Introduction to Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

October 13, 2019

Notes

Topics

Notes

Objectives

Objectives:

- Compute and interpret a CI for a population proportion.
- Determine the sample size required for the margin of error in a CI for a population proportion to be no bigger than some specified value.

Notes

Confidence Interval for a Population Proportion P (12.1)

The Sample Proportion $\hat{P}$ and the Population Proportion P

- The mean $\bar{x}$ (or μ) is appropriate for summarizing a ***quantitative*** (numerical) variable.
- For a ***qualitative*** (categorical) variable, the appropriate summary is the ***proportion*** of individuals in a category.

Notes

- The ***population proportion*** (a parameter) and ***sample proportion*** (a statistic) are defined on the next two slides.

Notes

Population Proportion: Consider a **population** in which each individual is classified according to a **qualitative** variable having **two categories**, **"success"** and **"failure"**, say.

The ***population proportion***, denoted p, is defined as:

$$p = \frac{\text{Number of "successes" in the population}}{N}$$

where N is the population size.

Notes

Sample Proportion: Now consider a **sample** of size n from the population just described.

The ***sample proportion***, denoted $\hat{p}$, is defined as:

$$\hat{p} = \frac{\text{Number of "successes" in the sample}}{n}$$

Notes

- The *sample proportion* $\hat{p}$ is used to **estimate** the *population proportion* p.

Notes

Example

Suppose a random **sample** of $n = 10$ students from a college was asked whether or not they smoke cigarettes (Yes or No), and that the resulting data are

Yes No Yes No Yes No No Yes No No

The **sample proportion** of smokers is

$$\hat{p} = \frac{4}{10} = \mathbf{0.4},$$

and so we'd **estimate** that the **true (unknown) proportion** p that smokes in the **population** is **0.4**, or **40%**.

Notes

- When the **sample proportion** $\hat{p}$ is to estimate a **population proportion** p, the **sampling error** is:

Sampling Error of the Sample Proportion:

$$\text{Sampling Error} = \hat{p} - p$$

Notes

The Sampling Distribution of the Sample Proportion $\hat{P}$

- The **sampling distribution** of $\hat{p}$ can be used to gauge how large the **sampling error** of $\hat{p}$ might be when estimating p.
- In the slides ahead, we'll see that:

  As long as the **sample size n is large**, the **sampling distribution of** $\hat{p}$ will be **approximately normal**.

Notes

Normality of $\hat{P}$ When n is Large

Normality of $\hat{P}$: If we take a **sample** of size n from a **population** of **"successes"** and **"failures"** whose proportion of "successes" is p, then as long as the **sample size** n is ***large***:

The $\hat{p}$ **distribution** will be (at least approximately) ***normal*** with mean $\mu_{\hat{p}}$ and standard deviation $\sigma_{\hat{p}}$, where

$$\mu_{\hat{p}} = p \qquad \text{and} \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Notes

- **Interpretation of $\mu_{\hat{p}}$ and $\sigma_{\hat{p}}$:**
  - $\mu_{\hat{p}}$ is the value that $\hat{p}$ takes, **on average**. Thus, because $\mu_{\hat{p}} = p$, **on average** the **sample proportion** equals the **population proportion**.
  - $\sigma_{\hat{p}}$ represents a **typical deviation** of $\hat{p}$ away from p, i.e. a typical **sampling error**. Thus, because $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$, the size of a **typical sampling error** is $\sqrt{p(1-p)/n}$.
- $\sqrt{p(1-p)/n}$ is often called the ***standard error*** of $\hat{p}$.

Notes

- The **standard error** of $\hat{p}$ **will be small** if either:
  1. The population proportion p **is close to 0 or 1** (i.e. the population is fairly ***homogeneous***).
  2. The sample size n **is large**.

  Under either of these conditions, $\hat{p}$ will be a **precise estimator** of p.

Notes

Confidence Interval for P

- We want to **estimate** p using a **CI**, which will be of the form

$$\hat{p} \pm \text{Margin of Error}$$

  On the slides ahead, we'll determine how big the **margin of error** would need to be for us to be **95% confident** that the interval will contain p.

  First, though, we'll look at an example.

Notes

Exercise

A June 28, 2016 report by the polling organization Marist states:

> *"A majority of Americans oppose legalizing the sale of human organs for transplant purposes."*

The conclusion was based on a survey of $n = 516$ adult Americans.

According to the report:

> "**55%** *of Americans do not think the sale of human organs for transplant purposes should be legal."*

The reported **margin of error** is ± 4.3 percentage points.

Notes

a) What is the **(unknown) population parameter** being investigated by the poll?

b) What's the value of the **sample statistic** being used to **estimate** the **(unknown) parameter**?

c) Using the **estimate** and **margin of error**, determine the **CI** for p.

d) The **level of confidence** isn't explicitly stated. What **level of confidence** was most likely used?

e) Based on the **CI** of part *c*, is the conclusion stated in the first quote above justified?

