Chapter 3

What Do
Samples Tell Us?

Lecture Slides

Case Study:
What Do Samples Tell Us? 1

According to the Centers for Disease Control and Prevention (CDC), there were 173 cases of measles reported between June 1 and May 29, 2015.

About 87% of the cases were related to 5 outbreaks during the same time period.

Blend Images/REB Images/Getty Images

Case Study: What Do Samples Tell Us? 2

The CDC also reported that the “United States experienced a record number of measles cases during 2014, with 668 cases from 27 states reported to CDC’s National Center for Immunization and Respiratory Diseases (NCIRD).”

Case Study:
What Do Samples Tell Us? 3

This is the greatest number of cases since measles elimination was documented in the United States in 2000.

According to the same report by the CDC, “the majority of people who got measles were unvaccinated.”

Vaccinating children against diseases like measles is controversial.

Case Study:
What Do Samples Tell Us? 4

A Gallup Poll conducted February 28–March 1, 2015, asked the following question:

“How important is it that parents get their children vaccinated—extremely important, very important, somewhat important, or not at all important?”

They found that 54% of respondents said, “extremely important” (down from 64% who responded to a similar Gallup Poll in 2001).

Can we trust this conclusion?

Case Study:
What Do Samples Tell Us? 5

Reading further, we find that Gallup talked with 1015 randomly selected adults to reach these conclusions.

The U.S. Census Bureau said that there were about 258 million adults in the United States in 2013.

How can 1015 people tell us about the opinions of 258 million people?

Is the 54% who feel that it is extremely important, in fact, the majority of Americans who feel this way? By the end of this chapter, you will learn the answers to these questions.

Parameters and Statistics

A ***parameter*** is a number that describes the ***population***.

- A parameter is a fixed number, but in practice we don’t know its value.

A ***statistic*** is a number that describes the ***sample***.

- The value of a statistic is known when we have taken a sample, but it can change from sample to sample.

We often use a statistic to estimate an unknown parameter.

Proportions 1

Now we'll talk about a specific parameter and statistic—using $\widehat{p}$ to estimate *p*.

$\hat{p}$ is the sample proportion (statistic) who have the trait/opinion of interest.
p is the population proportion (parameter) who have the trait/opinion of interest.

Proportions 2

A Columbia-based health club wants to estimate the proportion of Columbia residents who enjoy running. Let p = proportion of all Columbia residents who enjoy running

We decide to take an SRS of n = 100 Columbia residents.

$\hat{p}$ = proportion of residents in our sample who enjoy running.

Proportions 3

In our SRS of *n* = 100 Columbia residents, 17 said that they enjoy running. The sample proportion is

$$\hat{p} = \frac{17}{100} = 0.17 = 17\%$$

Suppose now that I take another SRS of Columbia residents of size *n* = 100 and 22 of them said that they enjoy running. Find p-hat.

$$\hat{p} = \frac{22}{100} = 0.22 = \ 22\%$$

Sampling Variability 1

Notice that we have two samples from the same population and our sample proportions are different from each other.

Question: Are statistics from the different samples (but drawn from the same population) going to be exactly the same? **No!**

Question: Is our statistic, calculated from any one sample, going to exactly match the population parameter it is attempting to estimate? **No!**

Sampling Variability 2

The fact that our statistics will not be the same from sample to sample is called **sampling variability** (because all samples are going to be a little different from each other).

We hope that as long as we have a good sampling scheme, we will be estimating the population parameter fairly well when we take our one shot at estimating it.

