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

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

sample σ

has approximately a standard normal distribution. This implies that

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

n เพิ่มขึ้น

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 μ ± (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

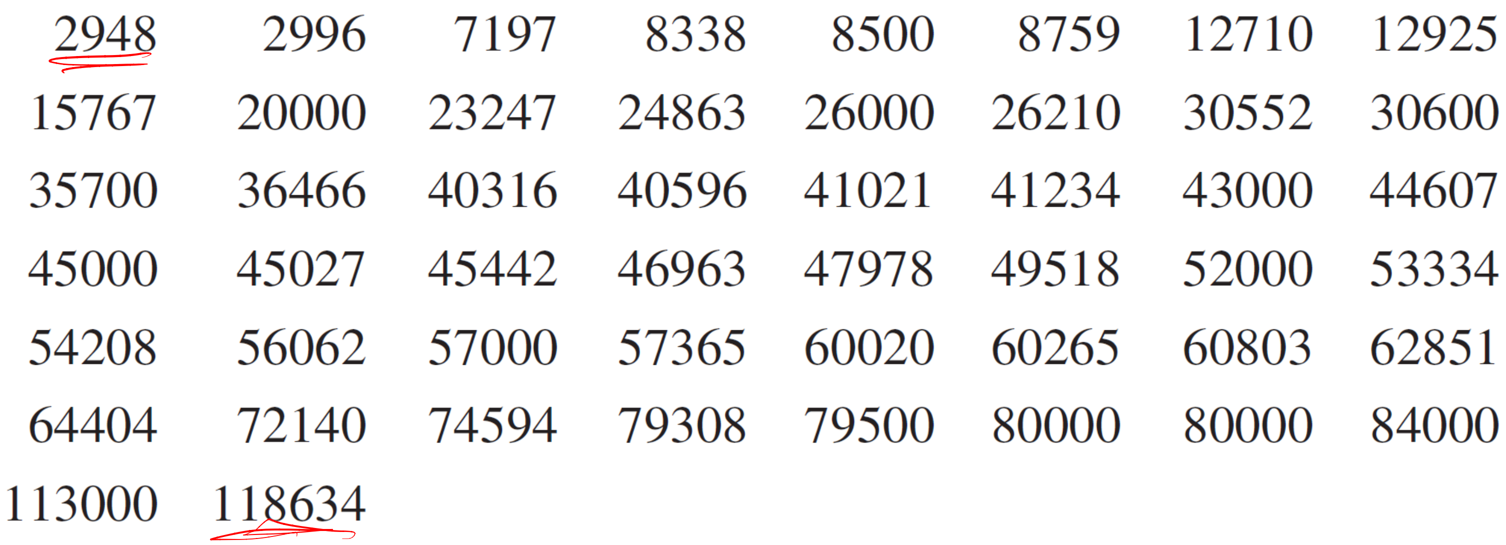
Haven’t you always wanted to own a Porsche? The author thought maybe he could afford a Boxster, the cheapest model. So he went to www.cars.com on Nov. 18, 2009, and found a total of 1113 such cars listed.

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:

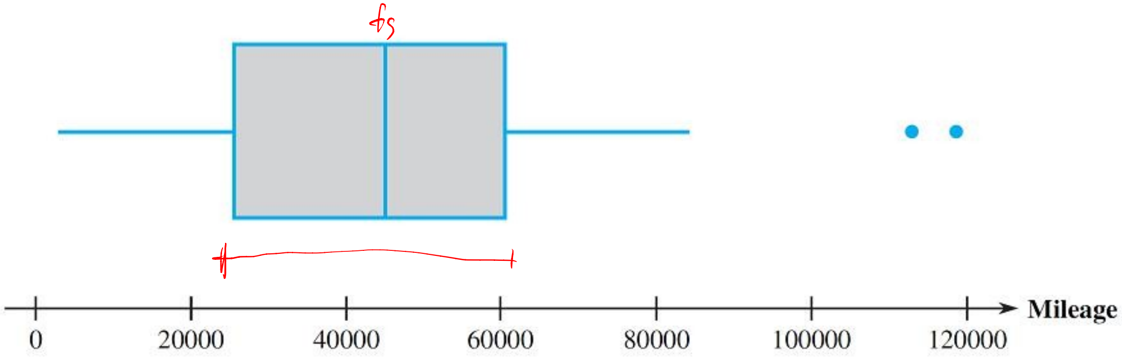


2948	2996	7197	8338	8500	8759	12710	12925
15767	20000	23247	24863	26000	26210	30552	30600
35700	36466	40316	40596	41021	41234	43000	44607
45000	45027	45442	46963	47978	49518	52000	53334
54208	56062	57000	57365	60020	60265	60803	62851
64404	72140	74594	79308	79500	80000	80000	84000
113000	118634						

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

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

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

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.

MT 0 discussion b

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.

