling is an attempt to reduce sampling error and ensure that at least some of each of the subgroups appears in the sample. After the strata are identified, units can be sampled randomly from each stratum. If the proportions of units selected from each subgroup for the sample are the same as the proportions of the subgroups in the population, the process is called proportionate stratified sampling. If not, it is called disproportionate stratified sampling.
- With systematic sampling, every kth item of the population is sampled until n units have been selected. Systematic sampling is used because of its convenience and ease of administration.
- Cluster or area sampling involves subdividing the population into nonoverlapping clusters or areas. Each cluster or area is a microcosm of the population and is usually heterogeneous within. A sample of clusters is randomly selected from the population. Individual units are then selected randomly from the clusters or areas to get the final sample. Cluster or area sampling is usually done to reduce costs. If a set of second clusters or areas is selected from the first set, the method is called two-stage sampling.
- Four types of nonrandom sampling were discussed: convenience, judgment, quota, and snowball. In convenience sampling, the researcher selects units from the population to be in the sample for convenience. In judgment sampling, units are selected according to the judgment of the researcher. Quota sampling is similar to stratified sampling, with the researcher identifying subclasses or strata. However, the researcher selects units from each stratum by some nonrandom technique until a specified quota from each stratum is filled. With snowball sampling, the researcher obtains additional sample members by asking current sample members for referral information.
- Sampling error occurs when the sample does not represent the population. With random sampling, sampling error occurs by chance. Nonsampling errors are all other research and analysis errors that occur in a study. They can include recording errors, input errors, missing data, and incorrect definition of the frame.
- According to the central limit theorem, if a population is normally distributed, the sample means for samples taken from that population also are normally distributed regardless of sample size. The central limit theorem also says that if the sample sizes are large ($n \geq 30$), the sample mean is approximately normally distributed regardless of the distribution shape of the population. This theorem is extremely useful because it enables researchers to analyze sample data by using the normal distribution for virtually any type of study in which means are an appropriate statistic, as long as the sample size is large enough. The central limit theorem states that sample proportions are normally distributed for large sample sizes.

KEY WORDS

1. **Central limit theorem:** According to the central limit theorem, if a population is normally distributed, the sample means for samples taken from that population also are normally distributed regardless of sample size. The central limit theorem also says that if the sample sizes are large ($n \geq 30$), the sample mean is approximately normally distributed regardless of the distribution shape of the population.
2. **Cluster (or area) sampling:** Cluster (or area) sampling involves dividing the population into nonoverlapping areas, or clusters. Clusters are tend to be internally heterogeneous.
3. **Convenience sampling:** In convenience sampling, elements for the sample are selected for the convenience of the researcher.

KEY WORDS

4. **Disproportionate stratified:** Disproportionate stratified random sampling occurs where the proportions of the strata in the sample are different from the proportions of the strata in the population.
5. **Finite correction factor:** for finite population, a statistical adjustment can be made to the z formula for sample mean, The adjustment is called Finite correction factor.
6. **Frame:** The final list or directory used to represent the population and from which the sample is drawn is called the frame.
7. **Judgment sampling:** In judgment sampling, units are selected according to the judgment of the researcher.
8. **Nonrandom sampling:** Nonrandom sampling is any sampling that is not random. In this method of sampling every unit of the population doesn't have the same probability of being getting selected.
9. **Nonrandom sampling techniques:** Sampling techniques used to select elements from the population by any mechanism that does not involve a random selection process are called nonrandom sampling techniques.
10. **Nonsampling errors:** All errors other than sampling errors are nonsampling errors. The many possible nonsampling errors include missing data, recording errors, input processing errors, and analysis errors.
11. **Proportionate stratified random:** Proportionate stratified random sampling occurs when the percentage of the sample taken from each stratum is proportionate to the percentage that each stratum is within the whole population.
12. **Sampling:** sampling is a method to select a subset from a population which have some properties of population itself.
13. **Quota sampling:** Quota sampling is similar to stratified sampling, with the researcher identifying subclasses or strata. However, the researcher selects units from each stratum by some nonrandom technique until a specified quota from each stratum is filled.
14. **Random sampling:** Random sampling occurs when each unit of the population has the same probability of being selected for the sample.
15. **Sample proportion:** Sample proportion is computed by dividing the frequency with which a given characteristic occurs in a sample by the number of items in the sample. We use this method if research results in countable items.
16. **Sampling error:** Sampling error occurs when the sample is not representative of the population.
17. **Simple random sampling:** In simple random sampling, every unit of the population is numbered. A table of random numbers or a random number generator is used to select n units from the population for the sample.
18. **Snowball sampling:** snowball sampling, the researcher obtains additional sample members by asking current sample members for referral information.

KEY WORDS

19. **Stratified random sampling:** In stratified random sampling the population is divided into nonoverlapping subpopulations called strata. The researcher then extracts a random sample from each of the subpopulations (strata).
20. **Systematic sampling:** With systematic sampling, every kth item is selected to produce a sample of size n from a population of size N.
21. **Two-stage sampling:** When clusters are too large, and a second set of clusters is taken from each original cluster then this technique is called two-stage sampling.

6.6 DESCRIPTIVE QUESTIONS

6.1. The mean of a population is 76 and the standard deviation is 14. The shape of the population is unknown. Determine the probability of each of the following occurring from this population.

(a) A random sample of size 35 yielding a sample mean of 79 or more

(b) A random sample of size 140 yielding a sample mean of between 74 and 77

(c) A random sample of size 219 yielding a sample mean of less than 76.5

6.2. Suppose the age distribution in a city is as follows.

Under 18	22%
18–25	18%
26–50	36%
51–65	10%
Over 65	14%

A researcher is conducting proportionate stratified random sampling with a sample size of 250. Approximately how many people should be sample from each stratum?

6.3. Determine a possible frame for conducting random sampling in each of the following studies.

(a) The average amount of overtime per week for production workers in a plastics company in Pennsylvania

(b) The average number of employees in all Ralphs supermarkets in Southern California

(c) A survey of commercial lobster catchers in Maine

6.4. A company has 1,250 employees, and you want to take a simple random sample of $n = 60$ employees. Explain how you would go about selecting this sample by using the table of random numbers. Are there numbers that you cannot use? Explain.

6.5. A survey of 2645 consumers by DDB Needham Worldwide of Chicago for public relations agency Porter/Novelli showed that how a company handles a crisis when at fault is one of the top influences in consumer buying decisions, with 73% claiming it is an influence. Quality of product was the number one influence, with 96% of consumers stating that quality influences their buying decisions. How a company handles complaints was number two, with 85% of consumers reporting it as an influence in their buying decisions. Suppose a random sample of 1100 consumers is taken and each is asked which of these three factors influence their buying decisions.

(a) What is the probability that more than 810 consumers claim that how a company handles a crisis when at fault is an influence in their buying decisions?

(b) What is the probability that fewer than 1030 consumers claim that quality of product is an influence in their buying decisions?

(c) What is the probability that between 82% and 84% of consumers claim that how a company handles complaints is an influence in their buying decisions?

6.6. A researcher is conducting a study of a *Fortune* 500 company that has factories, distribution centers, and retail outlets across the country. How can she use cluster or area sampling to take a random sample of employees of this firm?

6.7. The Aluminum Association reports that the average American uses 56.8 pounds of aluminum in a year. A random sample of 51 households is monitored for one year to determine aluminum usage. If the population standard deviation of annual usage is 12.3 pounds, what is the probability that the sample mean will be each of the following?

(a) More than 60 pounds

(b) More than 58 pounds

(c) Between 56 and 57 pounds

(d) Less than 55 pounds

(e) Less than 50 pounds

6.8. Direct marketing companies are turning to the Internet for new opportunities. A recent study by Gruppo, Levey, & Co. showed that 73% of all direct marketers conduct transactions on the Internet. Suppose a random sample of 300 direct marketing companies is taken.

(a) What is the probability that between 210 and 234 (inclusive) direct marketing companies are turning to the Internet for new opportunities?

(b) What is the probability that 78% or more of direct marketing companies are turning to the Internet for new opportunities?

(c) Suppose a random sample of 800 direct marketing companies is taken. Now what is the probability that 78% or more are turning to the Internet for new opportunities? How does this answer differ from the answer in part (b)? Why do the answers differ?

6.8. In a particular area of the Northeast, an estimated 75% of the homes use heating oil as the principal heating fuel during the winter. A random telephone survey of 150 homes is taken in an attempt to determine whether this figure is correct. Suppose 120 of the 150 homes surveyed use heating oil as the principal heating fuel. What is the probability of getting a sample proportion this large or larger if the population estimate is true?

6.7 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

6.1. $\mu = 76, \sigma = 14$

(a) $n = 35, P(\bar{x} \geq 79)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{79 - 76}{\frac{14}{\sqrt{35}}} = 1.27$$

from table A.5, area = .3980

$P(\bar{x} \geq 79) = .5000 - .3980 = \mathbf{.1020}$

(b) $n = 140, P(74 \leq \bar{x} \leq 77)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{74 - 76}{\frac{14}{\sqrt{140}}} = -1.69 \qquad z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{77 - 76}{\frac{14}{\sqrt{140}}} = 0.85$$

from table A.5, area for $z = -1.69$ is .4545

from table A.5, area for 0.85 is .3023

$P(74 \leq \bar{x} \leq 77) = .4545 + .3023 = \mathbf{.7568}$

(c) $n = 219, P(\bar{x} < 76.5)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{76.5 - 76}{\frac{14}{\sqrt{219}}} = 0.53$$

from table A.5, area = .2019

$P(\bar{x} < 76.5) = .5000 + .2019 = \mathbf{.7019}$

6.2.

Under 18	$250(.22) = \mathbf{55}$
18 – 25	$250(.18) = \mathbf{45}$
26 – 50	$250(.36) = \mathbf{90}$
51 – 65	$250(.10) = \mathbf{25}$
over 65	$250(.14) = \mathbf{35}$
	$n = 250$

6.3. (a) Roster of production employees secured from the human resources department of the company.

(b) Ralphs store records kept at the headquarters of their California division or merged files of store records from regional offices across the state.

(c) Membership list of Maine lobster catchers association.

6.4. Number the employees from 0001 to 1250. Randomly sample from the random number table until 60 different usable numbers are obtained. You cannot use numbers from 1251 to 9999.

6.5. $n = 1100$

(a) $x > 810, p = .73$

$$\hat{p} = \frac{x}{n} = \frac{810}{1100}$$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.7364 - .73}{\sqrt{\frac{(.73)(.27)}{1100}}} = 0.48$$

from table A.5, area = .1844

$$P(x > 810) = .5000 - .1844 = \mathbf{.3156}$$

(b) $x < 1030, p = .96,$

$$\hat{p} = \frac{x}{n} = \frac{1030}{1100} = .9364$$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.9364 - .96}{\sqrt{\frac{(.96)(.04)}{1100}}} = -3.99$$

from table A.5, area = .49997

$P(x < 1030) = .5000 - .49997 = \mathbf{.00003}$

(c) $p = .85$

$P(.82 \leq \hat{p} \leq .84)$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.82 - .85}{\sqrt{\frac{(.85)(.15)}{1100}}} = -2.79$$

from table A.5, area = .4974

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.84 - .85}{\sqrt{\frac{(.85)(.15)}{1100}}} = -0.93$$

from table A.5, area = .3238

$P(.82 \leq \hat{p} \leq .84) = .4974 - .3238 = \mathbf{.1736}$

6.6. Divide the factories into geographic regions and select a few factories to represent those regional areas of the country. Take a random sample of employees from each selected factory. Do the same for distribution centers and retail outlets. Divide the United States into regions of areas. Select a few areas. Take a random sample from each of the selected area distribution centers and retail outlets.

6.7. $\mu = 56.8 \quad n = 51 \quad \sigma = 12.3$

(a) $P(\bar{x} > 60)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{60 - 56.8}{\frac{12.3}{\sqrt{51}}} = 1.86$$

from Table A.5, Prob. = .4686

$P(\bar{x} > 60) = .5000 - .4686 = \mathbf{.0314}$

(b) $P(\bar{x} > 58)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{58 - 56.8}{\frac{12.3}{\sqrt{51}}} = 0.70$$

from Table A.5, Prob.= .2580

$P(\bar{x} > 58) = .5000 - .2580 = \mathbf{.2420}$

(c) $P(56 < \bar{x} < 57)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{56 - 56.8}{\frac{12.3}{\sqrt{51}}} = -0.46 \qquad z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{57 - 56.8}{\frac{12.3}{\sqrt{51}}} = 0.12$$

from Table A.5, Prob. for $z = -0.46$ is .1772

from Table A.5, Prob. for $z = 0.12$ is .0478

$P(56 < \bar{x} < 57) = .1772 + .0478 = \mathbf{.2250}$

(d) $P(\bar{x} < 55)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{55 - 56.8}{\frac{12.3}{\sqrt{51}}} = -1.05$$

from Table A.5, Prob.= .3531

$P(\bar{x} < 55) = .5000 - .3531 = \mathbf{.1469}$

(e) $P(\bar{x} < 50)$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{50 - 56.8}{\frac{12.3}{\sqrt{51}}} = -3.95$$

from Table A.5, Prob.= .5000

$P(\bar{x} < 50) = .5000 - .5000 = \mathbf{.0000}$

6.8. $p = .73, n = 300$

(a) $P(210 \le x \le 234)$:

$$\hat{p}_1 = \frac{x}{n} = \frac{210}{300} = .70 \qquad \hat{p}_2 = \frac{x}{n} = \frac{234}{300} = .78$$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.70 - .73}{\sqrt{\frac{(.73)(.27)}{300}}} = -1.17$$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.78 - .73}{\sqrt{\frac{(.73)(.27)}{300}}} = -1.95$$

from Table A.5, the area for $z = -1.17$ is .3790

the area for $z = 1.95$ is .4744

$P(210 \le x \le 234) = .3790 + .4744 = .8534$

(b) $P(\hat{p} \ge .78)$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.78 - .73}{\sqrt{\frac{(.73)(.27)}{300}}} = 1.95$$

from Table A.5, the area for $z = 1.95$ is .4744

$P(\hat{p} \ge .78) = .5000 - .4744 = .0256$

(c) $p = .73 \qquad n = 800 \qquad P(\hat{p} \ge .78)$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.78 - .73}{\sqrt{\frac{(.73)(.27)}{800}}} = 3.19$$

from Table A.5, the area for $z = 3.19$ is .4993

$P(\hat{p} \ge .78) = .5000 - .4993 = .0007$

6.9. $p = .75\ n = 150\ x = 120$

$P(\hat{p} \ge .80)$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.80 - .75}{\sqrt{\frac{(.75)(.25)}{150}}} = 1.41$$

from Table A.5, the area for $z = 1.41$ is .4207

$P(\hat{p} \ge .80) = .5000 - .4207 = .0793$

6.8 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
6.2 Sampling	1.	overregistered
	2.	random sampling
	3.	strata
	4.	False
	5.	False
6.3 Sampling Distribution of $\bar{x}$	6.	normal distribution
	7.	leptokurtic
	8.	False
	9.	True
	10.	True
6.4 Sampling Distribution of $\hat{p}$	11.	countable
	12.	standard error
	13.	False
	14.	True
	15.	True

7

CHAPTER

CORRELATION AND SIMPLE REGRESSION ANALYSIS

CONTENTS

INTRODUCTORY CASELET

NOTES

PREDICTING INTERNATIONAL MINIMUM WAGES BY THE PRICE OF A BIG MAC

In 120 countries we can have McDonald's burger and round about 68 million customers each day eats something from McDonald stores located all over the world. McDonald's operates 37,241 restaurants worldwide, employing more than 380,000 people as of the end of 2017.The McDonald's Corporation is the leading global foodservice retailer with more than 36,000 local restaurants serving nearly 69 million people in more than 100 countries each day. This global presence, in addition to its consistency in food offerings and restaurant operations, makes McDonald's a unique and attractive setting for economists to make salary and price comparisons around the world. Because the Big Mac hamburger is a standardized hamburger produced and sold in virtually every McDonald's around the world, the Economist, a weekly newspaper focusing on international politics and business news and opinion, was compiling as early as 1986 information about Big Mac prices as an indicator of exchange rates.

Building on this idea, researchers Ashenfelter and Jurajda proposed comparing wage rates across countries using the price of a Big Mac hamburger. The correlation between dollar price and minimum wages of employee can help many companies to know in which direction they are going and whether the price of food item and the salary are related or not. Shown below are Big Mac prices and minimum monthly wage figures (in U.S. dollars) for 25 countries.

Country	Dollar Price	Minimum Wages of Employee
Argentina	4.13	506.0 $
Australia	4.53	2,104.7 $
Belgium	4.62	1,874.0 $
Brazil	5.10	225.4 $
Canada	4.66	1,377.7 $
China	2.92	224.7 $
Germany	4.45	1,796.6 $
Greece	3.83	820.0 $
Hong Kong	2.46	821.9 $
India	2.76	54.5 $
Indonesia	2.40	137.8 $
Ireland	4.65	1,935.6 $
Israel	4.77	1,339.4 $
Japan	3.36	1,155.5 $
New Zealand	4.43	1,795.3 $
Pakistan	3.57	102.7 $
Poland	2.72	603.2 $
Portugal	3.71	811.5 $
Russia	2.28	85.1 $

INTRODUCTORY CASELET

Country	Dollar Price	Minimum Wages of Employee
Saudi Arabia	3.20	800.0 $
South Africa	2.26	291.6 $
South Korea	3.84	1,123.4 $
Spain	4.34	1,029.7 $
Sri Lanka	3.77	52.0 $
United States	5.30	1,256.7 $

LEARNING OBJECTIVES

The overall objective of Chapter 8 is to give you an understanding of bivariate linear regression analysis and correlation, thereby enabling you to:

- Calculate the Pearson product-moment correlation coefficient to determine if there is a correlation between two variables.
- Explain what regression analysis is and the concepts of independent and dependent variable.
- Calculate the slope and y-intercept of the least squares equation of a regression line and from those, determine the equation of the regression line.
- Calculate the standard error of the estimate using the sum of squares of error, and use the standard error of the estimate to determine the fit of the model.
- Calculate the coefficient of determination to measure the fit for regression models, and relate it to the coefficient of correlation.
- Calculate the residuals of regression line and test the assumptions of the regression model.

7.1 INTRODUCTION

In business, the key to decision making often lies in the understanding of the relationships between two or more variables. For example, a company in the distribution business may determine that there is a relationship between the price of crude oil and the company's transportation costs. Financial experts, in studying the behavior of the bond market, might find it useful to know if the interest rates on bonds are related to the prime interest rate set by the Federal Reserve. A marketing executive might want to know how strong the relationship is between advertising dollars and sales dollars for a product or a company.

In this chapter, we will study the concept of correlation and how it can be used to estimate the relationship between two variables. We will also explore simple regression analysis through which mathematical models can be developed to predict one variable by another. We will examine tools for testing the strength and predictability of regression models.

7.2 CORRELATION

Correlation is *a measure of the degree of relatedness of variables.* It can help a business researcher determine, for example, whether the stocks of two airlines rise and fall in any related manner. For a sample of pairs of data, correlation analysis can yield a numerical value that represents the degree of relatedness of the two stock prices over time. In the transportation industry, is a correlation evident between the price of transportation and the weight of the object being shipped? If so, how strong are the correlations? In economics, how strong is the correlation between the producer price index and the unemployment rate? In retail sales, are sales

related to population density, number of competitors, size of the store, amount of advertising, or other variables?

Several measures of correlation are available, the selection of which depends mostly on the level of data being analyzed. Ideally, researchers would like to solve for μ, the population coefficient of correlation. However, because researchers virtually always deal with sample data, this section introduces a widely used sample **coefficient of correlation**, r. This measure is applicable only if both variables being analyzed have at least an interval level of data.

The statistic r is the **Pearson product-moment correlation coefficient**, named after Karl Pearson (1857–1936), an English statistician who developed several coefficients of correlation along with other significant statistical concepts. The term r is a *measure of the linear correlation of two variables*. It is a number that ranges from −1 to 0 to +1, representing the strength of the relationship between the variables. An r value of +1 denotes a perfect positive relationship between two sets of numbers. An r value of −1 denotes a perfect negative correlation, which indicates an inverse relationship between two variables: as one variable gets larger, the other gets smaller. An r value of 0 means no linear relationship is present between the two variables.

Pearson Product-Moment Correlation Coefficient

$$r = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2}} = \frac{\Sigma xy - \dfrac{(\Sigma x\Sigma y)}{n}}{\sqrt{\left[\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}\right]\left[\Sigma y^2 - \dfrac{(\Sigma y)^2}{n}\right]}} \tag{7.1}$$

Figure 7.1 depicts five different degrees of correlation: (a) represents strong negative correlation, (b) represents moderate negative correlation, (c) represents moderate positive correlation, (d) represents strong positive correlation, and (e) contains no correlation.

What is the measure of correlation between the interest rate of federal funds and the commodities futures index? With data such as those shown in **Table 7.1**, which represent the values for interest rates of federal funds and commodities futures indexes for a sample of 12 days, a correlation coefficient, r, can be computed.

Examination of the formula for computing a Pearson product-moment correlation coefficient (7.1) reveals that the following values must be obtained to compute r: Σx, Σx^2, Σy, Σy^2, Σxy, and n. In correlation analysis, it does not matter which variable is designated x and which is designated y. For this example, the correlation coefficient is computed as shown in **Table 7.2**. The r value obtained ($r = .815$) represents a relatively strong positive relationship between interest rates and commodities futures index over this 12-day period.

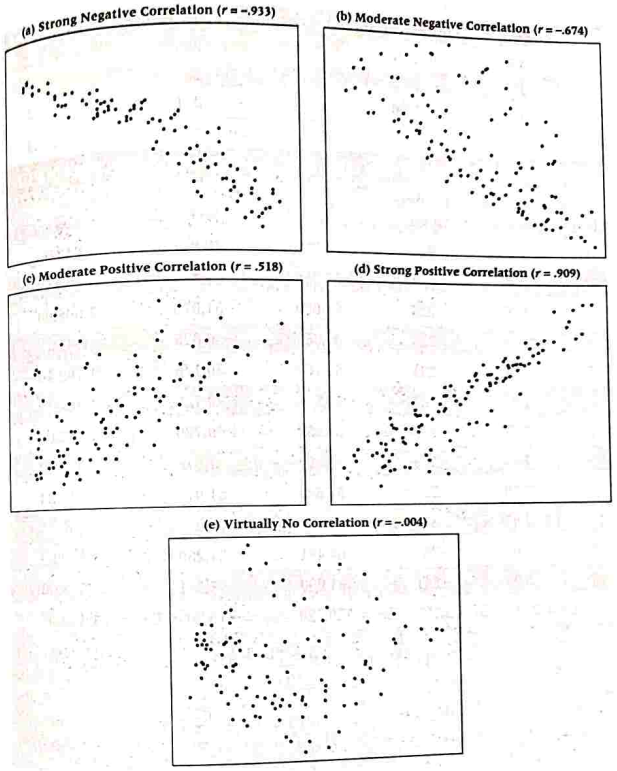


Figure 7.1: Five Correlations

TABLE 7.1: DATA FOR THE ECONOMICS EXAMPLE

Day	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226

(Continued)

Day	Interest Rate	Futures Index
10	8.07	235
11	8.03	233
12	8.00	241

TABLE 7.2: COMPUTATION OF r FOR THE ECONOMICS EXAMPLE

Day	Interest Rate x	Futures Index y	x^2	y^2	xy
1	7.43	221	55.205	48,841	1,642.03
2	7.48	222	55.950	49,284	1,660.56
3	8.00	226	64.000	51,076	1,808.00
4	7.75	225	60.063	50,625	1,743.75
5	7.60	224	57.760	50,176	1,702.40
6	7.63	223	58.217	49,729	1,701.49
7	7.68	223	58.982	49,729	1,712.64
8	7.67	226	58.829	51,076	1,733.42
9	7.59	226	57.608	51,076	1,715.34
10	8.07	235	65.125	55,225	1,896.45
11	8.03	233	64.481	54,289	1,870.99
12	8.00	241	64.000	58,081	1,928.00

$\Sigma x = 92.93 \quad \Sigma y = 2{,}725 \quad \Sigma x^2 = 720.220 \quad \Sigma y^2 = 619{,}207 \quad \Sigma xy = 21{,}115.07$

$$r = \frac{(21{,}115.07) - \frac{(92.93)(2725)}{12}}{\sqrt{\left[(720.22) - \frac{(92.93)^2}{12}\right]\left[(619{,}207) - \frac{(2725)^2}{12}\right]}} = .815$$

Figure 7.2 shows Excel output for this problem.

Excel Output

	Interest Rate	*Futures Index*
Interest Rate	1	
Futures Index	0.815	1

Figure 7.2: Excel Output for the Economics Example

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. Sample coefficient of correlation is denoted by _____.
2. In _____, every unit of the population has the same probability of being selected into the sample.
3. An r value of -1 denotes a perfect _____ correlation.

State whether the following statements are true/false:

4. The correlation between x and square root of x will be -1.
5. The correlation coefficient is a number that ranges from 0 to 1.

ACTIVITY

From www.bseindia.com you need to download the SENSEX index data for JAN-2017. You can also find line graph for the same. You need to predict what type of correlation SENSEX index has with BHEL data for the same duration.

7.3 INTRODUCTION TO SIMPLE REGRESSION ANALYSIS

Regression analysis is *the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables*. The most elementary regression model is called **simple regression** or **bivariate regression** involving two variables in which one variable is predicted by another variable. In simple regression, *the variable to be predicted* is called the **dependent variable** and is designated as y. The *predictor* is called the **independent variable**, or *explanatory variable*, and is designated as x. In simple regression analysis, only a straight-line relationship between two variables is examined. Nonlinear relationships and regression models with more than one independent variable can be explored by using multiple regression models.

Can the cost of flying a commercial airliner be predicted using regression analysis? If so, what variables are related to such cost? A few of the many variables that can potentially contribute are type of plane, distance, number of passengers, amount of luggage/freight, weather conditions, direction of destination, and perhaps even pilot skill. Suppose a study is conducted using only Boeing 737s traveling 500 miles on comparable routes during the same season of the year. Can the number of passengers predict the cost of flying such routes? It seems logical that more passengers result in more weight and more baggage, which could, in turn, result in increased fuel consumption and other costs. Suppose the data displayed in **Table 7.3** are the costs and associated number of passengers for twelve 500-mile commercial airline flights using Boeing 737s during the same season of the year. We will use these data to develop a regression model to predict cost by number of passengers.

Usually, the first step in simple regression analysis is to construct a **scatter plot** (or scatter diagram), discussed in Chapter 2. Graphing the data in this

TABLE 7.3: AIRLINE COST DATA

Number of Passengers	Cost ($1,000)
61	4.280
63	4.080
67	4.420
69	4.170
70	4.480
74	4.300
76	4.820
81	4.700
86	5.110
91	5.130
95	5.640
97	5.560

way yields preliminary information about the shape and spread of the data. **Figure 7.3** is an Excel scatter plot of the data in Table 7.3. **Figure 7.4** is a close-up view of the scatter plot. Try to imagine a line passing through the points. Is a linear fit possible? Would a curve fit the data better? The scatter plot gives some idea of how well a regression line fits the data. Later in the chapter, we present statistical techniques that can be used to determine more precisely how well a regression line fits the data.

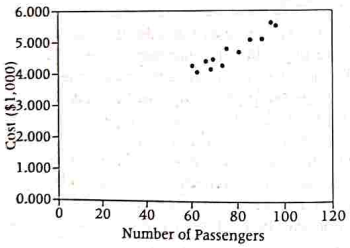


Figure 7.3: Excel Scatter Plot of Airline Cost Data

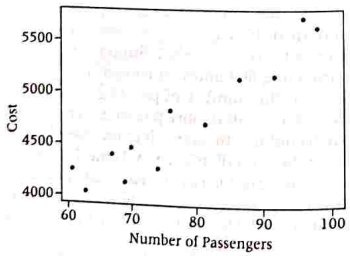


Figure 7.4: Close-Up Scatter Plot of Airline Cost Data

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

6. In simple regression, the variable to be predicted is called the _____.
7. _____ plot gives us an idea that by the data a linear fit is possible or not.

State whether the following statements are true/false:

8. Explanatory variable is designated as y.
9. Regression analysis is the process of constructing a mathematical model.
10. In simple regression analysis, only a straight-line relationship between two variables is examined.

ACTIVITY

https://data.gov.sg/dataset/credit-and-charge-card-statistics

From the above mentioned link, download the datasets. Now, out of all variables, decide which can be the dependent variable and which can be independent variable. Calculate the correlation between those variables which you have kept in dependent and independent list. Mention how does correlation create effect on independent and dependent variables.

7.4 DETERMINING THE EQUATION OF THE REGRESSION LINE

The first step in determining the equation of the regression line that passes through the sample data is to establish the equation's form. Several different types of equations of lines are discussed in algebra, finite math, or analytic geometry courses. Recall that among these equations of a line are the two-point form, the point-slope form, and the slope-intercept form. In regression analysis, researchers use the slope-intercept equation of a line. In math courses, the slope-intercept form of the equation of a line often takes the form

$$y = mx + b$$

where

m = slope of the line

$b = y$ intercept of the line

In statistics, the slope-intercept form of the equation of the regression line through the population points is

$$\hat{y} = \beta_0 + \beta_1 x$$

where

$\hat{y}$ = the predicted value of y

β_0 = the population y intercept

β_1 = the population slope

For any specific dependent variable value, y_i,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

x_i = the value of the independent variable for the ith value

y_i = the value of the dependent variable for the ith value

β_0 = the population y intercept

β_1 = the population slope

ϵ_i = the error of prediction for the ith value

Unless the points being fitted by the regression equation are in perfect alignment, the regression line will miss at least some of the points. In the preceding equation, ϵ_i represents the error of the regression line in fitting these points. If a point is on the regression line, $\epsilon_i = 0$.

These mathematical models can be either deterministic models or probabilistic models. **Deterministic models** are *mathematical models that produce an "exact" output for a given input.* For example, suppose the equation of a regression line is

$$y = 1.68 + 2.40x$$

For a value of $x = 5$, the exact predicted value of y is

$$y = 1.68 + 2.40(5) = 13.68$$

We recognize, however, that most of the time the values of y will not equal exactly the values yielded by the equation. Random error will occur in the prediction of the y values for values of x because it is likely that the variable x does not explain all the variability of the variable y. For example, suppose we are trying to predict the volume of sales (y) for a company through regression analysis by using the annual dollar amount of advertising (x) as the predictor. Although sales are often related to advertising, other factors related to sales are not accounted for by amount of advertising. Hence, a regression model to predict sales volume by amount of advertising probably involves some error. For this reason, in regression, we present the general model as a probabilistic model. A **probabilistic model** is *one that includes an error term that allows the y values to vary for any given value of x.*

A deterministic regression model is

$$y = \beta_0 + \beta_1 x$$

The probabilistic regression model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

$\beta_0 + \beta_1 x$ is the deterministic portion of the probabilistic model, $\beta_0 + \beta_1 x + \epsilon$. In a deterministic model, all points are assumed to be on the line and in all cases ϵ is zero.

Virtually all regression analyses of business data involve sample data, not population data. As a result, β_0 and β_1 are unattainable and must be estimated by using the sample statistics, b_0 and b_1. Hence the equation of the regression line contains the sample y intercept, b_0, and the sample slope, b_1.

Equation of the Simple Regression Line

$$\hat{y} = b_0 + b_1 x$$

where

b_0 = the sample y intercept

b_1 = the sample slope

To determine the equation of the regression line for a sample of data, the researcher must determine the values for b_0 and b_1. This process is sometimes referred to as least squares analysis. **Least squares analysis** is *a process whereby a regression model is developed by producing the minimum sum of the squared error values.* On the basis of this premise and calculus, a particular set of equations has been developed to produce components of the regression model.

Examine the regression line fit through the points in **Figure 7.5**. Observe that the line does not actually pass through any of the points. The vertical distance from each point to the line is the error of the prediction. In theory, an infinite number of lines could be constructed to pass through these points in some manner. The least squares regression line is the regression line that results in the smallest sum of errors squared.

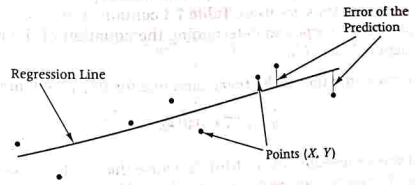


Figure 7.5: Plot of a Regression Line

Formula 7.2 is an equation for computing the value of the sample slope. Several versions of the equation are given to afford latitude in doing the computations.

Slope of the Regression Line

$$b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - n\bar{x}^2} = \frac{\Sigma xy - \dfrac{(\Sigma x)(\Sigma y)}{n}}{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}} \qquad (7.2)$$

The expression in the numerator of the slope Formula 7.2 appears frequently in this chapter and is denoted as SS_{xy}.

$$SS_{xy} = \Sigma(x-\bar{x})(y-\bar{y}) = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

The expression in the denominator of the slope Formula 7.2 also appears frequently in this chapter and is denoted as SS_{xx}.

$$SS_{xx} = \Sigma(x-\bar{x})^2 = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

With these abbreviations, the equation for the slope can be expressed as in Formula 7.3.

