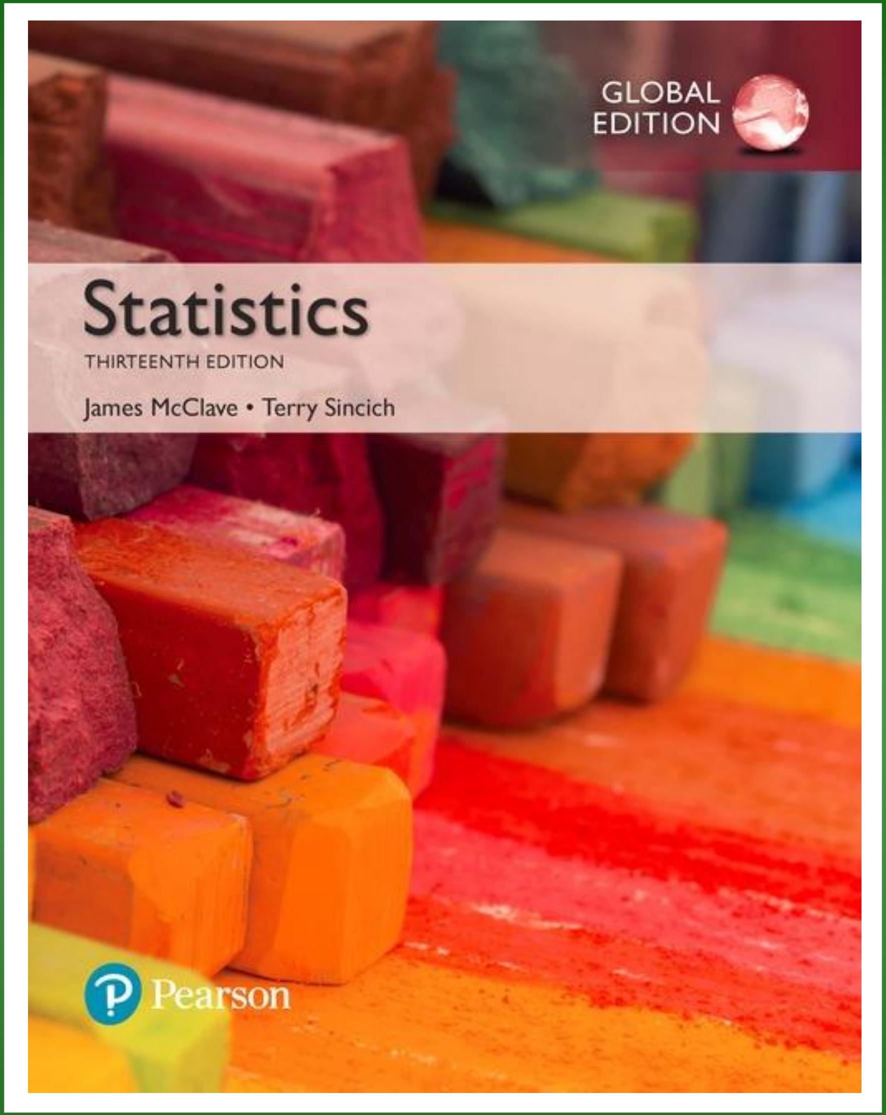


GLOBAL EDITION

Statistics

THIRTEENTH EDITION

James McClave • Terry Sincich

Pearson

Chapter 7

Inferences Based on a Single Sample

Estimation with Confidence Intervals

7.1

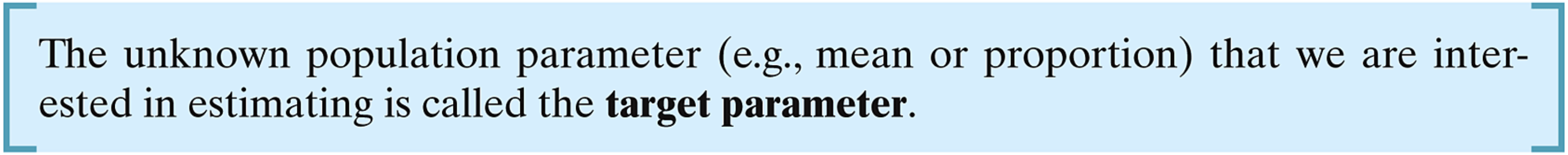
Identifying and Estimating the Target Parameter

Identifying and Estimating the Target Parameter

- Our goal is to estimate the value of an unknown population parameter,
  - a population mean / a proportion
- For example;
  - the **mean** gas mileage for a new car model
  - the **average** expected life of a flat-screen computer monitor,
  - the **proportion** of Iraq War veterans with post-traumatic stress syndrome

Definition

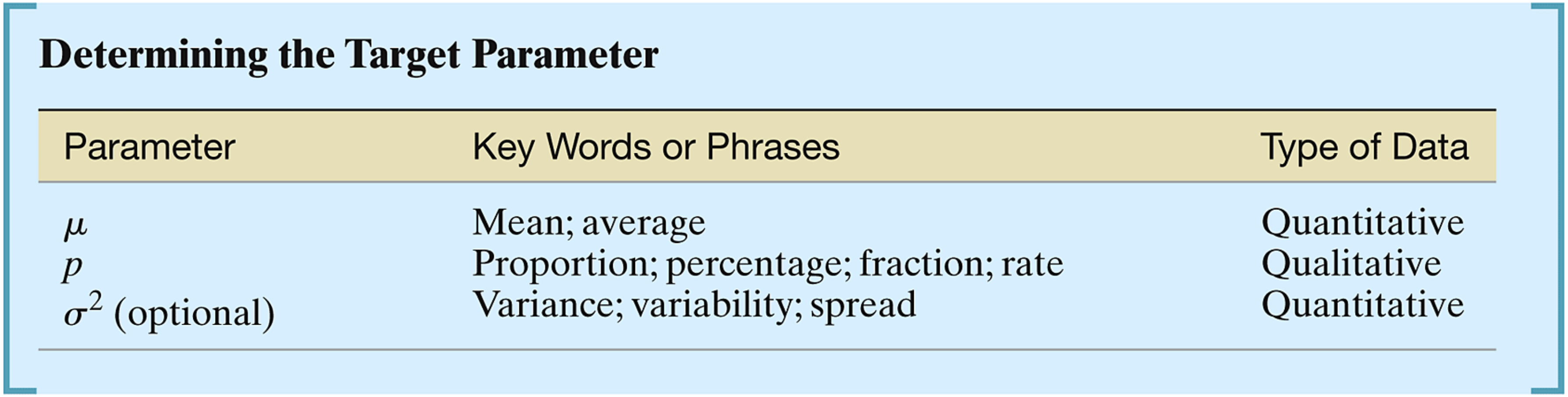
- Different techniques are used for estimating a mean or proportion, depending on whether a sample contains **a large** or **small number** of measurements.
- **Goal:** use the sample information to **estimate the population parameter** of interest and to **assess the reliability** of the estimate.



The unknown population parameter (e.g., mean or proportion) that we are interested in estimating is called the **target parameter**.

Procedure

- There are one or more **key words** in the statement of the problem that indicate the appropriate target parameter

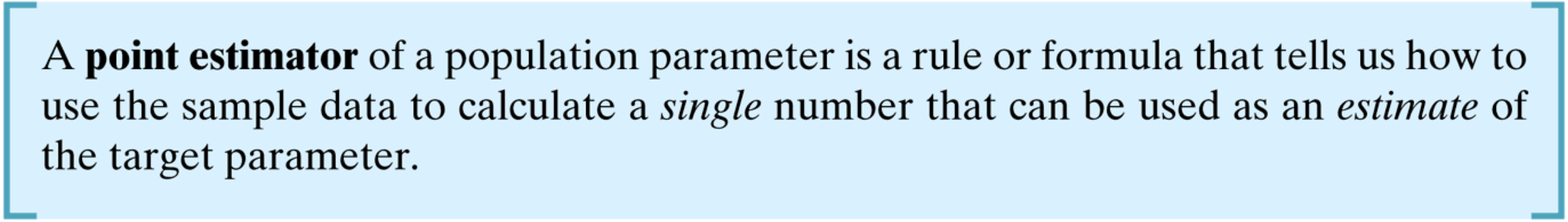


Determining the Target Parameter

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

Definition

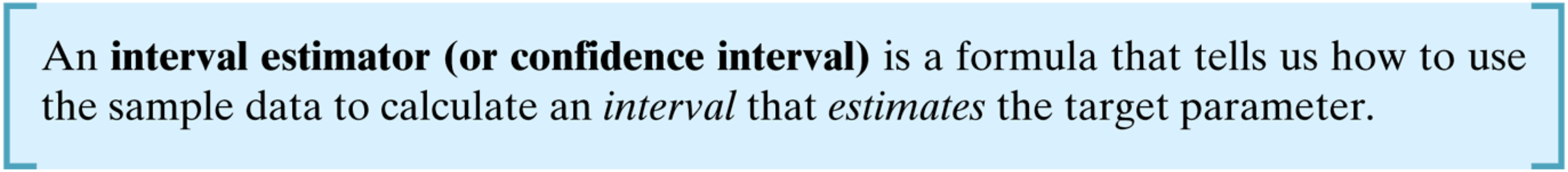
- A **single number** calculated from the sample that **estimates a target population parameter** is called a **point estimator**.
  - the sample mean, $\bar{x}$, to estimate the population mean μ
  - the sample proportion of successes, $\hat{p}$, is a point estimator for the binomial proportion p
  - the sample variance s^2 is a point estimator for the population variance σ^2



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.

Definition

- Also, we will attach a **measure of reliability** to our estimate by obtaining an **interval estimator**—a range of numbers that contains the target parameter with a **high degree of confidence**.
- For this reason the interval estimate is also called a **confidence interval**.



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.

7.2

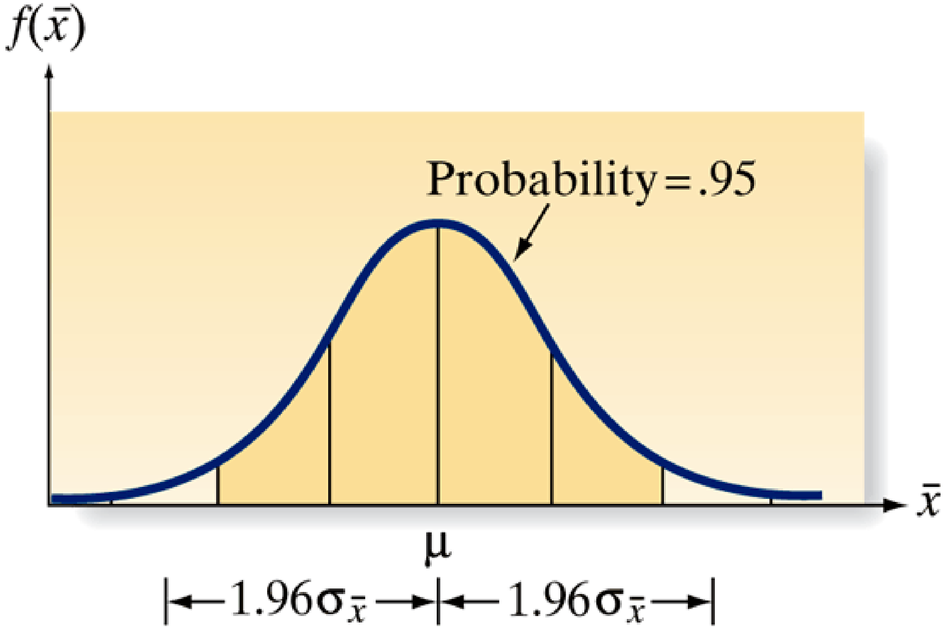
Confidence Interval for a Population Mean: Normal (*z*) Statistics

An conceptual example

- Suppose a large hospital wants to estimate the average length of time patients remain in the hospital.
- The hospital's target parameter is the population mean $\boldsymbol{\mu}$.
- To accomplish this objective, the hospital administrators plan to randomly sample 100 of all previous patients' records and to use the sample mean $\bar{\boldsymbol{x}}$ of the lengths of stay to estimate $\boldsymbol{\mu}$, the mean of all patients' visits.
- The sample mean $\bar{\boldsymbol{x}}$ represents a **point estimator** of the population mean $\boldsymbol{\mu}$.
- How can we assess the **accuracy** of this large-sample point estimator?

Figure 7.1 Sampling distribution of $\overline{x}$

- According to the Central Limit Theorem, the sampling distribution of the sample mean is approximately normal for large samples, as shown in Figure 7.1.
- Let us calculate the interval estimator:



$$\overline{x} \pm 1.96\sigma_{\overline{x}} = \overline{x} \pm \frac{1.96\sigma}{\sqrt{n}}$$

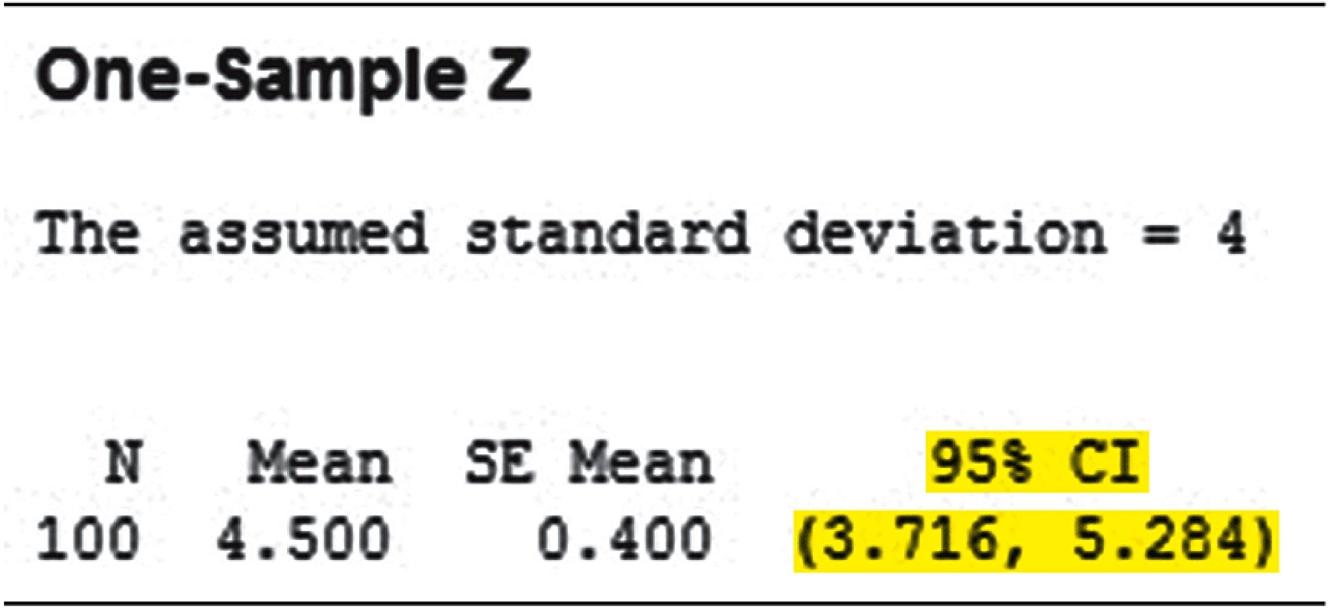
Problem

- The hospital randomly samples $\boldsymbol{n = 100}$ of its patients and finds that the sample mean length of stay is $\boldsymbol{\bar{x} = 4.5}$ days.
- Also, suppose it is known that the standard deviation of the length of stay for all hospital patients is $\boldsymbol{\sigma = 4}$ days.
- Use the interval estimator $\boldsymbol{\bar{x} \pm (1.96)\sigma_{\bar{x}}}$ to calculate a confidence interval for the target parameter, $\boldsymbol{\mu}$.

Figure 7.2 MINITAB Output Showing 95% Confidence Interval for μ, σ Known

- Substituting $\bar{\boldsymbol{x}} = \mathbf{4.5}$ and $\boldsymbol{\sigma} = \mathbf{4}$ into the interval estimator formula, we obtain:

$$\bar{x} \pm 1.96\sigma_{\bar{x}} = \bar{x} \pm (1.96)\sigma/\sqrt{n} = 4.5 \pm (1.96)(4/\sqrt{100}) = 4.5 \pm .78$$



```
One-Sample Z

The assumed standard deviation = 4


  N    Mean  SE Mean          95% CI
100   4.500    0.400  (3.716, 5.284)
```

What is a *large-sample*?

