Hypothesis Tests with the *t*-distribution

Slides developed by Mine Çetinkaya-Rundel of OpenIntro
Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.

Recap: Hypothesis Tests

- Identify the research question, formalize in terms of parameters.
- We want to make a decision on whether we think H_0 or H_1 is correct.
- Find a statistics, compute the distribution of the statistic under the null.
- p-value: The probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true.
- If the p-value is low (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence reject H_0.
- If the p-value is high (higher than the significance level, α) then it is pretty likely to observe the data even if the null hypothesis were true, so we do not reject H_0.
- We never accept H_0 since we're not in the business of trying to prove it!

Recap: Hypothesis Tests

- GOF tests:

  $H_0: \mathrm{p}_1 = \mathrm{p}_1^0, \mathrm{p}_2 = \mathrm{p}_2^0, \dots, \mathrm{p}_\mathrm{k} = \mathrm{p}_\mathrm{k}^0$

  $$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{df}$$

  df: Degrees of freedom, number of parameters allowed to vary freely

  For k categories, k-1, for independence tests, $(|X| - 1)\times(|Y| - 1)$

- For comparing proportions, we can use the CLT approximation and use

  $Z = \frac{\hat{p} - p_o}{SE} \sim N(0, 1)$

- This works for any other mean

Example: Student Grades

- Last year, the average grade of students the first midterm of applied statistics was 68.7.
  This year, the first 50 students that were graded have an average grade of 63.7 with a standard deviation of 12.
  Is student performance different this year?

- H_o: Average student performance is the same as last year
- H_1: Average student performance is different than last year

Example

- Last year, the average grade of students the first midterm of applied statistics was 68.7.
  This year, the first 50 students that were graded have an average grade of 63.7 with a standard deviation of 12.
  Is student performance different this year?

- $H_o: \mu = 68.7$

- $H_o: \mu \neq 68.7$

Conditions

- *Independence*: We are told to assume that cases (rows) are independent.
- *Normality* : The distribution of the sample mean is nearly normal (CLT)
  - Sample size / skew: Sample size is not very small, distribution of grades does not appear extremely skewed.

Z-test

- $Z = \frac{\overline{X_n} - \mu}{SE}$
- $SE = s/\sqrt{n}$

- Under the null, $Z \sim N(0, 1)$

- p-value: $P(|Z| > z||H_o) =$

Small sample size

- Assume that I only graded 10 midterm exams with an average of 59.2 and sample standard deviation 15

- *Independence*: We are told to assume that cases (rows) are independent.
- *Normality* : The distribution of the sample mean is nearly normal (CLT)
    - Sample size / skew: Sample size is very small, distribution of grades is hard to assess with such a small sample size.

So what do we do when the sample size is small?

Review: why do we need a large sample?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal
- While this is a helpful special case, it’s inherently difficult to verify normality in small data sets

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets

However, if we believe our original disstribution is normal, then

$$Z = \frac{\overline{X_n} - \mu}{SE = \sigma/\sqrt{n}} \sim N(0, 1)$$

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets

However, if we believe our original disstribution is normal, then

$$Z = \frac{\overline{X_n} - \mu}{SE = \sigma/\sqrt{n}} \sim N(0,1)$$

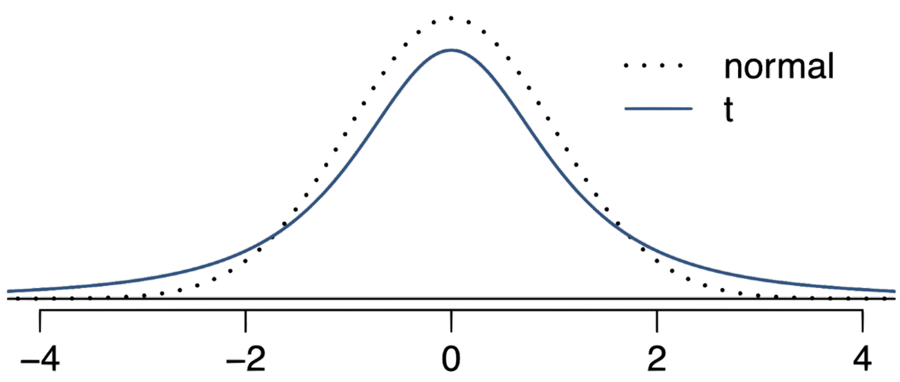
We do not know σ, and n is too small to get a reliable estimate using s.

