⊙ **Guided Practice 4.17** Create a 95% confidence interval for the average price difference between books at the UCLA bookstore and books on Amazon.[14]

In the textbook price example, we applied the t distribution. However, as we mentioned in the last section, the t distribution looks a lot like the normal distribution when the degrees of freedom are larger than about 30. In such cases, including this one, it would be reasonable to use the normal distribution in place of the t distribution.

4.3 Difference of two means

In this section we consider a difference in two population means, $\mu_1 - \mu_2$, under the condition that the data are not paired. Just as with a single sample, we identify conditions to ensure we can use the t distribution with a point estimate of the difference, $\bar{x}_1 - \bar{x}_2$.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the impact of pregnant womens' smoking habits on birth weights of newborns, and exploring whether there is statistically significant evidence that one variations of an exam is harder than another variation. This section is motivated by questions like "Is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?"

4.3.1 Confidence interval for a differences of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Table 4.15 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 in the study. A positive value corresponds to increased pumping capacity, which generally suggests a stronger recovery. Our goal will be to identify a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity relative to the control group.

A point estimate of the difference in the heart pumping variable can be found using the difference in the sample means:

$$\bar{x}_{esc} - \bar{x}_{control} = 3.50 - (-4.33) = 7.83$$

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

Table 4.15: Summary statistics of the embryonic stem cell study.

[14] Conditions have already verified and the standard error computed in Example 4.16. To find the interval, identify $t^\star_{72}$ (use $df = 70$ in the table, $t^\star_{70} = 1.99$) and plug it, the point estimate, and the standard error into the confidence interval formula:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 12.76 \pm 1.99 \times 1.67 \quad \rightarrow \quad (9.44, 16.08)$$

We are 95% confident that Amazon is, on average, between $9.44 and $16.08 cheaper than the UCLA bookstore for UCLA course books.

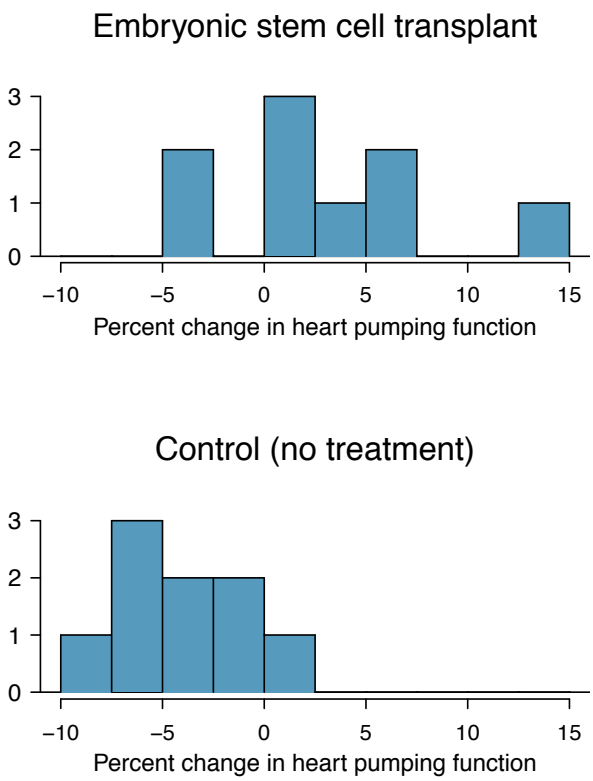


Figure 4.16: Histograms for both the embryonic stem cell group and the control group. Higher values are associated with greater improvement. We don't see any evidence of skew in these data; however, it is worth noting that skew would be difficult to detect with such a small sample.

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.

Example 4.18 Can the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$, be analyzed using the t distribution?

We check the two required conditions:

1. In this study, the sheep were independent of each other. Additionally, the distributions in Figure 4.16 don't show any clear deviations from normality, where we watch for prominent outliers in particular for such small samples. These findings imply each sample mean could itself be modeled using a t distribution.
2. The sheep in each group were also independent of each other.