- The interval $\bar{\boldsymbol{x}} \pm (\mathbf{1.96})\boldsymbol{\sigma}_{\bar{\boldsymbol{x}}}$ in Example 7.1 is called a large-sample 95% confidence interval for the population mean $\boldsymbol{\mu}$.
- The term large-sample refers to the sample being of sufficiently large size that we can apply the Central Limit Theorem and the **normal ($\boldsymbol{z}$) statistic** to determine the form of the sampling distribution of $\bar{\boldsymbol{x}}$.
- Empirical research suggests that a sample size $\boldsymbol{n}$ exceeding a value between 20 and 30 will usually yield sampling distribution of $\bar{\boldsymbol{x}}$ that is approximately normal.
- This result led many practitioners to adopt the rule of thumb that a sample size of $\boldsymbol{n} \geq \mathbf{30}$ is required to use large-sample confidence interval procedures.

What is a *large-sample*?

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

Problem

- Refer to previous example and the problem of estimating $\boldsymbol{\mu}$, the average length of stay of a hospital's patients.
- The lengths of stay for the $\boldsymbol{n = 100}$ sampled patients are shown in Table 7.1.
- Use the data to find a 95% confidence interval for $\boldsymbol{\mu}$ and interpret the result.

Table 7.1

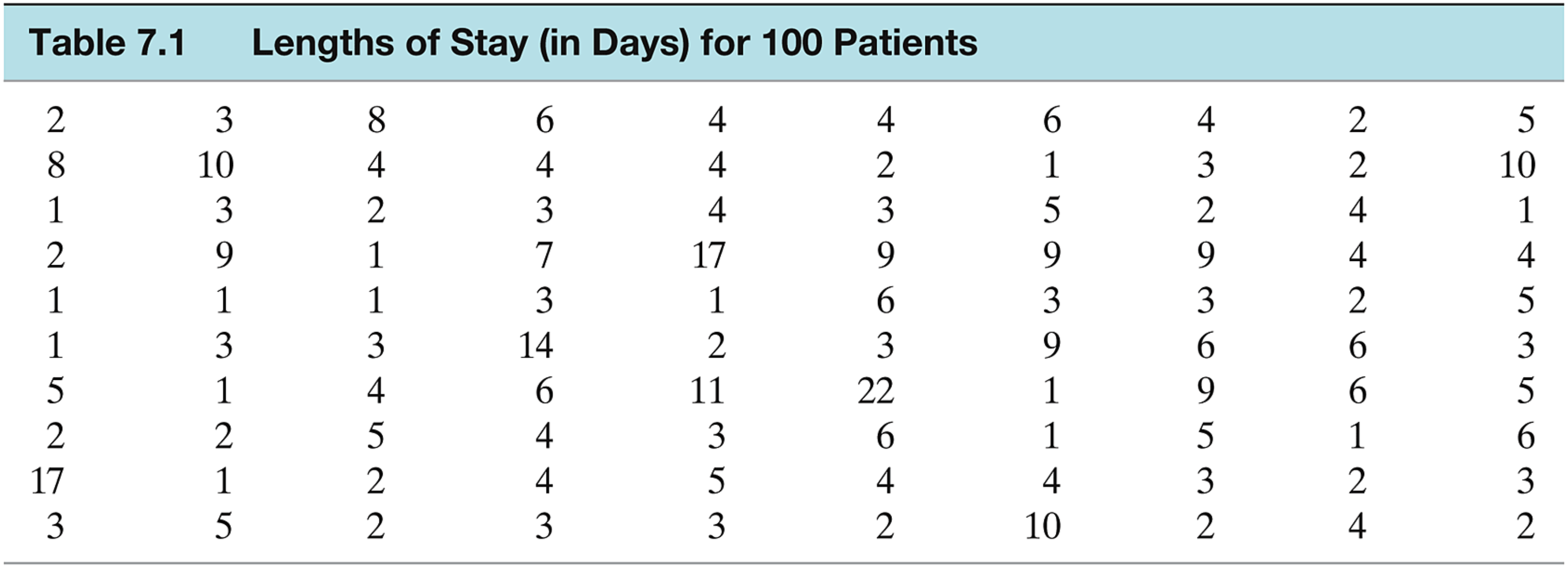
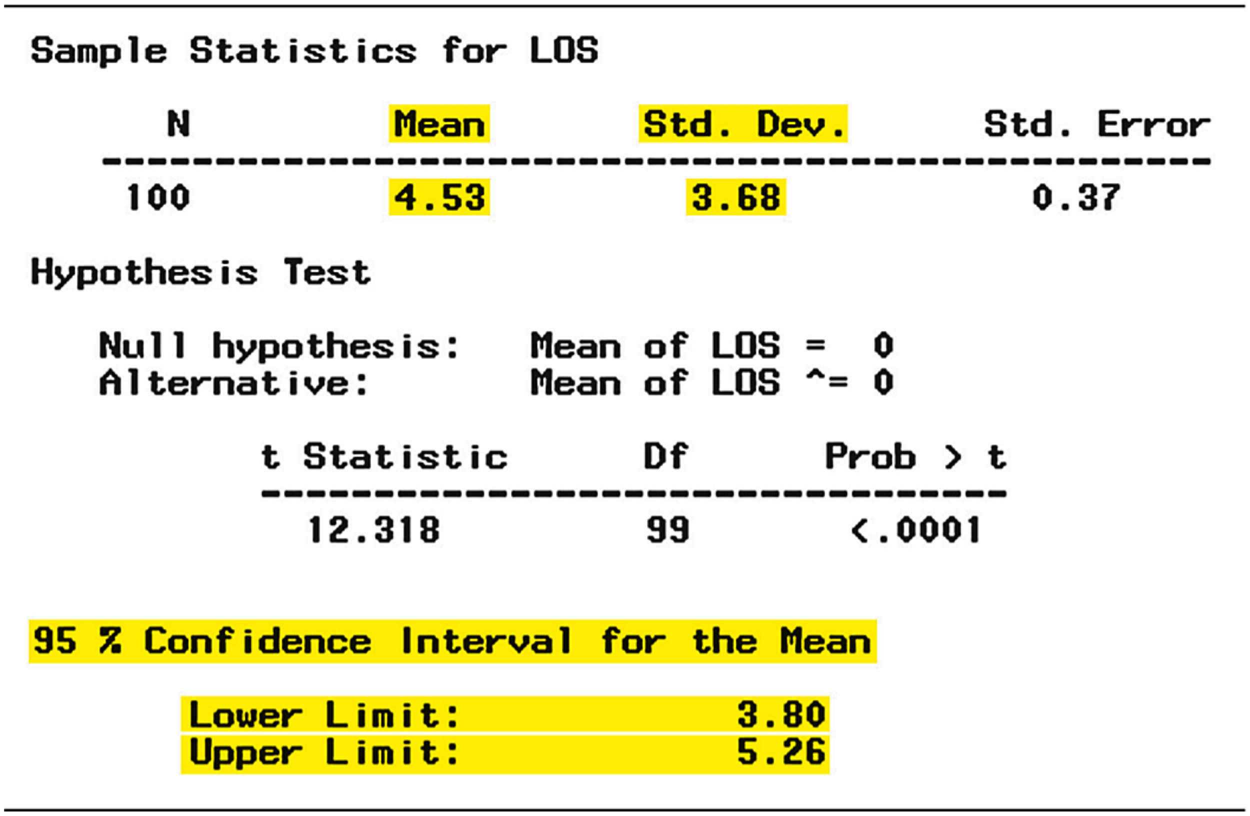


Table 7.1 Lengths of Stay (in Days) for 100 Patients

2	3	8	6	4	4	6	4	2	5
8	10	4	4	4	2	1	3	2	10
1	3	2	3	4	3	5	2	4	1
2	9	1	7	17	9	9	9	4	4
1	1	1	3	1	6	3	3	2	5
1	3	3	14	2	3	9	6	6	3
5	1	4	6	11	22	1	9	6	5
2	2	5	4	3	6	1	5	1	6
17	1	2	4	5	4	4	3	2	3
3	5	2	3	3	2	10	2	4	2

Figure 7.3 SAS printout with summary statistics and 95% confidence interval for data on 100 hospital stays



```
Sample Statistics for LOS

       N          Mean         Std. Dev.        Std. Error
    ---------------------------------------------------
      100         4.53           3.68              0.37

Hypothesis Test

    Null hypothesis:    Mean of LOS =  0
    Alternative:        Mean of LOS ^= 0

            t Statistic      Df       Prob > t
            -----------------------------------
              12.318         99        <.0001


95 % Confidence Interval for the Mean

           Lower Limit:                 3.80
           Upper Limit:                 5.26
```

Solution

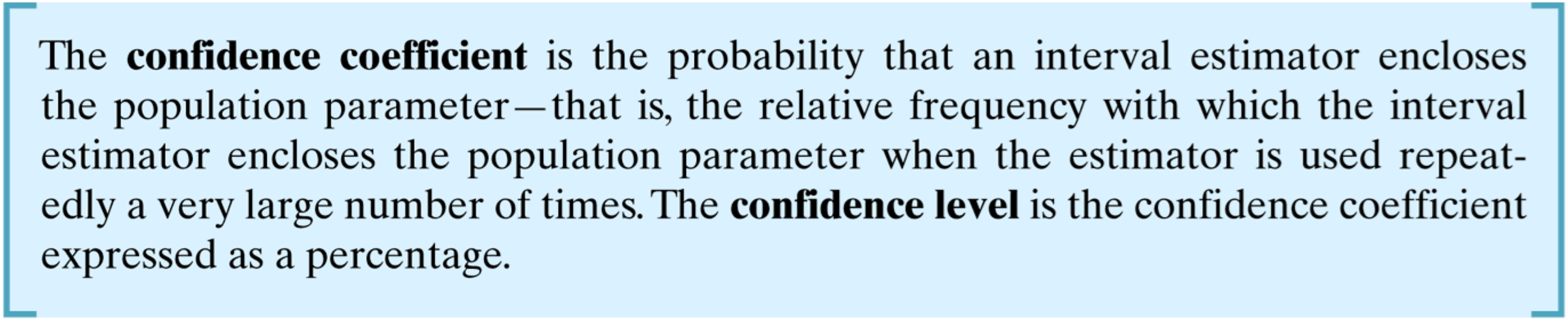
- The hospital almost surely does not know the true standard deviation, $\boldsymbol{\sigma}$, of the population of lengths of stay.
- However, since the sample size is large, we will use the sample standard deviation, $\boldsymbol{s}$, as an estimate for $\boldsymbol{\sigma}$ in the confidence interval formula.

$$\bar{x} \pm (1.96)\sigma/\sqrt{n} \approx \bar{x} \pm (1.96)s/\sqrt{n} = 4.53 \pm (1.96)(3.68)/\sqrt{100} = 4.53 \pm .72$$

- Or, $(\mathbf{3.81, 5.25})$. That is, we estimate the mean length of stay for all hospital patients to fall within the interval 3.81 to 5.25 days.

Definition

- Can we be sure that $\boldsymbol{\mu}$, the true mean, is in the interval from 3.81 to 5.25?
- We cannot be certain, but we can be reasonably confident that it is.
- This confidence is derived from the knowledge that if we were to draw repeated random samples of 100 measurements from this population and form the interval $\bar{x} \pm (1.96)\sigma_{\bar{x}}$ each time, 95% of the intervals would contain $\boldsymbol{\mu}$.



The **confidence coefficient** is the probability that an interval estimator encloses the population parameter—that is, the relative frequency with which the interval estimator encloses the population parameter when the estimator is used repeatedly a very large number of times. The **confidence level** is the confidence coefficient expressed as a percentage.

Figure 7.4 Confidence intervals for μ: 10 samples

- Now we have seen how an interval can be used to estimate a population mean.
- When we use an interval estimator, we can usually calculate the probability that the estimation process will result in an interval that contains the true value of the population mean.
- That is, the probability that the interval contains the parameter in repeated usage is usually known.
- **If our confidence level is 95%, then in the long run, 95% of our sample confidence intervals will contain μ.**

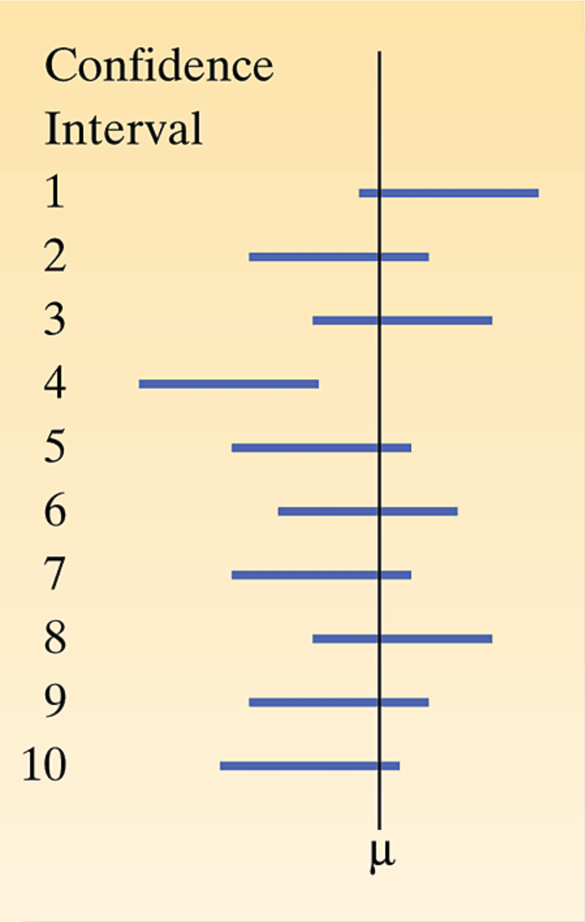
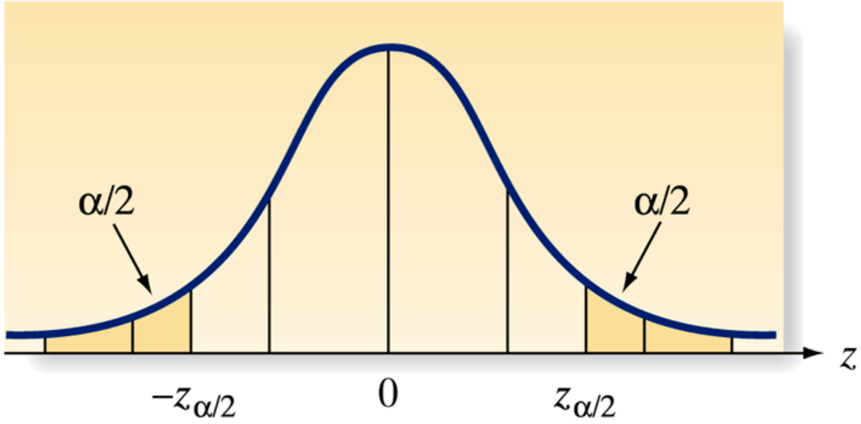


Figure 7.5 Locating $z_{\alpha/2}$ on the standard normal curve

- Suppose you wish to choose a confidence coefficient other than .95.
- We can construct a confidence interval with any desired confidence coefficient by increasing or decreasing the area (call it α) assigned to the tails of the sampling distribution.
- If we place the area $\alpha/2$ in each tail and if $z_{\alpha/2}$ is the z value such that $\alpha/2$ will lie to its right, then the confidence interval with confidence coefficient $(1-\alpha)$ is

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$$



Definition

The value z_α is defined as the value of the standard normal random variable z such that the area α will lie to its right. In other words, $P(z > z_\alpha) = \alpha$.

