Confidence Intervals for Proportions

“Far better an approximate answer to the right question, . . . than an exact answer to the wrong question.”

John Tukey

“The truth is seldom pure and never simple.”

Oscar Wilde

“They make things admirably plain, but one hard question will remain: If one hypothesis you lose, Another in its place you choose . . .”

James Russell Lowell, Credidimus Jovem Regnare

IN THE REAL WORLD  On the CNN evening news Wolfe Blitzer reports that, based on a CNN survey of registered voters, 52% of voters would vote for a particular candidate if the election were held today. He concludes the report with the statement that the estimate is “accurate to within 2.5 percentage points.” What does this statement concerning the accuracy of their survey results really mean?

Unit Objectives

At the conclusion of this chapter you will be able to:

☐ Estimate the unknown value of a population proportion and population mean by constructing confidence intervals based on the information contained in a single sample.

Reading Assignment

Chapter 19

Highlights from the Readings

OVERVIEW

→ → TO THIS POINT:
In the preceding chapters we learned that populations are characterized by numerical descriptive measures called parameters such as the (population) mean $\mu$, the (population) standard deviation $\sigma$, or a (population) proportion $p$. The value of a particular population descriptive measure is typically unknown; a reliable estimate of one of these unknown values is frequently needed. NASA may be very interested in an estimate of the reliability of a crucial component in the space shuttle; a candidate for an elective office may need an estimate of the proportion of voters that will vote for him/her to plan campaign strategy. Estimates of these unknown values are based on sample statistics computed from sample data. Since sample statistics vary in a random manner from sample to sample, estimates based on them will be subject to uncertainty. This uncertainty is reflected in the sampling distribution of a statistic.

WHAT'S NEXT → → :
This chapter puts the preceding material into practice. We estimate proportions based on the information in a single sample selected from the population of interest. Then, in a procedure that distinguishes statistics from mere guesswork, we use the sample distribution of a sample statistic to obtain a measure of reliability of our estimate.
Introduction

Statistical inference is the process by which we acquire information about populations from samples.

There are two procedures for making inferences:

i) Estimation
ii) Hypotheses testing

The objective of estimation is to estimate the unknown value of a population parameter on the basis of a sample statistic.

The objective of hypothesis testing is to weigh the evidence in the data against a null hypothesis $H_0$.

Estimation

There are two types of estimators:

i) Point Estimators
ii) Interval estimators

Point Estimator

A single value (or point) used to estimate the unknown value of a population parameter

What population parameters do we frequently need to estimate?

i) an unknown population proportion $p$
ii) an unknown population mean $\mu$

What are the best point estimators of the population proportion $p$?

• $\hat{p} = \frac{x}{n}$, the sample proportion of $x$ successes in a sample of size $n$, is the best point estimate of the population proportion $p$

Example: (Estimating an unknown population proportion $p$)

In a Rasmussen Reports poll on July 26-27, 2011, from a national random sample of 1,000 adults, 460 responded that they thought that most members of Congress are crooks.

$\hat{p} = \frac{460}{1000} = .46$ is the point estimate of the unknown population proportion $p$ of all adults that think most members of Congress are crooks.

⇒ Shortcoming of Point Estimates ⇐

• $\hat{p} = \frac{366}{610} = .60$, best estimate of population proportion $p$

BUT.....

How good is this best estimate?

No measure of reliability

Interval Estimator

A confidence interval is a range (or an interval) of values used to estimate the unknown value of a population parameter.
Confidence Intervals for a Population Proportion $p$

Interval estimators are constructed from *SAMPLING DISTRIBUTION MODELS*.

### Confidence Interval for a Population Proportion $p$

Recall that the best point estimator for a population proportion $p$ is the sample proportion $\hat{p} = \frac{x}{n}$.

The sampling distribution of $\hat{p} = \frac{x}{n}$ has the following characteristics:

$$E(\hat{p}) = p, \quad SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Recall the basic idea of the Central Limit Theorem:

the sampling distribution of the sample proportion $\hat{p}$ is approximately normal for large samples ($np \geq 10$ and $n(1-p) \geq 10$)

Therefore,

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

By the Central Limit Theorem, the sampling distribution model of $\hat{p}$ is normal.

Can use the above facts to estimate the unknown value of the population proportion $p$ with an interval.