Alternative Formula for Slope

$$b_1 = \frac{SS_{xy}}{SS_{xx}} \tag{7.3}$$

Formula 7.4 is used to compute the sample y intercept. The slope must be computed before the y intercept.

y Intercept of the Regression Line

$$b_0 = \bar{y} - b_1\bar{x} = \frac{\Sigma y}{n} - b_1\frac{(\Sigma x)}{n} \tag{7.4}$$

Formulas 7.2, 7.3, and 7.4 show that the following data are needed from sample information to compute the slope and intercept: Σx, Σy, Σx^2, and, Σxy, unless sample means are used. **Table 7.4** contains the results of solving for the slope and intercept and determining the equation of the regression line for the data in Table 7.3.

The least squares equation of the regression line for this problem is

$$\hat{y} = 1.57 + .0407x$$

The slope of this regression line is .0407. Because the x values were recoded for the ease of computation and are actually in $1,000 denominations, the slope is actually $40.70. One interpretation of the slope in this problem is that for every unit increase in x (every person added to the flight of the airplane), there is a $40.70 increase in the cost of the flight. The y-intercept is the point where the line crosses the y-axis (where x is zero). Sometimes in regression analysis, the y-intercept is meaningless in terms of the variables studied. However, in this problem, one interpretation of the y-intercept, which is 1.570 or $1,570, is that even if there were no passengers on the commercial flight, it would still cost $1,570. In other words, there are costs associated with a flight that carries no passengers.

Superimposing the line representing the least squares equation for this problem on the scatter plot indicates how well the regression line fits the data points, as shown in the Excel graph in **Figure 7.6**. The next several sections explore mathematical ways of testing how well the regression line fits the points.

TABLE 7.4: SOLVING FOR THE SLOPE AND THE y INTERCEPT OF THE REGRESSION LINE FOR THE AIRLINE COST EXAMPLE

Number of Passengers	Cost ($1,000)		
x	y	x^2	xy
61	4.280	3,721	261.080
63	4.080	3,969	257.040
67	4.420	4,489	296.140
69	4.170	4,761	287.730
70	4.480	4,900	313.600
74	4.300	5,476	318.200
76	4.820	5,776	366.320
81	4.700	6,561	380.700
86	5.110	7,396	439.460
91	5.130	8,281	466.830
95	5.640	9,025	535.800
97	5.560	9,409	539.320
$\Sigma x = 930$	$\Sigma y = 56.690$	$\Sigma x^2 = 73,764$	$\Sigma xy = 4462.220$

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 4462.22 - \frac{(930)(56.69)}{12} = 68.745$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 73,764 - \frac{(930)^2}{12} = 1689$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{68.745}{1689} = .0407$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{n} = \frac{56.69}{12} - (.0407)\frac{930}{12} = 1.57$$

$$\hat{y} = 1.57 + .0407x$$

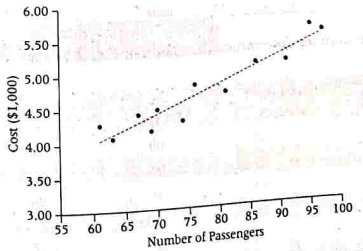


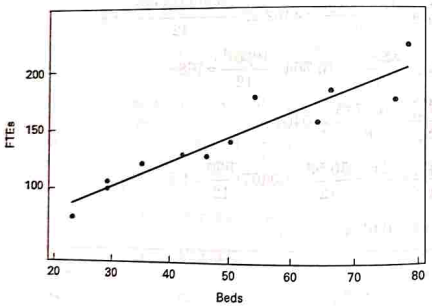
Figure 7.6: Excel Graph of Regression Line for the Airline Cost Example

DEMONSTRATION PROBLEM 7.1

A specialist in hospital administration stated that the number of FTEs (full-time employees) in a hospital can be estimated by counting the number of beds in the hospital (a common measure of hospital size). A healthcare business researcher decided to develop a regression model in an attempt to predict the number of FTEs of a hospital by the number of beds. She surveyed 12 hospitals and obtained the following data. The data are presented in sequence, according to the number of beds.

Number of Beds	FTEs	Number of Beds	FTEs
23	69	50	138
29	95	54	178
29	102	64	156
35	118	66	184
42	126	76	176
46	125	78	225

Solution: The following graph is a scatter plot of these data. Note the linear appearance of the data.



Next, the researcher determined the values of Σx, Σy, Σx^2, and Σxy.

Hospital	Number of Beds x	FTEs y	x^2	xy
1	23	69	529	1,587
2	29	95	841	2,755
3	29	102	841	2,958
4	35	118	1,225	4,130
5	42	126	1,764	5,292
6	46	125	2,116	5,750
7	50	138	2,500	6,900
8	54	178	2,916	9,612
9	64	156	4,096	9,984
10	66	184	4,356	12,144
11	76	176	5,776	13,376
12	78	225	6,084	17,550
	$\Sigma x = 592$	$\Sigma y = 1{,}692$	$\Sigma x^2 = 33{,}044$	$\Sigma xy = 92{,}038$

(Continued)

Using these values, the researcher solved for the sample slope (b_1) and the sample y-intercept (b_0).

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 92{,}038 - \frac{(592)(1692)}{12} = 8566$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33{,}044 - \frac{(592)^2}{12} = 3838.667$$

$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{8566}{3838.667} = 2.232$$

$$b_0 = \frac{\Sigma y}{n} - b_1\frac{\Sigma x}{12} = \frac{1692}{12} - (2.232)\frac{592}{12} = 30.888$$

The least squares equation of the regression line is

$$\hat{y} = 30.888 + 2.232x$$

The slope of the line, $b_1 = 2.232$, means that for every unit increase of x (every additional bed), y (number of FTEs) is predicted to increase by 2.232. Even though the y-intercept helps the researcher sketch the graph of the line by being one of the points on the line (0, 30.888), it has limited usefulness in terms of this solution because $x = 0$ denotes a hospital with no beds. On the other hand, it could be interpreted that a hospital has to have at least 31 FTEs to open its doors even with no patients—a sort of "fixed cost" of personnel.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

11. In linear regression, _____ are mathematical models that produce an "exact" output for a given input.
12. _____ is a process whereby a regression model is developed by producing the minimum sum of the squared error values.
13. In regression analysis, researchers use the slope-intercept equation which is the equation of _____.

State whether the following statements are true/false:

14. The horizontal distance from each point to the line is the error of the prediction.

SELF ASSESSMENT QUESTIONS

15. A probabilistic model in linear regression is one that includes an error term that allows the y values to vary for any given value of x.

ACTIVITY

https://vincentarelbundock.github.io/Rdatasets/csv/DAAG/ACF1.csv. Calculate the regression line (not by Excel data analysis tools). The regression line should be predicting "end time" through count variable. Also calculate the correlation between two variables.

7.5 STANDARD ERROR OF THE ESTIMATE

Residuals represent errors of estimation for individual points. With large samples of data, residual computations become laborious. Even with computers, a researcher sometimes has difficulty working through pages of residuals in an effort to understand the error of the regression model. An alternative way of examining the error of the model is the standard error of the estimate, which provides a single measurement of the regression error.

Because the sum of the residuals is zero, attempting to determine the total amount of error by summing the residuals is fruitless. This zero-sum characteristic of residuals can be avoided by squaring the residuals and then summing them.

Table 7.5 contains the airline cost data from Table 7.3, along with the residuals and the residuals squared. The *total of the residuals squared* column is called the **sum of squares of error (SSE)**.

TABLE 7.5: DETERMINING SSE FOR THE AIRLINE COST EXAMPLE

Number of Passengers x	Cost ($1,000) y	Residual $y - \hat{y}$	$(y - \hat{y})^2$
61	4.280	.227	.05153
63	4.080	−.054	.00292
67	4.420	.123	.01513
69	4.170	−.208	.04326
70	4.480	.061	.00372
74	4.300	−.282	.07952
76	4.820	.157	.02465
81	4.700	−.167	.02789
86	5.110	.040	.00160
91	5.130	−.144	.02074
95	5.640	.204	.04162
97	5.560	.042	.00176
		$\Sigma(y - \hat{y}) = -.001$	$\Sigma(y - \hat{y})^2 = .31434$
Sum of squares of error = SSE = .31434			

Sum of Squares of Error

$$SSE = \Sigma(y - \hat{y})^2$$

In theory, infinitely many lines can be fit to a sample of points. However, Formulas 7.2 and 7.4 produce a line of best fit for which the SSE is the smallest for any line that can be fit to the sample data. This result is guaranteed, because Formulas 7.2 and 7.4 are derived from calculus to minimize SSE. For this reason, the regression process used in this chapter is called *least squares* regression.

A computational version of the equation for computing SSE is less meaningful in terms of interpretation than $\Sigma(y - \hat{y})^2$ but it is usually easier to compute. The computational formula for SSE follows.

Computational Formula for SSE

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$$

For the airline cost example,

$$\Sigma y^2 = \Sigma[(4.280)^2 + (4.080)^2 + (4.420)^2 + (4.170)^2 + (4.480)^2 + (4.300)^2 + (4.820)^2 + (4.700)^2 + (5.110)^2 + (5.130)^2 + (5.640)^2 + (5.560)^2] = 270.9251$$

$$b_0 = 1.5697928$$

$$b_1 = .0407016$$

$$\Sigma y = 56.69$$

$$\Sigma xy = 4462.22$$

$$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$$

$$= 270.9251 - (1.5697928)(56.69) - (.0407016)(4462.22) = .31405$$

The slight discrepancy between this value and the value computed in Table 7.5 is due to rounding error.

The sum of squares error is in part a function of the number of pairs of data being used to compute the sum, which lessens the value of SSE as a measurement of error. A more useful measurement of error is the standard error of the estimate. The **standard error of the estimate**, denoted s_e, is *a standard deviation of the error of the regression model.* The standard error of the estimate follows.

Standard Error of the Estimate

$$s_e = \sqrt{\frac{SSE}{n-2}}$$

The standard error of the estimate for the airline cost example is

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{.31434}{10}} = .1773$$

How is the standard error of the estimate used? As previously mentioned, the standard error of the estimate is a standard deviation of error. Recall that if data are approximately normally distributed, the empirical rule states that about 68%

of all values are within $\mu \pm 1\sigma$ and that about 95% of all values are within $\mu \pm 2\sigma$. One of the assumptions for regression states that for a given x the error terms are normally distributed. Because the error terms are normally distributed, s_e is the standard deviation of error, and the average error is zero, approximately 68% of the error values (residuals) should be within $0 \pm 1s_e$ and 95% of the error values (residuals) should be within $0 \pm 2s_e$. By having knowledge of the variables being studied and by examining the value of s_e, the researcher can often make a judgment about the fit of the regression model to the data by using s_e. How can the s_e value for the airline cost example be interpreted?

The regression model here is used to predict airline cost by number of passengers. Note that the range of the airline cost data in Table 7.3 is from 4.08 to 5.64 ($4,080 to $5,640). The regression model for the data yields an s_e of .1773. An interpretation of s_e is that the standard deviation of error for the airline cost example is $177.30. If the error terms were normally distributed about the given values of x, approximately 68% of the error terms would be within ±$177.30 and 95% would be within ±2($177.30) = ±$354.60. Examination of the residuals reveals that 8 out of 12 (67%) of the residuals are within $\pm 1s_e$ and 100% of the residuals are within $2s_e$. The standard error of the estimate provides a single measure of error, which, if the researcher has enough background in the area being analyzed, can be used to understand the magnitude of errors in the model. In addition, some researchers use the standard error of the estimate to identify outliers. They do so by looking for data that are outside $\pm 2s_e$ or $\pm 3s_e$.

DEMONSTRATION PROBLEM 7.2

Compute the sum of squares of error and the standard error of the estimate for Demonstration Problem 7.1, in which a regression model was developed to predict the number of FTEs at a hospital by the number of beds.

Solution:

Hospital	Number of Beds x	FTES y	Residuals $y - \hat{y}$	$(y - \hat{y})^2$
1	23	69	−13.22	174.77
2	29	95	−.62	0.38
3	29	102	6.38	40.70
4	35	118	8.99	80.82
5	42	126	1.37	1.88
6	46	125	−8.56	73.27
7	50	138	−4.49	20.16
8	54	178	26.58	706.50
9	64	156	−17.74	314.71
10	66	184	5.80	33.64
11	76	176	−24.52	601.23
12	78	225	20.02	400.80
	$\Sigma x = 592$	$\Sigma y = 1692$	$\Sigma(y - \hat{y}) = -.01$	$\Sigma(y - \hat{y})^2 = 2448.86$

$$\text{SSE} = 2448.86$$

$$S_e = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{2448.86}{10}} = 15.65$$

The standard error of the estimate is 15.65 FTEs. An examination of the residuals for this problem reveals that 8 of 12 (67%) are within $\pm 1s_e$ and 100% are within $\pm 2s_e$. Is this size of error acceptable? Hospital administrators probably can best answer that question.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

16. Estimation of the _____ provides a single measurement of the regression error.
17. The total of the residuals squared column is called the _____.

State whether the following statements are true/false:

18. In regression, the line of best fit is the line for which the SSE is the highest.
19. Some researchers use the standard error of the estimate to identify outliers.
20. Standard error of the estimate is a standard deviation of error.

ACTIVITY

For the previous activity, you have drawn the regression lines and residuals. Now without the help of any software, try to calculate standard error and calculate SSE. Write these values in the notebook and try to visualise the line and residuals.

7.6 COEFFICIENT OF DETERMINATION

A widely used measure of fit for regression models is the **coefficient of determination**, or r^2. The coefficient of determination is *the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).*

The coefficient of determination ranges from 0 to 1. An r^2 of zero means that the predictor accounts for none of the variability of the dependent variable and that there is no regression prediction of y by x. An r^2 of 1 means perfect prediction of y by x and that 100% of the variability of y is accounted for by x. Of course, most r^2 values are between the extremes. The researcher must interpret whether a particular r^2 is high or low, depending on the use of the model and the context within which the model was developed.

In exploratory research where the variables are less understood, low values of r^2 are likely to be more acceptable than they are in areas of research where the parameters are more developed and understood. One NASA researcher who uses vehicular weight to predict mission cost searches for the regression models to have an r^2 of .90 or higher. However, a business researcher who is trying to develop a model to predict the motivation level of employees might be pleased to get an r^2 near .50 in the initial research.

The dependent variable, y, being predicted in a regression model has a variation that is measured by the sum of squares of y (SS_{yy}):

$$SS_{yy} = \Sigma(y - \bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

and is the sum of the squared deviations of the y values from the mean value of y. This variation can be broken into two additive variations: the *explained variation*, measured by the sum of squares of regression (SSR), and the *unexplained variation*, measured by the sum of squares of error (SSE). This relationship can be expressed in equation form as

$$SS_{yy} = SSR + SSE$$

If each term in the equation is divided by SS_{yy}, the resulting equation is

$$1 = \frac{SSR}{SS_{yy}} + \frac{SSE}{SS_{yy}}$$

The term r^2 is the proportion of the y variability that is explained by the regression model and represented here as

$$r^2 = \frac{SSR}{SS_{yy}}$$

Substituting this equation into the preceding relationship gives

$$1 = r^2 + \frac{SSE}{SS_{yy}}$$

Solving for r^2 yields Formula 7.5.

Coefficient of Determination

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{SSE}{\Sigma y^2 - \frac{(\Sigma y)^2}{n}} \qquad (7.5)$$

Note: $0 \le r^2 \le 1$

The value of r^2 for the airline cost example is solved as follows:

$$SSE = .31434$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 270.9251 - \frac{(56.69)^2}{12} = 3.11209$$

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{.31434}{3.11209} = .899$$

That is, 89.9% of the variability of the cost of flying a Boeing 737 airplane on a commercial flight is explained by variations in the number of passengers. This result also means that 10.1% of the variance in airline flight cost, y, is unaccounted for by x or unexplained by the regression model.

The coefficient of determination can be solved for directly by using

$$r^2 = \frac{SSR}{SS_{yy}}$$

It can be shown through algebra that

$$SSR = b_1^2 SS_{xx}$$

From this equation, a computational formula for r^2 can be developed.

Computational Formula for r^2

$$r^2 = \frac{b_1^2 SS_{xx}}{SS_{yy}}$$

For the airline cost example, $b_1 = .0407016$, $SS_{xx} = 1689$, and $SS_{yy} = 3.11209$. Using the computational formula for r^2 yields

$$r^2 = \frac{(.0407016)^2(1689)}{3.11209} = .899$$

DEMONSTRATION PROBLEM 7.3

Compute the coefficient of determination (r^2) for Demonstration Problem 7.1, in which a regression model was developed to predict the number of FTEs of a hospital by the number of beds.

Solution:

$$SSE = 2448.86$$

$$SS_{yy} = 260{,}136 - \frac{(1692)^2}{12} = 21{,}564$$

$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{2448.86}{21{,}564} = .886$$

This regression model accounts for 88.6% of the variance in FTEs, leaving only 11.4% unexplained variance.

Using $SS_{xx} = 3838.667$ and $b_1 = 2.232$ from Demonstration Problem 7.1, we can solve for r^2 with the computational formula:

$$r^2 = \frac{b_1^2 SS_{xx}}{SS_{yy}} = \frac{(2.232)^2(3838.667)}{21{,}564} = .886$$

Relationship Between r and r^2

Is r, the coefficient of correlation (introduced in Section 7.1), related to r^2, the coefficient of determination in linear regression? The answer is yes: r^2 equals $(r)^2$. The coefficient of determination is the square of the coefficient of correlation. In Demonstration Problem 7.1, a regression model was developed to predict FTEs by number of hospital beds. The r^2 value for the model was .886. Taking the square root of this value yields $r = .941$, which is the correlation between the sample number of beds and FTEs. A word of caution here:

Because r^2 is always positive, solving for r by taking $\sqrt{r^2}$ gives the correct magnitude of r but may give the wrong sign. The researcher must examine the sign of the slope of the regression line to determine whether a positive or negative relationship exists between the variables and then assign the appropriate sign to the correlation value.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

21. In _____research, low values of r^2 are likely to be more acceptable.
22. The coefficient of determination is the square of the _____.

State whether the following statements are true/false:

23. For coefficient of determination, an r^2 of 1 means perfect prediction of y by x.
24. The coefficient of determination ranges from –1 to 1.
25. Slope of the regression line is helpful to determine whether a positive or negative relationship exists between the variables.

ACTIVITY

In an umbrella making company, the sales is dependent on the advertisement they do in the year. Now they have made the equation of line with the help of one statistician. He gave them one model and will be going out of station for some work. The manager wants your help to understand the below-mentioned point.

Somebody told him that it is correlation between two values. Can you help him to calculate it? How can he calculate it? With linear regression, is it the coefficient of determination that is also equal to the square of the correlation between x and y scores?

For the previous month, the value of R^2 was 0. Explain to him what does that mean. Is it good for the company or not?

Last year, for the current month, the value of R^2 was 1. Explain the effect of this clearly to him.

7.7 INTERPRETING THE OUTPUT

Although manual computations can be done, most regression problems are analyzed by using a computer. In this section, computer output from Excel will be presented and discussed.

The Excel regression output, shown in **Figure 7.7** for Demonstration Problem 7.1, has all the essential regression features are present. The regression equation is found under Coefficients at the bottom of ANOVA. The slope or coefficient of x is 2.2315 and the y-intercept is 30.9125. The standard error of the estimate for the hospital problem is given as the fourth statistic under Regression Statistics at the top of the output, Standard Error = 15.6491. The r^2 value is given as 0.886 on the second line. The t test for the slope is found under t Stat near the bottom of the ANOVA section on the "Number of Beds" (x variable) row, $t = 8.83$. Adjacent to the t Stat is the p-value, which is the probability of the t statistic occurring by chance if the null hypothesis is true. For this slope, the probability shown is 0.000005. The ANOVA table is in the middle of the output with the F value having the same probability as the t statistic, 0.000005, and equaling t^2. The predicted values and the residuals are shown in the Residual Output section.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.942
R Square	0.886
Adjusted R Square	0.875
Standard Error	15.6491
Observations	12

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	19115.06322	19115.06	78.05	0.000005
Residual	10	2448.94	244.89		
Total	11	21564			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	30.9125	13.2542	2.33	0.041888
Number of Beds	2.2315	0.2526	8.83	0.000005

RESIDUAL OUTPUT

Observation	*Predicted FTEs*	*Residuals*
1	82.237	–13.237
2	95.626	–0.626
3	95.626	6.374
4	109.015	8.985
5	124.636	1.364
6	133.562	–8.562
7	142.488	–4.488
8	151.414	26.586
9	173.729	–17.729
10	178.192	5.808
11	200.507	–24.507
12	204.970	20.030

Figure 7.7: Excel Regression Output for Demonstration Problem 7.1

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

26. In many of the software, we find SS. Here SS stands for _____.
27. For ANOVA table, df stands for _____.

State whether the following statements are true/false:

28. t stat is the t test for the slope.
29. Most of the regression problems cannot be done through computers.
30. We can make regression equation through coefficient.

ACTIVITY

Download the financial statement of BHEL Co. Ltd. for the last 10 years from www.bseindia.com. Calculate the regression where sales and profit are independent and dependent variable, respectively.

7.8 SUMMARY

- Correlation measures the degree of relatedness of variables. The most well-known measure of correlation is the Pearson product-moment coefficient of correlation, r. This value ranges from -1 to 0 to $+1$. An r value of $+1$ is perfect positive correlation and an r value of -1 is perfect negative correlation. Positive correlation means that as one variable increases in value, the other variable tends to increase. Negative correlation means that as one variable increases in value, the other variable tends to decrease. For r values near zero, little or no correlation is present.
- Regression is a procedure that produces a mathematical model (function) that can be used to predict one variable by other variables. Simple regression is bivariate (two variables) and linear (only a line fit is attempted). Simple regression analysis produces a model that attempts to predict a y variable, referred to as the dependent variable, by an x variable, referred to as the independent variable. The general form of the equation of the simple regression line is the slope-intercept equation of a line. The equation of the simple regression model consists of a slope of the line as a coefficient of x and a y-intercept value as a constant.
- After the equation of the line has been developed, several statistics are available that can be used to determine how well the line fits the data. Using the historical data values of x, predicted values of y (denoted as $\hat{y}$) can be calculated by inserting values of x into the regression equation. The predicted values can then be compared to the actual values of y to determine how well the regression equation fits the known data. The difference between a specific y value and its associated predicted y value is called the residual or error of prediction. Examination of the residuals can offer insight into the magnitude of the errors produced by a model. In addition, residual analysis can be used to help determine whether the assumptions underlying the regression analysis have been met.
- A single value of error measurement called the standard error of the estimate, s_e, can be computed. The standard error of the estimate is the standard deviation of error of a model. The value of s_e can be used as a single guide to the magnitude of the error produced by the regression model as opposed to examining all the residuals.
- Another widely used statistic for testing the strength of a regression model is r^2, or the coefficient of determination. The coefficient of determination is the proportion of total variance of the y variable accounted for or predicted by x. The coefficient of determination ranges from 0 to 1. The higher the r^2 is, the stronger is the predictability of the model.
- Testing to determine whether the slope of the regression line is different from zero is another way to judge the fit of the regression model to the data. If the population slope of the regression line is not different from zero, the regression model is not adding significant predictability to the dependent variable. A t statistic is used to test the significance of the slope. The overall significance of the regression model can be tested using an F statistic. In simple regression, because only one predictor is present, this test accomplishes the same thing as the t test of the slope and $F = t^2$.
- One of the most prevalent uses of a regression model is to predict the values of y for given values of x. Recognizing that the predicted value is often not the same as the actual value, a confidence interval has been developed to yield a range within which the mean y value for a given x should fall. A prediction interval for a single y value for a given x value also is specified. This second interval is wider because it allows for the wide diversity of individual values, whereas the confidence interval for the mean y value reflects only the range of average y values for a given x.
- Time-series data are data that are gathered over a period of time at regular intervals. Developing the equation of a forecasting trend line for time-series data is a special case of simple regression analysis where the time factor is the predictor variable. The time variable can be in units of years, months, weeks, quarters, and others.

KEY WORDS

1. **Coefficient of determination:** The coefficient of determination is the proportion of variability of the dependent variable (y) accounted for or explained by the independent variable (x).
2. **Dependent variable:** In simple regression, the variable to be predicted is called the dependent variable and is designated as y.
3. **Independent variable:** The predictor is called the independent variable, or explanatory variable, and is designated as x.
4. **Least squares analysis:** Least squares analysis is a process whereby a regression model is developed by producing the minimum sum of the squared error values.
5. **Outliers:** Outliers are data points that lie apart from the rest of the points.
6. **Probabilistic model:** Probabilistic model is one that includes an error term that allows the y values to vary for any given value of x.
7. **Regression analysis:** Regression analysis is the process of constructing a mathematical model or function that can be used to predict or determine one variable by another variable or other variables.
8. **Scatter plot:** Scatter plot is a plot in which the values of two variables are plotted along two axes, the pattern of the resulting points revealing any correlation present. It is useful for regression analysis also.
9. **Simple regression:** The most elementary regression model is called simple regression or bivariate regression involving two variables in which one variable is predicted by another variable
10. **Standard error of the estimate (s_e):** The standard error of the estimate, denoted s_e, is a standard deviation of the error of the regression model.
11. **Sum of squares:** The total of the residuals squared is called the sum of squares of error (SSE).

7.9 DESCRIPTIVE QUESTIONS

7.1. Determine the value of the coefficient of correlation, r, for the following data.

X	4	6	7	11	14	17	21
Y	18	12	13	8	7	7	4

7.2. Determine the value of r for the following data.

X	158	296	87	110	436
Y	349	510	301	322	550

7.3. The following data are the claims (in $ millions) for BlueCross BlueShield benefits for nine states, along with the surplus (in $ millions) that the company had in assets in those states.

State	Claims	Surplus
Alabama	1,425	277
Colorado	273	100
Florida	915	120
Illinois	1,687	259
Maine	234	40
Montana	142	25
North Dakota	259	57
Oklahoma	258	31
Texas	894	141

Use the data to compute a correlation coefficient, r, to determine the correlation between claims and surplus.

7.4. The National Safety Council released the following data on the incidence rates for fatal or lost-worktime injuries per 100 employees for several industries in three recent years.

Industry	Year 1	Year 2	Year 3
Textile	.46	.48	.69
Chemical	.52	.62	.63
Communication	.90	.72	.81
Machinery	1.50	1.74	2.10
Services	2.89	2.03	2.46
Nonferrous metals	1.80	1.92	2.00
Food	3.29	3.18	3.17
Government	5.73	4.43	4.00

Compute r for each pair of years and determine which years are most highly correlated.

7.5. Sketch a scatter plot from the following data, and determine the equation of the regression line.

x	12	21	28	8	20
y	17	15	22	19	24

7.6. A corporation owns several companies. The strategic planner for the corporation believes dollars spent on advertising can to some extent be a predictor of total sales dollars. As an aid in long-term planning, she gathers the following sales and advertising information from several of the companies for 2016 ($ millions).

Advertising	Sales
12.5	148
3.7	55
21.6	338
60.0	994
37.6	541
6.1	89
16.8	126
41.2	379

Develop the equation of the simple regression line to predict sales from advertising expenditures using these data.

7.7. Is it possible to predict the annual number of business bankruptcies by the number of firm births (business starts) in the United States? The following data, published by the U.S. Small Business Administration, Office of Advocacy, are pairs of the number of business bankruptcies (1000s) and the number of firm births (10,000s) for a six-year period. Use these data to develop the equation of the regression model to predict the number of business bankruptcies by the number of firm births. Discuss the meaning of the slope.

Business Bankruptcies (1000)	Firm Births (10,000)
34.3	58.1
35.0	55.4
38.5	57.0
40.1	58.5
35.5	57.4
37.9	58.0

7.8. The Conference Board produces a Consumer Confidence Index (CCI) that reflects people's feelings about general business conditions, employment opportunities, and their own income prospects. Some researchers may feel that consumer confidence is a function of the median household income. Shown here are the CCIs for nine years and the median household incomes for the same nine years published by the U.S. Census Bureau. Determine the equation of the regression line

to predict the CCI from the median household income. Compute the standard error of the estimate for this model. Compute the value of r. Does median household income appear to be a good predictor of the CCI? Why or why not?

CCI	Median Household Income ($1,000)
116.8	37.415
91.5	36.770
68.5	35.501
61.6	35.047
65.9	34.700
90.6	34.942
100.0	35.887
104.6	36.306
125.4	37.005

7.10 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

7.1. $\Sigma x = 80$ $\Sigma x^2 = 1{,}148$ $\Sigma y = 69$

$\Sigma y^2 = 815$ $\Sigma xy = 624$ $n = 7$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

$$r = \frac{624 - \frac{(80)(69)}{7}}{\sqrt{\left[1{,}148 - \frac{(80)^2}{7}\right]\left[815 - \frac{(69)^2}{7}\right]}} = \frac{-164.571}{\sqrt{(233.714)(134.857)}}$$

$$r = \frac{-164.571}{177.533} = \mathbf{-0.927}$$

7.2. $\Sigma x = 1{,}087$ $\Sigma x^2 = 322{,}345$ $\Sigma y = 2{,}032$

$\Sigma y^2 = 878{,}686$ $\Sigma xy = 507{,}509$ $n = 5$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

$$r = \frac{507{,}509 - \frac{(1{,}087)(2{,}032)}{5}}{\sqrt{\left[322{,}345 - \frac{(1{,}087)^2}{5}\right]\left[878{,}686 - \frac{(2{,}032)^2}{5}\right]}}$$

$$r = \frac{65{,}752.2}{\sqrt{(86{,}031.2)(52{,}881.2)}} = \frac{65{,}752.2}{67{,}449.5} = \mathbf{.975}$$

7.3. $\Sigma x = 6{,}087$ $\Sigma x^2 = 6{,}796{,}149$

$\Sigma y = 1{,}050$ $\Sigma y^2 = 194{,}526$

$\Sigma xy = 1{,}130{,}483$ $n = 9$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

$$r = \frac{1{,}130{,}483 - \frac{(6{,}087)(1{,}050)}{9}}{\sqrt{\left[6{,}796{,}149 - \frac{(6{,}087)^2}{9}\right]\left[194{,}526 - \frac{(1{,}050)^2}{9}\right]}}$$

$$r = \frac{420{,}333}{\sqrt{(2{,}679{,}308)(72{,}026)}} = \frac{420{,}333}{439{,}294.705} = \mathbf{.957}$$

7.4. Correlation between Year 1 and Year 2:

$\Sigma x = 17.09$ $\Sigma x^2 = 58.7911$

$\Sigma y = 15.12$ $\Sigma y^2 = 41.7054$

$\Sigma xy = 48.97$ $n = 8$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}} =$$

$$r = \frac{48.97 - \frac{(17.09)(15.12)}{8}}{\sqrt{\left[58.7911 - \frac{(17.09)^2}{8}\right]\left[41.7054 - \frac{(15.12)^2}{8}\right]}} =$$

$$r = \frac{16.6699}{\sqrt{(22.28259)(13.1286)}} = \frac{16.6699}{17.1038} = \mathbf{.975}$$

Correlation between Year 2 and Year 3:

$\Sigma x = 15.12$ $\Sigma x^2 = 41.7054$

$\Sigma y = 15.86$ $\Sigma y^2 = 42.0396$

$\Sigma xy = 41.5934$ $n = 8$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

$$r = \frac{41.5934 - \frac{(15.12)(15.86)}{8}}{\sqrt{\left[41.7054 - \frac{(15.12)^2}{8}\right]\left[42.0396 - \frac{(15.86)^2}{8}\right]}}$$

$$r = \frac{11.618}{\sqrt{(13.1286)(10.59715)}} = \frac{11.618}{11.795} = .985$$

Correlation between Year 1 and Year 3:

$\Sigma x = 17.09$ $\Sigma x^2 = 58.7911$

$\Sigma y = 15.86$ $\Sigma y^2 = 42.0396$

$\Sigma xy = 48.5827$ $n = 8$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum y)^2}{n}\right]}}$$

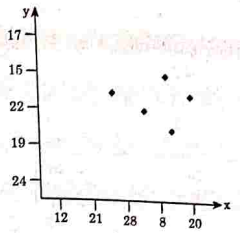
$$r = \frac{48.5827 - \frac{(17.09)(15.86)}{8}}{\sqrt{\left[58.7911 - \frac{(17.09)^2}{8}\right]\left[42.0396 - \frac{(15.86)^2}{8}\right]}}$$

$$r = \frac{14.702}{\sqrt{(22.2826)(10.5972)}} = \frac{14.702}{15.367} = .957$$

The years 2 and 3 are the most correlated with $r = .985$.