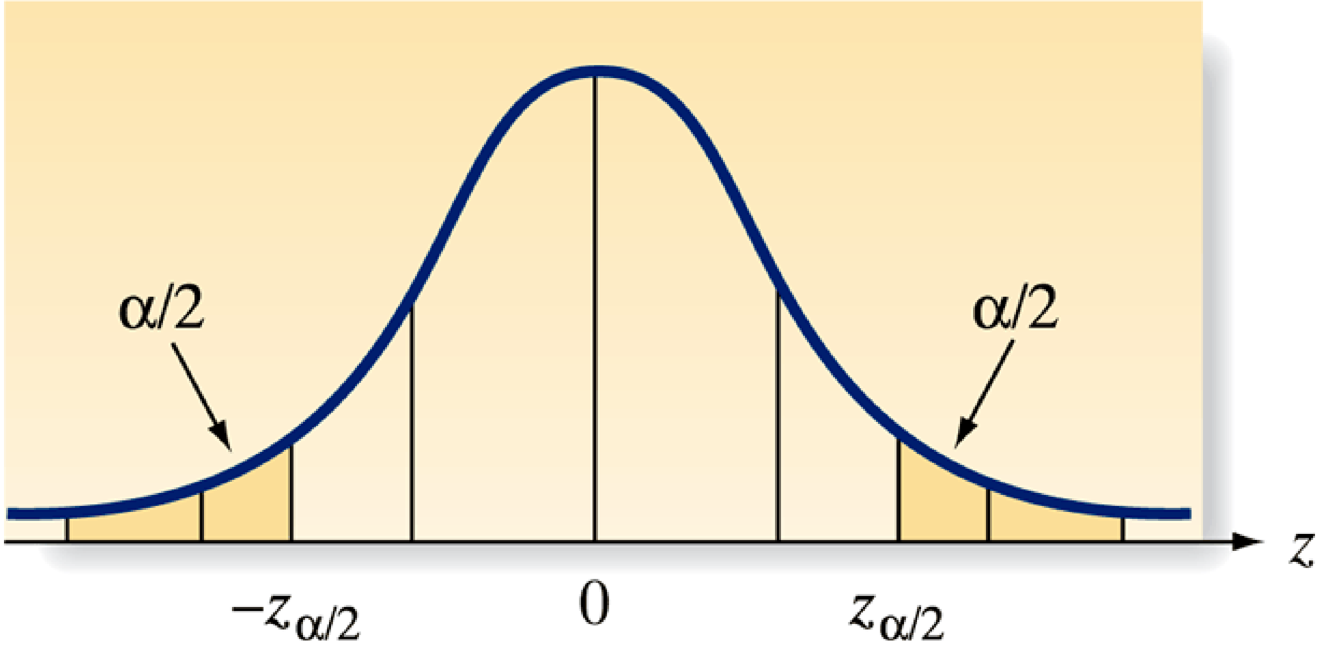


Figure 7.6 The z value ($z_{.05}$) corresponding to an area equal to .05 in the upper tail of the *z*-distribution

- To illustrate, for a confidence coefficient of $.90$, we have $(1-\alpha)=.90, \alpha=.10$, and $\frac{\alpha}{2}=.05$; $z_{.05}$ is the z value that locates area $.05$ in the upper tail of the sampling distribution.
- Since the total area to the right of the mean is $.5$, we find that $z_{.05}$ will be the z value corresponding to an area of $.5-.05 = .45$ to the right of the mean.
- $z_{.05} = 1.645$

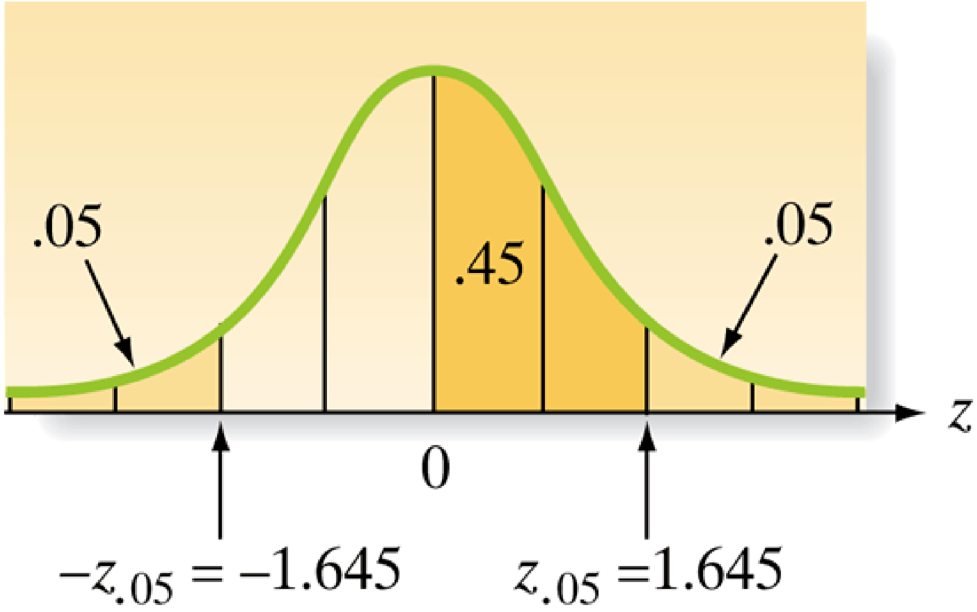
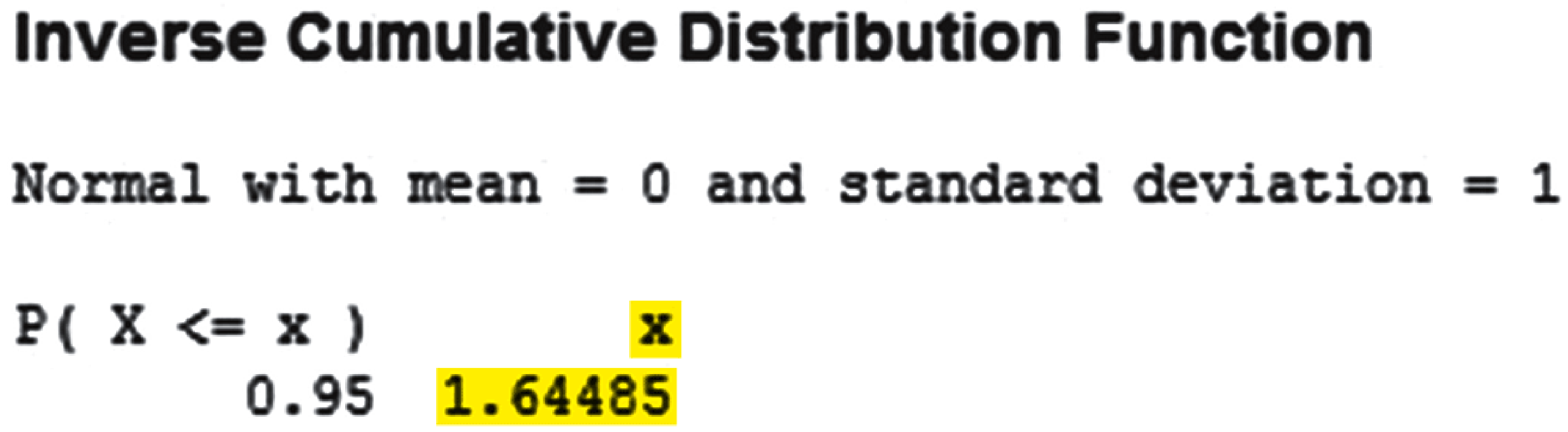


Figure 7.7 MINITAB output for Finding $z_{.05}$

- The same result may also be obtained using technology.
- The MINITAB printout in Figure 7.7 shows that the z-value that cuts off an upper tail area of $.05$ is approximately $z_{.05} = 1.645$.



```
Inverse Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1

P( X <= x )          x
       0.95    1.64485
```

Table 7.2

- Confidence coefficients used in practice usually range from $.90$ to $.99$.

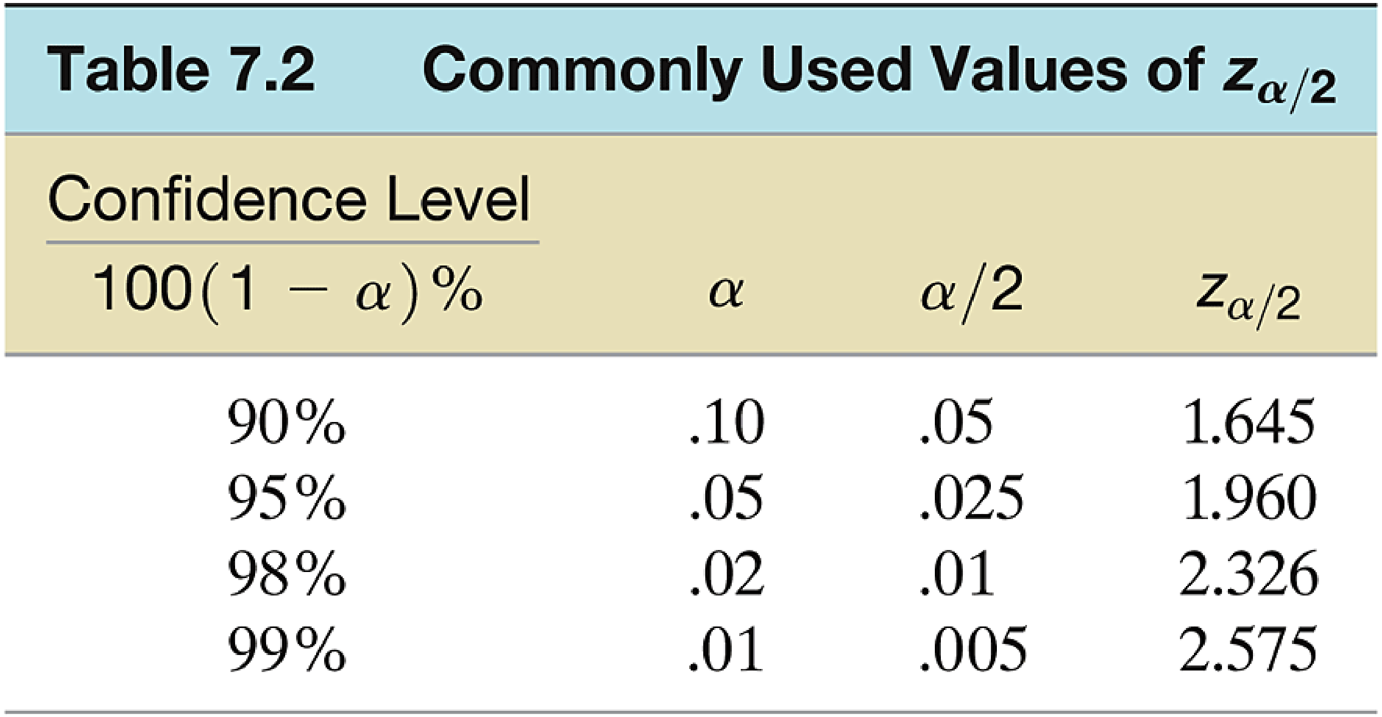
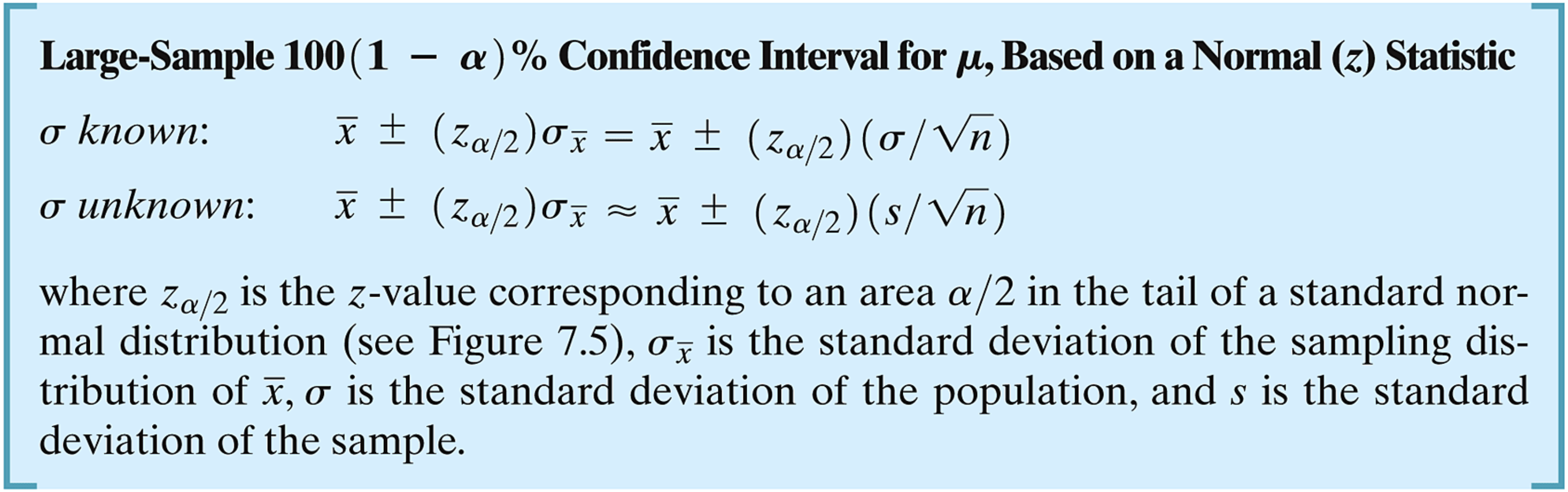


Table 7.2 Commonly Used Values of $z_{\alpha/2}$

Confidence Level $100(1-\alpha)\%$	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	1.645
95%	.05	.025	1.960
98%	.02	.01	2.326
99%	.01	.005	2.575

Procedure



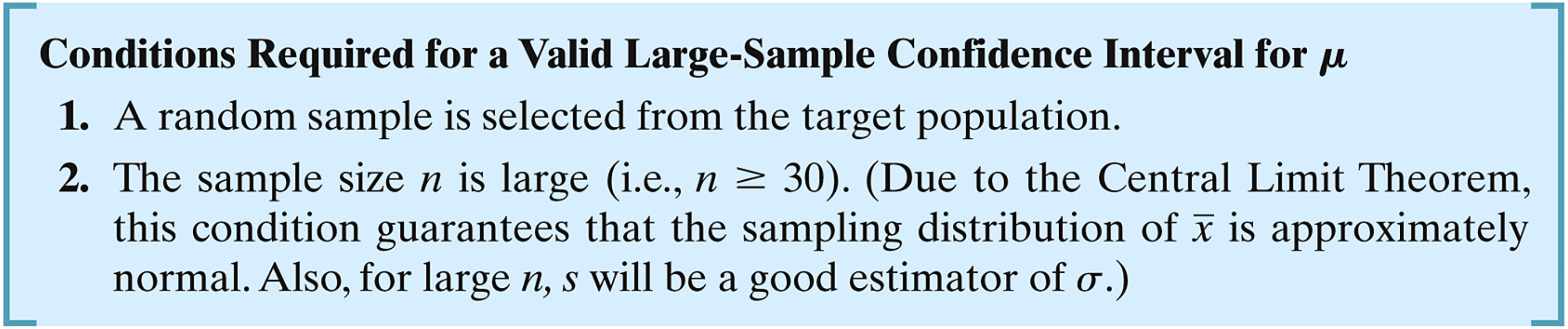
Large-Sample $100(1 - \alpha)\%$ Confidence Interval for μ, Based on a Normal (z) Statistic

σ *known*: $\bar{x} \pm (z_{\alpha/2})\sigma_{\bar{x}} = \bar{x} \pm (z_{\alpha/2})(\sigma/\sqrt{n})$

σ *unknown*: $\bar{x} \pm (z_{\alpha/2})\sigma_{\bar{x}} \approx \bar{x} \pm (z_{\alpha/2})(s/\sqrt{n})$

where $z_{\alpha/2}$ is the z-value corresponding to an area $\alpha/2$ in the tail of a standard normal distribution (see Figure 7.5), $\sigma_{\bar{x}}$ is the standard deviation of the sampling distribution of $\bar{x}$, σ is the standard deviation of the population, and s is the standard deviation of the sample.

