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*.

A Large-Sample Interval for μ

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 $\overline{X}$ has approximately a normal distribution whatever the nature of the population distribution.

It then follows that $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ has approximately a standard normal distribution, so that

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

A Large-Sample Interval for μ

We have know 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 – α)%. 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

Previously, there was randomness only in the numerator of *Z* by virtue of $\overline{X}$. In the new standardized variable, both $\overline{X}$ *and S* vary in value from one sample to another. So it might seem that the distribution of the new variable should be more spread out than the *z* curve to reflect the extra variation in the denominator. This is indeed true when *n* is small.

However, for large *n* the subsititution of *S* for σ adds little extra variability, so this variable also has approximately a standard normal distribution. Manipulation of the variable in a probability statement, as in the case of known σ, gives a general large-sample CI for μ.

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

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

is a **large-sample confidence interval for** μ with confidence level approximately 100(1 – α)%. This formula is valid regardless of the shape of the population distribution.

A Large-Sample Interval for μ

In words, the CI (7.8) is

point estimate of $\mu \pm$ (z critical value) (estimated standard error of the mean).

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

This is somewhat more conservative than the rule of thumb for the CLT because of the additional variability introduced by using S in place of σ.

Example 6

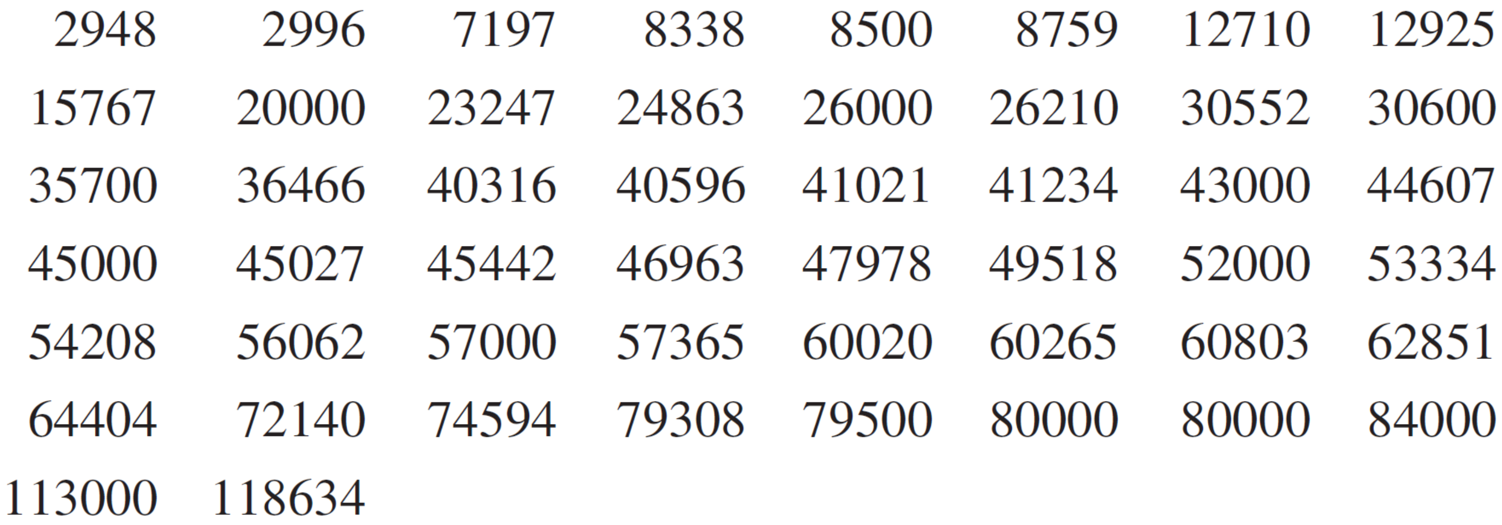
Haven’t you always wanted to own a Porsche? The author thought maybe he could afford a Boxster, the cheapest model. So he went to www.cars.com on Nov. 18, 2009, and found a total of 1113 such cars listed.

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:

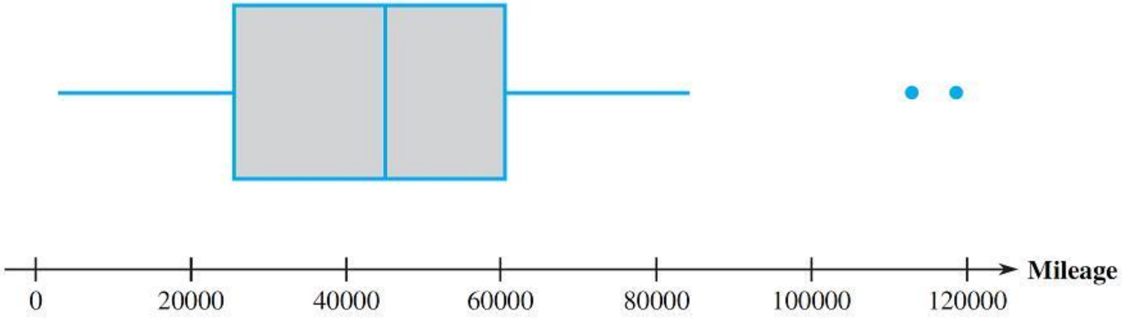


2948	2996	7197	8338	8500	8759	12710	12925
15767	20000	23247	24863	26000	26210	30552	30600
35700	36466	40316	40596	41021	41234	43000	44607
45000	45027	45442	46963	47978	49518	52000	53334
54208	56062	57000	57365	60020	60265	60803	62851
64404	72140	74594	79308	79500	80000	80000	84000
113000	118634						

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, f_s = 34,265.

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 and f_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 $z_{.025} = 1.96$, 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, $38{,}294.7 < \mu < 53{,}064.1$ 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.

Example 6

cont'd

Is the interval we've calculated one of the 95% that in the long run includes the parameter being estimated, or is it one of the "bad" 5% that does not do so? Without knowing the value of μ, we cannot tell.

Remember that the confidence level refers to the long run capture percentage when the formula is used repeatedly on various samples; it cannot be interpreted for a single sample and the resulting interval.