The *t* distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t* distribution.
- This distribution also has a bell shape, but its tails are thicker than the normal model's (extra uncertainty).
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution

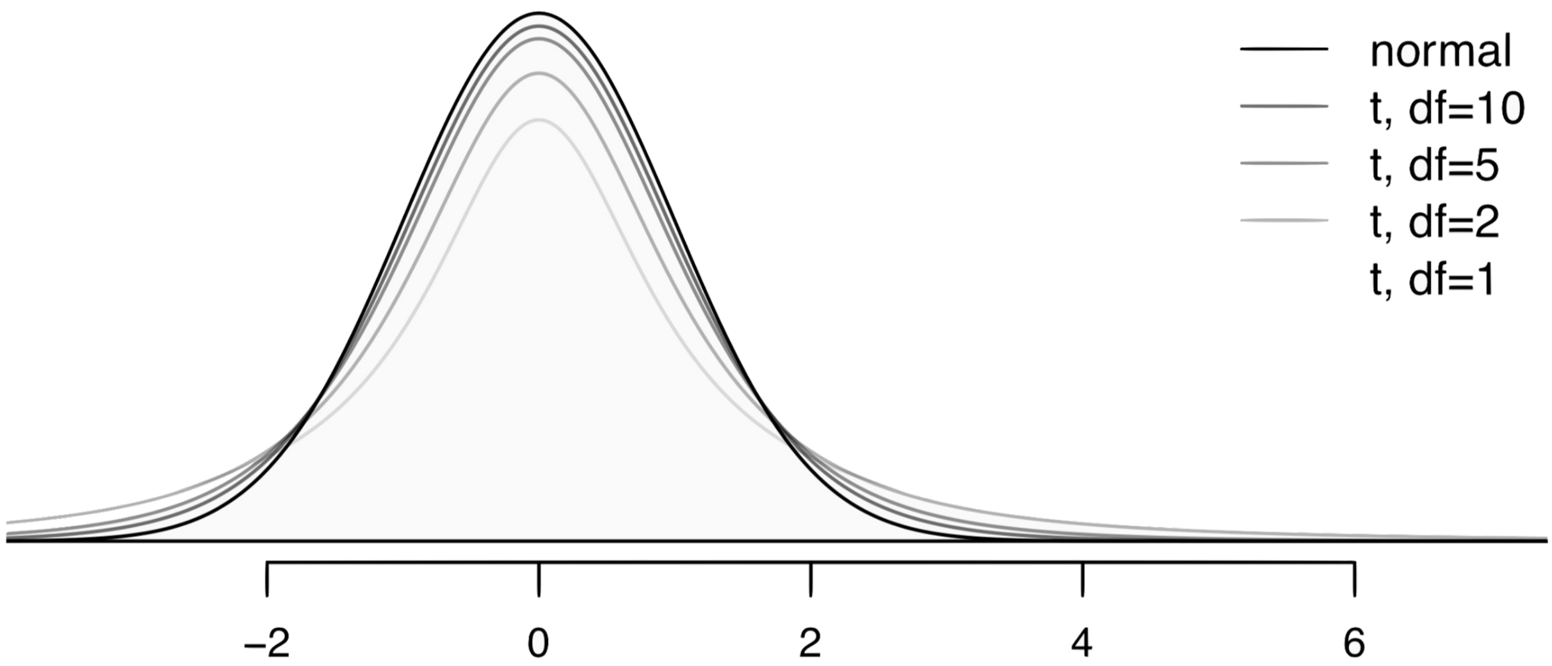
The *t* distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t* distribution.
- This distribution also has a bell shape, but its tails are thicker than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since *n* is small)



