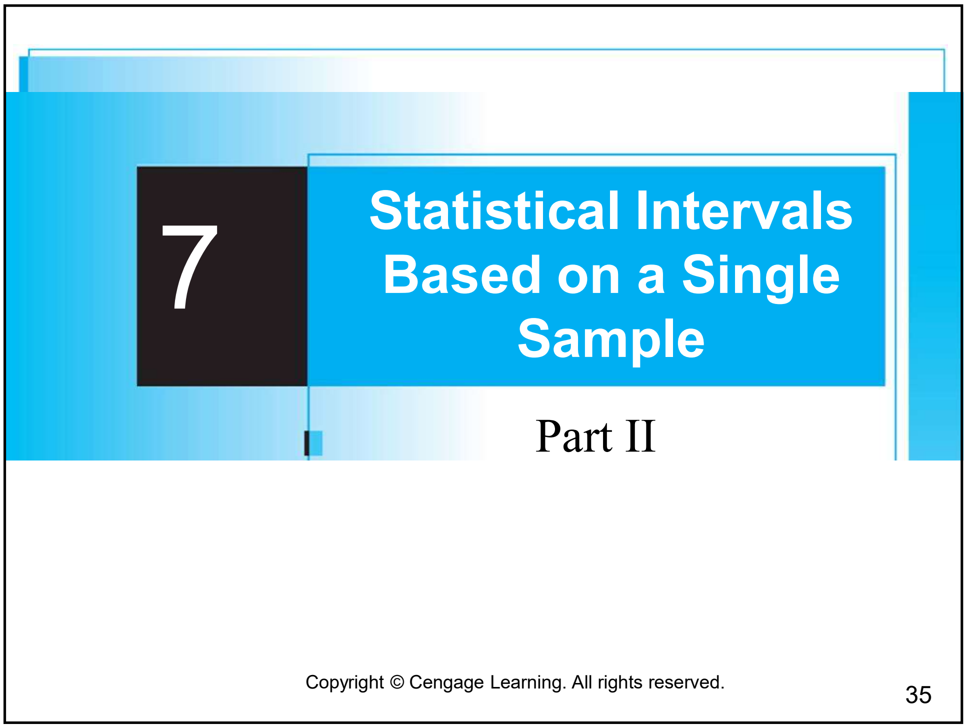
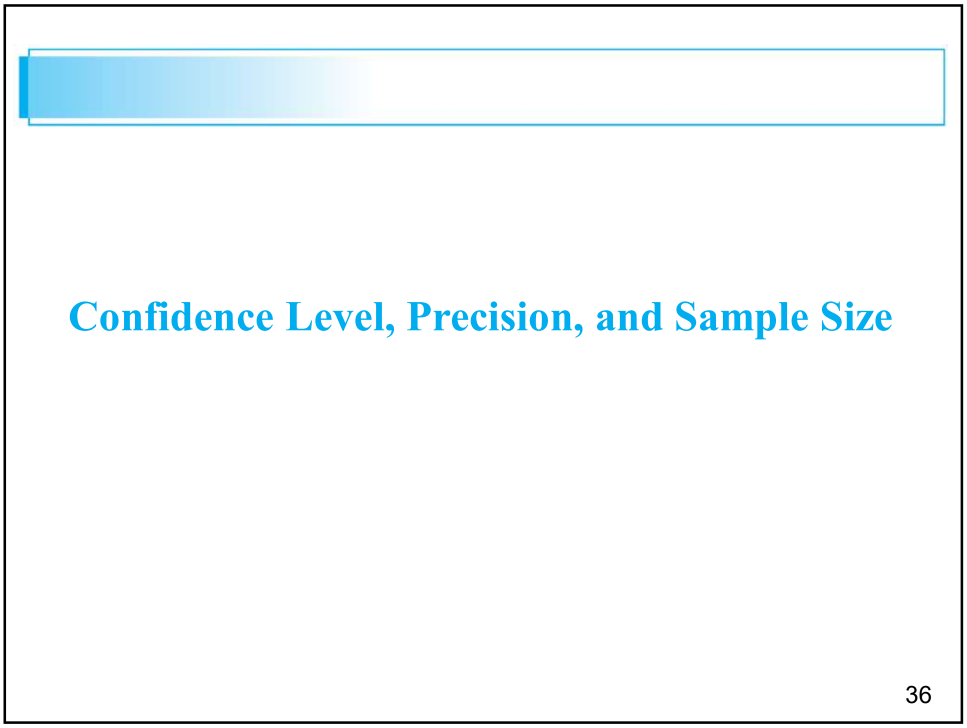


