Inference for Paired Data

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of OpenIntro and John McGreedy of Johns Hopkins University

Plan for Today

- Review small one sample means
- Inference for paired data

Warm Up

7.1 Identify the critical t. An independent random sample is selected from an approximately normal population with unknown standard deviation. Find the degrees of freedom and the critical t-value ($t^\star$) for the given sample size and confidence level.

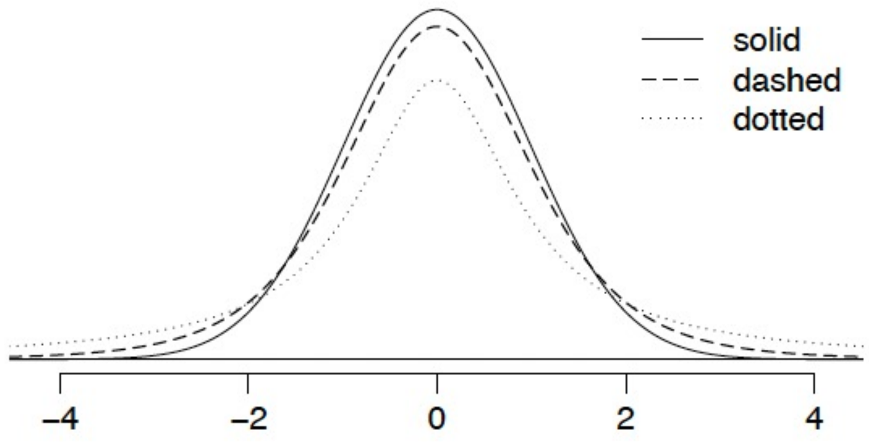
(a) $n = 6$, CL $= 90\%$

(b) $n = 21$, CL $= 98\%$

(c) $n = 29$, CL $= 95\%$

(d) $n = 12$, CL $= 99\%$

7.2 t-distribution. The figure on the right shows three unimodal and symmetric curves: the standard normal (z) distribution, the t-distribution with 5 degrees of freedom, and the t-distribution with 1 degree of freedom. Determine which is which, and explain your reasoning.



Warm Up

7.1

(a) $n = 6$, CL $= 90\%$, $df = 6 - 1 = 5$, $t_5^\star = 2.02$

(b) $n = 21$, CL $= 98\%$, $df = 21 - 1 = 20$, $t_{20}^\star = 2.53$

(c) $n = 29$, CL $= 95\%$, $df = 29 - 1 = 28$, $t_{28}^\star = 2.05$

(d) $n = 12$, CL $= 99\%$, $df = 12 - 1 = 11$, $t_{11}^\star = 3.11$

7.2 The dotted line is the t-distribution with 1 degree of freedom, the dashed line is the t-distribution with 5 degrees of freedom, and the solid line is the standard normal distribution. As the degrees of freedom increases the t-distribution approaches the normal distribution. Another valid justification is that lower the degrees of freedom, thicker the tails.

Warm Up

7.9 Find the mean. You are given the following hypotheses:

$$H_0 : \mu = 60$$
$$H_A : \mu \neq 60$$

We know that the sample standard deviation is 8 and the sample size is 20. For what sample mean would the p-value be equal to 0.05? Assume that all conditions necessary for inference are satisfied.

