STATISTICS for BUSINESS and ECONOMICS

14TH EDITION

James T. McClave • P. George Benson • Terry Sincich

APPLET CORRELATION

Applet	Concept Illustrated	Description	Applet Activity
Random numbers	Uses a random number generator to determine the experimental units to be included in a sample.	Generates random numbers from a range of integers specified by the user.	**1.1**, 44; **1.2**, 44; **3.6**, 198; **4.1**, 219; **4.2**, 219; **4.8**, 271
Sample from a population	Assesses how well a sample represents the population and the role that sample size plays in the process.	Produces random sample from population from specified sample size and population distribution shape. Reports mean, median, and standard deviation; applet creates plot of sample.	**4.4**, 233; **4.6**, 257; **4.7**, 271
Sampling distributions	Compares means and standard deviations of distributions; assesses effect of sample size; illustrates undbiasedness.	Simulates repeatedly choosing samples of a fixed size *n* from a population with specified sample size, number of samples, and shape of population distribution. Applet reports means, medians, and standard deviations; creates plots for both.	**5.1**, 310; **5.2**, 310
Long-run probability demonstrations illustrate the concept that theoretical probabilities are long-run experimental probabilities.			
Simulating probability of rolling a 6	Investigates relationship between theoretical and experimental probabilities of rolling 6 as number of die rolls increases.	Reports and creates frequency histogram for each outcome of each simulated roll of a fair die. Students specify number of rolls; applet calculates and plots proportion of 6s.	**3.1**, 162; **3.3**, 174; **3.4**, 175; **3.5**, 188
Simulating probability of rolling a 3 or 4	Investigates relationship between theoretical and experimental probabilities of rolling 3 or 4 as number of die rolls increases.	Reports outcome of each simulated roll of a fair die; creates frequency histogram for outcomes. Students specify number of rolls; applet calculates and plots proportion of 3s and 4s.	**3.3**, 174; **3.4**, 175
Simulating the probability of heads: fair coin	Investigates relationship between theoretical and experimental probabilities of getting heads as number of fair coin flips increases.	Reports outcome of each fair coin flip and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots proportion of heads.	**3.2**, 162; **4.2**, 219
Simulating probability of heads: unfair coin $(P(H) = .2)$	Investigates relationship between theoretical and experimental probabilities of getting heads as number of unfair coin flips increases.	Reports outcome of each flip for a coin where heads is less likely to occur than tails and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots the proportion of heads.	**4.3**, 233
Simulating probability of heads: unfair coin $(P(H) = .8)$	Investigates relationship between theoretical and experimental probabilities of getting heads as number of unfair coin flips increases.	Reports outcome of each flip for a coin where heads is more likely to occur than tails and creates a bar graph for outcomes. Students specify number of flips; applet calculates and plots the proportion of heads.	**4.3**, 233
Simulating the stock market	Theoretical probabilities are long run experimental probabilities.	Simulates stock market fluctuation. Students specify number of days; applet reports whether stock market goes up or down daily and creates a bar graph for outcomes. Calculates and plots proportion of simulated days stock market goes up.	**4.5**, 234
Mean versus median	Investigates how skewedness and outliers affect measures of central tendency.	Students visualize relationship between mean and median by adding and deleting data points; applet automatically updates mean and median.	**2.1**, 88; **2.2**, 88; **2.3**, 88

(*Continued*)

the population of 3,005 items would yield a mean GPF of at least 50.8%? Or, is it likely that two independent, random samples of sizes 134 and 119 would yield mean GPFs of at least 50.6% and 51.0%, respectively? (These were the questions posed to a statistician retained by the CPA firm.) Use the ideas of probability and sampling distributions to guide your analysis.

Prepare a professional document that presents the results of your analysis and gives your opinion regarding fraud. Be sure to describe the assumptions and methodology used to arrive at your findings.