7.5.

x	y
12	17
21	15
28	22
8	19
20	24



$\Sigma x = 89$ $\Sigma y = 97$ $\Sigma xy = 1{,}767$

$\Sigma x^2 = 1{,}833$ $\Sigma y^2 = 1{,}935$ $n = 5$

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{1{,}767 - \frac{(89)(97)}{5}}{1{,}833 - \frac{(89)^2}{5}} = \mathbf{0.162}$$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{n} = \frac{97}{5} - 0.162\frac{89}{5} = \mathbf{16.51}$$

$$\hat{y} = \mathbf{16.51 + 0.162\,x}$$

7.6.

(Advertising) x	(Sales) y
12.5	148
3.7	55
21.6	338
60.0	994
37.6	541
6.1	89
16.8	126
41.2	379

$\Sigma x = 199.5$ $\Sigma y = 2{,}670$ $\Sigma xy = 107{,}610.4$

$\Sigma x^2 = 7{,}667.15$ $\Sigma y^2 = 1{,}587{,}328$ $n = 8$

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{107{,}610.4 - \frac{(199.5)(2{,}670)}{8}}{7{,}667.15 - \frac{(199.5)^2}{8}} = \mathbf{15.240}$$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{n} = \frac{2{,}670}{8} - 15.24\frac{199.5}{8} = \mathbf{-46.292}$$

$$\hat{y} = \mathbf{-46.292 + 15.240\,x}$$

7.7.

Bankruptcies (y)	Firm Births (x)
34.3	58.1
35.0	55.4
38.5	57.0
40.1	58.5
35.5	57.4
37.9	58.0

$\Sigma x = 344.4$ $\Sigma y = 221.3$ $\Sigma x^2 = 19{,}774.78$

$\Sigma y^2 = 8188.41$ $\Sigma xy = 12{,}708.08$ $n = 6$

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{12{,}708.08 - \frac{(344.4)(221.3)}{6}}{19{,}774.78 - \frac{(344.4)^2}{6}}$$

$b_1 = 0.878$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{n} = \frac{221.3}{6} - (0.878)\frac{344.4}{6} = -13.503$$

$\hat{y} = -13.503 + 0.878\,x$

7.8.

CCI	Median Income
116.8	37.415
91.5	36.770
68.5	35.501
61.6	35.047
65.9	34.700
90.6	34.942
100.0	35.887
104.6	36.306
125.4	37.005

$\Sigma x = 323.573$ $\Sigma y = 824.9$ $\Sigma x^2 = 11{,}640.93413$

$\Sigma y^2 = 79{,}718.79$ $\Sigma xy = 29{,}804.4505$ $n = 9$

$$b_1 = \frac{SS_{xy}}{SS_x} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{29{,}804.4505 - \frac{(323.573)(824.9)}{9}}{11{,}640.93413 - \frac{(323.573)^2}{9}}$$

$b_1 = 19.2204$

$$b_0 = \frac{\sum y}{n} - b_1 \frac{\sum x}{n} = \frac{824.9}{9} - (19.2204)\frac{323.573}{9} = -599.3674$$

$\hat{y} = -599.3674 + 19.2204\,x$

$SSE = \Sigma y^2 - b_0 \Sigma y - b_1 \Sigma xy$

$= 79{,}718.79 - (-599.3674)(824.9) - 19.2204(29{,}804.4505) = 1283.13435$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{1283.13435}{7}} = 13.539$$

$$r^2 = 1 - \frac{SSE}{\sum y^2 - \frac{(\sum y)^2}{n}} = 1 - \frac{1283.13435}{79{,}718.79 - \frac{(824.9)^2}{9}} = .688$$

7.11 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
7.2 Correlation	1.	r
	2.	two
	3.	negative
	4.	False
	5.	False
7.3 Introduction to Simple Regression Analysis	6.	dependent variable
	7.	scatter
	8.	False
	9.	True
	10.	True
7.4 Determining the Equation of the Regression Line	11.	deterministic models
	12.	Least squares analysis
	13.	Line
	14.	False
	15.	True
7.5 Standard Error of the Estimate	16.	standard error
	17.	sum of squares of error
	18.	False
	19.	True
	20.	True
7.6 Coefficient of Determination	21.	exploratory
	22.	coefficient of correlation
	23.	True

Topics	Q. No.	Answers
	24.	False
	25.	True
7.8 Interpreting the Output	26.	sum of Square
	27.	degree of freedom
	28.	True
	29.	False
	30.	True

8

CHAPTER

MULTIPLE REGRESSION ANALYSIS

CONTENTS

LEARNING OBJECTIVES

This chapter presents the potential of multiple regression analysis as a tool in business decision making and its applications, thereby enabling you to:

- Elucidate how, by extending the simple regression model to a multiple regression model with two independent variables, it is possible to determine the multiple regression equation for any number of unknowns.
- Analyse significance tests of both the overall regression model and the regression coefficients.
- Compute the residual, standard error of the estimate, coefficient of multiple determination, and adjusted coefficient of multiple determination of a regression model.

8.1 INTRODUCTION

Simple regression analysis (discussed in Chapter 8) is bivariate linear regression in which one **dependent variable**, y, is predicted by one **independent variable**, x. Examples of simple regression applications include models to predict retail sales by population density, Dow Jones averages by prime interest rates, crude oil production by energy consumption, and CEO compensation by quarterly sales. However, in many cases, other independent variables, taken in conjunction with these variables, can make the regression model a better fit in predicting the dependent variable. For example, sales could be predicted by the size of store and number of competitors in addition to population density. A model to predict the Dow Jones average of 30 industrials could include, in addition to the prime interest rate, such predictors as yesterday's volume, the bond interest rate, and the producer price index. A model to predict CEO compensation could be developed by using variables such as company earnings per share, age of CEO, and size of company, in addition to quarterly sales. A model could perhaps be developed to predict the cost of outsourcing by such variables as unit price, export taxes, cost of money, damage in transit, and other factors. Each of these examples contains only one dependent variable, y, as with simple regression analysis. However, multiple independent variables, x (predictors), are involved. *Regression analysis with two or more independent variables or with at least one nonlinear predictor* is called **multiple regression** analysis.

8.2 THE MULTIPLE REGRESSION MODEL

Multiple regression analysis is similar in principle to simple regression analysis. However, it is more complex conceptually and computationally. The equation of the probabilistic simple regression model is

$$y = \beta_0 + \beta_1 x + \epsilon$$

where

y = the value of the dependent variable
β_0 = the population y intercept
β_1 = the population slope
ϵ = the error of prediction

Extending this notion to multiple regression gives the general equation for the probabilistic multiple regression model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \epsilon$$

where

y = the value of the dependent variable
β_0 = the regression constant
β_1 = the partial regression coefficient for independent variable 1
β_2 = the partial regression coefficient for independent variable 2
β_3 = the partial regression coefficient for independent variable 3
β_k = the partial regression coefficient for independent variable k
k = the number of independent variables

In multiple regression analysis, the dependent variable, y, is sometimes referred to as the **response variable**. The **partial regression coefficient** of an independent variable, β_1, *represents the increase that will occur in the value of y from a one-unit increase in that independent variable if all other variables are held constant.* The "full" (versus partial) regression coefficient of an independent variable is a coefficient obtained from the bivariate model (simple regression) in which the independent variable is the sole predictor of y. The partial regression coefficients occur because more than one predictor is included in a model. The partial regression coefficients are analogous to β_1, the slope of the simple regression model.

In actuality, the partial regression coefficients and the regression constant of a multiple regression model are population values and are unknown. In virtually all research, these values are estimated by using sample information. Shown here is the form of the equation for estimating y with sample information.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_k x_k$$

where

$\hat{y}$ = the predicted value of y
b_0 = the estimate of the regression constant
b_1 = the estimate of regression coefficient 1
b_2 = the estimate of regression coefficient 2
b_3 = the estimate of regression coefficient 3
b_k = the estimate of regression coefficient k
k = the number of independent variables

8.2.1 MULTIPLE REGRESSION MODEL WITH TWO INDEPENDENT VARIABLES (FIRST-ORDER)

The simplest multiple regression model is one constructed with two independent variables, where the highest power of either variable is 1 (first-order regression model). This regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The constant and coefficients are estimated from sample information, resulting in the following model.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Figure 8.1 is a three-dimensional graph of a series of points (x_1, x_2, y) representing values from three variables used in a multiple regression model to predict the sales price of a house by the number of square feet in the house and the age of the house. Simple regression models yield a line that is fit through data points in the xy plane. In multiple regression analysis, the resulting model produces a **response surface.** In the multiple regression model shown here with two independent first-order variables, the response surface is a **response plane.** The response plane for such a model is fit in a three-dimensional space (x_1, x_2, y).

If such a response plane is fit into the points shown in Figure 8.1, the result is the graph in **Figure 8.2.** Notice that most of the points are not on the plane. As in simple regression, an error in the fit of the model in multiple regression is usually present. The distances shown in the graph from the points to the response plane are the errors of fit, or residuals ($y - \hat{y}$). Multiple regression models with three or more independent variables involve more than three dimensions and are difficult to depict geometrically.

Observe in Figure 8.2 that the regression model attempts to fit a plane into the three-dimensional plot of points. Notice that the plane intercepts the y axis. Figure 8.2 depicts some values of y for various values of x_1 and x_2.

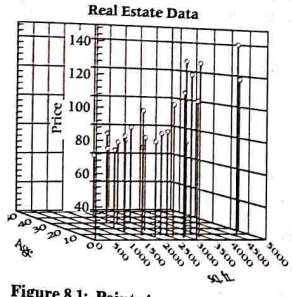


Figure 8.1: Points in a Sample Space

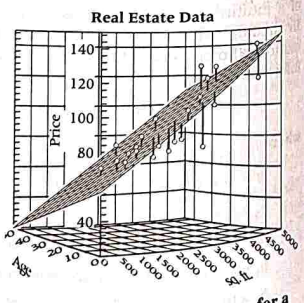


Figure 8.2: Response Plane for a First-Order Two-Predictor Multiple Regression Model

The error of the response plane (ϵ) in predicting or determining the y values is the distance from the points to the plane.

8.2.2 DETERMINING THE MULTIPLE REGRESSION EQUATION

The simple regression equations for determining the sample slope and intercept given in Chapter 8 are the result of using methods of calculus to minimize the sum of squares of error for the regression model. The procedure for developing these equations involves solving two simultaneous equations with two unknowns, b_0 and b_1. Finding the sample slope and intercept from these formulas requires the values of Σx, Σy, Σxy, and Σx^2.

The procedure for determining formulas to solve for multiple regression coefficients is similar. The formulas are established to meet an objective of *minimizing the sum of squares of error for the model.* Hence, the regression analysis shown here is referred to as **least squares analysis.** Methods of calculus are applied, resulting in $k + 1$ equations with $k + 1$ unknowns (b_0 and k values of b_i) for multiple regression analyses with k independent variables. Thus, a regression model with six independent variables will generate seven simultaneous equations with seven unknowns (b_0, b_1, b_2, b_3, b_4, b_5, b_6).

For multiple regression models with two independent variables, the result is three simultaneous equations with three unknowns (b_0, b_1, and b_2).

$$b_0 n + b_1 \Sigma x_1 + b_2 \Sigma x_2 = \Sigma y$$

$$b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 = \Sigma x_1 y$$

$$b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 = \Sigma x_2 y$$

The process of solving these equations by hand is tedious and time-consuming. Solving for the regression coefficients and regression constant in a multiple regression model with two independent variables requires Σx_1, Σx_2, Σy, Σx_1^2, Σx_2^2, $\Sigma x_1 x_2$, $\Sigma x_1 y$, and $\Sigma x_2 y$. In actuality, virtually all business researchers use computer statistical software packages to solve for the regression coefficients, the regression constant, and other pertinent information. In this chapter, we will discuss computer output and assume little or no hand calculation. The emphasis will be on the interpretation of the computer output.

8.2.3 A MULTIPLE REGRESSION MODEL

A real estate study was conducted in a small Louisiana city to determine what variables, if any, are related to the market price of a home. Several variables were explored, including the number of bedrooms, the number of bathrooms, the age of the house, the number of square feet of living space, the total number of square feet of space, and the number of garages. Suppose the researcher wants to develop a regression model to predict the market price of a home by two variables, "total number of square feet in the house" and "the age of the house." Listed in **Table 8.1** are the data for these three variables.

TABLE 8.1: REAL ESTATE DATA

Market Price ($1,000) y	Total Number of Square Feet x_1	Age of House (Years) x_2
63.0	1605	35
65.1	2489	45
69.9	1553	20
76.8	2404	32
73.9	1884	25
77.9	1558	14
74.9	1748	8
78.0	3105	10
79.0	1682	28
83.4	2470	30
79.5	1820	2
83.9	2143	6
79.7	2121	14
84.5	2485	9
96.0	2300	19
109.5	2714	4
102.5	2463	5
121.0	3076	7
104.9	3048	3
128.0	3267	6
129.0	3069	10
117.9	4765	11
140.0	4540	8

A number of statistical software packages can perform multiple regression analysis, including Excel. The output for the multiple regression analysis on the real estate data is given in **Figure 8.3**.

This output for regression analysis ends with "Regression Equation." From Figure 8.3, the regression equation for the real estate data in Table 8.1 is

$$\hat{y} = 57.4 + .0177x_1 - .666x_2$$

The regression constant, 57.4, is the y-intercept. The y-intercept is the value of $\hat{y}$ if both x_1 (number of square feet) and x_2 (age) are zero. In this example, a practical understanding of the y-intercept is meaningless. It makes little sense to say that a house containing no square feet ($x_1 = 0$) and no years of age ($x_2 = 0$) would cost $57,400. Note in Figure 8.2 that the response plane crosses the y-axis (price) at 57.4.

Regression Analysis: Price versus Square Feet, Age

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	8190	4094.9	28.63	0.000
Square Feet	1	4538	4538.5	31.73	0.000
Age	1	1222	1221.9	8.54	0.008
Error	20	2861	143.1		
Total	22	11051			

Model Summary

S	R-sq	R-sq(adj)
11.9604	74.11%	71.52%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	57.4	10.0	5.73	0.000
Square Feet	0.0177	0.0032	5.63	0.000
Age	−0.666	0.228	−2.92	0.008

Regression Equation

Price = 57.4 + 0.0177 Square Feet − 0.666 Age

Figure 8.3: Output of Regression for the Real Estate Example

The coefficient of x_1 (total number of square feet in the house) is .0177, which means that a one-unit increase in square footage would result in a predicted increase of .0177 · ($1,000) = $17.70 in the price of the home if age were held constant. All other variables being held constant, the addition of 1 square foot of space in the house results in a predicted increase of $17.70 in the price of the home.

The coefficient of x_2 (age) is −.666. The negative sign on the coefficient denotes an inverse relationship between the age of a house and the price of the house: the older the house, the lower the price. In this case, if the total number of square feet in the house is kept constant, a one-unit increase in the age of the house (1 year) will result in −.666 · ($1,000) = −$666, a predicted $666 drop in the price.

In examining the regression coefficients, it is important to remember that the independent variables are often measured in different units. It is usually not wise to compare the regression coefficients of predictors in a multiple regression model and decide that the variable with the largest regression coefficient is the best predictor. In this example, the two variables are in different units, square feet and years. Just because x_2 has the larger coefficient (.666) does not necessarily make x_2 the strongest predictor of y.

This regression model can be used to predict the price of a house in this small Louisiana city. If the house has 2500 square feet total and is 12 years old, $x_1 = 2500$ and $x_2 = 12$. Substituting these values into the regression model yields

$$\hat{y} = 57.4 + .0177x_1 - .666x_2$$
$$= 57.4 + .0177(2500) - .666(12) = 93.658$$

The predicted price of the house is $93,658. Figure 8.2 is a graph of these data with the response plane and the residual distances.

DEMONSTRATION PROBLEM 8.1

Since 1980, the prime interest rate in the United States has varied from less than 5% to over 15%. What factor in the U.S. economy seems to be related to the prime interest rate? Two possible predictors of the prime interest rate are the annual unemployment rate and the savings rate in the United States. Shown below are data for the annual prime interest rate for the even-numbered years over a 28-year period in the United States along with the annual unemployment rate and the annual average personal saving (as a percentage of disposable personal income). Use these data to develop a multiple regression model to predict the annual prime interest rate by the unemployment rate and the average personal saving. Determine the predicted prime interest rate if the unemployment rate is 6.5 and the average personal saving is 5.0.

Year	Prime Interest Rate	Unemployment Rate	Personal Saving
1986	8.33	7.0	8.2
1988	9.32	5.5	7.3
1990	10.01	5.6	7.0
1992	6.25	7.5	7.7
1994	7.15	6.1	4.8
1996	8.27	5.4	4.0
1998	8.35	4.5	4.3
2000	9.23	4.0	2.3
2002	4.67	5.8	2.4
2004	4.34	5.5	2.1
2006	7.96	4.6	0.7
2008	5.09	5.8	1.8
2010	3.25	9.6	5.8
2012	3.25	8.1	7.6
2014	3.25	6.2	4.8

Solution: The following output shows the results of analyzing the data by using the regression portion of Excel.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.820
R Square	0.672
Adjusted R Square	0.617
Standard Error	1.496
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	54.9835	27.4917	12.29	0.0012
Residual	12	26.8537	2.2378		
Total	14	81.8372			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	13.5786	1.728	7.86	0.0000
Unemployment Rates	−1.6622	0.337	−4.93	0.0003
Personal Savings	0.6586	0.199	3.31	0.0062

The regression equation is

$$\hat{y} = 13.5786 - 1.6622x_1 + 0.6586x_2$$

where:

$\hat{y}$ = prime interest rate
x_1 = unemployment rate
x_2 = personal saving

The model indicates that for every one-unit (1%) increase in the unemployment rate, the predicted prime interest rate decreases by 1.6622%, if personal saving is held constant. The model also indicates that for every one-unit (1%) increase in personal saving, the predicted prime interest rate increases by 0.6586%, if unemployment is held constant.

If the unemployment rate is 6.5 and the personal saving rate is 5.0, the predicted prime interest rate is 7.35%:

$$\hat{y} = 13.5786 - 1.6622(6.5) + 0.6586(5.0) = 6.07$$

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. Regression analysis with two or more independent variables or with at least one nonlinear predictor is called _____ regression analysis.
2. In multiple regression analysis, the dependent variable, y, is sometimes referred to as the _____ variable.
3. The _____ regression coefficient of an independent variable represents the increase that will occur in the value of y from a one-unit increase in that independent variable if all other variables are held constant.
4. The regression analysis is often referred to as _____ squares analysis.

State whether the following statements are true/false:

5. The simple regression equations for determining the sample slope and intercept are the result of using methods of calculus to minimize the sum of squares of error for the regression model.
6. The regression constant is the y-intercept.

ACTIVITY

Collect the prices of a stock of your choice listed in both BSE and NSE for 60 months, i.e., January 2012 to December 2017. Also collect the monthly Sensex and Nifty values over the same time frame. Use a computer to develop the equation of the regression model where stock price is the dependent variable and Sensex and Nifty values are independent variables.

8.3 RESIDUALS, STANDARD ERROR OF THE ESTIMATE, AND R^2

Three statistical tools for examining the strength of a regression model are the residuals, the standard error of the estimate, and the coefficient of multiple determination.

8.3.1 RESIDUALS

The **residual**, or error, of the regression model is *the difference between the y value and the predicted value, $\hat{y}$.*

$$\text{Residual} = y - \hat{y}$$

The residuals for a multiple regression model are solved for in the same manner as they are with simple regression. First, a predicted value, $\hat{y}$, is determined by entering the value for each independent variable for a given set of observations into the multiple regression equation and solving for $\hat{y}$.

Next, the value of $y - \hat{y}$ is computed for each set of observations. Shown here are the calculations for the residuals of the first set of observations from Table 8.1. The predicted value of y for $x_1 = 1605$ and $x_2 = 35$ is

$$\hat{y} = 57.4 + .0177(1605) - .666(35) = 62.499$$

Actual value of $y = 63.0$

$$\text{Residual} = y - \hat{y} = 63.0 - 62.499 = 0.501$$

All residuals for the real estate data and the regression model displayed in Table 8.1 and Figure 8.3 are displayed in **Table 8.2.**

An examination of the residuals in Table 8.2 can reveal some information about the fit of the real estate regression model. The business researcher can observe the residuals and decide whether the errors are small enough to support the accuracy of the model. The house price figures are in units of $1,000. Two of the 23 residuals are more than 20.00, or more than $20,000 off

TABLE 8.2: RESIDUALS FOR THE REAL ESTATE REGRESSION MODEL

y	$\hat{y}$	$y - \hat{y}$
63.0	62.499	.501
65.1	71.485	−6.385
69.9	71.568	−1.668
76.8	78.639	−1.839
73.9	74.097	−.197
77.9	75.653	2.247
74.9	83.012	−8.112
78.0	105.699	−27.699
79.0	68.523	10.477
83.4	81.139	2.261
79.5	88.282	−8.782
83.9	91.335	−7.435
79.7	85.618	−5.918
84.5	95.391	−10.891
96.0	85.456	10.544
109.5	102.774	6.726
102.5	97.665	4.835
121.0	107.183	13.817
104.9	109.352	−4.452
128.0	111.230	16.770
129.0	105.061	23.939
117.9	134.415	−16.515
140.0	132.430	7.570

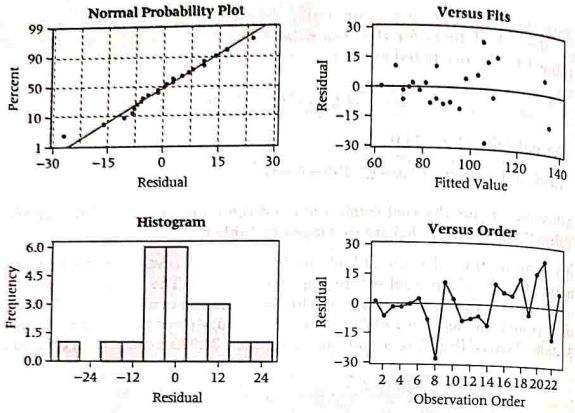


Figure 8.4: Residual Diagnosis for the Real Estate Example

in their prediction. On the other hand, two residuals are less than 1, or $1,000 off in their prediction.

Residuals are also helpful in locating outliers. **Outliers** are *data points that are apart, or far, from the mainstream of the other data.* They are sometimes data points that were mistakenly recorded or measured. Because every data point influences the regression model, outliers can exert an overly important influence on the model based on their distance from other points. In examining the residuals in Table 8.2 for outliers, the eighth residual listed is –27.699. This error indicates that the regression model was not nearly as successful in predicting house price on this particular house as it was with others (an error of more than $27,000). For whatever reason, this data point stands somewhat apart from other data points and may be considered an outlier.

Residuals are also useful in testing the assumptions underlying regression analysis. **Figure 8.4** displays diagnostic techniques for the real estate example. In the top right is a graph of the residuals.

8.3.2 SSE AND STANDARD ERROR OF THE ESTIMATE

One of the properties of a regression model is that the residuals sum to zero. In an effort to compute a single statistic that can represent the error in a regression analysis, the zero-sum property can be overcome by *squaring the residuals and then summing the squares.* Such an operation produces the **sum of squares of error (SSE).**

The formula for computing the sum of squares error (SSE) for multiple regression is the same as it is for simple regression.

$$SSE = \Sigma(y - \hat{y})^2$$

For the real estate example, SSE can be computed by squaring and summing the residuals displayed in Table 8.2.

$$\begin{aligned} SSE = [&(.501)^2 + (-6.385)^2 + (-1.668)^2 + (-1.839)^2 \\ &+ (-.197)^2 + (2.247)^2 + (-8.112)^2 + (-27.699)^2 \\ &+ (10.477)^2 + (2.261)^2 + (-8.782)^2 + (-7.435)^2 \\ &+ (-5.918)^2 + (-10.891)^2 + (10.544)^2 + (6.726)^2 \\ &+ (4.835)^2 + (13.817)^2 + (-4.452)^2 + (16.770)^2 \\ &+ (23.939)^2 + (-16.515)^2 + (7.570)^2] \\ = 2861.0 \end{aligned}$$

SSE can also be obtained directly from the multiple regression computer output by selecting the value of SS (sum of squares) listed beside error. Shown here is the ANOVA portion of the output displayed in Figure 8.3, which is the result of a multiple regression analysis model developed to predict house prices. Note that the SS for error shown in the ANOVA table equals the value of $\Sigma(y - \hat{y})^2$ just computed (2861.0).

SSE

Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	2	818.7	4094.9	28.63	.000
Error	20	2861.0	143.1		
Total	22	11050.7			

SSE has limited usage as a measure of error. However, it is a tool used to solve for other, more useful measures. One of those is the **standard error of the estimate,** s_e, which is essentially *the standard deviation of residuals (error) for the regression model.* An assumption underlying regression analysis is that the error terms are approximately normally distributed with a mean of zero. With this information and by the empirical rule, approximately 68% of the residuals should be within $\pm 1 s_e$ and 95% should be within $\pm 2 s_e$. This property makes the standard error of the estimate a useful tool in estimating how accurately a regression model is fitting the data.

The standard error of the estimate is computed by dividing SSE by the degrees of freedom of error for the model and taking the square root.

$$s_e = \sqrt{\frac{SSE}{n-k-1}}$$

where

n = number of observations

k = number of independent variables

The value of s_e can be computed for the real estate example as follows:

$$s_e = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{2861}{23-2-1}} = 11.96$$

The standard error of the estimate, s_e, is usually given as standard output from regression analysis by computer software packages. The output displayed in Figure 8.3 contains the standard error of the estimate for the real estate example.

$$S = 11.96$$

By the empirical rule, approximately 68% of the residuals should be within $\pm 1s_e = \pm 1(11.96) = \pm 11.96$. Because house prices are in units of $1,000, approximately 68% of the predictions are within ±1.96($1,000), or ±$11,960. Examining the output displayed in Table 8.2, 18/23, or about 78%, of the residuals are within this span. According to the empirical rule, approximately 95% of the residuals should be within $\pm 2s_e$, or $\pm 2(11.96) = \pm 23.92$. Further examination of the residual values in Table 8.2 shows that 21 of 23, or 91%, fall within this range. The business researcher can study the standard error of the estimate and these empirical rule–related ranges and decide whether the error of the regression model is sufficiently small to justify further use of the model.

8.3.3 COEFFICIENT OF MULTIPLE DETERMINATION (R^2)

The **coefficient of multiple determination (R^2)** is analogous to the coefficient of determination (r^2) discussed in Chapter 8. R^2 represents *the proportion of variation of the dependent variable, y, accounted for by the independent variables in the regression model.* As with r^2, the range of possible values for R^2 is from 0 to 1. An R^2 of 0 indicates no relationship between the predictor variables in the model and y. An R^2 of 1 indicates that 100% of the variability of y has been accounted for by the predictors. Of course, it is desirable for R^2 to be high, indicating the strong predictability of a regression model. The coefficient of multiple determination can be calculated by the following formula:

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

R^2 can be calculated in the real estate example by using the sum of squares regression (SSR), the sum of squares error (SSE), and sum of squares total (SSyy) from the ANOVA portion of Figure 8.3.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	8189.7	4094.9	28.63	.000
Error	20	2861.0	143.1		
Total	22	11050.7			

SSR → 8189.7; SSE → 2861.0; SS_{yy} → 11050.7

$$R^2 = \frac{SSR}{SS_{yy}} = \frac{8189.7}{11050.7} = .741$$

or

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{2861.0}{11050.7} = .741$$

In addition, virtually all statistical software packages print out R^2 as standard output with multiple regression analysis. A reexamination of Figure 8.3 reveals that R^2 is given as

$$R\text{-sq} = 74.1\%$$

This result indicates that a relatively high proportion of the variation of the dependent variable, house price, is accounted for by the independent variables in this regression model.

8.3.4 ADJUSTED R^2

As additional independent variables are added to a regression model, the value of R^2 cannot decrease, and in most cases it will increase. In the formulas for determining R^2,

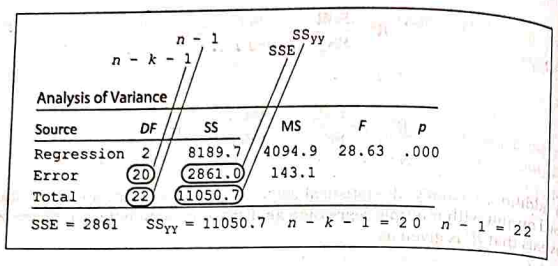
$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

the value of SS_{yy} for a given set of observations will remain the same as independent variables are added to the regression analysis because SS_{yy} is the sum of squares for the dependent variable. Because additional independent variables are likely to increase SSR at least by some amount, the value of R^2 will probably increase for any additional independent variables.

However, sometimes additional independent variables add no *significant* information to the regression model, yet R^2 increases. R^2 therefore may yield an inflated figure. Statisticians have developed an **adjusted R^2** *to take into consideration both the additional information each new independent variable brings to the regression model and the changed degrees of freedom of regression.* Many standard statistical computer packages now compute and report adjusted R^2 as part of the output. The formula for computing adjusted R^2 is

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-k-1)}{SS_{yy}/(n-1)}$$

The value of adjusted R^2 for the real estate example can be solved by using information from the ANOVA portion of the computer output in Figure 8.3.



$$\text{Adj.}R^2 = 1 - \left[\frac{2861/20}{11050.7/22}\right] = 1 - .285 = .715$$

The standard regression output in Figure 8.3 contains the value of the adjusted R^2 already computed. For the real estate example, this value is shown as

$$R\text{-sq(adj.)} = 71.5\%$$

A comparison of R^2 (.741) with the adjusted R^2 (.715) for this example shows that the adjusted R^2 reduces the overall proportion of variation of the dependent variable accounted for by the independent variables by a factor of .026, or 2.6%. The gap between the R^2 and adjusted R^2 tends to increase as non-significant independent variables are added to the regression model. As n increases, the difference between R^2 and adjusted R^2 becomes less.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

7. _____ are data points that are apart, or far, from the mainstream of the other data.
8. In an effort to compute a single statistic that can represent the error in a regression analysis, the _____ property can be overcome by squaring the residuals and then summing the squares.

State whether the following statements are true/false:

9. Standard error of the estimate is essentially the standard deviation of residuals (error) for the regression model.
10. R^2 represents the proportion of variation of the independent variable, x, accounted for by the dependent variable, y, in the regression model.

ACTIVITY

Use the data collected in the previous activity. On the basis of the regression outputs obtained, comment on the overall strength of the regression model using R^2 and adjusted R^2

8.4 INTERPRETING MULTIPLE REGRESSION COMPUTER OUTPUT

8.4.1 A REEXAMINATION OF THE MULTIPLE REGRESSION OUTPUT

Figure 8.5 shows again the multiple regression output for the real estate example. Many of the concepts discussed thus far in the chapter are highlighted. Note the following items:

1. The equation of the regression model
2. The ANOVA table with the F value for the overall test of the model
3. The t ratios, which test the significance of the regression coefficients
4. The value of SSE
5. The value of s_e
6. The value of R^2
7. The value of adjusted R^2

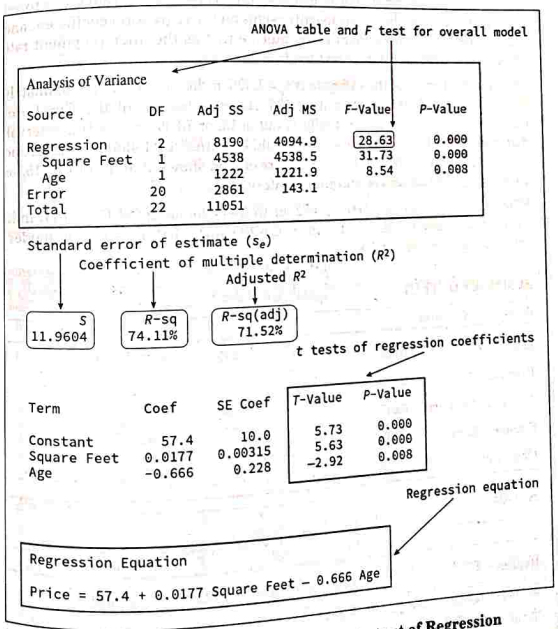


Figure 8.5: Annotated Version of the Output of Regression for the Real Estate Example

DEMONSTRATION PROBLEM 8.2

Discuss the Excel multiple regression output for Demonstration Problem 8.1. Comment on the F test for the overall significance of the model, the t tests of the regression coefficients, and the values of s_e, R^2, and adjusted R^2.

Solution: This multiple regression analysis was done to predict the prime interest rate using the predictors of unemployment and personal saving. The equation of the regression model was presented in the solution of Demonstration Problem 8.1. Shown here is the complete multiple regression output from the Excel analysis of the data.

The value of F for this problem is 12.29, with a p-value of .0012, which is significant at $\alpha = .01$. On the basis of this information, the null hypothesis is rejected for the overall test of significance. At least one of the predictor variables is **statistically significant**, and there is significant predictability of the prime interest rate by this model.

An examination of the t ratios reveals that unemployment rate is a significant predictor at $\alpha = .001$($t = -4.93$ with a p-value of .0003) and that personal savings is a significant predictor at $\alpha = .01$ ($t = 3.31$ with a p-value of .0062). The positive signs on the regression coefficient and the t value for personal savings indicate that as personal savings increase, the prime interest rate tends to get higher. On the other hand, the negative signs on the regression coefficient and the t value for unemployment rates indicate that as the unemployment rate increases, the prime interest rate tends to decrease.

