Chapter 1

Data Collection

1.4.2 Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students. All graduates in the last 5 years represent the *population*, and graduates who are selected for review are collectively called the *sample*. In general, we always seek to *randomly* select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted. For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates.

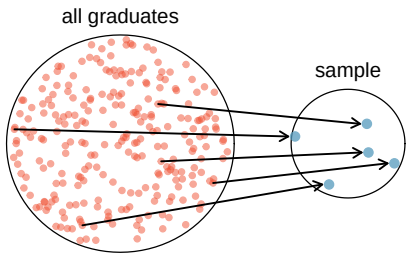

Figure 1.11: In this graphic, five graduates are randomly selected from the population to be included in the sample.

Why pick a sample randomly? Why not just pick a sample by hand? Consider the following scenario.

Example 1.16 Suppose we ask a student who happens to be majoring in nutrition to select several graduates for the study. What kind of students do you think she might collect? Do you think her sample would be representative of all graduates?

Perhaps she would pick a disproportionate number of graduates from health-related fields. Or perhaps her selection would be well-representative of the population. When selecting samples by hand, we run the risk of picking a *biased* sample, even if that bias is unintentional or difficult to discern.

If the student majoring in nutrition picked a disproportionate number of graduates from health-related fields, this would introduce selection bias into the sample. **Selection bias** occurs when some individuals of the population are inherently more likely to be included in the sample than others. In the example, this bias creates a problem because a degree in health-related fields might take more or less time to complete than a degree in other fields. Suppose that it takes longer. Since graduates from health-related fields would be more likely to be in the sample, the selection bias would cause her to *overestimate* the parameter.

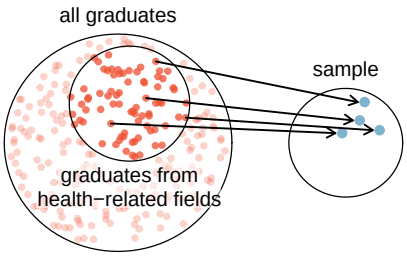

Figure 1.12: Instead of sampling from all graduates equally, a nutrition major might inadvertently pick graduates with health-related majors disproportionately often.

Sampling randomly resolves the problem of selection bias. The most basic random sample is called a **simple random sample**, and it is the equivalent of using a raffle to select cases. This means that each case in the population has an equal chance of being included and there is no implied connection between the cases in the sample.

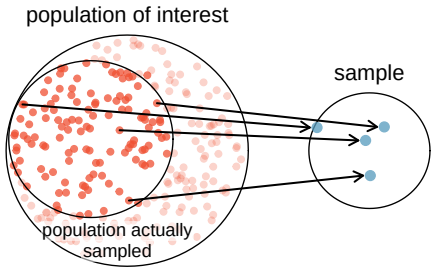

Figure 1.13: Due to the possibility of non-response, surveys studies may only reach a certain group within the population. It is difficult, and often times impossible, to completely fix this problem.

A common downfall is a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

Similarly, a **volunteer sample** is one in which people's responses are solicited and those who choose to participate, respond. This is a problem because those who choose to participate may tend to have different opinions than the rest of the population, resulting in a biased sample.

⊙ **Guided Practice 1.17** We can easily access ratings for products, sellers, and companies through websites. These ratings are based only on those people who go out of their way to provide a rating. If 50% of online reviews for a product are negative, do you think this means that 50% of buyers are dissatisfied with the product?[17]

The act of taking a random sample helps minimize bias; however, bias can crop up in other ways. Even when people are picked at random, e.g. for surveys, caution must be exercised if the **non-response** is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then it is unclear whether the results are **representative** of the entire population. This **non-response bias** can skew results.

Even if a sample has no selection bias and no non-response bias, there is an additional type of bias that often crops up and undermines the validity of results, known as response bias. **Response bias** refers to a broad range of factors that influence how a person responds, such as question wording, question order, and influence of the interviewer. This type of bias can be present even when we collect data from an entire population in what is called a **census**. Because response bias is often subtle, one must pay careful attention to how questions were asked when attempting to draw conclusions from the data.

● **Example 1.18** Suppose a high school student wants to investigate the student body's opinions on the food in the cafeteria. Let's assume that she manages to survey every student in the school. How might response bias arise in this context?

There are many possible correct answers to this question. For example, students might respond differently depending upon who asks the question, such as a school friend or someone who works in the cafeteria. The wording of the question could introduce response bias. Students would likely respond differently if asked "Do you like the food in the cafeteria?" versus "The food in the cafeteria is pretty bad, don't you think?"

[17] Answers will vary. From our own anecdotal experiences, we believe people tend to rant more about products that fell below expectations than rave about those that perform as expected. For this reason, we suspect there is a negative bias in product ratings on sites like Amazon. However, since our experiences may not be representative, we also keep an open mind should data on the subject become available.

TIP: Watch out for bias
Selection bias, non-response bias, and response bias can still exist within a random sample. Always ask how a sample was chosen, whether anyone failed to respond (and if so, how many people failed to respond), and critically examine the wording of the questions.

When there is no bias in a sample, increasing the sample size tends to increase the precision and reliability of the estimate. When a sample is biased, it may be impossible to decipher helpful information from the data, even if the sample is very large.

⊙ **Guided Practice 1.19** A researcher sends out questionnaires to 50 randomly selected households in a particular town asking whether or not they support the addition of a traffic light in their neighborhood. Because only 20% of the questionnaires are returned, she decides to mail questionnaires to 50 more randomly selected households in the same neighborhood. Comment on the usefulness of this approach.[18]

1.4.3 Simple, systematic, stratified, cluster, and multistage sampling

Almost all statistical methods for observational data rely on a sample being random and unbiased. When a sample is collected in a biased way, these statistical methods will not generally produce reliable information about the population.

The idea of a simple random sample was introduced in the last section. Here we provide a more technical treatment of this method and introduce four new random sampling methods: systematic, stratified, cluster, and multistage.[19] Figure 1.14 provides a graphical representation of simple versus systematic sampling while Figure 1.15 provides a graphical representation of stratified, cluster, and multistage sampling.

Simple random sampling is probably the most intuitive form of random sampling. Consider the salaries of Major League Baseball (MLB) players, where each player is a member of one of the league's 30 teams. For the 2010 season, N, the population size or total number of players, is 828. To take a simple random sample of $n = 120$ of these baseball players and their salaries, we could number each player from 1 to 828. Then we could randomly select 120 numbers between 1 and 828 (without replacement) using a

[18] The researcher should be concerned about non-response bias, and sampling more people will not eliminate this issue. Instead, she should make an effort to reach out to the households from the original sample that did not respond and solicit their feedback, possibly by going door-to-door.

[19] Systematic and Multistage sampling are not part of the AP syllabus.