Variable	Type	Description
MONTH	QL	Month in which item was sold in 1991
INVOICE	QN	Invoice number
SALES	QN	Sales price of item in dollars
PROFIT	QN	Profit amount of item in dollars
MARGIN	QN	Profit margin of item = (Profit/Sales) × 100%

Data Set: FIRE

6

CONTENTS

WHERE WE'VE BEEN

- Learned that populations are characterized by numerical descriptive measures called *parameters*
- Found that decisions about population parameters are based on *statistics* computed from the sample
- Discovered that *inferences* about parameters are subject to uncertainty and that this uncertainty is reflected in the *sampling distribution* of a statistic

WHERE WE'RE GOING

- Estimate a population parameter (means, proportion, or variance) based on a large sample selected from the population (**6.1**)
- Use the sampling distribution of a statistic to form a confidence interval for the population parameter (**6.2–6.4**, **6.6**, **6.7**)
- Show how to select the proper sample size for estimating a population parameter (**6.5**)

Inferences Based on a Single Sample

Estimation with Confidence Intervals

STATISTICS IN ACTION — Medicare Fraud Investigations

According to the Office of Inspector General website (https://oig.hhs.gov/fraud/strike-force/), the Medicare Fraud Strike Force—*established by the US Department of Justice (USDOJ) in March 2007—"uses data analytics and the combined resources of Federal, State, and local law enforcement entities to prevent and combat Medicare health care fraud, waste, and abuse. Strike Force teams have shut down health care fraud schemes around the country, arrested more than a thousand criminals, and recovered millions of taxpayer dollars."*

Below is just a sample of the press releases from the USDOJ during 2019 alone:

- Former Los Angeles-Area Physician Sentenced to Two Years in Federal Prison for Defrauding Medicare and Illegally Prescribing Opioid Drugs
- Detroit Home Health Owner Sentenced to Prison for Role in $1.5 Million Medicare Kickback Scheme
- Southern California Doctor Found Guilty in $12 Million Medicare Fraud and Device Adulteration Scheme
- Texas Physician Convicted in $16 Million Medicare Fraud Scheme
- Charges Brought Against 34 Individuals for Alleged West Coast Medicare and Medicaid Fraud Schemes Totaling $258 Million
- South Florida Health Care Facility Owner Sentenced To 20 Years in Prison for Role in Largest Health Care Fraud Scheme Ever Charged By the Department Of Justice

STATISTICS IN ACTION
(*continued*)

- New Jersey Doctor Pleads Guilty to $13 Million Conspiracy to Defraud Medicare with Telemedicine Orders of Orthotic Braces
- Owner and Chief Executive Officer of Telemedicine Company Pleads Guilty to $424 Million Conspiracy to Defraud Medicare and Receive Illegal Kickbacks in Exchange for Orders of Durable Medical Equipment
- Patient Recruiter Found Guilty in $1.3 Million Medicare Kickback Scheme
- Owner of Tampa-Area Medical Marketing Company Found Guilty in $2 Million Medicare Fraud Scheme
- South Florida Pill Mill Owner Sentenced to Prison for Role in $2.2 Million Medicare Fraud Scheme
- Owners of Los Angeles Home Health Agency Sentenced to Prison for Role in Health Care Fraud that Defrauded Medicare
- Michigan Doctor Pleads Guilty to Role in $2.5 Million Medicare Fraud Scheme
- Former CEO of Tennessee Pain Management Company Convicted for Role in Approximate $4 Million Medicare Kickback Scheme
- Former Administrator of Two Houston Home Health Companies Sentenced to Prison in $20 Million Medicare Fraud Scheme
- Owner of Washington, D.C.-Based Durable Medical Equipment Company Sentenced to Prison for Role in $9.8 Million Medicaid Fraud Scheme

One way in which Medicare fraud occurs is through the use of "upcoding," which refers to the practice of providers coding Medicare claims at a higher level of care than was actually provided to the patient. For example, suppose a particular kind of claim can be coded at three levels, where Level 1 is a routine office visit, Level 2 is a thorough examination involving advanced diagnostic tests, and Level 3 involves performing minor surgery. The amount of Medicare payment is higher for each increased level of claim. Thus, upcoding would occur if Level 1 services were billed at Level 2 or Level 3 payments, or if Level 2 services were billed at Level 3 payment.