The standard error of the estimate is $s_e = 1.496$, indicating that approximately 68% of the residuals are within ± 1.496. An examination of the Excel-produced residuals shows that actually 11 out of 15, or 73.3%, fall in this interval. Approximately 95% of the residuals should be within $\pm 2(1.496) = \pm 2.992$, and an examination of the Excel-produced residuals shows that 14 out of 15, or 93.3%, of the residuals are within this interval.

R^2 for this regression analysis is .672, or 67.2%. The adjusted R^2 is .617, indicating that there is some inflation in the R^2 value. Overall, there is modest predictability in this model.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.820
R Square	0.672
Adjusted R Square	0.617
Standard Error	1.496
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	54.9835	27.4917	12.29	0.0012
Residual	12	26.8537	2.2378		
Total	14	81.8372			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	13.5786	1.728	7.86	0.0000
Unemployment Rates	−1.6622	0.337	−4.93	0.0003
Personal Savings	0.6586	0.199	3.31	0.0062

RESIDUAL OUTPUT

Observation	*Predicted Prime Interest Rate*	*Residuals*
1	7.3441	0.9859
2	9.2446	0.0754
3	8.8808	1.1292
4	6.1837	0.0663
5	6.6008	0.5492
6	7.2374	1.0326
7	8.9309	−0.5809
8	8.4448	0.7852
9	5.5187	−0.8487
10	5.8198	−1.4798
11	6.3937	1.5663
12	5.1236	−0.0336
13	1.4418	1.8082
14	5.1206	−1.8706
15	6.4346	−3.1846

SELF ASSESSMENT QUESTIONS

Multiple Choice Questions:

Consider the following regression output:

Regression Statistics	
Multiple R	0.95067817
R Square	0.90378898
Adjusted R Square	0.87630011
Standard Error	5.73142152
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2160.055651	1080.028	32.87837	0.00027624
Residual	7	229.9443486	32.84919		
Total	9	2390			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	−8.68701467	9.515477247	−0.91294	0.391634	−31.1875429	13.8135136	−41.986268	24.612239
x_1	3.05672994	0.494424729	6.182397	0.000453	1.887601235	4.22585864	1.32649886	4.78696102
x_2	0.46171268	0.11055673	4.176251	0.004157	0.200287558	0.72313781	0.07482125	0.84860411

SELF ASSESSMENT QUESTIONS

11. Based on the given regression output, which of the following is most likely the regression equation for the model?

    A. $y = -8.687 + 3.057x_1 + 0.462x_2$

    B. $y = 8.687 - 3.057x_1 - 0.462x_2$

    C. $y = 0.913 - 6.182x_1 - 4.176x_2$

    D. $y = -0.913 + 6.182x_1 + 4.176x_2$

12. Based on the given regression output, which of the following is most likely true about the explanation power of the model?

    A. 95% the variation in y is explained by x_1 and x_2.

    B. 90% the variation in y is explained by x_1 and x_2.

    C. 5% the variation in y is explained by x_1 and x_2.

    D. 10% the variation in y is explained by x_1 and x_2.

13. Based on the given regression output, which of the following is most likely true of the impact of x_1 on y?

    A. y increases by 3.057 units for every one unit increase in x_1 keeping x_2 constant.

    B. y increases by 6.182 units for every one unit increase in x_1 keeping x_2 constant.

    C. y increases by 1.326 units for every one unit increase in x_1 keeping x_2 constant.

    D. y increases by 4.787 units for every one unit increase in x_1 keeping x_2 constant.

14. Based on the given regression output, which of the following is most likely true of the impact of x_2 on y?

    A. y increases by 4.176 units for every one unit increase in x_2 keeping x_1 constant.

    B. y increases by 0.462 unit for every one unit increase in x_2 keeping x_1 constant.

    C. y increases by 0.075 unit for every one unit increase in x_2 keeping x_1 constant.

    D. y increases by 0.849 unit for every one unit increase in x_2 keeping x_1 constant.

ACTIVITY

Horizon Transport Company is a transport company in southern India. A major portion of Horizon's business involves deliveries throughout its local area. To develop better work schedules, the managers want to estimate the total daily travel time (y) for their drivers. The managers believe that the total daily travel time would be closely related to the number of kilometres travelled in making the daily deliveries and the number of deliveries. The relevant data with kilometres travelled (x_1) and number of deliveries (x_2) are as follows:

Assignment	Kms (x_1)	Deliveries (x_2)	Time (Hours)
1	200	8	9.3
2	100	6	4.8
3	200	8	8.9
4	200	4	6.5
5	100	4	4.2
6	160	4	6.2
7	150	6	7.4
8	130	8	6
9	180	6	7.6
10	180	4	6.1

Follow the instructions to accomplish this activity.

Open a worksheet.

Enter the labels Assignment, Kilometres, Deliveries, and Time into cells A1:D1 of the worksheet.

Enter the assignment numbers into cells A2:A11 of the worksheet

Enter kilometres travelled into cells B2:B11 of the worksheet

Enter number of deliveries into cells C2:C11 of the worksheet.

Enter time spent into cells D2:D12 of the worksheet.

Now do the following steps to use the Regression tool for the multiple regression analysis.

Step 1: Select the "Data" menu

Step 2: Choose Data Analysis

Step 3: Choose Regression from the list of Analysis Tools

Step 4: When the Regression dialog box appears:

- ❑ Enter D1:D11 in the Input Y Range box
- ❑ Enter B1:C11 in the Input X Range box

ACTIVITY

- ❑ Select Labels
- ❑ Select Confidence Level
- ❑ Enter 99 in the Confidence Level box
- ❑ Select Output Range
- ❑ Enter A13 in the Output Range box
- ❑ Click OK

Hope Excel has generated regression output for you. Now interpret the regression output.

8.5 SUMMARY

- ❑ Multiple regression analysis is a statistical tool in which a mathematical model is developed in an attempt to predict a dependent variable by two or more independent variables or in which at least one predictor is nonlinear. Because doing multiple regression analysis by hand is extremely tedious and time-consuming, it is almost always done on a computer.
- ❑ Residuals, standard error of the estimate, and R^2 are also standard computer regression output with multiple regression. The coefficient of determination for simple regression models is denoted r^2, whereas for multiple regression it is R^2. The interpretation of residuals, standard error of the estimate, and R^2 in multiple regression is similar to that in simple regression. Because R^2 can be inflated with nonsignificant variables in the mix, an adjusted R^2 is often computed. Unlike R^2, adjusted R^2 takes into account the degrees of freedom and the number of observations.

KEY WORDS

1. **Adjusted R^2:** Sometimes additional independent variables add no significant information to the regression model, yet R^2 increases. To overcome this problem, statisticians have developed an adjusted R^2 which takes into consideration both the additional information each new independent variable brings to the regression model and the changed degrees of freedom of regression.
2. **Coefficient of multiple determination (R^2):** Coefficient of multiple determination represents the proportion of variation of the dependent variable, y, accounted for by the independent variables in the regression model.
3. **Dependent variable:** Dependent variable is a variable whose value is predicted in a regression model.
4. **Independent variable:** Independent variable is a variable which acts as a predictor in a regression model.

KEY WORDS

5. **Least squares analysis:** This analysis is concerned with minimizing the sum of squares of error in a regression model.
6. **Multiple regression:** Multiple regression is a regression analysis with two or more independent variables or with at least one nonlinear predictor.
7. **Outliers:** Outliers are data points that are apart, or far, from the mainstream of the other data.
8. **Partial regression coefficient:** Partial regression coefficient of an independent variable represents the increase that will occur in the value of the dependent variable from a one-unit increase in that independent variable if all other independent variables are held constant.
9. **Residual:** The residual of the regression model is the difference between the y value and the predicted y value.
10. **Response plane:** In a multiple regression model with two independent first-order variables, the response surface is a response plane which is fit in a three-dimensional space.
11. **Response surface:** Response surface explores the relationships between independent and dependent variables.
12. **Response variable:** Response variable is a variable whose value is predicted in a regression model.
13. **Standard error of the estimate (s_e):** This standard error is essentially the standard deviation of residuals (error) in a regression model.
14. **Sum of squares of error (SSE):** Sum of square of error is the sum of the squared residuals in a regression model.

8.6 DESCRIPTIVE QUESTIONS

8.1. Use the following data to develop a multiple regression model to predict y from x_1 and x_2. Discuss the output, including comments about the overall strength of the model, the significance of the regression coefficients, and other indicators of model fit.

y	x_1	x_2
198	29	1.64
214	71	2.81
211	54	2.22
219	73	2.70
184	67	1.57
167	32	1.63
201	47	1.99
204	43	2.14
190	60	2.04
222	32	2.93
197	34	2.15

8.2. Given here are the data for a dependent variable, y, and independent variables. Use these data to develop a regression model to predict y. Discuss the output.

y	x_1	x_2	x_3
14	51	16.4	56
17	48	17.1	64
29	29	18.2	53
32	36	17.9	41
54	40	16.5	60
86	27	17.1	55
117	14	17.8	71
120	17	18.2	48
194	16	16.9	60
203	9	18.0	77
217	14	18.9	90
235	11	18.5	67

8.3. The U.S. Bureau of Mines produces data on the price of minerals. Shown here are the average prices per year for several minerals over a decade. Use these data and multiple regression to produce a model to predict the average price of gold from the other variables. Comment on the results of the process.

Gold ($ per oz.)	Copper (cents per lb.)	Silver ($ per oz.)	Aluminum (cents per lb.)
161.1	64.2	4.4	39.8
308.0	93.3	11.1	61.0
613.0	101.3	20.6	71.6
460.0	84.2	10.5	76.0
376.0	72.8	8.0	76.0
424.0	76.5	11.4	77.8
361.0	66.8	8.1	81.0
318.0	67.0	6.1	81.0
368.0	66.1	5.5	81.0
448.0	82.5	7.0	72.3
438.0	120.5	6.5	110.1
382.6	130.9	5.5	87.8

8.4. The Shipbuilders Council of America in Washington, D.C., publishes data about private shipyards. Among the variables reported by this organization are the employment figures (per 1000), the number of naval vessels under construction, and the number of repairs or conversions done to commercial ships (in $ millions). Shown here are the data for these three variables over a seven-year period. Use the data to develop a regression model to predict private shipyard employment from number of naval vessels under construction and repairs or conversions of commercial ships. Comment on the regression model and its strengths and its weaknesses.

Commercial Ship		
Employment	Naval Vessels	Repairs or Conversions
133.4	108	431
177.3	99	1335
143.0	105	1419
142.0	111	1631
130.3	100	852
120.6	85	847
120.4	79	806

8.5. The U.S. Department of Agriculture publishes data annually on various selected farm products. Shown here are the unit production figures (in millions of bushels) for three farm products for 10 years during a 20-year period. Use these data and multiple regression analysis to predict corn production by the production of soybeans and wheat. Comment on the results.

Corn	Soybeans	Wheat
4152	1127	1352
6639	1798	2381
4175	1636	2420
7672	1861	2595
8876	2099	2424
8226	1940	2091
7131	1938	2108
4929	1549	1812
7525	1924	2037
7933	1922	2739

8.7 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

8.1. The regression model is:

$$\hat{y} = 137.27 + 0.0025\,x_1 + 29.206\,x_2$$

$F = 10.89$ with $p = .005$, $s_e = 9.401$, $R^2 = .731$, adjusted $R^2 = .664$. For x_1, $t = 0.01$ with $p = .99$ and for x_2, $t = 4.47$ with $p = .002$. This model has good predictability. The gap between R^2 and adjusted R^2 indicates that there may be a non-significant predictor in the model. The t values show x_1 has virtually no predictability and x_2 is a significant predictor of y.

8.2. The regression model is:

$$\hat{y} = 362.3054 - 4.745518\,x_1 - 13.89972\,x_2 + 1.874297\,x_3$$

$F = 16.05$ with $p = .001$, $s_e = 37.07$, $R^2 = .858$, adjusted $R^2 = .804$. For x_1, $t = -4.35$ with $p = .002$; for x_2, $t = -0.73$ with $p = .483$, for x_3, $t = 1.96$ with $p = .086$. Thus, only one of the three predictors, x_1, is a significant predictor in this model. This model has very good predictability ($R^2 = .858$). The gap between R^2 and adjusted R^2 underscores the fact that there are two non-significant predictors in this model.

8.3. The overall F for this model was 12.19 with a p-value of .002 which is significant at $\alpha = .01$. The t test for Silver is significant at $\alpha = .01$ ($t = 4.94$, $p = .001$). The t test for Aluminum yields a $t = 3.03$ with a p-value of .016 which is significant at $\alpha = .05$. The t test for Copper was insignificant with a p-value of .939. The value of R^2 was 82.1% compared to an adjusted R^2 of 75.3%. The gap between the two indicates the presence of some insignificant predictors (Copper). The standard error of the estimate is 53.44.

8.4. The regression model was:

Employment = 71.03 + 0.4620 Naval Vessels + 0.02082 Commercial

$F = 1.22$ with $p = .386$ (not significant)

$R^2 = .379$ and adjusted $R^2 = .068$

The low value of adjusted R^2 indicates that the model has very low predictability. Both t values are not significant ($t_{\text{Naval Vessels}} = 0.67$ with $p = .541$ and $t_{\text{Commercial}} = 1.07$ with $p = .345$). Neither predictor is a significant predictor of employment.

8.5. The regression model was:

Corn = –2718 + 6.26 Soybeans – 0.77 Wheat

$F = 14.25$ with a p-value of .003 which is significant at $\alpha = .01$

$s_e = 862.4$, $R^2 = 80.3\%$, adjusted $R^2 = 74.6\%$

One of the two predictors, Soybeans, yielded a t value that was significant at $\alpha = .01$ while the other predictor, Wheat was not significant ($t = -0.75$ with a p-value of .476).

8.8 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers	
8.2 The Multiple Regression Model	1.	multiple	
	2.	response	
	3.	partial	
	4.	least	
	5.	True	
	6.	True	
8.3 Residuals, Standard Error of the Estimate, and R^2	7.	Outliers	
	8.	zero-sum	
	9.	True	
	10.	False	
8.4 Interpreting Multiple Regression Computer Output	11.	A	As the coefficients of intercept, x_1 and x_2 are –8.687, 3.057, and 0.462 respectively, the regression equation is: $y = -8.687 + 3.057x_1 + 0.462x_2$
	12.	B	As the value of R^2 is 0.90, 90% the variation in y is explained by x_1 and x_2.
	13.	A	As the coefficient of x_1 is 3.057, y increases by 3.057 units for every one unit increase in x_1 keeping x_2 constant.
	14.	B	As the coefficient of x_2 is 0.462, y increases by 0.462 unit for every one unit increase in x_2 keeping x_1 constant.

9

CHAPTER

TIME-SERIES FORECASTING

CONTENTS

INTRODUCTORY CASELET

FORECASTING GDP

If ever there was a controversial icon from the statistics world, Gross domestic product or GDP is it. It measures income, but not equality, it measures growth, but not destruction, and it ignores values like social cohesion and the environment. Yet, governments, businesses, and probably most people swear by it. According to François Lequiller, head of national accounts at the OECD, part of the problem is that perhaps we expect too much from this trusty, though misunderstood, indicator.

If by growth you mean the expansion of output of goods and services, then GDP or preferably real GDP – which measures growth without the effects of inflation – is perfectly satisfactory. It has been built for this purpose. The letter P stands for "Product", the result of production. Gross domestic product is defined as the sum of all goods and services produced in a country over time, without double counting products used in other output. It is a comprehensive measure, covering the production of consumer goods and services, even government services and investment goods.

In this single number, you get an idea of whether the economy is expanding or contracting. Paul Samuelson, Nobel Laureate and author of many reference textbooks, once described GDP as "truly among the great inventions of the 20th century, a beacon that helps policymakers steer the economy toward key economic objectives".

The economy of India is a developing mixed economy. It is one of the 10 largest economies in the world today. The long-term growth prospective of the Indian economy is positive due to its young population, corresponding low dependency ratio, healthy savings and investment rates, and increasing integration into the global economy. India's GDP numbers (Rupees Billion) from the year 2011–2012 to 2017–2018 are as follows:

Year	Quarter	GDP
2011–12	Q1	19691.32
	Q2	19132.07
	Q3	20738.96
	Q4	21507.12
2012–13	Q1	20745.89
	Q2	20479.09
	Q3	21775.28
	Q4	22462.51
2013–14	Q1	22062.30
	Q2	21938.97
	Q3	23149.41
	Q4	23485.79
2014–15	Q1	23771.54
	Q2	23793.56
	Q3	24570.10
	Q4	24986.12

NOTES

INTRODUCTORY CASELET

Year	Quarter	GDP
2015–16	Q1	25630.13
	Q2	25812.39
	Q3	26395.26
	Q4	27195.71
2016–17	Q1	27750.63
	Q2	27681.67
	Q3	28213.65
	Q4	28830.35
2017–18	Q1	29296.01
	Q2	29405.96
	Q3	30109.42

GDP forecasts have a significant bearing on a country's policy decisions and welfare spending. Moreover, good GDP forecasts often attract FDI flows to a country, which in turn provides a boost to the actual GDP of that country. Now the question that arises is how to forecast the GDP numbers. Forecasts are often made based on the past data and the forecasting technique that produces the least forecasting error is considered the best.

Sources:

- http://oecdobserver.org/news/archivestory.php/aid/1518/Is_GDP_a_satisfactory_measure_of_growth_.html
- https://rbi.org.in/home.aspx
- https://dbie.rbi.org.in/DBIE/dbie.rbi?site=home

N O

LEARNING OBJECTIVES

This chapter discusses the general use of forecasting in business, several tools that are available for making business forecasts, the nature of time-series data, and the role of index numbers in business, thereby enabling you to:

- Compare among various measurements of forecasting error, including mean absolute deviation and mean square error, in order to assess which forecasting method to use.
- Delineate smoothing techniques for forecasting models, including naïve, simple average, moving average, weighted moving average, and exponential smoothing.
- Ascertain trend in time-series data by using linear regression trend analysis, quadratic model trend analysis, and Holt's two-parameter exponential smoothing method.
- Explicate seasonal effects of time-series data by using decomposition and Winters' three-parameter exponential smoothing method.
- Diagnose autocorrelation problem using the Durbin-Watson test and fix autocorrelation problem by adding independent variables and transforming variables, and taking advantage of autocorrelation with autoregression.

9.1 INTRODUCTION

Every day, **forecasting**—*the art or science of predicting the future*—is used in the decision-making process to help business people reach conclusions about buying, selling, producing, hiring, and many other actions. As an example, consider the following items:

- Market watchers predict a resurgence of stock values next year.
- Future brightens for wind power.
- Economist says other sectors to feel oil's decline.
- The baby care market will grow in the next decade.
- CEO says difficult times won't be ending soon for U.S. airline industry.
- Life insurance outlook fades.
- Increased competition from overseas businesses will result in significant layoffs in the U.S. computer chip industry.

How are these and other conclusions reached? What forecasting techniques are used? Are the forecasts accurate? In this chapter we discuss several forecasting techniques, how to measure the error of a forecast, and some of the problems that can occur in forecasting. In addition, this chapter will focus only on data that occur over time, time-series data.

Time-series data are *data gathered on a given characteristic over a period of time at regular intervals.* Time-series forecasting techniques attempt to account for changes over time by examining patterns, cycles, or trends, or using information about previous time periods to predict the outcome for a future time period. Time-series methods include naïve methods, averaging, smoothing, regression trend analysis, and the decomposition of the possible time-series factors, all of which are discussed in subsequent sections.

9.2 FORECASTING

Virtually all areas of business, including production, sales, employment, transportation, distribution, and inventory, produce and maintain time-series data. **Table 9.1** provides an example of time-series data released by the Office of Market Finance, U.S. Department of the Treasury. The table contains the bond yield rates of three-month Treasury Bills for a 17-year period.

Why does the average yield differ from year to year? Is it possible to use these time series data to predict average yields for year 18 or ensuing years? **Figure 9.1** is a graph of these data over time. Often graphical depiction of time-series data can give a clue about any trends, cycles, or relationships that might be present. Does the graph in Figure 9.1 show that bond yields are decreasing? Will next year's yield rate be lower or is a cycle occurring in these data that will result in an increase? To answer such questions, it is sometimes helpful to determine which of the four components of time-series data exist in the data being studied.

TABLE 9.1: BOND YIELDS OF THREE-MONTH TREASURY BILLS

Year	Average Yield
1	14.03%
2	10.69
3	8.63
4	9.58
5	7.48
6	5.98
7	5.82
8	6.69
9	8.12
10	7.51
11	5.42
12	3.45
13	3.02
14	4.29
15	5.51
16	5.02
17	5.07

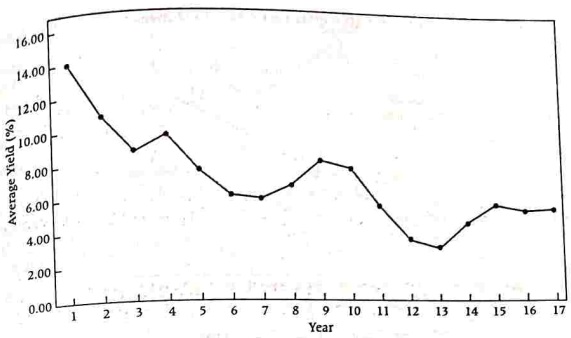


Figure 9.1: Excel Graph of Bond Yield Time-Series Data (new)

9.2.1 TIME-SERIES COMPONENTS

It is generally believed that time-series data are composed of four elements: trend, cycles, seasonal effects, and irregular fluctuations. Not all time-series data have all these elements. Consider **Figure 9.2,** which shows the effects of these time-series elements on data over a period of 13 years.

The long-term general direction of data is referred to as **trend.** Notice that even though the data depicted in Figure 9.2 move through upward and downward periods, the general direction or trend is increasing (denoted in Figure 9.2 by the line). **Cycles** are *patterns of highs and lows through which data move over time periods usually of more than a year.* Notice that the data in Figure 9.2 seemingly move through two periods or cycles of highs and lows over a 13-year period. Time-series data that do not extend over a long period of time may not have enough "history" to show **cyclical effects. Seasonal effects,** on the other hand, are *shorter cycles, which usually occur in time periods of less than one year.* Often seasonal effects are measured by the month, but they may occur by quarter, or may be measured in as small a time frame as a week or even a day. Note the seasonal effects shown in Figure 9.2 as up and down cycles, many of which occur during a 1-year period. **Irregular fluctuations** are *rapid changes or "bleeps" in the data, which occur in even shorter time frames than seasonal effects.* Irregular fluctuations can happen as often as day to day. They are subject to momentary change and are often unexplained. Note the irregular fluctuations in the data of Figure 9.2.

Observe again the bond yield data depicted in Figure 9.1. The general trend seems to move downward and contain two cycles. Each of the cycles traverses approximately 5 to 8 years. It is possible, although not displayed here, that seasonal periods of highs and lows within each year result in seasonal bond yields. In addition, irregular daily fluctuations of bond yield rates may occur but are unexplainable.

Time-series data that contain no trend, cyclical, or seasonal effects are said to be **stationary.** Techniques used to forecast stationary data analyze only the irregular fluctuation effects.

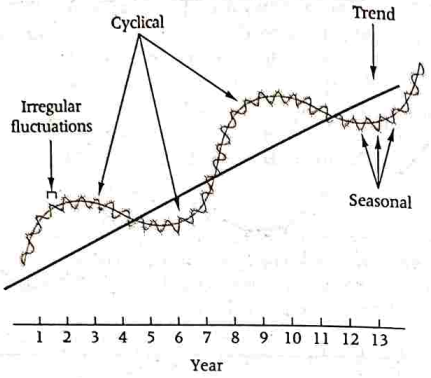


Figure 9.2: Time-Series Effects

9.2.2 THE MEASUREMENT OF FORECASTING ERROR

In this chapter, several forecasting techniques will be introduced that typically produce different forecasts. How does a decision maker know which forecasting technique is doing the best job in predicting the future? One way is to compare forecast values with actual values and determine the amount of **forecasting error** a technique produces. An examination of individual errors gives some insight into the accuracy of the forecasts. However, this process can be tedious, especially for large data sets, and often a single measurement of overall forecasting error is needed for the entire set of data under consideration. Any of several methods can be used to compute error in forecasting. The choice depends on the forecaster's objective, the forecaster's familiarity with the technique, and the method of error measurement used by the computer forecasting software. Several techniques can be used to measure overall error, including mean error (ME), mean absolute deviation (MAD), mean square error (MSE), mean percentage error (MPE), and mean absolute percentage error (MAPE). Here we will consider the mean absolute deviation (MAD) and the mean square error (MSE).

9.2.3 ERROR

The error of an individual forecast is *the difference between the actual value and the forecast of that value.*

Error of an Individual Forecast

$$e_t = x_t - F_t$$

where

e_t = the error of the forecast

x_t = the actual value

F_t = the forecast value

9.2.4 MEAN ABSOLUTE DEVIATION (MAD)

One measure of overall error in forecasting is the mean absolute deviation, MAD. The **mean absolute deviation (MAD)** is *the mean, or average, of the absolute values of the errors.* **Table 9.2** presents the nonfarm partnership tax returns (1000) in the United States over an 11-year period along with the forecast for each year and the error of the forecast. An examination of these data reveals that some of the forecast errors are positive and some are negative. In summing these errors in an attempt to compute an overall measure of error, the negative and positive values offset each other, resulting in an underestimation of the total error. The mean absolute deviation overcomes this problem by taking the absolute value of the error measurement, thereby analyzing the magnitude of the forecast errors without regard to direction.

TABLE 9.2: NONFARM PARTNERSHIP TAX RETURNS

Year	Actual	Forecast	Error
1	1,402	—	—
2	1,458	1,402	56.0
3	1,553	1,441.2	111.8
4	1,613	1,519.5	93.5
5	1,676	1,585.0	91.0
6	1,755	1,648.7	106.3
7	1,807	1,723.1	83.9
8	1,824	1,781.8	42.2
9	1,826	1,811.3	14.7
10	1,780	1,821.6	−41.6
11	1,759	1,792.5	−33.5

Mean Absolute Deviation

$$\text{MAD} = \frac{\Sigma |e_i|}{\text{Number of Forecasts}}$$

The mean absolute error can be computed for the forecast errors in Table 9.2 as follows.

$$\text{MAD} = \frac{|56.0| + |111.8| + |93.5| + |91.0| + |106.3| + |83.9| + |42.2| + |14.7| + |-41.6| + |-33.5|}{10}$$

$$= 67.45$$

9.2.5 MEAN SQUARE ERROR (MSE)

The **mean square error (MSE)** is another way to circumvent the problem of the canceling effects of positive and negative forecast errors. The MSE is *computed by squaring each error (thus creating a positive number) and averaging the squared errors.* The following formula states it more formally.

Mean Square Error

$$\text{MSE} = \frac{\Sigma e_i^2}{\text{Number of Forecasts}}$$

The mean square error can be computed for the errors shown in Table 9.2 as follows.

$$\text{MSE} = \frac{(56.0)^2 + (111.8)^2 + (93.5)^2 + (91.0)^2 + (106.3)^2 + (83.9)^2 + (42.2)^2 + (14.7)^2 + (-41.6)^2 + (-33.5)^2}{10} = 5,584.7$$

Selection of a particular mechanism for computing error is up to the forecaster. It is important to understand that different error techniques will yield different information. The business researcher should be informed enough about the various error measurement techniques to make an educated evaluation of the forecasting results.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

1. _____ is the art or science of predicting the future.
2. The long-term general direction of data is referred to as _____.
3. _____ are patterns of highs and lows through which data move over time periods usually of more than a year.
4. Time-series data that contain no trend, cyclical, or seasonal effects are said to be _____.

State whether the following statements are true/false:

5. Time-series data are data gathered on a given characteristic over a period of time at regular intervals.
6. It is generally believed that time-series data are composed of four elements: trend, cycles, seasonal effects, and irregular fluctuations.

ACTIVITY

Visit RBI website. Under "Statistics," go to "Database on Indian Economy" and click on "Exchange Rate" under "Indicators" to obtain the daily exchange rate of the Indian Rupee against US Dollar, Pound Sterling, Euro, and Japanese Yen from 20th December 2017 to 18th May 2018. Plot the time-series data obtained using MS Excel Graph.

9.3 SMOOTHING TECHNIQUES

Several techniques are available to forecast time-series data that are stationary or that include no significant trend, cyclical, or seasonal effects. These techniques are often referred to as **smoothing techniques** because they *produce forecasts based on "smoothing out" the irregular fluctuation effects in the time-series data.* Three general categories of smoothing techniques are presented here: (1) naïve forecasting models, (2) averaging models, and (3) exponential smoothing.

9.3.1 NAÏVE FORECASTING MODELS

Naïve forecasting models are *simple models in which it is assumed that the more recent time periods of data represent the best predictions or forecasts for future outcomes.* Naïve models do not take into account data trend, cyclical effects, or seasonality. For this reason, naïve models seem to work better with data that are reported on a daily or weekly basis or in situations that show no trend or seasonality. The simplest of the naïve forecasting methods is the model in which the forecast for a given time period is the value for the previous time period.

$$F_t = x_{t-1}$$

where

F_t = the forecast value for time period t

x_{t-1} = the value for time period $t-1$

As an example, if 532 pairs of shoes were sold by a retailer last week, this naïve forecasting model would predict that the retailer will sell 532 pairs of shoes this week. With this naïve model, the actual sales for this week will be the forecast for next week.

Observe the agricultural data in **Table 9.3** representing the total reported domestic rail, truck, and air shipments of bell peppers in the United States for a given year reported by the U.S. Department of Agriculture. **Figure 9.3** presents an Excel graph of these shipments over the 12-month period. From these data, we can make a naïve forecast of the total number of reported shipments of bell peppers for January of the next year by using the figure for December, which is 412.

Another version of the naïve forecast might be to use the number of shipments for January of the previous year as the forecast for January of next year, because the business researcher may believe a relationship exists between bell pepper shipments and the month of the year. In this case, the naïve forecast for next January from Table 9.3 is 336 (January of the previous year). The forecaster is free to be creative with the naïve forecast model method and search for other relationships or rationales within the limits of the time-series data that would seemingly produce a valid forecast.

9.3.2 AVERAGING MODELS

Many naïve model forecasts are based on the value of one time period. Often such forecasts become a function of irregular fluctuations of the data; as a result, the forecasts are "over-steered." Using averaging models, a forecaster enters information from several time periods into the forecast and "smoothes" the data. **Averaging models** are computed by *averaging data from several time periods and using the average as the forecast for the next time period.*

O T E S

TABLE 9.3: TOTAL REPORTED DOMESTIC SHIPMENTS OF BELL PEPPERS	
Month	**Shipments (millions of pounds)**
January	336
February	308
March	582
April	771
May	935
June	808
July	663
August	380
September	333
October	412
November	458
December	412

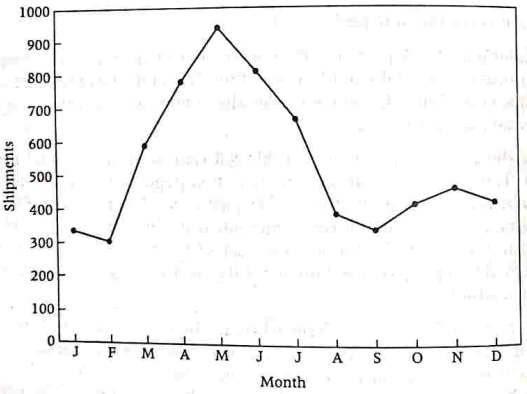


Figure 9.3: Excel Graph of Shipments of Bell Peppers over a 12-Month Period

Simple Averages

The most elementary of the averaging models is the **simple average model**. With this model, *the forecast for time period t is the average of the values for a given number of previous time periods,* as shown in the following equation:

$$F_t = \frac{X_{t-1} + X_{t-2} + X_{t-3} + \cdots + X_{t-n}}{n}$$

The data in **Table 9.4** provide the prices of natural gas in the United States for 3 years. **Figure 9.4** displays a graph of these data.