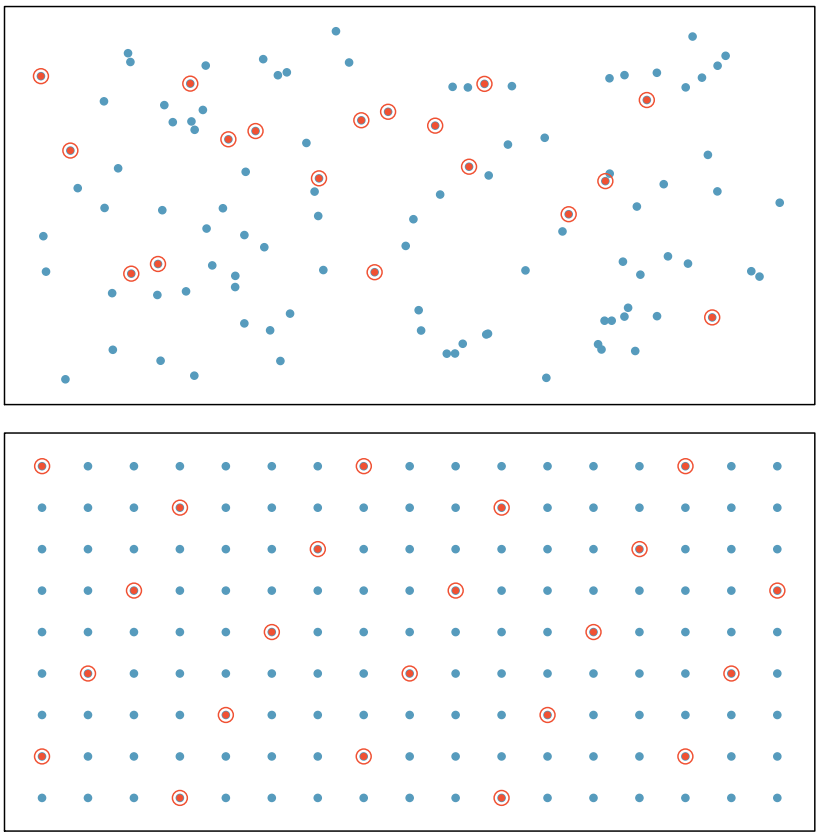

Figure 1.14: Examples of simple random sampling and systematic sampling. In the top panel, simple random sampling was used to randomly select 18 cases. In the lower panel, systematic random sampling was used to select every 7th individual.

random number generator or random digit table. The players with the selected numbers would comprise our sample.

Two properties are always true in a simple random sample:

1. Each case in the population has an equal chance of being included in the sample.

2. Each *group* of n cases has an equal chance of making up the sample.

The statistical methods in this book focus on data collected using simple random sampling. Note that Property 2 – that each group of n cases has an equal chance making up the sample – is not true for the remaining four sampling techniques. As you read each one, consider why.

Though less common than simple random sampling, **systematic sampling** is sometimes used when there exists a convenient list of all of the individuals of the population. Suppose we have a roster with the names of all the MLB players from the 2010 season. To take a systematic random sample, number them from 1 to 828. Select one random number between 1 and 828 and let that player be the first individual in the sample. Then, depending on the desired sample size, select every 10th number or 20th number, for example, to arrive at the sample.[20] If there are no patterns in the salaries based on the numbering then this could be a reasonable method.

Example 1.20 A systematic sample is not the same as a simple random sample. Provide an example of a sample that can come from a simple random sample but not from a systematic random sample.

Answers can vary. If we take a sample of size 3, then it is possible that we could sample players numbered 1, 2, and 3 in a simple random sample. Such a sample would be impossible from a systematic sample.

Sometimes there is a variable that is known to be associated with the quantity we want to estimate. In this case, a stratified random sample might be selected. **Stratified sampling** is a divide-and-conquer sampling strategy. The population is divided into groups called **strata**. The strata are chosen so that similar cases are grouped together and a sampling method, usually simple random sampling, is employed to select a certain number or a certain proportion of the whole within each stratum. In the baseball salary example, the 30 teams could represent the strata; some teams have a lot more money (we're looking at you, Yankees).

Example 1.21 For this baseball example, briefly explain how to select a stratified random sample of size $n = 120$.

Each team can serve as a stratum, and we could take a simple random sample of 4 players from each of the 30 teams, yielding a sample of 120 players.

Stratified sampling is inherently different than simple random sampling. For example, the stratified sampling approach described would make it impossible for the entire Yankees team to be included in the sample.

[20] If we want a sample of size $n = 138$, it would make sense to select every 6th player since $828/138 = 6$. Suppose we randomly select the number 810. Then player 810, 816, 822, 828, 6, 12, $\cdots$, 798, and 804 would make up the sample.

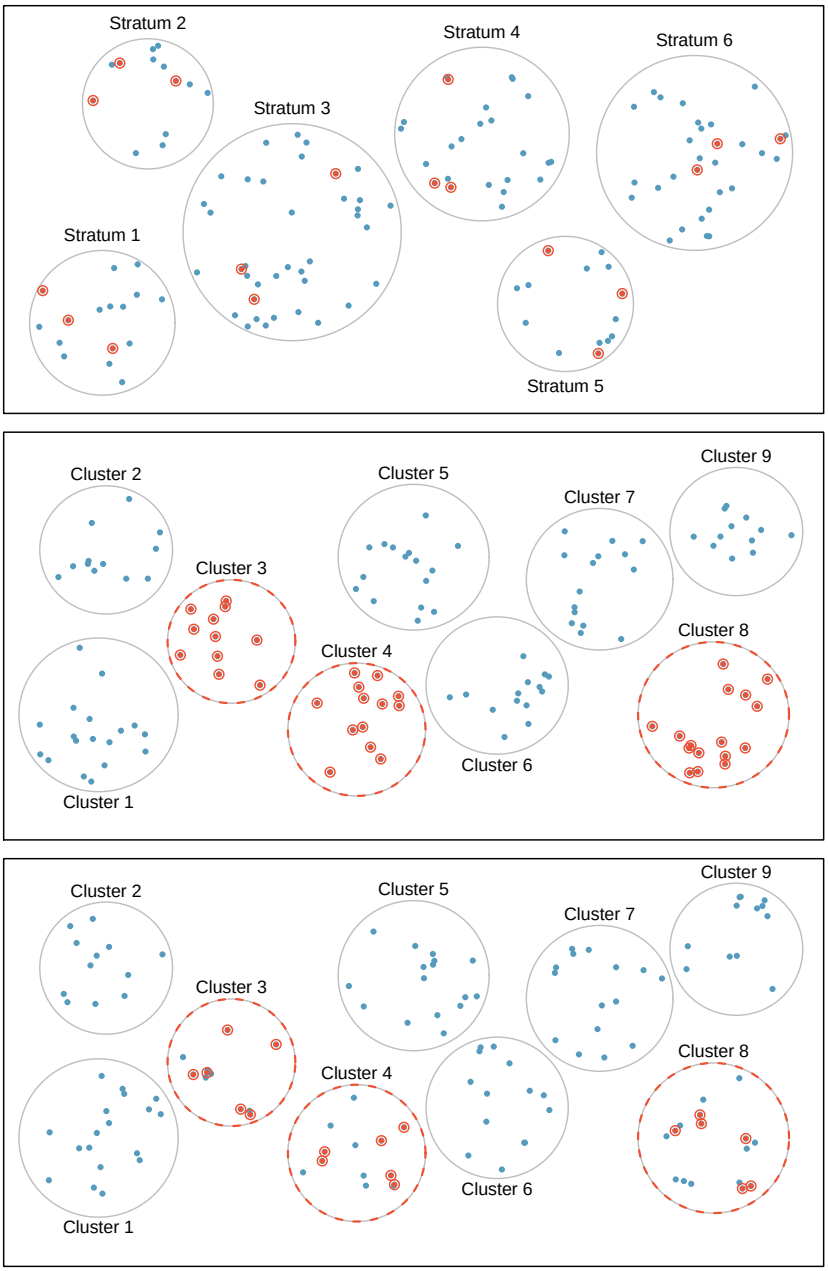

Figure 1.15: Examples of stratified, cluster, and multistage sampling. In the top panel, stratified sampling was used: cases were grouped into strata, and then simple random sampling was employed within each stratum. In the middle panel, cluster sampling was used, where data were binned into nine cluster and three clusters were randomly selected. In the bottom panel, multistage sampling was used. Data were binned into the nine clusters, three of the cluster were randomly selected, and then six cases were randomly sampled in each of the three selected clusters.

