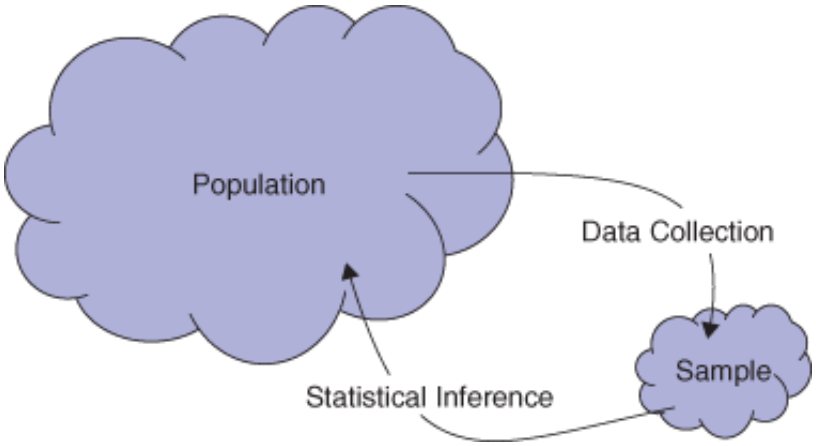
Print this page

3.1 SAMPLING DISTRIBUTIONS

In Chapter 1 we discuss data collection: methods for obtaining sample data from a population of interest. In this chapter we begin the process of going in the other direction: using the information in the sample to understand what might be true about the entire population. If all we see are the data in the sample, what conclusions can we draw about the population? How sure are we about the accuracy of those conclusions? Recall from Chapter 1 that this process is known as *statistical inference*.

Statistical Inference

Statistical inference is the process of drawing conclusions about the entire population based on the information in a sample.



Statistical inference uses sample data to understand a population

Population Parameters and Sample Statistics

To help identify whether we are working with the entire population or just a sample, we use the term *parameter* to identify a quantity measured for the population and *statistic* for a quantity measured for a sample.

Parameters vs Statistics

A **parameter** is a number that describes some aspect of a population.

A **statistic** is a number that is computed from the data in a sample.

As we saw in Chapter 2, although the name (such as “mean” or “proportion”) for a statistic and parameter is generally the same, we often use different notation to distinguish the two. For example, we use μ (mu) as a parameter to denote the mean for a population and $\bar{x}$ as a statistic for the mean of a

sample. Table 3.1 summarizes common notation for some population parameters and corresponding sample statistics. The notation for each should look familiar from Chapter 2.

Table 3.1 ***Notation for common parameters and statistics***

	Population Parameter	Sample Statistic
Mean	μ	$\bar{x}$
Standard deviation	σ	s
Proportion	p	$\hat{p}$
Correlation	ρ	r
Slope (regression)	β	b

Example 3.1

Proportion of College Graduates

The US Census states that 27.5% of US adults who are at least 25 years old have a college bachelor's degree or higher. Suppose that in a random sample of n=200 US residents who are 25 or older, 58 of them have a college bachelor's degree or higher. What is the population parameter? What is the sample statistic? Use correct notation for each answer.

Solution

The population parameter is the proportion with a bachelor's degree for *all* US adults who are at least 25 years old; it is p=0.275. The sample statistic is the proportion with a bachelor's degree for all people in the sample; it is $\hat{p} = 58 / 200 = 0.29$.

Sample Statistics as Point Estimates of Population Parameters

On April 29, 2011, Prince William married Kate Middleton (now Duchess Catherine) in London. The Pew Research Center reports that 34% of US adults watched some or all of the royal wedding.[1] Pew Research Center, "Too Much Coverage: Birth Certificate, Royal Wedding," http://www.pewresearch.org, May 3, 2011. How do we know that 34% of all US adults watched? Did anyone ask *you* if you watched it? In order to know for sure what proportion of US adults watched the wedding, we would need to ask *all* US adults whether or not they watched. This would be very difficult to do. As we will see, however, we can estimate the population parameter quite accurately with a sample statistic, as long as we use a random sample (as discussed in Chapter 1). In the case of the royal wedding, the estimate is based on a poll using a random sample of 1006 US adults.

In general, to answer a question about a population parameter *exactly*, we need to collect data from every individual in the population and then compute the quantity of interest. That is not feasible in most

settings. Instead, we can select a sample from the population, calculate the quantity of interest for the sample, and use this sample statistic to estimate the value for the whole population.

The value of a statistic for a particular sample gives a *point estimate* (or simply *estimate*) of the population parameter. If we only have the one sample and don't know the value of the population parameter, this point estimate is our best estimate of the true value of the population parameter.

Point Estimate

We use the statistic from a sample as a **point estimate** for a population parameter.

Example 3.2

Fuel economy information for cars is determined by the EPA (Environmental Protection Agency) by testing a sample of cars.[2]http://www.epa.gov/fueleconomy/data.htm. Based on a sample of $n=12$ Toyota Prius cars in 2012, the average fuel economy was 48.3 mpg. State the population and parameter of interest. Use the information from the sample to give the best estimate of the population parameter.

Solution

The population is all Toyota Prius cars manufactured in 2012. The population parameter of interest is μ, the mean fuel economy (mpg) for all 2012 Toyota Prius cars. For this sample, $\bar{x} = 48.3$. Unless we have additional information, the best point estimate of the population parameter is the sample statistic of 48.3. Notice that to find μ exactly, we would have to obtain information on the fuel economy for *every* 2012 Toyota Prius.

Example 3.3

For each of the questions below, identify the population parameter(s) of interest and the sample statistic we might use to estimate the parameter.

(a) What is the mean commute time for workers in a particular city?

(b) What is the correlation between the size of dinner bills and the size of tips at a restaurant?

(c) How much difference is there in the proportion of 30 to 39-year-old US residents who have only a cell phone (no land line phone) compared to 50 to 59-year-olds in the US?

Solution

(a) The relevant parameter is μ, the mean commute time for all people who work in the city. We estimate it using $\bar{x}$, the mean from a random sample of people who work in the city.

(b) The relevant parameter is ρ, the correlation between the bill amount and tip size for all dinner checks at the restaurant. We estimate it using r, the correlation from a random sample of dinner checks.

(c) The relevant quantity is $p_1 - p_2$, the difference in the proportion of all 30 to 39-year-old US residents with only a cell phone (p_1) and the proportion with the same property among all 50 to 59-year-olds

(p_2). We estimate it using $\hat{p}_1 - \hat{p}_2$, the difference in sample proportions computed from random samples taken in each age group.

Practice Problems 3.1A

Variability of Sample Statistics

We usually think of a parameter as a fixed value[3]In reality, a population may not be static and the value of a parameter might change slightly, for example, if a new person moves into a city. We assume that such changes are negligible when measuring a quantity for the entire population. while the sample statistic varies from sample to sample, depending on which cases are selected to be in the sample. We would like to know the value of the population parameter, but this usually cannot be computed directly because it is often very difficult or impossible to collect data from every member of the population. The sample statistic might vary depending on the sample, but at least we can compute its value.

In Example 3.3, we describe several situations where we might use a sample statistic to estimate a population parameter. How accurate can we expect the estimates to be? That is one of the fundamental questions of statistical inference. Because the value of the parameter is usually fixed but unknown, while the value of the statistic is known but varies depending on the sample, the key to addressing this question is to understand how the value of the sample statistic varies from sample to sample.