A simple 12-month average could be used to forecast the price of natural gas for June of year 3 from the data in Table 9.4 by averaging the values for June of year 2 through May of year 3 (the preceding 12 months).

$$F_{June,year\,3} = \frac{2.50+2.96+2.81+2.92+3.50+3.69+3.44+3.35+3.31+3.77+4.16+4.07}{12} = 3.37$$

TABLE 9.4: PRICES OF NATURAL GAS FUTURES ($)	
Time Frame	**Price of Natural Gas ($)**
January (year 1)	4.50
February	4.04
March	4.07
April	4.27
May	4.34
June	4.52
July	4.35
August	3.98
September	3.85
October	3.62
November	3.56
December	3.25
January (year 2)	2.71
February	2.53
March	2.30
April	2.05
May	2.49
June	2.50
July	2.96
August	2.81
September	2.92
October	3.50
November	3.69
December	3.44
January (year 3)	3.35
February	3.31
March	3.77
April	4.16
May	4.07

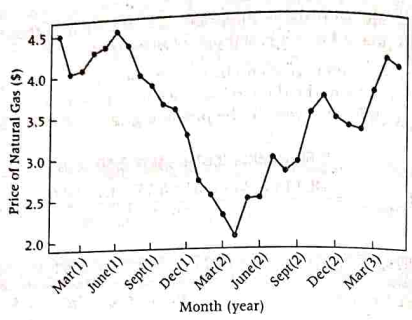


Figure 9.4: Graph of Natural Gas Futures Data

With this **simple average**, the forecast for the year 3 June price of natural gas is $3.37. Note that none of the previous 12-month figures equal this value and that this average is not necessarily more closely related to values early in the period than to those late in the period. The use of the simple average over 12 months tends to smooth the variations, or fluctuations, that occur during this time.

Moving Averages

Suppose we were to attempt to forecast the price of natural gas for July of year 3 by using averages as the forecasting method. would we still use the simple average for June of year 2 through may of year 3 as we did to forecast for June of year 3? instead of using the same 12 months' average used to forecast June of year 3, it would seem to make sense to use the 12 months prior to July of year 3 (July of year 2 through June of year 3) to average for the new forecast. suppose in June of year 3 the price of natural gas is $3.37. we could forecast July of year 3 with a new average that includes the same months used to forecast June of year 3, but without the value for June of year 2 and with the value of June of year 3 added.

$$F_{July,year\,3} = \frac{\begin{array}{c}2.96+2.81+2.92+3.50+3.69+3.44\\+3.35+3.31+3.77+4.16+4.07+3.37\end{array}}{12} = 3.45$$

Computing an average of the values from July of year 2 through June of year 3 produces a moving average, which can be used to forecast the price of natural gas for July of year 3. In computing this moving average, the earliest of the previous 12 values, June of year 2, is dropped and the most recent value, June of year 3, is included.

A **moving average** is *an average that is updated or recomputed for every new time period being considered.* The most recent information is utilized in each new moving average. This advantage is offset by the disadvantages that (1) it is difficult to choose the optimal length of time for which to compute the moving average, and (2) moving averages do not usually adjust for such time-series effects as trend, cycles, or seasonality. To determine the more optimal lengths for which to compute the moving averages, we would need to forecast with several different average lengths and compare the errors produced by them.

DEMONSTRATION PROBLEM 9.1

Shown here are shipments (in millions of dollars) for electric lighting and wiring equipment over a 12-month period. Use these data to compute a 4-month moving average for all available months.

Month	Shipments
January	1056
February	1345
March	1381
April	1191
May	1259
June	1361
July	1110
August	1334
September	1416
October	1282
November	1341
December	1382

Solution: The first moving average is

$$4\text{-Month Moving Average} = \frac{1056+1345+1381+1191}{4} = 1243.25$$

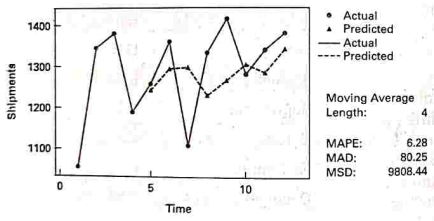
This first 4-month moving average can be used to forecast the shipments in May. Because 1259 shipments were actually made in May, the error of the forecast is

$$\text{Error}_{\text{May}} = 1259 - 1243.25 = 15.75$$

Shown next, along with the monthly shipments, are the 4-month moving averages and the errors of forecast when using the 4-month moving averages to predict the next month's shipments. The first moving average is displayed beside the month of May because it is computed by using January, February, March, and April and because it is being used to forecast the shipments for May. The rest of the 4-month moving averages and errors of forecast are as shown.

4-MONTH MOVING FORECAST			
Month	Shipments	Average	Error
January	1056	—	—
February	1345	—	—
March	1381	—	—
April	1191	—	—
May	1259	1243.25	15.75
June	1361	1294.00	67.00
July	1110	1298.00	−188.00
August	1334	1230.25	103.75
September	1416	1266.00	150.00
October	1282	1305.25	−23.25
November	1341	1285.50	55.50
December	1382	1343.25	38.75

The following graph shows the actual shipment values and the forecast shipment values based on the 4-month moving averages. Notice that the moving averages are "smoothed" in comparison with the individual data values. They appear to be less volatile and seem to be attempting to follow the general trend of the data.



Weighted Moving Averages

A forecaster may want to place more weight on certain periods of time than on others. For example, a forecaster might believe that the previous month's value is three times as important in forecasting as other months. *A moving average in which some time periods are weighted differently than others* is called a **weighted moving average.**

As an example, suppose a 3-month weighted average is computed by weighting last month's value by 3, the value for the previous month by 2, and the value for the month before that by 1. This weighted average is computed as

$$\bar{x}_{weighted} = \frac{3(M_{t-1}) + 2(M_{t-2}) + 1(M_{t-3})}{6}$$

where

M_{t-1} = last month's value
M_{t-2} = value for the previous month
M_{t-3} = value for the month before the previous month

Notice that the divisor is 6. With a weighted average, the divisor always equals the total number of weights. In this example, the value of M_{t-1} counts three times as much as the value for M_{t-3}.

DEMONSTRATION PROBLEM 9.2

Compute a 4-month weighted moving average for the electric lighting and wiring data from Demonstration Problem 9.1, using weights of 4 for last month's value, 2 for the previous month's value, and 1 for each of the values from the 2 months prior to that.

Solution: The first weighted average is

$$\frac{4(1191) + 2(1381) + 1(1345) + 1(1056)}{8} = 1240.875$$

This moving average is recomputed for each ensuing month. Displayed next are the monthly values, the weighted moving averages, and the forecast error for the data.

		4-Month Weighted Moving Average	
Month	Shipments	Forecast	Error
January	1056	—	—
February	1345	—	—
March	1381	—	—
April	1191	—	—
May	1259	1240.9	18.1
June	1361	1268.0	93.0
July	1110	1316.8	−206.8
August	1334	1201.5	132.5
September	1416	1272.0	144.0
October	1282	1350.4	−68.4
November	1341	1300.5	40.5
December	1382	1334.8	47.2

Note that in this problem the errors obtained by using the 4-month weighted moving average were greater than most of the errors obtained by using an unweighted 4-month moving average, as shown here.

Forecast Error, Unweighted 4-Month Moving Average	Forecast Error, Weighted 4-Month Moving Average
—	—
—	—
—	—
—	—
15.8	18.1
67.0	93.0
−188.0	−206.8
103.8	132.5
150.0	144.0
−23.3	−68.4
55.5	40.5
38.8	47.2

Larger errors with weighted moving averages are not always the case. The forecaster can experiment with different weights in using the weighted moving average as a technique. Many possible weighting schemes can be used.

9.3.3 EXPONENTIAL SMOOTHING

Another forecasting technique, **exponential smoothing**, is *used to weight data from previous time periods with exponentially decreasing importance in the forecast.* Exponential smoothing is accomplished by multiplying the actual value for the present time period, X_t, by a value between 0 and 1 (the exponential smoothing constant) referred to as α (not the same α used for a Type I error) and adding that result to the product of the present time period's forecast, F_t, and $(1 - \alpha)$. The following is a more formalized version.

Exponential Smoothing

$$F_{t+1} = \alpha \cdot X_t + (1-\alpha) \cdot F_t$$

where

F_{t+1} = the forecast for the next time period $(t + 1)$

F_t = the forecast for the present time period (t)

X_t = the actual value for the present time period

α = a value between 0 and 1 referred to as the exponential smoothing constant.

The value of α is determined by the forecaster. The essence of this procedure is that the new forecast is a combination of the present forecast and the present actual value. If α is chosen to be less than .5, less weight is placed on the actual value than on the forecast of that value. If α is chosen to be greater than .5, more weight is being put on the actual value than on the forecast value.

As an example, suppose the prime interest rate for a time period is 5% and the forecast of the prime interest rate for this time period was 6%. If the forecast of the prime interest rate for the next period is determined by exponential smoothing with $\alpha = .3$, the forecast is

$$F_{t+1} = (.3)(5\%) + (1.0 - .3)(6\%) = 5.7\%$$

Notice that the forecast value of 5.7% for the next period is weighted more toward the previous forecast of 6% than toward the actual value of 5% because α is .3. Suppose we use $\alpha = .7$ as the exponential smoothing constant. Then,

$$F_{t+1} = (.7)(5\%) + (1.0 - .7)(6\%) = 5.3\%$$

This value is closer to the actual value of 5% than the previous forecast of 6% because the exponential smoothing constant, α, is greater than .5.

To see why this procedure is called exponential smoothing, examine the formula for exponential smoothing again.

$$F_{t+1} = \alpha \cdot X_t + (1-\alpha) \cdot F_t$$

If exponential smoothing has been used over a period of time, the forecast for F_t will have been obtained by

$$F_t = \alpha \cdot X_{t-1} + (1-\alpha) \cdot F_{t-1}$$

Substituting this forecast value, F_t, into the preceding equation for F_{t+1} produces

$$\begin{aligned} F_{t+1} &= \alpha \cdot X_t + (1-\alpha)[\alpha \cdot X_{t-1} + (1-\alpha) \cdot F_{t-1}] \\ &= \alpha \cdot X_t + \alpha(1-\alpha) \cdot X_{t-1} + (1-\alpha)^2 F_{t-1} \end{aligned}$$

but

$$F_{t-1} = \alpha \cdot X_{t-2} + (1-\alpha) F_{t-2}$$

Substituting this value of F_{t-1} into the preceding equation for F_{t+1} produces

$$\begin{aligned} F_{t+1} &= \alpha \cdot X_t + \alpha(1-\alpha) \cdot X_{t-1} + (1-\alpha)^2 F_{t-1} \\ &= \alpha \cdot X_t + \alpha(1-\alpha) \cdot X_{t-1} + (1-\alpha)^2 [\alpha \cdot X_{t-2} + (1-\alpha) F_{t-2}] \\ &= \alpha \cdot X_t + \alpha(1-\alpha) \cdot X_{t-1} + \alpha(1-\alpha)^2 \cdot X_{t-2} + (1-\alpha)^3 F_{t-2} \end{aligned}$$

Continuing this process shows that the weights on previous-period values and forecasts include $(1 - \alpha)^n$ (exponential values). The following chart shows the values of α, $(1 - \alpha)$, $(1 - \alpha)^2$, and $(1 - \alpha)^3$ for three different values of α. Included is the value of $\alpha(1 - \alpha)^3$, which is the weight of the actual value for three time periods back. Notice the rapidly decreasing emphasis on values for earlier time periods. The impact of exponential smoothing on time-series data is to place much more emphasis on recent time periods. The choice of α determines the amount of emphasis.

α	$1-\alpha$	$(1-\alpha)^2$	$(1-\alpha)^3$	$\alpha(1-\alpha)^3$
.2	.8	.64	.512	.1024
.5	.5	.25	.125	.0625
.8	.2	.04	.008	.0064

Some forecasters use the computer to analyze time-series data for various values of α. By setting up criteria with which to judge the forecasting errors, forecasters can select the value of α that best fits the data.

The exponential smoothing formula

$$F_{t+1} = \alpha \cdot X_t + (1-\alpha) \cdot F_t$$

can be rearranged algebraically as

$$F_{t+1} = F_t + \alpha(X_t - F_t)$$

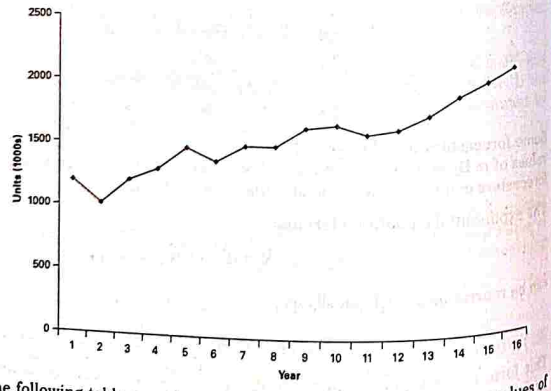
This form of the equation shows that the new forecast, F_{t+1}, equals the old forecast, F_t, plus an adjustment based on α times the error of the old forecast $(X_t - F_t)$. The smaller α is, the less impact the error has on the new forecast and the more the new forecast is like the old. It demonstrates the dampening effect of α on the forecasts.

DEMONSTRATION PROBLEM 9.3

The U.S. Census Bureau reports on the total units of new privately owned housing started over a 16-year recent period in the United States are given here. Use exponential smoothing to forecast the values for each ensuing time period. Work the problem using $\alpha = .2$, .5, and .8.

Year	Total Units (1000)
1	1193
2	1014
3	1200
4	1288
5	1457
6	1354
7	1477
8	1474
9	1617
10	1641
11	1569
12	1603
13	1705
14	1848
15	1956
16	2068

Solution: An Excel graph of these data is shown here.



The following table provides the forecasts with each of the three values of alpha. Note that because no forecast is given for the first time period, we cannot compute a forecast based on exponential smoothing for the second period. Instead, we use the actual value for the first period as the forecast for the second period to get started. As examples, the forecasts for the third, fourth, and fifth periods are computed for $\alpha = .2$ as follows.

$$F_3 = .2(1014) + .8(1193) = 1157.2$$

$$F_4 = .2(1200) + .8(1157.2) = 1165.8$$

$$F_5 = .2(1288) + .8(1165.8) = 1190.2$$

Year	Total Units (1000)	$\alpha = .2$		$\alpha = .5$		$\alpha = .8$	
		F	e	F	e	F	e
1	1193	—	—	—	—	—	—
2	1014	1193.0	−179.0	1193.0	−179.0	1193.0	−179.0
3	1200	1157.2	42.8	1103.5	96.5	1049.8	150.2
4	1288	1165.8	122.2	1151.8	136.2	1170.0	118.0
5	1457	1190.2	266.8	1219.9	237.1	1264.4	192.6
6	1354	1243.6	110.4	1338.4	15.6	1418.5	−64.5
7	1477	1265.7	211.3	1346.2	130.8	1366.9	110.1
8	1474	1307.9	166.1	1411.6	62.4	1455.0	19.0
9	1617	1341.1	275.9	1442.8	174.2	1470.2	146.8
10	1641	1396.3	244.7	1529.9	111.1	1587.6	53.4
11	1569	1445.2	123.8	1585.5	−16.5	1630.3	−61.3
12	1603	1470.0	133.0	1577.2	25.8	1581.3	21.7
13	1705	1496.6	208.4	1590.1	114.9	1598.7	106.3
14	1848	1538.3	309.7	1647.6	200.4	1683.7	164.3
15	1956	1600.2	355.8	1747.8	208.2	1815.1	140.9
16	2068	1671.4	396.6	1851.9	216.1	1927.8	140.2
		$\alpha = .2$		$\alpha = .5$		$\alpha = .8$	
	MAD:	209.8		128.3		111.2	
	MSE:	53,110.5		21,626.7		15,246.4	

Which value of alpha works best on the data? At the bottom of the preceding analysis are the values of two different measurements of error for each of the three different values of alpha. With each measurement of error, $\alpha = .8$ produces the smallest measurement of error. Observe from the Excel graph of the original data that the data are generally increasing. In exponential smoothing, the value of alpha is multiplied by the actual value and $1 - \alpha$ is multiplied by the forecast value to get the next forecast. Because the actual values are generally increasing, the exponential smoothing value with the largest alpha seems to be forecasting the best. In this case, by placing the greatest weight on the actual values, the new forecast seems to predict the new value better.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

7. A _____ average is an average that is updated or recomputed for every new time period being considered.
8. A moving average in which some time periods are weighted differently than others is called a _____ moving average.

State whether the following statements are true/false:

9. A moving average in which some time periods are weighted differently than others is called a weighted moving average.
10. Exponential smoothing is used to weight data from previous time periods with exponentially decreasing importance in the forecast.

ACTIVITY

Visit RBI website. Under "Statistics," go to "Database on Indian Economy" and click on "WPI-Monthly" under "Indicators" to obtain the monthly WPI from January 2017 to December 2018. Calculate 4-month weighted moving average forecasts and the corresponding errors.

9.4 TREND ANALYSIS

There are several ways to determine trend in time-series data, and one of the more prominent is by using regression analysis. In Section 12.9, we explored the use of simple regression analysis in determining the equation of a trend line. In time-series regression trend analysis, the response variable, Y, is the variable being forecast, and the independent variable, X, represents time.

Many possible trend fits can be explored with time-series data. In this section we examine only the linear model and the quadratic model because they are the easiest to understand and simplest to compute. Because seasonal effects can confound trend analysis, it is assumed here that no seasonal effects occur in the data or they were removed prior to determining the trend.

9.4.1 LINEAR REGRESSION TREND ANALYSIS

The data in **Table 9.5** represent 35 years of data on the average length of the workweek in Canada for manufacturing workers. A regression line can be fit to these data by using the time periods as the independent variable and length of workweek as the dependent variable. Because the time periods are consecutive, they can be entered as X along with the time-series data (Y) into a regression analysis. The linear model explored in this example is

$$Y_i = \beta_0 + \beta_1 X_{ti} + \epsilon_i$$

where

Y_i = data value for period i

X_{ti} = ith time period

TABLE 9.5: AVERAGE HOURS PER WEEK IN MANUFACTURING BY CANADIAN WORKERS

Time Period	Hours	Time Period	Hours
1	37.2	19	36.0
2	37.0	20	35.7
3	37.4	21	35.6
4	37.5	22	35.2
5	37.7	23	34.8
6	37.7	24	35.3
7	37.4	25	35.6
8	37.2	26	35.6
9	37.3	27	35.6
10	37.2	28	35.9
11	36.9	29	36.0
12	36.7	30	35.7
13	36.7	31	35.7
14	36.5	32	35.5
15	36.3	33	35.6
16	35.9	34	36.3
17	35.8	35	36.5
18	35.9		

Source: Data prepared by the U.S. Bureau of Labor Statistics, Office of Productivity and Technology.

Figure 9.5 shows the Excel regression output for this example. By using the coefficients of the X variable and intercept, the equation of the trend line can be determined to be

$$\hat{Y} = 37.4161 - .0614X_t$$

The slope indicates that for every unit increase in time period, X_t, a predicted decrease of .0614 occurs in the length of the average workweek in manufacturing. Because the workweek is measured in hours, the length of the average workweek decreases by an average of (.0614)(60 minutes) = 3.7 minutes each year in Canada in manufacturing. The Y intercept, 37.4161, indicates that in the year prior to the first period of these data the average workweek was 37.4161 hours.

The probability of the t ratio (.00000003) indicates that significant linear trend is present in the data. In addition, R^2 = .611 indicates considerable predictability in the model. Inserting the various period values (1, 2, 3, . . . , 35) into the preceding regression equation produces the predicted values of Y that are the trend. For example, for period 23 the predicted value is

$$\hat{Y} = 37.4161 - .0614(23) = 36.0 \text{ hours}$$

The model was developed with 35 periods (years). From this model, the average work-week in Canada in manufacturing for period 41 (the 41st year) can be forecast:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.782
R Square	0.611
Adjusted R Square	0.600
Standard Error	0.509
Observations	35

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	13.447	13.447	51.91	0.00000003
Residual	33	8.549	0.259		
Total	34	21.995			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	37.4161	0.1758	212.81	0.000000000
Year	−0.0614	0.0085	−7.20	0.00000003

Figure 9.5: Excel Regression Output for Hours Worked Using Linear Trend

$$\hat{Y} = 37.4161 - .0614(41) = 34.9 \text{ hours}$$

Figure 9.6 presents an Excel scatter plot of the average workweek lengths over the 35 periods (years). In this Excel plot, the trend line has been fitted through the points. Observe the general downward trend of the data, but also note the somewhat cyclical nature of the points. Because of this pattern, a forecaster might want to determine whether a quadratic model is a better fit for trend.

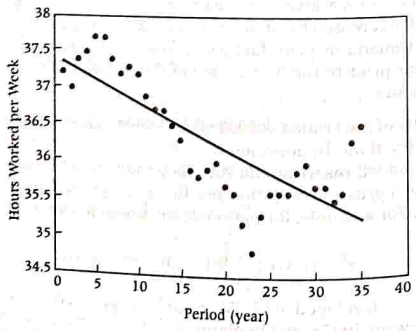


Figure 9.6: Graph for Hours Worked Using Linear Trend

9.4.2 REGRESSION TREND ANALYSIS USING QUADRATIC MODELS

In addition to linear regression, forecasters can explore using quadratic regression models to predict data by using the time-series periods. The quadratic regression model is

$$Y_i = \beta_0 + \beta_1 X_{ti} + \beta_2 X_{ti}^2 + \epsilon_i$$

where

Y_i = the time-series data value for period i

X_{ti} = the ith period

X_{ti}^2 = the square of the ith period

This model can be implemented in time-series trend analysis by using the time periods squared as an additional predictor. Thus, in the hours worked example, besides using $X_t = 1, 2, 3, 4, \ldots, 35$ as a predictor, we would also use $X_t^2 = 1, 4, 9, 16, \ldots, 1225$ as a predictor.

Table 9.6 provides the data needed to compute a quadratic regression trend model on the manufacturing workweek data. Note that the table includes the original data, the time periods, and the time periods squared.

TABLE 9.6: DATA FOR QUADRATIC FIT OF MANUFACTURING WORKWEEK EXAMPLE

Time Period	(Time Period)²	Hours	Time Period	(Time Period)²	Hours
1	1	37.2	19	361	36.0
2	4	37.0	20	400	35.7
3	9	37.4	21	441	35.6
4	16	37.5	22	484	35.2
5	25	37.7	23	529	34.8
6	36	37.7	24	576	35.3
7	49	37.4	25	625	35.6
8	64	37.2	26	676	35.6
9	81	37.3	27	729	35.6
10	100	37.2	28	784	35.9
11	121	36.9	29	841	36.0
12	144	36.7	30	900	35.7
13	169	36.7	31	961	35.7
14	196	36.5	32	1024	35.5
15	225	36.3	33	1089	35.6
16	256	35.9	34	1156	36.3
17	289	35.8	35	1225	36.5
18	324	35.9			

Source: Data prepared by the U.S. Bureau of Labor Statistics, Office of Productivity and Technology.

The Excel computer output for this quadratic trend regression analysis is shown in **Figure 9.7**. We see that the quadratic regression model produces an R^2 of .761 with both X_t and X^2_t in the model. The linear model produced an R^2 of .611 with X_t alone. The quadratic regression seems to add some predictability to the trend model. **Figure 9.8** displays an Excel scatter plot of the work week data with a second-degree polynomial fit through the data.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.873
R Square	0.761
Adjusted R Square	0.747
Standard Error	0.405
Observations	35

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	16.748	8.374	51.07	0.0000000001
Residual	32	5.247	0.164		
Total	34	21.995			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	38.1644	0.2177	175.34	0.000000
Year	−0.1827	0.0279	−6.55	0.000000
Year Sq.	0.0034	0.0008	4.49	0.000088

Figure 9.7: Excel Regression Output for Canadian Manufacturing Example with Quadratic Trend

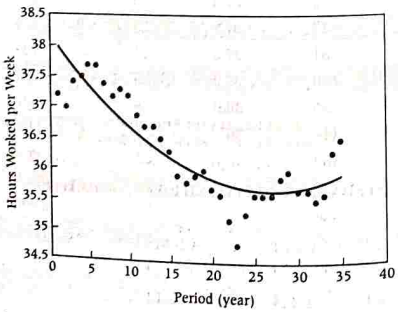


Figure 9.8: Excel Graph of Canadian Manufacturing Data with a Second-Degree Polynomial Fit

DEMONSTRATION PROBLEM 9.4

Following are data on the employed U.S. civilian labor force (100,000) for 1993 through 2014 obtained from the U.S. Bureau of Labor Statistics. Use regression analysis to fit a trend line through the data. Explore a quadratic regression trend also. Does either model do well? Compare the two models.

Year	Labor Force (100,000)
1993	120.26
1994	123.06
1995	124.90
1996	126.71
1997	129.56
1998	131.46
1999	133.49
2000	136.89
2001	136.93
2002	136.49
2003	137.74
2004	139.25
2005	141.73
2006	144.43
2007	146.05
2008	145.36
2009	139.90
2010	139.10
2011	139.90
2012	142.50
2013	143.93
2014	146.31

Solution: Recode the time periods as 1 through 22 and let that be *X*. Run the regression analysis with the labor force members as *Y*, the dependent variable, and the time period as the independent variable. Now square all the *X* values, resulting in 1, 4, 9, . . . , 400, 441, 484, and let those formulate a second predictor (X^2). Run the regression analysis to predict the number in the labor force with both the time period variable (X) and the (time period)^2 variable. The output for each of these regression analyses follows.

Regression Analysis: Labor Force (100,000) versus Year

Analysis of Variance					
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1001.2	1001.22	82.58	0.000
Error	20	242.5	12.12		
Total	21	1243.7			

Model Summary

S	R-sq	R-sq(adj)
3.48196	80.50%	79.53%

Coefficients

Term	Coef	SE Coef	*T*-Value	*P*-Value
Constant	124.41	1.54	80.95	0.000
Year	1.063	0.117	9.09	0.000

Regression Equation

Labor Force (100,000) = 124.41 + 1.063 Year

Regression Analysis: Labor Force (100,000) versus Year, Year Sq

Analysis of Variance

Source	DF	Adj SS	Adj MS	*F*-Value	*P*-Value
Regression	2	1151.25	575.625	118.30	0.000
Error	19	92.45	4.866		
Total	21	1243.70			

Model Summary

S	R-sq	R-sq(adj)
2.20589	92.57%	91.78%

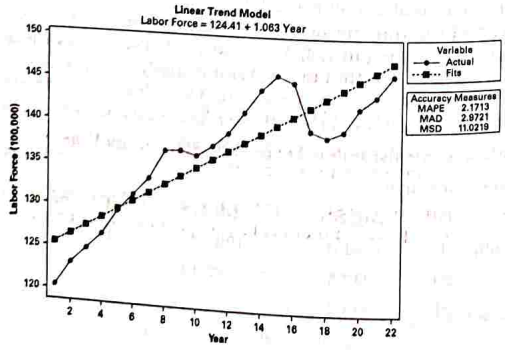
Coefficients

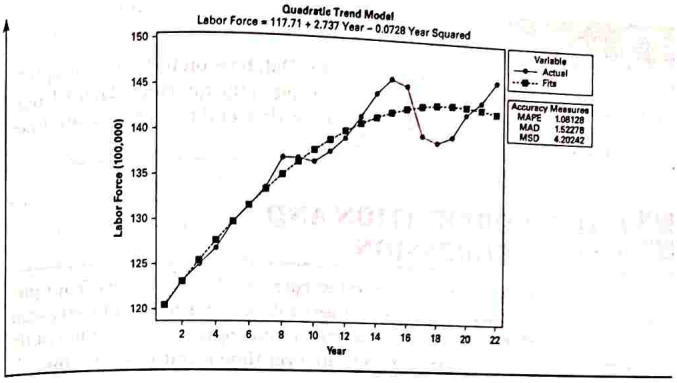
Term	Coef	SE Coef	*T*-Value	*P*-Value
Constant	117.71	1.55	75.96	0.000
Year	2.737	0.310	8.82	0.000
Year Sq	−0.0728	0.0131	−5.55	0.000

Regression Equation

Labor Force (100,000) = 117.71 + 2.737 Year − 0.0728 Year Sq

A comparison of the models shows that the linear model accounts for 80.50% of the variability of the labor force figures. The quadratic model increases the R^2 to 92.57%. Shown next are scatter plots of the data. First is the linear model, and then the quadratic model is presented. Note the considerable reduction in forecasting error when using the quadratic model.





9.4.3 HOLT'S TWO-PARAMETER EXPONENTIAL SMOOTHING METHOD

The exponential smoothing technique presented in Section 9.2 (single exponential smoothing) is appropriate to use in forecasting stationary time-series data but is ineffective in forecasting time-series data with a trend because the forecasts will lag behind the trend. However, another exponential smoothing technique, Holt's two-parameter exponential smoothing method, can be used for trend analysis. Holt's technique uses weights (β) to smooth the trend in a manner similar to the smoothing used in single exponential smoothing (α). Using these two weights and several equations, Holt's method is able to develop forecasts that include both a smoothing value and a trend value.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

11. In time-series regression trend analysis, the response variable, Y, is the variable being forecast, and the independent variable, X, represents _____.
12. Because _____ effects can confound trend analysis, it is assumed here that no seasonal effects occur in the data or they were removed prior to determining the trend.

State whether the following statements are true/false:

13. Holt's technique uses weights to smooth the trend in a manner similar to the smoothing used in single exponential smoothing.
14. In addition to linear regression, forecasters can explore using quadratic regression models to predict data by using the time-series periods.

ACTIVITY

Visit RBI website. Under "Statistics," go to "Database on Indian Economy" and click on "GDP" under "Indicators" to obtain the quarterly GDPs from 2012–2013 to 2017–2018. Plot the time-series data and fit a regression line using MS Excel.

9.5 AUTOCORRELATION AND AUTOREGRESSION

Data values gathered over time are often correlated with values from past time periods. This characteristic can cause problems in the use of regression in forecasting and at the same time can open some opportunities. One of the problems that can occur in regressing data over time is autocorrelation.

9.5.1 AUTOCORRELATION

Autocorrelation, or **serial correlation**, occurs in data *when the error terms of a regression forecasting model are correlated.* The likelihood of this occurring with business data increases over time, particularly with economic variables. Autocorrelation can be a problem in using regression analysis as the forecasting method because one of the assumptions underlying regression analysis is that the error terms are independent or random (not correlated). In most business analysis situations, the correlation of error terms is likely to occur as positive autocorrelation (positive errors are associated with positive errors of comparable magnitude and negative errors are associated with negative errors of comparable magnitude).

When autocorrelation occurs in a regression analysis, several possible problems might arise. First, the estimates of the regression coefficients no longer have the minimum variance property and may be inefficient. Second, the variance of the error terms may be greatly underestimated by the mean square error value. Third, the true standard deviation of the estimated regression coefficient may be seriously underestimated. Fourth, the confidence intervals and tests using the t and F distributions are no longer strictly applicable.

First-order autocorrelation results from correlation between the error terms of adjacent time periods (as opposed to two or more previous periods). If first-order autocorrelation is present, the error for one time period, e_t, is a function of the error of the previous time period, e_{t-1}, as follows:

$$e_t = \rho e_{t-1} + v_t$$

The first-order autocorrelation coefficient, ρ, measures the correlation between the error terms. It is a value that lies between –1 and 0 and +1, as does the coefficient of correlation discussed in Chapter 12. v_t is a normally distributed independent error term. If positive autocorrelation is present, the value of ρ is between 0 and +1. If the value of ρ is 0, $e_t = v_t$, which means there is no autocorrelation and e_t is just a random, independent error term.

One way to *test to determine whether autocorrelation is present in a time-series regression analysis* is by using the **Durbin-Watson test** for autocorrelation. Shown next is the formula for computing a Durbin-Watson test for autocorrelation.

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

where

n = the number of observations

Note from the formula that the Durbin-Watson test involves finding the difference between successive values of error $(e_t - e_{t-1})$. If errors are positively correlated, this difference will be smaller than with random or independent errors. Squaring this term eliminates the cancellation effects of positive and negative terms.

The null hypothesis for this test is that there is *no* autocorrelation. For a two-tailed test, the alternative hypothesis is that there *is* autocorrelation.

$$H_0: \rho = 0$$

$$H_a: \rho \neq 0$$

As mentioned before, most business forecasting autocorrelation is positive autocorrelation. In most cases, a one-tailed test is used.

$$H_0: \rho = 0$$

$$H_a: \rho > 0$$

In the Durbin-Watson test, D is the observed value of the Durbin-Watson statistic using the residuals from the regression analysis. A critical value for D can be obtained from the values of α, n, and k by using Table A.9 in the appendix, where α is the level of significance, n is the number of data items, and k is the number of predictors. Two Durbin-Watson tables are given in the appendix. One table contains values for $\alpha = .01$ and the other for $\alpha = .05$. The Durbin-Watson tables in Appendix A include values for d_U and d_L. These values range from 0 to 4. If the observed value of D is above d_U, we fail to reject the null hypothesis and there is no significant autocorrelation. If the observed value of D is below d_L, the null hypothesis is rejected and there is autocorrelation. Sometimes the observed statistic, D, is between the values of d_U and d_L. In this case, the Durbin-Watson test is inconclusive.