Since $\hat{p} \sim N \left( p, \sqrt{\frac{p(1-p)}{n}} \right)$, we can write

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = Z \quad \text{(standard normal r. v.)}$$
From the standard normal table,

\[ .95 = P(-1.96 \leq z \leq 1.96) \]

\[ = P \left( -1.96 \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96 \right) \]

\[ = P \left( \hat{p} - 1.96 \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96 \sqrt{\frac{p(1-p)}{n}} \right) \]

We have constructed a 95% confidence interval for \( p \):

\( \left( \hat{p} - 1.96 \sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96 \sqrt{\frac{p(1-p)}{n}} \right) \) which is usually written \( \hat{p} \pm 1.96 \sqrt{\frac{p(1-p)}{n}} \)

BUT...

we cannot calculate the above interval because we DO NOT KNOW THE VALUE OF \( p \).

So we use the estimate \( \hat{p} \) for \( p \) and calculate \( \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \)

A 95% confidence interval for a population proportion \( p \) can be calculated as follows:

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

Note that the confidence interval is

\[ \text{point estimate} \pm 1.96 \times \text{SE}(\hat{p}) \]

MARGIN OF ERROR

The extent of the interval on either side of \( \hat{p} \) is called the margin of error \( ME \)

We'll use the same approach for many other confidence intervals besides those for \( \hat{p} \). In general, confidence intervals look like

\[ \text{point estimate} \pm ME \]

EXAMPLE (Gallup polls)

Voter preference polls typically sample approximately 1600 voters; suppose \( \hat{p} = .52 \).

Then if we desire a 95% confidence interval for \( p \) we calculate...
Confidence intervals with confidence levels other than 95%

98% Confidence Interval For $p$

$$\hat{p} \pm 2.33 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Four Commonly Used Confidence Levels and Corresponding Critical Value

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>Critical $z$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>.90</td>
<td>1.645</td>
</tr>
<tr>
<td>.95</td>
<td>1.96</td>
</tr>
<tr>
<td>.98</td>
<td>2.33</td>
</tr>
<tr>
<td>.99</td>
<td>2.576</td>
</tr>
</tbody>
</table>

Confidence Intervals for $p$

- 90% $\hat{p} \pm 1.645 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- 95% $\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- 98% $\hat{p} \pm 2.33 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- 99% $\hat{p} \pm 2.576 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

**EXAMPLE** (Medication side effects)

Arthritis is a painful, chronic inflammation of the joints. An experiment on the side effects of pain relievers examined arthritis patients to find the proportion of patients who suffer side effects.

DATA: 440 subjects with chronic arthritis were given ibuprofen for pain relief; 23 subjects suffered from adverse side effects.

Calculate a 90% confidence interval for the population proportion $p$ of arthritis patients who suffer some “adverse symptoms.”

Confidence interval for $p$: $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$; $\hat{p} = \bar{X}$ = $1 - \hat{p}$ = $\bar{X}$ for 90% confidence,

$z^* = $

We are 90% confident that the interval (.034, .070) contains the true proportion of arthritis patients that experience some adverse symptoms when taking ibuprofen.
Impact of sample size

EXAMPLE
Recall the original example: \( n = 440, \hat{p} = 0.052; \)
90% confidence interval: \( 0.052 \pm 1.645 \sqrt{\frac{0.052(1-0.052)}{440}} = 0.052 \pm 0.018 \Rightarrow (0.034, 0.070) \)

Suppose that the same proportion \( \hat{p} = 0.052 \) had come from a sample of size \( n = 1000 \)
90% confidence interval: \( 0.052 \pm 1.645 \sqrt{\frac{0.052(1-0.052)}{1000}} = 0.052 \pm 0.007 \Rightarrow (0.045, 0.059) \)

\( n = 440 \) : width of 90% interval: \( 2 \times 0.018 = 0.036; \)
\( n = 1000 \) : width of 90% interval: \( 2 \times 0.007 = 0.014 \)

When the sample size is increased, the 90% confidence interval is narrower.

SUMMARY:
1. The higher the confidence level, the wider the interval.
2. Increasing the sample size \( n \) will make a confidence interval with the same confidence coefficient narrower (i.e., more precise).