Because both conditions are met, we can use the t distribution to model the difference of the two sample means.

Before we construct a confidence interval, we must calculate the standard error of the point estimate of the difference. For this, we use the following formula, where just as before we substitute the sample standard deviations into the formula:

$$SE_{\bar{x}_{esc}-\bar{x}_{control}} = \sqrt{\frac{\sigma_{esc}^2}{n_{esc}} + \frac{\sigma_{control}^2}{n_{control}}}$$
$$\approx \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Because we will use the t distribution, we also must 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 typically apply in the examples and guided practice.[15]

Distribution of a difference of sample means

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t distribution and the standard error

$$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{4.19}$$

when each sample mean can itself be modeled using a t distribution and the samples are independent. To calculate the degrees of freedom, use statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.

Example 4.20 Calculate a 95% confidence interval for the effect of ESCs on the change in heart pumping capacity of sheep after they've suffered a heart attack.

We will use the sample difference and the standard error for that point estimate from our earlier calculations:

$$\bar{x}_{esc} - \bar{x}_{control} = 7.83$$
$$SE = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Using $df = 8$, we can identify the appropriate $t^\star_{df} = t^\star_8$ for a 95% confidence interval as 2.31. Finally, we can enter the values into the confidence interval formula:

$$\text{point estimate} \pm z^\star SE \quad \rightarrow \quad 7.83 \pm 2.31 \times 1.95 \quad \rightarrow \quad (3.38, 12.38)$$

We are 95% confident that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack by 3.38% to 12.38%.

[15] 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. In this example, computer software would have provided us a more precise degrees of freedom of $df = 12.225$.

4.3.2 Hypothesis tests based on a difference in means

A data set called `baby_smoke` represents a random sample of 150 cases of mothers and their newborns in North Carolina over a year. Four cases from this data set are represented in Table 4.17. We are particularly interested in two variables: `weight` and `smoke`. The `weight` variable represents the weights of the newborns and the `smoke` variable describes which mothers smoked during pregnancy. We would like to know, is there convincing evidence that newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke? We will use the North Carolina sample to try to answer this question. The smoking group includes 50 cases and the nonsmoking group contains 100 cases, represented in Figure 4.18.

	fAge	mAge	weeks	weight	sexBaby	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

Table 4.17: Four cases from the `baby_smoke` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

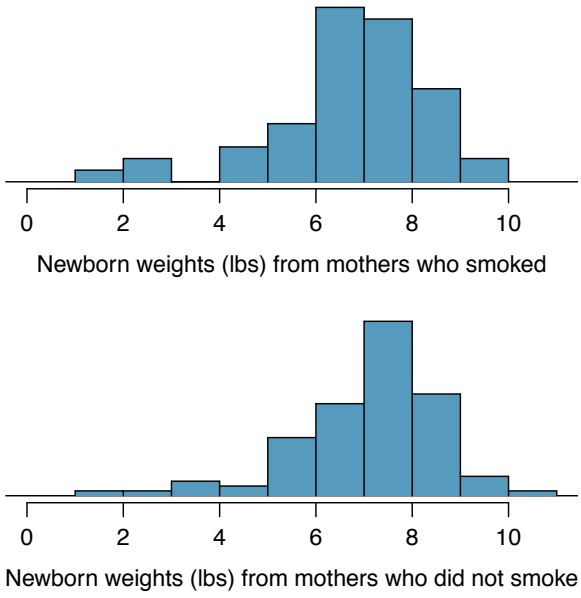


Figure 4.18: The top panel represents birth weights for infants whose mothers smoked. The bottom panel represents the birth weights for infants whose mothers who did not smoke. The distributions exhibit moderate-to-strong and strong skew, respectively.

Example 4.21 Set up appropriate hypotheses to evaluate whether there is a relationship between a mother smoking and average birth weight.

The null hypothesis represents the case of no difference between the groups.