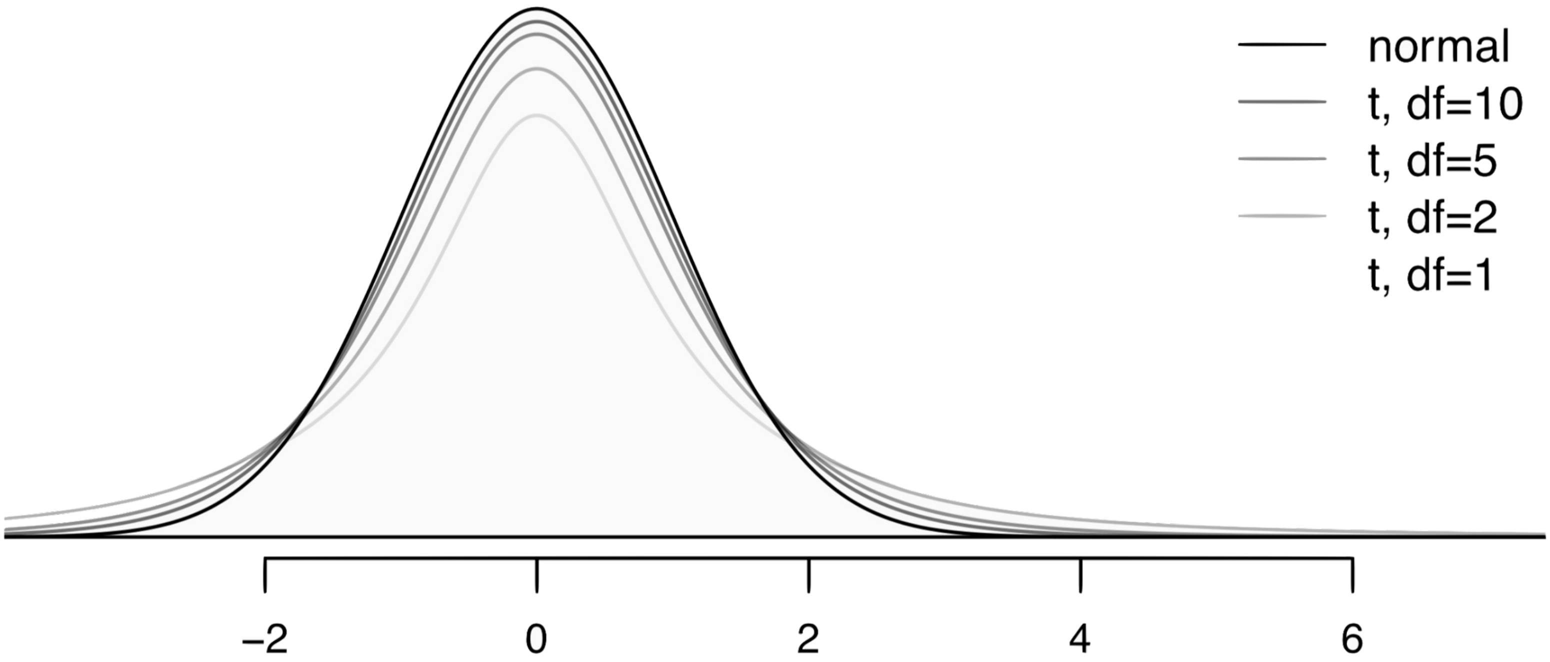
The *t* distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter: *degrees of freedom* (*df*).



The t distribution (cont.)