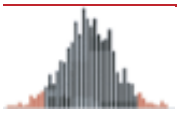
Consider the average fuel economy for 2012 Toyota Prius cars in Example 3.2. The average observed in the sample is $\bar{x} = 48.3$. Now suppose we were to take a new random sample of n=12 cars and calculate the sample average. A new sample average of $\bar{x} = 48.2$ (very close to 48.3!) would suggest low variability in the statistic from sample to sample, suggesting the original estimate of 48.3 is pretty accurate. However, a new sample average of 56.8 (pretty far from 48.3) would suggest high variability in the statistic from sample to sample, giving a large amount of uncertainty surrounding the original estimate.

Of course, it's hard to judge variability accurately from just two sample means. To get a better estimate of the variability in the means we should consider many more samples. One way to do this is to use computer simulations of samples from a known population, as illustrated in the following examples.

DATA 3.1 **Enrollment in Graduate Programs in Statistics**

Is a statistics graduate program in your future?

Graduate programs in statistics often pay their graduate students, which means that many graduate students in statistics are able to attend graduate school tuition free with an assistantship or fellowship. (This is one of the many wonderful things about studying statistics!) There are 82 US statistics doctoral programs (departments of statistics or biostatistics in the US reporting a PhD program) for which enrollment data were available.[4]Full list of the 82 Group IV departments was obtained at http://www.ams.org/profession/data/annual-survey/group_iv. Data on enrollment obtained primarily from *Assistantships and Graduate Fellowships in the Mathematical Sciences*, 2009, American Mathematical Society. The list does not include combined departments of mathematics and statistics and does not include departments that did not reply to the AMS survey. The dataset **StatisticsPhD** lists all these schools together with the total enrollment of full-time graduate students in each program in 2009.

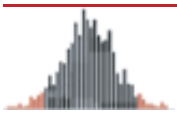


Example 3.4

What is the average full-time graduate student enrollment in US statistics doctoral programs in 2009? Use the correct notation for your answer.

Solution

Based on the data **StatisticsPhD**, the mean enrollment in 2009 is 53.54 full-time graduate students. Because this is the mean for the entire population of all US statistics doctoral programs for which data were available that year, we have μ=53.54 students.



Example 3.5

Select a random sample of 10 US statistics doctoral programs from the file **Statistics PhD** and compute the mean enrollment for your sample. Use the correct notation for your answer. Compare your answer to the population mean.

Solution

We use *StatKey*, other technology, or a random number table to select the sample. One sample is given in Table 3.2. We compute the mean enrollment for this sample to be $\bar{x} = 53.0$. The sample mean is similar to the population mean of μ=53.54 but is not exactly the same.

Table 3.2 ***A sample of full-time graduate student enrollment in 2009***

University	Department	Enrollment
Case Western Reserve University	Statistics	11
University of South Carolina	Biostatistics	45
Harvard University	Statistics	67
University of California-Riverside	Statistics	54
Medical University of South Carolina	Biostatistics	46
University of Nebraska	Statistics	44
New York University	Statistics	6
Columbia University	Statistics	196
University of Iowa	Biostatistics	35
Baylor University	Statistics	26

That the sample statistic in Example 3.5 does not exactly match the population parameter is not surprising: We don't expect to get exactly the mean of the entire population from every sample we choose, but we hope that our sample mean is close to the population mean.

A different random sample of 10 US statistics doctoral programs is shown in Table 3.3. The mean enrollment for this sample is $\bar{x} = 61.5$. Again, the sample mean is somewhat similar to the population mean of μ=53.54 but is not exactly the same. Note that this sample mean is also different from the sample mean found from the sample in Table 3.2.

Table 3.3 ***Another sample of full-time graduate student enrollment in 2009***

University	Department	Enrollment
University of Wisconsin	Statistics	116

Cornell University	Statistics	78
Yale University	Statistics	36
Iowa State University	Statistics	145
Boston University	Biostatistics	39
University of Nebraska	Statistics	44
University of Minnesota	Biostatistics	48
University of California-Los Angeles	Biostatistics	60
University of California-Davis	Statistics	34
Virginia Commonwealth University	Statistics	15

If everyone in your statistics class selects a random sample of size 10 from the population of US statistics doctoral programs and computes the sample mean, there will be many different results. Try it! (In fact, from a population of size 82, there are 2,139,280,241,670 different samples of size 10 that can be selected!) We expect these sample means to be clustered around the true population mean of μ=53.54. To see that this is so, we use *StatKey* or other technology to take 1000 random samples of size n=10 from our population and compute the sample mean in each case. A dotplot of the results is shown in Figure 3.1. The sample means of $\bar{x} = 53.0$ and $\bar{x} = 61.5$ from the two random samples above correspond to two of the dots in this dotplot.

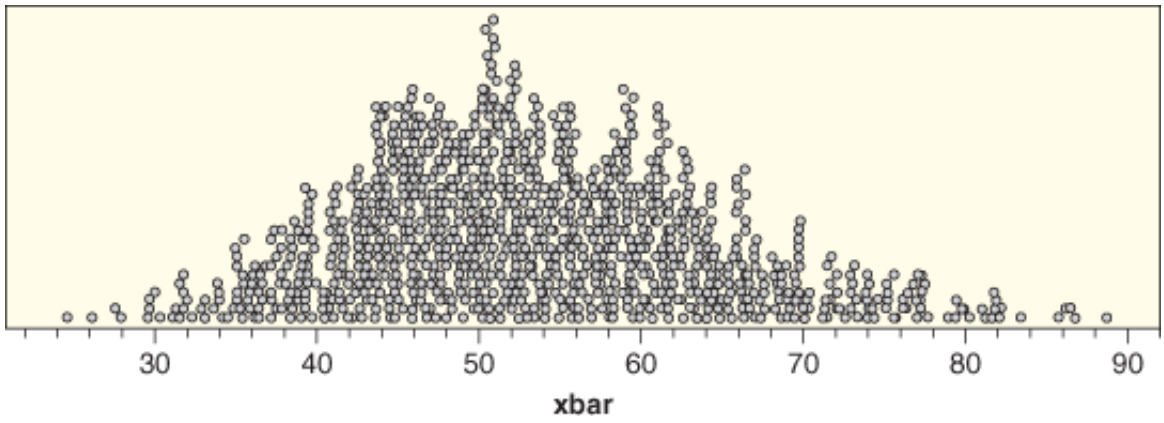


Figure 3.1 ***1000 means for samples of size n=10 from StatisticsPhD***

Notice in Figure 3.1 that we do indeed have many different values for the sample means, but the distribution of sample means is quite symmetric and centered approximately at the population mean of 53.54. From Figure 3.1 we see that most sample means for samples of size 10 fall between about 30 and 80. We will see that the bell-shaped curve seen in this distribution is very predictable. The distribution of sample statistics for many samples, such as those illustrated in Figure 3.1, is called a *sampling distribution*.

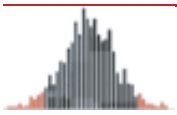
Sampling Distribution

A **sampling distribution** is the distribution of sample statistics computed for different samples of the same size from the same population.

A sampling distribution shows us how the sample statistic varies from sample to sample.

Figure 3.1 illustrates the sampling distribution for sample means based on samples of size 10 from the population of all statistics PhD programs. Of course, we don't show the means for all 2 trillion possible samples. However, the approximation based on 1000 samples should be sufficient to give us a good sense of the shape, center, and spread of the sampling distribution.

Sampling distributions apply to every statistic that we saw in Chapter 2 and lots more! We look next at a sampling distribution for a proportion.