Example 1.22 Stratified sampling is especially useful when the cases in each stratum are very similar *with respect to the outcome of interest*. Why is it good for cases within each stratum to be very similar?

We should get a more stable estimate for the subpopulation in a stratum if the cases are very similar. These improved estimates for each subpopulation will help us build a reliable estimate for the full population. For example, in a simple random sample, it is possible that just by random chance we could end up with proportionally too many Yankees players in our sample, thus overestimating the true average salary of all MLB players. A stratified random sample can assure proportional representation from each team.

Next, let's consider a sampling technique that randomly selects groups of people. **Cluster sampling** is much like simple random sampling, but instead of randomly selecting *individuals*, we randomly select groups or **clusters**. Unlike stratified sampling, cluster sampling is most helpful when there is a lot of case-to-case variability within a cluster but the clusters themselves don't look very different from one another. That is, we expect strata to be self-similar (homogeneous), while we expect clusters to be diverse (heterogeneous).

Sometimes cluster sampling can be a more economical random sampling technique than the alternatives. For example, if neighborhoods represented clusters, this sampling method works best when each neighborhood is very diverse. Because each neighborhood itself encompasses diversity, a cluster sample can reduce the time and cost associated with data collection, because the interviewer would need only go to some of the neighborhoods rather than to all parts of a city, in order to collect a useful sample.

Multistage sampling, also called **multistage cluster sampling**, is a two (or more) step strategy. The first step is to take a cluster sample, as described above. Then, instead of including all of the individuals in these clusters in our sample, a second sampling method, usually simple random sampling, is employed within each of the selected clusters. In the neighborhood example, we could first randomly select some number of neighborhoods and then take a simple random sample from just those selected neighborhoods. As seen in Figure 1.15, stratified sampling requires observations to be sampled from *every* stratum. Multistage sampling selects observations *only* from those clusters that were randomly selected in the first step.

It is also possible to have more than two steps in multistage sampling. Each cluster may be naturally divided into subclusters. For example, each neighborhood could be divided into streets. To take a three-stage sample, we could first select some number of clusters (neighborhoods), and then, within the selected clusters, select some number of subclusters (streets). Finally, we could select some number of individuals from each of the selected streets.

Example 1.23 Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals. However, multistage cluster sampling seems like a very good idea. First, we might randomly select half the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample and would still give us reliable information.

Caution: advanced sampling techniques require advanced methods
The methods of inference covered in this book generally only apply to simple random samples. More advanced analysis techniques are required for systematic, stratified, cluster, and multistage random sampling.

Chapter 4

Distributions of random variables

4.2 Sampling distribution of a sample mean

4.2.1 The mean and standard deviation of $\bar{x}$

In this section we consider a data set called `run10`, which represents all 16,924 runners who finished the 2012 Cherry Blossom 10 mile run in Washington, DC.[25] Part of this data set is shown in Table 4.16, and the variables are described in Table 4.17.

ID	time	age	gender	state
1	92.25	38.00	M	MD
2	106.35	33.00	M	DC
3	89.33	55.00	F	VA
4	113.50	24.00	F	VA
⋮	⋮	⋮	⋮	⋮
16923	122.87	37.00	F	VA
16924	93.30	27.00	F	DC

Table 4.16: Six observations from the `run10` data set.

variable	**description**
`time`	Ten mile run time, in minutes
`age`	Age, in years
`gender`	Gender (`M` for male, `F` for female)
`state`	Home state (or country if not from the US)

Table 4.17: Variables and their descriptions for the `run10` data set.

[24] First find the mean and standard deviation of $Y - X$. The mean of $Y - X$ is $\mu_{Y-X} = 5.8 - 5.6 = 0.2$. The standard deviation is $SD_{Y-X} = \sqrt{(0.13)^2 + (0.11)^2} = 0.170$. Then $Z = \frac{0-0.2}{0.170} = -1.18$ and $P(Z < -1.18) = .119$. There is an 11.9% chance that Friend 2 will complete the puzzle with a faster time than Friend 1.

[25] http://www.cherryblossom.org

ID	time	age	gender	state
1983	88.31	59	M	MD
8192	100.67	32	M	VA
11020	109.52	33	F	VA
⋮	⋮	⋮	⋮	⋮
1287	89.49	26	M	DC

Table 4.18: Four observations for the `run10Samp` data set, which represents a simple random sample of 100 runners from the 2012 Cherry Blossom Run.

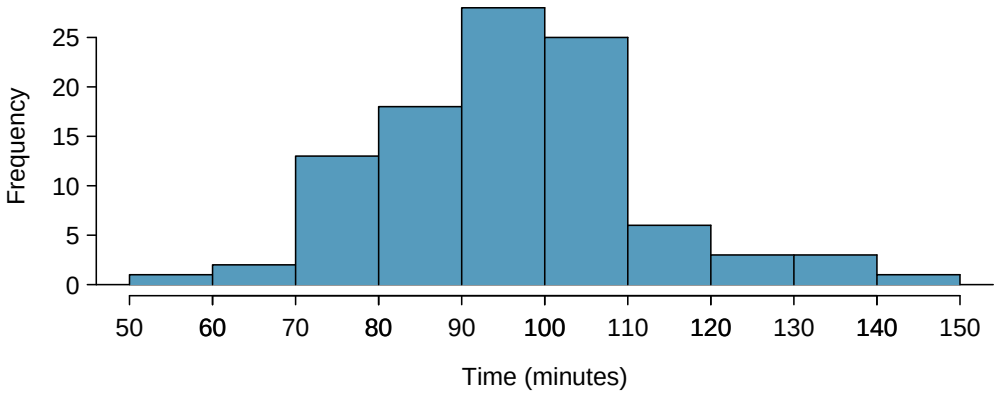

Figure 4.19: Histogram of `time` for a single sample of size 100. The average of the sample is in the mid-90s and the standard deviation of the sample $s \approx 16$ minutes.

These data are special because they include the results for the entire population of runners who finished the 2012 Cherry Blossom Run. We took a simple random sample of this population, which is represented in Table 4.18. A histogram summarizing the time variable in the `run10Samp` data set is shown in Figure 4.19.

From the random sample represented in `run10Samp`, we guessed the average time it takes to run 10 miles is 95.61 minutes. Suppose we take another random sample of 100 individuals and take its mean: 95.30 minutes. Suppose we took another (93.43 minutes) and another (94.16 minutes), and so on. If we do this many many times – which we can do only because we have the entire population data set – we can build up a **sampling distribution** for the sample mean when the sample size is 100, shown in Figure 4.20.

Sampling distribution
The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. It is useful to think of a particular point estimate as being drawn from such a distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.

The sampling distribution shown in Figure 4.20 is unimodal and approximately symmetric. It is also centered exactly at the true population mean: $\mu = 94.52$. Intuitively, this makes sense. The sample means should tend to "fall around" the population mean.