Procedure

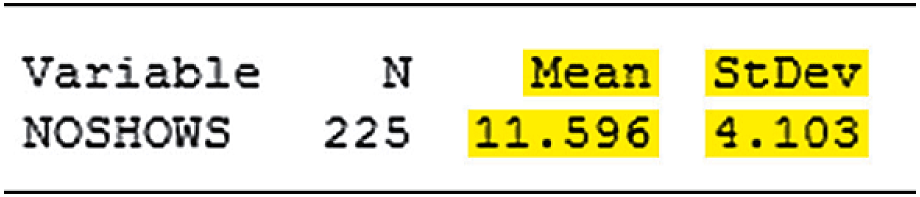


Conditions Required for a Valid Large-Sample Confidence Interval for μ

1. A random sample is selected from the target population.
2. The sample size n is large (i.e., $n \geq 30$). (Due to the Central Limit Theorem, this condition guarantees that the sampling distribution of $\bar{x}$ is approximately normal. Also, for large n, s will be a good estimator of σ.)

Problem

- Unoccupied seats on flights cause airlines to lose revenue. Suppose a large airline wants to estimate its average number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected, and the number of unoccupied seats is noted for each of the sampled flights. Descriptive statistics are given below. Estimate μ, the mean number of unoccupied seats per flight during the past year, using a 90% confidence interval.



```
Variable      N      Mean   StDev
NOSHOWS     225    11.596   4.103
```

Solution

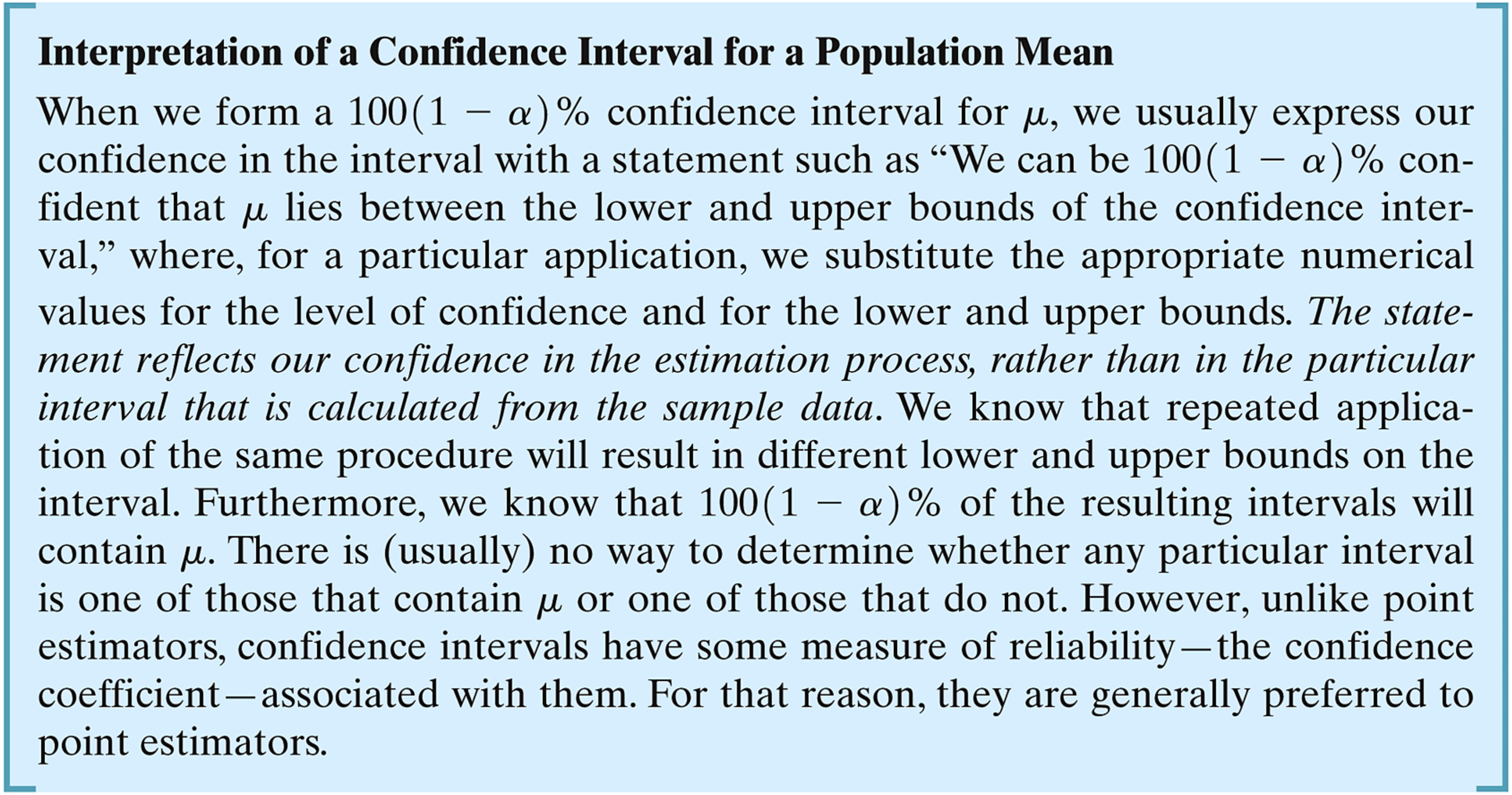
- The form of a large-sample 90% confidence interval for a population mean is:

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm z_{.05}\sigma_{\bar{x}} = \bar{x} \pm 1.645\left(\frac{\sigma}{\sqrt{n}}\right)$$

- Since we do not know the value of $\boldsymbol{\sigma}$, we use our best approximation—the sample standard deviation, $\boldsymbol{s = 4.1}$,

$$11.6 \pm 1.645\left(\frac{4.1}{\sqrt{225}}\right) = 11.6 \pm .45$$

Procedure



Interpretation of a Confidence Interval for a Population Mean

When we form a $100(1 - \alpha)\%$ confidence interval for μ, we usually express our confidence in the interval with a statement such as "We can be $100(1 - \alpha)\%$ confident that μ lies between the lower and upper bounds of the confidence interval," where, for a particular application, we substitute the appropriate numerical values for the level of confidence and for the lower and upper bounds. *The statement reflects our confidence in the estimation process, rather than in the particular interval that is calculated from the sample data*. We know that repeated application of the same procedure will result in different lower and upper bounds on the interval. Furthermore, we know that $100(1 - \alpha)\%$ of the resulting intervals will contain μ. There is (usually) no way to determine whether any particular interval is one of those that contain μ or one of those that do not. However, unlike point estimators, confidence intervals have some measure of reliability—the confidence coefficient—associated with them. For that reason, they are generally preferred to point estimators.

7.3

Confidence Interval for a Population Mean: Student’s *t*-Statistic

How to work with a small sample size?

- Suppose a pharmaceutical company must estimate the average increase in blood pressure of patients who take a certain new drug.
- Assume that only six patients (randomly selected from the population of all patients) can be used in the initial phase of human testing.
- The use of a *small sample* in making an inference about μ presents two immediate problems when we attempt to use the standard normal z as a test statistic.

Problems with z statistic - 1

- The shape of the sampling distribution of the sample mean $\bar{x}$ (and the z-statistic) now depends on the shape of the population that is sampled.
- We can no longer assume that the sampling distribution of $\bar{x}$ is approximately normal because the Central Limit Theorem ensures normality only for samples that are sufficiently large.

Solution to Problem 1

- The sampling distribution of $\overline{x}$ (and z) is exactly normal even for relatively small samples if the sampled population is normal.
- It is approximately normal if the sampled population is approximately normal.

Problems with z statistic - 2

- The population standard deviation $\boldsymbol{\sigma}$ is almost always unknown.
- Although it is still true that $\boldsymbol{\sigma_{\bar{x}} = \sigma/\sqrt{n}}$, the sample standard deviation $\boldsymbol{s}$ may provide a poor approximation for $\boldsymbol{\sigma}$ when the sample size is small.

Solution to Problem 2

- Instead of using the standard normal statistic

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

which requires knowledge of, or a good approximation to, σ, we define and use the statistic

random quantities

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

in which the sample standard deviation s replaces the population standard deviation σ.

Figure 7.9 Standard normal (*z*) distribution and *t*-distributions

- The actual amount of variability in the sampling distribution of t depends on the sample size n.
- A convenient way of expressing this dependence is to say that the t statistic has $(n - 1)$ degrees of freedom (df).

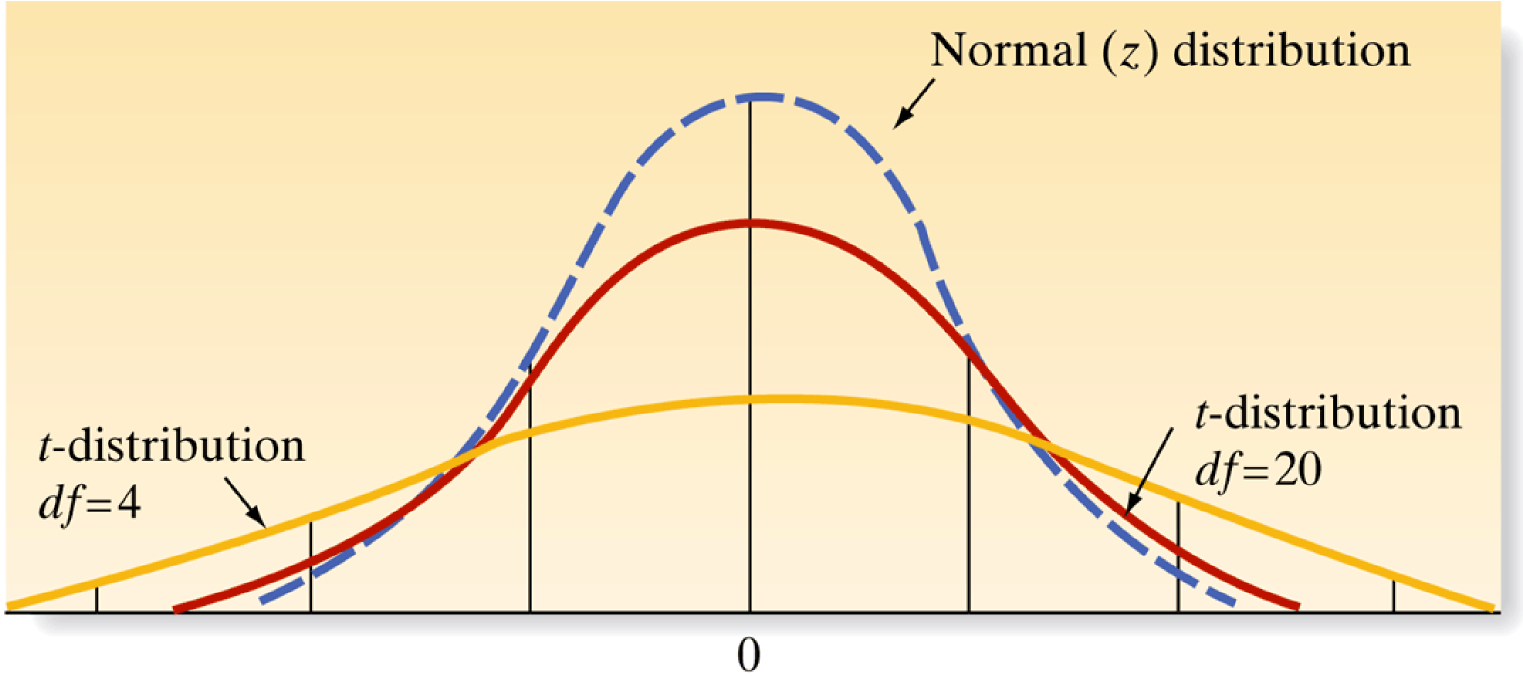
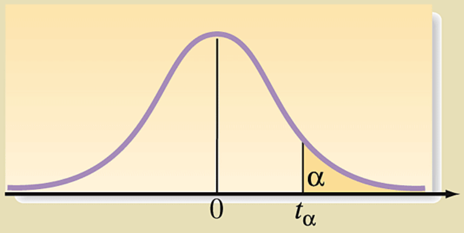


Table 7.3

- As the sample size $\boldsymbol{n}$ grows very large, $\boldsymbol{s}$ becomes closer to $\boldsymbol{\sigma}$ and thus $\boldsymbol{t}$ becomes closer in distribution to $\boldsymbol{z}$.
- when $\boldsymbol{df = 29}$, there is little difference between $\boldsymbol{z}$ and $\boldsymbol{t}$.

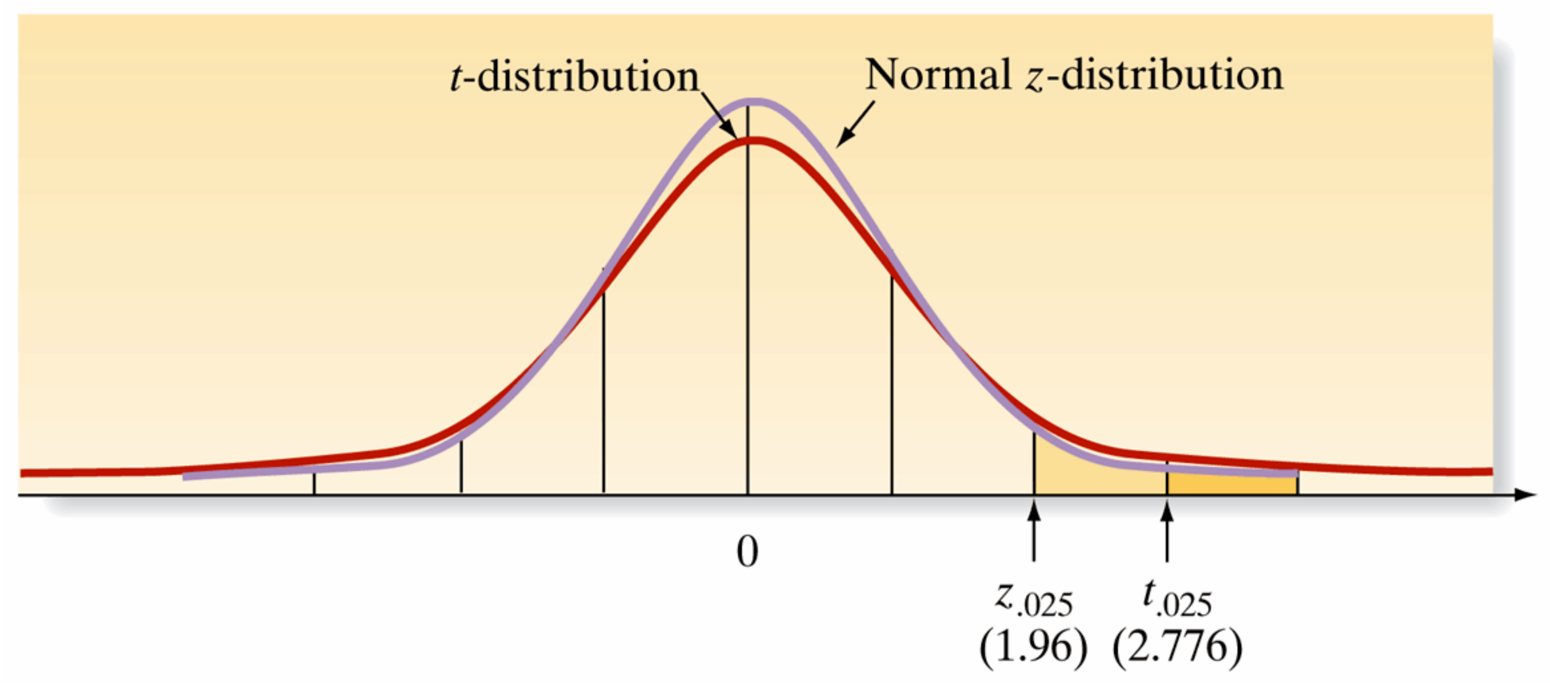
Table 7.3 Reproduction of Part of Table III in Appendix B



Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$	$t_{.0005}$
1	3.078	6.314	12.706	31.821	63.657	318.31	636.62
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Thus, we choose the arbitrary cutoff of $\boldsymbol{n = 30}$ $(\boldsymbol{df = 29})$ to distinguish between large-sample and small-sample inferential techniques when $\boldsymbol{\sigma}$ is unknown.

