## Section 7.3: Difference of Two Means

Diez, D. M., Çetinkaya-Rundel, M., Barr, C. D. (2019). OpenIntro Statistics (4th ed.). OpenIntro.
https://www.openintro.org/book/os/ CC BY-SA 3.0

STAT 1201
Introduction to Probability and Statistics

## 7.3 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, $\bar{x}_1 - \bar{x}_2$, and a new standard error formula. Other than these two differences, the details are almost identical to the one-mean procedures.

We apply these methods in three contexts: determining whether stem cells can improve heart function, exploring the relationship between pregnant womens' smoking habits and birth weights of newborns, and exploring whether there is statistically significant evidence that one variation 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?"

### 7.3.1 Confidence interval for a difference of means

Does treatment using embryonic stem cells (ESCs) help improve heart function following a heart attack? Figure 7.11 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. Figure 7.12 provides histograms of the two data sets. 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.

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

Figure 7.11: Summary statistics of the embryonic stem cell study.

The point estimate of the difference in the heart pumping variable is straightforward to find: it is the difference in the sample means.

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

For the question of whether we can model this difference using a $t$-distribution, we'll need to check new conditions. Like the 2-proportion cases, we will require a more robust version of independence so we are confident the two groups are also independent. Secondly, we also check for normality in each group separately, which in practice is a check for outliers.

#### 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

- *Independence, extended.* The data are independent within and between the two groups, e.g. the data come from independent random samples or from a randomized experiment.
- *Normality.* We check the outliers rules of thumb for each group separately.

The standard error may be computed as

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The official formula for the degrees of freedom is quite complex and is generally computed using software, so instead you may use the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom if software isn't readily available.

**EXAMPLE 7.21**

Can the $t$-distribution be used to make inference using the point estimate, $\bar{x}_{esc} - \bar{x}_{control} = 7.83$?

First, we check for independence. Because the sheep were randomized into the groups, independence within and between groups is satisfied.

Figure 7.12 does not reveal any clear outliers in either group. (The ESC group does look a bit more variability, but this is not the same as having clear outliers.)

With both conditions met, we can use the $t$-distribution to model the difference of sample means.

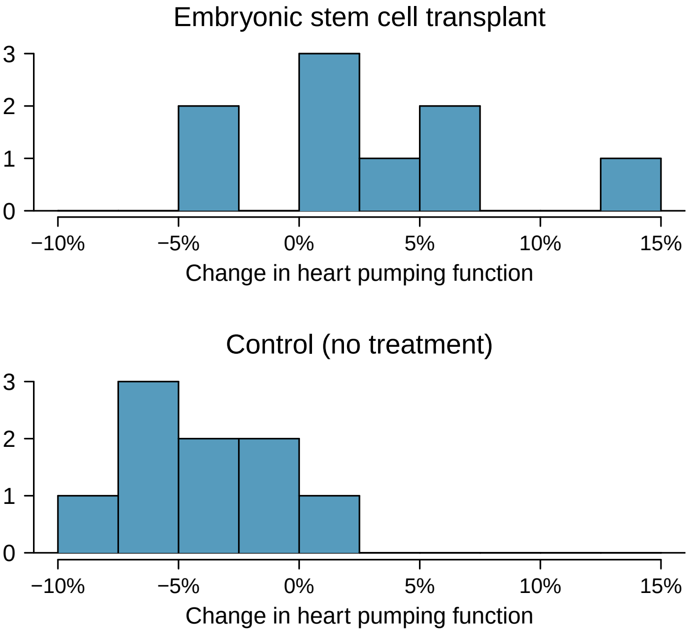


Figure 7.12: Histograms for both the embryonic stem cell and control group.

As with the one-sample case, we always compute the standard error using sample standard deviations rather than population standard deviations:

$$SE = \sqrt{\frac{s_{esc}^2}{n_{esc}} + \frac{s_{control}^2}{n_{control}}} = \sqrt{\frac{5.17^2}{9} + \frac{2.76^2}{9}} = 1.95$$

Generally, we use statistical software to find the appropriate degrees of freedom, or if software isn't available, we can use the smaller of $n_1 - 1$ and $n_2 - 1$ for the degrees of freedom, e.g. if using a $t$-table to find tail areas. For transparency in the Examples and Guided Practice, we'll use the latter approach for finding $df$; in the case of the ESC example, this means we'll use $df = 8$.

