When we compute the confidence interval for $\mu_1 - \mu_2$, the point estimate is the difference in sample means, the value $z^\star$ corresponds to the confidence level, and the standard error is computed from Equation (5.4) on page 217. While the point estimate and standard error formulas change a little, the framework for a confidence interval stays the same. This is also true in hypothesis tests for differences of means.

In a hypothesis test, we apply the standard framework and use the specific formulas for the point estimate and standard error of a difference in two means. The test statistic represented by the Z score may be computed as

$$Z = \frac{\text{point estimate} - \text{null value}}{SE}$$

When assessing the difference in two means, the point estimate takes the form $\bar{x}_1 - \bar{x}_2$, and the standard error again takes the form of Equation (5.4) on page 217. Finally, the null value is the difference in sample means under the null hypothesis. Just as in Chapter 4, the test statistic Z is used to identify the p-value.

5.2.5 Examining the standard error formula

The formula for the standard error of the difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean, $\bar{x}_1$, can be approximated by

$$SE_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}}$$

where s_1 and n_1 represent the sample standard deviation and sample size.

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.13}$$

This special relationship follows from probability theory.

⊙ **Exercise 5.14** Prerequisite: Section 2.4. We can rewrite Equation (5.13) in a different way:

$$SE_{\bar{x}_1 - \bar{x}_2}^2 = SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2$$

Explain where this formula comes from using the ideas of probability theory.[10]

5.3 One-sample means with the t distribution

The motivation in Chapter 4 for requiring a large sample was two-fold. First, a large sample ensures that the sampling distribution of $\bar{x}$ is nearly normal. We will see in Section 5.3.1 that if the population data are nearly normal, then $\bar{x}$ is also nearly normal regardless of the

[10] The standard error squared represents the variance of the estimate. If X and Y are two random variables with variances σ_x^2 and σ_y^2, then the variance of $X - Y$ is $\sigma_x^2 + \sigma_y^2$. Likewise, the variance corresponding to $\bar{x}_1 - \bar{x}_2$ is $\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2$. Because $\sigma_{\bar{x}_1}^2$ and $\sigma_{\bar{x}_2}^2$ are just another way of writing $SE_{\bar{x}_1}^2$ and $SE_{\bar{x}_2}^2$, the variance associated with $\bar{x}_1 - \bar{x}_2$ may be written as $SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2$.

sample size. The second motivation for a large sample was that we get a better estimate of the standard error when using a large sample. The standard error estimate will not generally be accurate for smaller sample sizes, and this motivates the introduction of the t distribution, which we introduce in Section 5.3.2.

We will see that the t distribution is a helpful substitute for the normal distribution when we model a sample mean $\bar{x}$ that comes from a small sample. While we emphasize the use of the t distribution for small samples, this distribution may also be used for means from large samples.

5.3.1 The normality condition

We use a special case of the Central Limit Theorem to ensure the distribution of the sample means will be nearly normal, regardless of sample size, provided the data come from a nearly normal distribution.

Central Limit Theorem for normal data
The sampling distribution of the mean is nearly normal when the sample observations are independent and come from a nearly normal distribution. This is true for any sample size.

While this seems like a very helpful special case, there is one small problem. It is inherently difficult to verify normality in small data sets.

Caution: Checking the normality condition
We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from. For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

You may relax the normality condition as the sample size goes up. If the sample size is 10 or more, slight skew is not problematic. Once the sample size hits about 30, then moderate skew is reasonable. Data with strong skew or outliers require a more cautious analysis.

5.3.2 Introducing the t distribution

The second reason we previously required a large sample size was so that we could accurately estimate the standard error using the sample data. In the cases where we will use a small sample to calculate the standard error, it will be useful to rely on a new distribution for inference calculations: the t distribution. A t distribution, shown as a solid line in Figure 5.10, has a bell shape. However, its tails are thicker than the normal model's. This means observations are more likely to fall beyond two standard deviations from the mean than under the normal distribution.[11] These extra thick tails are exactly the correction we need to resolve the problem of a poorly estimated standard error.