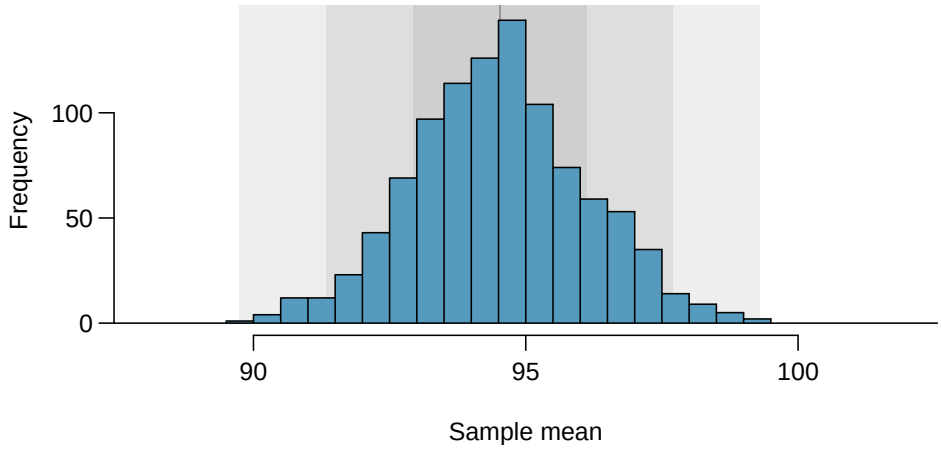

Figure 4.20: A histogram of 1000 sample means for run time, where the samples are of size $n = 100$. This histogram approximates the true sampling distribution of the sample mean, with mean $\mu_{\bar{x}}$ and standard deviation $\sigma_{\bar{x}}$.

We can see that the sample mean has some variability around the population mean, which can be quantified using the standard deviation of this distribution of sample means: $\sigma_{\bar{x}} = 1.59$. The standard deviation of the sample mean tells us how far the typical estimate is away from the actual population mean, 94.52 minutes. It also describes the typical **error** of a single estimate.

Standard deviation of an estimate
The standard deviation associated with an estimate describes the typical error or uncertainty associated with the estimate.

Example 4.32 Looking at Figures 4.19 and 4.20, we see that the standard deviation of the sample mean with $n = 100$ is much smaller than the standard deviation of a single sample. Interpret this statement and explain why it is true.

The variation from one sample mean to another sample mean is much smaller than the variation from one individual to another individual. This makes sense because when we average over 100 values, the large and small values tend to cancel each other out. While many individuals have a time under 90 minutes, it would be unlikely for the *average* of 100 runners to be less than 90 minutes.

Guided Practice 4.33 (a) Would you rather use a small sample or a large sample when estimating a parameter? Why? (b) Using your reasoning from (a), would you expect a point estimate based on a small sample to have smaller or larger standard deviation than a point estimate based on a larger sample?[26]

[26] (a) Consider two random samples: one of size 10 and one of size 1000. Individual observations in the small sample are highly influential on the estimate while in larger samples these individual observations would more often average each other out. The larger sample would tend to provide a more accurate estimate. (b) If we think an estimate is better, we probably mean it typically has less error. Based on (a), our intuition suggests that a larger sample size corresponds to a smaller standard deviation.

When considering how to calculate the standard deviation of a sample mean, there is one problem: there is no obvious way to estimate this from a single sample. However, statistical theory provides a helpful tool to address this issue.

In the sample of 100 runners, the standard deviation of the sample mean is equal to one-tenth of the population standard deviation: $1.59 = 15.93/10$. In other words, the standard error of the sample mean based on 100 observations is equal to

$$SD_{\bar{x}} = \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{15.93}{\sqrt{100}} = 1.59$$

where σ_x is the standard deviation of the individual observations. This is no coincidence. We can show mathematically that this equation is correct when the observations are independent using the probability tools of Section 3.4.

Computing SD for the sample mean

Given n independent observations from a population with standard deviation σ, the standard deviation of the sample mean is equal to

$$SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4.34}$$

A reliable method to ensure sample observations are independent is to conduct a simple random sample consisting of less than 10% of the population.

⊙ **Guided Practice 4.35** The average of the runners' ages is 35.05 years with a standard deviation of $\sigma = 8.97$. A simple random sample of 100 runners is taken. (a) What is the standard deviation of the sample mean? (b) Would you be surprised to get a sample of size 100 with an average of 36 years?[27]

⊙ **Guided Practice 4.36** (a) Would you be more trusting of a sample that has 100 observations or 400 observations? (b) We want to show mathematically that our estimate tends to be better when the sample size is larger. If the standard deviation of the individual observations is 10, what is our estimate of the standard deviation of the mean when the sample size is 100? What about when it is 400? (c) Explain how your answer to (b) mathematically justifies your intuition in part (a).[28]

[27] (a) Use Equation (4.34) with the population standard deviation to compute the standard deviation of the sample mean: $SD_{\bar{y}} = 8.97/\sqrt{100} = 0.90$ years. (b) It would not be surprising. 36 years is about 1 standard deviation from the true mean of 35.05. Based on the 68, 95 rule, we would get a sample mean at least this far away from the true mean approximately $100\% - 68\% = 32\%$ of the time.

[28] (a) Extra observations are usually helpful in understanding the population, so a point estimate with 400 observations seems more trustworthy. (b) The standard deviation of the mean when the sample size is 100 is given by $SD_{100} = 10/\sqrt{100} = 1$. For 400: $SD_{400} = 10/\sqrt{400} = 0.5$. The larger sample has a smaller standard deviation of the mean. (c) The standard deviation of the mean of the sample with 400 observations is lower than that of the sample with 100 observations. The standard deviation of $\bar{x}$ describes the typical error, and since it is lower for the larger sample, this mathematically shows the estimate from the larger sample tends to be better – though it does not guarantee that every large sample will provide a better estimate than a particular small sample.

4.2.2 Examining the Central Limit Theorem

When sampling from a population that is normally distributed, the distribution of a sample mean is normal. Even when the population values are skewed, the normal model for the sample mean tends to be very good when the sample consists of at least 30 independent observations. The Central Limit Theorem provides the theory that allows us to model the sample mean using the normal distribution.

Central Limit Theorem, informal definition
The distribution of $\bar{x}$ approaches the normal distribution as n increases. Generally, if the sample size $n \geq 30$, the distribution $\bar{x}$ will be well approximated by the normal distribution, even for skewed populations.

The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. We will investigate three cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *normal* distribution. These distributions are shown in the top panels of Figure 4.21. The uniform distribution is symmetric, and the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers).

The left panel in the $n = 2$ row represents the sampling distribution of $\bar{x}$ if it is the sample mean of two observations from the uniform distribution shown. The dashed line represents the closest approximation of the normal distribution. Similarly, the center and right panels of the $n = 2$ row represent the respective distributions of $\bar{x}$ for data from exponential and log-normal distributions.

Guided Practice 4.37 Examine the distributions in each row of Figure 4.21. What do you notice about the normal approximation for each sampling distribution as the sample size becomes larger?[29]

Example 4.38 Would the normal approximation be good in all applications where the sample size is at least 30?

Yes, the sampling distributions when $n = 30$ all look very much like the normal distribution.

However, the more non-normal a population distribution, the larger a sample size seems is necessary for the sampling distribution to look nearly normal.

TIP: With larger n, the sampling distribution of $\bar{x}$ becomes more normal
As the sample size increases, the normal model for $\bar{x}$ becomes more reasonable. We can also relax our condition on skew when the sample size is very large.

[29] The normal approximation becomes better as larger samples are used. However, in the case when the population is normally distributed, the normal distribution of the sample mean is normal for all sample sizes.