Example 3.6

In Example 3.1 we see that 27.5% of US adults at least 25 years old have a college bachelor's degree or higher. Investigate the behavior of sample proportions from this population by using *StatKey* or other technology to simulate lots of random samples of size n=200 when the population proportion is p=0.275. Describe the shape, center, and spread of the distribution of sample proportions.

Solution

Figure 3.2 illustrates the sampling distribution of proportions for 1000 samples, each of size n=200 when p=0.275. We see that the sampling distribution of simulated $\hat{p}$ values is relatively symmetric, centered around the population proportion of p=0.275, ranges from about 0.175 to 0.38, and again has the shape of a bell-shaped curve. Note that the sample statistic $\hat{p} = 0.29$ mentioned in Example 3.1 is just one of the dots in this dotplot.

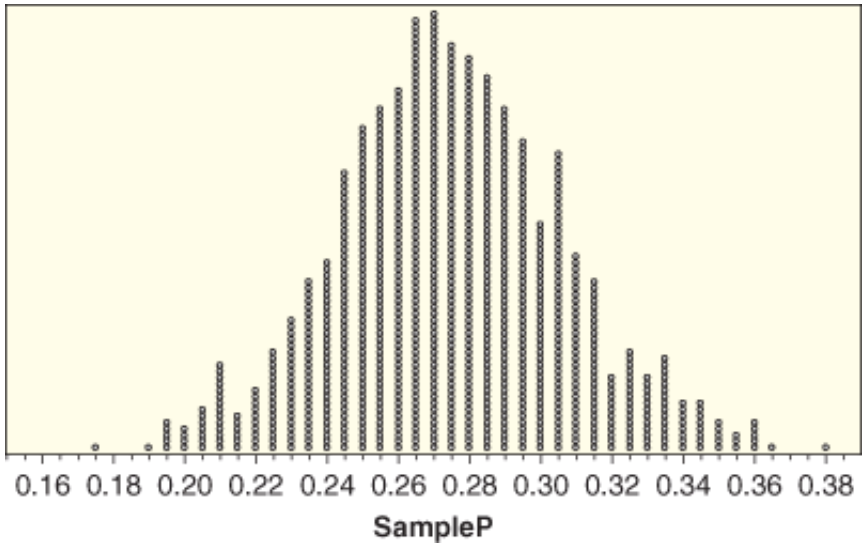


Figure 3.2 ***Sample proportions when n=200 and p=0.275***

The distributions of sample proportions in Figure 3.2 and sample means in Figure 3.1 have a similar shape. Both are symmetric, bell-shaped curves centered at the population parameter. As we will see, this is a very common pattern and can often be predicted with statistical theory. If samples are randomly selected and the sample size is large enough, the corresponding sample statistics will often have a symmetric, bell-shaped distribution centered at the value of the parameter. In later chapters we formalize the idea of a bell-shaped distribution and elaborate on how large a sample size is "large enough."

Shape and Center of a Sampling Distribution

For most of the parameters we consider, if samples are randomly selected and the sample size is large enough, the sampling distribution will be symmetric and bell-shaped and centered at the value of the population parameter.

Practice Problems 3.1B

Measuring Sampling Variability: The Standard Error

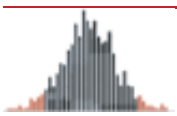
What we really care about is the *spread* of the sampling distribution (the variability of the statistic from sample to sample). Knowing how much a statistic varies from sample to sample is key in helping us know how accurate an estimate is.

One measure of variability associated with the sample statistic can be found by computing the standard deviation of the sample statistics in a sampling distribution. Although this is no different from the standard deviation of sample values we saw in Chapter 2, the standard deviation of a sample statistic is so important that it gets its own name: the *standard error* of the statistic. The different name helps to distinguish between the variability in the sample statistics and the variability among the values within a particular sample. We think of the standard error as a "typical" distance between the sample statistics and the population parameter.

Standard Error

The **standard error** of a statistic, denoted SE, is the standard deviation of the sample statistic.

In situations such as the mean graduate program enrollment in Example 3.5 and the proportion of college graduates in Example 3.6 where we can simulate values of a statistic for many samples from a population, we can estimate the standard error of the statistic by finding the usual standard deviation of the simulated statistics.



Example 3.7

Use *StatKey* or other technology to estimate the standard error for the sampling distributions of the following:

(a) Mean enrollment in statistics PhD programs in samples of size 10 (as in Example 3.5)

(b) Proportion of college graduates in samples of size 200 (as in Example 3.6)

Solution

The standard error is the standard deviation of all the simulated sample statistics. In *StatKey*, this standard deviation is given in the upper right corner of the box containing the sampling distribution. With other technology, once we have the sampling distribution we find the standard deviation of the values in the same way as in Chapter 2.

(a) For the 1000 means for simulated samples of $n=10$ statistics program enrollments shown in Figure 3.1, we find the standard deviation of the 1000 means to be 10.9 so we have $SE=10.9$.
(b) For the 1000 proportions of college graduates in simulated samples of size 200 shown in Figure 3.2, we find the standard deviation of the 1000 proportions to be 0.03, so we have $SE=0.03$.

Since these standard errors are estimated from a set of random simulations, the values might change slightly from one simulation to another.

Recall from Section 2.3 that when distributions are relatively symmetric and bell-shaped, the 95% rule tells us that approximately 95% of the data values fall within two standard deviations of the mean. Applying the 95% rule to sampling distributions, we see that about 95% of the sample statistics will fall within two standard errors of the mean. This allows us to get a rough estimate of the standard error directly from the dotplot of the sampling distribution, even if we don't have the individual values for each dot.

Example 3.8

Use the 95% rule to estimate the standard error for the following sampling distributions:

(a) Mean enrollment in statistics PhD programs in samples of size 10 (from Figure 3.1)

(b) Proportion of college graduates in samples of size 200 (from Figure 3.2)

Solution

(a) In Figure 3.1, we see that the middle 95% of sample means appear to range from about 34 to about 78. This should span about two standard errors below the mean and two standard errors above the mean. We estimate the standard error to be about $(78-34)/4=11$.
(b) In Figure 3.2, we see that the middle 95% of sample proportions appear to range from about 0.21 to 0.34, or about 0.065 above and below the mean of $p=0.275$. We estimate the standard error to be about $0.065/2=0.0325$.

These rough estimates from the graphs match what we calculated in Example 3.7.

A low standard error means statistics vary little from sample to sample, so we can be more certain that

our sample statistic is a reasonable point estimate. In Section 3.2, we will learn more about how to use the standard error to quantify the uncertainty in a point estimate.

Practice Problems 3.1C

The Importance of Sample Size

Example 3.9

In Example 3.1, we learn that the population proportion of college graduates in the US is $p=0.275$, and Figure 3.2 shows the sampling distribution for the sample proportion of college graduates when repeatedly taking samples of size $n=200$ from the population. How does this distribution change for other sample sizes? Figure 3.3 shows the distributions of sample proportions for many (simulated) random samples of size $n=50$, $n=200$, and $n=1000$. Discuss the effect of sample size on the center and variability of the distributions.

