⊙ **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:

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

[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.

Because we will use the t distribution, we will need to identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will apply in the examples and exercises.[21]

Using the t distribution for a difference in means

The t distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t distribution and (2) the samples are independent. We estimate the standard error of the difference of two means using Equation (5.27).

5.4.2 Two sample t test

Summary statistics for each exam version are shown in Table 5.19. The teacher would like to evaluate whether this difference is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Version	n	$\bar{x}$	s	min	max
A	30	79.4	14	45	100
B	27	74.1	20	32	100

Table 5.19: Summary statistics of scores for each exam version.

⊙ **Exercise 5.28** Construct a two-sided hypothesis test to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, might be due to chance.[22]

⊙ **Exercise 5.29** To evaluate the hypotheses in Exercise 5.28 using the t distribution, we must first verify assumptions. (a) Does it seem reasonable that the scores are independent within each group? (b) What about the normality condition for each group? (c) Do you think scores from the two groups would be independent of each other (i.e. the two samples are independent)?[23]

After verifying the conditions for each sample and confirming the samples are independent of each other, we are ready to conduct the test using the t distribution. In this case, we are estimating the true difference in average test scores using the sample data, so the point estimate is $\bar{x}_A - \bar{x}_B = 5.3$. The standard error of the estimate can be calculated using Equation (5.27):

$$SE = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{14^2}{30} + \frac{20^2}{27}} = 4.62$$

[21] This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this *df* method.

[22] Because the teacher did not expect one exam to be more difficult prior to examining the test results, she should use a two-sided hypothesis test. H_0: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. H_A: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

[23] (a) It is probably reasonable to conclude the scores are independent. (b) The summary statistics suggest the data are roughly symmetric about the mean, and it doesn't seem unreasonable to suggest the data might be normal. Note that since these samples are each nearing 30, moderate skew in the data would be acceptable. (c) It seems reasonable to suppose that the samples are independent since the exams were handed out randomly.

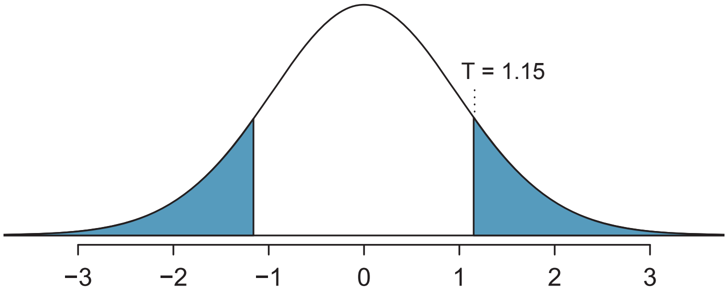

Figure 5.20: The t distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

Finally, we construct the test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE} = \frac{(79.4 - 74.1) - 0}{4.62} = 1.15$$

If we have a computer handy, we can identify the degrees of freedom as 45.97. Otherwise we use the smaller of $n_1 - 1$ and $n_2 - 1$: $df = 26$.

⊙ **Exercise 5.30** Identify the p-value, shown in Figure 5.20. Use $df = 26$.[24]

In Exercise 5.30, we could have used $df = 45.97$. However, this value is not listed in the table. In such cases, we use the next lower degrees of freedom (unless the computer also provides the p-value). For example, we could have used $df = 45$ but not $df = 46$.

⊙ **Exercise 5.31** Do embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 5.21 contains summary statistics for an experiment to test ESCs in sheep that had a heart attack. Each of these sheep was randomly assigned to the ESC or control group, and the change in their hearts' pumping capacity was measured. A positive value generally corresponds to increased pumping capacity, which suggests a stronger recovery.

(a) Set up hypotheses that will be used to test whether there is convincing evidence that ESCs actually increase the amount of blood the heart pumps. (b) Check conditions for using the t distribution for inference with the point estimate $\bar{x}_1 - \bar{x}_2$. To assist in this assessment, the data are presented in Figure 5.22.[25]

[24] We examine row $df = 26$ in the t table. Because this value is smaller than the value in the left column, the p-value is larger than 0.200 (two tails!). Because the p-value is so large, we do not reject the null hypothesis. That is, the data do not convincingly show that one exam version is more difficult than the other, and the teacher should not be convinced that she should add points to the Version B exam scores.

[25] (a) We first setup the hypotheses:

H_0: The stem cells do not improve heart pumping function. $\mu_{esc} - \mu_{control} = 0$.

H_A: The stem cells do improve heart pumping function. $\mu_{esc} - \mu_{control} > 0$.

(b) Because the sheep were randomly assigned their treatment and, presumably, were kept separate from one another, the independence assumption is reasonable for each sample as well as for between samples. The data are very limited, so we can only check for obvious outliers in the raw data in Figure 5.22. Since the distributions are (very) roughly symmetric, we will assume the normality condition is acceptable. Because the conditions are satisfied, we can apply the t distribution.

