CAAP Statistics - Lec14

Jul 27, 2022

Review

- Point estimates and sampling variability
  - What is sampling distribution
  - Central Limit Theorem
- Confidence intervals for a proportion
  - Interpreting the confidence interval
- Hypothesis testing for a proportion
  - Null hypothesis vs. Alternative hypothesis
  - Decision Error(Type I error, Type II error)

Learning Objectives

- One-sample mean with the t-distribution
  - Sampling distribution of **sample mean,** $\bar{x}$
  - Is the average sleeping hour for the students in this class equal to 7 hours?
- Paired Data: $x_1 - y_1, x_2 - y_2, x_3 - y_3 \ldots$
  - Is there a price difference in **Trader Joe's** and **Wholefoods**?
- Difference in two means: $\bar{x} - \bar{y}$
  - Malaria vaccine case: Are **the death rates** from two groups different?

One-sample mean with the *t*-distribution

Review of Hypothesis Testing

1. Set the Hypotheses

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \neq \mu_0 \text{ or } H_A : \mu < \mu_0 \text{ or } H_A : \mu > \mu_0$$

2. Check assumptions/conditions
   - Independence
   - Normality
     - The distribution is normally distributed
     - Sample size is large enough that we can apply CLT
3. Calculate a test-statistic and a p-value

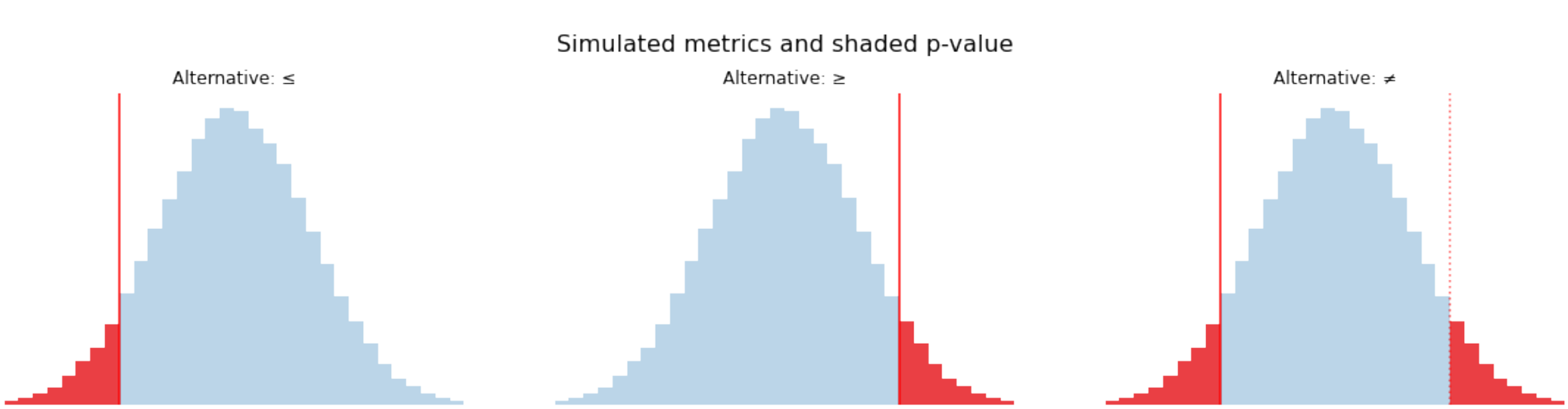
$$z = \frac{\bar{x} - \mu_0}{SE} \text{ where } SE = \sigma/\sqrt{n}$$

4. Make a decision
   - If p-value < α, reject H_0
   - If p-value > α, do not reject H_0

P-value depends on “alternative hypotheses”

- **p-value** is probability of having more extreme value than our observed value under null distribution

- The direction of extremity depends on the alternative hypothesis.



Source: https://notes.dsc10.com/05-hypothesis_testing/1_hypothesis_tests.html

Review: what is the role of “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

- **If the population distribution is normal**, CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size.

The normality condition

- **If the population distribution is normal**, CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size.
- **If the sample size is large enough**, by CLT, the sampling distributions will be nearly normal.

The normality condition

- **If the population distribution is normal**, CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size.
- **If the sample size is large enough**, by CLT, the sampling distributions will be nearly normal.
- **What if the sample size is small?**
  - It is important to not only examine the data but also think about where the data come from. For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

The t distribution

- If we can use normal distribution, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
- Instead, we use $SE = \sigma/\sqrt{n} \approx s/\sqrt{n}$. The extra uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.

The t distribution

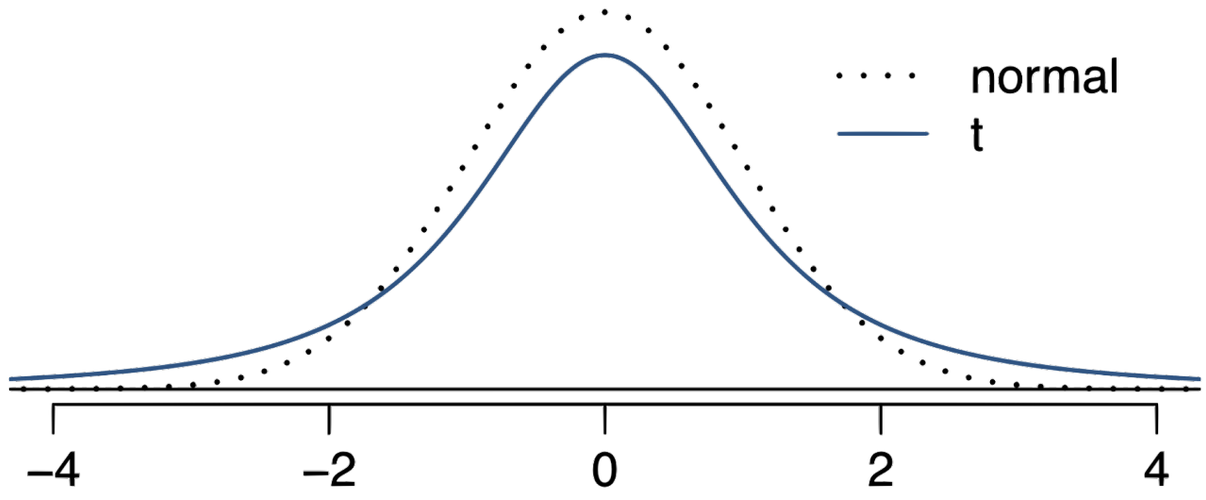
- If we can use normal distribution, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
- Instead, we use $SE = \sigma/\sqrt{n} \approx s/\sqrt{n}$. The extra uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.
- This distribution also has a bell shape, but its tails are thicker than the normal model's.