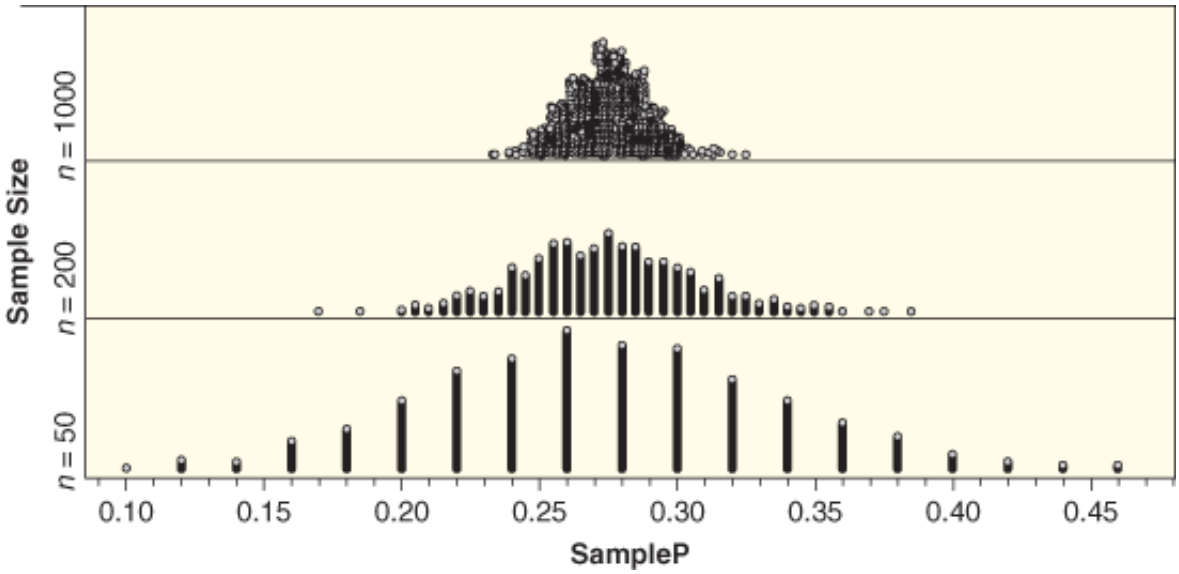


Figure 3.3 ***What effect does sample size have on the distributions?***

Solution

The center appears to be close to the population proportion of $p=0.275$ in all three distributions, but the variability is quite different. As the sample size increases, the variability decreases and a sample proportion is likely to be closer to the population proportion. In other words, as the sample size increases, the standard error decreases.

We see in Example 3.9 that the larger the sample size the lower the variability in the sampling distribution, so the smaller the standard error of the sample statistic. This makes sense: A larger sample allows us to collect more information and estimate a population parameter more precisely. If the sample were the entire population, then the sample statistic would match the population parameter exactly and the sampling distribution would be one stack of dots over a single value!

Sample Size Matters!

As the sample size increases, the variability of sample statistics tends to decrease and sample statistics tend to be closer to the true value of the population parameter.

Example 3.10

Here are five possible standard errors for proportions of college graduates using different size samples:

$$SE = 0.005 \quad SE = 0.014 \quad SE = 0.032 \quad SE = 0.063 \quad SE = 0.120$$

For each of the three sample sizes shown in Figure 3.3, use the 95% rule to choose the most appropriate standard error from the five options listed.

Solution

Since each of the distributions is centered near p=0.275, we consider the interval $0.275 \pm 2 \cdot SE$ and see which standard error gives an interval that contains about 95% of the distribution of simulated $\hat{p}$'s shown in Figure 3.3.

n=1000: It appears that SE=0.014 is the best choice, since the interval on either side of p=0.275 would go from $0.275 \pm 2(0.014)$ which is 0.247 to 0.303. This looks like a reasonable interval to contain the middle 95% of the values in the dotplot shown in the top panel of Figure 3.3, when the sample size is n=1000.

n=200: It appears that SE=0.032 is the best choice, since the interval on either side of p=0.275 would go from $0.275 \pm 2(0.032)$ which is 0.211 to 0.339. This looks like a reasonable interval to contain the middle 95% of the values in the dotplot shown in the middle panel of Figure 3.3, when the sample size is n=200.

n=50: It appears that SE=0.063 is the best choice, since the interval on either side of p=0.275 would go from $0.275 \pm 2(0.063)$ which is 0.149 to 0.401. This looks like a reasonable interval to contain the middle 95% of the values in the dotplot shown in the bottom panel of Figure 3.3, when the sample size is n=50.

The standard error of SE=0.005 is too small for any of these plots, and SE=0.120 would give an interval that is too large.

We see again in Example 3.10 that as the sample size increases, the standard error decreases, so the sample statistic generally becomes a better estimate of the population parameter.

Importance of Random Sampling

So far, the sampling distributions we have looked at have all been centered around the population parameter. It is important that samples were selected at random in each of these cases. Too often this is overlooked. Random sampling will generally yield a sampling distribution centered around the

population parameter, but, as we learned in Section 1.2, non-random samples may be biased, in which case the sampling distribution may be centered around the wrong value.

Example 3.11

Suppose that students at one college study, on average, 15 hours a week. Two different students, Judy and Mark, are curious about sampling. They each sample $n=50$ students many times, ask each student the number of hours they study a week, and record the mean of each sample. Judy takes many random samples of 50 students from the entire student body, while Mark takes many samples of 50 students by asking students in the library. The sampling distributions generated by Mark and Judy are shown with different colors in Figure 3.4. Which set of sample means (red or black) were produced by Judy? Why did Mark and Judy get such different results?

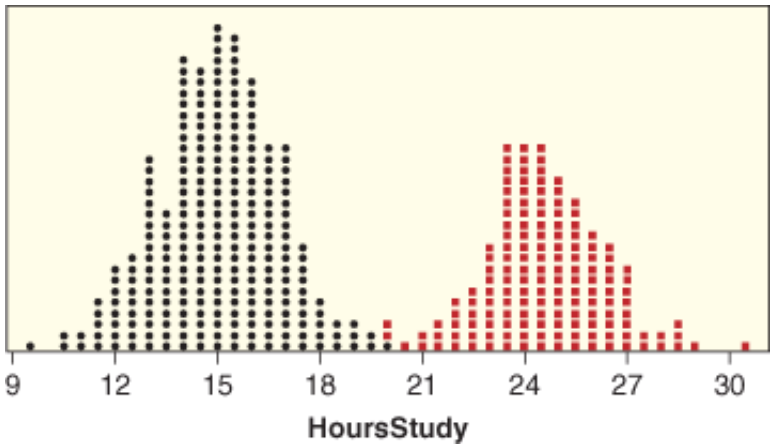


Figure 3.4 ***Sample means: Which color shows a biased sampling method?***

Solution

Judy was utilizing random sampling, so we expect her sample means to be centered around the true average of 15 hours a week. Therefore, we can conclude that her sample means correspond to the black dots. Mark chose to take a convenient sampling approach, rather than take a random sample. Due to this fact his samples are not representative of the population (students sampled in the library are likely to study more), so his sample means are biased to overestimate the average number of hours students study.

Inference Caution

Statistical inference is built on the assumption that samples are drawn randomly from a population. Collecting the sample data in a way that biases the results can lead to false conclusions about the population.

In this section, we've learned that statistics vary from sample to sample, and that a sample statistic can be used as a point estimate for an unknown fixed population parameter. However, a sample statistic will

usually not match the population parameter exactly, and a key question is how accurate we expect our estimate to be. We explore this by looking at many statistics computed from many samples of the same size from the same population, which together form a sampling distribution. The standard deviation of the sampling distribution, called the standard error, is a common way of measuring the variability of a statistic. Knowing how much a statistic varies from sample to sample will allow us to determine the uncertainty in our estimate, a concept we will explore in more detail in the next section.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Distinguish between a population parameter and a sample statistic, recognizing that a parameter is fixed while a statistic varies from sample to sample
- Compute a point estimate for a parameter using an appropriate statistic from a sample
- Recognize that a sampling distribution shows how sample statistics tend to vary
- Recognize that statistics from random samples tend to be centered at the population parameter
- Estimate the standard error of a statistic from its sampling distribution
- Explain how sample size affects a sampling distribution

Exercises for Section 3.1

SKILL BUILDER 1

In Exercises 3.1 to 3.5, state whether the quantity described is a parameter or a statistic and give the correct notation.