As an example, consider **Table 9.7**, which contains crude oil production and natural gas withdrawal data for the United States over a 25-year period published by the Energy Information Administration in its Annual Energy Review. A regression line can be fit through these data to determine whether the amount of natural gas withdrawals can be predicted by the amount of crude oil production. The resulting errors of prediction can be tested by the Durbin-Watson statistic for the presence of significant positive autocorrelation by using $\alpha = .05$. The hypotheses are

$$H_0: \rho = 0$$

$$H_a: \rho > 0$$

TABLE 9.7: U.S. CRUDE OIL PRODUCTION AND NATURAL GAS WITHDRAWALS OVER A 25-YEAR PERIOD

Year	Crude Oil Production (1000s)	Natural Gas Withdrawals from Natural Gas Wells (1000s)
1	8.597	17.573
2	8.572	17.337
3	8.649	15.809
4	8.688	14.153
5	8.879	15.513
6	8.971	14.535
7	8.680	14.154
8	8.349	14.807
9	8.140	15.467
10	7.613	15.709
11	7.355	16.054
12	7.417	16.018
13	7.171	16.165
14	6.847	16.691
15	6.662	17.351
16	6.560	17.282
17	6.465	17.737
18	6.452	17.844
19	6.252	17.729
20	5.881	17.590
21	5.822	17.726
22	5.801	18.129
23	5.746	17.795
24	5.681	17.819
25	5.430	17.739

The following regression equation was obtained by means of a statistical software.

$$\text{Natural Gas Withdrawals} = 22.7372 - 0.8507\,\text{Crude Oil Production}$$

Using the values for crude oil production (X) from Table 9.7 and the regression equation shown here, predicted values of Y (natural gas withdrawals) can be computed. From the predicted values and the actual values, the errors of prediction for each time interval, e_t, can be calculated. **Table 9.8** shows the values of $\hat{Y}$, e_t, e_t^2, $(e_t - e_{t-1})$, and $(e_t - e_{t-1})^2$ for this example. Note that the first predicted value of Y is

$$\hat{Y}_1 = 22.7372 - 0.8507(8.597) = 15.4237$$

The error for year 1 is

$$\text{Actual}_1 - \text{Predicted}_1 = 17.573 - 15.4237 = 2.1493$$

TABLE 9.8: PREDICTED VALUES AND ERROR TERMS FOR THE CRUDE OIL PRODUCTION AND NATURAL GAS WITHDRAWAL DATA

Year	$\hat{Y}$	e_t	e_t^2	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$
1	15.4237	2.1493	4.6195	—	—
2	15.4450	1.8920	3.5797	−0.2573	0.0662
3	15.3795	0.4295	0.1845	−1.4625	2.1389
4	15.3463	−1.1933	1.4240	−1.6228	2.6335
5	15.1838	0.3292	0.1084	1.5225	2.3180
6	15.1056	−0.5706	0.3256	−0.8998	0.8096
7	15.3531	−1.1991	1.4378	−0.6285	0.3950
8	15.6347	−0.8277	0.6851	0.3714	0.1379
9	15.8125	−0.3455	0.1194	0.4822	0.2325
10	16.2608	−0.5518	0.3045	−0.2063	0.0426
11	16.4803	−0.4263	0.1817	0.1255	0.0158
12	16.4276	−0.4096	0.1678	0.0167	0.0003
13	16.6368	−0.4718	0.2226	−0.0622	0.0039
14	16.9125	−0.2215	0.0491	0.2503	0.0627
15	17.0698	0.2812	0.0791	0.5027	0.2527
16	17.1566	0.1254	0.0157	−0.1558	0.0243
17	17.2374	0.4996	0.2496	0.3742	0.1400
18	17.2485	0.5955	0.3546	0.0959	0.0092
19	17.4186	0.3104	0.0963	−0.2851	0.0813
20	17.7342	−0.1442	0.0208	−0.4546	0.2067
21	17.7844	−0.0584	0.0034	0.0858	0.0074
22	17.8023	0.3267	0.1067	0.3851	0.1483
23	17.8491	−0.0541	0.0029	−0.3808	0.1450
24	17.9044	−0.0854	0.0073	−0.0313	0.0010
25	18.1179	−0.3789	0.1436	−0.2935	0.0861
		$\Sigma e_t^2 = 14.4897$		$\Sigma(e_t - e_{t-1})^2 = 9.9589$	

The value of $e_t - e_{t-1}$ for year 1 and year 2 is computed by subtracting the error for year 1 from the error of year 2.

$$e_{\text{year2}} - e_{\text{year1}} = 1.8920 - 2.1493 = -0.2573$$

The Durbin-Watson statistic can now be computed:

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{9.9589}{14.4897} = 0.6873$$

Because we used a simple linear regression, the value of k is 1. The sample size, n, is 25, and $\alpha = .05$. The critical values in Table A.9 are

$$d_L = 1.29 \text{ and } d_U = 1.45$$

Because the computed D statistic, 0.6873, is less than the value of $d_L = 1.29$, the null hypothesis is rejected. A positive autocorrelation is present in this example.

9.5.2 WAYS TO OVERCOME THE AUTOCORRELATION PROBLEM

Several approaches to data analysis can be used when autocorrelation is present. One uses additional independent variables and another transforms the independent variable.

Addition of Independent Variables

Often the reason autocorrelation occurs in regression analyses is that one or more important predictor variables have been left out of the analysis. For example, suppose a researcher develops a regression forecasting model that attempts to predict sales of new homes by sales of used homes over some period of time. Such a model might contain significant autocorrelation. The exclusion of the variable "prime mortgage interest rate" might be a factor driving the autocorrelation between the other two variables. Adding this variable to the regression model might significantly reduce the autocorrelation.

Transforming Variables

When the inclusion of additional variables is not helpful in reducing autocorrelation to an acceptable level, transforming the data in the variables may help to solve the problem. One such method is the **first-differences approach**. With the first-differences approach, *each value of X is subtracted from each succeeding time period value of X*; these "differences" become the new and transformed X variable. The same process is used to transform the Y variable. The regression analysis is then computed on the transformed X and transformed Y variables to compute a new model that is hopefully free of significant autocorrelation effects.

Another way is to generate new variables by using the percentage changes from period to period and regressing these new variables. A third way is to use autoregression models.

9.5.3 AUTOREGRESSION

A forecasting technique that takes advantage of the relationship of values (Y_t) to previous-period values (Y_{t-1}, Y_{t-2}, Y_{t-3}, . . .) is called autoregression. **Autoregression** is *a multiple regression technique in which the independent variables are time-lagged versions of the dependent variable,* which means we try to predict a value of Y from values of Y from previous time periods. The independent variable can be lagged for one, two, three, or more time periods. An autoregressive model containing independent variables for three time periods looks like this:

$$\hat{Y} = b_0 + b_1 Y_{t-1} + b_2 Y_{t-2} + b_3 Y_{t-3}$$

As an example, we shall attempt to predict the volume of natural gas withdrawal, displayed in Table 9.12, by using data lagged for both one and two time periods. The data used in this analysis are displayed in **Table 9.9**. Using Excel, a multiple regression model is developed to predict the values of Y_t by the values of Y_{t-1} and Y_{t-2}. The results appear in **Figure 9.9**. Note that the regression analysis does not use data from years 1 and 2 of Table 9.9 because there are no values for the two lagged variables for one or both of those years.

The autoregression model is

$$Y_t = 2.4081 + 0.9678 Y_{t-1} - 0.1128 Y_{t-2}$$

The relatively high value of R^2 (74.6%) and relatively small value of s_e (0.693) indicate that this regression model has fairly strong predictability. Interestingly, the one-period lagged variable is quite significant ($t = 4.36$ with

TABLE 9.9: TIME-LAGGED NATURAL GAS DATA

Year	Natural Gas Withdrawal Y_t	One Period Lagged Y_{t-1} (X_1)	Two Period Lagged Y_{t-2} (X_2)
1	17.573	—	—
2	17.337	17.573	—
3	15.809	17.337	17.573
4	14.153	15.809	17.337
5	15.513	14.153	15.809
6	14.535	15.513	14.153
7	14.154	14.535	15.513
8	14.807	14.154	14.535
9	15.467	14.807	14.154
10	15.709	15.467	14.807
11	16.054	15.709	15.467
12	16.018	16.054	15.709
13	16.165	16.018	16.054
14	16.691	16.165	16.018
15	17.351	16.691	16.165
16	17.282	17.351	16.691
17	17.737	17.282	17.351
18	17.844	17.737	17.282
19	17.729	17.844	17.737
20	17.590	17.729	17.844
21	17.726	17.590	17.729
22	18.129	17.726	17.590
23	17.795	18.129	17.726
24	17.819	17.795	18.129
25	17.739	17.819	17.795

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.864
R Square	0.746
Adjusted R Square	0.721
Standard Error	0.693
Observations	23

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	28.3203	14.1602	29.44	0.0000011
Residual	20	9.6187	0.4809		
Total	22	37.9390			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.4081	1.9608	1.23	0.2337
Lagged 1	0.9678	0.2221	4.36	0.0003
Lagged 2	−0.1128	0.2239	−0.50	0.6201

Figure 9.9: Excel Autoregression Result for Natural Gas Withdrawal Data

a *p*-value of 0.0003), but the two-period lagged variable is not significant ($t = -0.50$ with a *p*-value of 0.6201), indicating the presence of first-order autocorrelation.

Autoregression can be a useful tool in locating seasonal or cyclical effects in time series data. For example, if the data are given in monthly increments, autoregression using variables lagged by as much as 12 months can search for the predictability of previous monthly time periods. If data are given in quarterly time periods, autoregression of up to four periods removed can be a useful tool in locating the predictability of data from previous quarters. When the time periods are in years, lagging the data by yearly periods and using autoregression can help in locating cyclical predictability.

SELF ASSESSMENT QUESTIONS

Fill in the blanks:

15. ——— occurs in data when the error terms of a regression forecasting model are correlated.
16. ——— is a multiple regression technique in which the independent variables are time-lagged versions of the dependent variable.

SELF ASSESSMENT QUESTIONS

17. With the first-differences approach, each value of X is subtracted from each succeeding time period value of X; these "differences" become the new and transformed X variable.
18. Serial correlation occurs in data when the error terms of a regression forecasting model are uncorrelated.
19. One way to test to determine whether autocorrelation is present in a time-series regression analysis is by using the Durbin-Watson test for autocorrelation.

ACTIVITY

Take the monthly Nifty values from January 2010 to December 2017. Use the data to create a regression forecasting model using the first-differences data transformation.

9.6 SUMMARY

- One way to establish the validity of a forecast is to examine the forecasting error. The error of a forecast is the difference between the actual value and the forecast value. Computing a value to measure forecasting error can be done in several different ways. This chapter presents mean absolute deviation and mean square error for this task.
- Regression analysis with either linear or quadratic models can be used to explore trend. Regression trend analysis is a special case of regression analysis in which the dependent variable is the data to be forecast and the independent variable is the time periods numbered consecutively from 1 to k, where k is the number of time periods. For the quadratic model, a second independent variable is constructed by squaring the values of the first independent variable, and both independent variables are included in the analysis.
- One group of time-series forecasting methods contains smoothing techniques. Among these techniques are naïve models, averaging techniques, and simple exponential smoothing. These techniques do much better if the time series data are stationary or show no significant trend or seasonal effects. Naïve forecasting models are models in which it is assumed that the more recent time periods of data represent the best predictions or forecasts for future outcomes.
- Simple averages use the average value for some given length of previous time periods to forecast the value for the next period. Moving averages are time period averages that are revised for each time period by including the most recent value(s) in the computation of the average and deleting the value or values that are farthest away from the present time period. A special case of the moving average is the weighted moving average, in which different weights are placed on the values from different time periods.

- Simple (single) exponential smoothing is a technique in which data from previous time periods are weighted exponentially to forecast the value for the present time period. The forecaster has the option of selecting how much to weight more recent values versus those of previous time periods.
- Autocorrelation or serial correlation occurs when the error terms from forecasts are correlated over time. In regression analysis, this effect is particularly disturbing because one of the assumptions is that the error terms are independent. One way to test for autocorrelation is to use the Durbin-Watson test. There are a number of methods that attempt to overcome the effects of autocorrelation on the data.
- Autoregression is a forecasting technique in which time-series data are predicted by independent variables that are lagged versions of the original dependent variable data. A variable that is lagged one period is derived from values of the previous time period. Other variables can be lagged two or more periods.

KEY WORDS

1. **Autocorrelation:** Autocorrelation occurs in data when the error terms of a regression forecasting model are correlated.
2. **Autoregression:** Autoregression is a multiple regression technique in which the independent variables are time-lagged versions of the dependent variable.
3. **Averaging models:** These are computed by averaging data from several time periods and using the average as the forecast for the next time period.
4. **Cycles:** Cycles are patterns of highs and lows through which data move over time periods usually of more than a year.
5. **Cyclical effects:** These effects are shown by time-series data that extend over a long period of time with enough "history" to show.
6. **Decomposition:** is a technique for isolating the effects of seasonality.
7. **Durbin-Watson test:** Durbin-Watson is a test to determine whether autocorrelation is present in a time-series regression analysis.
8. **Error of an individual forecast:** is the difference between the actual value and the forecast of that value.
9. **Exponential smoothing:** This is used to weight data from previous time periods with exponentially decreasing importance in the forecast.
10. **First-difference approach:** This approach requires each value of X is subtracted from each succeeding time period value of X and these "differences" become the new and transformed X variable.
11. **Forecasting:** is the art or science of predicting the future.
12. **Forecasting error:** This error is the difference between the actual value and the forecast of that value.

KEY WORDS

13. **Mean absolute deviation (MAD):** Mean absolute deviation is the mean, or average, of the absolute values of the errors.
14. **Mean squared error (MSE):** This error is computed by squaring each error (thus creating a positive number) and averaging the squared errors.
15. **Moving average:** Moving average is an average that is updated or recomputed for every new time period being considered.
16. **Naïve forecasting methods:** These are simple models in which it is assumed that the more recent time periods of data represent the best predictions or forecasts for future outcomes.
17. **Serial correlation:** This correlation occurs in data when the error terms of a regression forecasting model are correlated.
18. **Simple average:** Simple average is the total value of all the observations divided by the number of observations.
19. **Simple average model:** With this model, the forecast for time period *t* is the average of the values for a given number of previous time periods.
20. **Smoothing techniques:** These techniques are used to produce forecasts based on "smoothing out" the irregular fluctuation effects in the time-series data.
21. **Stationary:** Time-series data that contain no trend, cyclical, or seasonal effects are said to be stationary.
22. **Time-series data:** These are data that have been gathered at regular intervals over a period of time.
23. **Trend:** Trend is the long-term general direction of data.
24. **Weighted moving average:** This is a moving average in which some time periods are weighted differently than others.

9.7 DESCRIPTIVE QUESTIONS

9.1. Following are the average yields of long-term new corporate bonds over a several-month period published by the Office of Market Finance of the U.S. Department of the Treasury.

Month	Yield	Month	Yield
1	10.08	13	7.91
2	10.05	14	7.73
3	9.24	15	7.39
4	9.23	16	7.48
5	9.69	17	7.52
6	9.55	18	7.48
7	9.37	19	7.35

Month	Yield	Month	Yield
8	8.55	20	7.04
9	8.36	21	6.88
10	8.59	22	6.88
11	7.99	23	7.17
12	8.12	24	7.22

(a) Explore trends in these data by using regression trend analysis. How strong are the models? Is the quadratic model significantly stronger than the linear trend model?

(b) Use a 4-month moving average to forecast values for each of the ensuing months.

(c) Use simple exponential smoothing to forecast values for each of the ensuing months. Let $\alpha = .3$ and then let $\alpha = .7$. Which weight produces better forecasts?

(d) Compute MAD for the forecasts obtained in parts (b) and (c) and compare the results.

9.2. The following data contain the quantity (million pounds) of U.S. domestic fish caught annually over a 25-year period as published by the National Oceanic and Atmospheric Administration.

(a) Use a 3-year moving average to forecast the quantity of fish for the years 4 through 25 for these data. Compute the error of each forecast and then determine the mean absolute deviation of error for the forecast.

(b) Use exponential smoothing and $\alpha = .2$ to forecast the data from 4 through 25. Let the forecast for 2 equal the actual value for 1. Compute the error of each forecast and then determine the mean absolute deviation of error for the forecast.

(c) Compare the results obtained in parts (a) and (b) using MAD. Which technique seems to perform better? Why?

Year	Quantity	Year	Quantity
1	6,137	14	9,089
2	7,019	15	8,876
3	7,391	16	9,290
4	8,750	17	9,250
5	9,816	18	9,315
6	9,644	19	9,424
7	9,951	20	9,379
8	9,971	21	9,180
9	10,089	22	9,026
10	9,693	23	7,953
11	9,380	24	7,875
12	9,615	25	7,994
13	8,992		

9.3. Given below are data on the number of business establishments (millions) and the self-employment rate (%) released by the Small Business Administration, Office of Advocacy, for a 21-year period of U.S. business activity. Develop a regression model to predict the self-employment rate by the number of business establishments. Use this model to predict the self-employment rate for a year in which there are 7.0 (million) business establishments. Discuss the strength of the regression model. Use these data and the regression model to compute a Durbin-Watson test to determine whether significant autocorrelation is present. Let alpha be .05.

Number of Establishments (millions)	Self-Employment Rate (%)
4.54317	8.1
4.58651	8.0
4.63396	8.1
5.30679	8.2
5.51772	8.2
5.70149	8.0
5.80697	7.9
5.93706	8.0
6.01637	8.2
6.10692	8.1
6.17556	8.0
6.20086	8.1
6.31930	7.8
6.40123	8.0
6.50907	8.1
6.61272	7.9
6.73848	7.8
6.89487	7.7
6.94182	7.5
7.00844	7.2
7.07005	6.9

9.4. The U.S. Department of Commerce publishes data on industrial machinery and equipment. Shown here are the shipments (in $ billions) of industrial machinery and equipment from the first quarter of year 1 through the fourth quarter of year 6. Use these data to determine the seasonal indexes for the data through time-series decomposition methods. Use the four-quarter centered moving average in the computations.

Time Period	Industrial Machinery and Equipment Shipments
1st quarter (year 1)	54.019
2nd quarter	56.495
3rd quarter	50.169
4th quarter	52.891

9.5. The Board of Governors of the Federal Reserve System publishes data on mortgage debt outstanding by type of property and holder. The following data give the amounts of residential nonfarm debt (in $ billions) held by savings institutions in the United States over a 10-year period. Use these data to develop an autoregression model with a one-period lag. Discuss the strength of the model.

Year	Debt
1	529
2	554
3	559
4	602
5	672
6	669
7	600
8	538
9	490
10	470

9.6. Shown here are data from the Investment Company Institute on Total Net Assets and Total Number of Shareholder Accounts of money market funds over a period of 27 years. Use these data to develop a regression model to forecast Total Net Assets by the Total Number of Shareholder Accounts. Total Net Assets is given in billions of dollars and Total Number of Shareholder Accounts is given in millions. Conduct a Durbin-Watson test on the data and the regression model to determine whether significant autocorrelation is present. Let $\alpha = .01$.

Year	Total Net Assets ($ billions)	Total Number of Shareholder Accounts (millions)
1986	292.2	16.3
1987	316.1	17.7
1988	338.0	18.6
1989	428.1	21.3
1990	498.3	23.0
1991	542.4	23.6
1992	546.2	23.6
1993	565.3	23.6
1994	611.0	25.4
1995	753.0	30.1
1996	901.8	32.2
1997	1,058.9	35.6
1998	1,351.7	38.8
1999	1,613.1	43.6
2000	1,845.2	48.1
2001	2,285.3	47.2
2002	2,265.1	45.4
2003	2,040.0	41.2
2004	1,901.3	37.6
2005	2,026.8	36.8
2006	2,338.5	37.1
2007	3,085.8	39.1
2008	3,832.2	38.1
2009	3,315.9	33.5
2010	2,803.9	30.3
2011	2,691.4	28.7
2012	2,693.5	27.9
2013	2,718.3	26.1

9.7. The purchasing-power value figures for the minimum wage in dollars for the years 1 through 18 are shown here. Use these data and exponential smoothing to develop forecasts for the years 2 through 18. Try $\alpha = .1$, .5, and .8, and compare the results using MAD. Discuss your findings. Select the value of alpha that worked best and use your exponential smoothing results to predict purchasing power for year 19.

Year	Purchasing Power	Year	Purchasing Power
1	$6.04	10	$4.34
2	5.92	11	4.67
3	5.57	12	5.01
4	5.40	13	4.86
5	5.17	14	4.72
6	5.00	15	4.60
7	4.91	16	4.48
8	4.73	17	4.86
9	4.55	18	5.15

9.8. Shown below is the Excel output for a regression analysis to predict the number of business bankruptcy filings over a 16-year period by the number of consumer bankruptcy filings. How strong is the model? Note the residuals. Compute a Durbin-Watson statistic from the data and discuss the presence of autocorrelation in this model.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.529
R Square	0.280
Adjusted R Square	0.228
Standard Error	8179.84
Observations	16

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	364069877.4	364069877.4	5.44	0.0351
Residual	14	936737379.6	66909812.8		
Total	15	1300807257			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	75532.43621	4980.08791	15.17	0.0000
Year	−0.01574	0.00675	−2.33	0.0351

RESIDUAL OUTPUT

Observation	*Predicted Bus. Bankruptcies*	*Residuals*
1	70638.58	−1338.6
2	71024.28	−8588.3
3	71054.61	−7050.6
4	70161.99	1115.0
5	68462.72	12772.3
6	67733.25	14712.8
7	66882.45	−3029.4
8	65834.05	−2599.1
9	64230.61	622.4
10	61801.70	9747.3
11	61354.16	9288.8
12	62738.76	−434.8
13	63249.36	−10875.4
14	61767.01	−9808.0
15	57826.69	−4277.7
16	54283.80	−256.8

9.8 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

9.1. (a) The linear model: Yield = 9.96 – 0.14 Month

$F = 219.24$ $p = .000$ $R^2 = 90.9$ $s_e = .3212$

The quadratic model: Yield = 10.4 – 0.252 Month + .00445 Month²

$F = 176.21$ $p = .000$ $R^2 = 94.4\%$ $s_e = .2582$

In the quadratic model, both t ratios are significant, for x: $t = -7.93$, $p = .000$ and for x^2d: $t = 3.61$, $p = .002$.

The linear model is a strong model. The quadratic term adds some predictability but has a smaller t ratio than does the linear term.

(b)

x	F	e
10.08	–	–
10.05	–	–
9.24	–	–
9.23	–	–
9.69	9.65	.04
9.55	9.55	.00
9.37	9.43	.06
8.55	9.46	.91
8.36	9.29	.93
8.59	8.96	.37
7.99	8.72	.73
8.12	8.37	.25
7.91	8.27	.36
7.73	8.15	.42
7.39	7.94	.55
7.48	7.79	.31
7.52	7.63	.11
7.48	7.53	.05
7.35	7.47	.12
7.04	7.46	.42
6.88	7.35	.47
6.88	7.19	.31
7.17	7.04	.13
7.22	6.99	.23

$\Sigma|e| = 6.77$

$$\text{MAD} = \frac{6.77}{20} = .3385$$

NOTES

(c)

x	$\alpha = .3$ F	$\lvert e \rvert$	$\alpha = .7$ F	$\lvert e \rvert$
10.08	–	–	–	–
10.05	10.08	.03	10.08	.03
9.24	10.07	.83	10.06	.82
9.23	9.82	.59	9.49	.26
9.69	9.64	.05	9.31	.38
9.55	9.66	.11	9.58	.03
9.37	9.63	.26	9.56	.19
8.55	9.55	1.00	9.43	.88
8.36	9.25	.89	8.81	.45
8.59	8.98	.39	8.50	.09
7.99	8.86	.87	8.56	.57
8.12	8.60	.48	8.16	.04
7.91	8.46	.55	8.13	.22
7.73	8.30	.57	7.98	.25
7.39	8.13	.74	7.81	.42
7.48	7.91	.43	7.52	.04
7.52	7.78	.26	7.49	.03
7.48	7.70	.22	7.51	.03
7.35	7.63	.28	7.49	.14
7.04	7.55	.51	7.39	.35
6.88	7.40	.52	7.15	.27
6.88	7.24	.36	6.96	.08
7.17	7.13	.04	6.90	.27
7.22	7.14	.08	7.09	.13
		$\Sigma\lvert e \rvert = 10.06$		$\Sigma\lvert e \rvert = 5.97$

$$\text{MAD}_{\alpha=.3} = \frac{10.06}{23} = .4374 \qquad \text{MAD}_{\alpha=.7} = \frac{5.97}{23} = .2596$$

$\alpha = .7$ produces better forecasts based on MAD.

(d) MAD for b) .3385, c) .4374 and .2596. Exponential smoothing with $\alpha = .7$ produces the lowest error (.2596 from part c).

9.2. **(a)** Moving average **(b)** $\alpha = .2$

Year	Quantity	F	$\lvert e \rvert$	F	$\lvert e \rvert$
1	6137				
2	7019			6137.00	
3	7391			6313.40	
4	8750	6849.00	1901.00	6528.92	2221.08
5	9816	7720.00	2096.00	6973.14	2842.86
6	9644	8652.33	991.67	7541.71	2102.29
7	9951	9403.33	547.67	7962.17	1988.83
8	9971	9803.67	167.33	8359.93	1611.07
9	10089	9855.33	233.67	8682.15	1406.85
10	9693	10,003.67	310.67	8963.52	729.48
11	9380	9917.67	537.67	9109.42	270.59
12	9615	9720.67	105.67	9163.53	451.47
13	8992	9562.67	570.67	9253.82	261.82
14	9089	9329.00	240.00	9201.46	112.46
15	8876	9232.00	356.00	9178.97	302.97
16	9290	8985.67	304.33	9118.38	171.63
17	9250	9085.00	165.00	9152.70	97.30
18	9315	9138.67	176.33	9172.16	142.84
19	9424	9285.00	139.00	9200.73	223.27
20	9379	9329.67	49.33	9245.38	133.62
21	9180	9372.67	192.67	9272.11	92.11
22	9026	9327.67	301.67	9253.68	227.68
23	7953	9195.00	1242.00	9208.15	1255.15
24	7875	8719.67	844.67	8957.12	1082.12
25	7994	8284.67	290.67	8740.69	746.69
		$\Sigma\lvert e \rvert = 11{,}763.69$			$\Sigma\lvert e \rvert = 18{,}474.18$

$$\text{MAD}_{\text{moving average}} = \frac{\sum|e|}{\text{number forecasts}} = \frac{11{,}763.69}{22} = 534.71$$

$$\text{MAD}_{\alpha=.2} = \frac{\sum|e|}{\text{number forecasts}} = \frac{18{,}474.18}{22} = 839.74.$$

(c) The three-year moving average produced a smaller MAD (534.71) than did exponential smoothing with $\alpha = .2$ (MAD = 838.74). Using MAD as the criterion, the three-year moving average was a better forecasting tool than the exponential smoothing with $\alpha = .2$.

9.3. $\hat{y} = 9.5382 - 0.2716x$

$\hat{y}(7) = 7.637$

$R^2 = 40.2\%$ $F = 12.78, p = .002$

$s_e = 0.264862$

Durbin-Watson:

$n = 21$ $k = 1$ $\alpha = .05$

$D = 0.44$

$d_L = 1.22$ and $d_U = 1.42$

Since $D = 0.44 < d_L = 1.22$, the decision is to reject the null hypothesis. There is significant autocorrelation.

9.4.

	Qtr	TSCI	4qrtot	8qrtot	TC	SI	TCI	T
Year 1	1	54.019						
	2	56.495						
			213.574					
	3	50.169		425.044	53.131	94.43	51.699	53.722
			211.470					
	4	52.891		421.546	52.693	100.38	52.341	55.945
			210.076					
Year 2	1	51.915		423.402	52.925	98.09	52.937	58.274
			213.326					
	2	55.101		430.997	53.875	102.28	53.063	60.709
			217.671					
	3	53.419		440.490	55.061	97.02	55.048	63.249
			222.819					
	4	57.236		453.025	56.628	101.07	56.641	65.895
			230.206					
Year 3	1	57.063	237.160	467.366	58.421	97.68	58.186	68.646
	2	62.488		480.418	60.052	104.06	60.177	71.503
			243.258					
	3	60.373		492.176	61.522	98.13	62.215	74.466
			248.918					
	4	63.334		503.728	62.966	100.58	62.676	77.534
			254.810					
Year 4	1	62.723		512.503	64.063	97.91	63.957	80.708
			257.693					
	2	68.380		518.498	64.812	105.51	65.851	83.988
			260.805					
	3	63.256		524.332	65.542	96.51	65.185	87.373
			263.527					
	4	66.446		526.685	65.836	100.93	65.756	90.864
			263.158					
Year 5	1	65.445		526.305	65.788	99.48	66.733	94.461
			263.147					
	2	68.011		526.720	65.840	103.30	65.496	98.163
			263.573					
	3	63.245		521.415	65.177	97.04	65.174	101.971
			257.842					
	4	66.872		511.263	63.908	104.64	66.177	105.885
			253.421					
Year 6	1	59.714		501.685	62.711	95.22	60.889	109.904
			248.264					
	2	63.590		491.099	61.387	103.59	61.238	114.029
	3	58.088						
	4	61.443						

Quarter	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Index
1		98.09	97.68	97.91	99.48	95.22	97.89
2		102.28	104.06	105.51	103.30	103.59	103.65
3	94.43	97.02	98.13	96.51	97.04		96.86
4	100.38	101.07	100.58	100.93	104.64		100.86
Total							399.26

Adjust the seasonal indexes by: $\frac{400}{399.26} = 1.00185343$

Adjusted seasonal indexes:

Quarter	Index
1	98.07
2	103.84
3	97.04
4	101.05
Total	400.00

9.5. $R^2 = 55.8\%$ $F = 8.83$ with $p = .021$

$s_e = 50.18$

This model with a lag of one year has modest predictability. The overall F is significant at $\alpha = .05$ but not at $\alpha = .01$.

9.6. The regression equation is:

Assets = –532 + 68.0 Accounts

$R^2 = 34.3\%$ $s_e = 878.575$

$D = 0.08$

For $n = 27$ and $\alpha = .01$, $d_L = 1.10$ and $d_U = 1.24$.

Since $D = 0.09 < d_L = 1.10$ the null hypothesis is rejected. There is significant autocorrelation in this model.

9.7.

	$\alpha = .1$		$\alpha = .5$		$\alpha = .8$			
Year	PurPwr	F	|e|	F	|e|	F	|e|
1	6.04							
2	5.92	6.04	.12	6.04	.12	6.04	.12
3	5.57	6.03	.46	5.98	.41	5.94	.37
4	5.40	5.98	.58	5.78	.38	5.64	.24
5	5.17	5.92	.75	5.59	.42	5.45	.28
6	5.00	5.85	.85	5.38	.38	5.23	.23
7	4.91	5.77	.86	5.19	.28	5.05	.14
8	4.73	5.68	.95	5.05	.32	4.94	.21
9	4.55	5.59	1.04	4.89	.34	4.77	.22
10	4.34	5.49	1.15	4.72	.38	4.59	.25
11	4.67	5.38	.71	4.53	.14	4.39	.28
12	5.01	5.31	.30	4.60	.41	4.61	.40
13	4.86	5.28	.42	4.81	.05	4.93	.07
14	4.72	5.24	.52	4.84	.12	4.87	.15
15	4.60	5.19	.59	4.78	.18	4.75	.15
16	4.48	5.13	.65	4.69	.21	4.63	.15
17	4.86	5.07	.21	4.59	.27	4.51	.35
18	5.15	5.05	.10	4.73	.42	4.79	.36
			$\Sigma\|e\| = 10.26$		$\Sigma\|e\| = 4.83$		$\Sigma\|e\| = 3.97$

$$MAD_1 = \frac{\sum|e|}{N} = \frac{10.26}{17} = .60$$

$$MAD_2 = \frac{\sum|e|}{N} = \frac{4.83}{17} = .28$$

$$MAD_3 = \frac{\sum|e|}{N} = \frac{3.97}{17} = .23$$

The smallest mean absolute deviation error is produced using $\alpha = .8$.

The forecast for year 19 is: $F(19) = (.8)(5.15) + (.2)(4.79) = 5.08$

9.8. The model is: Bankruptcies = 75,532.436 – 0.016 Year

Since $R^2 = .28$ and the adjusted $R^2 = .23$, this is a weak model.

e_t	$e_t - e_{t-1}$	$(e_t - e_{t-1})^2$	e_t^2
–1,338.58			1,791,796
–8,588.28	–7,249.7	52,558,150	73,758,553
–7,050.61	1,537.7	2,364,521	49,711,101
1,115.01	8,165.6	66,677,023	1,243,247
12,772.28	11,657.3	135,892,643	163,131,136
14,712.75	1,940.5	3,765,540	216,465,013
–3,029.45	–17,742.2	314,785,661	9,177,567
–2,599.05	430.4	185,244	6,755,061
622.39	3,221.4	10,377,418	387,369
9,747.30	9,124.9	83,263,800	95,009,857
9,288.84	– 458.5	210,222	86,282,549
–434.76	– 9,723.6	94,548,397	189,016
–10,875.36	–10,440.6	109,006,128	118,273,455
–9,808.01	1,067.4	1,139,343	96,197,060
–4,277.69	5,530.3	30,584,218	18,298,632
– 256.80	4,020.9	16,167,637	65,946

$\sum(e_t - e_{t-1})^2 = 921,525,945$ $\sum e_t^2 = 936,737,358$

$$D = \frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2} = \frac{921,525,945}{936,737,358} = 0.98$$

For $n = 16$, $\alpha = .05$, $d_L = 1.10$ and $d_U = 1.37$

Since $D = 0.98 < d_L = 1.10$, the decision is to reject the null hypothesis and conclude that there is significant autocorrelation.