Figure 7.10 The $t_{.025}$ value in a t-distribution with 4 df, and the corresponding $z_{.025}$ value



Problem

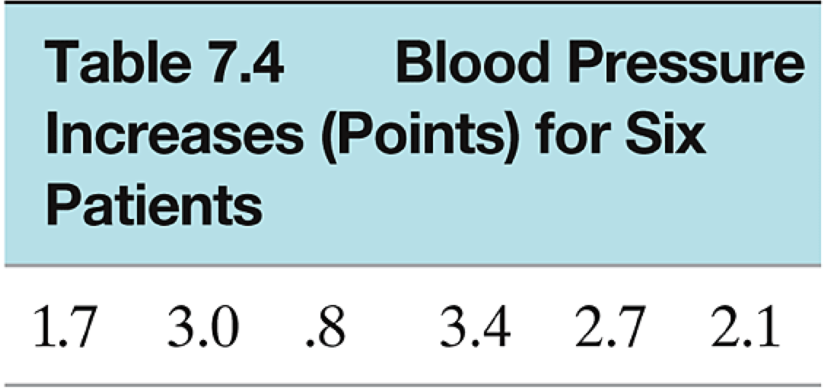


Table 7.4 Blood Pressure Increases (Points) for Six Patients

1.7	3.0	.8	3.4	2.7	2.1

- Consider the pharmaceutical company that desires an estimate of the mean increase in blood pressure of patients who take a new drug.
- The blood pressure increases (measured in points) for the $n = 6$ patients in the human testing phase are shown in Table 7.4.
- Use this information to construct a 95% confidence interval for μ, the mean increase in blood pressure associated with the new drug for all patients in the population.

Solution

- First, note that we are dealing with a sample too small to assume, by the Central Limit Theorem, that the sample mean $\bar{x}$ is approximately normally distributed.
- That is, we do not get the normal distribution of $\bar{x}$ “automatically” from the Central Limit Theorem when the sample size is small.
- Instead, we must *assume* that the measured variable, in this case the increase in blood pressure, is normally distributed in order for the distribution of $\bar{x}$ to be normal.

Solution

- Second, unless we are fortunate enough to know the population standard deviation $\boldsymbol{\sigma}$, which in this case represents the standard deviation of all the patients' increases in blood pressure when they take the new drug, we cannot use the standard normal $\boldsymbol{z}$-statistic to form our confidence interval for $\boldsymbol{\mu}$.
- Instead, we must use the $\boldsymbol{t}$-distribution, with $(\boldsymbol{n}-\mathbf{1})$ degrees of freedom.

Solution

- In this case, $\boldsymbol{n-1=5}$ df, and the $\boldsymbol{t}$-value is found in Table 7.3 to be $\boldsymbol{t_{.025}=2.571}$ with 5 df
- Recall that the large-sample confidence interval would have been of the form

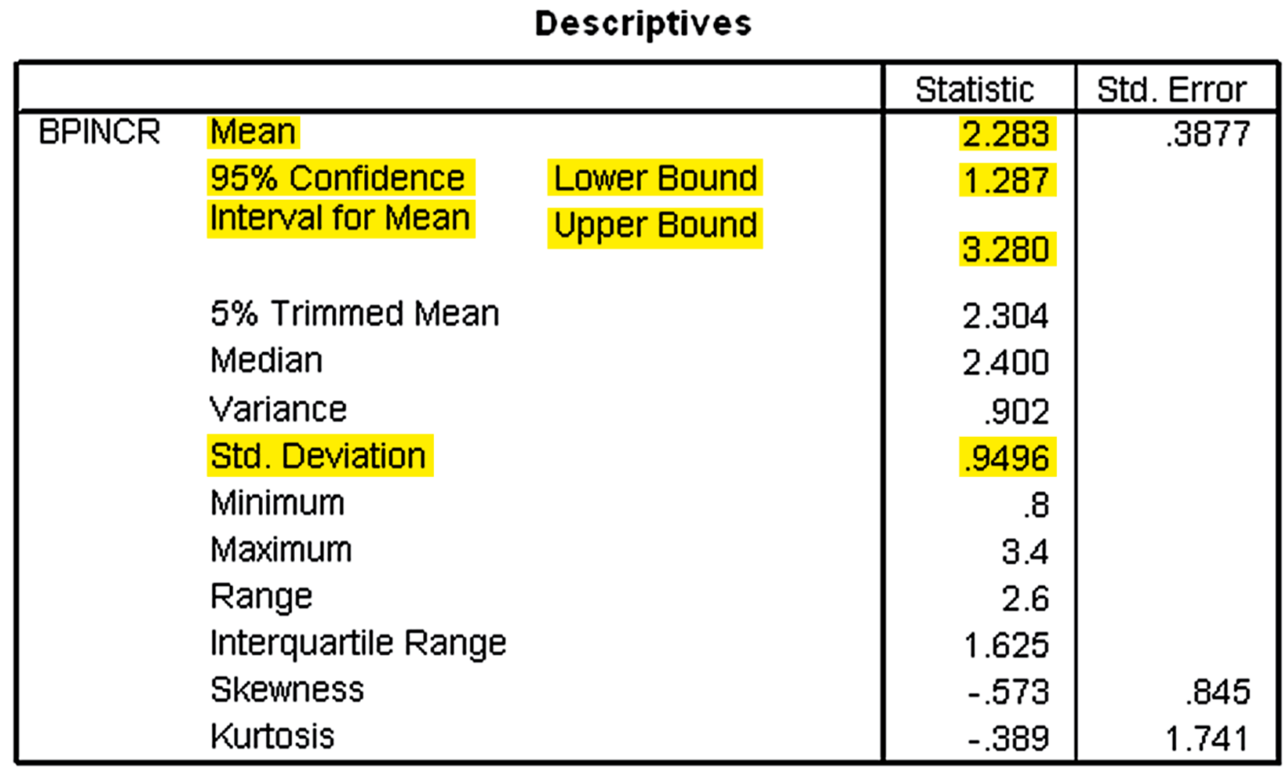
$$\overline{x} \pm z_{\alpha/2}\sigma_{\overline{x}} = \overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \overline{x} \pm z_{.025}\frac{\sigma}{\sqrt{n}}$$

where 95% is the desired confidence level.

- To form the interval for a small sample from *a normal distribution, we simply substitute* $\boldsymbol{t}$ *for* $\boldsymbol{z}$ *and* $\boldsymbol{s}$ *for* $\boldsymbol{\sigma}$ *in the preceding formula, yielding*

$$\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Figure 7.11 SPSS confidence interval for mean blood pressure increase



Descriptives

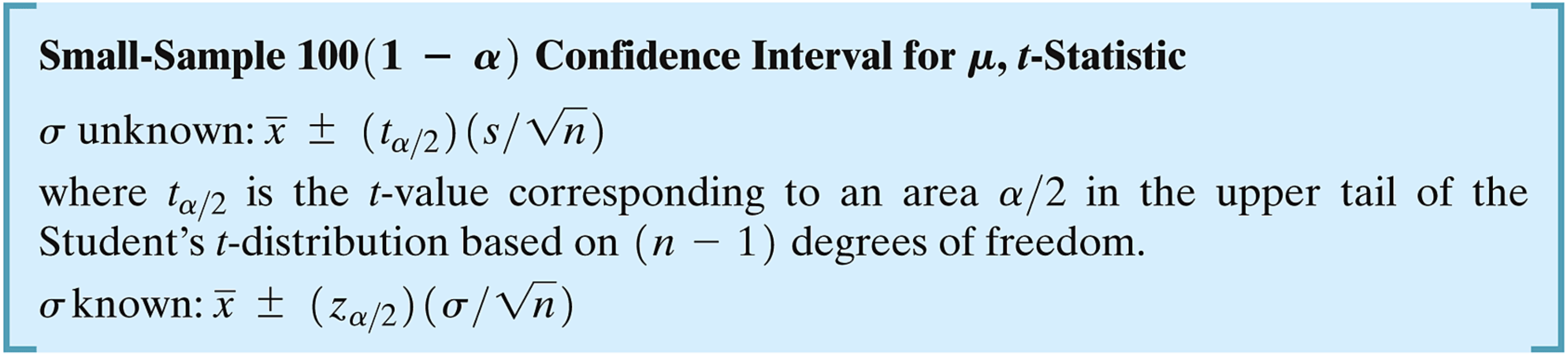
			Statistic	Std. Error
BPINCR	Mean		2.283	.3877
	95% Confidence Interval for Mean	Lower Bound	1.287	
		Upper Bound	3.280	
	5% Trimmed Mean		2.304	
	Median		2.400	
	Variance		.902	
	Std. Deviation		.9496	
	Minimum		.8	
	Maximum		3.4	
	Range		2.6	
	Interquartile Range		1.625	
	Skewness		-.573	.845
	Kurtosis		-.389	1.741

- $\boldsymbol{t_{.025} = 2.571}$ with 5 df

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

$$2.283 \pm (2.571)\left(\frac{.950}{\sqrt{6}}\right) = 2.283 \pm .997$$

Procedure



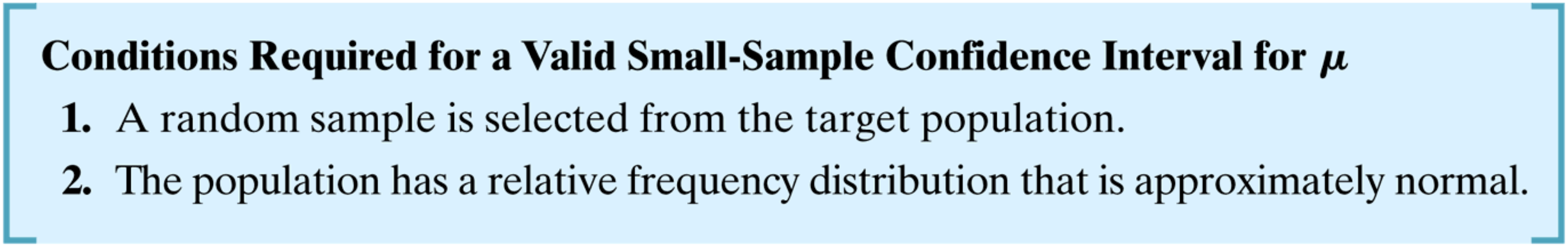
Small-Sample $100(1 - \alpha)$ Confidence Interval for μ, t-Statistic

σ unknown: $\bar{x} \pm (t_{\alpha/2})(s/\sqrt{n})$

where $t_{\alpha/2}$ is the t-value corresponding to an area $\alpha/2$ in the upper tail of the Student's t-distribution based on $(n - 1)$ degrees of freedom.

σ known: $\bar{x} \pm (z_{\alpha/2})(\sigma/\sqrt{n})$

Definition



Conditions Required for a Valid Small-Sample Confidence Interval for μ

1. A random sample is selected from the target population.
2. The population has a relative frequency distribution that is approximately normal.

Problem

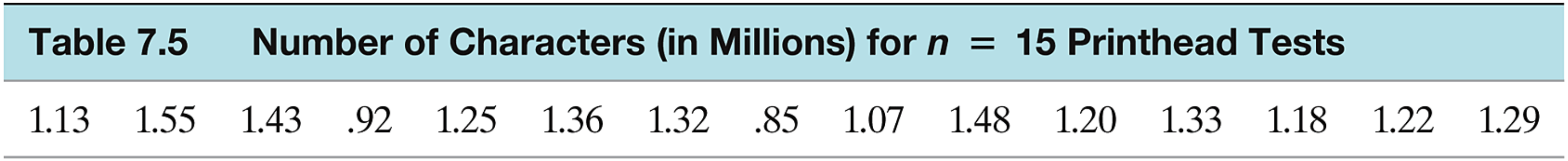


Table 7.5 Number of Characters (in Millions) for $n = 15$ Printhead Tests

1.13	1.55	1.43	.92	1.25	1.36	1.32	.85	1.07	1.48	1.20	1.33	1.18	1.22	1.29

- Some quality control experiments require destructive sampling (i.e., the test to determine whether the item is defective destroys the item) in order to measure a particular characteristic of the product.
- The cost of destructive sampling often dictates small samples.
- Suppose a manufacturer of printers for personal computers wishes to estimate the mean number of characters printed before the printhead fails.
- The printer manufacturer tests $n = 15$ printheads and records the number of characters printed until failure for each.

Figure 7.12 MINITAB printout with descriptive statistics and 99% confidence interval for Example 7.5

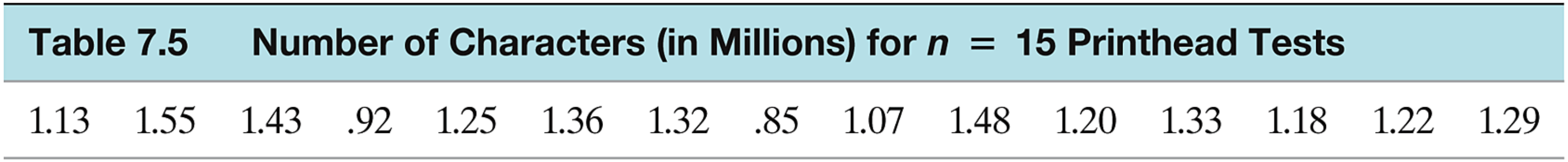
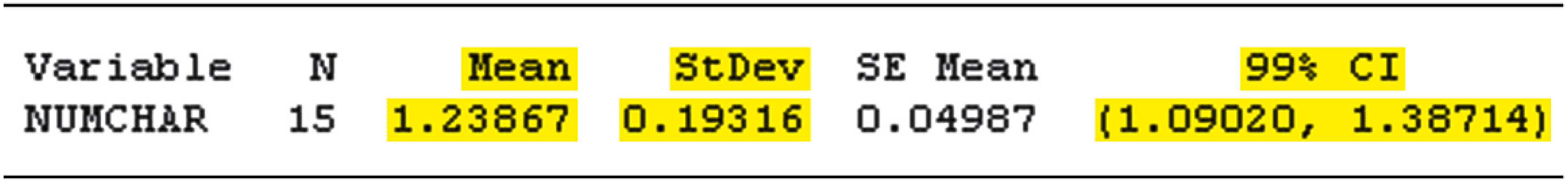


Table 7.5 Number of Characters (in Millions) for *n* = 15 Printhead Tests

1.13	1.55	1.43	.92	1.25	1.36	1.32	.85	1.07	1.48	1.20	1.33	1.18	1.22	1.29



```
Variable    N       Mean      StDev   SE Mean          99% CI
NUMCHAR    15    1.23867    0.19316   0.04987   (1.09020, 1.38714)
```

a) Form a 99% confidence interval for the mean number of characters printed before the printhead fails. Interpret the result.

b) What assumption is required for the interval you found in part ***a*** to be valid? Is that assumption reasonably satisfied?