Notes

Exercise

The Marist Poll described in the last exercise also reported the following result:

> *"In assessing the moral dimension of this debate,* ***49%*** *of U.S. residents believe it is wrong for someone to sell their organs, such as a kidney, to a transplant patient who can afford to pay the price."*

a) For this part of the survey, what is the **(unknown) population parameter** of interest?

b) What's the value of the **sample statistic** being used to **estimate** the **(unknown) parameter**?

Notes

c) Using the **estimate** and **margin of error** (± 4.3), determine the **CI** for p.

d) Based on the **CI**, would it be reasonable to conclude that **fewer than half** of all Americans believe it's wrong for someone to sell their organs?

Notes

- To see how the **margin of error** is determined, recall that $\hat{p}$ follows a **normal** distribution with **mean** and **standard error**

$$\mu_{\hat{p}} = p \qquad \text{and} \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

Notes

- Thus the **standardized version** of $\hat{p}$,

$$z = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}},$$

follows a **standard normal** distribution ...

... and therefore will lie between

$$-z_{0.025} = -1.96 \qquad \text{and} \qquad z_{0.025} = 1.96$$

95% of the time when we take a sample of size n.

Notes

- In other words,

$$-1.96 \le \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \le 1.96$$

95% of the time.

We'll "solve" for p:

Multiplying through by $\sqrt{p(1-p)/n}$, we can rewrite this as

$$-1.96\sqrt{\frac{p(1-p)}{n}} \le \hat{p} - p \le 1.96\sqrt{\frac{p(1-p)}{n}}$$

which says the ***sampling error*** **won't be bigger than** $\mathbf{1.96\sqrt{p(1-p)/n}}$ **95% of the time**.

Notes

- (cont'd)

Subtracting $\hat{p}$ from all three terms gives

$$-\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \le -p \le -\hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Multiplying each of the three terms above by -1 (which changes the direction of the inequalities) gives

$$\hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}} \ge p \ge \hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Finally, reordering the terms, we get that **95% of the time**,

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}.$$

Notes

- (cont'd)

Thus, we can be **95% confident that p will be between**

$$\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \qquad \text{and} \qquad \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}$$

Unfortunately, we ***can't*** use

$$1.96\sqrt{\frac{p(1-p)}{n}}$$

as the margin of error in our CI because it depends on the **unknown** population proportion p.

Instead, we plug in the **estimate** $\hat{p}$ of p.

Notes

95% One-Proportion z CI for P: A 95% confidence interval for the **unknown** population **proportion** of "successes" p is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

and the **margin of error** is

$$\text{Margin of Error} = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

(These are valid when the sample is from a population of **"successes"** and **"failures"** and the sample size n is **large**.)

Notes

- For **other levels of confidence**, we replace 1.96 by the appropriate z ***critical value***:

Commonly Used Z Critical Values:

$z_{0.05}$	$= 1.645$	for a 90% level of confidence
$z_{0.025}$	$= 1.96$	for a 95% level of confidence
$z_{0.005}$	$= 2.58$	for a 99% level of confidence

These $z_{\alpha/2}$ values are obtained from Table II.

Notes

One-Proportion z CI for p: The ***one-proportion z confidence interval*** for the **unknown** population **proportion** of "successes" p is

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

with ***margin of error***

$$\text{Margin of Error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $z_{\alpha/2}$ is a z **critical value** and α is either **0.10**, **0.05**, or **0.01**, depending on the level of confidence (see the next slide).

Notes

α	$= \mathbf{0.10}$	for a **90%** level of confidence (so $1-\alpha = \mathbf{0.90}$, $\alpha/2 = \mathbf{0.05}$, and $z_{\mathbf{0.05}} = \mathbf{1.645}$)
α	$= \mathbf{0.05}$	for a **95%** level of confidence (so $1-\alpha = \mathbf{0.95}$, $\alpha/2 = \mathbf{0.025}$, and $z_{\mathbf{0.025}} = \mathbf{1.96}$)
α	$= \mathbf{0.01}$	for a **99%** level of confidence (so $1-\alpha = \mathbf{0.99}$, $\alpha/2 = \mathbf{0.005}$, and $z_{\mathbf{0.005}} = \mathbf{2.58}$)

(The CI and margin of error are valid when the sample is from a population of **"successes"** and **"failures"** and the sample size n is **large**.)

Notes

Example

A June 2-7, 2015 Gallup poll of **1,527** adults, aged 18 and older, living in the U.S. found that **1,390** (or **91%**) of those surveyed would vote for a Hispanic for president.

The **sample proportion** is

$$\hat{p} = \frac{1,390}{1,527} = \mathbf{0.91},$$

and this is the **estimate** of p, the true (unknown) population proportion of U.S. residents that would vote for a Hispanic.