7 Statistical Intervals Based on a Single Sample

Part II



Confidence Level, Precision, and Sample Size

Confidence Level, Precision, and Sample Size

- Why settle for a confidence level of 95% when a level of 99% is achievable? Because the price paid for the higher confidence level is a wider interval.
- Since the 95% interval extends $1.96 \cdot \sigma/\sqrt{n}$ to each side of $\bar{x}$, the width of the interval is $2(1.96) \cdot \sigma/\sqrt{n} = 3.92 \cdot \sigma/\sqrt{n}$.
- Similarly, the width of the 99% interval is $2(2.58) \cdot \sigma/\sqrt{n} = 5.16 \cdot \sigma/\sqrt{n}$.
- That is, we have more confidence in the 99% interval precisely because it is wider.
- The higher the desired degree of confidence, the wider the resulting interval will be.

Confidence Level, Precision, and Sample Size

- If we think of the width of the interval as specifying its precision or accuracy, then the confidence level (or reliability) of the interval is inversely related to its precision.
- A highly reliable interval estimate may be imprecise in that the endpoints of the interval may be far apart, whereas a precise interval may entail relatively low reliability.
- Thus it cannot be said unequivocally that a 99% interval is to be preferred to a 95% interval; the gain in reliability entails a loss in precision.

Confidence Level, Precision, and Sample Size

- An appealing strategy is to specify both the desired confidence level and interval width and then determine the necessary **sample size**.

Example 7.4

- Extensive monitoring of a computer time-sharing system has suggested that response time to a particular editing command is normally distributed with standard deviation 25 millisec.
- A new operating system has been installed, and we wish to estimate the true average response time μ for the new environment.
- Assuming that response times are still normally distributed with $\sigma = 25$, **what sample size** is necessary to ensure that the resulting 95% CI has a **width** of (at most) 10?

Example 7.4 (cont.)

cont'd

- The sample size n must satisfy

$$10 = 2 \cdot (1.96)(\frac{25}{\sqrt{n}})$$

- Rearranging this equation gives

$$\sqrt{n} = 2 \cdot \frac{(1.96)(25)}{10} = 9.80$$

- So

$$n = (9.80)^2 = 96.04$$

- Since n must be an integer, a sample size of 97 is required.

Confidence Level, Precision, and Sample Size

❑ A general formula for the sample size n necessary to ensure an interval width w is obtained from equating w to $2 \cdot z_{\alpha/2} \cdot \sigma/\sqrt{n}$ and solving for n.

❑ The sample size necessary for the CI (7.5) to have a width w is

$$n = \left(z_{\alpha/2} \cdot \frac{\sigma}{w}\right)^2$$

❑ The smaller the desired width w, the larger n must be.

❑ In addition, n is an increasing function of σ (more population variability necessitates a larger sample size) and of the confidence level $100(1-\alpha)$ (as α decreases, $z_{\alpha/2}$ increases).

Confidence Level, Precision, and Sample Size

- The half-width $1.96\,\sigma/\sqrt{n})$ of the 95% CI is sometimes called the **bound on the error of estimation** associated with a 95% confidence level.

- That is, with 95% confidence, the point estimate $\bar{x}$ will be no farther than this from μ.

- Before obtaining data, an investigator may wish to determine a sample size for which a particular value of the bound is achieved.

43

Confidence Level, Precision, and Sample Size

- For example, with μ representing the average fuel efficiency (mpg) for all cars of a certain type, the objective of an investigation may be to estimate μ to within 1 mpg with 95% confidence.