Solution

- For this small sample ($\boldsymbol{n = 15}$), we use the $\boldsymbol{t}$-statistic to form the confidence interval.
- We use a confidence coefficient of $\mathbf{0.99}$ and $\boldsymbol{n - 1 = 14}$ df to find $\boldsymbol{t_{\alpha/2}}$ in Table III: → $\boldsymbol{t_{\alpha/2} = t_{.005} = 2.977}$

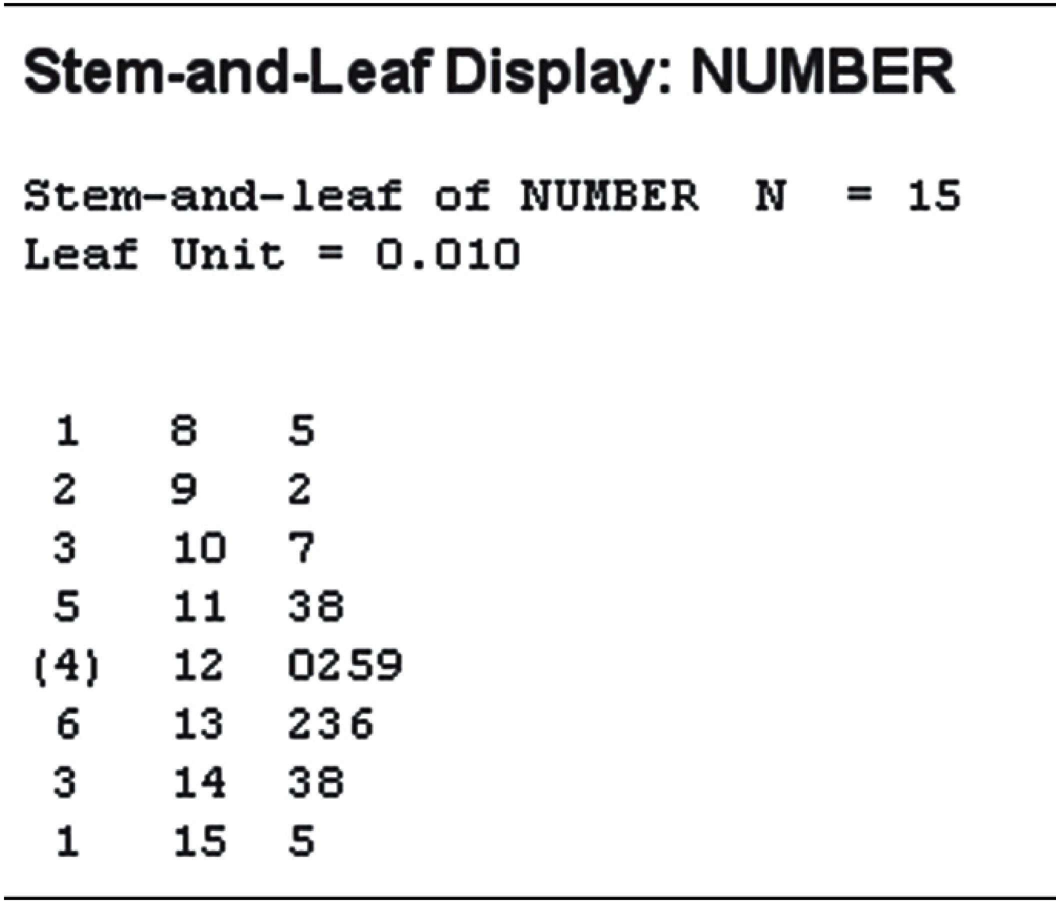
$$\bar{x} \pm t_{.005}\left(\frac{s}{\sqrt{n}}\right) = 1.24 \pm 2.977\left(\frac{.19}{\sqrt{15}}\right)$$

$$= 1.24 \pm .15 \text{ or } (1.09, 1.39)$$

- The manufacturer can be 99% confident that the printhead has a mean life of between 1.09 and 1.39 million characters. If the manufacturer advertises that the mean life of its printheads is (at least) 1 million characters, the interval would support such a claim.

Figure 7.13 MINITAB stem-and-leaf display of data in Table 7.5

- Since *n* is small, we must assume that the number of characters printed before the printhead fails is a random variable from a normal distribution.
- One way to check this assumption is to graph the distribution of data in Table 7.5.



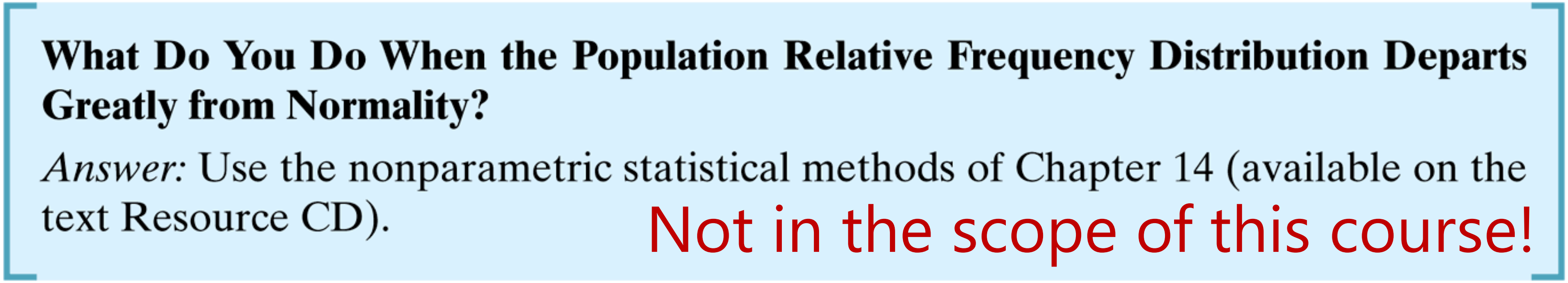
Stem-and-Leaf Display: NUMBER

```
Stem-and-leaf of NUMBER  N  = 15
Leaf Unit = 0.010


 1    8   5
 2    9   2
 3   10   7
 5   11   38
(4)  12   0259
 6   13   236
 3   14   38
 1   15   5
```

Procedure

- Although many phenomena do have approximately normal distributions, it is also true that many random phenomena have distributions that are not normal or even mound shaped.
- Empirical evidence acquired over the years has shown that the t-distribution is rather insensitive to moderate departures from normality.



What Do You Do When the Population Relative Frequency Distribution Departs Greatly from Normality?

Answer: Use the nonparametric statistical methods of Chapter 14 (available on the text Resource CD).

Not in the scope of this course!

7.4

Large-Sample Confidence Interval for a Population Proportion

Concept

- Estimating the percentage of some group with a certain characteristic
  - the **percentage** of people in favor of the president’s welfare-reform program,
  - the **fraction** of voters in favor of a certain candidate,
  - the **fraction** of customers who favor a particular brand of wine,
  - the **proportion** of people who smoke cigarettes

Problem

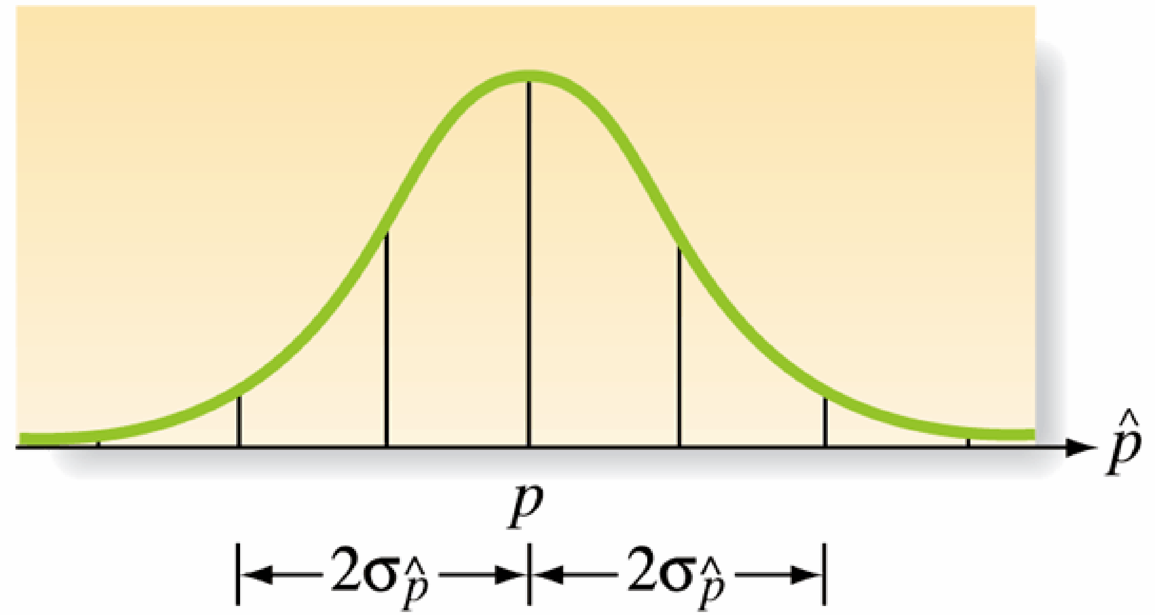
- Public-opinion polls are conducted regularly to estimate the fraction of U.S. citizens who trust the president.
- Suppose 1,000 people are randomly chosen and 637 answer that they trust the president.
- How would you estimate the true fraction of all U.S. citizens who trust the president?

Solution

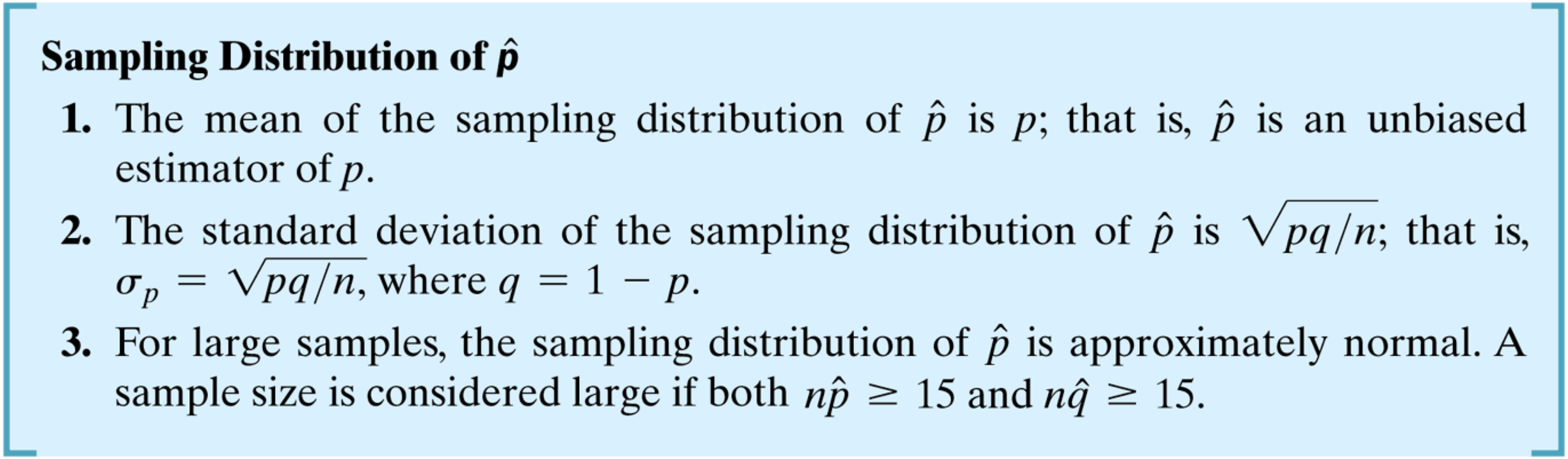
- What we have really asked is how you would estimate the probability $\boldsymbol{p}$ of success in a binomial experiment in which $\boldsymbol{p}$ is the probability that a person chosen trusts the president.
- One logical method of estimating $\boldsymbol{p}$ for the population is to use the proportion of successes in the sample. That is, we can estimate $\boldsymbol{p}$ by calculating

$$\hat{p} = \frac{\text{Number of people sampled who trust the president}}{\text{Number of people sampled}} \qquad \hat{p} = \frac{637}{1{,}000} = .637$$

Figure 7.14 Sampling distribution of $\hat{p}$



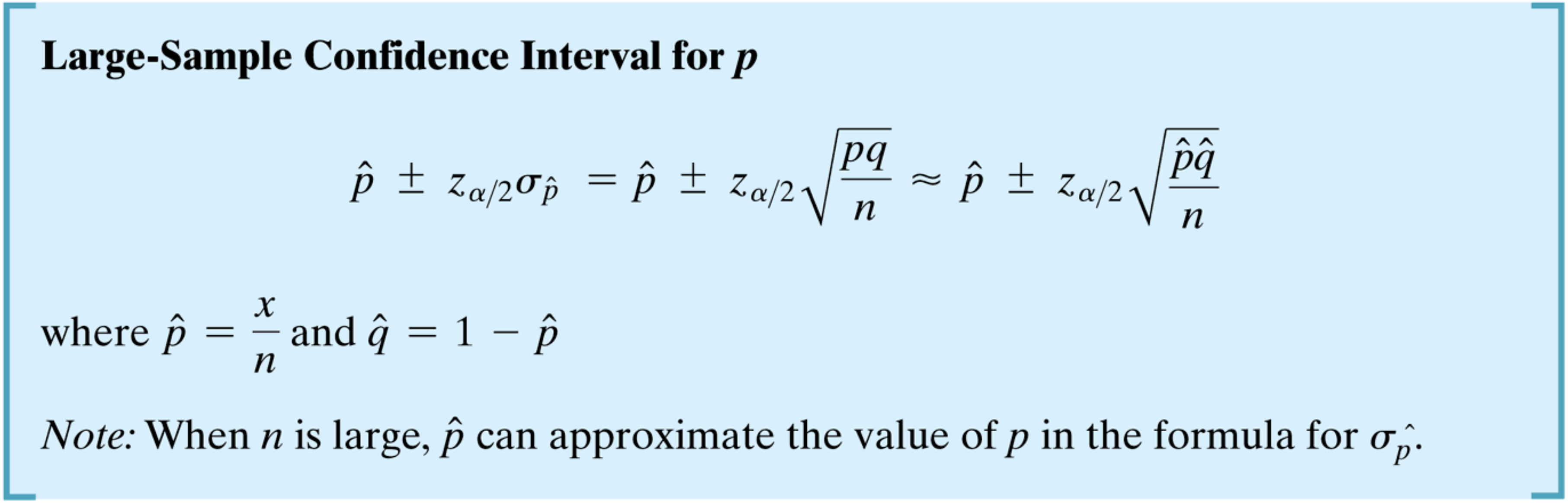
Procedure



Sampling Distribution of $\hat{p}$

1. The mean of the sampling distribution of $\hat{p}$ is p; that is, $\hat{p}$ is an unbiased estimator of p.
2. The standard deviation of the sampling distribution of $\hat{p}$ is $\sqrt{pq/n}$; that is, $\sigma_p = \sqrt{pq/n}$, where $q = 1 - p$.
3. For large samples, the sampling distribution of $\hat{p}$ is approximately normal. A sample size is considered large if both $n\hat{p} \geq 15$ and $n\hat{q} \geq 15$.

Procedure

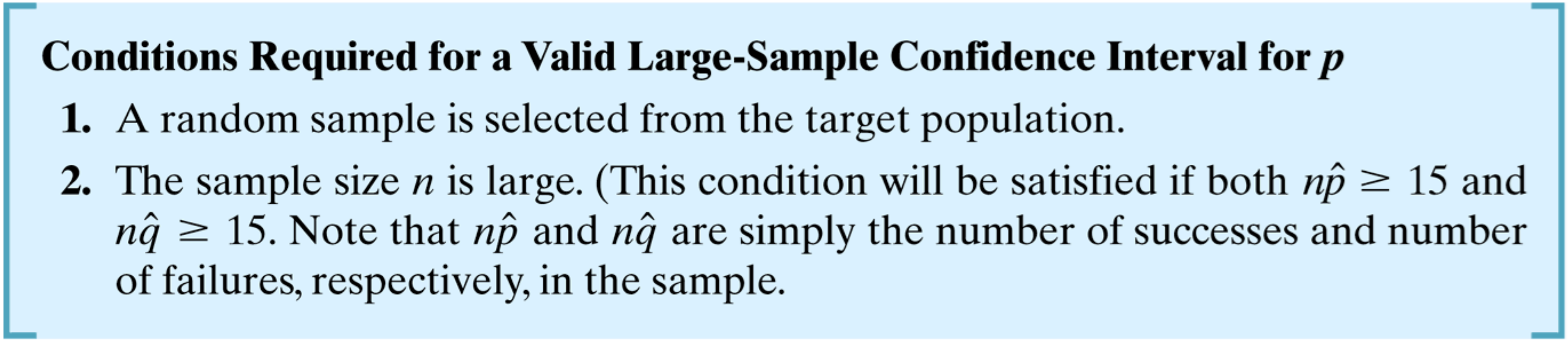


Large-Sample Confidence Interval for *p*

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{pq}{n}} \approx \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p} = \frac{x}{n}$ and $\hat{q} = 1 - \hat{p}$

Note: When n is large, $\hat{p}$ can approximate the value of p in the formula for $\sigma_{\hat{p}}$.



Conditions Required for a Valid Large-Sample Confidence Interval for *p*

1. A random sample is selected from the target population.
2. The sample size n is large. (This condition will be satisfied if both $n\hat{p} \geq 15$ and $n\hat{q} \geq 15$. Note that $n\hat{p}$ and $n\hat{q}$ are simply the number of successes and number of failures, respectively, in the sample.