9.9 ANSWERS AND HINTS

ANSWERS FOR SELF ASSESSMENT QUESTIONS

Topics	Q. No.	Answers
9.2 Forecasting	1.	Forecasting
	2.	trend
	3.	Cycles
	4.	stationary
	5.	True
	6.	True

Topics	Q. No.	Answers	
9.3 Smoothing Techniques	7.	moving	
	8.	weighted	
	9.	True	
	10.	True	
9.4 Trend Analysis	11.	time	
	12.	seasonal	
	13.	True	
	14.	True	
9.5 Autocorrelation and Autoregression	15.	Autocorrelation	
	16.	Autoregression	
	17.	True	
	18.	False	Serial correlation occurs in data when the error terms of a regression forecasting model are correlated.
	19.	True	

10

CHAPTER

DECISION ANALYSIS

CONTENTS

LEARNING OBJECTIVES

This chapter describes how to use decision analysis to improve management decisions, thereby enabling you to:

- Make decisions under certainty by constructing a decision table.
- Make decisions under uncertainty using the maximax criterion, the maximin criterion, the Hurwicz criterion, and minimax regret.
- Make decisions under risk by constructing decision trees, calculating expected monetary value and expected value of perfect information, and analyzing utility.

10.1 THE DECISION TABLE AND DECISION MAKING UNDER CERTAINTY

Many decision analysis problems can be viewed as having three variables: decision alternatives, states of nature, and payoffs.

Decision alternatives are *the various choices or options available to the decision maker in any given problem situation.* On most days, financial managers face the choices of whether to invest in blue chip stocks, bonds, commodities, certificates of deposit, money markets, annuities, and other investments. Construction decision makers must decide whether to concentrate on one building job today, spread out workers and equipment to several jobs, or not work today. In virtually every possible business scenario, decision alternatives are available. A good decision maker identifies many options and effectively evaluates them.

States of nature are *the occurrences of nature that can happen after a decision is made that can affect the outcome of the decision and over which the decision maker has little or no control.* These states of nature can be literally natural atmospheric and climatic conditions or they can be such things as the business climate, the political climate, the worker climate, or the condition of the marketplace, among many others. The financial investor faces such states of nature as the prime interest rate, the condition of the stock market, the international monetary exchange rate, and so on. A construction company is faced with such states of nature as the weather, wildcat strikes, equipment failure, absenteeism, and supplier inability to deliver on time. States of nature are usually difficult to predict but are important to identify in the decision-making process.

The **payoffs** of a decision analysis problem are *the benefits or rewards that result from selecting a particular decision alternative.* Payoffs are usually given in terms of dollars. In the financial investment industry, for example, the payoffs can be small, modest, or large, or the investment can result in a loss. Most business decisions involve taking some chances with personal or company money in one form or another. Because for-profit businesses are looking for a return on the dollars invested, the payoffs are extremely important for a successful manager. The trick is to determine which decision alternative to take in order to generate the greatest payoff. Suppose a CEO is examining various environmental decision alternatives. Positive payoffs could include increased market share, attracting and retaining quality employees, consumer appreciation, and governmental support. Negative payoffs might take the form of fines and penalties, lost market share, and lawsuit judgments.

Decision Table

The concepts of decision alternatives, states of nature, and payoffs can be examined jointly by using a **decision table**, or **payoff table. Table 10.1** shows the structure of a decision table. On the left side of the table are the various decision alternatives, denoted by d_i. Along the top row are the states of nature, denoted by s_j. In the middle of the table are the various payoffs for each decision alternative under each state of nature, denoted by P_{ij}.

As an example of a decision table, consider the decision dilemma of the investor shown in **Table 10.2**. The investor is faced with the decision of where and how to invest $10,000 under several possible states of nature.

The investor is considering four decision alternatives.

1. Invest in the stock market
2. Invest in the bond market

TABLE 10.1: DECISION TABLE

		State of Nature				
		s_1	s_2	s_3	$\cdots$	s_n
Decision Alternative	d_1	$P_{1,1}$	$P_{1,2}$	$P_{1,3}$	$\cdots$	$P_{1,n}$
	d_2	$P_{2,1}$	$P_{2,2}$	$P_{2,3}$	$\cdots$	$P_{2,n}$
	d_3	$P_{3,1}$	$P_{3,2}$	$P_{3,3}$	$\cdots$	$P_{3,n}$
	.	.	.	.	.	
	.	.	.	.	$\cdots$	.
	.	.	.	.	$\cdots$	.
	d_m	P_{m1}	P_{m2}	P_{m3}	$\cdots$	P_{mn}

where
s_j = state of nature
d_i = decision alternative
P_{ij} = payoff for decision i under state j

TABLE 10.2: YEARLY PAYOFFS ON AN INVESTMENT OF $10,000

		State of the Economy		
		Stagnant	*Slow Growth*	*Rapid Growth*
Investment Decision Alternative	*Stocks*	–$500	$700	$2,200
	Bonds	–$100	$600	$900
	CDs	$300	$500	$750
	Mixture	–$200	$650	$1,300

3. Invest in government certificates of deposit (CDs)
4. Invest in a mixture of stocks and bonds

Because the payoffs are in the future, the investor is unlikely to know ahead of time what the state of nature will be for the economy. However, the table delineates three possible states of the economy.

1. A stagnant economy
2. A slow-growth economy
3. A rapid-growth economy

The matrix in Table 10.2 lists the payoffs for each possible investment decision under each possible state of the economy. Notice that the largest payoff comes with a stock investment under a rapid-growth economic scenario, with a payoff of $2,200 per year on an investment of $10,000. The lowest payoff occurs for a stock investment during stagnant economic times, with an annual loss of $500 on the $10,000 investment.

Decision Making Under Certainty

The most elementary of the decision-making scenarios is **decision making under certainty.** In making decisions under certainty, *the states of nature are known.* The decision maker needs merely to examine the payoffs under different decision alternatives and select the alternative with the largest payoff. In the preceding example involving the $10,000 investment, if it is known that the economy is going to be stagnant, the investor would select the decision alternative of CDs, yielding a payoff of $300. Indeed, each of the other three decision alternatives would result in a loss under stagnant economic conditions. If it is known that the economy is going to have slow growth, the investor would choose stocks as an investment, resulting in a $700 payoff. If the economy is certain to have rapid growth, the decision maker should opt for stocks, resulting in a payoff of $2,200. Decision making under certainty is almost the trivial case.

SELF ASSESSMENT QUESTIONS

Fill in the Blanks

1. In decision analysis, decision-making scenarios are divided into three categories: decision-making under _____, decision-making under _____, and decision-making under _____.
2. Many decision analysis problems can be viewed as having three variables: 1. _____, 2. _____, and 3. _____.
3. Occurrences of nature that can happen after a decision has been made that can effect the outcome of the decision and over which the decision-maker has little or no control are called _____.

State whether the following statements are true/false:

4. In a decision analysis problem, variables (such as general macroeconomic conditions) which are not under the decision maker's control are called prior probabilities.

SELF ASSESSMENT QUESTIONS

5. In a decision analysis problem, variables (such as investing in common stocks or corporate bonds) which are under the decision maker's control are called decision alternatives.
6. In a decision analysis problem, variables (such as benefits or rewards that result from investments in common stocks or corporate bonds and from a new product launch) which result from selecting a particular decision alternative are called posterior probabilities.

ACTIVITY

A local B2B logistics company has an option to deliver 2000, 3000, and 4000 packets daily with its existing delivery and supply chain network. The profit realized after deducting all the costs associated with procurement and delivery on each packet is $10.

Prepare a payoff matrix when on any given day, the logistics company receives service request from 0, 2000, 3000, and 4000 businesses for delivery of packets.

10.2 DECISION MAKING UNDER UNCERTAINTY

In making decisions under certainty, the decision maker knows for sure which state of nature will occur, and he or she bases the decision on the optimal payoff available under that state. **Decision making under uncertainty** occurs *when it is unknown which states of nature will occur and the probability of a state of nature occurring is also unknown.* Hence, the decision maker has virtually no information about which state of nature will occur, and he or she attempts to develop a strategy based on payoffs.

Several different approaches can be taken to making decisions under uncertainty. Each uses a different decision criterion, depending on the decision maker's outlook. Each of these approaches will be explained and demonstrated with a decision table. Included are the maximax criterion, maximin criterion, Hurwicz criterion, and minimax regret.

In section 10.1, we discussed the decision dilemma of the financial investor who wants to invest $10,000 and is faced with four decision alternatives and three states of nature. The data for this problem were given in Table 10.2. In decision making under certainty, we selected the optimal payoff under each state of the economy and then, on the basis of which state we were certain would occur, selected a decision alternative. Shown next are techniques to use when we are uncertain which state of nature will occur.

10.2.1 MAXIMAX CRITERION

The **maximax criterion** approach is an optimistic approach in which the decision maker bases action on a notion that the best things will happen.

The decision maker *isolates the maximum payoff under each decision alternative and then selects the decision alternative that produces the highest of these maximum payoffs.* The name "maximax" means selecting the maximum overall payoff from the maximum payoffs of each decision alternative. Consider the $10,000 investment problem. The maximum payoff is $2,200 for stocks, $900 for bonds, $750 for CDs, and $1,300 for the mixture of investments. The maximax criterion approach requires that the decision maker select the maximum payoff of these four.

		State of the Economy			
		Stagnant	*Slow Growth*	*Rapid Growth*	*Maximum*
Investment Decision Alternative	*Stocks*	–$500	$700	$2,200	$2,200
	Bonds	–$100	$600	$900	$900
	CDs	$300	$500	$750	$750
	Mixture	–$200	$650	$1,300	$1,300
	maximum of {$2,200, $900, $750, $1,300} = $2,200				

Because the maximax criterion results in $2,200 as the optimal payoff, the decision alternative selected is the stock alternative, which is associated with the $2,200.

10.2.2 MAXIMIN CRITERION

The **maximin criterion** approach to decision making is a pessimistic approach. The assumption is that the worst will happen and attempts must be made to minimize the damage. The decision maker starts by examining the payoffs under each decision alternative and selects the worst, or minimum, payoff that can occur under that decision. Then the decision maker *selects the maximum or best payoff of those minimums selected under each decision alternative.* Thus, the decision maker has maximized the minimums. In the investment problem, the minimum payoffs are –$500 for stocks, –$100 for bonds, $300 for CDs, and –$200 for the mixture of investments. With the maximin criterion, the decision maker examines the minimum payoffs for each decision alternative given in the last column and selects the maximum of those values.

		State of the Economy			
		Stagnant	*Slow Growth*	*Rapid Growth*	*Minimum*
Investment Decision Alternative	*Stocks*	–$500	$700	$2,200	–$500
	Bonds	–$100	$600	$900	–$100
	CDs	$300	$500	$750	$300
	Mixture	–$200	$650	$1,300	–$200
	maximum of {–$500, –$100, $300, –$200} = $300				

The decision is to invest in CDs because that investment alternative yields the highest, or maximum, payoff under the worst-case scenario.

10.2.3 HURWICZ CRITERION

The Hurwicz criterion is an approach somewhere between the maximax and the maximin approaches. The **Hurwicz criterion** approach selects the maximum and the minimum payoff from each decision alternative. A value called alpha (not the same as the probability of a Type I error), which is between 0 and 1, is selected as a weight of optimism. The nearer alpha is to 1, the more optimistic is the decision maker. The use of alpha values near 0 implies a more pessimistic approach. The maximum payoff under each decision alternative is multiplied by alpha and the minimum payoff (pessimistic view) under each decision alternative is multiplied by $1 - \alpha$ (weight of pessimism). These weighted products are summed for each decision alternative, resulting in a weighted value for each decision alternative. The maximum weighted value is selected, and the corresponding decision alternative is chosen.

Following are the data for the investment example, along with the minimum and maximum values.

		State of the Economy				
		Stagnant	*Slow Growth*	*Rapid Growth*	*Minimum*	*Maximum*
Investment Decision Alternative	*Stocks*	–$500	$700	$2,200	–$500	$2,200
	Bonds	–$100	$600	$900	–$100	$900
	CDs	$300	$500	$750	$300	$750
	Mixture	–$200	$650	$1,300	–$200	$1,300

Suppose we are more optimistic than pessimistic and select $\alpha = .7$ for the weight of optimism. The calculations of weighted values for the decision alternative follow.

Stocks	($2,200)(.7) + (–$500)(.3)	= $1,390
Bonds	($900)(.7) + (–$100)(.3)	= $ 600
CDs	($750)(.7) + ($300)(.3)	= $ 615
Mixture	($1,300)(.7) + (–$200)(.3)	= $ 850

The Hurwicz criterion leads the decision maker to choose the maximum of these values, $1,390. The result under the Hurwicz criterion with $\alpha = .7$ is to choose stocks as the decision alternative. An advantage of the Hurwicz criterion is that it allows the decision maker the latitude to explore various weights of optimism. A decision maker's outlook might change from scenario to scenario and from day to day. In this case, if we had been fairly pessimistic and chosen an alpha of .2, we would have obtained the following weighted values.

Stocks	($2,200)(.2) + (–$500)(.8)	= $ 40
Bonds	($900)(.2) + (–$100)(.8)	= $100
CDs	($750)(.2) + ($300)(.8)	= $390
Mixture	($1,300)(.2) + (–$200)(.8)	= $100

Under this scenario, the decision maker would choose the CD option because it yields the highest weighted payoff ($390) with $\alpha = .2$.

Table 10.3 displays the payoffs obtained by using the Hurwicz criterion for various values of alpha for the investment example. The circled values are the optimum payoffs and represent the decision alternative selection for that value of alpha. Note that for $\alpha = .0, .1, .2$, and .3, the decision is to invest in CDs. For $\alpha = .4$ to 1.0, the decision is to invest in stocks.

Figure 10.1 shows graphically the weighted values for each decision alternative over the possible values of alpha. The thicker line segments represent the maximum of these under each value of alpha. Notice that the graph reinforces the choice of CDs for $\alpha = .0, .1, .2, .3$ and the choice of stocks for $\alpha = .4$ through 1.0.

TABLE 10.3: DECISION ALTERNATIVES FOR VARIOUS VALUES OF ALPHA

		Stocks		Bonds		CDs		Mixture	
		Max.	Min.	Max.	Min.	Max.	Min.	Max.	Min.
α	$1-\alpha$	2,200	−500	900	−100	750	300	1,300	−200
.0	1.0		−500		−100	(300)			−200
.1	.9		−230		0	(345)			−50
.2	.8		40		100	(390)			100
.3	.7		310		200	(435)			250
.4	.6	(580)			300		480		400
.5	.5	(850)			400		525		550
.6	.4	(1,120)			500		570		700
.7	.3	(1,390)			600		615		850
.8	.2	(1,660)			700		660		1,000
.9	.1	(1,930)			800		705		1,150
1.0	.0	(2,200)			900		750		1,300

Note: Circled values indicate the choice for the given value of alpha.

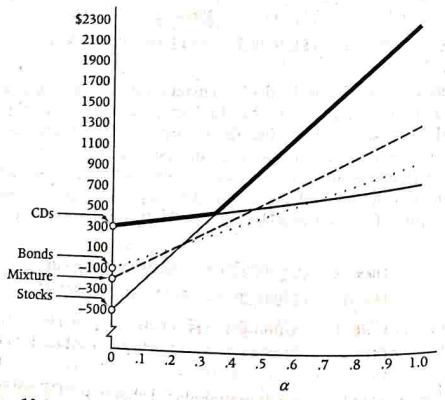


Figure 10.1: Graph of Hurwicz Criterion Selections for Various Values of Alpha

Between $\alpha = .3$ and $\alpha = .4$, there is a point at which the line for weighted payoffs for CDs intersects the line for weighted payoffs for stocks. By setting the alpha expression with maximum and minimum values of the CD investment equal to that of the stock investment, we can solve for the alpha value at which the intersection occurs. At this value of alpha, the weighted payoffs of the two investments under the Hurwicz criterion are equal, and the decision maker is indifferent as to which one he or she chooses.

$$\begin{aligned} \text{Stocks Weighted Payoff} &= \text{CDs Weighted Payoff} \\ 2{,}200(\alpha) + (-500)(1-\alpha) &= 750(\alpha) + (300)(1-\alpha) \\ 2{,}200\alpha - 500 + 500\alpha &= 750\alpha + 300 - 300\alpha \\ 2{,}250\alpha &= 800 \\ \alpha &= .3555 \end{aligned}$$

At $\alpha = .3555$, both stocks and CDs yield the same payoff under the Hurwicz criterion. For values less than $\alpha = .3555$, CDs are the chosen investment. For $\alpha > .3555$, stocks are the chosen investment. Neither bonds nor the mixture produces the optimum payoff under the Hurwicz criterion for any value of alpha. Notice that in Figure 10.1 the dark line segments represent the optimum solutions. The lines for both bonds and the mixture are beneath these optimum line segments for the entire range of α. In another problem with different payoffs, the results might be different.

10.2.4 MINIMAX REGRET

The strategy of **minimax regret** is based on lost opportunity. Lost opportunity occurs when a decision maker loses out on some payoff or portion of a payoff because he or she chose the wrong decision alternative. For example, if a decision maker selects decision alternative d_i, which pays $200, and the selection of alternative d_j would have yielded $300, the opportunity loss is $100.

$$\$300 - \$200 = \$100$$

In analyzing decision-making situations under uncertainty, an analyst can transform a decision table (payoff table) into an **opportunity loss table,** which can be used to apply the minimax regret criterion. Repeated here is the $10,000 investment decision table.

		State of the Economy		
		Stagnant	*Slow Growth*	*Rapid Growth*
Investment Decision Alternative	*Stocks*	−$500	$700	$2,200
	Bonds	−$100	$600	$900
	CDs	$300	$500	$750
	Mixture	−$200	$650	$1,300

Suppose the state of the economy turns out to be stagnant. The optimal decision choice would be CDs, which pay off $300. Any other decision would lead to an opportunity loss. The opportunity loss for each decision alternative other than CDs can be calculated by subtracting the decision alternative payoff from $300.

Stocks	$300 − (−$500) = $800
Bonds	$300 − (−$100) = $400
CDs	$300 − ($300) = $0
Mixture	$300 − (−$200) = $500

The opportunity losses for the slow-growth state of the economy are calculated by subtracting each payoff from $700, because $700 is the maximum payoff that can be obtained from this state; any other payoff is an opportunity loss. These opportunity losses follow.

Stocks	$700 − ($700) = $0
Bonds	$700 − ($600) = $100
CDs	$700 − ($500) = $200
Mixture	$700 − ($650) = $50

The opportunity losses for a rapid-growth state of the economy are calculated similarly.

Stocks	$2,200 − ($2,200) = $0
Bonds	$2,200 − ($900) = $1,300
CDs	$2,200 − ($750) = $1,450
Mixture	$2,200 − ($1,300) = $900

Replacing payoffs in the decision table with opportunity losses produces the opportunity loss table, as shown in **Table 10.4**.

TABLE 10.4: OPPORTUNITY LOSS TABLE

		State of the Economy		
		Stagnant	*Slow Growth*	*Rapid Growth*
Investment Decision Alternative	*Stocks*	$800	$0	$0
	Bonds	$400	$100	$1,300
	CDs	$0	$200	$1,450
	Mixture	$500	$50	$900

After the opportunity loss table is determined, the decision maker examines the lost opportunity, or regret, under each decision, and selects the maximum regret for consideration. For example, if the investor chooses stocks, the maximum regret or lost opportunity is $800. If the investor chooses bonds, the maximum regret is $1,300. If the investor chooses CDs, the maximum regret is $1,450. If the investor selects a mixture, the maximum regret is $900.

In making a decision based on a minimax regret criterion, the decision maker examines the maximum regret under each decision alternative and selects the minimum of these. The result is the stocks option, which has the minimum regret of $800. An investor who wants to minimize the maximum regret under the various states of the economy will choose to invest in stocks under the minimax regret strategy.

DEMONSTRATION PROBLEM 10.1

A manufacturing company is faced with a capacity decision. Its present production facility is running at nearly maximum capacity. Management is considering the following three capacity decision alternatives.

1. No expansion
2. Add on to the present facility
3. Build a new facility

The managers believe that if a large increase occurs in demand for their product in the near future, they will need to build a new facility to compete and capitalize on more efficient technological and design advances. However, if demand does not increase, it might be more profitable to maintain the present facility and add no capacity. A third decision alternative is to add on to the present facility, which will suffice for a moderate increase in demand and will be cheaper than building an entirely new facility. A drawback of adding to the old facility is that if there is a large demand for the product, the company will be unable to capitalize on new technologies and efficiencies, which cannot be built into the old plant.

The following decision table shows the payoffs (in $ millions) for these three decision alternatives for four different possible states of demand for the company's product (less demand, same demand, moderate increase in demand, and large increase in demand). Use these data to determine which decision alternative would be selected by the maximax criterion and the maximin criterion. Use $\alpha = .4$ and the Hurwicz criterion to determine the decision alternative. Calculate an opportunity loss table and determine the decision alternative by using the minimax regret criterion.

		State of Demand			
		Less	No Change	Moderate Increase	Large Increase
Capacity Decision	No Expansion	−$3	$2	$3	$6
	Add On	−$40	−$28	$10	$20
	Build a New Facility	−$210	−$145	−$5	$55

Solution: The maximum and minimum payoffs under each decision alternative follow.

	Maximum	Minimum
No Expansion	$ 6	−$ 3
Add On	$20	−$ 40
Build a New Facility	$55	−$210

Using the maximax criterion, the decision makers select the maximum of the maximum payoffs under each decision alternative. This value is the maximum of {$6, $20, $55} = $55, or the selection of the decision alternative of building a new facility and maximizing the maximum payoff ($55).

Using the maximin criterion, the decision makers select the maximum of the minimum payoffs under each decision alternative. This value is the maximum of {−$3, −$40, −$210} = −$3. They select the decision alternative of no expansion and maximize the minimum payoff (−$3).

Following are the calculations for the Hurwicz criterion with $\alpha = .4$.

No Expansion	$6(.4) + (-\$3)(.6)$	=	$0.60
Add On	$20(.4) + (-\$40)(.6)$	=	−$16.00
Build a New Facility	$55(.4) + (-\$210)(.6)$	=	−$104.00

Using the Hurwicz criterion, the decision makers would select no expansion as the maximum of these weighted values ($.60).

Following is the opportunity loss table for this capacity choice problem. Note that each opportunity loss is calculated by taking the maximum payoff under each state of nature and subtracting each of the other payoffs under that state from that maximum value.

		State of Demand			
		Less	No Change	Moderate Increase	Large Increase
Capacity Decision	No Expansion	$0	$0	$7	$49
	Add On	$37	$30	$0	$35
	Build a New Facility	$207	$147	$15	$0

Using the minimax regret criterion on this opportunity loss table, the decision makers first select the maximum regret under each decision alternative.

Decision Alternative	Maximum Regret
No Expansion	49
Add On	37
Build a New Facility	207

Next, the decision makers select the decision alternative with the minimum regret, which is to add on, with a regret of $37.

SELF ASSESSMENT QUESTIONS

Fill in the Blanks

7. Examine the decision table shown below:

		State of Nature			
		1	2	3	4
Decision Alternative	1	−50	−25	75	125
	2	10	15	20	25
	3	−20	−5	10	20

The selected decision alternative using a Maximax criterion is _____ and the optimal payoff is _____.

8. Use the decision table from Question 7. The selected decision alternative using a Maximin criterion is _____ and the payoff for this is _____. Suppose Hurwicz criterion is used to select a decision alternative and α is .3. The selected decision alternative is _____ and the payoff is _____. However, if α is .8, the selected decision alternative is _____ and the payoff is _____.

9. Use the decision table from Question 7 to construct an Opportunity Loss table. Using this table and a Minimax Regret criterion, the selected decision alternative is _____ and the minimum regret is _____.

State whether the following statements are true/false:

10. In a decision-making scenario, if it is not known which of the states of nature will occur and further if the probabilities of occurrence of the states are also unknown the scenario is called decision-making under double risk.

11. In a decision-making under uncertainty scenario, the decision maker chooses the decision alternative that has the minimum expected (i.e., probability-weighted) payoff among all the available alternatives.

12. In a decision-making under uncertainty scenario, the decision maker attempts to develop a strategy based on payoffs since virtually no information is available about which state of nature will occur.

ACTIVITY

A domestic airline company is struggling against decline in its sales revenue. To counter this problem, it came up with four alternatives: (D_1) operating new routes; (D_2) increased frequency on the existing routes; (D_3) advertisements; and (D_4) sale and booking offers. Owing to these decision alternatives, the expected increase in the sales revenue.could be: (S_1) 30%; (S_2) 22%; (S_3) 14% and (S_4) 6%.

The following table presents the yearly revenue (in million $) which the airline would get if any of the four strategies are selected.

Analyze this case under the approaches (*except Hurwicz approach*) of decision making with uncertainty and suggest which alternative shall be chosen by the decision maker. Assume $\alpha = 0.2$

Annual Revenue of the Domestic Airline.

		States of Nature			
		S_1	S_2	S_3	S_4
Decision Alternatives	D_1	$50	$35	$20	$10
	D_2	$70	$20	$50	$20
	D_3	$20	$60	$60	$40
	D_4	$60	$40	$80	$50

10.3 DECISION MAKING UNDER RISK

In Section 10.1 we discussed making decisions in situations where it is certain which states of nature will occur. In section 10.2, we examined several strategies for making decisions when it is uncertain which state of nature will occur. In this section we examine decision making under risk. **Decision making under risk** occurs *when it is uncertain which states of nature will occur but the probability of each state of nature occurring has been determined.* Using these probabilities, we can develop some additional decision-making strategies.

In preceding sections, we discussed the dilemma of how best to invest $10,000. Four investment decision alternatives were identified and three states of the economy seemed possible (stagnant economy, slow-growth economy, and rapid-growth economy). Suppose we determine that there is a .25 probability of a stagnant economy, a .45 probability of a slow-growth economy, and a .30 probability of a rapid-growth economy. In a decision table, or payoff table, we place these probabilities next to each state of nature. **Table 10.5** is a decision table for the investment example shown in Table 10.2 with the probabilities given in parentheses.

Decision Trees

Another way to depict the decision process is through the use of decision trees. **Decision trees** have a □ node to represent decision alternatives and a O node to represent states of nature. If probabilities are available for states of nature, they are assigned to the line segment following the state-of-nature node symbol, O. Payoffs are displayed at the ends of the decision tree limbs. **Figure 10.2** is a decision tree for the financial investment example given in Table 10.5.

TABLE 10.5: DECISION TABLE WITH STATE OF NATURE PROBABILITIES

		State of the Economy		
		Stagnant (.25)	*Slow Growth (.45)*	*Rapid Growth (.30)*
Investment Decision Alternative	*Stocks*	−$500	$700	$2,200
	Bonds	−$100	$600	$ 900
	CDs	$300	$500	$ 750
	Mixture	−$200	$650	$1,300

Expected Monetary Value (EMV)

One strategy that can be used in making decisions under risk is the **expected monetary value (EMV)** approach. A person who uses this approach is sometimes referred to as an **EMVer.** The expected monetary value of each decision alternative is calculated by multiplying the probability of each state of nature by the state's associated payoff and summing these products across the states of nature for each decision alternative,

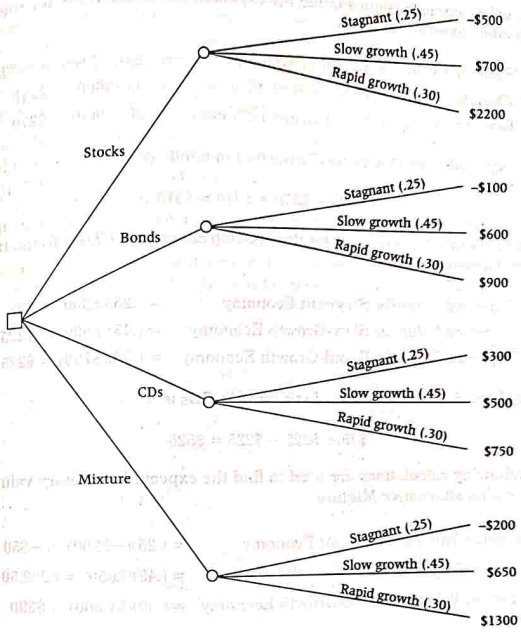


Figure 10.2: Decision Tree for the Investment Example

producing an expected monetary value for each decision alternative. The decision maker compares the expected monetary values for the decision alternatives and selects the alternative with the highest expected monetary value.

As an example, we can compute the expected monetary value for the $10,000 investment problem displayed in Table 10.5 and Figure 10.2 with the associated probabilities. We use the following calculations to find the expected monetary value for the decision alternative *Stocks*.

Expected Value for Stagnant Economy = (.25)(−$500) = −$125
Expected Value for Slow-Growth Economy = (.45)($700) = $315
Expected Value for Rapid-Growth Economy = (.30)($2,200) = $660

The expected monetary value of investing in stocks is

$$-\$125 + \$315 + \$660 = \$850$$

The calculations for determining the expected monetary value for the decision alternative *Bonds* follow.

Expected Value for Stagnant Economy = (.25)(−$100) = −$25
Expected Value for Slow-Growth Economy = (.45)($600) = $270
Expected Value for Rapid-Growth Economy = (.30)($900) = $270

The expected monetary value of investing in bonds is

$$-\$25 + \$270 + \$270 = \$515$$

The expected monetary value for the decision alternative *CDs* is found by the following calculations.

Expected Value for Stagnant Economy = (.25)($300) = $75
Expected Value for Slow-Growth Economy = (.45)($500) = $225
Expected Value for Rapid-Growth Economy = (.30)($750) = $225

The expected monetary value of investing in CDs is

$$\$75 + \$225 + \$225 = \$525$$

The following calculations are used to find the expected monetary value for the decision alternative *Mixture*.

Expected Value for Stagnant Economy = (.25)(−$200) = −$50
Expected Value for Slow-Growth Economy = (.45)($650) = $292.50
Expected Value for Rapid-Growth Economy = (.30)($1,300) = $390

The expected monetary value of investing in a mixture is

$$-\$50 + \$292.50 + \$390 = \$632.50$$

A decision maker using expected monetary value as a strategy will choose the maximum of the expected monetary values computed for each decision alternative.

$$\text{Maximum of } \{\$850, \$515, \$525, \$632.5\} = \$850$$

The maximum of the expected monetary values is $850, which is produced from a stock investment. An EMVer chooses to invest in stocks on the basis of this information.

This process of expected monetary value can be depicted on decision trees like the one in Figure 10.2. Each payoff at the end of a branch of the tree is multiplied by the associated probability of that state of nature. The resulting products are summed across all states for a given decision choice, producing an expected monetary value for that decision alternative. These expected monetary values are displayed on the decision tree at the chance or state-of-nature nodes, ○.

The decision maker observes these expected monetary values. The optimal expected monetary value is the one selected and is displayed at the decision node in the tree, □. The decision alternative pathways leading to lesser, or nonoptimal, monetary values are marked with a double vertical line symbol, ||, to denote rejected decision alternatives. **Figure 10.3** depicts the EMV analysis on the decision tree in Figure 10.2.

The strategy of expected monetary value is based on a long-run average. If a decision maker could "play this game" over and over with the probabilities and payoffs remaining the same, he or she could *expect* to earn an average of $850 in the long run by choosing to invest in stocks. The reality is that for any *one* occasion, the investor will earn payoffs of either −$500, $700, or $2,200 on a stock investment, depending on which state of the economy occurs. The investor will not earn $850 at any *one* time on this decision, but he or she could average a profit of $850 if the investment

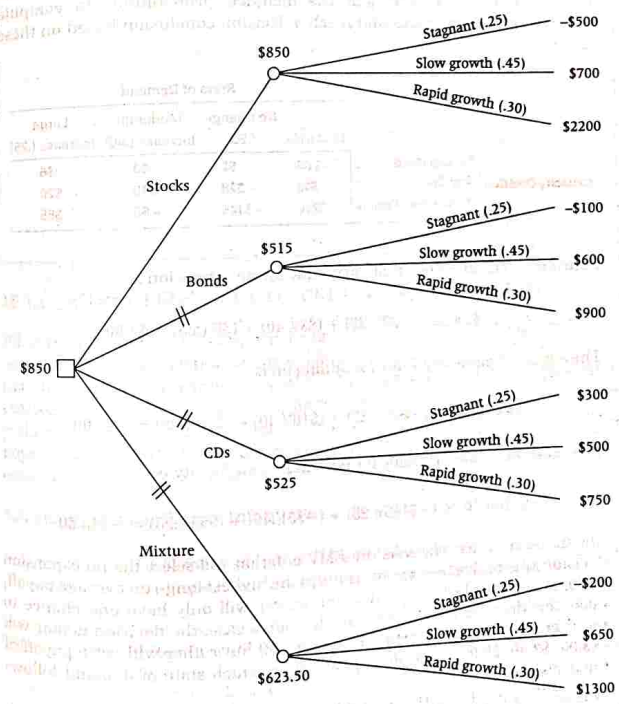


Figure 10.3: Expected Monetary Value for the Investment Example

continued through time. With an investment of this size, the investor will potentially have the chance to make this decision several times. Suppose, on the other hand, an investor has to decide whether to spend $5 million to drill an oil well. Expected monetary values might not mean as much to the decision maker if he or she has only enough financial support to make this decision once.