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 $\theta = \mu$, $\hat{\mu} = \overline{X}$ is an unbiased estimator whose distribution is approximately normal when *n* is large and $\sigma_{\hat{\mu}} = \sigma_{\overline{X}} = \sigma/\sqrt{n}$. 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 $\theta = \mu$).

A General Large-Sample Confidence Interval

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.

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)

A Confidence Interval for a Population Proportion

A Confidence Interval for a Population Proportion

Let *p* denote the proportion of “successes” in a population, where *success* identifies an individual or object that has a specified property (e.g., individuals who graduated from college, computers that do not need warranty service, etc.).

A random sample of *n* individuals is to be selected, and *X* is the number of successes in the sample. Provided that *n* is small compared to the population size, *X* can be regarded as a binomial rv with $E(X) = np$ and

$$\sigma_X = \sqrt{np(1-p)}$$

Furthermore, if both $np \geq 10$ and $nq \geq 10$, ($q = 1 - p$), *X* has approximately a normal distribution.

A Confidence Interval for a Population Proportion

The natural estimator of *p* is $\hat{p}$ = *X*/*n*, the sample fraction of successes. Since $\hat{p}$ is just *X* multiplied by the constant 1/*n*, $\hat{p}$ also has approximately a normal distribution. As we know that, *E*($\hat{p}$) = *p* (unbiasedness) and $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.

The standard deviation $\sigma_{\hat{p}}$ involves the unknown parameter *p*. Standardizing $\hat{p}$ by subtracting *p* and dividing by $\sigma_{\hat{p}}$ then implies that

[illegible]

pop

$$P\left(-z_{\alpha/2} < \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1-\alpha$$

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

$$\sqrt{np(1-p)}/\sqrt{n}$$

A Confidence Interval for a Population Proportion

Proceeding as suggested in the subsection “Deriving a Confidence Interval”, the confidence limits result from replacing each < by = and solving the resulting quadratic equation for *p.* This gives the two roots

$$p = \frac{\hat{p} + z_{\alpha/2}^2/2n}{1 + z_{\alpha/2}^2/n} \pm z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n}$$

$$= \tilde{p} \pm z_{\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n}$$

A Confidence Interval for a Population Proportion

Proposition

Let $\tilde{p} = \dfrac{\hat{p} + z_{\alpha/2}^2/2n}{1 + z_{\alpha/2}^2/n}$. Then a **confidence interval for a population proportion *p*** with confidence level approximately 100(1 – α) % is

$$\tilde{p} \pm z_{\alpha/2} \frac{\sqrt{\hat{p}\hat{q}/n + z_{\alpha/2}^2/4n^2}}{1 + z_{\alpha/2}^2/n} \qquad \textbf{(7.10)}$$

A Confidence Interval for a Population Proportion

where $\hat{q} = 1 - \hat{p}$ and, as before, the – in (7.10) corresponds to the lower confidence limit and the + to the upper confidence limit.

This is often referred to as the *score CI* for *p*.

A Confidence Interval for a Population Proportion

If the sample size *n* is very large, then $z^2/2n$ is generally quite negligible (small) compared to $\hat{p}$ and z^2/n is quite negligible compared to 1, from which $\tilde{p} \approx \hat{p}$. In this case $z^2/4n^2$ is also negligible compared to pq/n (n^2 is a much larger divisor than is *n*); as a result, the dominant term in the ± expression is $z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}$ and the score interval is approximately

ประมาณค่า

$$\boxed{\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}} \quad \text{จำ*} \qquad \mathbf{(7.11)}$$

This latter interval has the general form $\hat{\theta} \pm z_{\alpha/2}\hat{\sigma}_{\hat{\theta}}$ of a large-sample interval suggested in the last subsection.

A Confidence Interval for a Population Proportion

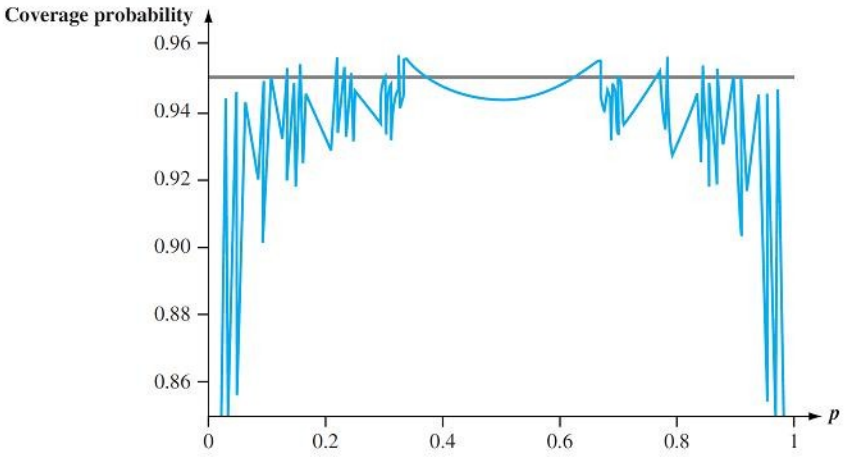
The approximate CI (7.11) is the one that for decades has appeared in introductory statistics textbooks. It clearly has a much simpler and more appealing form than the score CI. So why bother with the latter?

