Advanced Statistics DS2003 (BDS-4A) Lecture 14

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

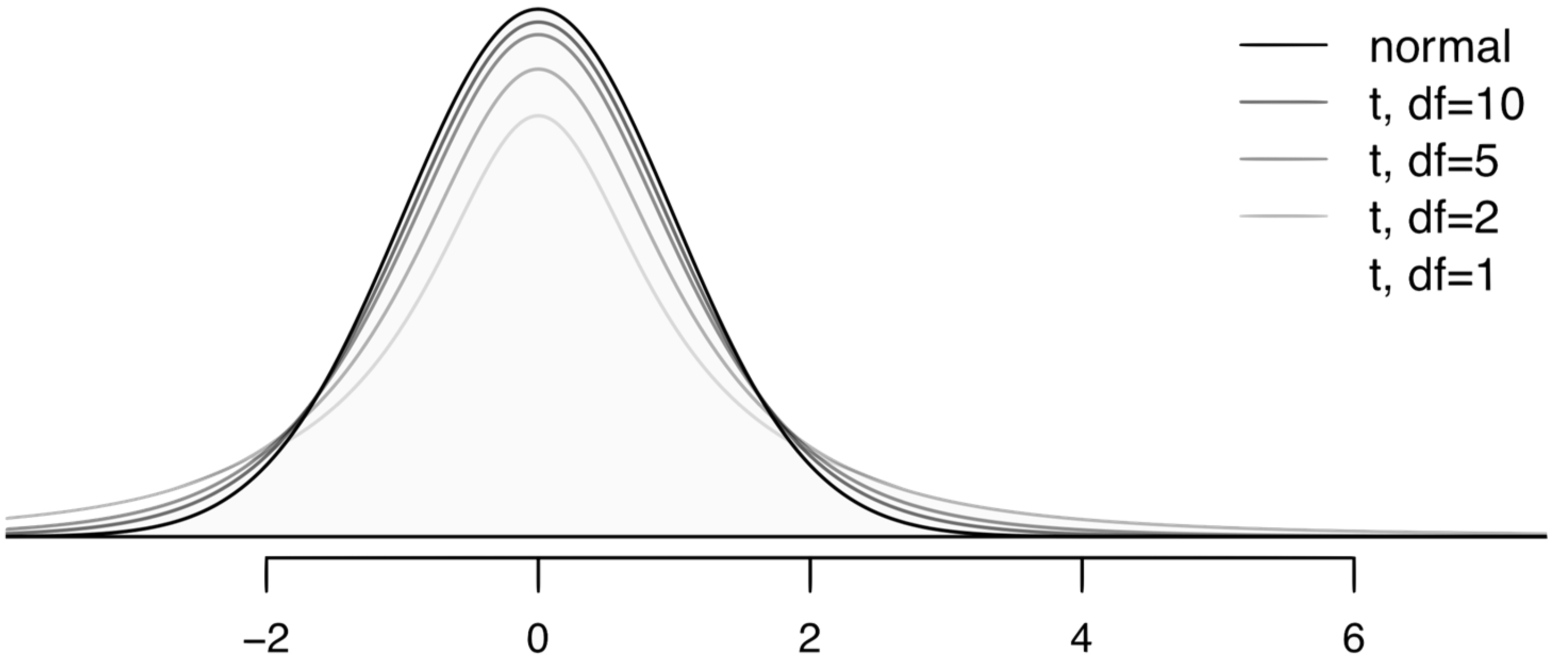
07 April, 2022

Previous Lecture

- One sample mean with the t-distribution
  - Example: traffic flow on Friday 13th v. traffic flow on Friday 6th
  - For finding the confidence interval, we need t* instead of z*
  - Making conclusions from hypothesis tests (but not making inferences about reasons for a significant result)

The t distribution (continued)

- 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*

Complete R code (https://rdrr.io/snippets)

```
num_diff <- c(698, 1104, 1037, 1889, 1911, 2416, 2761, 4382, 1839, 321)
print(paste("Sample mean = ", mean(num_diff)))
print(paste("Standard deviation of sample = ", sd(num_diff)))
std_error = sd(num_diff)/sqrt(length(num_diff))
print(paste("Standard error of sample = ", std_error))
t_stat_diff = (mean(num_diff) - 0)/std_error
print(paste("T statistic = ", t_stat_diff))
print(paste("p-value = ", 2 * pt(t_stat_diff, df = length(num_diff)-1, lower.tail = FALSE)))
```

