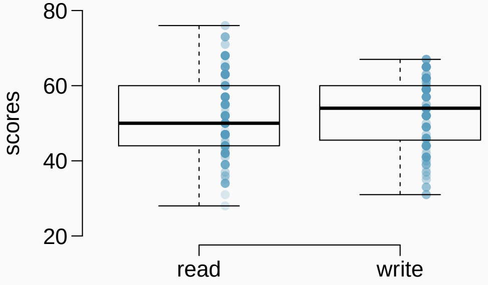
Announcements

- Homework
  - Next weeks homework posted, due on Monday
  - In-class lab 8 due Monday as well
- Next in-class lab will be next Wednesday
- I'll be sending out more information this weekend about the final projects
- Physics Tea today at 3pm!
- Polling: `rembold-class.ddns.net`

Reading and Writing...

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?



Independence?

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

C) Impossible to say

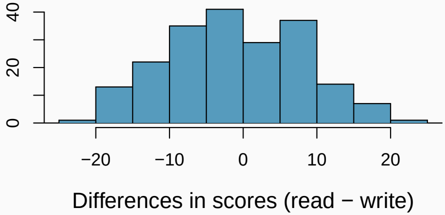
Analyzing paired data

- 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.

$$\mu_{diff}$$

- ***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?

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?

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: There is no difference between the average reading and writing score.

$$\mu_{diff} = 0$$

H_A: There is a difference between the average reading and writing score.

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

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 student.

(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% of 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 student.

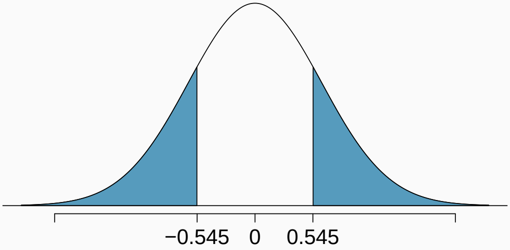
(b) The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

(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% of all students, we can assume that the sampling distribution of the average difference will be nearly normal.

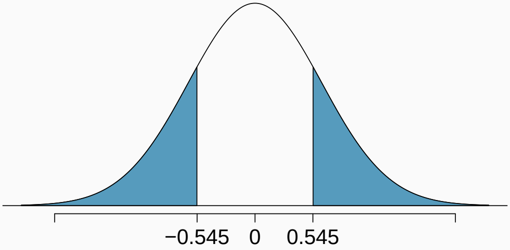
Calculating the test-statistic and the p-value

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-statistic and the p-value

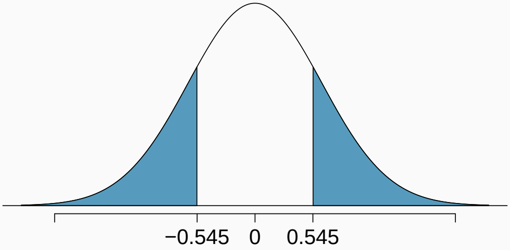
The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.



$$
\begin{aligned}
T &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} \\
&= \frac{-0.545}{0.628} = -0.87 \\
df &= 200 - 1 = 199
\end{aligned}
$$

Calculating the test-statistic and the p-value

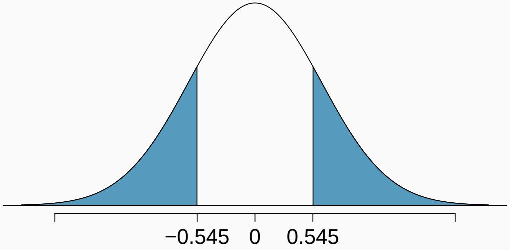
The observed average difference between the two scores is -0.545 points and the standard deviation of the difference is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams? Use $\alpha = 0.05$.



$$
\begin{aligned}
T &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} \\
&= \frac{-0.545}{0.628} = -0.87 \\
df &= 200 - 1 = 199 \\
p - value &= 0.1927 \times 2 = 0.3854
\end{aligned}
$$

Calculating the test-statistic 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 $\alpha = 0.05$.



$$
\begin{aligned}
T &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} \\
&= \frac{-0.545}{0.628} = -0.87 \\
df &= 200 - 1 = 199 \\
p - value &= 0.1927 \times 2 = 0.3854
\end{aligned}
$$

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

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?

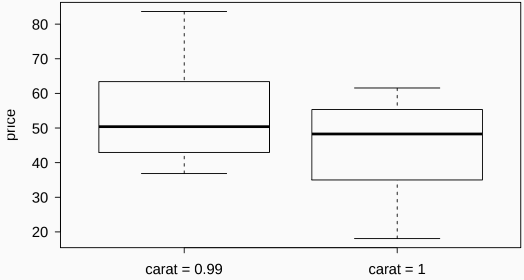
(a) yes
(b) no
(c) cannot tell from the information given

$$\begin{aligned} -0.545 \pm 1.97 \frac{8.887}{\sqrt{200}} &= -0.545 \pm 1.97 \times 0.628 \\ &= -0.545 \pm 1.24 \\ &= (-1.785, 0.695) \end{aligned}$$

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

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

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) $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$
   $H_A : \bar{x}_{pt99} < \bar{x}_{pt100}$

Test statistic

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.

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

where

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

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

Test statistic (cont.)

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

Here:

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

Test statistic (cont.)

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

Here:

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

Test statistic (cont.)

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

Here:

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

Test statistic (cont.)

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

Here:

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

p-value

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

$$T = -2.508$$

(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	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

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	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

p-value

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

$$T = -2.508$$

(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	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

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.

Confidence interval

Calculate the interval, and interpret it in context.

Confidence interval

Calculate the interval, and interpret it in context.

$$\text{point estimate} \pm ME$$

Confidence interval

Calculate the interval, and interpret it in context.

$$\text{point estimate} \pm ME$$

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

Confidence interval

Calculate the interval, and interpret it in context.

$$\text{point estimate} \pm ME$$

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

Confidence interval

Calculate the interval, and interpret it in context.

$$\text{point estimate} \pm ME$$

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

Confidence interval

Calculate the interval, and interpret it in context.

$$\text{point estimate} \pm ME$$

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

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:

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

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:

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

- Confidence interval:

$$\text{point estimate} \pm t_{df}^{\star} \times SE$$