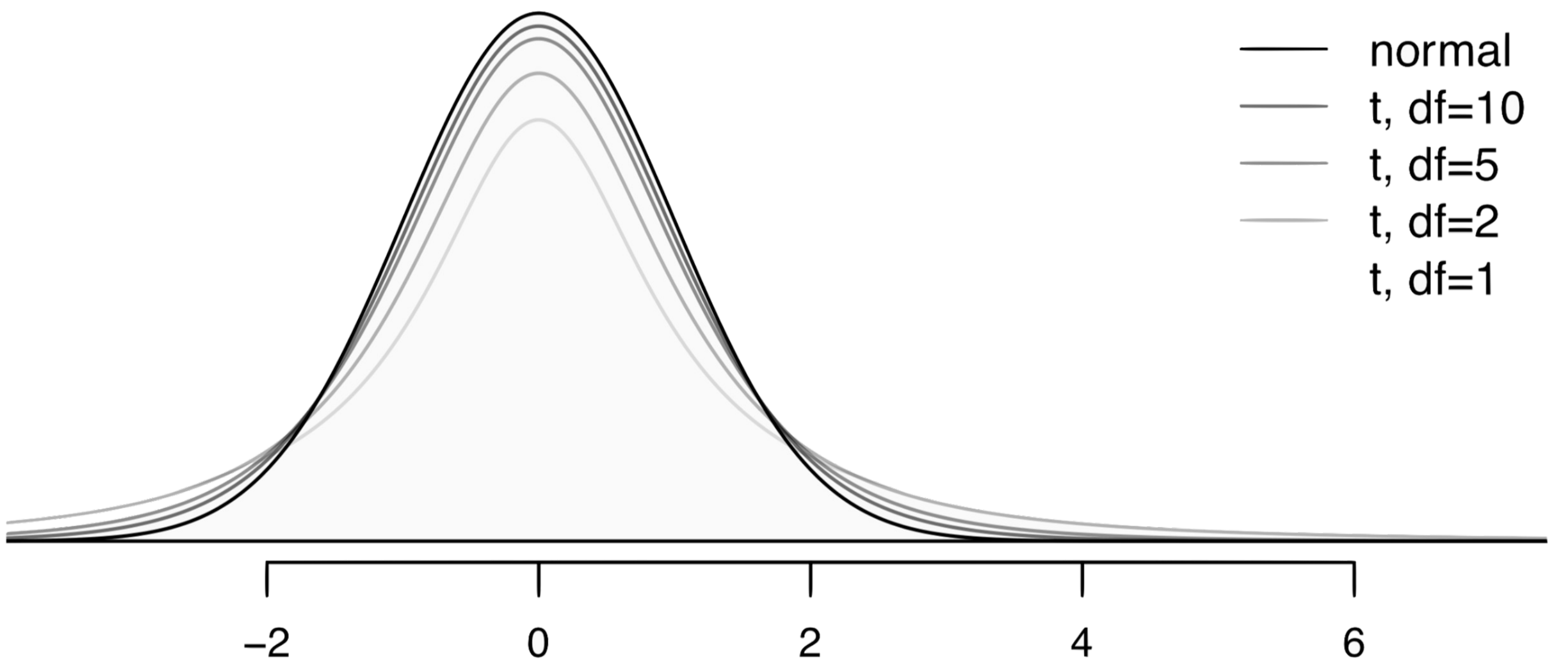
- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter: *degrees of freedom* (*df*).



What happens to the shape of the t distribution as *df* increases?

The t distribution (cont.)

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter: *degrees of freedom* (*df*).



What happens to the shape of the t distribution as *df* increases?

Approaches normal

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample (*n* < 30) mean is the T statistic with *df* = *n* - 1

$$T_{df} = \frac{\overline{X_n} - \mu}{s/\sqrt{n}}$$

Under the null, $T_{df} \sim t_{n-1}$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample (n < 30) mean is the T statistic with $df = n - 1$

$$T_{df} = \frac{\overline{X_n} - \mu}{s/\sqrt{n}}$$

in context...

n = 10

$\overline{X_n}$ = 59.2

$s = 15$

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the t distribution

$$t_{df} = -2.00$$

$$P(\left|T_{df}\right| > |t_{df}|) =$$

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the t distribution
- Using R:

```
> 2 * pt(2.00, df = 9, lower.tail = FALSE)
[1] 0.0008022394
```

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the t distribution
- Old-school

		0.100	0.050	0.025	0.010	0.005
	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	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the t distribution
- Old-school

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	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17

Conclusion of the test

What is the conclusion of this hypothesis test?

Conclusion of the test

What is the conclusion of this hypothesis test?

Since the p-value is not quite low, we cannot reject the null hypothesis that the performance is the same

Confidence interval for a small sample mean

- Confidence intervals are always of the form

$$point\ estimate \pm ME$$

Confidence interval for a small sample mean

- Confidence intervals are always of the form

$$point\ estimate \pm ME$$

- ME is always calculated as the product of a critical value and SE

Confidence interval for a small sample mean

- Confidence intervals are always of the form

$$point\ estimate \pm ME$$

- ME is always calculated as the product of a critical value and SE
- Since small sample means follow a ***t*** distribution (and not a z distribution), the critical value is a ***t****** (as opposed to a z*?).

$$point\ estimate \pm t^*\ x\ SE$$

Finding the critical value ($t^\star$)

Using R:

```
> qt(p = 0.975, df = 9)
[1] 2.262157
```

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval

$\overline{x_n} = 59.2, \quad s = 15, \quad n = 10, \quad SE = 4.74$

A. 59.2 ± 1.96 x 4.74
B. 59.2 ± 2.26 x 4.74
C. 59.2 ± 2.26 x 15

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval?