Output:

```
[1] "Sample mean =  1835.8"
[1] "Standard deviation of sample =  1176.01386991065"
[1] "Standard error of sample =  371.888238886661"
[1] "T statistic =  4.93642930331951"
[1] "p-value =  0.000806184359845231"
```

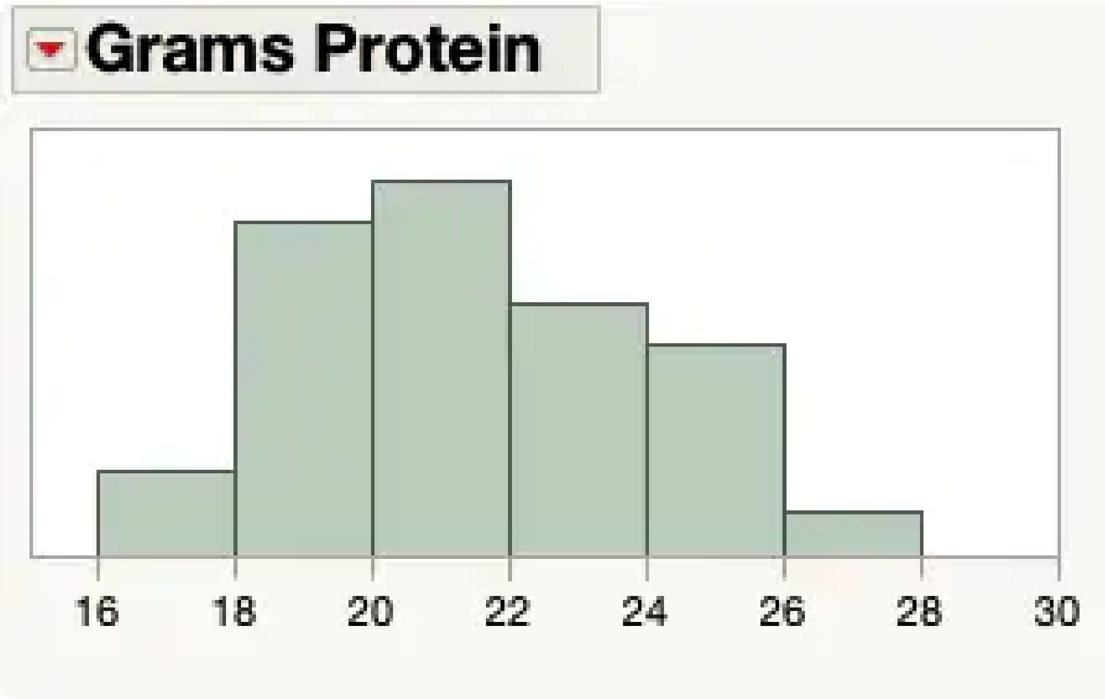


Figure 1: Histogram and summary statistics

Mean = 21.399
St_dev = 2.54186
St_error = 2.54186/sqrt(31) = 0.4565
DF = 31-1 = 30

R code:

```
t_stat = (21.399-20)/0.4565
print(2 * pt(t_stat, df = 30, lower.tail = FALSE))
```

Output:
0.004578209 (p-value indicates we can reject the null hypothesis, it is less than 0.05)

Reject the claim that the energy bar contains 20 grams of protein

The t-distribution

- **What is the *t*-distribution?**
    - The *t*-distribution describes the standardized distances of sample means to the population mean when the population standard deviation is not known, and the observations come from a normally distributed population.
- **Is the *t*-distribution the same as the Student's *t*-distribution?**
    - Yes
- **What's the key difference between the *t*- and z-distributions?**
    - The standard normal or z-distribution assumes that you know the population standard deviation. The *t*-distribution is based on the sample standard deviation.

t-Distribution vs. normal distribution

The *t*-distribution is similar to a normal distribution. It has a precise mathematical definition. Instead of diving into complex math, let’s look at the useful properties of the *t*-distribution and why it is important in analyses.