3.1 Average household income for all houses in the US, using data from the US Census.

ANSWER ⊕

WORKED SOLUTION ⊕

3.2 Correlation between height and weight for players on the 2010 Brazil World Cup team, using data from all 23 players on the roster.

3.3 Proportion of people who use an electric toothbrush, using data from a sample of 300 adults.

ANSWER ⊕

WORKED SOLUTION ⊕

3.4 Proportion of registered voters in a county who voted in the last election, using data from the county voting records.

3.5 Average number of television sets per household in North Carolina, using data from a sample of 1000 households.

ANSWER ⊕

WORKED SOLUTION ⊕

SKILL BUILDER 2

In Exercises 3.6 to 3.11, give the correct notation for the quantity described and give its value.

3.6 Proportion of families in the US who were homeless in 2010. The number of homeless families[5]Luo, M., "*Number of Families in Shelters Rises*," *New York Times*, September 12, 2010. in 2010 was about 170,000 while the total number of families is given in the 2010 Census at 78 million.

3.7 Average enrollment in charter schools in Illinois. In 2010, there were 95 charter schools in the state of Illinois[6]Data obtained from www.uscharterschools.org. and the total number of students attending the charter schools was 30,795.

ANSWER ⊕

WORKED SOLUTION ⊕

3.8 Proportion of US adults who own a cell phone. In a survey of 2252 US adults, 82% said they had a cell phone.[7]"Spring Change Assessment Survey 2010," Princeton Survey Research Associates International, 6/4/10, accessed via "Cell Phones and American Adults," Amanda Lenhart, Pew Research Center's Internet and American Life Project, accessed at http://pewinternet.org/Reports/2010/Cell-Phones-and-American-Adults/Overview.aspx.

3.9 Correlation between age and heart rate for patients admitted to an Intensive Care Unit. Data from the 200 patients included in the file **ICUAdmissions** gives a correlation of 0.037.

ANSWER ⊕

WORKED SOLUTION ⊕

3.10 Mean number of cell phone calls made or received per day by cell phone users. In a survey of 1917 cell phone users, the mean was 13.10 phone calls a day.[8]"Spring Change Assessment Survey 2010," Princeton Survey Research Associates International, 6/4/10, accessed via "Cell Phones and American Adults," Amanda Lenhart, Pew Research Center's Internet and American Life Project, accessed at http://pewinternet.org/Reports/2010/Cell-Phones-and-American-Adults/Overview.aspx.

3.11 Correlation between points and penalty minutes for the 24 players with at least one point on the 2009-2010 Ottawa Senators[9]Data obtained from http://senators.nhl.com/club/stats.htm. NHL hockey team. The data are given in Table 3.4 and the full data are available in the file **OttawaSenators**.

Table 3.4 ***Points and penalty minutes for the 2009-2010 Ottawa Senators NHL team***

Points	71	57	53	49	48	34	32	29	28	26	26	26
Pen mins	22	20	59	54	34	18	38	20	28	121	53	24
Points	24	22	18	16	14	14	13	13	11	5	3	3

Pen mins	45	175	16	20	20	38	107	22	190	40	12	14

ANSWER

WORKED SOLUTION

SKILL BUILDER 3

Exercises 3.12 to 3.15 refer to the sampling distributions given in Figure 3.5. In each case, estimate the value of the population parameter and estimate the standard error for the sample statistic.

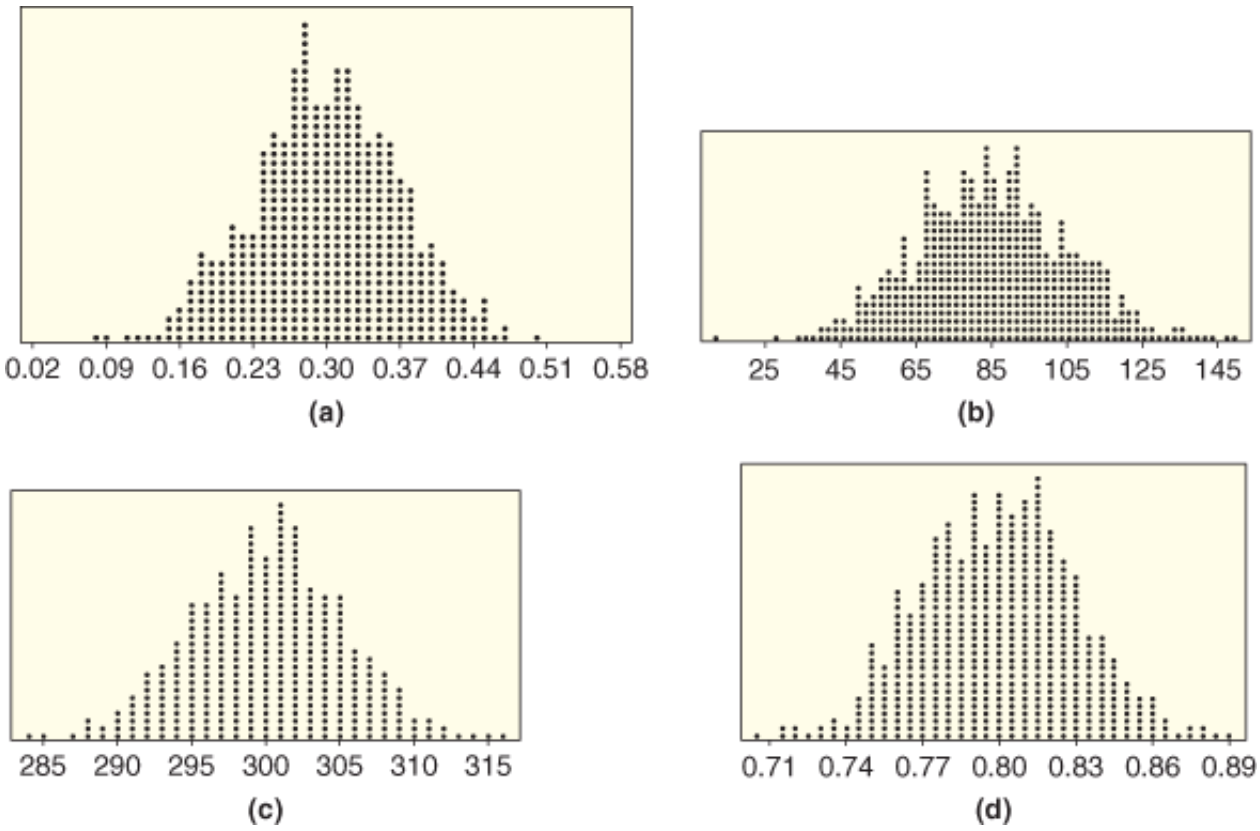


Figure 3.5 ***Four sampling distributions***

3.12 Figure 3.5(a) shows sample proportions from samples of size $n=40$ from a population.

3.13 Figure 3.5(b) shows sample means from samples of size $n=30$ from a population.

ANSWER

WORKED SOLUTION

3.14 Figure 3.5(c) shows sample means from samples of size $n=100$ from a population.

3.15 Figure 3.5(d) shows sample proportions from samples of size $n=180$ from a population.

ANSWER

WORKED SOLUTION

SKILL BUILDER 4

Exercises 3.16 to 3.19 refer to the sampling distributions given in Figure 3.5. Several possible values are

given for a sample statistic. In each case, indicate whether each value is (i) reasonably likely to occur from a sample of this size, (ii) unusual but might occur occasionally, or (iii) extremely unlikely to ever occur.