The t distribution

- If we can use normal distribution, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
- Instead, we use $SE = \sigma/\sqrt{n} \approx s/\sqrt{n}$. The extra 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 2 SDs from the mean than under the normal distribution

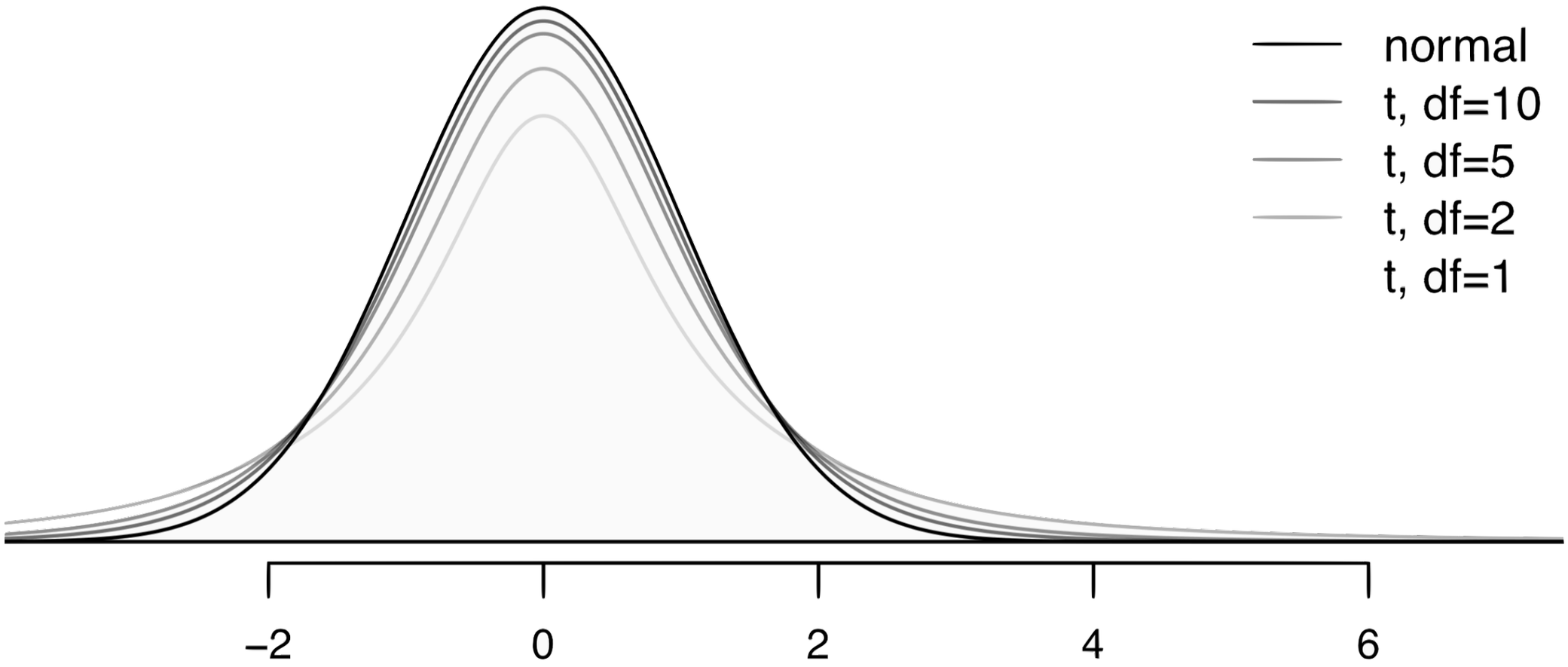
The t distribution

- If we can use normal distribution, $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
- Instead, we use $SE = \sigma/\sqrt{n} \approx s/\sqrt{n}$. The extra 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 2 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)



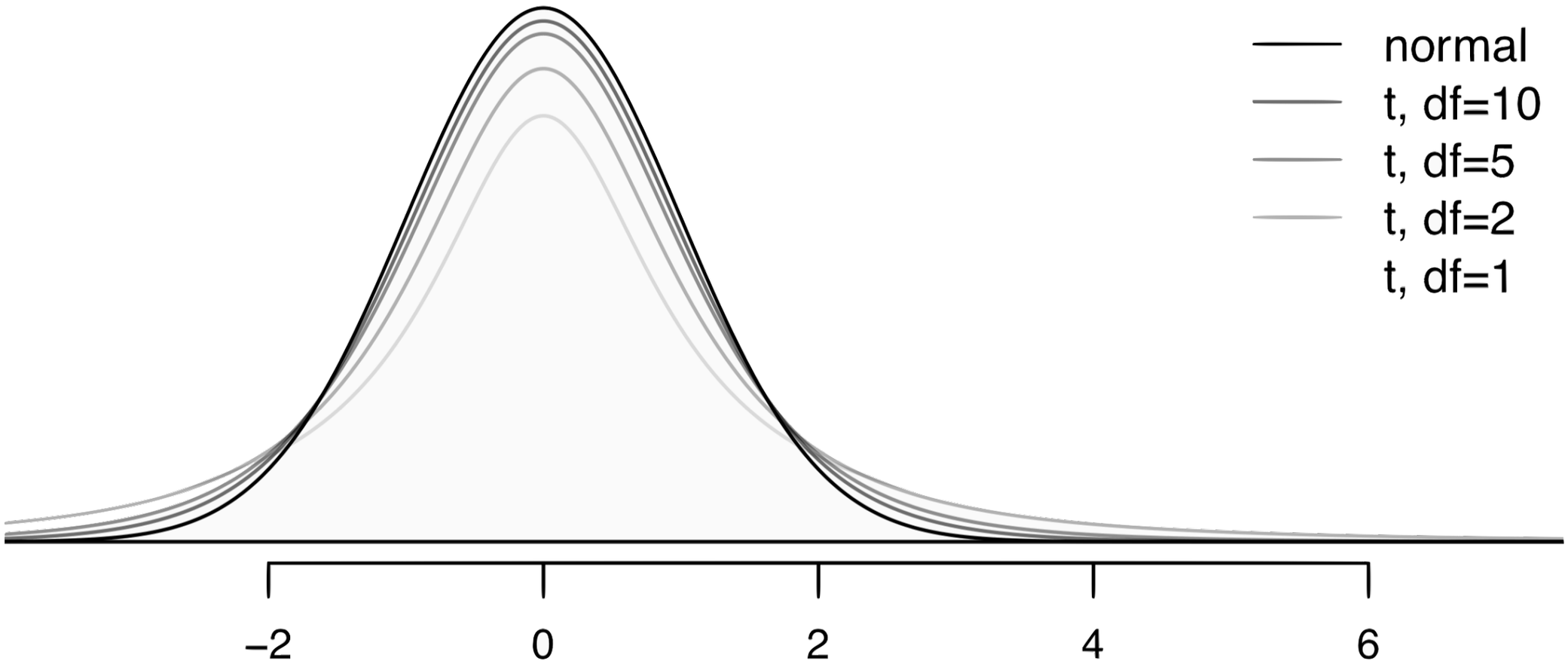
The *t* distribution (cont.)