The t distribution, always centered at zero, has a single parameter: degrees of freedom. The **degrees of freedom (df)** describe the precise form of the bell-shaped t distribution.

[11] The standard deviation of the t distribution is actually a little more than 1. However, it is useful to always think of the t distribution as having a standard deviation of 1 in all of our applications.

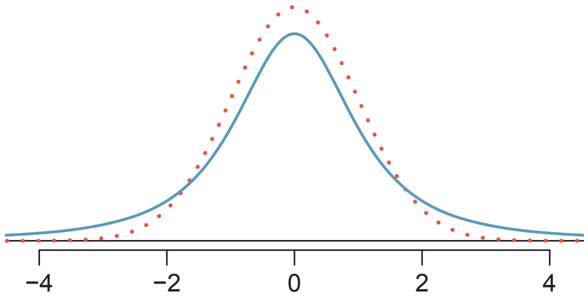

Figure 5.10: Comparison of a t distribution (solid line) and a normal distribution (dotted line).

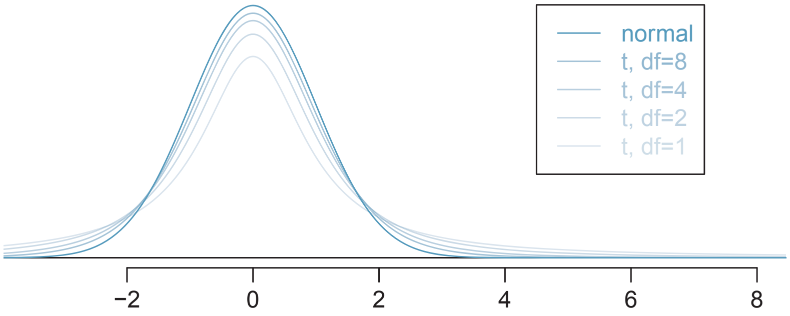

Figure 5.11: The larger the degrees of freedom, the more closely the t distribution resembles the standard normal model.

Several t distributions are shown in Figure 5.11. When there are more degrees of freedom, the t distribution looks very much like the standard normal distribution.

Degrees of freedom (df)
The degrees of freedom describe the shape of the t distribution. The larger the degrees of freedom, the more closely the distribution approximates the normal model.

When the degrees of freedom is about 30 or more, the t distribution is nearly indistinguishable from the normal distribution. In Section 5.3.3, we relate degrees of freedom to sample size.

We will find it very useful to become familiar with the t distribution, because it plays a very similar role to the normal distribution during inference for small samples of numerical data. We use a **t table**, partially shown in Table 5.12, in place of the normal probability table for small sample numerical data. A larger table is presented in Appendix B.2 on page 410.

Each row in the t table represents a t distribution with different degrees of freedom. The columns correspond to tail probabilities. For instance, if we know we are working with the t distribution with $df = 18$, we can examine row 18, which is **highlighted** in

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	**1.33**	**1.73**	**2.10**	**2.55**	**2.88**
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

Table 5.12: An abbreviated look at the t table. Each row represents a different t distribution. The columns describe the cutoffs for specific tail areas. The row with *df* $= 18$ has been **highlighted**.

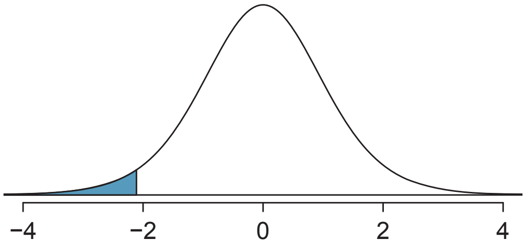

Figure 5.13: The t distribution with 18 degrees of freedom. The area below -2.10 has been shaded.

Table 5.12. If we want the value in this row that identifies the cutoff for an upper tail of 10%, we can look in the column where *one tail* is 0.100. This cutoff is 1.33. If we had wanted the cutoff for the lower 10%, we would use -1.33. Just like the normal distribution, all t distributions are symmetric.