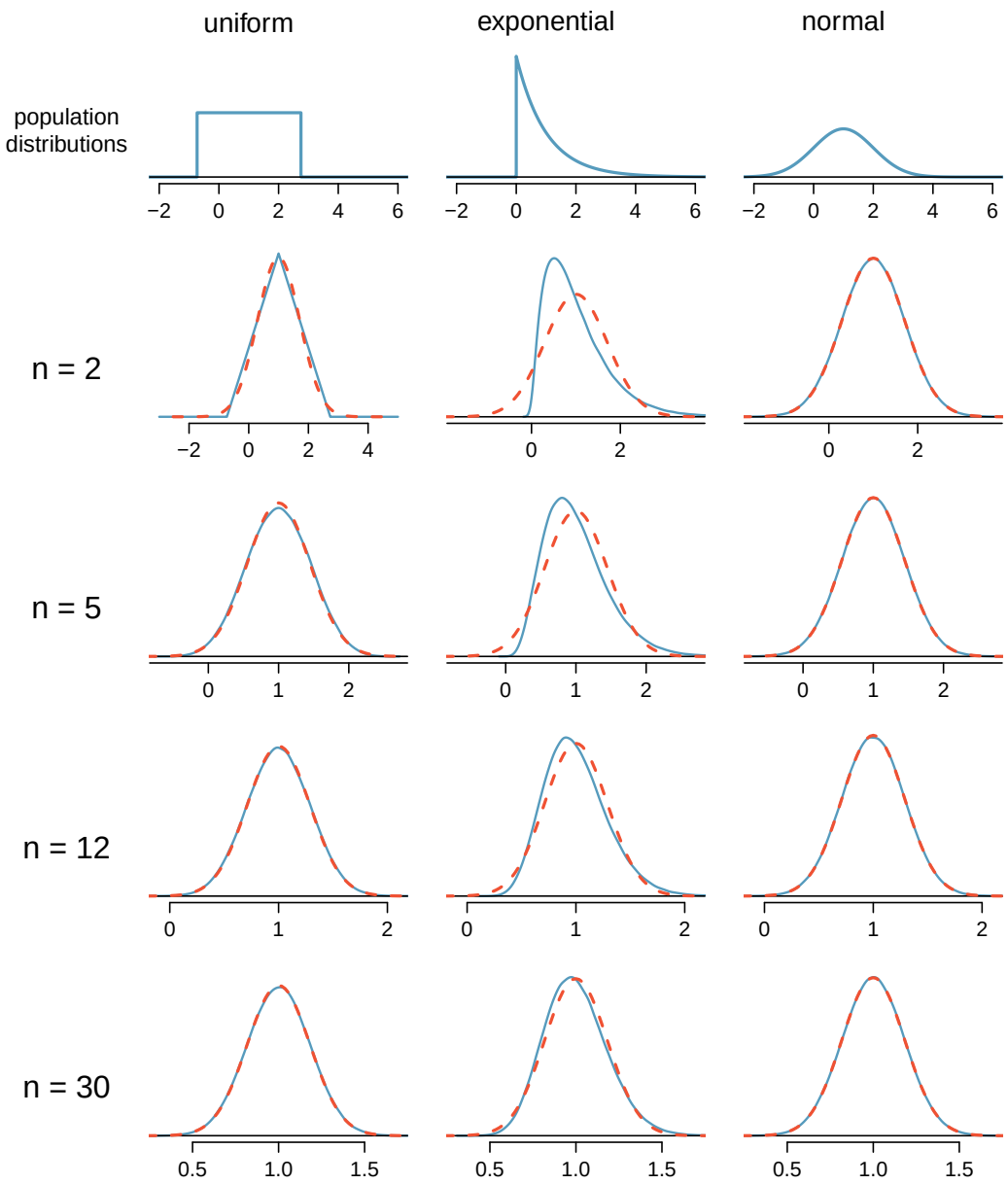

Figure 4.21: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

Example 4.39 Figure 4.22 shows a histogram of 50 observations. These represent winnings and losses from 50 consecutive days of a professional poker player. Can the normal approximation be applied to the sample mean, 90.69?

We should consider each of the required conditions.

(1) These are referred to as **time series data**, because the data arrived in a particular sequence. If the player wins on one day, it may influence how she plays the next. To make the assumption of independence we should perform careful checks on such data. While the supporting analysis is not shown, no evidence was found to indicate the observations are not independent.

(2) The sample size is 50, which is pretty large.

(3) There are two outliers, one very extreme, which suggests the data are very strongly skewed or very distant outliers may be common for this type of data. Outliers can play an important role and affect the distribution of the sample mean and the estimate of the standard error.

Since we should be skeptical of the independence of observations and the very extreme upper outlier poses a challenge, we should not use the normal model for the sample mean of these 50 observations. If we can obtain a much larger sample, perhaps several hundred observations, then the concerns about skew and outliers would no longer apply.

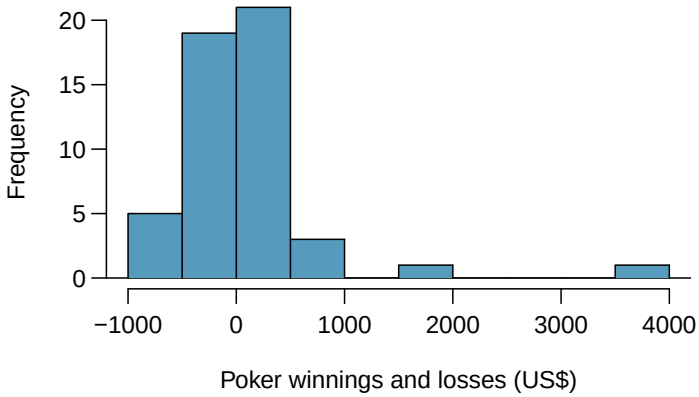

Figure 4.22: Sample distribution of poker winnings. These data include some very clear outliers. These are problematic when considering the normality of the sample mean. For example, outliers are often an indicator of very strong skew.

Caution: Examine data structure when considering independence
Some data sets are collected in such a way that they have a natural underlying structure between observations, e.g. when observations occur consecutively. Be especially cautious about independence assumptions regarding such data sets.

Caution: Watch out for strong skew and outliers
Strong skew is often identified by the presence of clear outliers. If a data set has prominent outliers, or such observations are somewhat common for the type of data under study, then it is useful to collect a sample with many more than 30 observations if the normal model will be used for $\bar{x}$. There are no simple guidelines for what sample size is big enough for all situations, so proceed with caution when working in the presence of strong skew or more extreme outliers.

4.2.3 Normal approximation for the sampling distribution of $\bar{x}$

We have seen that when the sample size is at least 30 and the observations are independent, we can apply the normal model for the sampling distribution of $\bar{x}$.

Example 4.40 In the 2012 Cherry Blossom 10 mile run, the average time for all of the runners is 94.52 minutes with a standard deviation of 8.97 minutes. The distribution of run times is approximately normal. Find the probabiliy that a randomly selected runner completes the run in less than 90 minutes.

Because the distribution of run times is approximately normal, we can use normal approximation.

$$Z = \frac{x - \mu}{\sigma} = \frac{90 - 94.52}{8.97} = -0.504$$
$$P(Z < -0.504) = 0.3072$$

There is a 30.72% probability that a randomly selected runner will complete the run in less than 90 minutes.

Example 4.41 Find the probabiliy that the average of 20 runners is less than 90 minutes.