**EXAMPLE 7.22**

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 that we computed earlier calculations:

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

Using $df = 8$, we can identify the critical value of $t^\star_8 = 2.31$ for a 95% confidence interval. Finally, we can enter the values into the confidence interval formula:

$$\text{point estimate} \ \pm \ t^\star \times SE \quad \rightarrow \quad 7.83 \ \pm \ 2.31 \times 1.95 \quad \rightarrow \quad (3.32, 12.34)$$

We are 95% confident that embryonic stem cells improve the heart's pumping function in sheep that have suffered a heart attack by 3.32% to 12.34%.

As with past statistical inference applications, there is a well-trodden procedure.

**Prepare.** Retrieve critical contextual information, and if appropriate, set up hypotheses.

**Check.** Ensure the required conditions are reasonably satisfied.

**Calculate.** Find the standard error, and then construct a confidence interval, or if conducting a hypothesis test, find a test statistic and p-value.

**Conclude.** Interpret the results in the context of the application.

The details change a little from one setting to the next, but this general approach remain the same.

## 7.3.2 Hypothesis tests for the difference of two means

A data set called `ncbirths` 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 Figure 7.13. 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.

| | fage | mage | weeks | weight | sex | 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 |

Figure 7.13: Four cases from the `ncbirths` data set. The value "NA", shown for the first two entries of the first variable, indicates that piece of data is missing.

E

### EXAMPLE 7.23

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 $\mu_n$ represents non-smoking mothers and $\mu_s$ 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 model the difference in sample means using the $t$-distribution.

- Because the data come from a simple random sample, the observations are independent, both within and between samples.
- With both data sets over 30 observations, we inspect the data in Figure 7.14 for any particularly extreme outliers and find none.

Since both conditions are satisfied, the difference in sample means may be modeled using a $t$-distribution.

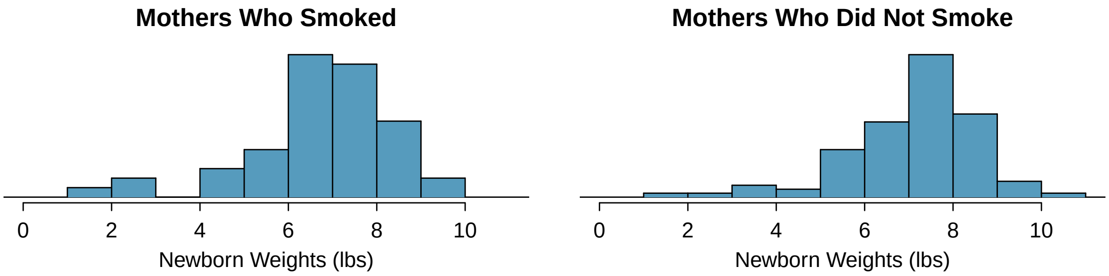


Figure 7.14: The left panel represents birth weights for infants whose mothers smoked. The right panel represents the birth weights for infants whose mothers who did not smoke.

G

### GUIDED PRACTICE 7.24

The summary statistics in Figure 7.15 may be useful for this Guided Practice.[13]

(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).

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

Figure 7.15: Summary statistics for the ncbirths data set.

[13] (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 calculated using the standard error formula:

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

E

### EXAMPLE 7.25

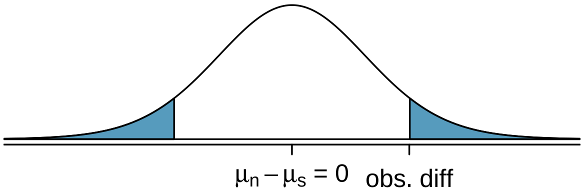
Complete the hypothesis test started in Example 7.23 and Guided Practice 7.24. Use a significance level of $\alpha = 0.05$. For reference, $\bar{x}_n - \bar{x}_s = 0.40$, $SE = 0.26$, and the sample sizes were $n_n = 100$ and $n_s = 50$.