● **Example 5.15** What proportion of the t distribution with 18 degrees of freedom falls below -2.10?

Just like a normal probability problem, we first draw the picture in Figure 5.13 and shade the area below -2.10. To find this area, we identify the appropriate row: *df* $= 18$. Then we identify the column containing the absolute value of -2.10; it is the third column. Because we are looking for just one tail, we examine the top line of the table, which shows that a one tail area for a value in the third row corresponds to 0.025. About 2.5% of the distribution falls below -2.10. In the next example we encounter a case where the exact t value is not listed in the table.

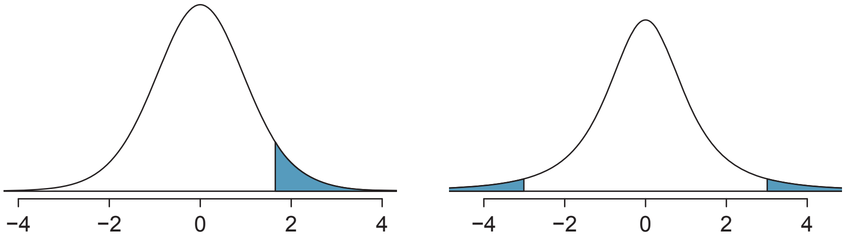

Figure 5.14: Left: The t distribution with 20 degrees of freedom, with the area above 1.65 shaded. Right: The t distribution with 2 degrees of freedom, with the area further than 3 units from 0 shaded.

● **Example 5.16** A t distribution with 20 degrees of freedom is shown in the left panel of Figure 5.14. Estimate the proportion of the distribution falling above 1.65.

We identify the row in the t table using the degrees of freedom: $df = 20$. Then we look for 1.65; it is not listed. It falls between the first and second columns. Since these values bound 1.65, their tail areas will bound the tail area corresponding to 1.65. We identify the one tail area of the first and second columns, 0.050 and 0.10, and we conclude that between 5% and 10% of the distribution is more than 1.65 standard deviations above the mean. If we like, we can identify the precise area using statistical software: 0.0573.

● **Example 5.17** A t distribution with 2 degrees of freedom is shown in the right panel of Figure 5.14. Estimate the proportion of the distribution falling more than 3 units from the mean (above or below).

As before, first identify the appropriate row: $df = 2$. Next, find the columns that capture 3; because $2.92 < 3 < 4.30$, we use the second and third columns. Finally, we find bounds for the tail areas by looking at the two tail values: 0.05 and 0.10. We use the two tail values because we are looking for two (symmetric) tails.

⊙ **Exercise 5.18** What proportion of the t distribution with 19 degrees of freedom falls above -1.79 units?[12]

5.3.3 The t distribution as a solution to the standard error problem

When estimating the mean and standard error from a small sample, the t distribution is a more accurate tool than the normal model. This is true for both small and large samples.

TIP: When to use the t distribution
Use the t distribution for inference of the sample mean when observations are independent and nearly normal. You may relax the nearly normal condition as the sample size increases. For example, the data distribution may be moderately skewed when the sample size is at least 30.

[12] We find the shaded area *above* -1.79 (we leave the picture to you). The small left tail is between 0.025 and 0.05, so the larger upper region must have an area between 0.95 and 0.975.

To proceed with the t distribution for inference about a single mean, we must check two conditions.

Independence of observations. We verify this condition just as we did before. We collect a simple random sample from less than 10% of the population, or if it was an experiment or random process, we carefully check to the best of our abilities that the observations were independent.

Observations come from a nearly normal distribution. This second condition is difficult to verify with small data sets. We often (i) take a look at a plot of the data for obvious departures from the normal model, and (ii) consider whether any previous experiences alert us that the data may not be nearly normal.