DEMONSTARATION PROBLEM 10.2

Recall the capacity decision scenario presented in Demonstration Problem 10.1. Suppose probabilities have been determined for the states of demand such that there is a .10 probability that demand will be less, a .25 probability that there will be no change in demand, a .40 probability that there will be a moderate increase in demand, and a .25 probability that there will be a large increase in demand. Use the data presented in the problem, which are restated here, and the included probabilities to compute expected monetary values and reach a decision conclusion based on these findings.

		State of Demand			
		Less (.10)	No Change (.25)	Moderate Increase (.40)	Large Increase (.25)
Capacity Decision	No Expansion	−$3	$2	$3	$6
	Add On	−$40	−$28	$10	$20
	Build a New Facility	−$210	−$145	−$5	$55

Solution: The expected monetary value for no expansion is

$$(-\$3)(.10) + (\$2)(.25) + (\$3)(.40) + (\$6)(.25) = \$2.90$$

The expected monetary value for adding on is

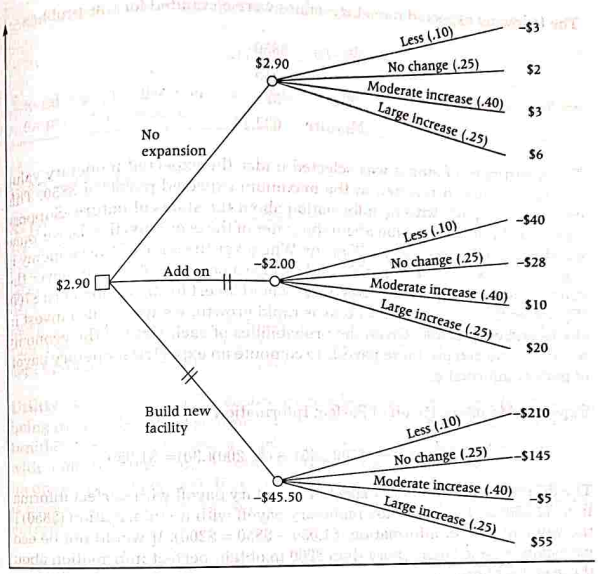
$$(-\$40)(.10) + (-\$28)(.25) + (\$10)(.40) + (\$20)(.25) = -\$2.00$$

The expected monetary value for building a new facility is

$$(-\$210)(.10) + (-\$145)(.25) + (-\$5)(.40) + (\$55)(.25) = -\$45.50$$

The decision maker who uses the EMV criterion will select the no-expansion decision alternative because it results in the highest long-run average payoff, $2.90. It is possible that the decision maker will only have one chance to make this decision at this company. In such a case, the decision maker will not average $2.90 for selecting no expansion but rather will get a payoff of −$3.00, $2.00, $3.00, or $6.00, depending on which state of demand follows the decision.

This analysis can be shown through the use of a decision tree.



10.3.1 EXPECTED VALUE OF PERFECT INFORMATION

What is the value of knowing which state of nature will occur and when? The answer to such a question can provide insight into how much it is worth to pay for market or business research. The **expected value of perfect information** is *the difference between the payoff that would occur if the decision maker knew which states of nature would occur and the expected monetary payoff from the best decision alternative when there is no information about the occurrence of the states of nature.*

Expected Value of Perfect Information

= Expected Monetary Payoff with Perfect Information
−Expected Monetary Value without Information

As an example, consider the $10,000 investment example with the probabilities of states of nature shown.

		State of the Economy		
		Stagnant (.25)	Slow Growth (.45)	Rapid Growth (.30)
Investment Decision Alternative	Stocks	−$500	$700	$2,200
	Bonds	−$100	$600	$900
	CDs	$300	$500	$750
	Mixture	−$200	$650	$1,300

The following expected monetary values were computed for this problem.

Stocks	$850
Bonds	515
CDs	525
Mixture	632.50

The investment of stocks was selected under the expected monetary value strategy because it resulted in the maximum expected payoff of $850. This decision was made with no information about the states of nature. Suppose we could obtain information about the states of the economy; that is, we *know* which state of the economy will occur. Whenever the state of the economy is stagnant, we would invest in CDs and receive a payoff of $300. Whenever the state of the economy is slow growth, we would invest in stocks and earn $700. Whenever the state of the economy is rapid growth, we would also invest in stocks and earn $2,200. Given the probabilities of each state of the economy occurring, we can use these payoffs to compute an expected monetary payoff of perfect information.

Expected Monetary Payoff of Perfect Information

$$= (\$300)(.25) + (\$700)(.45) + (\$2{,}200)(.30) = \$1{,}050$$

The difference between this expected monetary payoff with perfect information ($1,050) and the expected monetary payoff with no information ($850) is the value of perfect information ($1,050 − $850 = $200). It would not be economically wise to spend more than $200 to obtain perfect information about these states of nature.

DEMONSTRATION PROBLEM 10.3

Compute the value of perfect information for the capacity problem discussed in Demonstration Problems 10.1 and 10.2. The data are shown again here.

		State of Demand			
		Less (.10)	No Change (.25)	Moderate Increase (.40)	Large Increase (.25)
Capacity Decision	No Expansion	−$3	$2	$3	$6
	Add On	−$40	−$28	$10	$20
	Build a New Facility	−$210	−$145	−$5	$55

Solution: The expected monetary value (payoff) under no information computed in Demonstration Problem 10.2 was $2.90 (recall that all figures are in $ millions). If the decision makers had perfect information, they would select no expansion for the state of less demand, no expansion for the state of no change, add on for the state of moderate increase, and build a new facility for the state of large increase. The expected payoff of perfect information is computed as

$$(-\$3)(.10) + (\$2)(.25) + (\$10)(.40) + (\$55)(.25) = \$17.95$$

The expected value of perfect information is

$$\$17.95 - \$2.90 = \$15.05$$

In this case, the decision makers might be willing to pay up to $15.05 ($ million) for perfect information.

10.3.2 UTILITY

As pointed out in the preceding section, expected monetary value decisions are based on long-run averages. Some situations do not lend themselves to expected monetary value analysis because these situations involve relatively large amounts of money and one-time decisions. Examples of these one-time decisions might be drilling an oil well, building a new production facility, merging with another company, ordering 100 new 737s, or buying a professional sports franchise. In analyzing the alternatives in such decisions, a concept known as utility can be helpful.

Utility is the degree of pleasure or displeasure a decision maker has in being involved in the outcome selection process given the risks and opportunities available. Suppose a person has the chance to enter a contest with a 50–50 chance of winning $100,000. If the person wins the contest, he or she wins $100,000. If the person loses, he or she receives $0. There is no cost to enter this contest. The expected payoff of this contest for the entrant is

$$(\$100{,}000)(.50) + (\$0)(.50) = \$50{,}000$$

In thinking about this contest, the contestant realizes that he or she will never get $50,000. The $50,000 is the long-run average payoff if the game is played over and over. Suppose contest administrators offer the contestant $30,000 not to play the game. Would the player take the money and drop out of the contest? Would a certain payoff of $30,000 outdraw a .50 chance at $100,000? The answer to this question depends, in part, on the person's financial situation and on his or her propensity to take risks. If the contestant is a multimillionaire, he or she might be willing to take big risks and even refuse $70,000 to drop out of the contest, because $70,000 does not significantly increase his or her worth. On the other hand, a person on welfare who is offered $20,000 not to play the contest might take the money because $20,000 is worth a great deal to him or her. In addition, two different people on welfare might have different risk-taking profiles. One might be a risk taker who, in spite of a need for money, is not willing to take less than $70,000 or $80,000 to pull out of a contest. The same could be said for the wealthy person.

Utility theory provides a mechanism for determining whether a person is a risk taker, a risk avoider, or an EMVer for a given decision situation. Consider the contest just described. A person receives $0 if he or she does not win the contest and $100,000 if he or she does win the contest. How much money would it take for a contestant to be indifferent between participating in the contest and dropping out? Suppose we examine three possible contestants, X, Y, and Z.

X is indifferent between receiving $20,000 and a .50 chance of winning the contest. For any amount more than $20,000, X will take the money and not play the game. As we stated before, a .50 chance of winning yields an expected payoff of $50,000. Suppose we increase the chance of winning to .80, so that the expected monetary payoff is $80,000. Now X is indifferent between receiving $50,000 and playing the game and will drop out of the game for any amount more than $50,000. In virtually all cases, X is willing to take less money than the expected payoff to quit the game. X is referred to as a **risk avoider.** Many of us are risk avoiders. For this reason, we pay insurance companies to cover our personal lives, our homes, our businesses, our cars, and so on, even when we know the odds are in the insurance companies' favor. We see the potential to lose at such games as unacceptable, so we bail out of the games for less than the expected payoff and pay out more than the expected cost to avoid the game.

Y, on the other hand, loves such contests. It would take about $70,000 to get Y not to play the game with a .50 chance of winning $100,000, even though the expected payoff is only $50,000. Suppose Y were told that there was only a .20 chance of winning the game. How much would it take for Y to become indifferent to playing? It might take $40,000 for Y to be indifferent, even though the expected payoff for a .20 chance is only $20,000. Y is a **risk taker** and enjoys playing risk-taking games. It always seems to take more than the expected payoff to get Y to drop out of the contest.

Z is an EMVer. Z is indifferent between receiving $50,000 and having a .50 chance of winning $100,000. To get Z out of the contest if there is only a .20 chance of winning, the contest directors would have to offer Z about $20,000 (the expected value). Likewise, if there were an .80 chance of winning, it would take about $80,000 to get Z to drop out of the contest. Z makes a decision by going with the long-run averages even in one-time decisions.

Figure 10.4 presents a graph with the likely shapes of the utility curves for X, Y, and Z. This graph is constructed for the game using the payoff range of $0 to $100,000; in-between values can be offered to the players in an effort to buy them out of the game. These units are displayed along what is normally the x axis. Along the y axis are the probabilities of winning the game, ranging from .0 to 1.0. A straight line through the middle of the values represents the EMV responses. If a person plays the game with a. 50 chance of winning, he or she is indifferent to taking $50,000 not to play and playing. For .20, it is $20,000. For .80, it is $80,000.

Notice in the graph that where the chance of winning is .50, contestant X is willing to drop out of the game for $20,000. This point, ($20,000, .50), is above the EMV line. When the chance is .20, X will drop out for $5,000; for a chance of .80, X will drop out for $50,000. Both of these points, ($5,000, .20) and ($50,000, .80), are above the EMV line also.

Y, in contrast, requires $80,000 to be indifferent to a .50 chance of winning. Hence, the point ($80,000, .50) is below the EMV line. Contest officials will have to offer Y at least $40,000 for Y to become indifferent to a .20 chance of winning. This point, ($40,000, .20), also is below the EMV line.

X is a risk avoider and Y is a risk taker. Z is an EMVer. In the utility graph in Figure 10.4, the risk avoider's curve is above the EMV line and the risk taker's curve is below the line.

As discussed earlier in the chapter, in making decisions under uncertainty risk takers might be more prone to use the maximax criterion and risk avoiders might be more prone to use the maximin criterion. The Hurwicz criterion allows the user to introduce his or her propensity toward risk into the analysis by using alpha.

Much information has been compiled and published about utility theory. The objective here is to give you a brief introduction to it through this example, thus enabling you to see that there are risk takers and risk avoiders along with EMVers. A more detailed treatment of this topic is beyond the scope of this text.

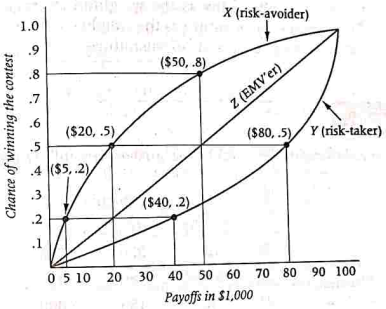


Figure 10.4: Risk Curves for Game Players

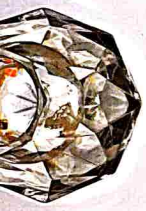
SELF ASSESSMENT QUESTIONS

Fill in the Blanks

13. With decision trees, the decision alternatives are depicted by a _____ node and the states of nature are represented by a _____ node.

14. The decision table presented in question 7 has been reproduced below with probabilities assigned to the states of nature:

		State of Nature			
		1(.20)	2(.35)	3(.40)	4(.05)
Decision Alternative	1	–50	–25	75	125
	2	10	15	20	25
	3	–20	–5	10	20



X is indifferent between receiving $20,000 and a .50 chance of winning the contest. For any amount more than $20,000, X will take the money and not play the game. As we stated before, a .50 chance of winning yields an expected payoff of $50,000. Suppose we increase the chance of winning to .80, so that the expected monetary payoff is $80,000. Now X is indifferent between receiving $50,000 and playing the game and will drop out of the game for any amount more than $50,000. In virtually all cases, X is willing to take less money than the expected payoff to quit the game. X is referred to as a **risk avoider.** Many of us are risk avoiders. For this reason, we pay insurance companies to cover our personal lives, our homes, our businesses, our cars, and so on, even when we know the odds are in the insurance companies' favor. We see the potential to lose at such games as unacceptable, so we bail out of the games for less than the expected payoff and pay out more than the expected cost to avoid the game.

Y, on the other hand, loves such contests. It would take about $70,000 to get Y not to play the game with a .50 chance of winning $100,000, even though the expected payoff is only $50,000. Suppose Y were told that there was only a .20 chance of winning the game. How much would it take for Y to become indifferent to playing? It might take $40,000 for Y to be indifferent, even though the expected payoff for a .20 chance is only $20,000. Y is a **risk taker** and enjoys playing risk-taking games. It always seems to take more than the expected payoff to get Y to drop out of the contest.

Z is an EMVer. Z is indifferent between receiving $50,000 and having a .50 chance of winning $100,000. To get Z out of the contest if there is only a .20 chance of winning, the contest directors would have to offer Z about $20,000 (the expected value). Likewise, if there were an .80 chance of winning, it would take about $80,000 to get Z to drop out of the contest. Z makes a decision by going with the long-run averages even in one-time decisions.

Figure 10.4 presents a graph with the likely shapes of the utility curves for X, Y, and Z. This graph is constructed for the game using the payoff range of $0 to $100,000; in-between values can be offered to the players in an effort to buy them out of the game. These units are displayed along what is normally the x axis. Along the y axis are the probabilities of winning the game, ranging from .0 to 1.0. A straight line through the middle of the values represents the EMV responses. If a person plays the game with a. 50 chance of winning, he or she is indifferent to taking $50,000 not to play and playing. For .20, it is $20,000. For .80, it is $80,000.

Notice in the graph that where the chance of winning is .50, contestant X is willing to drop out of the game for $20,000. This point, ($20,000, .50), is above the EMV line. When the chance is .20, X will drop out for $5,000; for a chance of .80, X will drop out for $50,000. Both of these points, ($5,000, .20) and ($50,000, .80), are above the EMV line also.

Y, in contrast, requires $80,000 to be indifferent to a .50 chance of winning. Hence, the point ($80,000, .50) is below the EMV line. Contest officials will have to offer Y at least $40,000 for Y to become indifferent to a .20 chance of winning. This point, ($40,000, .20), also is below the EMV line.

X is a risk avoider and Y is a risk taker. Z is an EMVer. In the utility graph in Figure 10.4, the risk avoider's curve is above the EMV line and the risk taker's curve is below the line.

As discussed earlier in the chapter, in making decisions under uncertainty risk takers might be more prone to use the maximax criterion and risk avoiders might be more prone to use the maximin criterion. The Hurwicz criterion allows the user to introduce his or her propensity toward risk into the analysis by using alpha.

Much information has been compiled and published about utility theory. The objective here is to give you a brief introduction to it through this example, thus enabling you to see that there are risk takers and risk avoiders along with EMVers. A more detailed treatment of this topic is beyond the scope of this text.

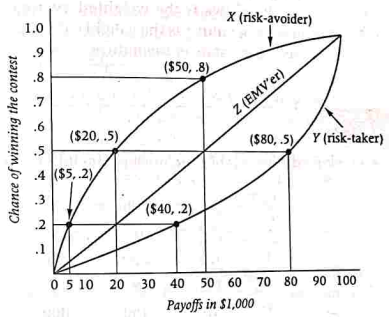


Figure 10.4: Risk Curves for Game Players

SELF ASSESSMENT QUESTIONS

Fill in the Blanks

13. With decision trees, the decision alternatives are depicted by a _____ node and the states of nature are represented by a _____ node.

14. The decision table presented in question 7 has been reproduced below with probabilities assigned to the states of nature:

		State of Nature			
		1(.20)	2(.35)	3(.40)	4(.05)
Decision Alternative	1	–50	–25	75	125
	2	10	15	20	25
	3	–20	–5	10	20

SELF ASSESSMENT QUESTIONS

The expected monetary value of selecting decision alternative 1 is _____. The expected monetary value of selecting decision alternative 2 is _____. The expected monetary value of selecting decision alternative 3 is _____.

State whether the following statements are true/false:

15. In a decision-making scenario, if it is not known which of the states of nature will occur but the probabilities of occurrence of the states are known the scenario is called decision-making under risk.
16. In a decision-making under risk scenario, the expected monetary value of a decision alternative is the arithmetic average of the payoffs to the decision alternative in each state of the nature.
17. In a decision-making under risk scenario, the expected monetary value of a decision alternative is the weighted average (using the probability of each state of nature as the weight) of the payoffs to the decision alternative in each state of the nature.

ACTIVITY

Consider the following decision table and answer the following

		States of Nature		
		S_1(0.5)	S_2(0.3)	S_3(0.2)
Decision Alternatives	D_1	150	250	350
	D_2	120	180	400
	D_3	75	150	600

Draw a decision tree for the payoff table.

10.4 SUMMARY

- Decision analysis is a branch of quantitative management in which mathematical and statistical approaches are used to assist decision makers in reaching judgments about alternative opportunities. Three types of decisions are (1) decisions made under certainty, (2) decisions made under uncertainty, and (3) decisions made with risk. Several aspects of the decision-making situation are decision alternatives, states of nature, and payoffs. Decision alternatives are the options open to decision makers from which they can choose. States of nature are situations or conditions that arise after the decision has been made over which the decision maker has no control. The payoffs are the gains or losses that the decision maker will reap from various decision alternatives. These three aspects (decision alternatives, states of nature, and payoffs) can be displayed in a decision table or payoff table.
- Decision making under certainty is the easiest of the three types of decisions to make. In this case, the states of nature are known, and the decision maker merely selects the decision alternative that yields the highest payoff.
- Decisions are made under uncertainty when the likelihoods of the states of nature occurring are unknown. Four approaches to making decisions under uncertainty are maximax criterion, maximin criterion, Hurwicz criterion, and minimax regret. The maximax criterion is an optimistic approach based on the notion that the best possible outcomes will occur. In this approach, the decision maker selects the maximum possible payoff under each decision alternative and then selects the maximum of these. Thus, the decision maker is selecting the maximum of the maximums.
- The maximin criterion is a pessimistic approach. The assumption is that the worst case will happen under each decision alternative. The decision maker selects the minimum payoffs under each decision alternative and then picks the maximum of these as the best solution. Thus, the decision maker is selecting the best of the worst cases, or the maximum of the minimums.
- The Hurwicz criterion is an attempt to give the decision maker an alternative to maximax and maximin that is somewhere between an optimistic and a pessimistic approach. With this approach, decision makers select a value called alpha between 0 and 1 to represent how optimistic they are. The maximum and minimum payoffs for each decision alternative are examined. The alpha weight is applied to the maximum payoff under each decision alternative and $1-\alpha$ is applied to the minimum payoff. These two weighted values are combined for each decision alternative, and the maximum of these weighted values is selected.
- Minimax regret is calculated by examining opportunity loss. An opportunity loss table is constructed by subtracting each payoff from the maximum payoff under each state of nature. This step produces a lost opportunity under each state. The maximum lost opportunity from each decision alternative is determined from the opportunity table. The minimum of these values is selected, and the corresponding decision alternative is chosen. In this way, the decision maker has reduced or minimized the regret, or lost opportunity.
- In decision making with risk, the decision maker has some prior knowledge of the probability of each occurrence of each state of nature. With these probabilities, a weighted payoff referred to as expected monetary value (EMV) can be calculated for each decision alternative. A person who makes decisions based on these EMVs is called an EMVer. The expected monetary value is essentially the average payoff that would occur if the decision process were to be played out over a long period of time with the probabilities holding constant.
- The expected value of perfect information can be determined by comparing the expected monetary value if the states of nature are known to the expected monetary value. The difference in the two is the expected value of perfect information.
- Utility refers to a decision maker's propensity to take risks. People who avoid risks are called risk avoiders. People who are prone to take risks are referred to as risk takers. People who use EMV generally fall between these two categories. Utility curves can be sketched to ascertain or depict a decision maker's tendency toward risk.

KEY WORDS

1. **Decision alternatives:** The various choices or options available to the decision maker in any given problem situation.
2. **Decision analysis:** A category of quantitative business techniques particularly targeted at clarifying and enhancing the decision-making process.
3. **Decision making under certainty:** A decision-making situation in which the states of nature are known.
4. **Decision making under risk:** A decision-making situation in which it is uncertain which states of nature will occur but the probability of each state of nature occurring has been determined.
5. **Decision making under uncertainty:** A decision-making situation in which the states of nature that may occur are unknown and the probability of a state of nature occurring is also unknown.
6. **Decision table:** A matrix that displays the decision alternatives, the states of nature, and the payoffs for a particular decision-making problem; also called a payoff table.
7. **Decision trees:** A flowchart-like depiction of the decision process that includes the various decision alternatives, the various states of nature, and the payoffs.
8. **EMVer:** A person who uses an expected monetary value (EMV) approach to making decisions under risk.
9. **Expected monetary value (EMV):** A value of a decision alternative computed by multiplying the probability of each state of nature by the state's associated payoff and summing these products across the states of nature.
10. **Expected value of perfect information:** The difference between the expected monetary payoff that would occur if the decision maker knew which states of nature would occur and the payoff from the best decision alternative when there is no information about the occurrence of the states of nature.
11. **Expected value of sample information:** The difference between the expected monetary value with information and the expected monetary value without information.
12. **Hurwicz criterion:** An approach to decision making in which the maximum and minimum payoffs selected from each decision alternative are used with a weight, between 0 and 1 to determine the alternative with the maximum weighted average. The higher the value of, the more optimistic is the decision maker.
13. **Maximax criterion:** An optimistic approach to decision making under uncertainty in which the decision alternative is chosen according to which alternative produces the maximum overall payoff of the maximum payoffs from each alternative.
14. **Maximin criterion:** A pessimistic approach to decision making under uncertainty in which the decision alternative is chosen according to which alternative produces the maximum overall payoff among the minimum payoffs from each alternative.

KEY WORDS

15. **Minimax regret:** A decision-making strategy in which the decision maker determines the lost opportunity for each decision alternative and selects the decision alternative with the minimum of lost opportunity or regret.
16. **Opportunity loss table:** A decision table constructed by subtracting all payoffs for a given state of nature from the maximum payoff for that state of nature and doing this for all states of nature; displays the lost opportunities or regret that would occur for a given decision alternative if that particular state of nature occurred.
17. **Payoffs:** The benefits or rewards that result from selecting a particular decision alternative.
18. **Payoff table:** A matrix that displays the decision alternatives, the states of nature, and the payoffs for a particular decision-making problem; also called a decision table.
19. **Risk avoider:** A decision maker who avoids risk whenever possible and is willing to drop out of a game when given the chance even when the payoff is less than the expected monetary value.
20. **Risk taker:** A decision maker who enjoys taking risk and will not drop out of a game unless the payoff is more than the expected monetary value.
21. **States of nature:** The occurrences of nature that can happen after a decision has been made that can affect the outcome of the decision and over which the decision maker has little or no control.

10.5 DESCRIPTIVE QUESTIONS

10.1. Use the decision table given here to complete parts (a) through (d).

		State of Nature		
		s_1	s_2	s_3
Decision Alternative	d_1	250	175	−25
	d_2	110	100	70
	d_3	390	140	−80

(a) Use the maximax criterion to determine which decision alternative to select.

(b) Use the maximin criterion to determine which decision alternative to select.

(c) Use the Hurwicz criterion to determine which decision alternative to select. Let $\alpha = .3$ and then let $\alpha = .8$ and compare the results.

(d) Compute an opportunity loss table from the data Use this table and a minimax regret criterion to determine which decision alternative to select.

10.2. Election results can affect the payoff from certain types of investments. Suppose a brokerage firm is faced with the prospect of investing $20 million a few weeks before the national election for president of the

United States. They feel that if a Republican is elected, certain types of investments will do quite well; but if a Democrat is elected, other types of investments will be more desirable. To complicate the situation, an independent candidate, if elected, is likely to cause investments to behave in a different manner. Following are the payoffs for different investments under different political scenarios. Use the data to reach a conclusion about which decision alternative to select. Use both the maximax and maximin criteria and compare the answers.

		Election Winner		
		Republican	Democrat	Independent
Investment	A	60	15	−25
	B	10	25	30
	C	−10	40	15
	D	20	25	5

10.3. The introduction of a new product into the marketplace is quite risky. The percentage of new product ideas that successfully make it into the marketplace is as low as 1%. Research and development costs must be recouped, along with marketing and production costs. However, if a new product is warmly received by customers, the payoffs can be great. Following is a payoff table (decision table) for the production of a new product under different states of the market. Notice that the decision alternatives are to not produce the product at all, produce a few units of the product, and produce many units of the product. The market may be not receptive to the product, somewhat receptive to the product, and very receptive to the product.

(a) Use this matrix and the Hurwicz criterion to reach a decision. Let $\alpha = .6$.

(b) Determine an opportunity loss table from this payoff table and use minimax regret to reach a decision.

		State of the Market		
		Not Receptive	Somewhat Receptive	Very Receptive
Production Alternative	Don't Produce	−50	−50	−50
	Produce Few	−200	300	400
	Produce Many	−600	100	1000

10.4. A home buyer is completing application for a home mortgage. The buyer is given the option of "locking in" a mortgage loan interest rate or waiting 60 days until closing and locking in a rate on the day of closing. The buyer is not given the option of locking in at any time in between. If the buyer locks in at the time of application and interest rates go down, the loan will cost the buyer $150 per month more (–$150 payoff) than it would have if he or she had waited and locked in later. If the buyer locks in at the time of application and interest rates go up, the buyer has saved money by locking in at a lower rate. The amount saved under this condition is a payoff of +$200. If the buyer does not lock in at application and rates go up, he or she must pay more interest on the mortgage loan; the payoff is –$250. If the buyer does not lock in at application and rates go down, he or she has reduced the interest amount and the payoff is +$175. If the rate does not change at all, there is a $0 payoff for locking in at the time of application and also a $0 payoff for not locking in at that time. There is a probability of .65 that the interest rates will rise by the end of the 60-day period, a .30 probability that they will fall, and $\alpha = .05$ probability that they will remain constant. Construct a decision table from this information.

Compute the expected monetary values from the table and reach a conclusion about the decision alternatives. Compute the value of perfect information.

10.5. A person has a chance to invest $50,000 in a business venture. If the venture works, the investor will reap $200,000. If the venture fails, the investor will lose his money. It appears that there is about $\alpha = .50$ probability of the venture working. Using this information, answer the following questions.

(a) What is the expected monetary value of this investment?

(b) If this person decides not to undertake this venture, is he an EMVer, a risk avoider, or a risk taker? Why?

(c) You would have to offer at least how much money to get a risk taker to quit pursuing this investment?

10.6. Shown here is a decision table from a business situation. The decision maker has an opportunity to purchase sample information in the form of a forecast. With the sample information, the prior probabilities can be revised Also shown are the probabilities of forecasts from the sample information for each state of nature. Use this information to answer parts (a) through (d).

		State of Nature	
		s_1(.30)	s_2(.70)
Alternative	d_1	$350	−$100
	d_2	−$200	$325

		State of Nature	
		s_1	s_2
Forecast	s_1	.90	.25
	s_2	.10	.75

(a) Compute the expected monetary value of this decision without sample information.

(b) Compute the expected monetary value of this decision with sample information.

(c) Use a tree diagram to show the decision options in parts (a) and (b).

(d) Calculate the value of the sample information.

10.6 SOLUTIONS FOR DESCRIPTIVE QUESTIONS

10.1.

	S_1	S_2	S_3	Max	Min
d_1	250	175	–25	250	–25
d_2	110	100	70	110	70
d_3	390	140	–80	390	–80

(a) Max {250, 110, 390} = **390**

Decision: Select d_3

(b) Max {–25, 70, –80} = **70**

Decision: Select d_2

(c) For $\alpha = .3$

d_1: .3(250) + .7(–25) = **57.5**

d_2: .3(110) + .7(70) = **82**

d_3: .3(390) + .7(–80) = **61**

Decision: Select d_2

For $\alpha = .8$

d_1: .8(250) + .2(–25) = **195**

d_2: .8(110) + .2(70) = **102**

d_3: .8(390) + .2(–80) = **296**

Decision: Select d_3

Comparing the results for the two different values of alpha, with a more pessimist point-of-view ($\alpha = .3$), the decision is to select d_2 and the payoff is 82. Selecting by using a more optimistic point-of-view ($\alpha = .8$) results in choosing d_3 with a higher payoff of 296.

(d) The opportunity loss table is:

	S_1	S_2	S_3	Max
d_1	140	0	95	140
d_2	280	75	0	280
d_3	0	35	150	150

The minimax regret = min {140, 280, 150} = **140**

Decision: Select d_1 to minimize the regret.

10.2.

	R	D	I	Max	Min
A	60	15	–25	60	–25
B	10	25	30	30	10
C	–10	40	15	40	–10
D	20	25	5	25	5

Maximax = Max {60, 30, 40, 25} = **60**

Decision: Select A

Maximin = Max {–25, 10, –10, 5} = **10**

Decision: Select B

10.3.

	Not	Somewhat	Very	Max	Min
None	–50	–50	–50	–50	–50
Few	–200	300	400	400	–200
Many	–600	100	1000	1000	–600

(a) For Hurwicz criterion using $\alpha = .6$:

Max {[.6(–50) + .4(–50)], [.6(400) + .4(–200)],

[.6(1000) + .4(–600)]} = {–50, –160, 360} = **360**

Decision: Select "Many"

(b) Opportunity Loss Table:

	Not	Somewhat	Very	Max
None	0	0 350	1050	1050
Few	150	0	600	600
Many	550	200	0	550

Minimax regret = Min {1050, 600, 550} = **550**

Decision: Select "Many"

10.4.

	Down(.30)	Up(.65)	No Change(.05)	EMV
Lock-In	–150	200	0	85
No	175	–250	0	–110

Decision: Based on the highest EMV(85), "Lock-In"

Expected Payoff with Perfect Information =

175(.30) + 200(.65) + 0(.05) = **182.5**

Expected Value of Perfect Information = 182.5 – 85 = **97.5**

10.5.

(a) EMV = 200,000(.5) + (–50,000)(.5) = **75,000**

(b) Risk Avoider because the EMV is more than the investment (75,000 > 50,000)

(c) You would have to offer more than 75,000 which is the expected value.

10.6.

(a)

	S_1(.30)	S_2(.70)	EMV
d_1	350	–100	35
d_2	–200	325	167.5

Decision: Based on EMV,

maximum {35, 167.5} = **167.5**

(b) & (c) For Forecast S_1:

	Prior	Cond.	Joint	Revised
S_1	.30	.90	.27	.6067
S_2	.70	.25	.175	.3933
			$F(S_1)$ = .445	

For Forecast S_2:

	Prior	Cond.	Joint	Revised
S_1	.30	.10	.030	.054
S_2	.70	.75	.525	.946
			$F(S_2)$ = .555	

EMV with Sample Information = **241.63**

(d) Value of Sample Information = 241.63 – 167.5 = **74.13**

10.7 ANSWERS AND HINTS

Topics	Q. No.	Answers
10.1 The Decision Table and Decision Making Under Certainty	1.	Certainty, Uncertainty, Risk
	2.	Decision Alternatives, States of Nature, Payoffs
	3.	States of Nature, Payoffs, Decision Alternatives
	4.	False
	5.	False
	6.	False
10.2 Decision Making Under Uncertainty	7.	1, 125
	8.	2, 10, 2, 14.5, 1, 90
	9.	1,60
	10.	False
	11.	False
10.3 Decision Making Under Risk	12.	True
	13.	`, !
	14.	17.5, 16.5, -0.75
	15.	True
	16.	False
	17.	True

NMIMS Global Access - School for Continuing Education