- More generally, if we wish to estimate μ to within an amount B (the specified bound on the error of estimation) with $100(1-\alpha)\,\%$ confidence, the necessary sample size results from replacing $2/w$ by $1/B$ in the formula in the preceding box.

44

Deriving a Confidence Interval

Deriving a Confidence Interval

- Let $X_1, X_2, \dots, X_n$ denote the sample on which the CI for a parameter θ is to be based.
- Suppose a random variable satisfying the following two properties can be found:

1) The variable depends functionally on both $X_1, \dots, X_n$ and θ.
2) The probability distribution of the variable does not depend on θ or on any other unknown parameters.

Deriving a Confidence Interval

- Let $h(X_1, X_2, \ldots, X_n;\ \theta)$ denote this random variable.
- For example, if the population distribution is normal with known σ and $\theta = \mu$, the variable

$$h(X_1, \ldots, X_n; \mu) = (\bar{X} - \mu)/(\sigma/\sqrt{n})$$

- satisfies both properties; it clearly depends functionally on μ, yet has the standard normal probability distribution, which does not depend on μ.
- In general, the form of the h function is usually suggested by examining the distribution of an appropriate estimator $\hat{\theta}$

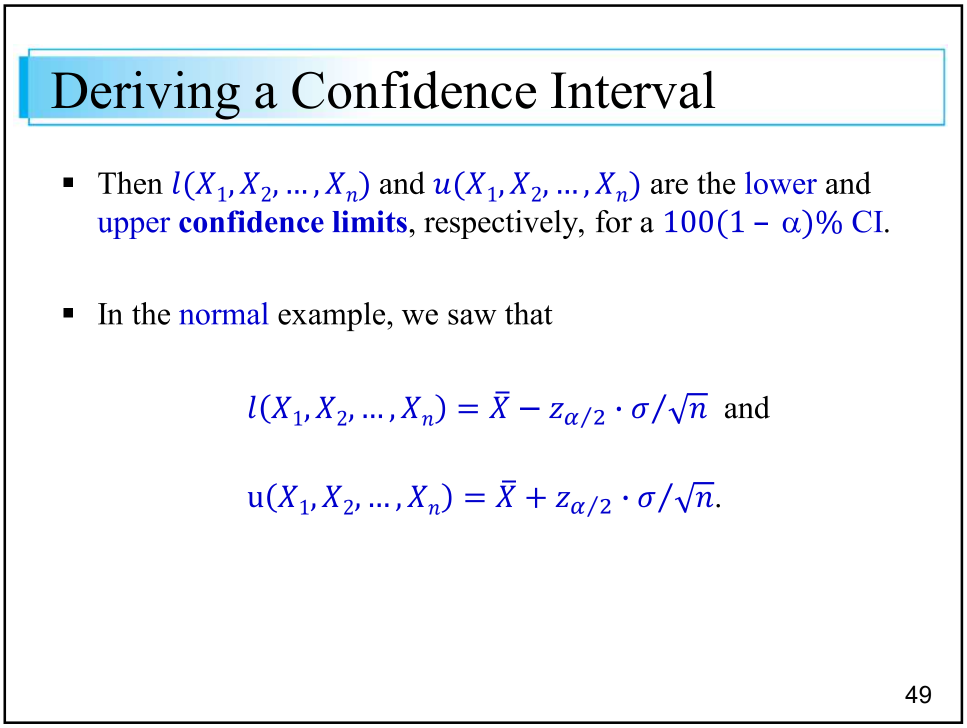
Deriving a Confidence Interval

- For any α between 0 and 1, constants a and b can be found to satisfy

$$\mathrm{P}(\mathrm{a} < h(X_1, \ldots, X_n; \theta) < b) = 1 - \alpha \ldots\ldots\ldots\ldots\ldots (7.6)$$

- Because of the second property, a and b do not depend on θ.
- In the normal example, $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$.
- Now suppose that the inequalities in (7.6) can be manipulated to isolate θ, giving the equivalent probability statement

$$P(l(X_1, X_2, \ldots, X_n) < \theta < u(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$