- Like the normal distribution, the *t*-distribution has a smooth shape.
- Like the normal distribution, the *t*-distribution is symmetric. If you think about folding it in half at the mean, each side will be the same.
- Like a standard normal distribution (or z-distribution), the *t*-distribution has a mean of zero.
- The normal distribution assumes that the population standard deviation is known. The *t*-distribution does not make this assumption.
- The *t*-distribution is defined by the *degrees of freedom*. These are related to the sample size.
- The *t*-distribution is most useful for small sample sizes, when the population standard deviation is not known, or both.
- As the sample size increases, the *t*-distribution becomes more similar to a normal distribution.

Comparing t and Z distributions

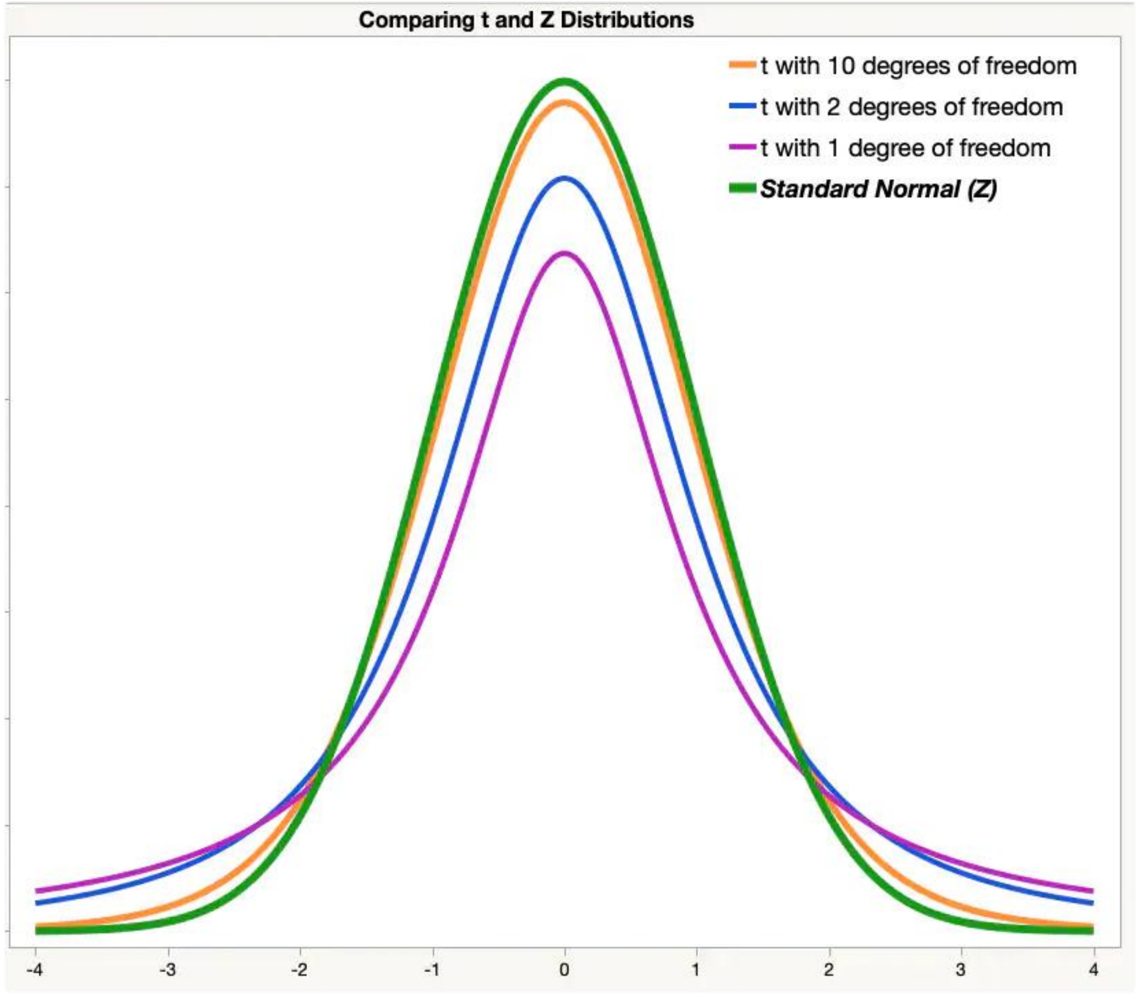


Figure 1: Three t-distributions and a standard normal (z-) distribution.