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



The t distribution (cont.)

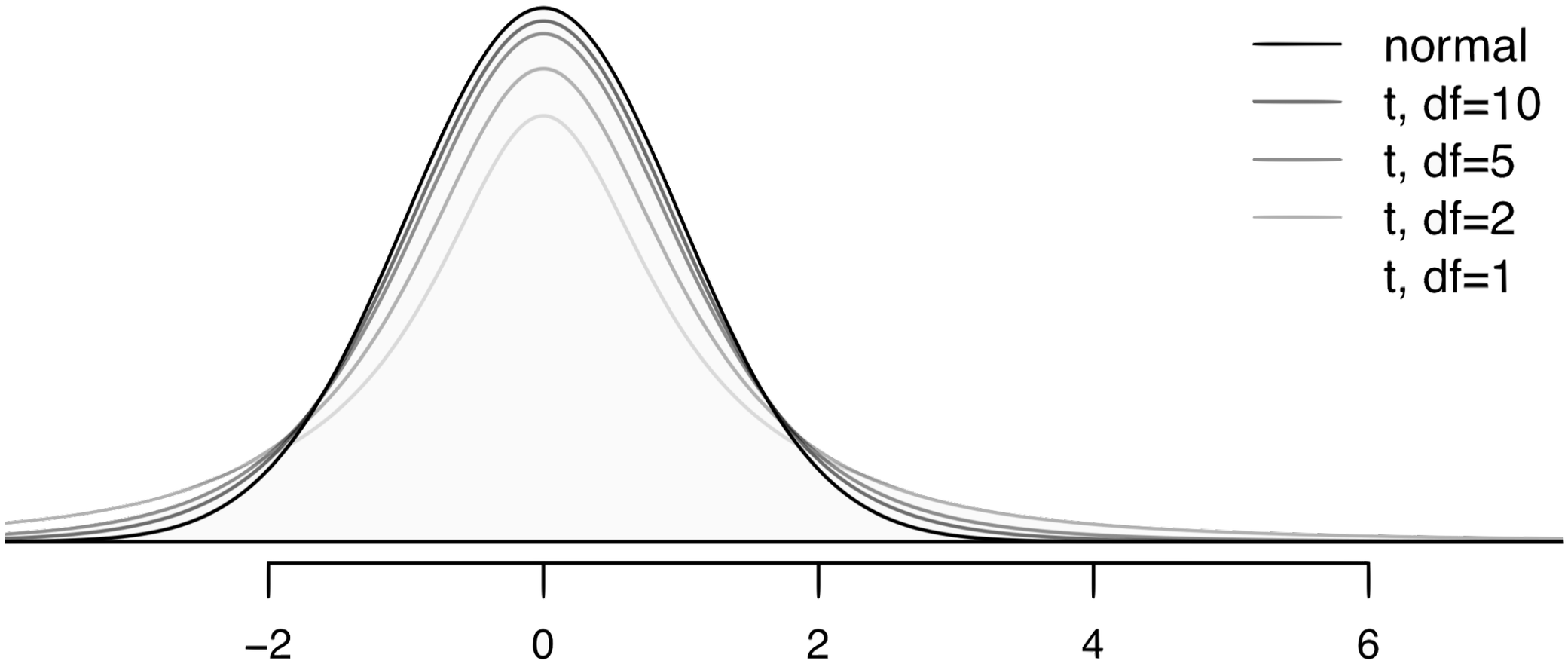
- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter: *degrees of freedom* (*df*).



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

Sleep habits of New Yorkers

New York is known as “the city that never sleeps”. A random sample of **25** New Yorkers were asked how much sleep they get per night. Statistical summaries of these data are shown below. The point estimate suggests New Yorkers sleep less than 8 hours a night on average.

n	$\bar{x}$	s	min	max
25	7.73	0.77	6.17	9.78

Corresponding hypotheses would be..

- $H_0 : \mu = 8$ hours
- $H_A : \mu < 8$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

Note: Null value is 0 because in the null hypothesis we set $\mu = 0$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

Point estimate = $\bar{x} = 7.73$

Note: Null value is 0 because in the null hypothesis we set $\mu = 0$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

Point estimate = $\bar{x} = 7.73$

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{0.77}{\sqrt{25}} = 0.154$$

Note: Null value is 0 because in the null hypothesis we set $\mu = 0$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

Point estimate = $\bar{x} = 7.73$

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{0.77}{\sqrt{25}} = 0.154$$

$$\text{T} = \frac{\bar{x} - \mu_0}{SE} = \frac{7.73 - 8}{0.154} = -1.753$$

Note: Null value is 0 because in the null hypothesis we set $\mu = 0$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

in context...