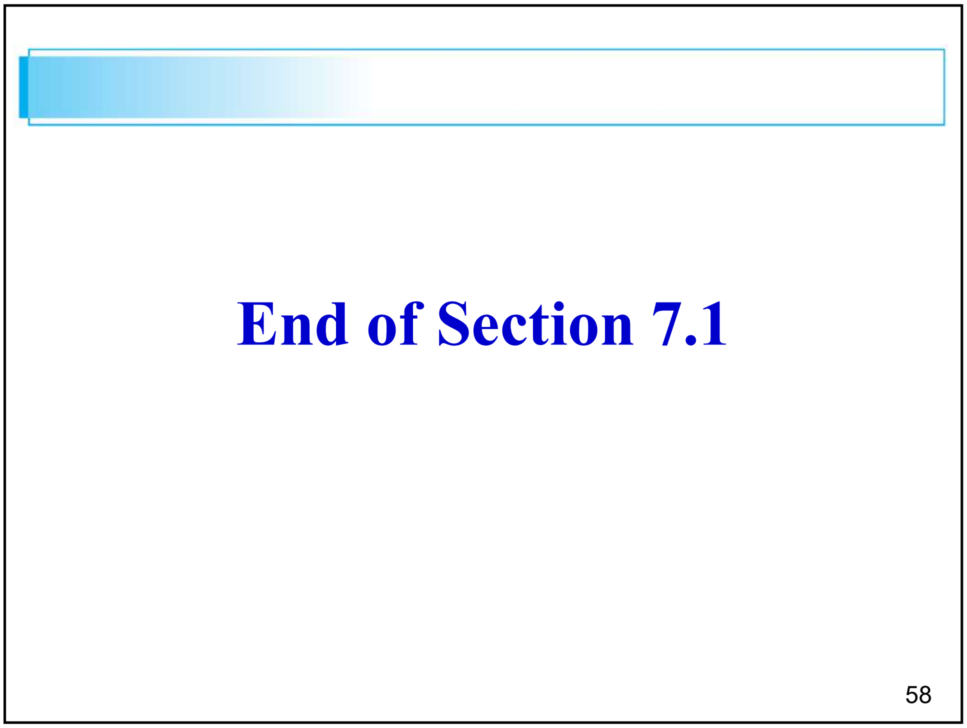
Deriving a Confidence Interval

- Then $l(X_1, X_2, \ldots, X_n)$ and $u(X_1, X_2, \ldots, X_n)$ are the lower and upper **confidence limits**, respectively, for a $100(1 - \alpha)\%$ CI.