Warm Up

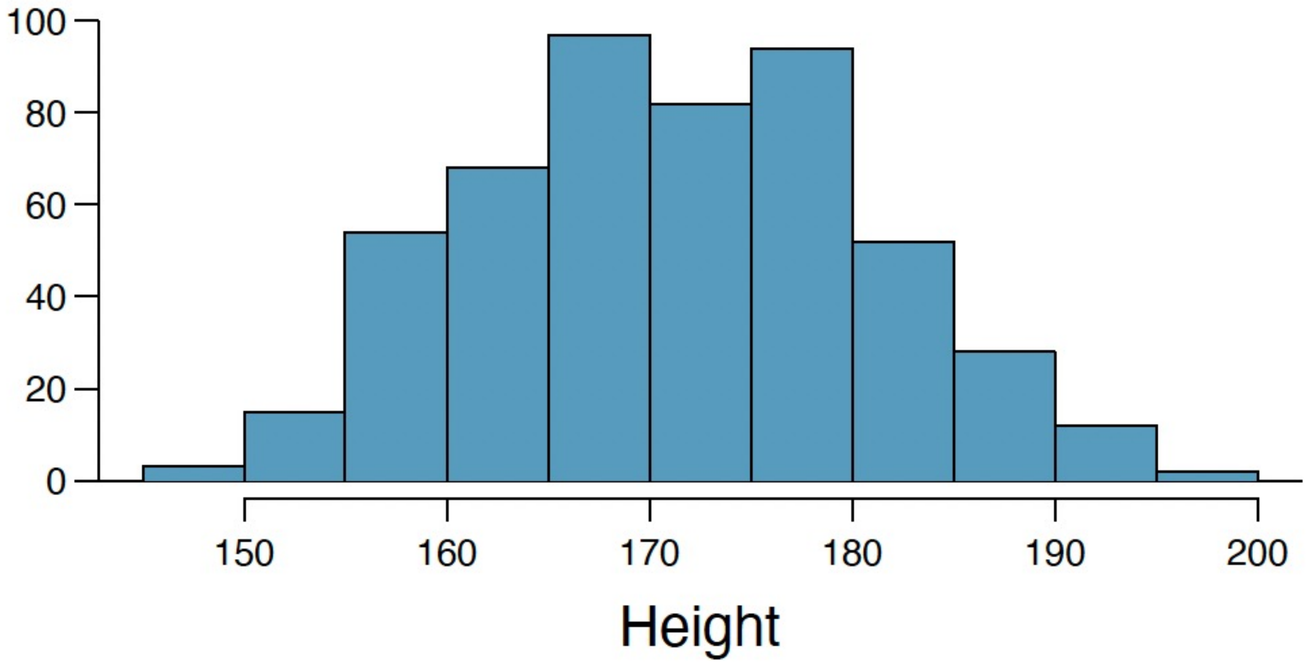
7.9 For the single tails to each be 0.025 at $n - 1 = 20 - 1 = 19$ degrees of freedom, T score must equal to be either -2.09 or +2.09. Then, either:

$$-2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 56.26$$

$$2.09 = \frac{\bar{x} - 60}{\frac{8}{\sqrt{20}}} \rightarrow \bar{x} = 63.74$$

Warm Up

7.8 Heights of adults. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.[8]



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

(a) What is the point estimate for the average height of active individuals? What about the median?

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

Warm Up

7.8

(a) Use the sample mean to estimate the population mean: 171.1. Likewise, use the sample median to estimate the population median: 170.3.

(b) Use the sample standard deviation (9.4) and sample IQR ($177.8 - 163.8 = 14$).

(c) In order to determine if 180 cm or 155 cm are considered unusual observations we need to calculate how many standard deviations away from the mean this observation is, i.e. calculate the Z-score.

$$Z = \frac{180 - 171.1}{9.4} = 0.95 \qquad Z = \frac{155 - 171.1}{9.4} = -1.71$$

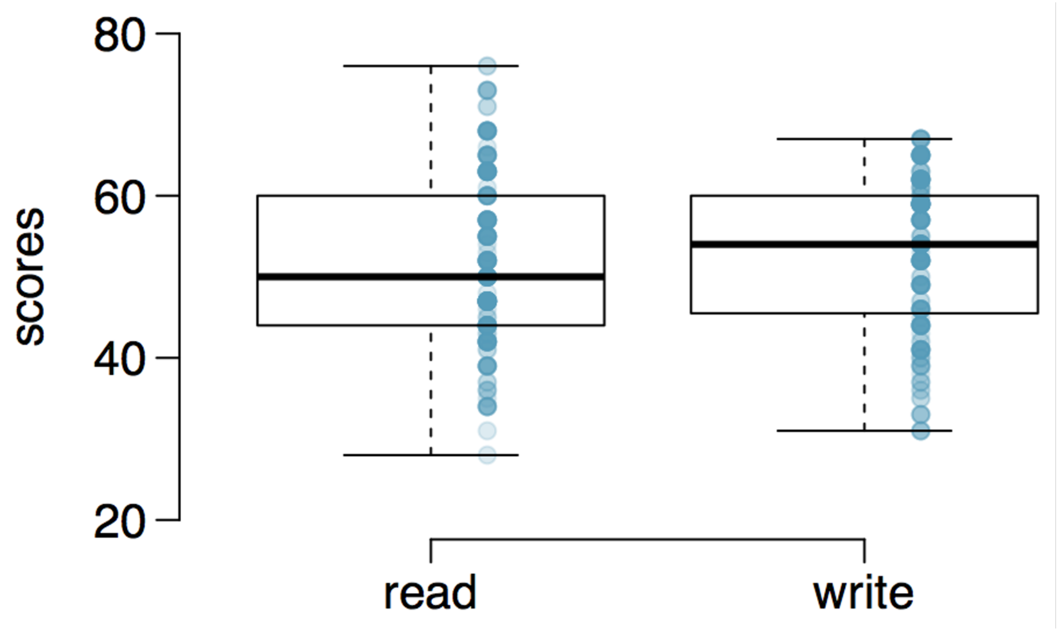
Neither of these observations is more than two standard deviations away from the mean, so neither would be considered unusual.