Table 7.6

- The approximation for p does not have to be especially accurate because the value of $\sqrt{pq}$ needed for the confidence interval is relatively insensitive to changes in p. Therefore, we can use $\hat{p}$ to approximate p.

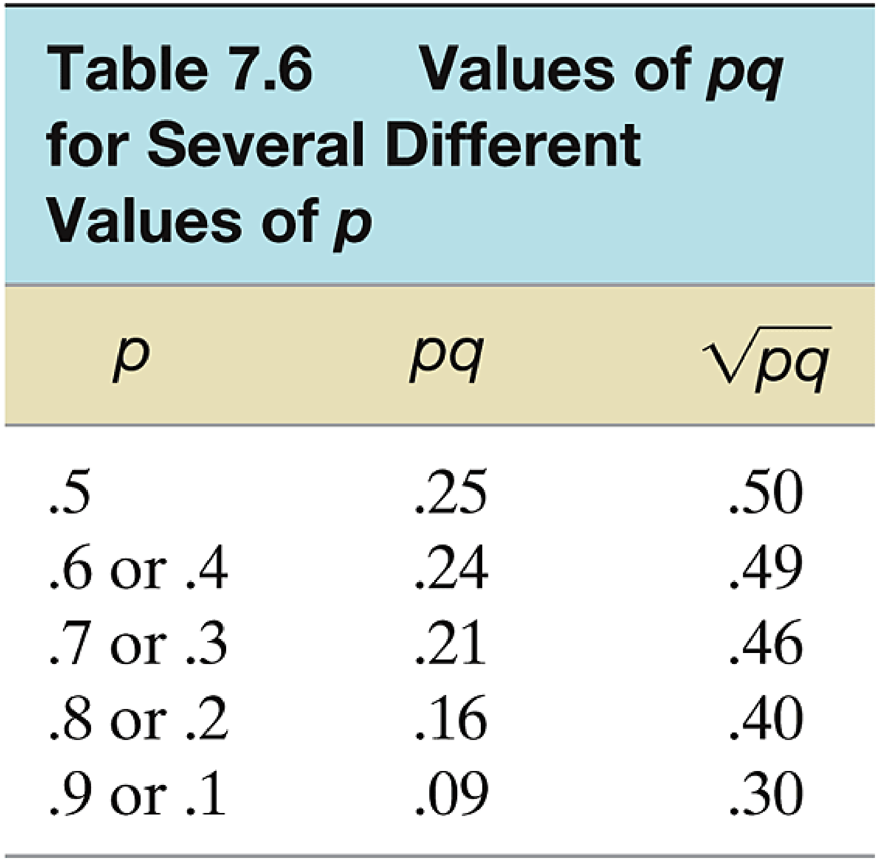


Table 7.6 Values of *pq* for Several Different Values of *p*

p	pq	$\sqrt{pq}$
.5	.25	.50
.6 or .4	.24	.49
.7 or .3	.21	.46
.8 or .2	.16	.40
.9 or .1	.09	.30

Definition

- Thus, if 637 of 1,000 U.S. citizens say they trust the president, a 95% confidence interval for the proportion of all U.S. citizens who trust the president is

$$\hat{p} \pm z_{\alpha/2}\sigma_{\hat{p}} = .637 \pm 1.96\sqrt{\frac{pq}{1{,}000}}$$

$$\begin{aligned}\hat{p} \pm 1.96\sqrt{pq/1{,}000} &\approx \hat{p} \pm 1.96\sqrt{\hat{p}\hat{q}/1{,}000} \\ &= .637 \pm 1.96\sqrt{(.637)(.363)/1{,}000} = .637 \pm .030 \\ &= (.607, .667)\end{aligned}$$

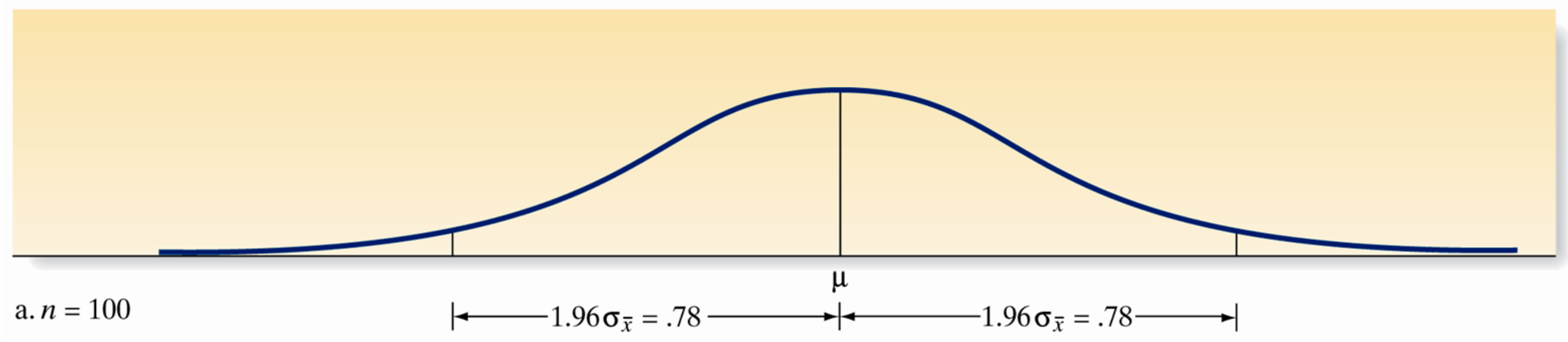
7.5

Determining the Sample Size

Estimating a Population Mean

- Consider Example 7.1, in which we estimated the mean length of stay for patients in a large hospital. A sample of 100 patients' records produced the 95% confidence interval $\boldsymbol{\bar{x} \pm (1.96)\sigma_{\bar{x}} = 4.5 \pm .78}$
- Suppose we want to estimate $\boldsymbol{\mu}$ to within $\mathbf{.25}$ day with 95% confidence.
- That is, we want to narrow the width of the confidence interval from $\mathbf{1.56}$ days to $\mathbf{0.50}$ day.
- **How much will the sample size have to be increased to accomplish this?**

Figure 7.16 Relationship between sample size and width of confidence interval: hospital-stay example



$$1.96\sigma_{\bar{x}} = .25 \text{ or, equivalently, } 1.96\left(\frac{\sigma}{\sqrt{n}}\right) = .25$$

$$1.96\left(\frac{\sigma}{\sqrt{n}}\right) = 1.96\left(\frac{4}{\sqrt{n}}\right) = .25$$

$$\sqrt{n} = \frac{1.96(4)}{.25} = 31.36$$

$$n = (31.36)^2 = 983.45$$

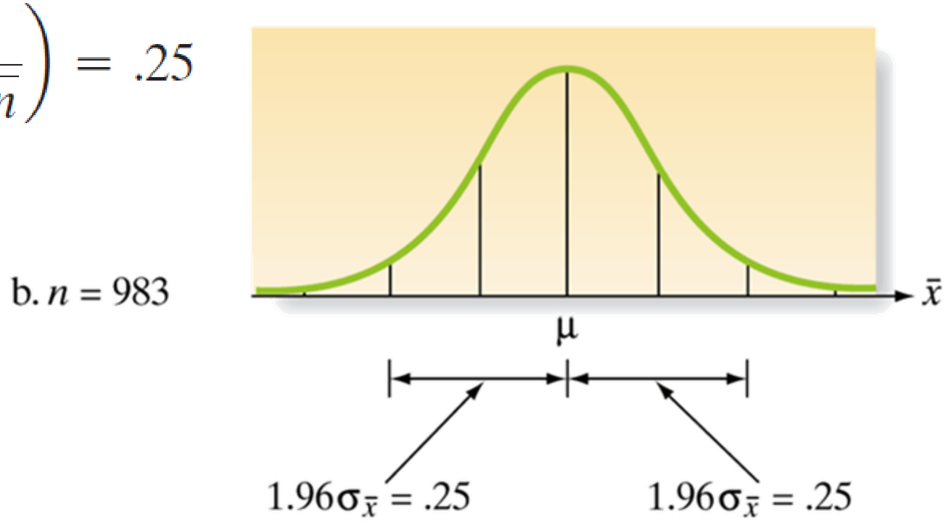
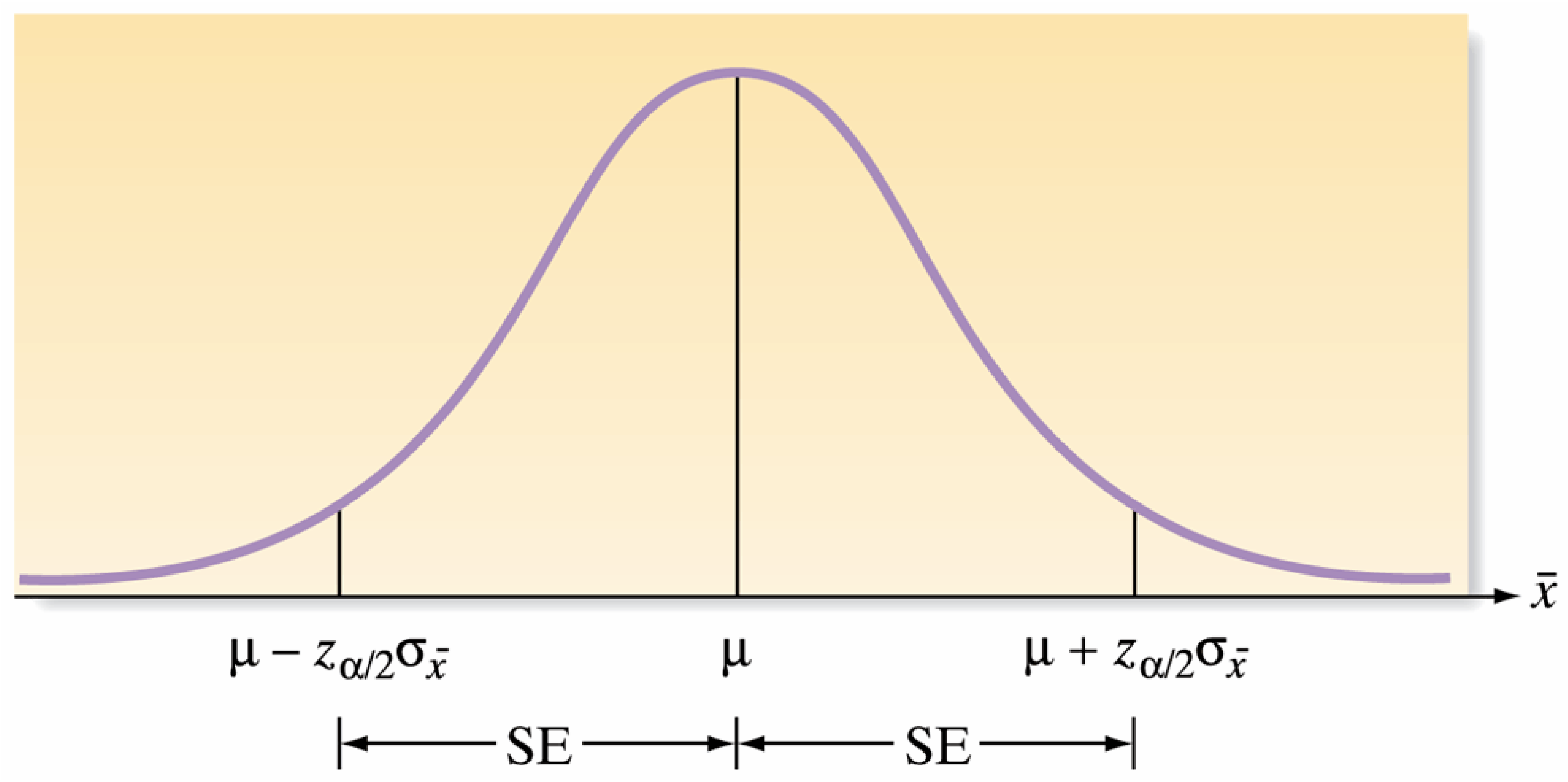
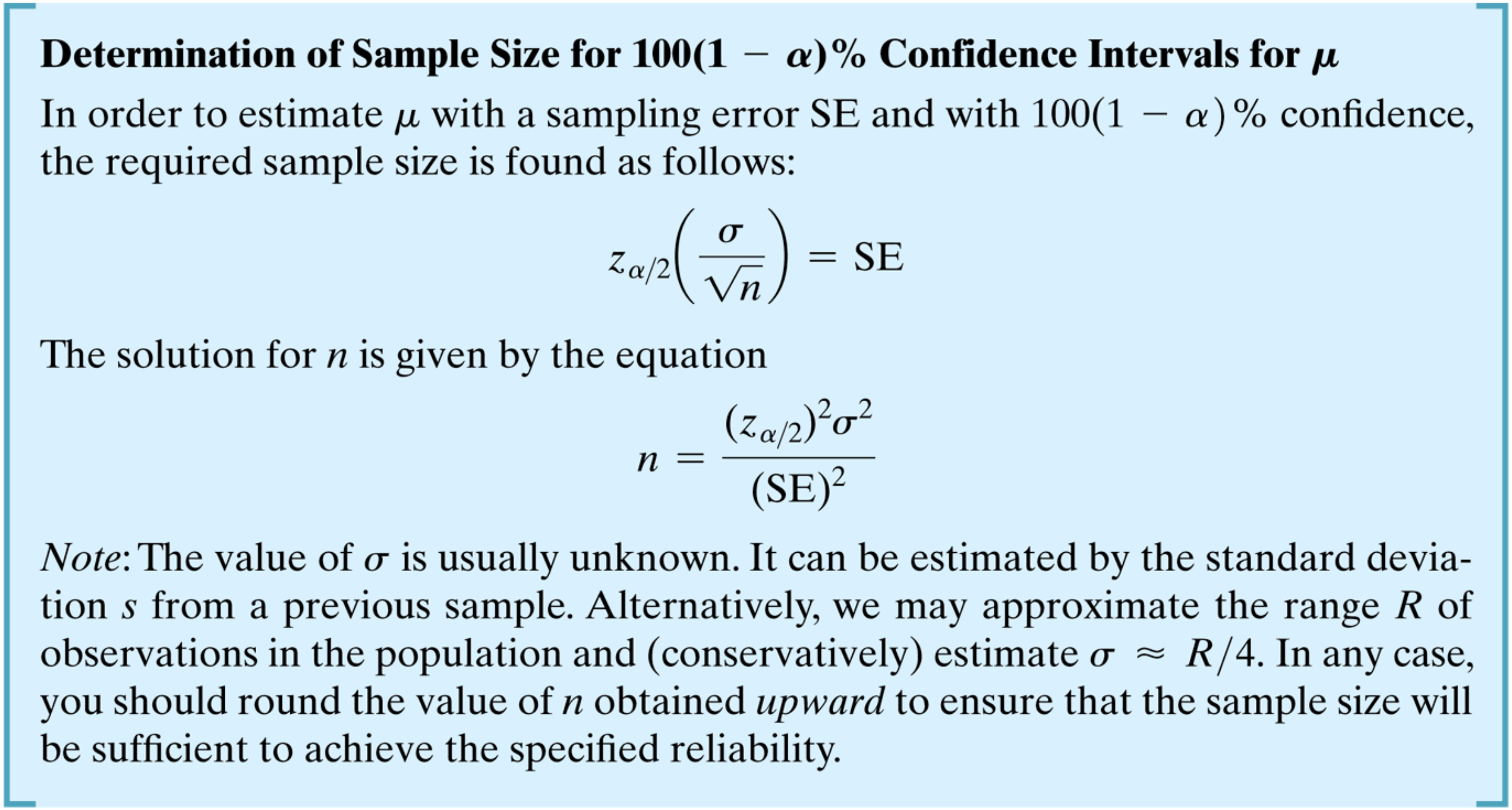


Figure 7.17 Specifying the sampling error SE as the half-width of a confidence interval


$\bar{x}$
$\mu - z_{\alpha/2}\sigma_{\bar{x}}$
μ
$\mu + z_{\alpha/2}\sigma_{\bar{x}}$
SE
SE

Procedure



Determination of Sample Size for 100(1 − α)% Confidence Intervals for μ

In order to estimate μ with a sampling error SE and with 100(1 − α)% confidence, the required sample size is found as follows:

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = \text{SE}$$

The solution for n is given by the equation

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{(\text{SE})^2}$$

Note: The value of σ is usually unknown. It can be estimated by the standard deviation s from a previous sample. Alternatively, we may approximate the range R of observations in the population and (conservatively) estimate $\sigma \approx R/4$. In any case, you should round the value of n obtained *upward* to ensure that the sample size will be sufficient to achieve the specified reliability.