3.16 Using the sampling distribution shown in Figure 3.5(a), how likely are these sample proportions:

(a) $\hat{p} = 0.1$

(b) $\hat{p} = 0.35$

(c) $\hat{p} = 0.6$

3.17 Using the sampling distribution shown in Figure 3.5(b), how likely are these sample means:

(a) $\bar{x} = 70$

ANSWER ⊕

WORKED SOLUTION ⊕

(b) $\bar{x} = 100$

ANSWER ⊕

WORKED SOLUTION ⊕

(c) $\bar{x} = 140$

ANSWER ⊕

WORKED SOLUTION ⊕

3.18 Using the sampling distribution shown in Figure 3.5(c), how likely are these sample means:

(a) $\bar{x} = 250$

(b) $\bar{x} = 305$

(c) $\bar{x} = 315$

3.19 Using the sampling distribution shown in Figure 3.5(d), how likely are these sample proportions:

(a) $\hat{p} = 0.72$

ANSWER ⊕

WORKED SOLUTION ⊕

(b) $\hat{p} = 0.88$

ANSWER ⊕

WORKED SOLUTION ⊕

(c) $\hat{p} = 0.95$

ANSWER ⊕

WORKED SOLUTION ⊕

3.20 Customized Home Pages

A random sample of n=1675 Internet users in the US in January 2010 found that 469 of them have

customized their web browser's home page to include news from sources and on topics that particularly interest them.[10]Purcell, Rainie, Mitchell, Rosenthal, and Olmstead, "*Understanding the Participatory News Consumer*," Pew Research Center, March 1, 2010, http://www.pewinternet.org/Reports/2010/Online-News.aspx. State the population and parameter of interest. Use the information from the sample to give the best estimate of the population parameter. What would we have to do to calculate the value of the parameter exactly?

3.21 Laptop Computers

A survey conducted in May of 2010 asked 2252 adults in the US "Do you own a laptop computer?" The number saying yes was 1238. What is the best estimate for the proportion of US adults owning a laptop computer? Give notation for the quantity we are estimating, notation for the quantity we are using to make the estimate, and the value of the best estimate. Be sure to clearly define any parameters in the context of this situation.

ANSWER ⊕

WORKED SOLUTION ⊕

3.22 Florida Lakes

Florida has over 7700 lakes.[11]www.stateofflorida.com/florquicfac.html. We wish to estimate the correlation between the pH levels of all Florida lakes and the mercury levels of fish in the lakes. We see in Data 2.4 that the correlation between these two variables for a sample of $n=53$ of the lakes is -0.575.

(a) Give notation for the quantity we are estimating, notation for the quantity we use to make the estimate, and the value of the best estimate.

(b) Why is an estimate necessary here? What would we have to do to calculate the exact value of the quantity we are estimating?

3.23 Topical Painkiller Ointment

The use of topical painkiller ointment or gel rather than pills for pain relief was approved just within the last few years in the US for prescription use only.[12]Tarkan, L., "*Topical Gel Catches up with Pills for Relief*," *The New York Times*, September 6, 2010. Insurance records show that the average copayment for a month's supply of topical painkiller ointment for regular users is \$30. A sample of $n=75$ regular users found a sample mean copayment of \$27.90.

(a) Identify each of 30 and 27.90 as a parameter or a statistic and give the appropriate notation for each.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) If we take 1000 samples of size $n=75$ from the population of all copayments for a month's supply of topical painkiller ointment for regular users and plot the sample means on a dotplot, describe the shape you would expect to see in the plot and where it would be centered.

ANSWER ⊕

WORKED SOLUTION ⊕

(c) How many dots will be on the dotplot you described in part b? What will each dot represent?

ANSWER ⊕

WORKED SOLUTION ⊕

3.24 Average Household Size

The latest US Census lists the average household size for all households in the US as 2.61. (A household is all people occupying a housing unit as their primary place of residence.) Figure 3.6 shows possible distributions of means for 1000 samples of household sizes. The scale on the horizontal axis is the same in all four cases.

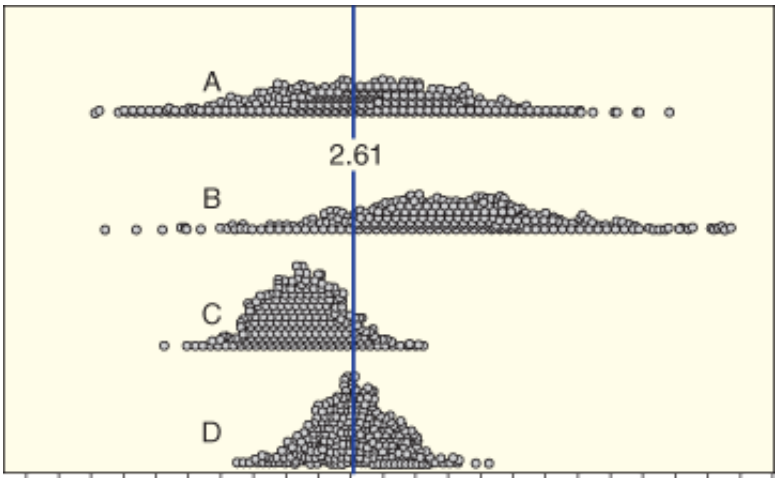


Figure 3.6 ***Sets of 1000 sample means***

(a) Assume that two of the distributions show results from 1000 random samples, while two others show distributions from a sampling method that is biased. Which two dotplots appear to show samples produced using a biased sampling method? Explain your reasoning. Pick one of the distributions that you listed as biased and describe a sampling method that might produce this bias.

(b) For the two distributions that appear to show results from random samples, suppose that one comes from 1000 samples of size $n=100$ and one comes from 1000 samples of size $n=500$. Which distribution goes with which sample size? Explain.

3.25 Proportion of US Residents Less than 25 Years Old

The US Census indicates that 35% of US residents are less than 25 years old. Figure 3.7 shows possible sampling distributions for the proportion of a sample less than 25 years old, for samples of size $n=20$, $n=100$, and $n=500$.

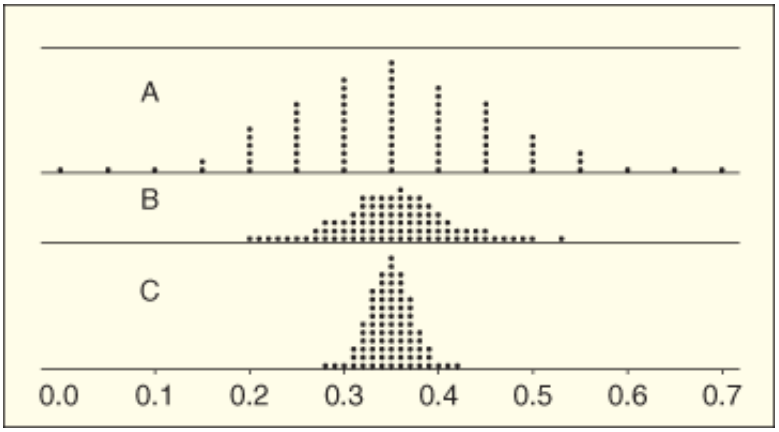


Figure 3.7 ***Match the dotplots with the sample size***

(a) Which distribution goes with which sample size?