$\overline{x_n} = 59.2, \quad s = 15, \quad n = 10, \quad SE = 4.74$

A. 59.2 ± 1.96 x 4.74
B. 59.2 ± 2.26 x 4.74
C. 59.2 ± 2.26 x 15

Recap: Inference using the t-distribution

- If σ is unknown, use the t-distribution with $SE = \frac{s}{\sqrt{n}}$

Recap: Inference using the t-distribution

- If σ is unknown, use the t-distribution with $SE = \frac{s}{\sqrt{n}}$
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, n < 10% of population)
  - no extreme skew

Recap: Inference using the t-distribution

- If $\boldsymbol{\sigma}$ is unknown, use the $\boldsymbol{t}$-distribution with $SE = \frac{s}{\sqrt{n}}$
- Conditions:
    - independence of observations (often verified by a random sample, and if sampling without replacement, *n* < 10% of population)
    - no extreme skew
- Hypothesis Testing:

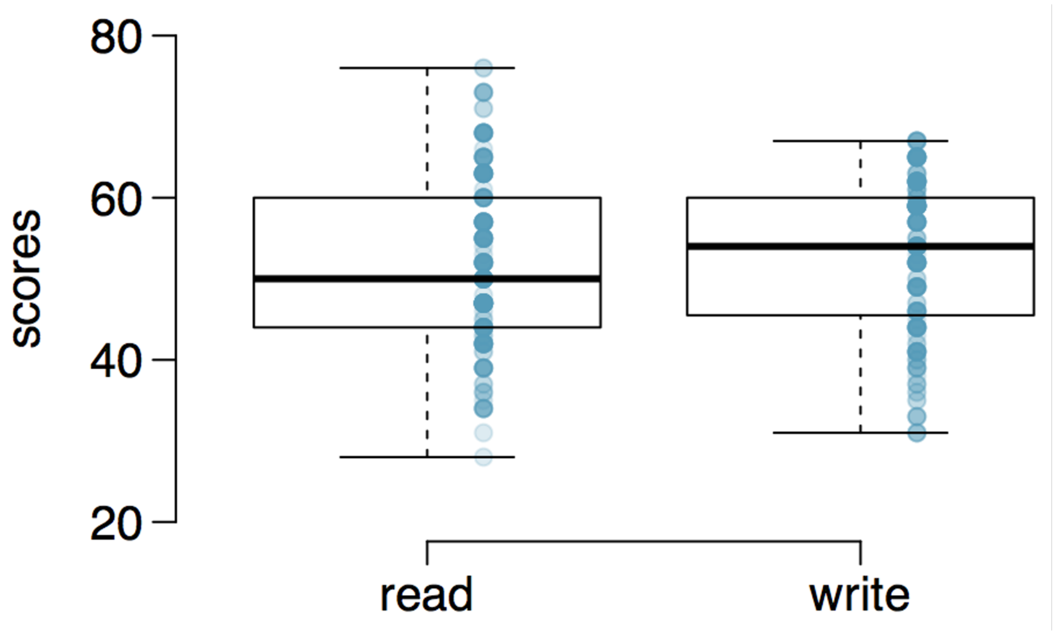
$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = n - 1$$

Paired Data

Slides developed by Mine Çetinkaya-Rundel of OpenIntro
Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.

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

The same students took a reading and writing test and their scores are shown below. 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

The same students took a reading and writing test and their scores are shown below. 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

diff = read - write

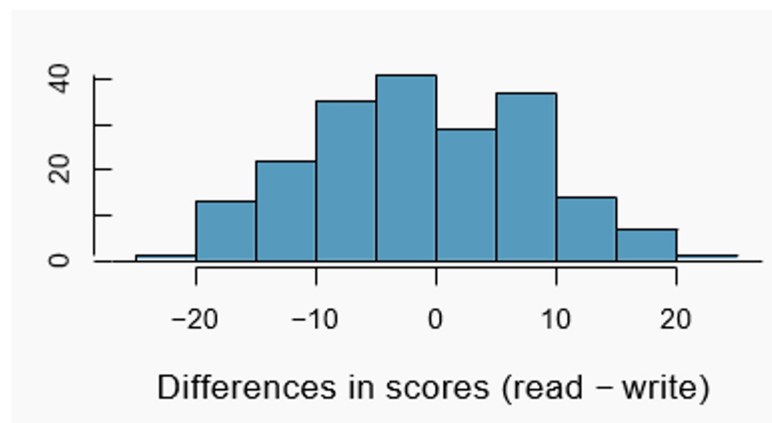
Analyzing paired data

- 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

$$\text{diff} = \text{read} - \text{write}$$

- 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: Reading and writing scores for each student are on average the same.

$$\mu_{diff} = 0$$

H_A : : Reading and writing scores for each student are on average the same.

