
Lecture 13: Confidence Interval for a Population Proportion

STAT 515 — Statistical Methods I

Dr. Yen-Yi Ho

Department of Statistics, University of South Carolina

1. Large-Sample Confidence Interval (CI) for a Proportion p
2. The Adjusted (Agresti–Coull) Confidence Interval
3. Exercises on CIs
4. Determining the needed sample size

7.4 Large-Sample CI for a Proportion p

By the Central Limit Theorem,

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1).$$

$100(1 - \alpha)\%$ Confidence Interval for p

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \approx \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Requires a large sample to ensure the validity of the interval:

$$n\hat{p} = x \geq 5 \quad \text{and} \quad n(1 - \hat{p}) = n - x \geq 5$$

(page 484 in Chap. 10.2).

Example: Candidate Support in South Carolina

Voter Poll

A candidate wants an estimate of what their current support is among all of the registered voters in South Carolina. A random sample of 200 registered voters shows that 94 would currently vote for them. Construct a 95% confidence interval for the percent of support.

Define the parameter: p is the true proportion of registered voters who support the candidate.

$$\hat{p} = \frac{x}{n} = \frac{94}{200} = .47; \quad \text{check } n\hat{p} = 94 \geq 5 \text{ and } n(1 - \hat{p}) = 106 \geq 5. \quad \checkmark$$

$$.47 \pm 1.96 \sqrt{\frac{.47(1 - .47)}{200}} = .47 \pm .069 = (.401, .539)$$

Example: High School Seniors

In-State Education

In a random sample of 250 high school seniors in a Midwestern state, 168 of them said that they intended to continue their education at an in-state college or university. Find a 95% confidence interval for the corresponding true proportion.

Define the parameter: p is the true proportion of high school seniors who intend to continue in-state education.

$$\hat{p} = \frac{168}{250} = .672; \quad \text{check } n\hat{p} = 168 \geq 5 \text{ and } n(1 - \hat{p}) = 82 \geq 5. \quad \checkmark$$

$$.672 \pm 1.96 \sqrt{\frac{.672(1 - .672)}{250}} = .672 \pm .0582 = (.614, .730)$$

1. Large-Sample Confidence Interval (CI) for a Proportion p
2. The Adjusted (Agresti–Coull) Confidence Interval
3. Exercises on CIs
4. Determining the needed sample size

The Adjusted (Agresti–Coull) Confidence Interval

Caution: unless the sample size n is extremely large, the large-sample procedure performs **poorly** when p is near 0 or near 1.

Adjusted $100(1 - \alpha)\%$ CI for p

$$\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{\tilde{n}}}, \quad \tilde{p} = \frac{\tilde{x}}{\tilde{n}}$$

where

$$\tilde{x} = x + 0.5 z_{\frac{\alpha}{2}}^2 \quad (\approx x + 2 \text{ for } 95\%), \quad \tilde{n} = n + z_{\frac{\alpha}{2}}^2 \quad (\approx n + 4 \text{ for } 95\%)$$

For 95%: $z_{.025} = 1.96 \approx 2$, so $0.5z_{.025}^2 \approx 2$ and $z_{.025}^2 \approx 4$.

Agresti, A. & Coull, B. A. (1998). Approximate Is Better than “Exact” for Interval Estimation of Binomial Proportions. *The American Statistician*, 52, 119–126.

Example: Victims of Violent Crime

True Odds: How Risk Affects Your Everyday Life (Walsh, 1997)

The probability of being the victim of a violent crime is less than .01. Suppose that, in a random sample of 200 Americans, 3 were victims of a violent crime. Use a 95% confidence interval to estimate the true proportion of Americans who were victims of a violent crime.

$n = 200$, $x = 3$, $\hat{p} = \frac{3}{200} = .015$, so $n\hat{p} = 3 \not\geq 5$ — **the sample is not large enough** for the ordinary interval.