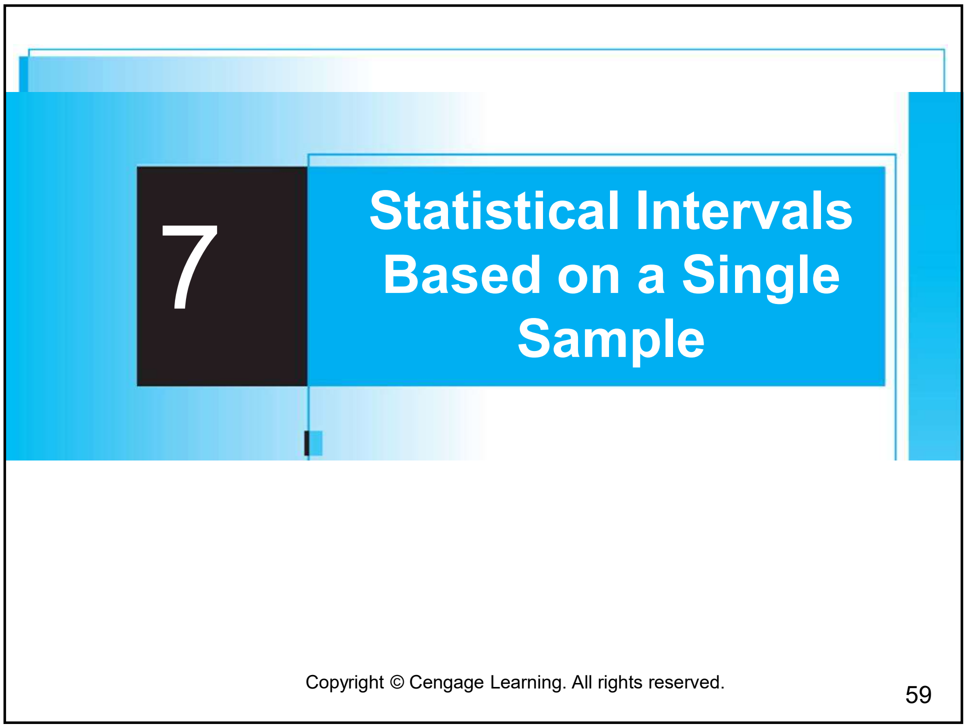
- In the normal example, we saw that

$$l(X_1, X_2, \ldots, X_n) = \bar{X} - z_{\alpha/2} \cdot \sigma/\sqrt{n} \text{ and}$$

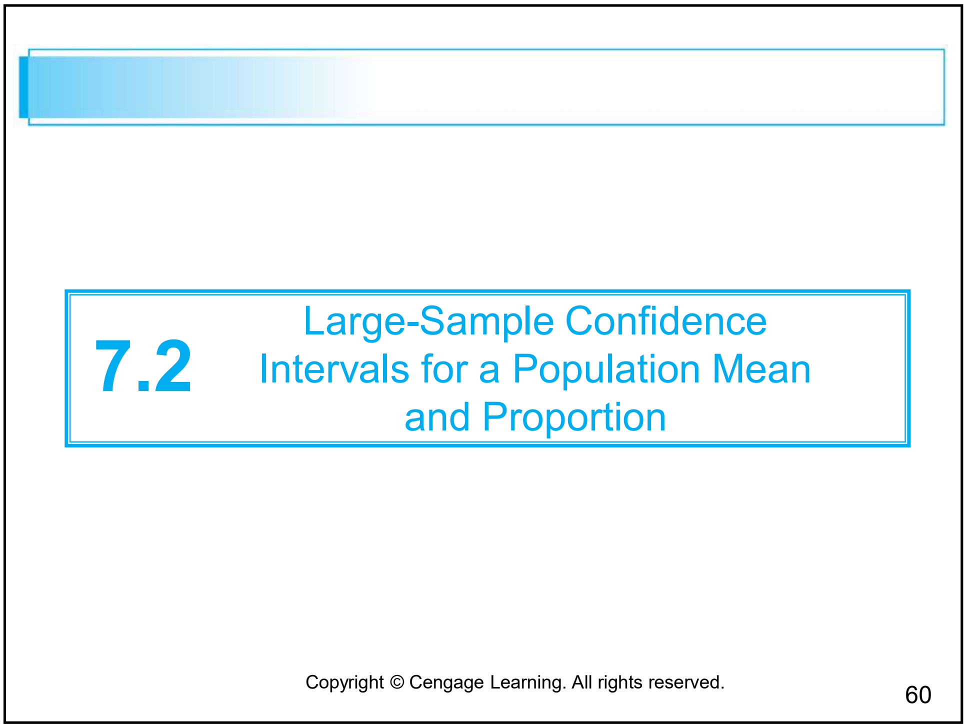
$$\mathrm{u}(X_1, X_2, \ldots, X_n) = \bar{X} + z_{\alpha/2} \cdot \sigma/\sqrt{n}.$$



End of Section 7.1



7 Statistical Intervals Based on a Single Sample



7.2 Large-Sample Confidence Intervals for a Population Mean and Proportion

Large-Sample Confidence Intervals for a Population Mean and Proportion

- Earlier we have come across the CI for μ which assumed that the population distribution is normal with the value of σ known.
- We now present a large-sample CI whose validity does not require these assumptions.
- After showing how the argument leading to this interval generalizes to yield other large-sample intervals, we focus on an interval for a population proportion p.

