Chapter 7: Inference for Numerical Data

Math 140 · Fall '21

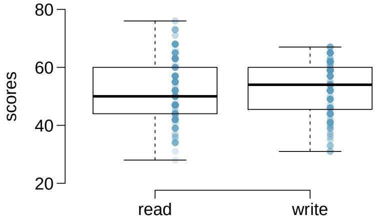
Based on content in OpenIntro Stats, 4th Ed

Hitchman

November 3, 2021

Section 7.2
Paired t-test

Q: *200 students were randomly sampled from a "High School and Beyond survey." Each student had taken 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?*



The same students took the reading and writing tests and their scores are shown below. Are the reading and writing scores 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

The same students took the reading and writing tests and their scores are shown below. Are the reading and writing scores 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

No, the scores are linked by the student. Each student took the reading and writing tests.

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.

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

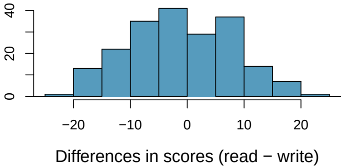
Analyzing paired data

- 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

Q: *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

Q: *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

Q: *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

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

Setting the hypotheses

Q: *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

Q: 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.

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

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.*

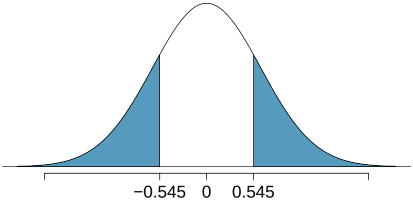
(b) The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

(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.

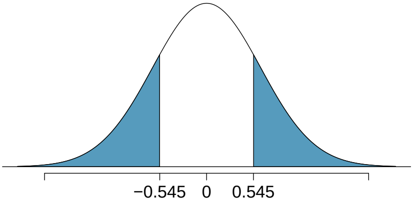
Calculating the test-statistic and the p-value

Q: *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

Q: *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* $\alpha = 0.05$.

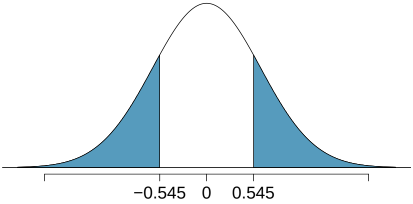


$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

Calculating the test-statistic and the p-value

Q: *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* $\alpha = 0.05$.



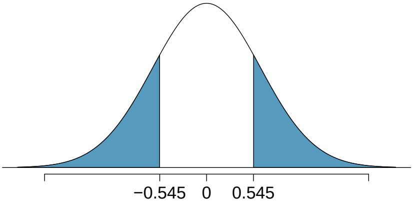
$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

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

Calculating the test-statistic and the p-value

Q: *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* $\alpha = 0.05$.



$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} = \frac{-0.545}{0.628} = -0.87$$

$$df = 200 - 1 = 199$$

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

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

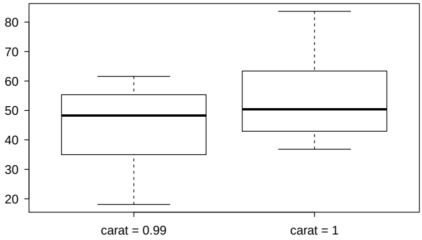
$$\begin{aligned} -0.545 \pm 1.97 \frac{8.887}{\sqrt{200}} &= -0.545 \pm 1.97 \times 0.628 \\ &= -0.545 \pm 1.24 \\ &= (-1.785, 0.695) \end{aligned}$$

Section 7.3
Inference on 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:* The difference between the average point price of all 0.99 carat diamonds and the average point price of all 1 carat diamonds.

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

Parameter and point estimate

- *Parameter of interest:* The difference between the average point price of all 0.99 carat diamonds and the average point price of all 1 carat diamonds.

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

- *Point estimate:* The difference between the average point price of sampled 0.99 carat and the average point price of sampled 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) $H_0 : \bar{x}_{pt99} = \bar{x}_{pt100}$
$H_A : \bar{x}_{pt99} < \bar{x}_{pt100}$

Conditions

Which of the following does not need to be satisfied in order to conduct this hypothesis 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 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 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 pt99	*1 carat* pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

in context...

Test statistic (cont.)

	0.99 carat pt99	*1 carat* 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 statistic (cont.)

	0.99 carat pt99	*1 carat* 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 statistic (cont.)

	0.99 carat pt99	*1 carat* 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 statistic (cont.)

	0.99 carat pt99	*1 carat* 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 statistic (cont.)

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

(a) 22
(b) 23
(c) 30
(d) 29
(e) 52

Test statistic (cont.)

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

(a) *22*
(b) 23
(c) 30
(d) 29
(e) 52

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

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

Synthesis

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

Synthesis

Q: *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.*

Summary: Inference on two means

- If σ_1 or σ_2 is unknown, difference between two 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:

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