To calculate confidence intervals for \( p \) using EXCEL, see our class web page http://www.stat.ncsu.edu/people/reland/courses/st305/, click on Lecture Handouts in left column, and in the Chapter 19 material click on the file “Excel Spreadsheet for Calculating Confidence Intervals for \( p \) and \( 1-p^2 \); StatCrunch, go to http://statcrunch.stat.ncsu.edu, click on Stat > Proportions > One sample; TI 83/84: see Calculator Appendix, p. 14.

EXAMPLE. A random sample of 1000 NCSU students found that 50 students strongly favored the current lottery system for awarding tickets to football and men's basketball games. Find a 95% confidence interval for \( p \), the proportion of NCSU students that strongly favor the current lottery system.

SOLUTION \( \hat{p} = \frac{50}{1000} = 0.05 \), so \( \hat{q} = 0.95 \) and the confidence interval is
\[
0.05 \pm 1.96 \sqrt{\frac{0.05(0.95)}{1000}} =
\]

IMPORTANT: INTERPRETING CONFIDENCE INTERVALS

- **Previous example**: \( 0.05 \pm 0.014 \Rightarrow (0.036, 0.064) \)
- **Correct Interpretation**
  We are 95% confident that the interval from 0.036 to 0.064 actually does contain the true value of \( p \).
  This means that if we were to select many different samples of size 1000 and construct a 95% CI from each sample, 95% of the resulting intervals would contain the value of the population proportion \( p \). (.036, .064) is one such interval. (Note that 95% refers to the procedure we used to construct the interval; it does not refer to the population proportion \( p \)).

- **Wrong**
  There is a 95% chance that the population proportion \( p \) falls between .036 and .064. (Note that \( p \) is not random, it is a fixed but unknown number)
Determining Sample Size

**TO ESTIMATE A POPULATION PROPORTION \( p \):**

**Question:**
If you want to estimate a population proportion \( p \) with a \( C\% \) confidence interval with margin of error \( ME \) on the error of estimation, how large does the sample size \( n \) need to be?

In terms of the margin of error \( ME \), the confidence interval for \( p \) can be expressed as
\[ \hat{p} \pm ME; \]

The confidence interval for \( p \) is
\[ \hat{p} \pm z^* \sqrt{\frac{p(1-p)}{n}} \]
where \( z^* \) is the critical value for the specified confidence level \( C \).

Therefore, set
\[ z^* \sqrt{\frac{p(1-p)}{n}} = ME \]
and solve for the sample size \( n \):

\[ n = \frac{(z^*)^2 p(1-p)}{(ME)^2} \]

However, we don’t know \( p \); if prior information is available concerning the value of \( p \), use that estimate to calculate \( n \); if no prior information about \( p \) is available, to obtain a conservative estimate of the required sample size, use \( p = \frac{1}{2} \).

**EXAMPLES**

1. **Estimating proportion of crimes that use firearms.**
The U. S. Crime Commission wants to estimate \( p = \) the proportion of crimes in which firearms are used to within .02 with 90% confidence. Data from previous years shows that \( p \) is about .6.

\[ n = \frac{(z^*)^2 p(1-p)}{(ME)^2}; \ ME = .02; p \text{ estimated from previous data to be approximately } .6; 90\% \]

\[ CI \Rightarrow z^* = 1.645. \]

\[ n = \]

**EXAMPLES** (continued)

2. **Estimating proportion of customers that will purchase a new product.**
The Curdle Dairy Co. wants to estimate the proportion \( p \) of customers that will purchase its new broccoli-flavored ice cream. Curdle wants to be 90% confident that they have estimated \( p \) to within .03. How many customers should they sample?
METHOD 1: no knowledge concerning value of $p$
Let $p = .5$. This results in the largest $n$ needed for a 90% confidence interval of the form $\hat{p} \pm .03$.
If the proportion $p$ does not equal .5, the actual margin of error ME will be narrower than .03 using the sample size $n$ obtained by this formula:

$$n = \frac{z^2 + ps(1-p)}{(ME)^2}; \quad ME = .03; \quad 90\% \text{ CI } \Rightarrow z^* = 1.645 \quad \Rightarrow \quad n = \frac{(1.645)^2 + p(.5)(.5)}{(.03)^2} = 751.67 \uparrow 752$$

METHOD 2: knowledge of approximate $p$.
Suppose from previous marketing data
$p \approx .2$.
$$n = \frac{(1.645)^2 + .2(.8)}{(.03)^2} = 481.07 \uparrow 482.$$