Adjusted: $\tilde{x} = x + 2 = 5$, $\tilde{n} = n + 4 = 204$, $\tilde{p} = \frac{5}{204} = .025$:

$$.025 \pm 1.96 \sqrt{\frac{.025(1 - .025)}{204}} = .025 \pm .021 = (.004, .046)$$

We are 95% confident that the true proportion is between .4% and 4.6%.

Summary: Which Interval Do I Use?

Parameter	Interval	Condition
μ (small n)	$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$	normal population; robust
μ ($n \geq 30$)	$\bar{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	CLT; no normality needed
σ^2	$\left(\frac{(n-1)s^2}{\chi_{\text{upper}}^2}, \frac{(n-1)s^2}{\chi_{\text{lower}}^2} \right)$	normal population; not robust (Chap 7)
p	$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$np \geq 5$ and $np(1-p) \geq 5$
p (adjusted)	$\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}}$	small n , or p near 0 or 1

Remember

Higher confidence $\Rightarrow$ wider interval. Larger $n \Rightarrow$ narrower interval.

1. Large-Sample Confidence Interval (CI) for a Proportion p
2. The Adjusted (Agresti–Coull) Confidence Interval
3. Exercises on CIs
4. Determining the needed sample size

Confidence intervals used in these exercises

- Mean, σ *known*: $\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$
- Mean, σ *unknown*: $\bar{x} \pm t_{\alpha/2, df=n-1} \left(\frac{s}{\sqrt{n}} \right)$
- Proportion: $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$

All three require a **random sample** from the target population.

Exercise 4: An Opinion Poll

Problem 4

USA Today reported on the results of an opinion poll in which adults were asked what one thing they are most likely to do when they are home sick with a cold or the flu. In the survey, 63% said that they are most likely to sleep and 18% said that they would watch television.

- (a) Assuming that the sample size is 1,000 for this poll, construct a 95% confidence interval for the true percentage of all adults who would choose to sleep when they are at home sick.
- (b) If the true percentage of adults who would choose to sleep when they are at home sick is 70%, would you be surprised? Explain.

Exercise 4(a): Check the Conditions

Parameter and estimate

p : the true percentage (proportion) of all adults who would choose to sleep when at home sick.

Sample proportion: $\hat{p} = 63\% = .63$, $n = 1000$.

Conditions for the large-sample proportion CI:

$$n\hat{p} \geq 5 : 1000(.63) = 630 \geq 5 \quad \checkmark$$

$$n\hat{q} \geq 5 : 1000(1 - .63) = 370 \geq 5 \quad \checkmark$$

Exercise 4(a): The Computation

A 95% CI for p

$$\hat{p} \pm z_{.025} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = .63 \pm 1.96 \sqrt{\frac{.63(.37)}{1000}} = (.60, .66)$$

Interpretation. We are 95% confident that the interval (60%, 66%) covers the true percentage of all adults who would choose to sleep when at home sick - meaning that if we repeated this sampling procedure many times, 95% of the intervals we built would contain p .

Exercise 4(b): Would 70% Be Surprising?

Answer

Yes. We would doubt that the true percentage is 70%, because 70% is **outside** the 95% confidence interval (60%, 66%).

- A confidence interval gives the range of **plausible** values for the parameter.
- A value far outside the interval is not well supported by the data.
- This is exactly the logic we will formalize as a **hypothesis test** later in this lecture.

1. Large-Sample Confidence Interval (CI) for a Proportion p
2. The Adjusted (Agresti–Coull) Confidence Interval
3. Exercises on CIs
4. Determining the needed sample size

Section 2

7.5 Determining the Needed Sample Size

7.5 Determining the Needed Sample Size

- One important element of a designed experiment is to determine the size of the sample.
- How do we determine the sample size related to the **reliability** of a confidence interval? We want the resulting confidence interval to achieve a certain **confidence level** and, at the same time, a certain required **precision**.