(d) No, sample point estimates only estimate the population parameter, and they vary from one sample to another. Therefore we cannot expect to get the same mean and standard deviation with each random sample.

(e) We use the standard error of the mean to measure the variability in means of random samples of same size taken from a population. The variability in the means of random samples is quantified by the standard error. Based on this sample, $SE_{\bar{x}} = \frac{9.4}{\sqrt{507}} = 0.417$.

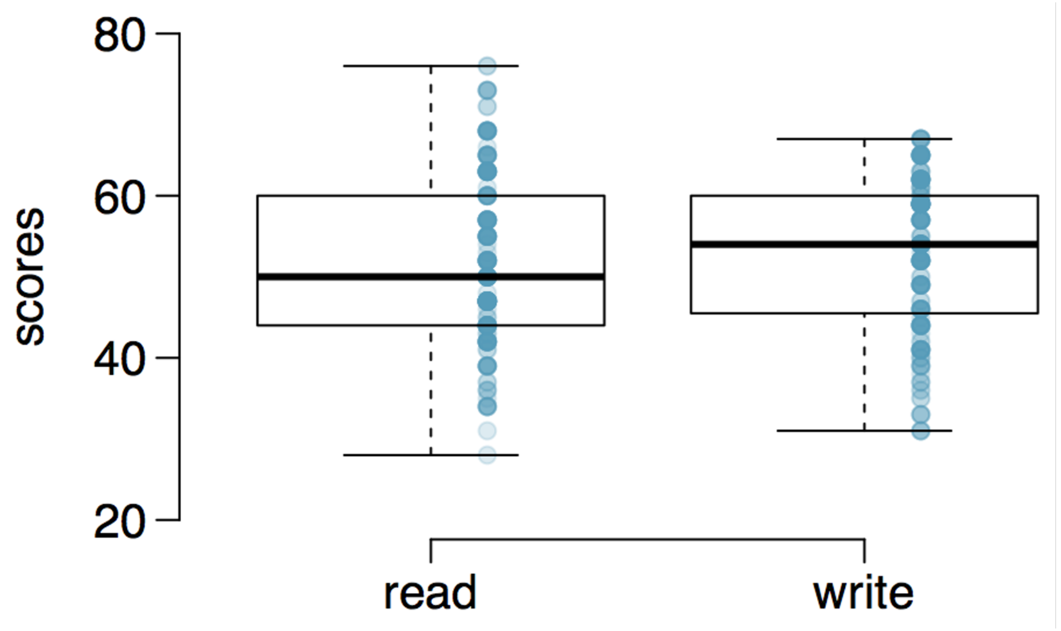
Paired observations

200 observations were randomly sampled from the High School and Beyond survey. The same students took a reading and writing test and their scores are shown below. At a first glance, does there appear to be a difference between the average reading and writing test score?



Paired observations

200 observations were randomly sampled from the High School and Beyond survey. The same students took a reading and writing test and their scores are shown below. At a first glance, does there appear to be a difference between the average reading and writing test score?



No, most of the scores are concentrated in the same range.

Paired observations

Are the reading and writing scores of each student independent of each other?