Point estimate = $\bar{x} = 7.73$

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{0.77}{\sqrt{25}} = 0.154$$

$$\text{T} = \frac{\bar{x} - \mu_0}{SE} = \frac{7.73 - 8}{0.154} = -1.753$$

d.f = 25-1 = 24

Note: Null value is 0 because in the null hypothesis we set $\mu = 0$

Finding the p-value

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

Finding the p-value

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

```
> pt(-1.753, df = 24)
[1] 0.04618422
```

Finding the p-value

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

```
> pt(-1.753, df = 24)
[1] 0.04618422
```

- Or when these aren't available, we can use a *t*-table

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 **smaller** than $\alpha = 0.05$, we conclude that the data provide enough evidence that New Yorkers actually sleep less than 8 hours. (We reject the null hypothesis)

Confidence interval for a small sample mean

- Confidence intervals are always of the form:

point estimate ± ME

Confidence interval for a small sample mean

- Confidence intervals are always of the form:

point estimate ± ME

- Margin of Error 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 ± ME

- Margin of Error 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^*?).

Confidence interval for a small sample mean

- Confidence intervals are always of the form:

point estimate ± ME

- Margin of Error 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 $\pm\ t^* \times SE$

Finding the critical value (t★)

Using R:

```
> qt(0.95, df = 24)
[1] 1.710882
```

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Synthesis

Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 8.

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 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
- Hypothesis Testing:

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

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
- Hypothesis Testing:

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

- Confidence interval:

$$point\ estimate \pm t^*_{df} \times SE$$

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
- Hypothesis Testing:

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

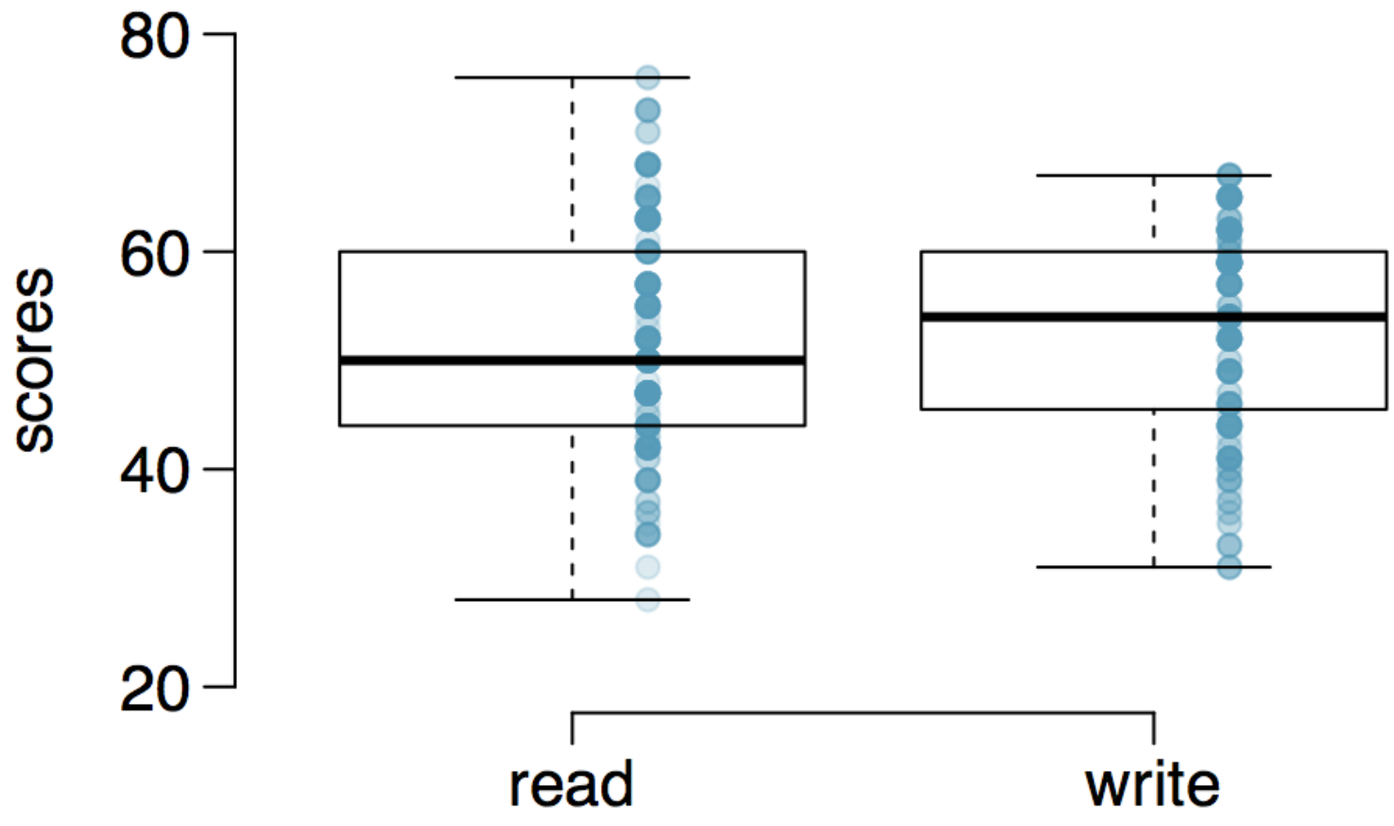
- Confidence interval:

$$point\ estimate \pm t^*_{df} \times SE$$

Paired Data

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

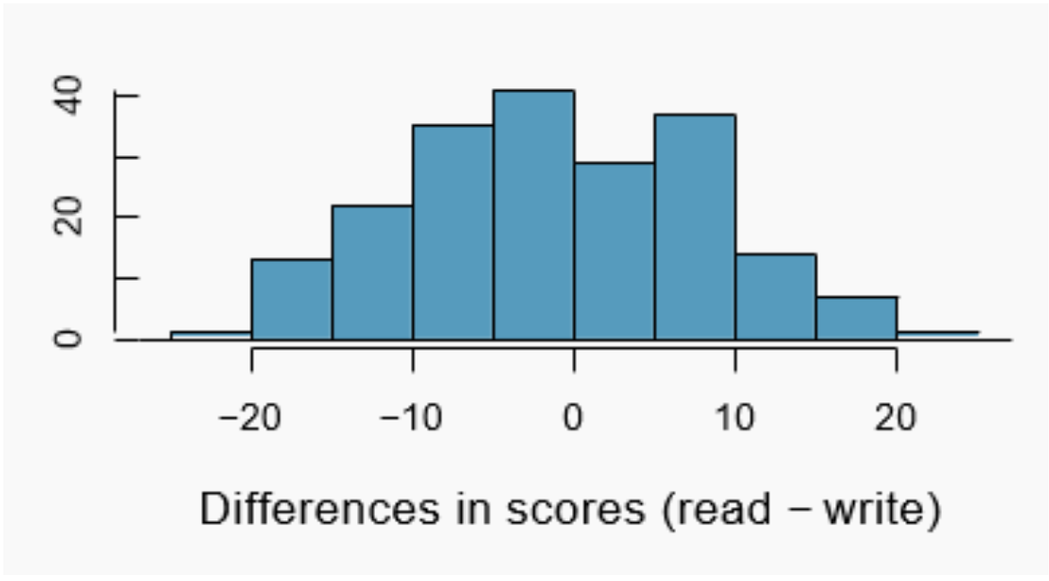
Analyzing paired data

- 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 : Average reading and writing scores are equal.

$$\mu_{diff} = 0$$

H_A : Average reading and writing 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

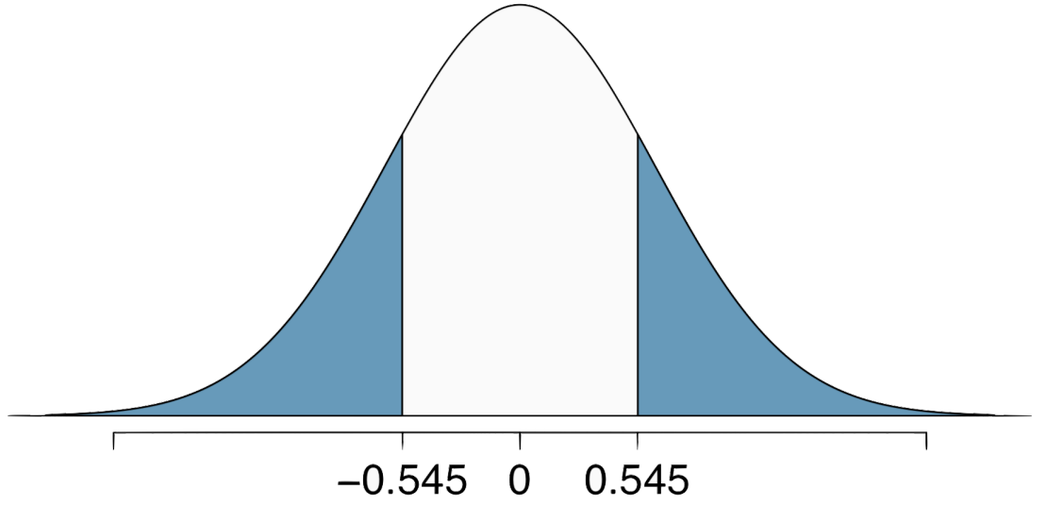
B. The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test

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

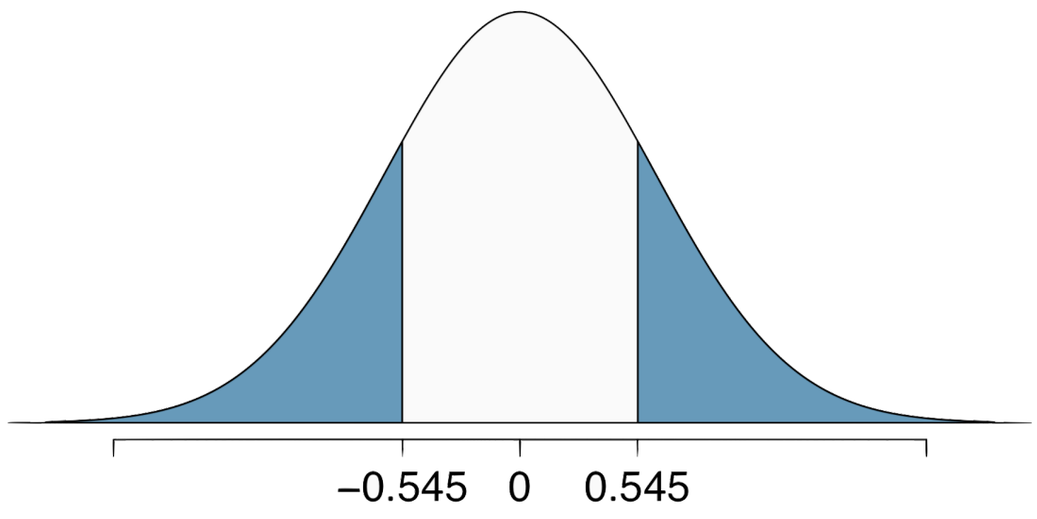
Calculating the test-statistics and the p-value