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

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

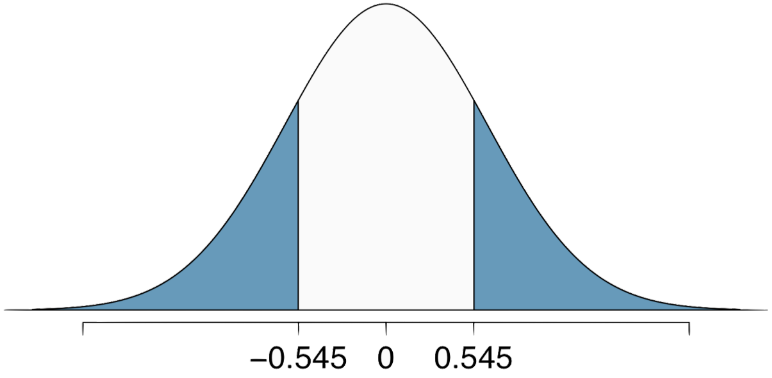
Checking assumptions & conditions

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

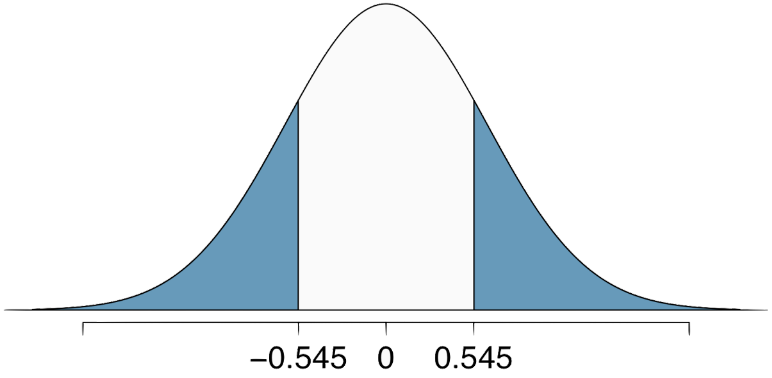
Calculating the test-statistics and the p-value

The observed average difference between the two scores 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 $\alpha = 0.05$



Calculating the test-statistics and the p-value

The observed average difference between the two scores 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



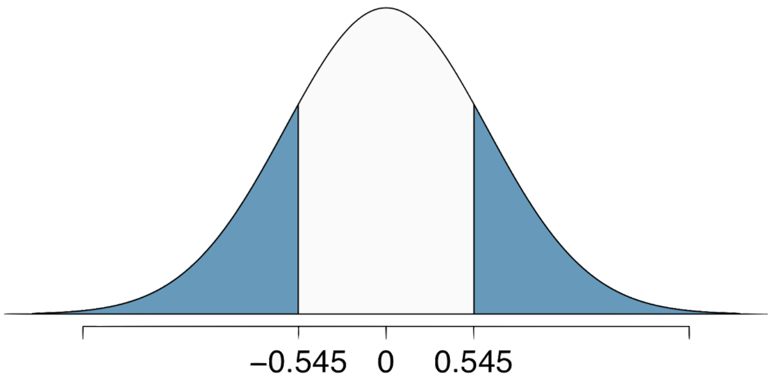
$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}}$$

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

$$df = 200 - 1 = 199$$

Calculating the test-statistics and the p-value

The observed average difference between the two scores 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



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

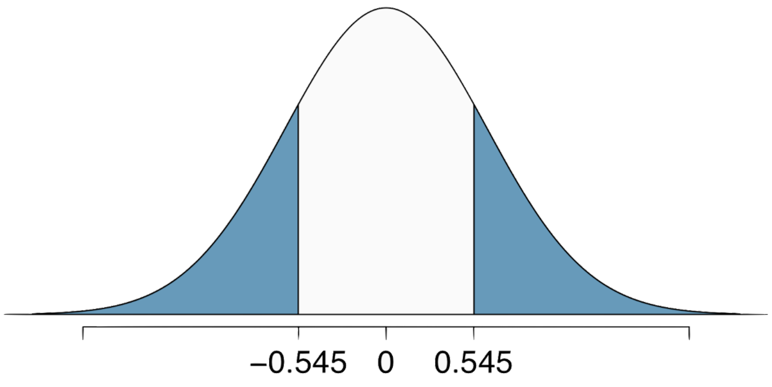
$$T = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

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

Calculating the test-statistics and the p-value

The observed average difference between the two scores 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