61

A Large-Sample Interval for μ

62

A Large-Sample Interval for μ

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a population having a mean μ and standard deviation σ.
- Provided that n is large, the Central Limit Theorem (CLT) implies that $\bar{X}$ has approximately a normal distribution whatever the nature of the population distribution.
- It then follows that

$$Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$$

has approximately a standard normal distribution, so that

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

A Large-Sample Interval for μ

- We have known that an argument parallel yields $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ as a large-sample CI for μ with a confidence level of *approximately* $100(1 - \quad)\%$.
- That is, when n is large, the CI for μ given previously remains valid whatever the population distribution, provided that the qualifier “approximately” is inserted in front of the confidence level.
- A practical difficulty with this development is that computation of the CI requires the value of σ, which will rarely be known.
- Consider the standardized variable

$$(\bar{X} - \mu)/(S/\sqrt{n}),$$

in which the sample standard deviation S has replaced σ.

A Large-Sample Interval for μ

Proposition

If n is sufficiently large, the standardized variable

$$Z = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has approximately a standard normal distribution.

This implies that

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \qquad \textbf{(7.8)}$$

is a **large-sample confidence interval for** μ with confidence level approximately 100(1 – ▯)%.

This formula is valid regardless of the shape of the population distribution.

A Large-Sample Interval for μ $\overline{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

In words, the CI (7.8) is

point estimate of μ ± (z critical value) (estimated standard error of the mean).

Generally speaking, $n > 40$ will be sufficient to justify the use of this interval.

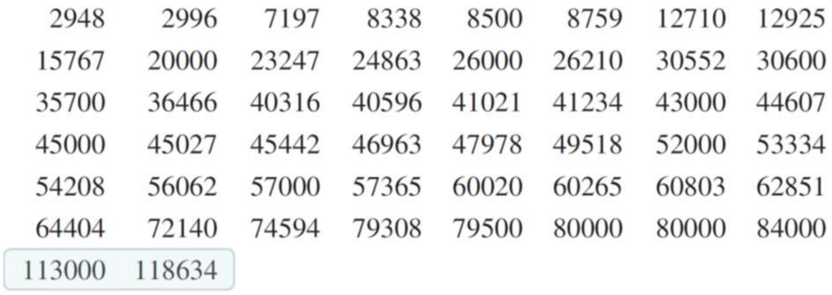
Example 6

- Haven’t you always wanted to own a Porsche?
- The author thought maybe he could afford a Boxster, the cheapest model.
- So he went to www.cars.com on Nov. 18, 2009, and found a total of 1113 such cars listed.
- Asking prices ranged from $3499 to $130,000 (the latter price was one of only two exceeding $70,000).
- The prices depressed him, so he focused instead on odometer readings (miles).

Example 6

cont’d

Here are reported readings for a sample of 50 of these Boxsters:

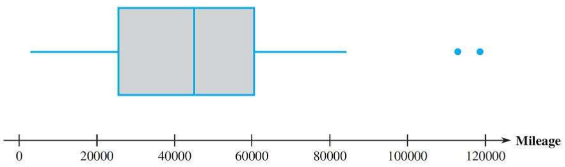


2948	2996	7197	8338	8500	8759	12710	12925
15767	20000	23247	24863	26000	26210	30552	30600
35700	36466	40316	40596	41021	41234	43000	44607
45000	45027	45442	46963	47978	49518	52000	53334
54208	56062	57000	57365	60020	60265	60803	62851
64404	72140	74594	79308	79500	80000	80000	84000
113000	118634						

Example 6

cont'd

- A boxplot of the data (Figure 7.5) shows that, except for the two outliers at the upper end, the distribution of values is reasonably symmetric (in fact, a normal probability plot exhibits a reasonably linear pattern, though the points corresponding to the two smallest and two largest observations are somewhat removed from a line fit through the remaining points).