When examining a sample mean and estimated standard error from a sample of n independent and nearly normal observations, we use a t distribution with $n - 1$ degrees of freedom (df). For example, if the sample size was 19, then we would use the t distribution with $df = 19 - 1 = 18$ degrees of freedom and proceed exactly as we did in Chapter 4, except that *now we use the t table.*

5.3.4 One sample t confidence intervals

Dolphins are at the top of the oceanic food chain, which causes dangerous substances such as mercury to concentrate in their organs and muscles. This is an important problem for both dolphins and other animals, like humans, who occasionally eat them. For instance, this is particularly relevant in Japan where school meals have included dolphin at times.

Figure 5.15: A Risso's dolphin.

Photo by Mike Baird (http://www.bairdphotos.com/).

Here we identify a confidence interval for the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan.[13] The data are summarized in Table 5.16. The minimum and maximum observed values can be used to evaluate whether or not there are obvious outliers or skew.

[13] Taiji was featured in the movie *The Cove*, and it is a significant source of dolphin and whale meat in Japan. Thousands of dolphins pass through the Taiji area annually, and we will assume these 19 dolphins represent a simple random sample from those dolphins. Data reference: Endo T and Haraguchi K. 2009. High mercury levels in hair samples from residents of Taiji, a Japanese whaling town. Marine Pollution Bulletin 60(5):743-747.

n	$\bar{x}$	s	minimum	maximum
19	4.4	2.3	1.7	9.2

Table 5.16: Summary of mercury content in the muscle of 19 Risso's dolphins from the Taiji area. Measurements are in μg/wet g (micrograms of mercury per wet gram of muscle).

Example 5.19 Are the independence and normality conditions satisfied for this data set?

The observations are a simple random sample and consist of less than 10% of the population, therefore independence is reasonable. The summary statistics in Table 5.16 do not suggest any skew or outliers; all observations are within 2.5 standard deviations of the mean. Based on this evidence, the normality assumption seems reasonable.

In the normal model, we used $z^\star$ and the standard error to determine the width of a confidence interval. We revise the confidence interval formula slightly when using the t distribution:

$$\bar{x} \ \pm \ t^\star_{df} SE$$

$t^\star_{df}$
Multiplication factor for
t conf. interval

The sample mean and estimated standard error are computed just as before ($\bar{x} = 4.4$ and $SE = s/\sqrt{n} = 0.528$). The value $t^\star_{df}$ is a cutoff we obtain based on the confidence level and the t distribution with df degrees of freedom. Before determining this cutoff, we will first need the degrees of freedom.

Degrees of freedom for a single sample
If the sample has n observations and we are examining a single mean, then we use the t distribution with $df = n - 1$ degrees of freedom.

In our current example, we should use the t distribution with $df = 19 - 1 = 18$ degrees of freedom. Then identifying $t^\star_{18}$ is similar to how we found $z^\star$.

- For a 95% confidence interval, we want to find the cutoff $t^\star_{18}$ such that 95% of the t distribution is between -$t^\star_{18}$ and $t^\star_{18}$.
- We look in the t table on page 224, find the column with area totaling 0.05 in the two tails (third column), and then the row with 18 degrees of freedom: $t^\star_{18} = 2.10$.

Generally the value of $t^\star_{df}$ is slightly larger than what we would get under the normal model with $z^\star$.

Finally, we can substitute all our values into the confidence interval equation to create the 95% confidence interval for the average mercury content in muscles from Risso's dolphins that pass through the Taiji area:

$$\bar{x} \ \pm \ t^\star_{18} SE \quad \rightarrow \quad 4.4 \ \pm \ 2.10 \times 0.528 \quad \rightarrow \quad (3.29, 5.51)$$

We are 95% confident the average mercury content of muscles in Risso's dolphins is between 3.29 and 5.51 μg/wet gram. This is above the Japanese regulation level of 0.4 μg/wet gram.

Finding a t confidence interval for the mean

Based on a sample of n independent and nearly normal observations, a confidence interval for the population mean is

$$\bar{x} \ \pm\ t^{\star}_{df} SE$$

where $\bar{x}$ is the sample mean, $t^{\star}_{df}$ corresponds to the confidence level and degrees of freedom, and SE is the standard error as estimated by the sample.