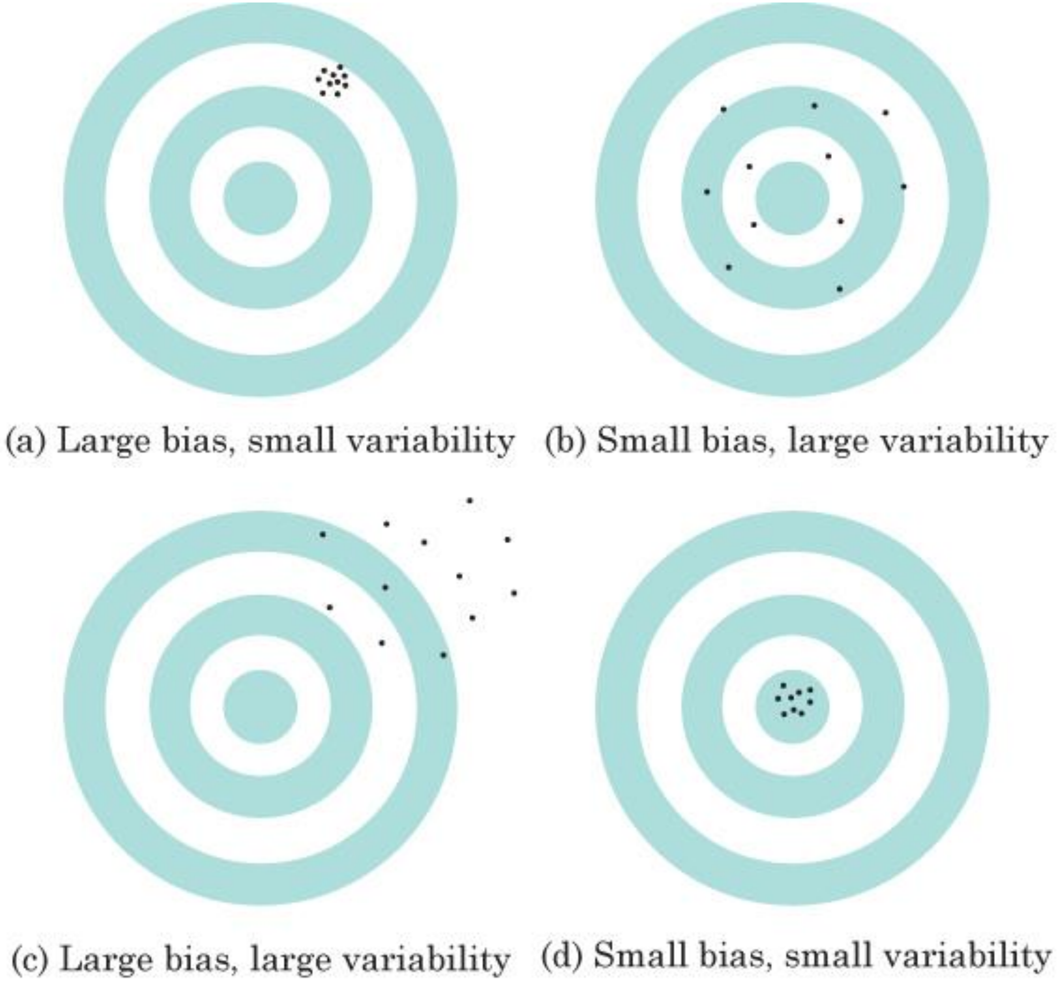
Bias and Variability 1

Bias is consistent, repeated deviation of the statistic from the parameter in the same direction when we take many samples.

Variability describes how spread out the values of the statistic are when we take many samples.

Large variability means the result of sampling is not repeatable. A good sampling method has small bias and small variability.

Bias and Variability 2



(a) Large bias, small variability

(b) Small bias, large variability

(c) Large bias, large variability

(d) Small bias, small variability

Figure 3.3

Moore/Notz, *Statistics: Concepts and Controversies*, 9e, © 2017 W. H. Freeman and Company