First of all, suppose we use $z_{.025}$ = 1.96 in the traditional formula (7.11). Then our *nominal* confidence level (the one we think we're buying by using that *z* critical value) is approximately 95%.

So before a sample is selected, the probability that the random interval includes the actual value of *p* (i.e., the *coverage probability*) should be about .95.

A Confidence Interval for a Population Proportion

But as Figure 7.6 shows for the case *n* = 100, the actual coverage probability for this interval can differ considerably from the nominal probability .95, particularly when *p* is not close to .5 (the graph of coverage probability versus *p* is very jagged because the underlying binomial probability distribution is discrete rather than continuous).



Actual coverage probability for the interval (7.11) for varying values of p when *n* = 100

Figure 7.6

A Confidence Interval for a Population Proportion

This is generally speaking a deficiency of the traditional interval—the actual confidence level can be quite different from the nominal level even for reasonably large sample sizes.

Recent research has shown that the score interval rectifies this behavior—for virtually all sample sizes and values of *p*, its actual confidence level will be quite close to the nominal level specified by the choice of $z_{\alpha/2}$.

This is due largely to the fact that the score interval is shifted a bit toward .5 compared to the traditional interval.

A Confidence Interval for a Population Proportion

In particular, the midpoint $\tilde{p}$ of the score interval is always a bit closer to .5 than is the midpoint $\hat{p}$ of the traditional interval. This is especially important when *p* is close to 0 or 1.

In addition, the score interval can be used with nearly all sample sizes and parameter values.

A Confidence Interval for a Population Proportion

It is thus not necessary to check the conditions $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ that would be required were the traditional interval employed.

So rather than asking when *n* is large enough for (7.11) to yield a good approximation to (7.10), our recommendation is that the score CI should *always* be used.

The slight additional tediousness of the computation is outweighed by the desirable properties of the interval.

Example 8

The article “Repeatability and Reproducibility for Pass/Fail Data” (*J. of Testing and Eval.*, 1997: 151–153) reported that in *n* = 48 trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette.

Let *p* denote the long-run proportion of all such trials that would result in ignition. A point estimate for *p* is $\hat{p}$ = 16/48 = .333. A confidence interval for *p* with a confidence level of approximately 95% is

$$\overset{\tilde{p}}{\frac{.333 + (1.96)^2/96}{1 + (1.96)^2/48}} \pm (1.96)\frac{\sqrt{(.333)(.667)/48 + (1.96)^2/9216}}{1 + (1.96)^2/48}$$

Example 8

cont'd

$$= .345 \pm .129$$

$$= (.216, .474)$$

This interval is quite wide because a sample size of 48 is not at all large when estimating a proportion.

The traditional interval is

$$.333 \pm 1.96\sqrt{(.333)(.667)/48} = .333 \pm .133$$

$$= (.200, .466)$$

Example 8

cont’d

These two intervals would be in much closer agreement were the sample size substantially larger.

A Confidence Interval for a Population Proportion

Equating the width of the CI for *p* to a prespecified width *w* gives a quadratic equation for the sample size *n* necessary to give an interval with a desired degree of precision. Suppressing the subscript in $z_{\alpha/2}$, the solution is

$$n = \frac{2z^2\hat{p}\hat{q} - z^2w^2 \pm \sqrt{4z^4\hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^2z^4}}{w^2} \qquad \textbf{(7.12)}$$

Neglecting the terms in the numerator involving w^2 gives

$$\boxed{n \approx \frac{4z^2\hat{p}\hat{q}}{w^2}}$$

$$W = 2 \times Z_{\frac{\alpha}{2}} \times \left(\frac{S}{\sqrt{n}}\right)$$

A Confidence Interval for a Population Proportion

This latter expression is what results from equating the width of the traditional interval to *w*.

These formulas unfortunately involve the unknown $\hat{p}$. The most conservative approach is to take advantage of the fact that $\hat{p}\hat{q}[= \hat{p}(1 - \hat{p})]$ is a maximum when $\hat{p}$ = .5. Thus if $\hat{p}$ = $\hat{q}$ = .5 is used in (7.12), the width will be at most *w* regardless of what value of $\hat{p}$ results from the sample.

Alternatively, if the investigator believes strongly, based on prior information, that $p \leq p_0 \leq .5$, then p_0 can be used in place of $\hat{p}$. A similar comment applies when $p \geq p_0 \geq .5$