⊙ **Exercise 5.20** The FDA's webpage provides some data on mercury content of fish.[14] Based on a sample of 15 croaker white fish (Pacific), a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent. Based on the summary statistics of the data, do you have any objections to the normality condition of the individual observations?[15]

● **Example 5.21** Estimate the standard error of $\bar{x} = 0.287$ ppm using the data summaries in Exercise 5.20. If we are to use the t distribution to create a 90% confidence interval for the actual mean of the mercury content, identify the degrees of freedom we should use and also find $t^{\star}_{df}$.

The standard error: $SE = \frac{0.069}{\sqrt{15}} = 0.0178$. Degrees of freedom: $df = n - 1 = 14$.

Looking in the column where two tails is 0.100 (for a 90% confidence interval) and row $df = 14$, we identify $t^{\star}_{14} = 1.76$.

⊙ **Exercise 5.22** Using the results of Exercise 5.20 and Example 5.21, compute a 90% confidence interval for the average mercury content of croaker white fish (Pacific).[16]

5.3.5 One sample t tests

An SAT preparation company claims that its students' scores improve by over 100 points on average after their course. A consumer group would like to evaluate this claim, and they collect data on a random sample of 30 students who took the class. Each of these students took the SAT before and after taking the company's course, and so we have a difference in scores for each student. We will examine these differences $x_1 = 57$, $x_2 = 133$, ..., $x_{30} = 140$ as a sample to evaluate the company's claim. (This is *paired data*, so we analyze the score differences; for a review of the ideas of paired data, see Section 5.1.) The distribution of the differences, shown in Figure 5.17, has mean 135.9 and standard deviation 82.2. Do these data provide convincing evidence to back up the company's claim?

⊙ **Exercise 5.23** Set up hypotheses to evaluate the company's claim. Use μ_{diff} to represent the true average difference in student scores.[17]

[14] http://www.fda.gov/food/foodborneillnesscontaminants/metals/ucm115644.htm

[15] There are no obvious outliers; all observations are within 2 standard deviations of the mean. If there is skew, it is not evident. There are no red flags for the normal model based on this (limited) information, and we do not have reason to believe the mercury content is not nearly normal in this type of fish.

[16] $\bar{x} \ \pm\ t^{\star}_{14} SE \ \rightarrow\ 0.287 \ \pm\ 1.76 \times 0.0178 \ \rightarrow\ (0.256, 0.318)$. We are 90% confident that the average mercury content of croaker white fish (Pacific) is between 0.256 and 0.318 ppm.

[17] This is a one-sided test. H_0: student scores do not improve by more than 100 after taking the company's course. $\mu_{diff} = 100$ (we always write the null hypothesis with an equality). H_A: students scores improve by more than 100 points on average after taking the company's course. $\mu_{diff} > 100$.

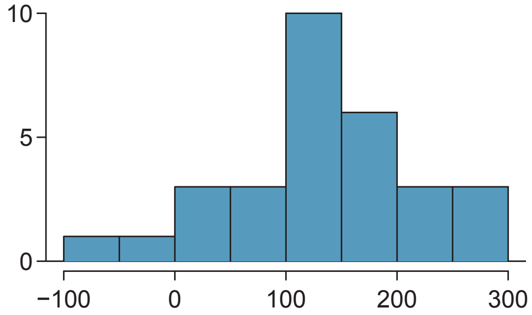

Figure 5.17: Sample distribution of improvements in SAT scores after taking the SAT course. The distribution is approximately symmetric.