Variability of p-hat 1

1000 of size n = 100

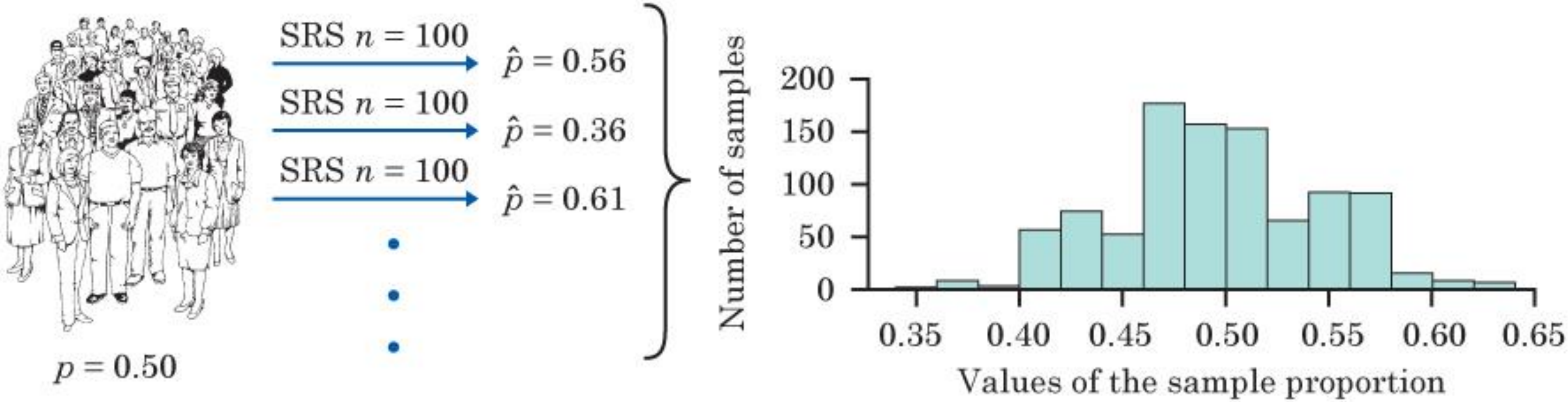


Figure 3.1
Moore/Notz, *Statistics: Concepts and Controversies*, 9e, © 2017 W. H. Freeman and Company

Variability of p-hat 2

1000 of size n = 1015

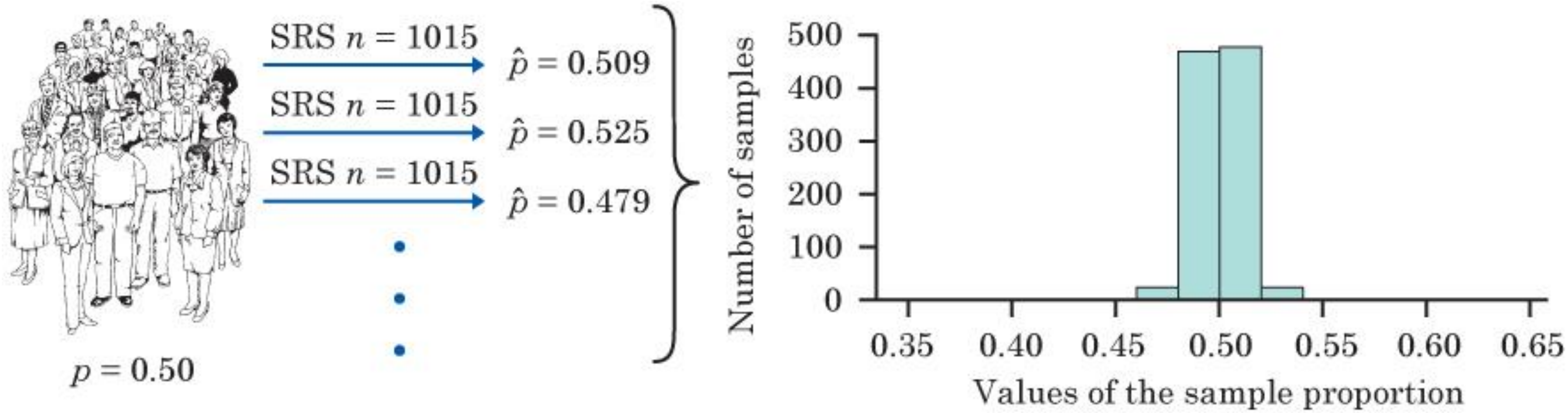


Figure 3.2
Moore/Notz, *Statistics: Concepts and Controversies*, 9e, © 2017 W. H. Freeman and Company

Notice that with **larger samples** (1015 vs. 100), there is a lot **less variability,** but the distribution is still centered at p = 0.50 (so **p-hat is unbiased for p).**

Reducing Bias and Variability

To reduce bias, use random sampling.

To reduce variability of your statistic when sampling with an SRS, use a larger sample. You can make the variability as small as you want by taking a large enough sample.

Large random samples almost always give an estimate that is close to the truth (population parameter).

Example: Smartphone Usage

Source: news.gallup.com, April 17–May 18, 2015, in the report *Most U.S. Smartphone Owners Check Phone at Least Hourly,* stated that 52% ± 1% of American smartphone owners check their devices several times an hour or more frequently.

Where does the plus or minus 1% come from?

This is called margin of error (MOE).

Margin of Error

The margin of error (MOE) is a value that quantifies the uncertainty in our estimate.

When using the sample proportion to estimate the population proportion, the MOE is a measure of how close we believe the sample proportion is to the population proportion. We usually report this through a confidence interval. (More on this coming soon.)