ANSWER

WORKED SOLUTION

(b) If we use a proportion $\hat{p}$, based on a sample of size $n=20$, to estimate the population parameter $p=0.35$, would it be very surprising to get an estimate that is off by more than 0.10 (that is, the sample proportion is less than 0.25 or greater than 0.45)? How about with a sample of size $n=100$? How about with a sample of size $n=500$?

ANSWER

WORKED SOLUTION

(c) Repeat part b if we ask about the sample proportion being off by just 0.05 or more.

ANSWER

WORKED SOLUTION

(d) Using parts b and 3.25, comment on the effect that sample size has on the accuracy of an estimate.

ANSWER

WORKED SOLUTION

3.26 Mix It Up for Better Learning

In preparing for a test on a set of material, is it better to study one topic at a time or to study topics mixed together? In one study,[13]Rohrer, D. and Taylor, K., "*The Effects of Interleaved Practice*," *Applied Cognitive Psychology*, 2010;24(6): 837-848. a sample of fourth graders were taught four equations. Half of the children learned by studying repeated examples of one equation at a time, while the other half studied mixed problem sets that included examples of all four types of calculations grouped together. A day later, all the students were given a test on the material. The students in the mixed practice group had an average grade of 77, while the students in the one-at-a-time group had an average grade of 38. What is the best estimate for the difference in the average grade between fourth-grade students who study mixed problems and those who study each equation independently? Give notation (as a difference with a minus sign) for the quantity we are trying to estimate, notation for the quantity that gives the best

estimate, and the value of the best estimate. Be sure to clearly define any parameters in the context of this situation.

3.27 What Proportion of Adults and Teens Text Message?

A study of $n=2252$ adults age 18 or older found that 72% of the cell phone users send and receive text messages.[14]Lenhart, A.,"*Cell Phones and American Adults*," Pew Research Center's Internet and American Life Project, accessed at http://pewinternet.org/Reports/2010/Cell-Phones-and-American-Adults/Overview.aspx. A study of $n=800$ teens age 12 to 17 found that 87% of the teen cell phone users send and receive text messages. What is the best estimate for the difference in the proportion of cell phone users who use text messages, between adults (defined as 18 and over) and teens? Give notation (as a difference with a minus sign) for the quantity we are trying to estimate, notation for the quantity that gives the best estimate, and the value of the best estimate. Be sure to clearly define any parameters in the context of this situation.

ANSWER ⊕

WORKED SOLUTION ⊕

3.28 Hollywood Movies

Data 2.7 introduces the dataset **HollywoodMovies2011**, which contains information on all the 136 movies that came out of Hollywood in 2011.[15]McCandless, D., "*Most Profitable Hollywood Movies*," "Information is Beautiful," davidmccandless.com, accessed January 2012. One of the variables is the budget (in millions of dollars) to make the movie. Figure 3.8 shows two boxplots. One represents the budget data for one random sample of size $n=30$. The other represents the values in a sampling distribution of 1000 means of budget data from samples of size 30.

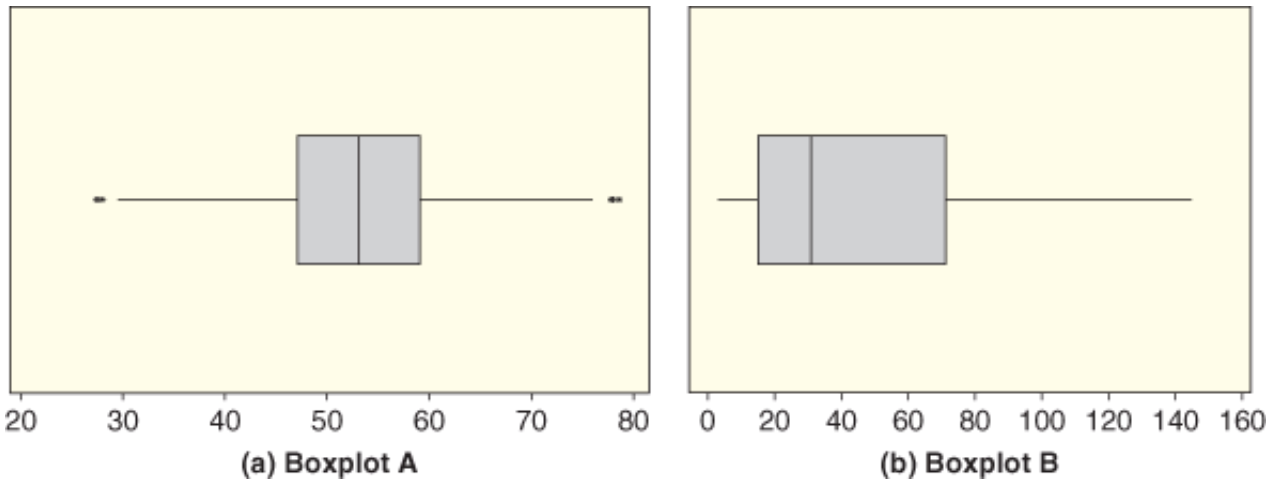


Figure 3.8 ***One sample and one sampling distribution: Which is which?***

(a) Which is which? Explain.

(b) From the boxplot showing the data from one random sample, what does one value in the sample represent? How many values are included in the data to make the boxplot? Estimate the minimum and maximum values. Give a rough estimate of the mean of the values and use appropriate notation for your answer.

(c) From the boxplot showing the data from a sampling distribution, what does one value in the sampling distribution represent? How many values are included in the data to make the boxplot? Estimate the minimum and maximum values. Give a rough estimate of the value of the population parameter and use appropriate notation for your answer.

3.29 Defective Screws

Suppose that 5% of the screws a company sells are defective. Figure 3.9 shows sample proportions from two sampling distributions: One shows samples of size 100, and the other shows samples of size 1000.

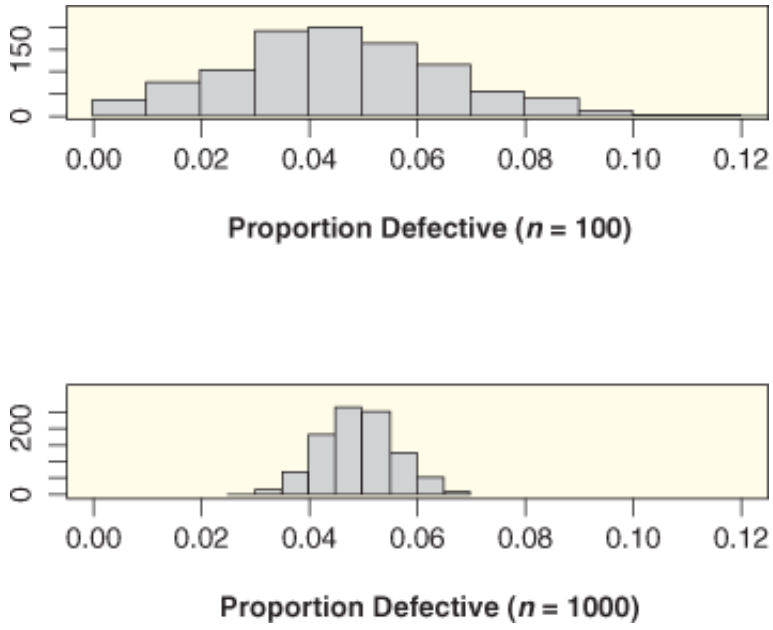


Figure 3.9 ***Sampling distributions for n=100 and n=1000 screws***

(a) What is the center of both distributions?

ANSWER ⊕

WORKED SOLUTION ⊕

(b) What is the approximate minimum and maximum of each distribution?