Here, $n = 20 < 30$, but the distribution of the population, that is, the distribution of run times is stated to be approximately normal. Because of this, the sampling distribution will be normal for any sample size.

$$SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8.97}{\sqrt{20}} = 2.01$$
$$Z = \frac{\bar{x} - \mu}{\sigma} = \frac{90 - 94.52}{2.01} = -2.25$$
$$P(Z < -0.504) = 0.0123$$

There is a 1.23% probability that the average run time of 20 randomly selected runners will be less than 90 minutes.

Example 4.42 The average of all the runners' ages is 35.05 years with a standard deviation of $\sigma = 8.97$. The distribution of age is somewhat skewed. What is the probability that a randomly selected runner is older than 37 years?

Because the distribution of age is skewed and is not normal, we cannot use normal approximation for this problem. In order to answer this question, we would need to look at all of the data.

⊙ **Guided Practice 4.43** What is the probability that the average of 50 randomly selected runners is greater than 37 years?[30]

TIP: Remember to divide by $\sqrt{n}$
When finding the probability that an *average* or mean is greater or less than a particular value, remember to divide the standard deviation of the data by $\sqrt{n}$ to calculate the correct SD.

[30] Because $n = 50 \geq 30$, the sampling distribution of the mean is approximately normal, so we can use normal approximation for this problem. The mean is given as 35.05 years.

$$SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{8.97}{\sqrt{50}} = 1.27 \qquad z = \frac{\bar{x} - \mu}{SD_{\bar{x}}} = \frac{37 - 35.05}{1.27} = 1.535 \qquad P(Z > 37) = 0.062$$

There is a 6.2% chance that the average age of 50 runners will be greater than 37.

Chapter 5

Foundation for inference

5.2 Confidence intervals

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. In addition to supplying a point estimate of a parameter, a next logical step would be to provide a plausible *range of values* for the parameter.

5.2.1 Capturing the population parameter

A plausible range of values for the population parameter is called a **confidence interval**. Using only a point estimate is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net. We can throw a spear where we saw a fish, but we will probably miss. On the other hand, if we toss a net in that area, we have a good chance of catching the fish.

If we report a point estimate, we probably will not hit the exact population parameter. On the other hand, if we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.

⊙ **Guided Practice 5.7** If we want to be very confident we capture the population parameter, should we use a wider interval or a smaller interval?[2]

[2] If we want to be more confident we will capture the fish, we might use a wider net. Likewise, we use a wider confidence interval if we want to be more confident that we capture the parameter.

5.2.2 Constructing a 95% confidence interval

A point estimate is our best guess for the value of the parameter, so it makes sense to build the confidence interval around that value. The standard error, which is a measure of the uncertainty associated with the point estimate, provides a guide for how large we should make the confidence interval.

Constructing a 95% confidence interval

When the sampling distribution of a point estimate can reasonably be modeled as normal, the point estimate we observe will be within 1.96 standard errors of the true value of interest about 95% of the time. Thus, a **95% confidence interval** for such a point estimate can be constructed:

$$\text{point estimate} \pm 1.96 \times SE \tag{5.8}$$

We can be **95% confident** this interval captures the true value.

⊙ **Guided Practice 5.9** Compute the area between -1.96 and 1.96 for a normal distribution with mean 0 and standard deviation 1. [3]

● **Example 5.10** The point estimate from the smoking example was 15%. In the next chapters we will determine when we can apply a normal model to a point estimate. For now, assume that the normal model is reasonable. The standard error for this point estimate was calculated to be $SE = 0.04$. Construct a 95% confidence interval.

$$\begin{aligned} \text{point estimate} &\pm 1.96 \times SE \\ 0.15 &\pm 1.96 \times 0.04 \\ (0.0716 &\text{ , } 0.2284) \end{aligned}$$

We are 95% confident that the true proportion of smokers in this population is between 7.16% and 22.84%.

● **Example 5.11** Based on the confidence interval above, is there evidence that a smaller proportion smoke in this county than in the state as a whole? The proportion that smoke in the state is known to be 0.20.

While the point estimate of 0.15 is lower than 0.20, this deviation is likely due to random chance. Because the confidence interval *includes* the value 0.20, 0.20 is a reasonable value for the proportion of smokers in the county. Therefore, based on this confidence interval, we do not have evidence that a smaller proportion smoke in the county than in the state.

In Section 1.1 we encountered an experiment that examined whether implanting a stent in the brain of a patient at risk for a stroke helps reduce the risk of a stroke. The results from the first 30 days of this study, which included 451 patients, are summarized in Table 5.1. These results are surprising! The point estimate suggests that patients who received stents may have a *higher* risk of stroke: $p_{trmt} - p_{ctrl} = 0.090$.

[3] We will leave it to you to draw a picture. The Z scores are $Z_{left} = -1.96$ and $Z_{right} = 1.96$. The area between these two Z scores is 0.9500. This is where "1.96" comes from in the 95% confidence interval formula.

	stroke	no event	Total
treatment	33	191	224
control	13	214	227
Total	46	405	451

Table 5.1: Descriptive statistics for 30-day results for the stent study.

Example 5.12 Consider the stent study and results. The conditions necessary to ensure the point estimate $p_{trmt} - p_{ctrl} = 0.090$ is nearly normal have been verified for you, and the estimate's standard error is $SE = 0.028$. Construct a 95% confidence interval for the change in 30-day stroke rates from usage of the stent.

The conditions for applying the normal model have already been verified, so we can proceed to the construction of the confidence interval:

$$\begin{aligned} \text{point estimate} &\pm 1.96 \times SE \\ 0.090 &\pm 1.96 \times 0.028 \\ (0.035 &, 0.145) \end{aligned}$$

We are 95% confident that implanting a stent in a stroke patient's brain. Since the entire interval is greater than 0, it means the data provide statistically significant evidence that the stent used in the study *increases* the risk of stroke, contrary to what researchers had expected before this study was published!

We can be 95% confident that a 95% confidence interval contains the true population parameter. However, confidence intervals are imperfect. About 1-in-20 (5%) properly constructed 95% confidence intervals will fail to capture the parameter of interest. Figure 5.2 shows 25 confidence intervals for a proportion that were constructed from simulations where the true proportion was $p = 0.3$. However, 1 of these 25 confidence intervals happened not to include the true value.

Guided Practice 5.13 In Figure 5.2, one interval does not contain the true proportion, $p = 0.3$. Does this imply that there was a problem with the simulations run?[4]

5.2.3 Changing the confidence level

Suppose we want to consider confidence intervals where the confidence level is somewhat higher than 95%: perhaps we would like a confidence level of 99%.

Example 5.14 Would a 99% confidence interval be wider or narrower than a 95% confidence interval?

Using a previous analogy: if we want to be more confident that we will catch a fish, we should use a wider net, not a smaller one. To be 99% confidence of capturing the true value, we must use a wider interval. On the other hand, if we want an interval with lower confidence, such as 90%, we would use a narrower interval.

[4]No. Just as some observations occur more than 1.96 standard deviations from the mean, some point estimates will be more than 1.96 standard errors from the parameter. A confidence interval only provides a plausible range of values for a parameter. While we might say other values are implausible based on the data, this does not mean they are impossible.