Problem

Suppose the manufacturer of official NFL footballs uses a machine to inflate the new balls to a pressure of 13.5 pounds. When the machine is properly calibrated, the mean inflation pressure is 13.5 pounds, but uncontrollable factors cause the pressures of individual footballs to vary randomly from about 13.3 to 13.7 pounds. For quality control purposes, the manufacturer wishes to estimate the mean inflation pressure to within .025 pound of its true value with a 99% confidence interval. What sample size should be specified for the experiment?

Solution

- We desire a 99% confidence interval that estimates $\boldsymbol{\mu}$ with a sampling error of SE = .025 pound. For a 99% confidence interval, we have $z_{\alpha/2} = z_{.005} = 2.575$. To estimate $\boldsymbol{\sigma}$, we note that the range of observations is $R = 13.7 - 13.3 = .4$ and we use $\sigma \approx \frac{R}{4} = .1$
- Next, we employ the formula derived in the box to find the sample size n:

$$n = \frac{(z_{\alpha/2})^2\sigma^2}{(\mathrm{SE})^2} \approx \frac{(2.575)^2(.1)^2}{(.025)^2} = 106.09$$

Estimating a Population Proportion

- In Example 7.6, a pollster used a sample of 1,000 U.S. citizens to calculate a 95% confidence interval for the proportion who trust the president, obtaining the interval $.637 \pm .03$. Suppose the pollster wishes to estimate more precisely the proportion who trust the president, say, to within $.015$ with a 95% confidence interval.
- The pollster wants a confidence interval for $\boldsymbol{p}$ with a sampling error $\boldsymbol{SE} = \mathbf{.015}$.

Figure 7.18 Specifying the sampling error SE of a confidence interval for a population proportion *p*

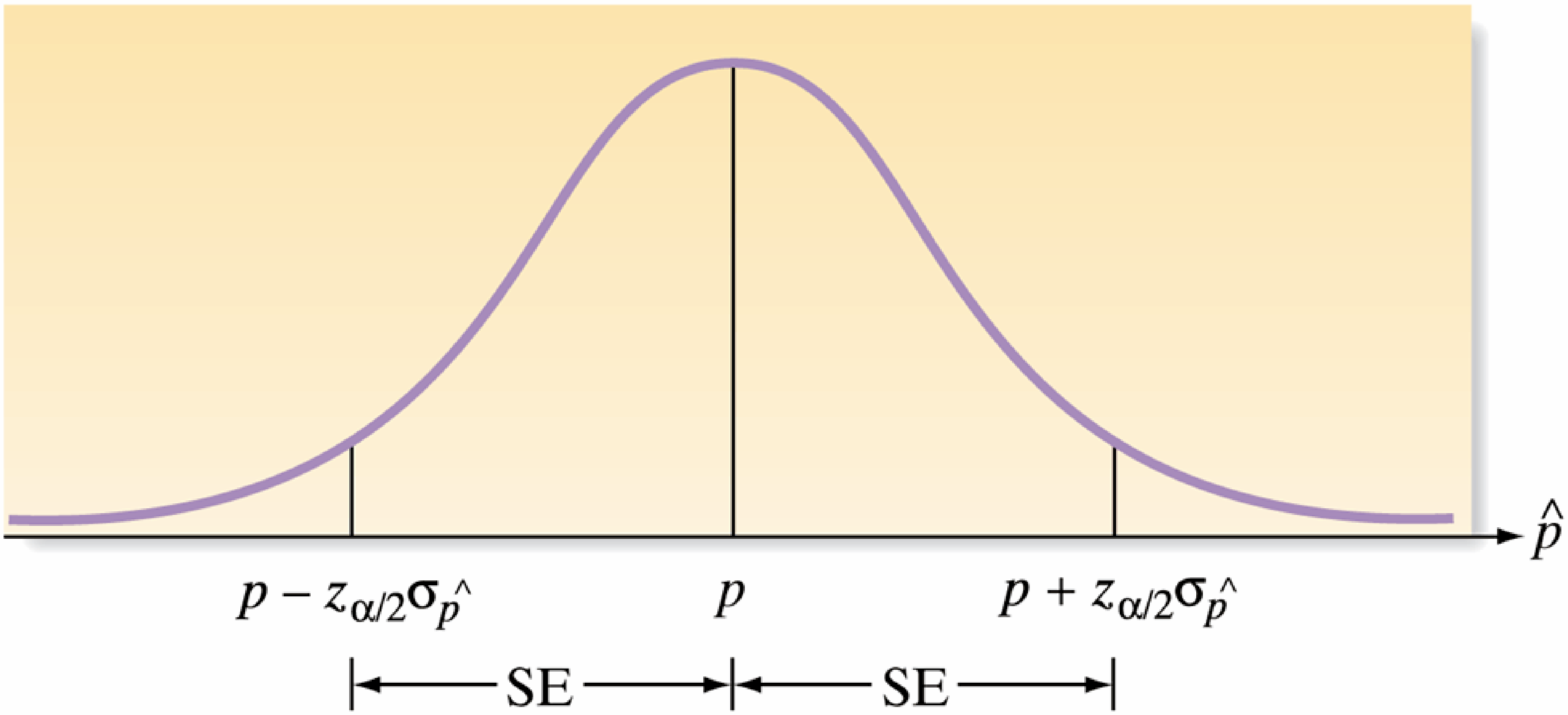

$\hat{p}$
$p - z_{\alpha/2}\sigma_{\hat{p}}$
p
$p + z_{\alpha/2}\sigma_{\hat{p}}$
SE
SE

Figure 7.18 Specifying the sampling error SE of a confidence interval for a population proportion *p*

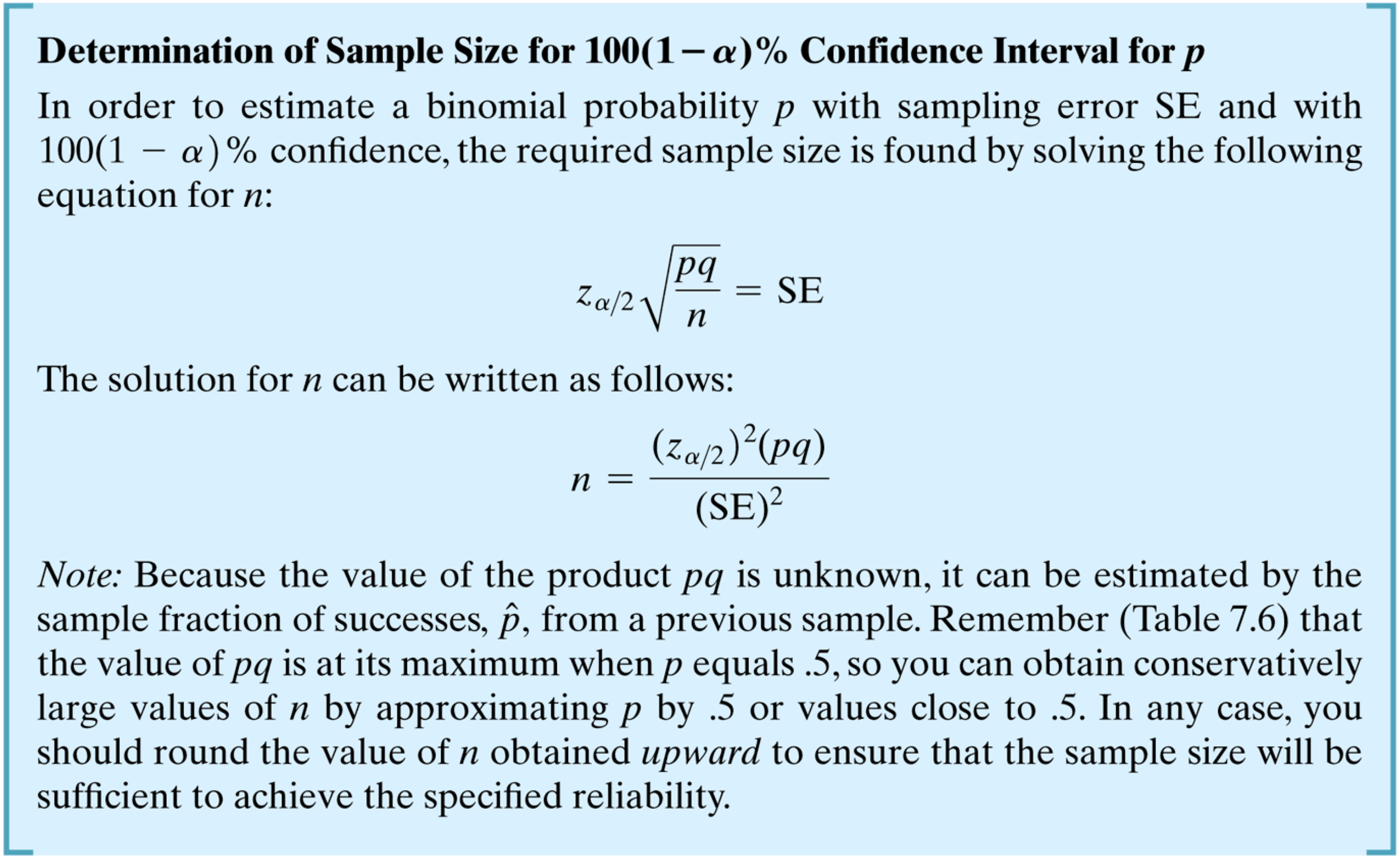
- The sample size required to generate such an interval is found by solving the following equation for n:

$$z_{\alpha/2}\sigma_{\hat{p}} = \text{SE} \qquad \text{or} \qquad z_{\alpha/2}\sqrt{\frac{pq}{n}} = .015$$

$$1.96\sqrt{\frac{(.60)(.40)}{n}} = .015$$

$$n = \frac{(1.96)^2(.60)(.40)}{(.015)^2} = 4{,}097.7 \approx 4{,}098$$

Procedure



Determination of Sample Size for 100(1 − *α*)% Confidence Interval for *p*

In order to estimate a binomial probability p with sampling error SE and with $100(1 - \alpha)\%$ confidence, the required sample size is found by solving the following equation for n:

$$z_{\alpha/2}\sqrt{\frac{pq}{n}} = \text{SE}$$

The solution for n can be written as follows:

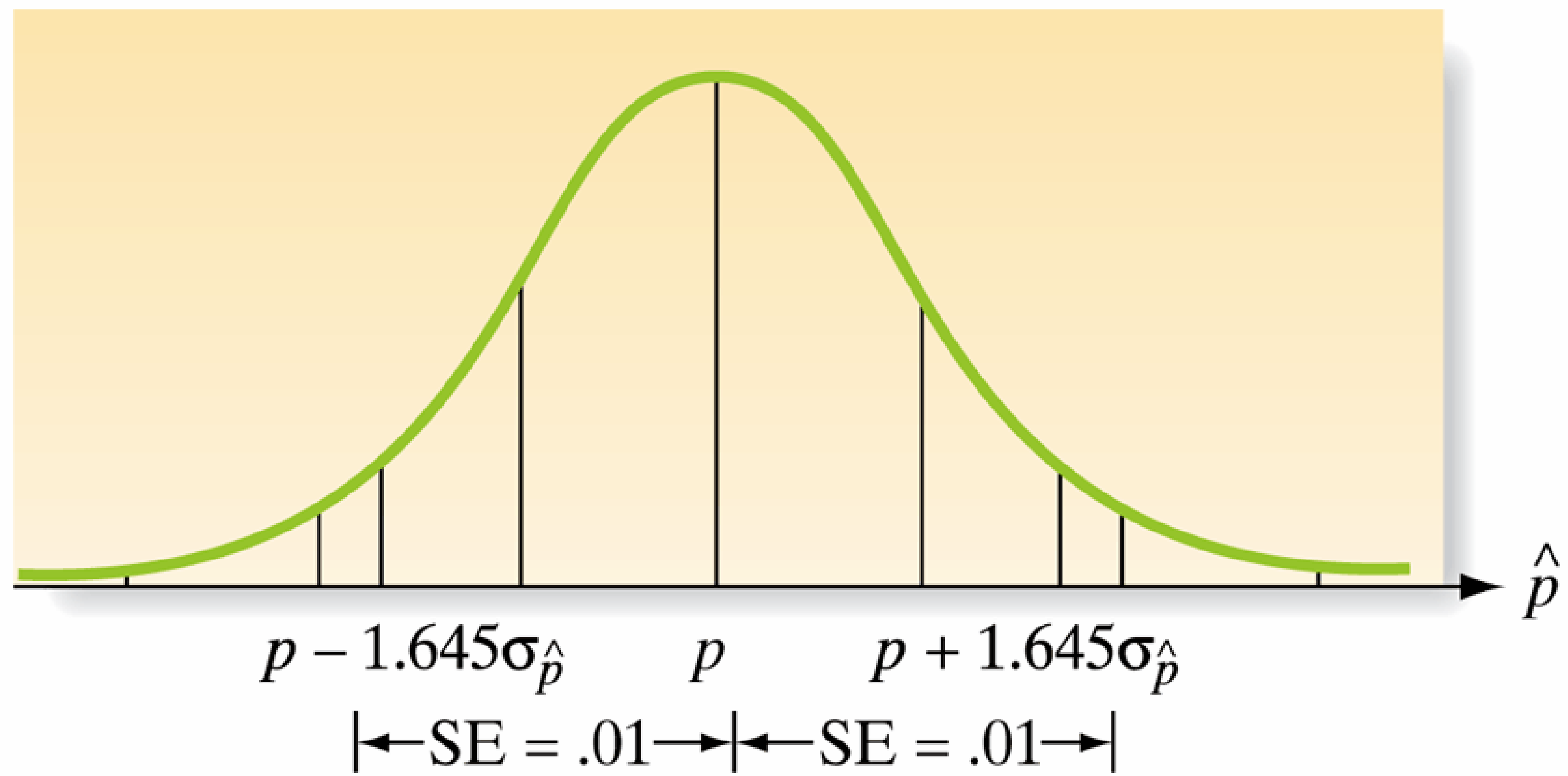
$$n = \frac{(z_{\alpha/2})^2(pq)}{(\text{SE})^2}$$

Note: Because the value of the product pq is unknown, it can be estimated by the sample fraction of successes, $\hat{p}$, from a previous sample. Remember (Table 7.6) that the value of pq is at its maximum when p equals .5, so you can obtain conservatively large values of n by approximating p by .5 or values close to .5. In any case, you should round the value of n obtained *upward* to ensure that the sample size will be sufficient to achieve the specified reliability.

Problem

A cellular telephone manufacturer that entered the post-regulation market quickly has an initial problem with excessive customer complaints and consequent returns of cell phones for repair or replacement. The manufacturer wants to estimate the magnitude of the problem in order to design a quality control program. How many cellular telephones should be sampled and checked in order to estimate the fraction defective, $\boldsymbol{p}$, to within $\mathbf{.01}$ with $\mathbf{90\%}$ confidence?

Figure 7.19 Specified reliability for estimate of fraction defective in Example 7.10



Solution

- The equation for the sample size n requires an estimate of the product $\boldsymbol{pq}$.
- We could most conservatively estimate $\boldsymbol{pq} =.\mathbf{25}$ (i.e., use $\boldsymbol{p} = .\mathbf{5}$), but this estimate may be too conservative.
- By contrast, a value of $.\mathbf{1}$, corresponding to 10% defective, will probably be conservatively large for this application. The solution is therefore

$$n = \frac{(z_{\alpha/2})^2(pq)}{(SE)^2} = \frac{(1.645)^2(.1)(.9)}{(.01)^2} = 2{,}435.4 \approx 2{,}436$$