Calculating Margin of Error

Use the sample proportion, $\hat{p}$, from an SRS of size *n* to estimate an unknown population proportion *p*.

For 95% confidence:

$$\text{MOE} \approx \frac{1}{\sqrt{\text{n}}}$$

We'll refine this approximate MOE formula to something a little more precise in Chapter 21.

Example: Margin of Error

A CNN Poll interviewed 1000 people. What is the margin of error for 95% confidence?

$$MOE \approx \frac{1}{\sqrt{1000}} = \frac{1}{31.6228} = 0.0316 = 3.16\%$$

If the sample size is 100, what is the margin of error for 95% confidence?

$$MOE \approx \frac{1}{\sqrt{100}} = \frac{1}{10} = 0.10 = 10\%$$

MOE: What Is It?

“Margin of error plus or minus 4 percentage points” is shorthand for this statement:

If we took many samples using the same method we used to get this one sample, 95% of the samples would give a result within plus or minus 4 percentage points of the truth about the population.

Confidence Interval

Use MOE to calculate an interval that we think includes the parameter form for most confidence intervals:

$$\text{estimate} \pm \text{MOE}$$

Approximately 95% confidence interval for *p*

$$\hat{p} \pm \frac{1}{\sqrt{\text{n}}}$$

Confidence Statements

A ***confidence statement*** interprets a confidence interval and has two parts: a ***margin of error*** and a ***level of confidence***.

Margin of error says how close the statistic lies to the parameter.

Level of confidence says what percentage of all possible samples results in a confidence interval which contains the true parameter.

Example: Smartphone Usage (continued)

52% plus or minus 1% of American smartphone owners check their devices several times an hour or more frequently.

Compute and interpret the confidence interval for $\hat{p}$ = 52% with MOE 1%.

CI is 52% ± 1%,

so 51% < *p* < 53%.

We are 95% confident that the percent of all American smartphone owners who check their devices several times an hour or more frequently is between 51% and 53%.

-OR-

We are 95% confident that between 51% and 53% of all American smartphone owners check their devices several times an hour or more frequently.

Example: Higher Education in the United States

http://pewsocialtrends.org/2011/05/15/is-college-worth-it/

This May 2011 Pew Research survey finds that 57% of the 2142 adult Americans polled think that “the higher education system in the United States fails to provide students good value for the money they and their families spend.” Using the quick formula for MOE, **compute and interpret** a 95% confidence interval for *p*.

$\hat{p}$ **= 57% with MOE =** $\frac{1}{\sqrt{2142}} = 0.022 \xrightarrow{\text{yields}} 2.2\%$

CI is computed using 57% ± 2.2%,

so confidence interval is 54.8% < p **< 59.2%.**

We are 95% confident that between 54.8 and 59.2% of all adult Americans think the higher education system in the United States fails to provide students good value for the money they and their families spend.

Population Size Doesn’t Matter

The variability of a statistic from an SRS does not depend on the size of the population as long as the population is at least 100 times larger than the sample.

Suppose we take a sample of size 2527 from a population of 300,000. Then we take a sample of 2527 from a population of 1,000,000. Which sample statistic has more variability?

Neither. The sample sizes are the same. Population size does not affect the variability of the statistic.

Statistics in Summary 1

The purpose of sampling is to use a sample to gain information about a population. We often use a sample **statistic** to estimate the value of a population **parameter**.

One big idea: To describe how trustworthy a sample is, ask, “What would happen if we took a large number of samples from the same population?” If almost all samples would give a result close to the truth, we can trust our one sample even though we can’t be certain that it is close to the truth.

Statistics in Summary 2

In planning a sample survey, first aim for small **bias** by using random sampling and avoiding bad sampling methods such as voluntary response. Next, choose a large enough random sample to reduce the **variability** of the result. Using a large random sample guarantees that almost all samples will give accurate results.

Statistics in Summary 3

To say how accurate our conclusions about the population are, make a **confidence statement**. News reports often mention only the margin of error. Most often this **margin of error** is for **95% confidence**. That is, if we chose many samples, the truth about the population would be within the margin of error 95% of the time.

Statistics in Summary 4

We can roughly approximate the margin of error for 95% confidence based on a simple random sample of size *n* by the formula 1/√n. As this formula suggests, only the size of the sample, not the size of the population, matters. This is true as long as the population is much larger (at least 20 times larger) than the sample.