A boxplot of the odometer reading data from Example 6

Figure 7.5

Example 6

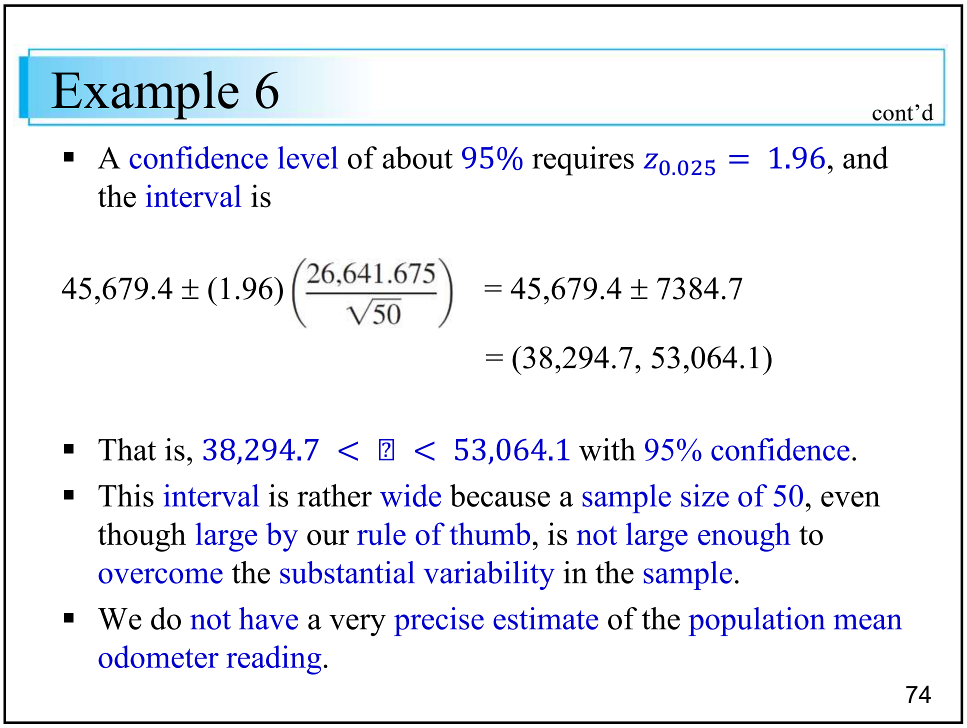
cont'd

- Summary quantities include

  $n = 50$, $\bar{x} = 45{,}679.4$,
  $\tilde{x} = 45{,}013.5$, $s = 26{,}641.675$.

- The mean and median are reasonably close (if the two largest values were each reduced by 30,000, the mean would fall to 44,479.4, while the median would be unaffected).

- The boxplot and the magnitudes of s relative to the mean and median both indicate a substantial amount of variability.