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



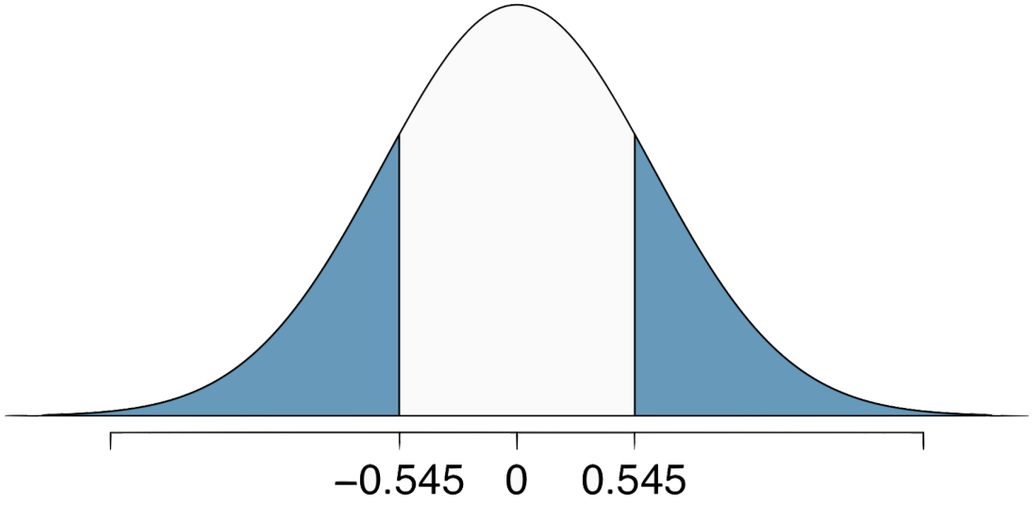
$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}}$$

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

$$df = 200 - 1 = 199$$

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 $\alpha = 0.05$



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

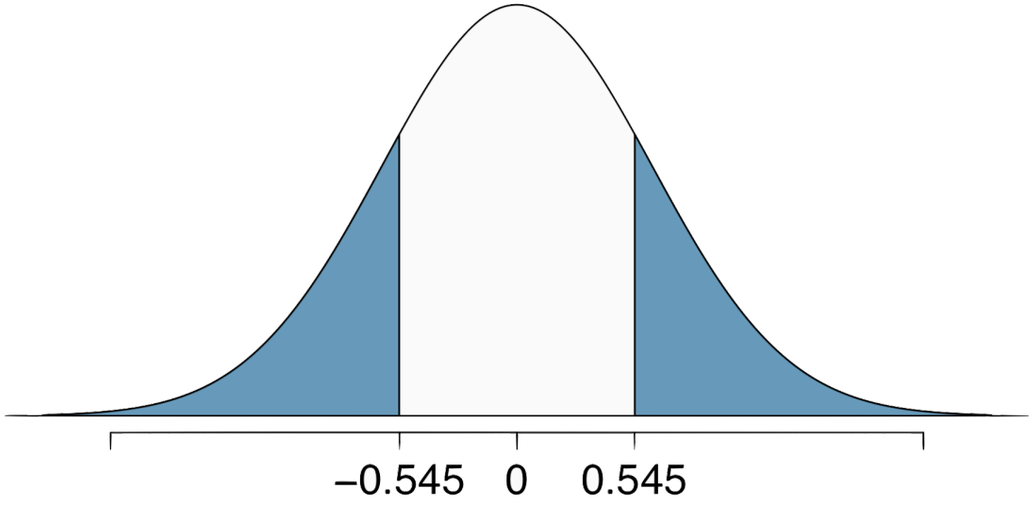
$$T = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