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

The diagram shows the formula for a confidence interval, $\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$, enclosed in a red rectangular box. Two blue arrows point from below towards the formula. The left arrow points to the $z_{\alpha/2}$ term and is labeled "confidence level". The right arrow points to the standard error term $\left(\frac{\sigma}{\sqrt{n}} \right)$ and is labeled "precision (half width)".

confidence level precision (half width)

Needed Sample Size for Estimating a Population Mean

Sampling error SE

Denote the desired **half width** of the confidence interval as SE — the allowed *Sampling Error* of the sample mean from the true mean — and let the confidence level be $1 - \alpha$.

Needed sample size for estimating the population mean

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

If the population standard deviation σ is unknown, plug in an estimate of it, such as the sample standard deviation s .

Example

Say we expect σ to be approximately 15. Approximately how large does n need to be to have the 95% confidence interval for μ be ± 1 ?

Here $SE = 1$ and $z_{.025} = 1.96$. Set the half width equal to 1:

$$z_{.025} \frac{\sigma}{\sqrt{n}} = 1 \implies n = \left(\frac{z_{.025} \cdot \sigma}{1} \right)^2 = \left(\frac{1.96(15)}{1} \right)^2 = 864.36 \approx \mathbf{865}$$

Always round the needed sample size up.

Needed sample size for a proportion

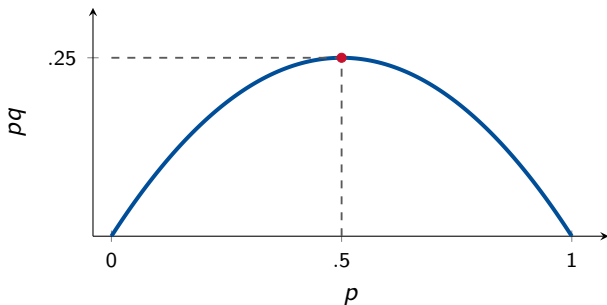
In order to estimate the population proportion p with sampling error SE and with $100(1 - \alpha)\%$ confidence level, the needed sample size is

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

- Because the value of the product pq is unknown, it can be estimated by a sample proportion $\hat{p}$ from a previous study.
- We can also obtain a **conservatively large** sample size by plugging in the **maximum value of pq , which is 0.25**.

Why Is the Maximum of pq Equal to 0.25?

Since $q = 1 - p$, the product is $pq = p(1 - p)$, a downward parabola that is maximized at $p = .5$, where $pq = (.5)(.5) = .25$.



Example: Proportion Within $\pm .03$ with 95% Confidence

Example

What sample size is needed for 95% confidence for a population proportion to be $\pm .03$?

Start from the CI $\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n}$ and set the half width equal to the required sampling error:

$$z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} = .03 = SE$$

$$n = \frac{(z_{\alpha/2})^2 pq}{(.03)^2} = \frac{(z_{.025})^2 pq}{(.03)^2} = \frac{1.96^2 (.25)}{(.03)^2} = 1067.11 \approx \mathbf{1068}$$

using the maximum value of pq , which is .25.

Exercise 7: Defective Cellular Telephones

Problem 7

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 to within .01 with 90% confidence? (For this application, in general 10% defective can already be severe.)

Exercise 7: The Computation

Given: $SE = .01$, $1 - \alpha = .90 \Rightarrow z_{.05} = 1.645$, and we use the prior information $p = .1$, so $q = .9$.

Needed sample size

$$n = \frac{z_{\alpha/2}^2 pq}{SE^2} = \frac{z_{.05}^2 (.1)(.9)}{(.01)^2} = \frac{1.645^2 (.1)(.9)}{(.01)^2} = \mathbf{2436}$$

- Here the prior value $p = .1$ is used rather than the conservative $pq = .25$; since $pq = (.1)(.9) = .09 < .25$, this gives a **smaller** required sample size.