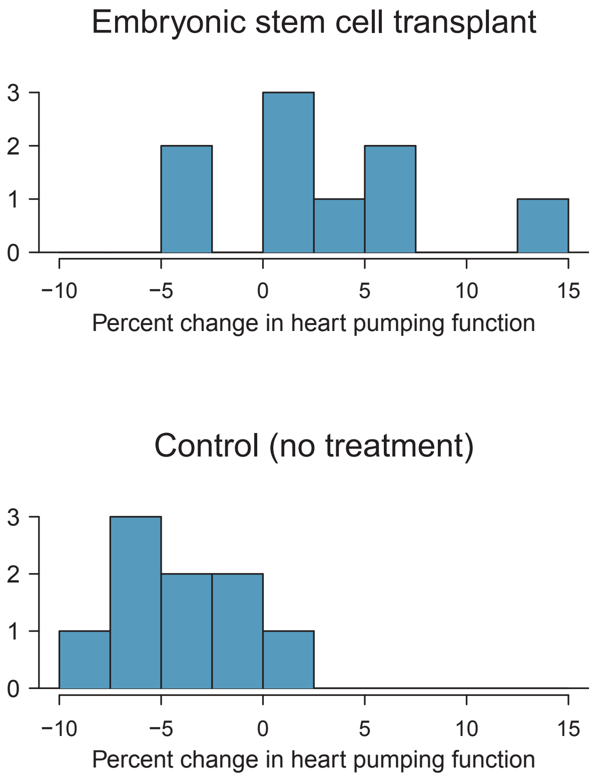

Figure 5.22: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

	n	$\bar{x}$	s
ESCs	9	3.50	5.17
control	9	-4.33	2.76

Table 5.21: Summary statistics for the embryonic stem cell data set.

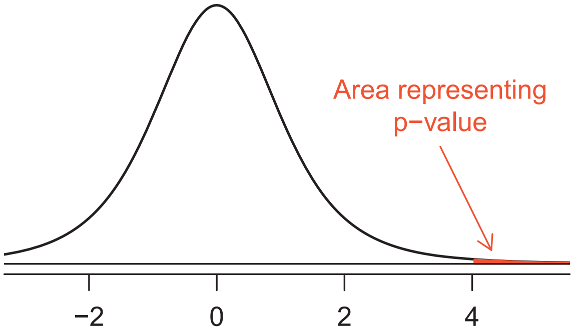

Figure 5.23: Distribution of the sample difference of the test statistic if the null hypothesis was true. The shaded area, hardly visible in the right tail, represents the p-value.

● **Example 5.32** Use the data from Table 5.21 and $df = 8$ to evaluate the hypotheses for the ESC experiment described in Exercise 5.31.

First, we compute the sample difference and the standard error for that point estimate:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83$$

$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

The p-value is depicted as the shaded slim right tail in Figure 5.23, and the test statistic is computed as follows:

$$T = \frac{7.83 - 0}{1.95} = 4.02$$

We use the smaller of $n_1 - 1$ and $n_2 - 1$ (each are the same) for the degrees of freedom: $df = 8$. Finally, we look for $T = 4.02$ in the t table; it falls to the right of the last column, so the p-value is smaller than 0.005 (one tail!). Because the p-value is less than 0.005 and therefore also smaller than 0.05, we reject the null hypothesis. The data provide convincing evidence that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack.

5.4.3 Two sample t confidence interval

The results from the previous section provided evidence that ESCs actually help improve the pumping function of the heart. But how large is this improvement? To answer this question, we can use a confidence interval.

⊙ **Exercise 5.33** In Exercise 5.31, you found that the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$, has a standard error of 1.95. Using $df = 8$, create a 99% confidence interval for the improvement due to ESCs.[26]

5.4.4 Pooled standard deviation estimate (special topic)

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make our t distribution approach slightly more precise by using a pooled standard deviation.

The **pooled standard deviation** of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s_1 and s_2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s_{pooled}^2 = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s_{pooled}^2 in place of s_1^2 and s_2^2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2$$

The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger degrees of freedom parameter for the t distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$.

Caution: Pooling standard deviations should be done only after careful research

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

[26] We know the point estimate, 7.83, and the standard error, 1.95. We also verified the conditions for using the t distribution in Exercise 5.31. Thus, we only need identify $t_8^\star$ to create a 99% confidence interval: $t_8^\star = 3.36$. The 99% confidence interval for the improvement from ESCs is given by

$$\text{point estimate} \ \pm\ t_8^\star SE \quad \rightarrow \quad 7.83 \ \pm\ 3.36 \times 1.95 \quad \rightarrow \quad (1.33, 14.43)$$

That is, we are 99% confident that the true improvement in heart pumping function is somewhere between 1.33% and 14.43%.