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

(a) Yes

(b) No

Paired observations

Are the reading and writing scores of each student independent of each other?

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

(a) Yes

(b) No

Analyzing paired data

- When two sets of observations have this special correspondence (**not independent**), they are said to be *paired*

Analyzing paired data

- When two sets of observations have this special correspondence (**not independent**), they are said to be *paired*
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations

  - For example, in the reading and writing test case:

    diff = read - write

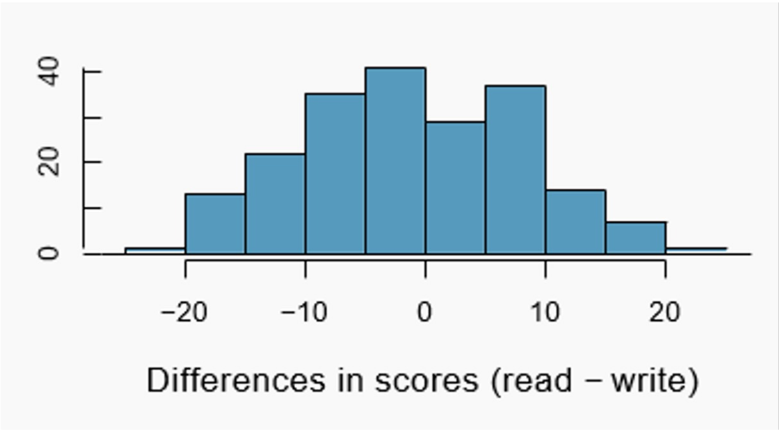
Analyzing paired data

- When two sets of observations have this special correspondence (not independent), they are said to be *paired*
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations

diff = read - write

- It is important that we always subtract using a consistent order

	id	read	write	diff
1	70	57	52	5
2	86	44	33	11
3	141	63	44	19
4	172	47	52	-5
:	:	:	:	:
200	137	63	65	-2



Parameter and point estimate

- *Parameter of interest*: Average difference between the reading and writing scores of all high school students

$$\mu_{diff}$$

Parameter and point estimate

- *Parameter of interest*: Average difference between the reading and writing scores of all high school students

$$\mu_{diff}$$

- *Point estimate*: Average difference between the reading and writing scores of sampled high school students

$$\bar{x}_{diff}$$

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

0

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

0

What are the hypotheses for testing if there is a difference between the average reading and writing scores?

Setting the hypotheses

If in fact there was no difference between the scores on the reading and writing exams, what would you expect the average difference to be?

0

What are the hypotheses for testing if there is a difference between the average reading and writing scores?

H_0 : Average reading and writing exam scores are not different.

$$\mu_{diff} = 0$$

H_A : Average reading and writing exam scores are different.

$$\mu_{diff} \neq 0$$

Nothing new here

- The analysis is no different than what we have done before
- We have data from one sample: differences.
- We are testing to see if the average difference is different than 0.

Checking assumptions & conditions

Which of the following is true?

A. Since students are sampled randomly and are less than 10% of all high school students, we can assume that the difference between the reading and writing scores of one student in the sample is independent of another
B. The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test
C. In order for differences to be random we should have sampled with replacement
D. Since students are sampled randomly and are less than 10% all students, we can assume that the sampling distribution of the average difference will be nearly normal

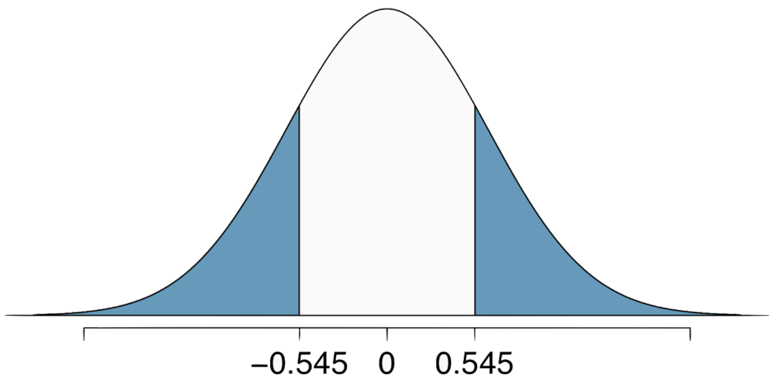
Checking assumptions & conditions

Which of the following is true?