The USDOJ relies on sound statistical methods to help identify Medicare fraud. Once the USDOJ has determined that possible upcoding has occurred, it next seeks to further investigate whether it is the result of legitimate practice (perhaps the provider is a specialist giving higher levels of care) or the result of fraudulent action on the part of the provider. To further its investigation, the USDOJ will next ask a statistician to select a sample of the provider's claims. For example, the statistician might determine that a random sample of 52 claims from the 1,000 claims in question will provide a sufficient sample to estimate the overcharge reliably. The USDOJ then asks a health care expert to audit each of the medical files corresponding to the sampled claims and determine whether the level of care matches the level billed by the provider, and, if not, to determine what level should have been billed. Once the audit has been completed, the USDOJ will calculate the overcharge.

In this chapter, we present a recent Medicare fraud case investigated by the USDOJ. Results for the audit of 52 sampled claims, with the amount paid for each claim, the amount disallowed by the auditor, and the amount that should have been paid for each claim, are saved in the **MFRAUD** file.* Knowing that a total of $103,500 was paid for the 1,000 claims, the USDOJ wants to use the sample results to extrapolate the overpayment amount to the entire population of 1,000 claims.

STATISTICS IN ACTION REVISITED

- Estimating the Mean Overpayment **(p. 345)**
- Estimating the Coding Error Rate **(p. 353**
- Determining Sample Size **(p. 360)**

Data Set: MFRAUD

*Data provided (with permission) from Info Tech, Inc., Gainesville, Florida.

6.1 Identifying and Estimating the Target Parameter

In this chapter, our goal is to estimate the value of an unknown population parameter, such as a population mean or a proportion from a binomial population. For example, we might want to know the mean gas mileage for a new car model, the average expected life of a flat-screen computer monitor, or the proportion of dot-com companies that fail within a year of start-up.

You'll see that different techniques are used for estimating a mean or proportion, depending on whether a sample contains a large or small number of measurements. Nevertheless, our objectives remain the same. We want to use the sample information to estimate the population parameter of interest (called the **target parameter**) and assess the reliability of the estimate.

> The unknown population parameter (e.g., mean or proportion) that we are interested in estimating is called the **target parameter.** In general, the target parameter is denoted by the symbol, θ.

Often, there are one or more key words in the statement of the problem that indicate the appropriate target parameter. Some key words associated with the two parameters covered in this section are listed in the following table.

Determining the Target Parameter

Parameter (θ)	Key Words or Phrases	Type of Data
μ	Mean; average	Quantitative
p	Proportion; percentage; fraction; rate	Qualitative
σ^2 (optional)	Variance; variability; spread	Quantitative

For the examples given above, the words *mean* in *mean gas mileage* and *average* in *average life expectancy* imply that the target parameter is the population mean, μ. The word *proportion* in *proportion of dot-com companies that fail within one year of start-up* indicates that the target parameter is the binomial proportion, p.

In addition to key words and phrases, the type of data (quantitative or qualitative) collected is indicative of the target parameter. With quantitative data, you are likely to be estimating the mean or variance of the data. With qualitative data with two outcomes (success or failure), the binomial proportion of successes is likely to be the parameter of interest.

A single number calculated from the sample that estimates a target population parameter is called a **point estimator.** For example, we'll use the sample mean, $\bar{x}$, to estimate the population mean μ. Consequently, $\bar{x}$ is a point estimator. Similarly, we'll learn that the sample proportion of successes, denoted $\hat{p}$, is a point estimator for the binomial proportion p and that the sample variance s^2 is a point estimator for the population variance σ^2. Also, we will attach a measure of reliability to our estimate by obtaining an **interval estimator**—a range of numbers that contain the target parameter with a high degree of confidence. For this reason the interval estimate is also called a **confidence interval.**

> A **point estimator** of a population parameter is a rule or formula that tells us how to use the sample data to calculate a *single* number that can be used as an *estimate* of the target parameter.