Common Rule of Thumb

- A common rule of thumb is that for a *sample size of at least 30*, one can use the z-distribution in place of a *t*-distribution.

The Paired *t*-Test

- **What is the paired *t*-test?**
  - The paired *t*-test is a method used to test whether the mean difference between pairs of measurements is zero or not.
- **When can I use the test?**
  - You can use the test when your data values are paired measurements. For example, you might have before-and-after measurements for a group of people. Also, the distribution of differences between the paired measurements should be normally distributed.
- **What are some other names for the paired *t*-test?**
  - The paired *t*-test is also known as the dependent samples *t*-test, the paired-difference *t*-test, the matched pairs *t*-test and the repeated-samples *t*-test.
- **What if my data isn’t nearly normally distributed?**
  - If your sample sizes are very small, you might not be able to test for normality. You might need to rely on your understanding of the data. Or, you can perform a *nonparametric* test that doesn’t assume normality.

Using the paired *t*-test

- **What do we need?**
  - For the paired *t*-test, we need two variables. One variable defines the pairs for the observations. The second variable is a measurement. Sometimes, we already have the paired differences for the measurement variable. Other times, we have separate variables for “before” and “after” measurements for each pair and need to calculate the differences.
- We also have an idea, or hypothesis, that the differences between pairs is zero. Here is an example:
  - We measure weights of people in a program to quit smoking. For each person, we have the weight at the start and end of the program. We want to know if the mean weight change for people in the program is zero or not.

Paired *t*-test assumptions

To apply the paired t-test to test for differences between paired measurements, the following assumptions need to hold:

- Subjects must be independent. Measurements for one subject do not affect measurements for any other subject.
- Each of the paired measurements must be obtained from the same subject. For example, the before-and-after weight for a smoker in the example above must be from the same person.
- The measured differences are normally distributed.

- An instructor wants to use two exams in her classes next year. This year, she gives both exams to the students.
- She wants to know if the exams are equally difficult and wants to check this by looking at the differences between scores.
- If the mean difference between scores for students is “close enough” to zero, she will make a practical conclusion that the exams are equally difficult.

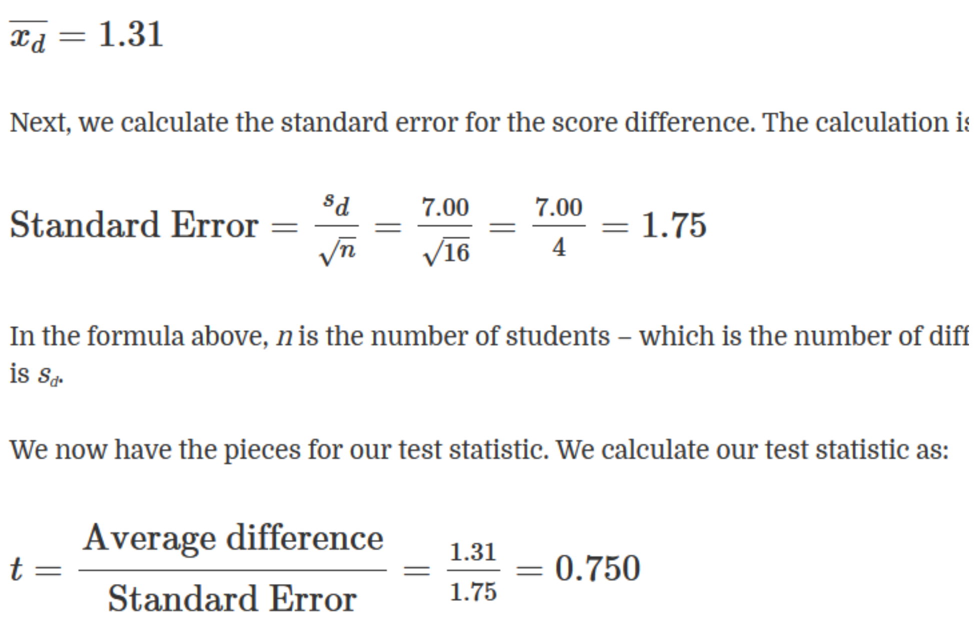
Here is the data:

Student	Exam 1 Score	Exam 2 Score	Difference
Bob	63	69	
Nina	65	65	
Tim	56	62	
Kate	100	91	
Alonzo	88	78	
Jose	83	87	
Nikhil	77	79	
Julia	92	88	
Tohru	90	85	
Michael	84	92	
Jean	68	69	
Indra	74	81	
Susan	87	84	
Allen	64	75	
Paul	71	84	
Edwina	88	82	

Student	Exam 1 Score	Exam 2 Score	Difference
Bob	63	69	6
Nina	65	65	0
Tim	56	62	6
Kate	100	91	-9
Alonzo	88	78	-10
Jose	83	87	4
Nikhil	77	79	2
Julia	92	88	-4
Tohru	90	85	-5
Michael	84	92	8
Jean	68	69	1
Indra	74	81	7
Susan	87	84	-3
Allen	64	75	11
Paul	71	84	13
Edwina	88	82	-6

- An instructor wants to use two exams in her classes next year. This year, she gives both exams to the students.
- She wants to know if the exams are equally difficult and wants to check this by looking at the differences between scores.
- If the mean difference between scores for students is “close enough” to zero, she will make a practical conclusion that the exams are equally difficult.

Here is the data:



$$\overline{x_d} = 1.31$$

Next, we calculate the standard error for the score difference. The calculation is

$$\text{Standard Error} = \frac{s_d}{\sqrt{n}} = \frac{7.00}{\sqrt{16}} = \frac{7.00}{4} = 1.75$$

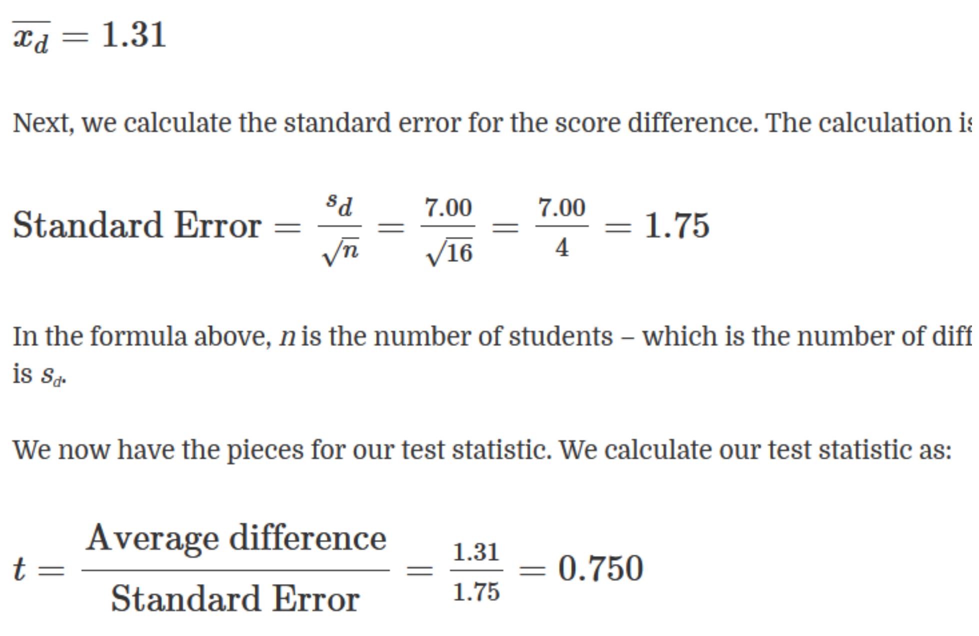
In the formula above, n is the number of students – which is the number of diff

is s_d.