H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke. In statistical notation: $\mu_n - \mu_s = 0$, where μ_n represents non-smoking mothers and μ_s represents mothers who smoked.

H_A: There is some difference in average newborn weights from mothers who did and did not smoke ($\mu_n - \mu_s \neq 0$).

We check the two conditions necessary to apply the t distribution to the difference in sample means. (1) Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent. Additionally, while each distribution is strongly skewed, the sample sizes of 50 and 100 would make it reasonable to model each mean separately using a t distribution. The skew is reasonable for these sample sizes of 50 and 100. (2) The independence reasoning applied in (1) also ensures the observations in each sample are independent. Since both conditions are satisfied, the difference in sample means may be modeled using a t distribution.

	`smoker`	`nonsmoker`
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

Table 4.19: Summary statistics for the `baby_smoke` data set.

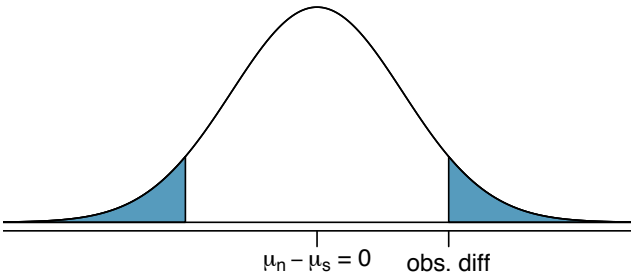
Guided Practice 4.22 The summary statistics in Table 4.19 may be useful for this exercise. (a) What is the point estimate of the population difference, $\mu_n - \mu_s$? (b) Compute the standard error of the point estimate from part (a).[16]

Example 4.23 Draw a picture to represent the p-value for the hypothesis test from Example 4.21.

To depict the p-value, we draw the distribution of the point estimate as though H_0 were true and shade areas representing at least as much evidence against H_0 as what was observed. Both tails are shaded because it is a two-sided test.

[16] (a) The difference in sample means is an appropriate point estimate: $\bar{x}_n - \bar{x}_s = 0.40$. (b) The standard error of the estimate can be estimated using Equation (4.19):

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}} = \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26$$



Example 4.24 Compute the p-value of the hypothesis test using the figure in Example 4.23, and evaluate the hypotheses using a significance level of $\alpha = 0.05$.

We start by computing the T score:

$$T = \frac{0.40 - 0}{0.26} = 1.54$$

Next, we compare this value to values in the t table in Appendix C.2 on page 342, where we use the smaller of $n_n - 1 = 99$ and $n_s - 1 = 49$ as the degrees of freedom: $df = 49$. The T score falls between the first and second columns in the $df = 49$ row of the t table, meaning the two-tailed p-value falls between 0.10 and 0.20 (reminder, find tail areas along the top of the table). This p-value is larger than the significance value, 0.05, so we fail to reject the null hypothesis. There is insufficient evidence to say there is a difference in average birth weight of newborns from North Carolina mothers who did smoke during pregnancy and newborns from North Carolina mothers who did not smoke during pregnancy.

Guided Practice 4.25 Does the conclusion to Example 4.24 mean that smoking and average birth weight are unrelated?[17]

Guided Practice 4.26 If we made a Type 2 Error and there is a difference, what could we have done differently in data collection to be more likely to detect such a difference?[18]

4.3.3 Case study: two versions of a course exam

An instructor decided to run two slight variations of the same exam. Prior to passing out the exams, she shuffled the exams together to ensure each student received a random version. Summary statistics for how students performed on these two exams are shown in Table 4.20. Anticipating complaints from students who took Version B, she would like to evaluate whether the difference observed in the groups is so large that it provides convincing evidence that Version B was more difficult (on average) than Version A.

Guided Practice 4.27 Construct a hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance.[19]

[17] Absolutely not. It is possible that there is some difference but we did not detect it. If there is a difference, we made a Type 2 Error. Notice: we also don't have enough information to, if there is an actual difference difference, confidently say which direction that difference would be in.