> An **interval estimator** (or **confidence interval**) is a formula that tells us how to use the sample data to calculate an *interval* that *estimates* the target parameter.

We consider confidence intervals for estimating a population mean in Sections 6.2 and Section 6.3. Confidence intervals for a population proportion are presented in

Section 6.4. In Section 6.5, we show how to determine the sample sizes necessary for reliable estimates of the target parameters based on simple random sampling. Optional Section 6.6 presents a method to apply when the sample size is large relative to the population size. Finally, in optional Section 6.7 we discuss estimation of a population variance.

6.2 Confidence Interval for a Population Mean: Normal (z) Statistic

Suppose a large bank wants to estimate the average amount of money owed by its delinquent debtors (i.e., debtors who are more than 2 months behind in payment). To accomplish this objective, the bank plans to randomly sample 100 of its delinquent accounts and use the sample mean, $\bar{x}$, of the amounts overdue to estimate μ, the mean for *all* delinquent accounts. Because the sample mean $\bar{x}$ represents a single number estimator, it is the *point estimator* of the target parameter μ. How can we assess the accuracy of this point estimator?

According to the Central Limit Theorem, the sampling distribution of the sample mean is approximately normal for large samples, as shown in Figure 6.1. Let us calculate the interval estimator:

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

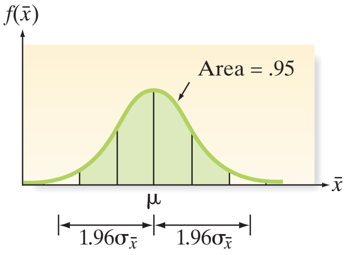


Figure 6.1
Sampling distribution of $\bar{x}$

That is, we form an interval from 1.96 standard deviations below the sample mean to 1.96 standard deviations above the mean. *Prior to drawing the sample,* what are the chances that this interval will enclose μ, the population mean?

To answer this question, refer to Figure 6.1. If the 100 measurements yield a value of $\bar{x}$ that falls between the two lines on either side of μ (i.e., within 1.96 standard deviations of μ), then the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will contain μ; if $\bar{x}$ falls outside these boundaries, the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will not contain μ. From Section 5.3, we know that the area under the normal curve (the sampling distribution of $\bar{x}$) between these boundaries is exactly .95. Thus, the probability that a randomly selected interval, $\bar{x} \pm 1.96\sigma_{\bar{x}}$, will contain μ is equal to .95.

EXAMPLE 6.1

Estimating the Mean, σ Known—Delinquent Debtors

Problem Consider the large bank that wants to estimate the average amount of money owed by its delinquent debtors, μ. The bank randomly samples $n = 100$ of its delinquent accounts and finds that the sample mean amount owed is $\bar{x} = \$230$. Also, suppose it is known that the standard deviation of the amount owed for all delinquent accounts is $\sigma = \$90$. Use the interval estimator $\bar{x} \pm 1.96\sigma_{\bar{x}}$ to calculate a confidence interval for the target parameter, μ.

Solution Substituting $\bar{x} = 230$ and $\sigma = 90$ into the interval estimator formula, we obtain:

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm (1.96)\sigma/\sqrt{n} = 230 \pm (1.96)(90/\sqrt{100}) = 230 \pm 17.64$$

Or, (212.36, 247.64). We can also obtain this confidence interval using statistical software, as shown (highlighted) on the Minitab printout, Figure 6.2.

One-Sample Z

Descriptive Statistics

N	Mean	SE Mean	95% CI for μ
100	230.00	9.00	(212.36, 247.64)

μ: mean of Sample
Known standard deviation = 90

Figure 6.2
Minitab output showing 95% confidence interval for μ, σ known

Look Back Because we know the probability that the interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ will contain μ is .95, we call the interval estimator a 95% *confidence interval* for μ.