Notes

Now we'll estimate the nationwide proportion p using a **95% CI**:

$$\begin{aligned}\hat{p} \pm z_{0.025}\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} &= 0.91 \pm 1.96 \times \sqrt{\tfrac{0.91(1-0.91)}{1{,}527}} \\ &= 0.91 \pm 0.01 \\ &= \mathbf{(0.90,\ \ 0.92)}\end{aligned}$$

This gives a range plausible values for p, and we can be **95% confident** that p is in the interval somewhere.

For example, based on the **CI**, it's **not plausible** that p is as small as **0.85** because this value isn't in the interval.

Notes

The **margin of error** in the **estimate**, $\hat{p} = 0.91$, of the nationwide proportion p is **0.01**, or **1 percentage point**.

We **interpret** this value as a measure of **how reliable** the estimate is.

More precisely, it tells us that the **sampling error** is **not likely** to be **larger** than **0.01**.

Notes

If we use a **99% confidence level**, the **CI** for p is

$$\begin{aligned}\hat{p} \pm z_{0.005}\sqrt{\tfrac{\hat{p}(1-\hat{p})}{n}} &= 0.91 \pm 2.58 \times \sqrt{\tfrac{0.91(1-0.91)}{1{,}527}} \\ &= 0.91 \pm 0.02 \\ &= \mathbf{(0.89,\ \ 0.93)}\end{aligned}$$

Note that the **99%** CI is **wider** than the **95%** CI.

Notes

Exercise

The Gallup poll from the last example also found that **886** (or **58%**) of the **1,527** adults surveyed would vote for an atheist for president.

a) **Compute** and **interpret** a **95% CI** for the **true (unknown) nationwide proportion** p that would vote for an atheist.

b) **Is it plausible**, based on the CI, that that p is **as large as 0.65**? Explain.

c) How big is the **margin of error** in the estimate, $\hat{p} = 0.58$, of p? How do you **interpret** this value?

d) If a **99% CI** was computed instead, would the interval be **wider** or **narrower** than the **95% CI**?

Notes

(Optional Section) Determining the Required Sample Size for Estimating p (12.1)

- Recall that the **margin of error** in the **CI** for a population proportion p is

$$\text{Margin of Error} \;=\; z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,.$$

We can make the **margin of error as small as we want** by using a **large enough** n.

Notes

- Suppose we want the **margin of error** to **no bigger than** some value E.

Solving

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \;\le\; E$$

for n gives

$$n \;\ge\; \hat{p}(1-\hat{p})\left(\frac{z_{\alpha/2}}{E}\right)^2.$$

But the right side depends on the value of the $\hat{p}$, which isn't known yet (because the sample hasn't been taken).

Instead, we plug in a ***guess*** $\hat{p}_g$ for $\hat{p}$, which gives the ***required sample size*** (see the next slide).

Notes

Sample Size for Estimating μ: The sample size required for the margin of error in a CI for p to be no bigger than some value E is

$$n \;\ge\; \hat{p}_g(1-\hat{p}_g)\left(\frac{z_{\alpha/2}}{E}\right)^2,$$

which we **round up** to the nearest integer.

Above, $\hat{p}_g$ is a guess for the value of $\hat{p}$, either:

1. $\hat{p}_g$ is a guessed value for $\hat{p}$ (e.g. based on prior studies)

   or

2. $\hat{p}_g = 0.5$ is used as the guess for $\hat{p}$.

Notes

- The second option (using $\hat{p}_g = 0.5$) is **conservative** (i.e. **safe**) in that the resulting n will be ***as large or larger*** than what's actually needed, and so the resulting **margin of error** is **guaranteed** to be ***as small or smaller*** than E.

Notes

Example

We want to conduct a poll to **estimate** the **true (unknown) proportion** of Colorado voters p that plan to vote for the Democrat in the next presidential election.

If we want the **margin of error** in a **95% CI** for p to be **0.03** (i.e. **3 percentage points**), **how big should the sample size be**?

Notes

Using the guess $\hat{p}_g = 0.5$ in the sample size calculation, we'd need a **sample size**

$$\begin{aligned} n &\geq \hat{p}_g(1-\hat{p}_g)\left(\frac{z_{0.025}}{E}\right)^2 \\ &= 0.5 \times (1-0,5)\left(\frac{1.96}{0.03}\right)^2 \\ &= \mathbf{1067.1}, \end{aligned}$$

which we **round up** to $n = \mathbf{1,068}$.

Notes

Exercise

Suppose instead that in our poll of Colorado voters, we only need the margin of error to be **0.04** (i.e. **4 percentage points**).

How big should the sample size be?

Use the guess $\hat{p}_g = 0.5$ in the sample size calculation.

Notes