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

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 $\alpha = 0.05$



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

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

$$df = 200 - 1 = 199$$

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

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

Hypothesis Testing ↔ Confidence Interval

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

Hypothesis Testing ↔ Confidence Interval

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

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

$$= -0.545 \pm 1.24$$

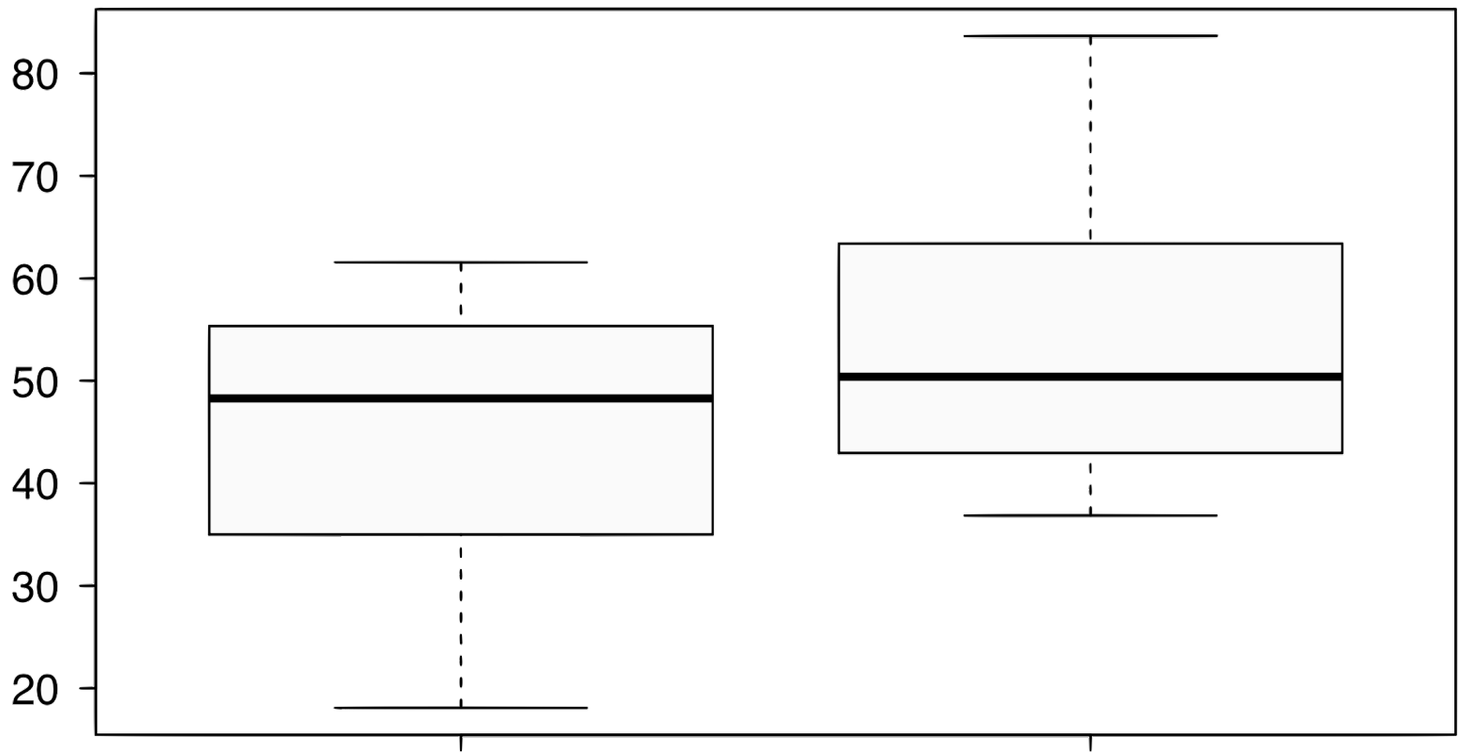
$$= (-1.785, 0.695)$$

Difference in two means

Diamonds

- Weights of diamonds are measured in carats
- 1 carat = 100 points, 0.99 carats = 99 points, etc.
- The difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but does the price of a 1 carat diamond tend to be higher than the price of a 0.99 diamond?
- We are going to test to see if there is a difference between the average prices of 0.99 and 1 carat diamonds
- In order to be able to compare equivalent units, we divide the prices of 0.99 carat diamonds by 99 and 1 carat diamonds by 100, and compare the average point prices

Data



	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

Note: These data are a random sample from the diamonds data set in ggplot2 R package.

Parameter and point estimate

- *Parameter of interest*: Average difference between the point prices of *all* 0.99 carat and 1 carat diamonds

$$\mu_{pt99} - \mu_{pt100}$$

Parameter and point estimate

- *Parameter of interest*: Average difference between the point prices of *all* 0.99 carat and 1 carat diamonds

$$\mu_{pt99} - \mu_{pt100}$$

- *Point estimate*: Average difference between the point prices of *sampled* 0.99 carat and 1 carat diamonds

$$\bar{x}_{pt99} - \bar{x}_{pt100}$$

Hypotheses

Which of the following is the correct set of hypotheses for testing if the average point price of 1 carat diamonds (pt100) is higher than the average point price of 0.99 carat diamonds (pt99)?

A. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} \neq \mu_{pt100}$

B. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} > \mu_{pt100}$

C. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} < \mu_{pt100}$

D. $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$

$H_A : \bar{x}_{pt99} \neq \bar{x}_{pt100}$

Hypotheses

Which of the following is the correct set of hypotheses for testing if the average point price of 1 carat diamonds (pt100) is higher than the average point price of 0.99 carat diamonds (pt99)?

A. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} \neq \mu_{pt100}$

B. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} > \mu_{pt100}$

C. $H_0 : \mu_{pt99} = \mu_{pt100}$

$H_A : \mu_{pt99} < \mu_{pt100}$

D. $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$

$H_A : \bar{x}_{pt99} \neq \bar{x}_{pt100}$

Conditions

Which of the following does not need to be satisfied in order to conduct **this hypothesis test(t-test)** using theoretical methods?

A. Point price of one 0.99 carat diamond in the sample should be independent of another, and the point price of one 1 carat diamond should independent of another as well
B. Point prices of 0.99 carat and 1 carat diamonds in the sample should be independent.
C. Distributions of point prices of 0.99 and 1 carat diamonds should not be extremely skewed
D. Both sample sizes should be at least 30

Conditions

Which of the following does not need to be satisfied in order to conduct **this hypothesis test(t-test)** using theoretical methods?

A. Point price of one 0.99 carat diamond in the sample should be independent of another, and the point price of one 1 carat diamond should independent of another as well
B. Point prices of 0.99 carat and 1 carat diamonds in the sample should be independent.
C. Distributions of point prices of 0.99 and 1 carat diamonds should not be extremely skewed
D. Both sample sizes should be at least 30

Test statistics

Test statistic for inference on the difference of two small sample means

The test statistic for inference on the difference of two means where σ_1 and σ_2 are unknown is the *T* statistic.

where

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{and} \qquad df = min(n_1 - 1, n_2 - 1)$$

Note: The calculation of the *df* is actually much more complicated(Welch–Satterthwaite equation). For simplicity we'll use the above formula to estimate the true df when conducting the analysis by hand

Test statistics (cont.)

	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

In context...

Test statistics (cont.)

	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

In context...

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

Test statistics (cont.)

	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

In context...

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}}$$

Test statistics (cont.)

	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

In context...

$$
\begin{aligned}
T &= \frac{\text{point estimate} - \text{null value}}{SE} \\
&= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}} \\
&= \frac{-8.93}{3.56}
\end{aligned}
$$

Test statistics (cont.)

	0.99 carat	*1 carat*
	pt99	pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

In context...

$$
\begin{aligned}
T &= \frac{\text{point estimate} - \text{null value}}{SE} \\
&= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}} \\
&= \frac{-8.93}{3.56} \\
&= -2.508
\end{aligned}
$$

Test statistics (cont.)

Which of the following is the correct *df* for this hypothesis test?

A. 22
B. 23
C. 30
D. 29
E. 52

Test statistics (cont.)

Which of the following is the correct *df* for this hypothesis test?

A. 22
B. 23
C. 30
D. 29
E. 52

$$df = \min(n_{pt99} - 1, n_{pt100} - 1)$$
$$= \min(23 - 1, 30 - 1)$$
$$= \min(22, 29)$$

p-value

Which of the following is the correct p-value for this hypothesis test?

$$T = -2.508 \qquad df = 22$$

A. between 0.005 and 0.01
B. between 0.01 and 0.025
C. between 0.02 and 0.05
D. between 0.01 and 0.02

p-value

Which of the following is the correct p-value for this hypothesis test?

$$T = -2.508 \qquad df = 22$$

A. between 0.005 and 0.01
B. between 0.01 and 0.025
C. between 0.02 and 0.05
D. between 0.01 and 0.02

```
> pt(q = -2.508, df = 22)
[1] 0.0100071
```

Synthesis

What is the conclusion of the hypothesis test? How (if at all) would this conclusion change your behavior if you went diamond shopping?

Synthesis

What is the conclusion of the hypothesis test? How (if at all) would this conclusion change your behavior if you went diamond shopping?