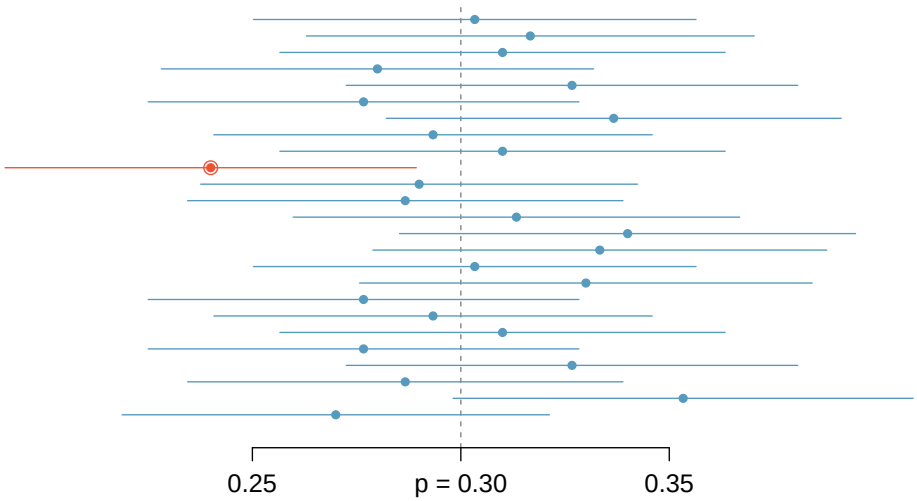

Figure 5.2: Twenty-five samples of size $n = 300$ were simulated when $p = 0.30$. For each sample, a confidence interval was created to try to capture the true proportion p. However, 1 of these 25 intervals did not capture $p = 0.30$.

The 95% confidence interval structure provides guidance in how to make intervals with new confidence levels. Below is a general 95% confidence interval for a point estimate that comes from a nearly normal distribution:

$$\text{point estimate} \ \pm \ 1.96 \times SE \tag{5.15}$$

There are three components to this interval: the point estimate, "1.96", and the standard error. The choice of $1.96 \times SE$ was based on capturing 95% of the distribution since the estimate is within 1.96 standard deviations of the true value about 95% of the time. The choice of 1.96 corresponds to a 95% confidence level.

⊙ **Guided Practice 5.16** If X is a normally distributed random variable, how often will X be within 2.58 standard deviations of the mean?[5]

To create a 99% confidence interval, change 1.96 in the 95% confidence interval formula to be 2.58. Guided Practice 5.16 highlights that 99% of the time a normal random variable will be within 2.58 standard deviations of its mean. This approach – using the Z scores in the normal model to compute confidence levels – is appropriate when the point estimate is associated with a normal distribution and we can properly compute the standard error. Thus, the formula for a 99% confidence interval is

$$\text{point estimate} \ \pm \ 2.58 \times SE \tag{5.17}$$

Figure 5.3 provides a picture of how to identify $z^\star$ based on a confidence level.

[5]This is equivalent to asking how often the Z score will be larger than -2.58 but less than 2.58. (For a picture, see Figure 5.3.) There is ≈ 0.99 probability that the unobserved random variable X will be within 2.58 standard deviations of the mean.

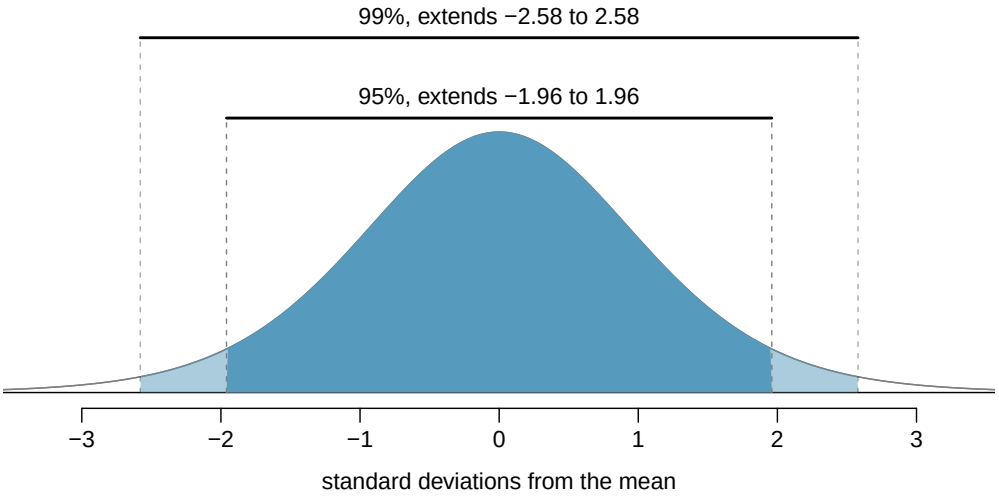

Figure 5.3: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

⊙ **Guided Practice 5.18** Create a 99% confidence interval for the impact of the stent on the risk of stroke using the data from Example 5.12. The point estimate is 0.090, and the standard error is $SE = 0.028$. It has been verified for you that the point estimate can reasonably be modeled by a normal distribution.[6]

Confidence interval for any confidence level

If the point estimate follows the normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate} \pm z^\star \times SE$$

where $z^\star$ corresponds to the confidence level selected.

Finding the value of $z^\star$ that corresponds to a particular confidence level is most easily accomplished by using a new table, called the t table. For now, what is noteworthy about this table is that the bottom row corresponds to confidence levels. The numbers inside the table are the critical values, but which row should we use? Later in this book, we will see that a t curve with infinite degrees of freedom corresponds to the normal curve. For this reason, when finding using the t table to find the appropriate $z^\star$, always use row ∞.

[6]Since the necessary conditions for applying the normal model have already been checked for us, we can go straight to the construction of the confidence interval: point estimate $\pm$ $2.58 \times SE \rightarrow (0.018, 0.162)$. We are 99% confident that implanting a stent in the brain of a patient who is at risk of stroke increases the risk of stroke within 30 days by a rate of 0.018 to 0.162 (assuming the patients are representative of the population).

	one tail	0.100	0.050	0.025	0.010	0.005
df	1	3.078	6.314	12.71	31.82	63.66
	2	1.886	2.920	4.303	6.965	9.925
	3	1.638	2.353	3.182	4.541	5.841
	⋮	⋮	⋮	⋮	⋮	
	1000	1.282	1.646	1.962	2.330	2.581
	∞	1.282	1.645	1.960	2.326	2.576
Confidence level C		80%	90%	95%	98%	99%

Table 5.4: An abbreviated look at the t table. The columns correspond to confidence levels. Row ∞ corresponds to the normal curve.

TIP: Finding $z^\star$ for a particular confidence level
We select $z^\star$ so that the area between -$z^\star$ and $z^\star$ in the normal model corresponds to the confidence level. Use the t table at row ∞ to find the critical value $z^\star$.

⊙ **Guided Practice 5.19** In Example 5.12 we found that implanting a stent in the brain of a patient at risk for a stroke *increased* the risk of a stroke. The study estimated a 9% increase in the number of patients who had a stroke, and the standard error of this estimate was about $SE = 2.8\%$ or 0.028. Compute a 90% confidence interval for the effect. Note: the conditions for normality had earlier been confirmed for us.[7]

The normal approximation is crucial to the precision of these confidence intervals. The next two chapters provides detailed discussions about when the normal model can safely be applied to a variety of situations. When the normal model is not a good fit, we will use alternate distributions that better characterize the sampling distribution.

5.2.4 Margin of error

The confidence intervals we have encountered thus far have taken the form

$$\text{point estimate} \ \pm \ z^* \times SE$$

Confidence intervals are also often reported as

$$\text{point estimate} \ \pm \ \text{margin of error}$$

For example, instead of reporting an interval as $0.09 \ \pm \ 1.645 \times 0.028$ or $(0.044, 0.136)$, it could be reported as $0.09 \ \pm \ 0.046$.