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Give a rough estimate of the standard error in each case.

ANSWER ⊕

WORKED SOLUTION ⊕

(d) Suppose you take one more sample in each case. Would a sample proportion of 0.08 (that is, 8% defective in the sample) be reasonably likely from a sample of size 100? Would it be reasonably likely from a sample of size 1000?

ANSWER ⊕

WORKED SOLUTION ⊕

3.30 Number of Screws in a Box

A company that sells boxes of screws claims that a box of its screws contains on average 50 screws $(\mu, =, 50)$. Figure 3.10 shows a distribution of sample means collected from many simulated random samples of size 10 boxes.

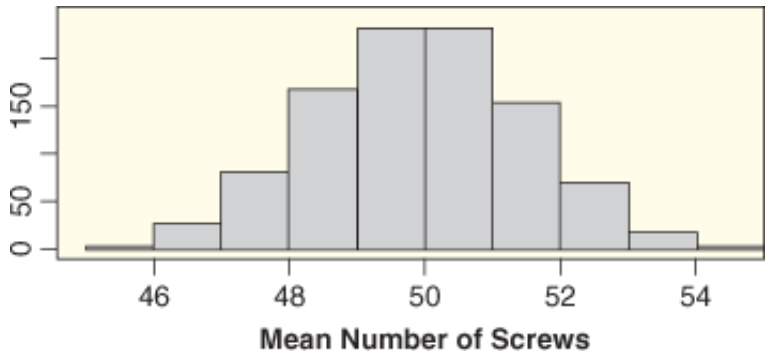


Figure 3.10 ***Sampling distribution for average number of screws in a box***

(a) For a random sample of 10 boxes, is it unlikely that the sample mean will be more than 2 screws different from μ? What about more than 5? 10?

(b) If you bought a random sample of 10 boxes at the hardware store and the mean number of screws per box was 42, would you conclude that the company's claim $(\mu, =, 50)$ is likely to be incorrect?

(c) If you bought a random box at the hardware store and it only contained 42 screws, would you conclude that the company's claim is likely to be incorrect?

3.31 Average Points for a Hockey Player

Table 3.4 gives the number of points for all 24 players on the Ottawa Senators NHL hockey team, also available in the dataset **OttawaSenators**.

(a) Use *StatKey*, other technology, or a random number table to select a random sample of 5 of the 24 *Points* values. Indicate which values you've selected and compute the sample mean.

ANSWER ⊕

WORKED SOLUTION ⊕

(b) Repeat part a by taking a second sample and calculating the mean.

ANSWER ⊕

WORKED SOLUTION ⊕

(c) Find the mean for the entire population of these 24 players. Use correct notation for your answer. Comment on the accuracy of using the sample means found in parts a and b to estimate the population mean.

ANSWER ⊕

WORKED SOLUTION ⊕

(d) Give a rough sketch of the sampling distribution if we calculate many sample means taking samples of size $n=5$ from this population of *Points* values. What shape will it have and where will it be centered?

ANSWER

WORKED SOLUTION

3.32 Time to Finish in 2008 Olympic Men's Marathon

In the 2008 Olympic Men's Marathon, 76 athletes finished the race. Their times are stored in the file **OlympicMarathon**. Use the times stored in the *Minutes* column.

(a) Use *StatKey*, other technology, or a random number table to randomly select 10 values. Indicate which values you've selected and compute the sample mean.

(b) Repeat part a by taking a second sample and calculating the mean. Make a mini-dotplot by plotting the two sample means on a dotplot.

(c) Find the mean for the entire population of 76 race times. Use correct notation for your answer. Comment on the accuracy of using the sample means found in parts a and b to estimate the population mean.

(d) Suppose we take many samples of size $n=10$ from this population of values and plot the mean for each sample on a dotplot. Describe the shape and center of the result. Draw a rough sketch of a possible distribution for these means.

3.33 A Sampling Distribution for Average Points for a Hockey Player

Use *StatKey* or other technology to generate a sampling distribution of sample means using a sample size of $n=5$ from the *Points* values in Table 3.4, which gives the number of points for all 24 players on the Ottawa Senators NHL hockey team, also available in the dataset **OttawaSenators**. What are the smallest and largest sample means in the distribution? What is the standard deviation of the sample means (in other words, what is the standard error?)

ANSWER

WORKED SOLUTION

3.34 A Sampling Distribution for Time to Finish in 2008 Olympic Men's Marathon

Use *StatKey* or other technology to generate a sampling distribution of sample means using a sample size of $n=10$ from the population of all times to finish the 2008 Olympic Men's Marathon, available in the *Minutes* column of the file **OlympicMarathon**. What are the smallest and largest sample means in the distribution? What is the standard deviation of the sample means (in other words, what is the standard error?)

3.35 Gender in the Rock and Roll Hall of Fame

From its founding through 2012, the Rock and Roll Hall of Fame has inducted 273 groups or individuals. Forty-one of the inductees have been female or have included female members.[16]Rock and Roll Hall of Fame website: rockhall.com/inductees. The full dataset is available in **RockandRoll**.

(a) What proportion of inductees have been female or have included female members? Use the correct notation with your answer.

ANSWER

WORKED SOLUTION

(b) If we took many samples of size 50 from the population of all inductees and recorded the proportion female or with female members for each sample, what shape do we expect the distribution of sample proportions to have? Where do we expect it to be centered?

ANSWER

WORKED SOLUTION

3.36 Performers in the Rock and Roll Hall of Fame

From its founding through 2012, the Rock and Roll Hall of Fame has inducted 273 groups or individuals, and 181 of the inductees have been performers while the rest have been related to the world of music in some way other than as a performer. The full dataset is available in **RockandRoll**.

(a) What proportion of inductees have been performers? Use the correct notation with your answer.

(b) If we took many samples of size 50 from the population of all inductees and recorded the proportion who were performers for each sample, what shape do we expect the distribution of sample proportions to have? Where do we expect it to be centered?

3.37 A Sampling Distribution for Gender in the Rock and Roll Hall of Fame

Exercise 3.35 tells us that 41 of the 273 inductees to the Rock and Roll Hall of Fame have been female or have included female members. The data are given in **Rockand Roll**. Using all inductees as your population:

(a) Use *StatKey* or other technology to take many random samples of size $n=10$ and compute the sample proportion that are female or with female members. What is the standard error for these sample proportions? What is the value of the sample proportion farthest from the population proportion of $p=0.150$? How far away is it?

ANSWER

WORKED SOLUTION

(b) Repeat part a using samples of size $n=20$.

ANSWER

WORKED SOLUTION

(c) Repeat part a using samples of size $n=50$.

ANSWER

WORKED SOLUTION

(d) Use your answers to parts a, b, and c to comment on the effect of increasing the sample size on the accuracy of using a sample proportion to estimate the population proportion.

ANSWER

WORKED SOLUTION

3.38 A Sampling Distribution for Performers in the Rock and Roll Hall of Fame

Exercise 3.36 tells us that 181 of the 273 inductees to the Rock and Roll Hall of Fame have been performers. The data are given in **RockandRoll**. Using all inductees as your population:

(a) Use *StatKey* or other technology to take many random samples of size $n=10$ and compute the sample proportion that are performers. What is the standard error of the sample proportions? What is the value of the sample proportion farthest from the population proportion of $p=0.663$? How far away is it?

(b) Repeat part a using samples of size $n=20$.

(c) Repeat part a using samples of size $n=50$.

(d) Use your answers to parts a, b, and c to comment on the effect of increasing the sample size on the accuracy of using a sample proportion to estimate the population proportion.