- p-value is small so reject H_0. The data provide convincing evidence to suggest that the point price of 0.99 carat diamonds is lower than the point price of 1 carat diamonds
- Maybe buy a 0.99 carat diamond? It looks like a 1 carat, but is significantly cheaper

Equivalent confidence level

What is the equivalent confidence level for a one-sided hypothesis test at $\alpha = 0.05$?

A. 90%
B. 92.5%
C. 95%
D. 97.5%

Critical value

What is the appropriate t^* for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

A. 1.32
B. 1.72
C. 2.07
D. 2.82

Critical value

What is the appropriate t^* for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

A. 1.32
B. 1.72
C. 2.07
D. 2.82

```
> qt(p = 0.95, df = 22)
[1] 1.717144
```

Confidence interval

Calculate the interval, and interpret it in context

$$point\ estimate \pm ME$$

Confidence interval

Calculate the interval, and interpret it in context

$$point\ estimate \pm ME$$

$$\left(\bar{x}_{pt99} - \bar{x}_{pt1}\right) \pm t^*_{df} \times SE = (44.50\ \ - 53.43) \pm 1.72 \times 3.56$$

Confidence interval

Calculate the interval, and interpret it in context

$$point\ estimate \pm ME$$

$$\left(\bar{x}_{pt99} - \bar{x}_{pt1}\right) \pm t^*_{df} \times SE = (44.50\ \ - 53.43) \pm 1.72 \times 3.56$$

$$= -8.93 \pm 6.12$$

Confidence interval

Calculate the interval, and interpret it in context

$$point\ estimate \pm ME$$

$$\begin{aligned}
\left(\bar{x}_{pt99} - \bar{x}_{pt1}\right) \pm t^*_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\
&= -8.93 \pm 6.12 \\
&= (-15.05, -2.81)
\end{aligned}$$

Confidence interval

Calculate the interval, and interpret it in context

$$point\ estimate \pm ME$$

$$\begin{aligned}
\left(\bar{x}_{pt99} - \bar{x}_{pt1}\right) \pm t^*_{df} \times SE &= (44.50\ \ - 53.43) \pm 1.72 \times 3.56 \\
&= -8.93 \pm 6.12 \\
&= (-15.05, -2.81)
\end{aligned}$$

We are 90% confident that the average point price of a 0.99 carat diamond is $15.05 to $2.81 lower than the average point price of a 1 carat diamond

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, n < 10% of population) and between groups
  - no extreme skew in either group

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, n < 10% of population) and between groups
  - no extreme skew in either group
- Hypothesis testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = \min(n_1 - 1, n_2 - 1)$$

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a ***t***-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
- Conditions:
  - independence within groups (often verified by a random sample, and if sampling without replacement, n < 10% of population) and between groups
  - no extreme skew in either group
- Hypothesis testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = \min(n_1 - 1, n_2 - 1)$$

- Confidence interval: $point\ estimate \pm t^*_{df} \times SE$

p-value

Which of the following is the correct p-value for this hypothesis test?

$$T = -2.508 \qquad df = 22$$

A. between 0.005 and 0.01
B. between 0.01 and 0.025
C. between 0.02 and 0.05
D. between 0.01 and 0.02

one tail		0.100	0.050	0.025	0.010
two tails		0.200	0.100	0.050	0.020
df	21	1.32	1.72	2.08	2.52
	22	1.32	1.72	2.07	2.51
	23	1.32	1.71	2.07	2.50
	24	1.32	1.71	2.06	2.49
	25	1.32	1.71	2.06	2.49

p-value

Which of the following is the correct p-value for this hypothesis test?

$$T = -2.508 \qquad df = 22$$

A. between 0.005 and 0.01
B. between 0.01 and 0.025
C. between 0.02 and 0.05
D. between 0.01 and 0.02

one tail	0.100	0.050	*0.025*	*0.010*
two tails	0.200	0.100	0.050	0.020
df 21	1.32	1.72	2.08	2.52
22	1.32	1.72	*2.07*	*2.51*
23	1.32	1.71	2.07	2.50
24	1.32	1.71	2.06	2.49
25	1.32	1.71	2.06	2.49

Critical value

What is the appropriate t^* for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

A. 1.32
B. 1.72
C. 2.07
D. 2.82

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 21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79

Critical value

What is the appropriate t^* for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

A. 1.32
B. 1.72
C. 2.07
D. 2.82

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 21	1.32	1.72	2.08	2.52	2.83
22	1.32	*1.72*	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79

Let’s discuss!

Sample size and pairing

Determine if the following statement is true or false, and if false, explain your reasoning: If comparing means of two groups with equal sample sizes, always use a paired test.

Hen eggs

The distribution of the number of eggs laid by a certain species of hen during their breeding period has a mean of 35 eggs with a standard deviation of 18.2. Suppose a group of researchers randomly samples 45 hens of this species, counts the number of eggs laid during their breeding period, and records the sample mean. They repeat this 1,000 times, and build a distribution of sample means.

(a) What is this distribution called?

(b) Would you expect the shape of this distribution to be symmetric, right skewed, or left skewed? Explain your reasoning.

(c) Suppose the researchers' budget is reduced and they are only able to collect random samples of 10 hens. The sample mean of the number of eggs is recorded, and we repeat this 1,000 times, and build a new distribution of sample means. How will the variability of this new distribution compare to the variability of the original distribution?

Online communication

A study suggests that the average college student spends 10 hours per week communicating with others online. You believe that this is an underestimate and decide to collect your own sample for a hypothesis test. You randomly sample 60 students from your dorm and find that on average they spent 13.5 hours a week communicating with others online. A friend of yours, who offers to help you with the hypothesis test, comes up with the following set of hypotheses. Indicate any errors you see.

$H_0 : \bar{x} < 10$ hours

$H_A : \bar{x} > 13.5$ hours

Tomorrow is R Session!

- The first half of the lecture will be R session
- The second half of the lecture will be the time for the project discussion
  - Objective: Exploratory Data Analysis
    - Try to summarize the variable
    - Conduct the preliminary(eyeball) investigation for your hypotheses
- Try to digest the project as much as possible during the lecture time!