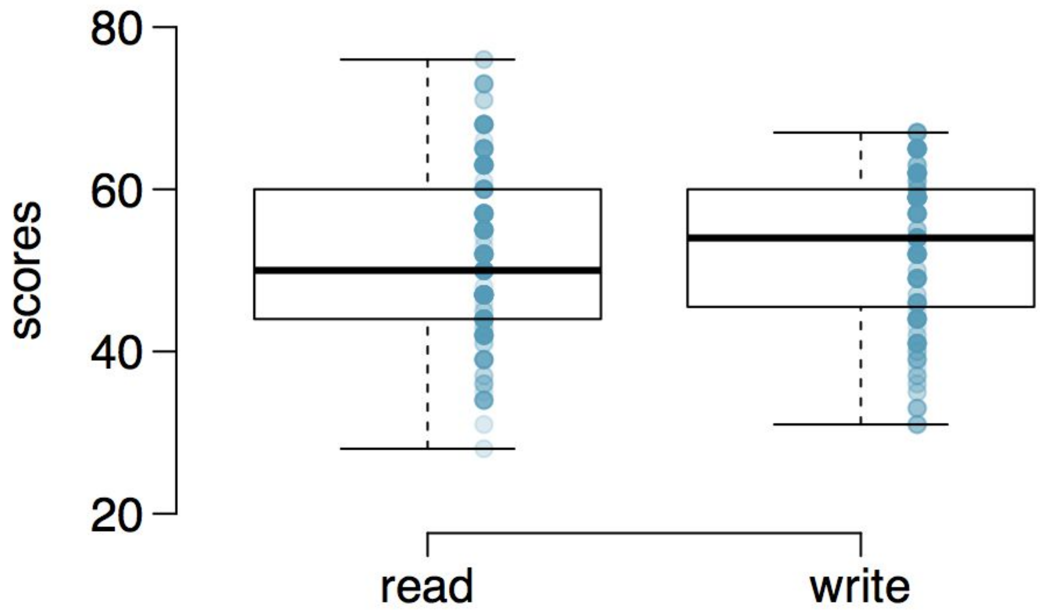
Section 7.2
Inference for paired data

Stats 7 Summer Session II 2022

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

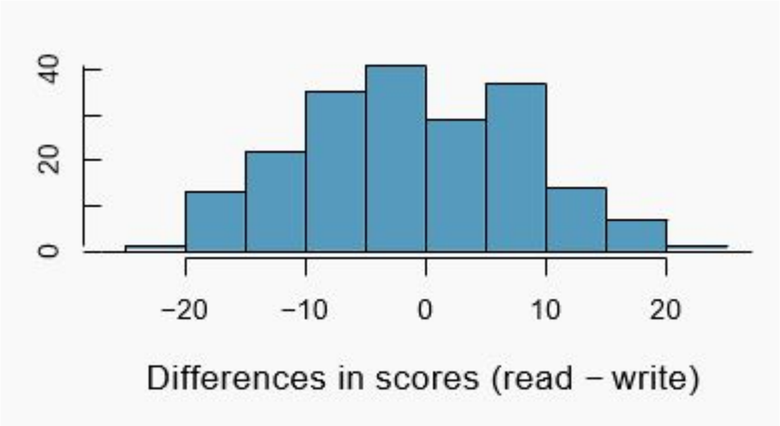
Analyzing paired data

- 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

$$\mu_{diff}$$

- *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?

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

- Once we have the differences between reading and writing scores for each student we use those as our observations
- 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.

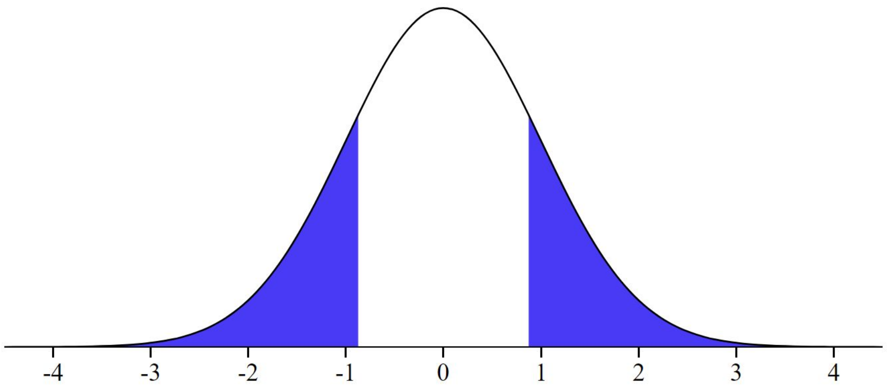
Checking assumptions & conditions

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. 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 α = 0.05



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

$$df = n - 1 = 200 - 1 = 199$$

```
> 2 * pt(-0.87, df = 199)
[1] 0.3853486
```

Since p-value > 0.05, fail to reject.

The data do not provide convincing evidence (p-value = 0.3853 > α = 0.05) 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

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

Let's calculate it. First we need the t^* that corresponds to 95% confidence for the t-distribution with 199 degrees of freedom.

```
> qt((1 - 0.95) / 2, df = 199)
[1] -1.971957
```

Now we can calculate our confidence interval. Note one we have computed the differences the scenario is the same as a one-sample mean.

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

$$-0.545 \pm 1.97 * \frac{8.887}{\sqrt{199}}$$

$$-0.545 \pm 1.241$$

$$(-1.79, 0.70)$$