[18] We could have collected more data. If the sample sizes are larger, we tend to have a better shot at finding a difference if one exists.

[19] 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. $\mu_A - \mu_B \neq 0$.

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

Table 4.20: Summary statistics of scores for each exam version.

⊙ **Guided Practice 4.28** To evaluate the hypotheses in Guided Practice 4.27 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 / skew condition for observations in each group? (c) Do you think scores from the two groups would be independent of each other, i.e. the two samples are independent?[20]

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 as

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

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 $n_2 - 1$: $df = 26$.

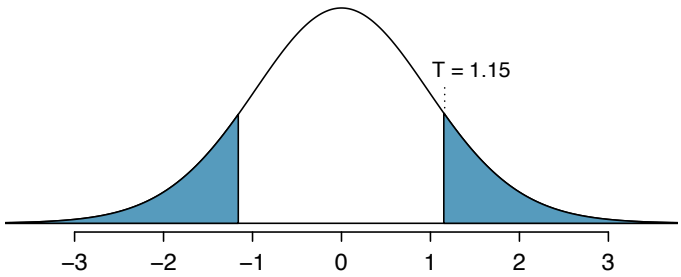


Figure 4.21: The t distribution with 26 degrees of freedom. The shaded right tail represents values with $T \geq 1.15$. Because it is a two-sided test, we also shade the corresponding lower tail.

[20] (a) It is probably reasonable to conclude the scores are independent, provided there was no cheating. (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.

Example 4.29 Identify the p-value using $df = 26$ and provide a conclusion in the context of the case study.

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.

4.3.4 Summary for inference using the t distribution

Hypothesis tests. When applying the t distribution for a hypothesis test, we proceed as follows:

- Write appropriate hypotheses.
- Verify conditions for using the t distribution.
  - One-sample or differences from paired data: the observations (or differences) must be independent and nearly normal. For larger sample sizes, we can relax the nearly normal requirement, e.g. slight skew is okay or sample sizes of 15, moderate skew for sample sizes of 30, and strong skew for sample sizes of 60.
  - For a difference of means when the data are not paired: each sample mean must separately satisfy the one-sample conditions for the t distribution, and the data in the groups must also be independent.
- Compute the point estimate of interest, the standard error, and the degrees of freedom. For df, use $n - 1$ for one sample, and for two samples use either statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.
- Compute the T score and p-value.
- Make a conclusion based on the p-value, and write a conclusion in context and in plain language so anyone can understand the result.

Confidence intervals. Similarly, the following is how we generally computed a confidence interval using a t distribution:

- Verify conditions for using the t distribution. (See above.)
- Compute the point estimate of interest, the standard error, the degrees of freedom, and $t^\star_d f$.
- Calculate the confidence interval using the general formula, point estimate $\pm t^\star_{df} SE$.
- Put the conclusions in context and in plain language so even non-statisticians can understand the results.

4.3.5 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 the 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:

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

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s^2_{pooled} 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.

4.4 Comparing many means with ANOVA (special topic)

Sometimes we want to compare means across many groups. We might initially think to do pairwise comparisons; for example, if there were three groups, we might be tempted to compare the first mean with the second, then with the third, and then finally compare the second and third means for a total of three comparisons. However, this strategy can be treacherous. If we have many groups and do many comparisons, it is likely that we will eventually find a difference just by chance, even if there is no difference in the populations.

In this section, we will learn a new method called **analysis of variance (ANOVA)** and a new test statistic called F. ANOVA uses a single hypothesis test to check whether the means across many groups are equal:

H_0: The mean outcome is the same across all groups. In statistical notation, $\mu_1 = \mu_2 = \cdots = \mu_k$ where μ_i represents the mean of the outcome for observations in category i.

H_A: At least one mean is different.

Generally we must check three conditions on the data before performing ANOVA:

- the observations are independent within and across groups,
- the data within each group are nearly normal, and
- the variability across the groups is about equal.