A. *Since students are sampled randomly and are less than 10% of all high school students, we can assume that the difference between the reading and writing scores of one student in the sample is independent of another*
B. The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test
C. In order for differences to be random we should have sampled with replacement
D. Since students are sampled randomly and are less than 10% all students, we can assume that the sampling distribution of the average difference will be nearly normal

Calculating the test-statistics and the p-value

The observed average difference between the two scores for the 200 observations is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use α = 0.05



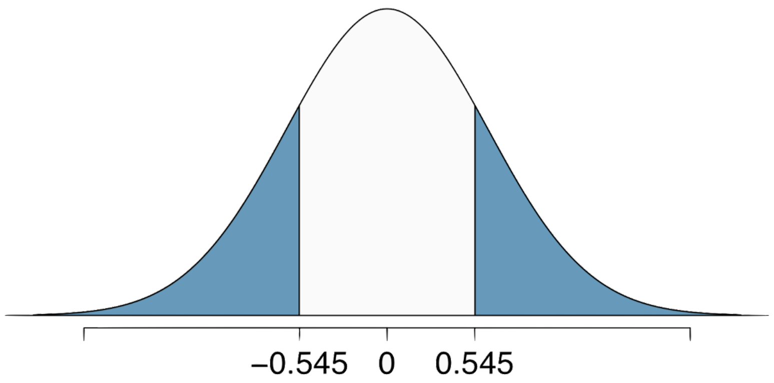
$$T = \frac{obs - \mu}{SE}$$

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$df = n - 1$$

Calculating the test-statistics and the p-value

The observed average difference between the two scores for the 200 observations is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use α = 0.05



$$T = \frac{obs - \mu}{SE}$$

$$SE = \frac{\sigma}{\sqrt{n}}$$

$$df = n - 1$$

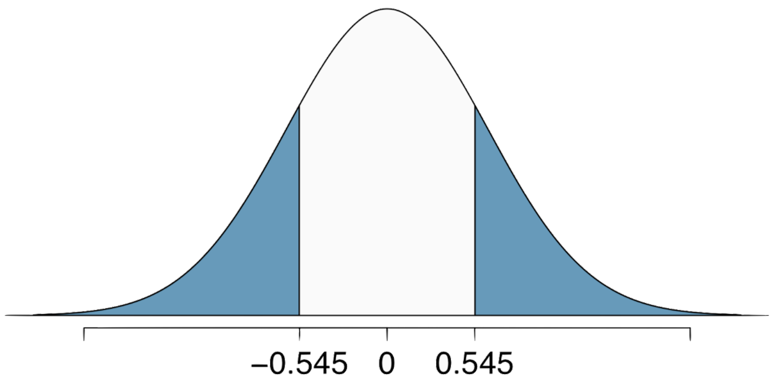
$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}}$$

$$T = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

Calculating the test-statistics and the p-value

The observed average difference between the two scores for the 200 observations is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use α = 0.05



$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}}$$

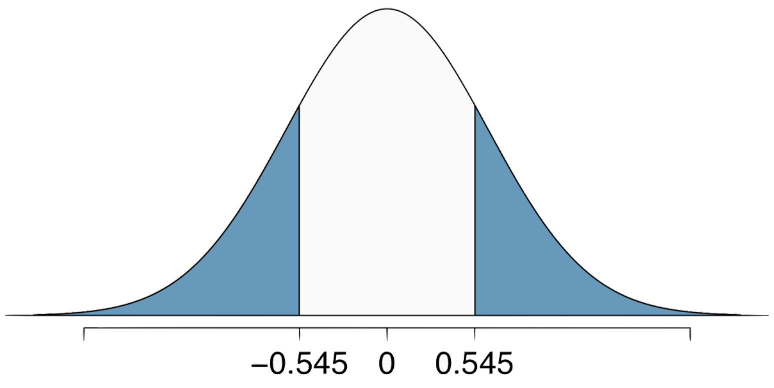
$$T = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

$$p - value = 0.1927 \times 2 = 0.3854$$

Calculating the test-statistics and the p-value

The observed average difference between the two scores for the 200 observations is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use α = 0.05



$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}}$$

$$T = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

$$p - value = 0.1927 \ \times 2 = 0.3854$$

Since p-value > 0.05, fail to reject, the data do not provide convincing evidence of a difference between the average reading and writing scores

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

A. Probability that the average scores on the reading and writing exams are equal
B. Probability that the average scores on the reading and writing exams are different
C. Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0
D. Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true

Interpretation of p-value

Which of the following is the correct interpretation of the p-value?