---

We can find the test statistic for this test using the values from Guided Practice 7.24:

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

The p-value is represented by the two shaded tails in the following plot:



We find the single tail area using software (or the $t$-table in Appendix C.2). We'll use the smaller of $n_n - 1 = 99$ and $n_s - 1 = 49$ as the degrees of freedom: $df = 49$. The one tail area is 0.065; doubling this value gives the two-tail area and p-value, 0.135.

The p-value is larger than the significance value, 0.05, so we do not 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.

G

### GUIDED PRACTICE 7.26

We've seen much research suggesting smoking is harmful during pregnancy, so how could we fail to reject the null hypothesis in Example 7.25? [14]

G

### GUIDED PRACTICE 7.27

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 the difference?[15]

Public service announcement: while we have used this relatively small data set as an example, larger data sets show that women who smoke tend to have smaller newborns. In fact, some in the tobacco industry actually had the audacity to tout that as a *benefit* of smoking:

> *It's true. The babies born from women who smoke are smaller, but they're just as healthy as the babies born from women who do not smoke. And some women would prefer having smaller babies.*
>
> - Joseph Cullman, Philip Morris' Chairman of the Board
> on CBS' *Face the Nation*, Jan 3, 1971

Fact check: the babies from women who smoke are not actually as healthy as the babies from women who do not smoke.[16]

[14] It is possible that there is a difference but we did not detect it. If there is a difference, we made a Type 2 Error.

[15] 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. In fact, this is exactly what we would find if we examined a larger data set!

[16] You can watch an episode of John Oliver on *Last Week Tonight* to explore the present day offenses of the tobacco industry. Please be aware that there is some adult language: youtu.be/6UsHHOCH4q8.

### 7.3.3 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 Figure 7.16. 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.

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

Figure 7.16: Summary statistics of scores for each exam version.

**GUIDED PRACTICE 7.28**

Construct hypotheses to evaluate whether the observed difference in sample means, $\bar{x}_A - \bar{x}_B = 5.3$, is due to chance. We will later evaluate these hypotheses using $\alpha = 0.01$.[17]

**GUIDED PRACTICE 7.29**

To evaluate the hypotheses in Guided Practice 7.28 using the $t$-distribution, we must first verify conditions.[18]

(a) Does it seem reasonable that the scores are independent?

(b) Any concerns about outliers?

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

[17] $H_0$: the exams are equally difficult, on average. $\mu_A - \mu_B = 0$. $H_A$: one exam was more difficult than the other, on average. $\mu_A - \mu_B \neq 0$.

[18] (a) Since the exams were shuffled, the "treatment" in this case was randomly assigned, so independence within and between groups is satisfied. (b) The summary statistics suggest the data are roughly symmetric about the mean, and the min/max values don't suggest any outliers of concern.

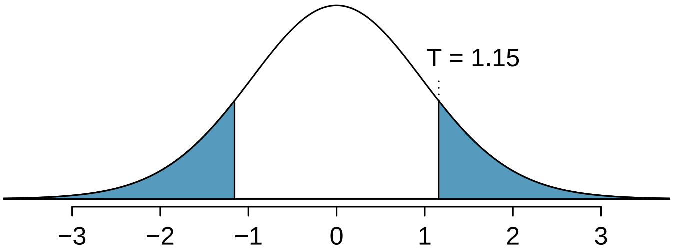


Figure 7.17: The $t$-distribution with 26 degrees of freedom and the p-value from exam example represented as the shaded areas.

E

**EXAMPLE 7.30**

Identify the p-value depicted in Figure 7.17 using $df = 26$, and provide a conclusion in the context of the case study.

---

Using software, we can find the one-tail area (0.13) and then double this value to get the two-tail area, which is the p-value: 0.26. (Alternatively, we could use the $t$-table in Appendix C.2.)

In Guided Practice 7.28, we specified that we would use $\alpha = 0.01$. Since the p-value is larger than $\alpha$, 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.

## 7.3.4 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 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 very 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 $\bar{x}_1 - \bar{x}_2$, if the standard deviations of the two groups are indeed equal.

**POOL STANDARD DEVIATIONS ONLY AFTER CAREFUL CONSIDERATION**

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.