7 Statistical Intervals Based on a Single Sample

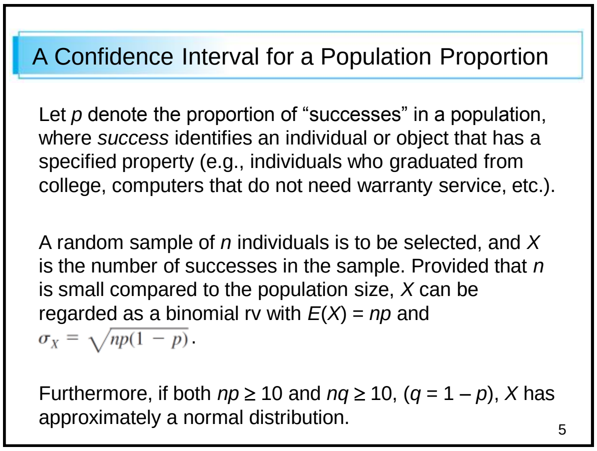
7.2 Large-Sample Confidence Intervals for a Population Mean and Proportion

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 Confidence Interval for a Population Proportion

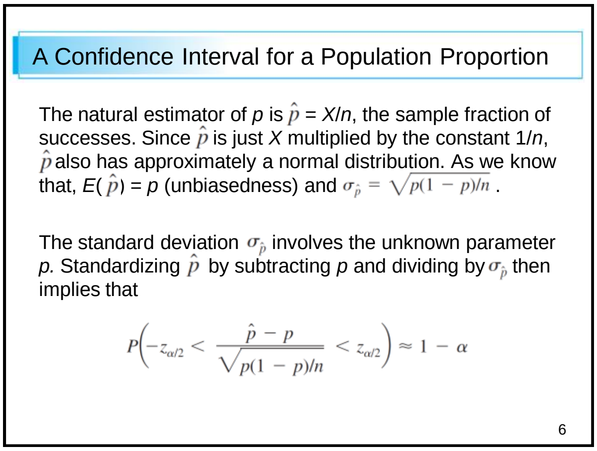


A Confidence Interval for a Population Proportion

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* ≥ 10 and *nq* ≥ 10, (*q* = 1 – *p*), *X* has approximately a normal distribution.

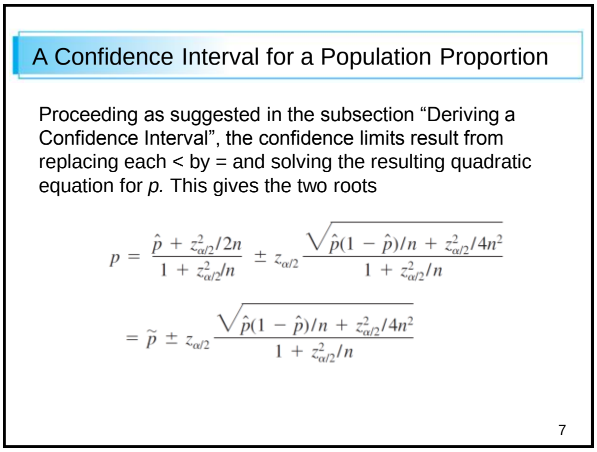


A Confidence Interval for a Population Proportion

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

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

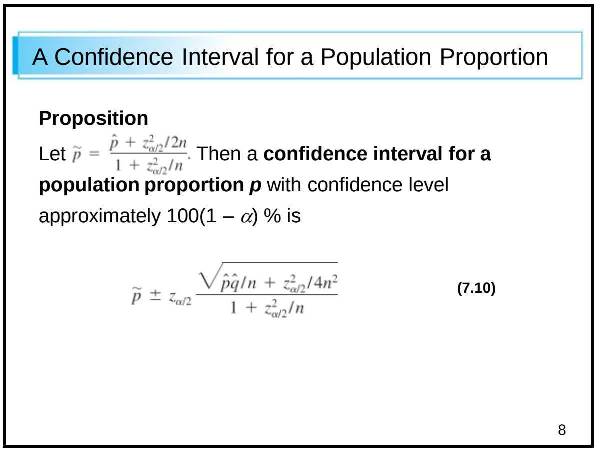


A Confidence Interval for a Population Proportion

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

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

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



A Confidence Interval for a Population Proportion

Proposition

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

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

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

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} \qquad \textbf{(7.11)}$$

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

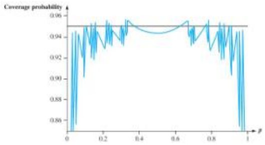
The approximate CI (7.11) is the one that for decades has appeared in introductory statistics textbooks. It clearly has a much simpler and more appealing form than the score CI. So why bother with the latter?

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

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

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

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

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 $p \geq p_0 \geq .5$

One-Sided Confidence Intervals (Confidence Bounds)

One-Sided Confidence Intervals (Confidence Bounds)

The confidence intervals discussed thus far give both a lower confidence bound *and* an upper confidence bound for the parameter being estimated.

In some circumstances, an investigator will want only one of these two types of bounds.

For example, a psychologist may wish to calculate a 95% upper confidence bound for true average reaction time to a particular stimulus, or a reliability engineer may want only a lower confidence bound for true average lifetime of components of a certain type.

One-Sided Confidence Intervals (Confidence Bounds)

Because the cumulative area under the standard normal curve to the left of 1.645 is .95,

$$P\left(\frac{\overline{X}-\mu}{S/\sqrt{n}} < 1.645\right) \approx .95$$

Manipulating the inequality inside the parentheses to isolate μ on one side and replacing rv’s by calculated values gives the inequality $\mu > \overline{x} – 1.645s/\sqrt{n}$; the expression on the right is the desired lower confidence bound.

One-Sided Confidence Intervals (Confidence Bounds)

Starting with $P(–1.645 < Z) \approx .95$ and manipulating the inequality results in the upper confidence bound. A similar argument gives a one-sided bound associated with any other confidence level.

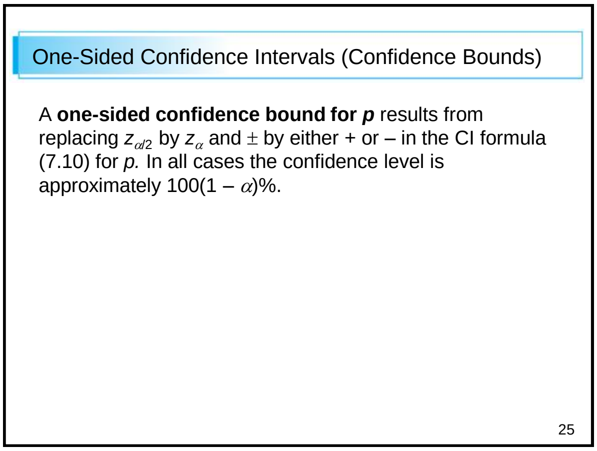
Proposition

A **large-sample upper confidence bound for** μ is

$$\mu < \overline{x} + z_\alpha \cdot \frac{s}{\sqrt{n}}$$

and a **large-sample lower confidence bound for** μ is

$$\mu > \overline{x} - z_\alpha \cdot \frac{s}{\sqrt{n}}$$



One-Sided Confidence Intervals (Confidence Bounds)

A **one-sided confidence bound for *p*** results from replacing $z_{\alpha/2}$ by z_{α} and ± by either + or – in the CI formula (7.10) for *p.* In all cases the confidence level is approximately $100(1 - \alpha)\%$.