A. Probability that the average scores on the reading and writing exams are equal
B. Probability that the average scores on the reading and writing exams are different
C. *Probability of obtaining a random sample of 200 students where the average difference between the reading and writing scores is at least 0.545 (in either direction), if in fact the true average difference between the scores is 0*
D. Probability of incorrectly rejecting the null hypothesis if in fact the null hypothesis is true

HT $\leftrightarrow$ CI

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

A. yes
B. no
C. cannot tell from the information given

HT $\leftrightarrow$ CI

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

A. *yes*
B. no
C. cannot tell from the information given

HT $\leftrightarrow$ CI

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

A. *yes*
B. no
C. cannot tell from the information given

Calculate the CI

$$CI = \bar{x} \pm t^{*} * SE$$

HT $\leftrightarrow$ CI

Suppose we were to construct a 95% confidence interval for the average difference between the reading and writing scores. Would you expect this interval to include 0?

A. *yes*
B. no
C. cannot tell from the information given

Calculate the CI

$$CI = \bar{x} \pm t^* * SE$$

$$-0.545 \pm 1.97 \frac{8.887}{\sqrt{200}} = -0.545 \pm 1.87 \times 0.628$$

$$= -0.545 \pm 1.24$$

$$= (-1.785, 0.695)$$

More Practice

We sampled 201 UCLA courses. Of those, 68 required books could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.8, where prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	M10	35.96	32.40	3.56

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

What are the null and alternative hypotheses for the research question: Are books on Amazon less expensive than books at the UCLA bookstore?

More Practice

We sampled 201 UCLA courses. Of those, 68 required books could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.8, where prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	M10	35.96	32.40	3.56

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

What are the null and alternative hypotheses for the research question: Are books on Amazon less expensive than books at the UCLA bookstore?

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.

H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

More Practice

We sampled 201 UCLA courses. Of those, 68 required books could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.8, where prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	M10	35.96	32.40	3.56

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.

H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Complete the hypothesis test

More Practice

We sampled 201 UCLA courses. Of those, 68 required books could be found on Amazon. A portion of the data set from these courses is shown in Figure 7.8, where prices are in US dollars.

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	M10	35.96	32.40	3.56

n_{diff}	$\bar{x}_{diff}$	s_{diff}
68	3.58	13.42

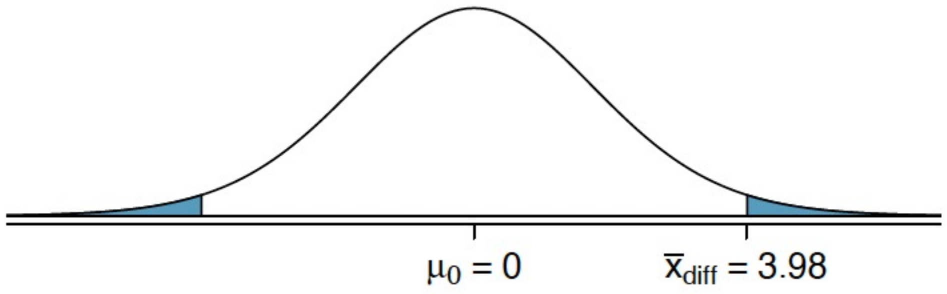
H_0: $\mu_{diff} = 0$. There is no difference in the average textbook price.

H_A: $\mu_{diff} \neq 0$. There is a difference in average prices.

Complete the hypothesis test

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63 \quad ; \; df = 68 - 1 = 67$$

$$T = \frac{\bar{x}_{diff} - 0}{SE_{\bar{x}_{diff}}} = \frac{3.58 - 0}{1.63} = 2.20$$



$p = 0.0156, for\ two\ tails\ p = 0.0156 * 2 = 0.0312$

Reject the null hypothesis.