

Probability and Statistics

Nels Grevstad

Metropolitan State University of Denver

ngrevsta@msudenver.edu

April 15, 2019

Nels Grevstad

CI for a Population Proportion p

Topics

- 1 CI for a Population Proportion p
 - CI for a Population Proportion p
 - Sampling Distribution of \hat{P}
 - Sample Size Determination

Nels Grevstad

CI for a Population Proportion p

Objectives

Objectives:

- Calculate and interpret a CI for a population proportion p when n is large.
- Determine the sample size n for attaining a desired CI width.

Nels Grevstad

CI for a Population Proportion p

CI for a Population Proportion p
Sampling Distribution of \hat{P}
Sample Size Determination

CI for a Population Proportion p (7.2)

- Consider a population of **successes** and **failures**, and let p denote the **population proportion** of successes.
- Suppose our goal is to **estimate** p using a random sample from the population.

Nels Grevstad

Notes

Notes

Notes

Notes

- The **point estimator** of p is the **sample proportion**, denoted \hat{P} .

Sample Proportion:

$$\hat{P} = \frac{\text{\# of Successes in the Sample}}{\text{Sample Size}} = \frac{X}{n}$$

where

X = The number of successes in the sample.

Nels Grevstad

Example

A random sample of $n = 10$ patrons at a restaurant were asked whether they smoke cigarettes (Yes or No). Here are the data.

Yes No Yes No Yes No No Yes No No

The **sample proportion** of smokers is

$$\hat{p} = \frac{4}{10} = 0.4.$$

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

Nels Grevstad

- The **sampling error** of the sample proportion is

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

- It's preferable to estimate p using a **confidence interval** because its **width** will reflect how big the **sampling error** might be.
- To derive the CI, we'll need the **sampling distribution** of \hat{P} .

Nels Grevstad

Sampling Distribution of the Sample Proportion \hat{P} (7.2)

- The numerator X in

$$\hat{P} = \frac{X}{n},$$

is the **number of successes** among the n individuals in the sample.

X can be regarded as a **binomial**(n, p) random variable.

Thus $E(X) = np$ and $V(X) = np(1 - p)$, and so ...

$$E(\hat{P}) = \frac{1}{n}E(X) = p$$

and

$$V(\hat{P}) = \frac{1}{n^2}V(X) = \frac{p(1 - p)}{n}.$$

Nels Grevstad

Notes

Notes

Notes

Notes

Mean and Variance of \hat{P} : For a **random sample** from a population of **successes** and **failures** whose proportion of successes is p , the **sampling distribution of \hat{P}** has mean $\mu_{\hat{p}}$ given by

$$\mu_{\hat{p}} = E(\hat{P}) = p$$

and variance $\sigma_{\hat{p}}^2$ given by

$$\sigma_{\hat{p}}^2 = V(\hat{P}) = \frac{p(1-p)}{n}.$$

- Because $E(\hat{P}) = p$, \hat{P} is an **unbiased** estimator of p .
- The standard deviation

$$SD(\hat{P}) = \sqrt{\frac{p(1-p)}{n}}$$

is called the **standard error of \hat{P}** . It represents a **typical deviation** of \hat{P} away from p .

- The **standard error** will be **small** if either:
 1. The population proportion p is close to 0 or 1, or
 2. The sample size n is **large**

- The following is a consequence of the **Normal Approximation to the Binomial**:

Proposition

Normality of \hat{P} : For a **random sample** from a population of **successes** and **failures** whose proportion of successes is p , if **n is large**,

$$\hat{P} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \quad (\text{approximately}).$$

Thus the **standardized** version of \hat{P} follows a **standard normal** distribution, i.e.

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1) \quad (\text{approximately}).$$

Notes

Notes

Notes

Notes

One-Sample z CI for p (7.2)

- To derive **95% CI for p** , note that

$$0.95 \approx P\left(-1.96 < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < 1.96\right)$$

$$\vdots$$

$$= P\left(\hat{P} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{P} + 1.96\sqrt{\frac{p(1-p)}{n}}\right),$$

Nels Grevstad

CI for a Population Proportion p CI for a Population Proportion p
Sampling Distribution of \hat{P}
Sample Size Determination

- Thus with probability 0.95, the (unknown) population proportion p will lie in the interval

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

But this interval involves the unknown p . To get around this, we'll plug in the estimate \hat{P} for p .

This leads to the following CI.

Nels Grevstad

CI for a Population Proportion p CI for a Population Proportion p
Sampling Distribution of \hat{P}
Sample Size Determination**Large-Sample $100(1 - \alpha)\%$ Confidence Interval for p :**

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

This is called the **one-sample z confidence interval for p** .

It's valid when the sample is from a population of successes and failures and n is large.

Nels Grevstad

CI for a Population Proportion p CI for a Population Proportion p
Sampling Distribution of \hat{P}
Sample Size Determination

- In practice, n is **large** enough for the **one-sample z CI for p** to be valid as long as

$$n\hat{P} \geq 10 \quad \text{and} \quad n(1 - \hat{P}) \geq 10.$$

i.e. as long as there are at least **10 successes** and at least **10 failures** in the sample.

Nels Grevstad

Notes

Notes

Notes

Notes

Example

A June, 2016 Marist poll of $n = 516$ adult Americans found that **284** (or **55%**) oppose legalizing the sale of human organs for transplant purposes.

The **sample proportion** is

$$\hat{p} = \frac{284}{516} = 0.55,$$

and this is the **point estimate** of p , the true (unknown) proportion of all Americans that oppose legalizing the sale of organs.

Nels Grevstad

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

$$\begin{aligned}\hat{P} \pm z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} &= 0.55 \pm 1.96 \sqrt{\frac{0.55(1-0.55)}{516}} \\ &= 0.55 \pm 0.04 \\ &= (0.51, 0.59)\end{aligned}$$

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

Nels Grevstad

Sample Size Determination (7.2)

- The **width** of the CI for p is

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

so if we want the **width** to be no bigger than w , the **sample size** n needs to be big enough that:

$$2 \left(z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} \right) = w.$$

Solving for n leads to the following sample size.

Nels Grevstad

Sample Size for a Desired CI Width: The width of the $100(1-\alpha)\%$ confidence interval for p will be approximately w when

$$n = \left(\frac{2z_{\alpha/2}}{w} \right)^2 \hat{P}(1-\hat{P})$$

In practice we plug in **0.5** (or an **educated guess**) for \hat{P} .

- Using 0.5 for \hat{P} is conservative (leads to a bigger sample size than is actually needed) because $\hat{P}(1-\hat{P})$ is maximized between 0 and 1 when $\hat{P} = 0.5$.

Nels Grevstad

Notes

Notes

Notes

Notes

Example

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

We want the **width** of a **95% confidence interval for p** to be **0.06** irrespective of the value of \hat{P} .

How big should n be?

Nels Grevstad

We'd need

$$n = \left(\frac{2 \cdot 1.96}{0.06} \right)^2 (0.5)(1 - 0.5) = 1,067.1$$

which we **round up** to $n = 1,068$.

Nels Grevstad

Notes

Notes

Notes

Notes