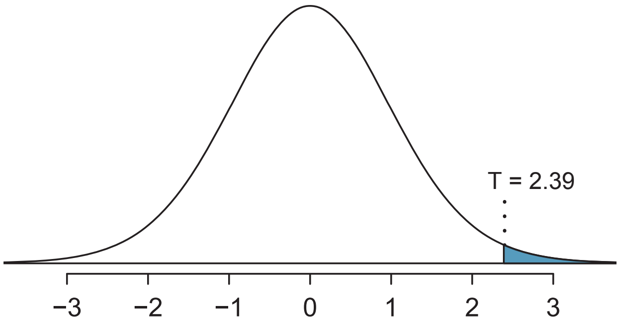

Figure 5.18: The t distribution with 29 degrees of freedom.

⊙ **Exercise 5.24** Are the conditions to use the t distribution method satisfied?[18]

Just as we did for the normal case, we standardize the sample mean using the Z score to identify the test statistic. However, we will write T instead of Z, because we have a small sample and are basing our inference on the t distribution:

T
T score
(like Z score)

$$T = \frac{\bar{x} - \text{null value}}{SE} = \frac{135.9 - 100}{82.2/\sqrt{30}} = 2.39$$

If the null hypothesis was true, the test statistic T would follow a t distribution with $df = n - 1 = 29$ degrees of freedom. We can draw a picture of this distribution and mark the observed T, as in Figure 5.18. The shaded right tail represents the p-value: the probability of observing such strong evidence in favor of the SAT company's claim, if the average student improvement is really only 100.

[18] This is a random sample from less than 10% of the company's students (assuming they have more than 300 former students), so the independence condition is reasonable. The normality condition also seems reasonable based on Figure 5.17. We can use the t distribution method. Note that we could use the normal distribution. However, since the sample size ($n = 30$) just meets the threshold for reasonably estimating the standard error, it is advisable to use the t distribution.

⊙ **Exercise 5.25** Use the t table in Appendix B.2 on page 410 to identify the p-value. What do you conclude?[19]

⊙ **Exercise 5.26** Because we rejected the null hypothesis, does this mean that taking the company's class improves student scores by more than 100 points on average?[20]

5.4 The t distribution for the difference of two means

It is also useful to be able to compare two means for small samples. For instance, a teacher might like to test the notion that two versions of an exam were equally difficult. She could do so by randomly assigning each version to students. If she found that the average scores on the exams were so different that we cannot write it off as chance, then she may want to award extra points to students who took the more difficult exam.

In a medical context, we might investigate whether embryonic stem cells can improve heart pumping capacity in individuals who have suffered a heart attack. We could look for evidence of greater heart health in the stem cell group against a control group.

In this section we use the t distribution for the difference in sample means. We will again drop the minimum sample size condition and instead impose a strong condition on the distribution of the data.

5.4.1 Sampling distributions for the difference in two means

In the example of two exam versions, the teacher would like to evaluate whether there is convincing evidence that the difference in average scores between the two exams is not due to chance.

It will be useful to extend the t distribution method from Section 5.3 to apply to a difference of means:

$$\bar{x}_1 - \bar{x}_2 \qquad \text{as a point estimate for} \qquad \mu_1 - \mu_2$$

Our procedure for checking conditions mirrors what we did for large samples in Section 5.2. First, we verify the small sample conditions (independence and nearly normal data) for each sample separately, then we verify that the samples are also independent. For instance, if the teacher believes students in her class are independent, the exam scores are nearly normal, and the students taking each version of the exam were independent, then we can use the t distribution for inference on the point estimate $\bar{x}_1 - \bar{x}_2$.

The formula for the standard error of $\bar{x}_1 - \bar{x}_2$, introduced in Section 5.2, also applies to small samples:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.27}$$

[19] We use the row with 29 degrees of freedom. The value $T = 2.39$ falls between the third and fourth columns. Because we are looking for a single tail, this corresponds to a p-value between 0.01 and 0.025. The p-value is guaranteed to be less than 0.05 (the default significance level), so we reject the null hypothesis. The data provide convincing evidence to support the company's claim that student scores improve by more than 100 points following the class.

[20] This is an observational study, so we cannot make this causal conclusion. For instance, maybe SAT test takers tend to improve their score over time even if they don't take a special SAT class, or perhaps only the most motivated students take such SAT courses.