[7] We must find $z^\star$ such that 90% of the distribution falls between -$z^\star$ and $z^\star$ in the standard normal model. Using the t table with a confidence level of 90% at row ∞ gives 1.645. Thus $z^\star = 1.645$. The 90% confidence interval can then be computed as

$$\begin{aligned} \text{point estimate} \ &\pm \ z^\star \times SE \\ 0.09 \ &\pm \ 1.645 \times 0.028 \\ (0.044 \ &, \ 0.136) \end{aligned}$$

That is, we are 90% confident that implanting a stent in a stroke patient's brain increased the risk of stroke within 30 days by 4.4% to 13.6%.

The **margin of error** is the distance between the point estimate and the lower or upper bound of a confidence interval.

Margin of error
A confidence interval can be written as point estimate ± margin of error.
For a confidence interval for a proportion, the margin of error is $z^\star \times SE$.

⊙ **Guided Practice 5.20** To have a smaller margin or error, should one use a larger sample or a smaller sample?[8]

⊙ **Guided Practice 5.21** What is the margin of error for the confidence interval: (0.035, 0.145)?[9]

[8]Intuitively, a larger sample should tend to yield less error. We can also note that n, the sample size is in the denominator of the SE formula, so a n goes up, the SE and thus the margin of error go down.

[9]Because we both add and subtract the margin of error to get the confidence interval, the margin of error is *half* of the width of the interval. $(0.145 - 0.035)/2 = 0.055$.

Appendix A

End of chapter exercise solutions

1 Data collection

1.15 The estimate will be biased, and it will tend to overestimate the true family size. For example, suppose we had just two families: the first with 2 parents and 5 children, and the sec- ond with 2 parents and 1 child. Then if we draw one of the six children at random, 5 times out of 6 we would sample the larger family

1.17 (a) No, this is an observational study. (b) This statement is not justified; it implies a casual association between sleep disorders and bullying. However, this was an observational study. A better conclusion would be "School children identified as bullies are more likely to suffer from sleep disorders than non-bullies."

1.19 (a) Experiment, as the treatment was as- signed to each patient. (b) Response: Duration of the cold. Explanatory: Treatment, with 4 levels: *placebo*, *1g*, *3g*, *3g with additives*. (c) Pa- tients were blinded. (d) Double-blind with re- spect to the researchers evaluating the patients, but the nurses who briefly interacted with pa- tients during the distribution of the medica- tion were not blinded. We could say the study was partly double-blind. (e) No. The patients were randomly assigned to treatment groups and were blinded, so we would expect about an equal number of patients in each group to not adhere to the treatment.

1.21 (a) Experiment. (b) Treatment is exercise twice a week. Control is no exercise. (c) Yes, the blocking variable is age. (d) No. (e) This is an experiment, so a causal conclusion is reasonable. Since the sample is random, the conclusion can be generalized to the population at large. How- ever, we must consider that a placebo effect is possible. (f) Yes. Randomly sampled people should not be required to participate in a clinical trial, and there are also ethical concerns about the plan to instruct one group not to partici- pate in a healthy behavior, which in this case is exercise.

4.33 This is the same as checking that the av- erage bag weight of the 10 bags is greater than 46 lbs. $SD_{\bar{x}} = \frac{3.2}{\sqrt{10}} = 1.012$; $z = \frac{46-45}{1.012} = 0.988$; $P(z > 0.988) = 0.162 = 16.2\%$.

4.35 (a) No. The cards are not independent. For example, if the first card is an ace of clubs, that implies the second card cannot be an ace of clubs. Additionally, there are many possible categories, which would need to be simplified. (b) No. There are six events under considera- tion. The Bernoulli distribution allows for only two events or categories. Note that rolling a die could be a Bernoulli trial if we simply to two events, e.g. rolling a 6 and not rolling a 6, though specifying such details would be neces- sary.

4.37 (a) $(1 - 0.471)^2 \times 0.471 = 0.1318$.
(b) $0.471^3 = 0.1045$. (c) $\mu = 1/0.471 = 2.12$,
$\sigma = \sqrt{2.38} = 1.54$. (d) $\mu = 1/0.30 = 3.33$,
$\sigma = 2.79$. (e) When p is smaller, the event is rarer, meaning the expected number of trials be- fore a success and the standard deviation of the waiting time are higher.

4.39 (a) $0.875^2 \times 0.125 = 0.096$. (b) $\mu = 8$, $\sigma = 7.48$.

4.41 (a) $\mu = 35$, $\sigma = 3.24$. (b) Yes. $Z = 3.09$. Since 45 is more than 2 standard deviations from the mean, it would be considered unusual. Note that the normal model is not required to apply this rule of thumb. (c) Using a normal model: 0.0010. This does indeed appear to be an un- usual observation. If using a normal model with a 0.5 correction, the probability would be calculated as 0.0017.

4.43 Want to find the probabiliy that there will be more than 1,786 enrollees. Using the normal model: 0.0537. With a 0.5 correction: 0.0559.

5 Foundation for inference

5.7 (a) False. Confidence intervals provide a range of plausible values, and sometimes the truth is missed. A 95% confidence interval "misses" about 5% of the time. (b) True. Notice that the description focuses on the true popu- lation value. (c) True. If we examine the 95% confidence interval computed in Exercise 5.5, we can see that 50% is not included in this interval. This means that in a hypothesis test, we would reject the null hypothesis that the proportion is 0.5. (d) False. The standard error describes the uncertainty in the overall estimate from natural fluctuations due to randomness, not the uncer- tainty corresponding to individuals' responses.

5.9 The subscript $_{pr}$ corresponds to provoca- tive and $_{con}$ to conservative. (a) $H_0 : p_{pr} = p_{con}$. $H_A : p_{pr} \neq p_{con}$. (b) -0.35. (c) The left tail for the p-value is calculated by adding up the two left bins: 0.005 + 0.015 = 0.02. Doubling the one tail, the p-value is 0.04. (Students may have approximate results, and a small number of students may have a p-value of about 0.05.) Since the p-value is low, we reject H_0. The data provide strong evidence that people react differ- ently under the two scenarios.

5.11 The primary concern is confirmation bias. If researchers look only for what they suspect to be true using a one-sided test, then they are for- mally excluding from consideration the possibil- ity that the opposite result is true. Additionally, if other researchers believe the opposite possibility might be true, they would be very skeptical of the one-sided test.

5.13 (a) $H_0 : p = 0.69$. $H_A : p \neq 0.69$. (b) $\hat{p} = \frac{17}{30} = 0.57$. (c) The success-failure condition is not satisfied; note that it is appropriate to use the null value ($p_0 = 0.69$) to compute the expected number of successes and failures. (d) Answers may vary. Each student can be represented with a card. Take 100 cards, 69 black cards representing those who follow the news about Egypt and 31 red cards represent- ing those who do not. Shuffle the cards and draw with replacement (shuffling each time in between draws) 30 cards representing the 30 high school students. Calculate the proportion of black cards in this sample, $\hat{p}_{sim}$, i.e. the pro- portion of those who follow the news in the sim- ulation. Repeat this many times (e.g. 10,000 times) and plot the resulting sample propor- tions. The p-value will be two times the propor- tion of simulations where $\hat{p}_{sim} \leq 0.57$. (Note: we would generally use a computer to perform these simulations.) (e) The p-value is about 0.001 + 0.005 + 0.020 + 0.035 + 0.075 = 0.136, meaning the two-sided p-value is about 0.272. Your p-value may vary slightly since it is based on a visual estimate. Since the p-value is greater than 0.05, we fail to reject H_0. The data do not provide strong evidence that the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.