Example 6

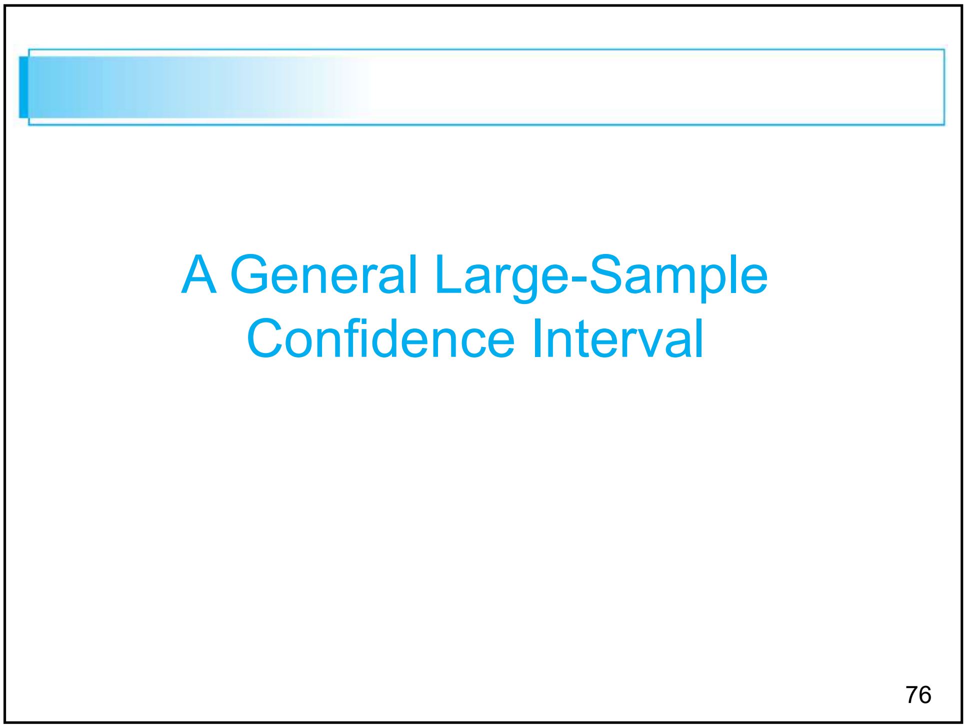
cont'd

- A confidence level of about 95% requires $z_{0.025} = 1.96$, and the interval is

$$45{,}679.4 \pm (1.96)\left(\frac{26{,}641.675}{\sqrt{50}}\right) = 45{,}679.4 \pm 7384.7$$

$$= (38{,}294.7,\ 53{,}064.1)$$

- That is, 38,294.7 < ▯ < 53,064.1 with 95% confidence.
- This interval is rather wide because a sample size of 50, even though large by our rule of thumb, is not large enough to overcome the substantial variability in the sample.
- We do not have a very precise estimate of the population mean odometer reading.



A General Large-Sample Confidence Interval

A General Large-Sample Confidence Interval

The large-sample intervals $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$ and $\bar{x} \pm z_{\alpha/2} \cdot S/\sqrt{n}$ are special cases of a general large-sample CI for a parameter θ.

Suppose that $\hat{\theta}$ is an estimator satisfying the following properties:

(1) It has approximately a normal distribution;

(2) it is (at least approximately) unbiased; and

(3) an expression for $\sigma_{\hat{\theta}}$, the standard deviation of $\hat{\theta}$, is available.

A General Large-Sample Confidence Interval

- For example, in the case $\theta = \mu$, $\hat{\mu} = \bar{X}$ is an unbiased estimator whose distribution is approximately normal when n is large and $\sigma_{\hat{\mu}} = \sigma_{\bar{X}} = \sigma/\sqrt{n}$.
- Standardizing $\hat{\theta}$ yields the rv $Z = (\hat{\theta} - \theta)/\ \sigma_{\hat{\theta}}$, which has approximately a standard normal distribution.
- This justifies the probability statement

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1 - \alpha \qquad \textbf{(7.9)}$$

- Suppose first that $\sigma_{\hat{\theta}}$ does not involve any unknown parameters (e.g., known σ in the case $\theta = \mu$).
- Then replacing each < in (7.9) by = results in $\theta = \hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$, so the lower and upper confidence limits are $\hat{\theta} - z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$ and $\hat{\theta} + z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$, respectively.

A General Large-Sample Confidence Interval

- Now suppose that $\sigma_{\hat{\theta}}$ does not involve θ but does involve at least one other unknown parameter.
- Let $S_{\hat{\theta}}$ be the estimate of $\sigma_{\hat{\theta}}$ obtained by using estimates in place of the unknown parameters (e.g., $S/\sqrt{n}$ estimates $\sigma/\sqrt{n}$).
- Under general conditions (essentially that $S_{\hat{\theta}}$ be close to $\sigma_{\hat{\theta}}$ for most samples), a valid CI is $\hat{\theta} \pm z_{\alpha/2} \cdot S_{\hat{\theta}}$.
- The large-sample interval $\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$ is an example.

A General Large-Sample Confidence Interval

- Finally, suppose that $\sigma_{\hat{\theta}}$ does involve the unknown θ.
- This is the case, for example, when $\theta = p$, a population proportion.
- Then $(\hat{\theta} - \theta)/\ \sigma_{\hat{\theta}} = z_{\alpha/2}$ can be difficult to solve.
- An approximate solution can often be obtained by replacing θ in $\sigma_{\hat{\theta}}$ by its estimate $\hat{\theta}$.
- This results in an estimated standard deviation $S_{\hat{\theta}}$, and the corresponding interval is again $\hat{\theta} \pm z_{\alpha/2} \cdot S_{\hat{\theta}}$.
- In words, this CI is a
- point estimate of θ ± (z critical value) (estimated standard error of the estimator)