Now Work Exercise 6.3a

The interval $\bar{x} \pm 1.96\sigma_{\bar{x}}$ in Example 6.1 is called a *large-sample* 95% confidence interval for the population mean μ. The term *large-sample* refers to the sample being of sufficiently large size that we can apply the Central Limit Theorem and the normal (z) statistic to determine the form of the sampling distribution of $\bar{x}$. Empirical research suggests that a sample size n exceeding a value between 20 and 30 will usually yield a sampling distribution of $\bar{x}$ that is approximately normal. This result led many practitioners to adopt the rule of thumb that a sample size of $n \geq 30$ is required to use large-sample confidence interval procedures. Keep in mind, though, that 30 is not a magical number and, in fact, is quite arbitrary.

Also, *note that the large-sample interval estimator requires knowing the value of the population standard deviation,* σ. In most (if not nearly all) practical business applications, however, the value of σ will be unknown. For large samples, the fact that σ is unknown poses only a minor problem because the sample standard deviation s provides a very good approximation to σ*. The next example illustrates the more realistic large-sample confidence interval procedure.

EXAMPLE 6.2

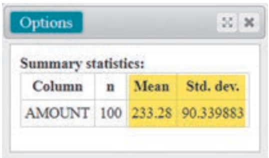
Estimating the Mean, σ Unknown–Delinquent Debtors

Problem Refer to Example 6.1 and the problem of estimating μ, the average amount of money owed by a bank's delinquent debtors. The overdue amounts for the $n = 100$ delinquent accounts are shown in Table 6.1. Use the data to find a 95% confidence interval for μ and interpret the result.

Table 6.1 Overdue Amounts (in Dollars) for 100 Delinquent Accounts

195	243	132	133	209	400	142	312	221	289
221	162	134	275	355	293	242	458	378	148
278	222	236	178	202	222	334	208	194	135
363	221	449	265	146	215	113	229	221	243
512	193	134	138	209	207	206	310	293	310
237	135	252	365	371	238	232	271	121	134
203	178	180	148	162	160	86	234	244	266
119	259	108	289	328	331	330	227	162	354
304	141	158	240	82	17	357	187	364	268
368	274	278	190	344	157	219	77	171	280

Data Set: OVRDUE



Column	n	Mean	Std. dev.
AMOUNT	100	233.28	90.339883

Figure 6.3a

StatCrunch summary statistics for overdue amounts

Solution The large bank almost surely does not know the true standard deviation, σ, of the population of overdue amounts. However, because the sample size is large, we will use the sample standard deviation, s, as an estimate for σ in the confidence interval formula. A StatCrunch printout of summary statistics for the sample of 100 overdue amounts is shown in Figure 6.3a. From the shaded portion of the printout, we find $\bar{x} = 233.28$ and $s = 90.34$. Substituting these values into the interval estimator formula, we obtain:

$$\bar{x} \pm (1.96)\sigma/\sqrt{n} \approx \bar{x} \pm (1.96)s/\sqrt{n} = 233.28 \pm (1.96)(90.34)/\sqrt{100} = 233.28 \pm 17.71$$

or (215.57, 250.99). That is, we estimate the mean amount of delinquency for all accounts to fall within the interval \$215.57 to \$250.99. This confidence interval is highlighted on the StatCrunch output shown in Figure 6.3b.

Look Back To obtain the StatCrunch output shown in Figure 6.3b, we input the value of $\sigma \approx s = 90.34$ on the appropriate menu screen. This approximation is reasonable since the sample size ($n = 100$) is large. However, if we do not assume that $\sigma \approx s$, then the resulting confidence interval will differ from the one shown in Figure 6.3b. The 95% confidence interval for μ for this case is shown (highlighted) at the bottom of the XLSTAT printout, Figure 6.3c. Note that the endpoints of the interval (215.35, 251.21) vary slightly from those

*In the terminology of Chapter 5, it can be shown that s is an *unbiased* estimator of σ.