We now have the pieces for our test statistic. We calculate our test statistic as:

$$t = \frac{\text{Average difference}}{\text{Standard Error}} = \frac{1.31}{1.75} = 0.750$$

Student	Exam 1 Score	Exam 2 Score	Difference
Bob	63	69	6
Nina	65	65	0
Tim	56	62	6
Kate	100	91	-9
Alonzo	88	78	-10
Jose	83	87	4
Nikhil	77	79	2
Julia	92	88	-4
Tohru	90	85	-5
Michael	84	92	8
Jean	68	69	1
Indra	74	81	7
Susan	87	84	-3
Allen	64	75	11
Paul	71	84	13
Edwina	88	82	-6



$$\overline{x_d} = 1.31$$

Next, we calculate the standard error for the score difference. The calculation i

$$\text{Standard Error} = \frac{s_d}{\sqrt{n}} = \frac{7.00}{\sqrt{16}} = \frac{7.00}{4} = 1.75$$

In the formula above, n is the number of students – which is the number of diff is s_d.

We now have the pieces for our test statistic. We calculate our test statistic as:

$$t = \frac{\text{Average difference}}{\text{Standard Error}} = \frac{1.31}{1.75} = 0.750$$

T(alpha=0.05, DF=15) = 2.131

- Since 0.75 < 2.131, we cannot reject the H_0

Another Example, try by yourself……

Student	Pre-module Score	Post-module Score	Difference
1	18	22	
2	21	25	
3	16	17	
4	22	24	
5	19	16	
6	24	29	
7	17	20	
8	21	23	
9	23	19	
10	18	20	
11	14	15	
12	16	15	
13	16	18	
14	19	26	
15	18	18	
16	20	24	
17	12	18	
18	22	25	
19	15	19	
20	17	26	

Student	Pre-module Score	Post-module Score	Difference
1	18	22	
2	21	25	
3	16	17	
4	22	24	
5	19	16	
6	24	29	
7	17	20	
8	21	23	
9	23	19	
10	18	20	
11	14	15	
12	16	15	
13	16	18	
14	19	26	
15	18	18	
16	20	24	
17	12	18	
18	22	25	
19	15	19	
20	17	26	

$$\bar{d} = 2.05$$

$$s_d = 2.837$$

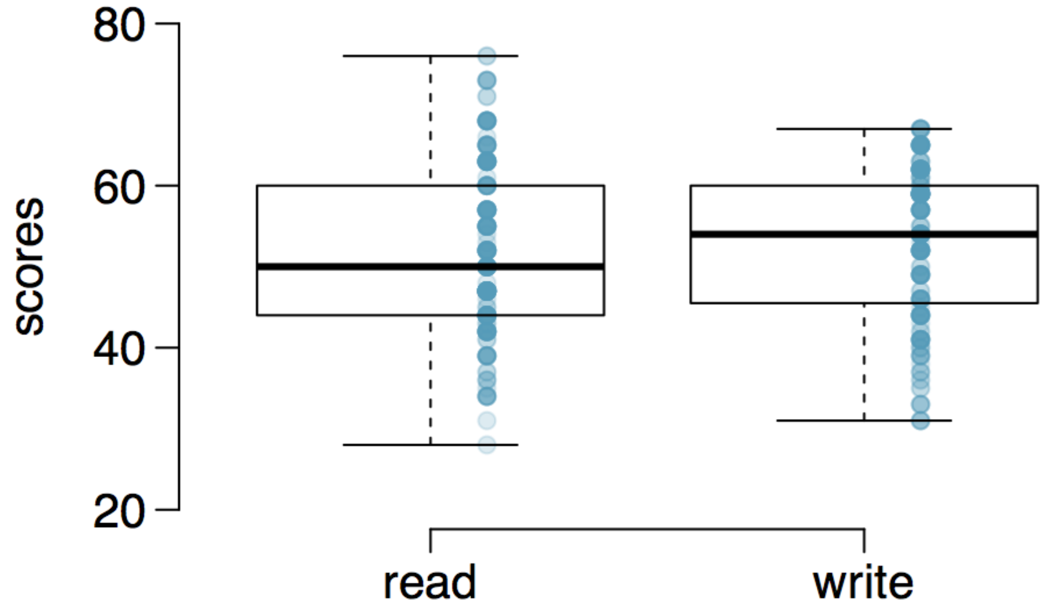
$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.837}{\sqrt{20}} = 0.634$$

Paired Data

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

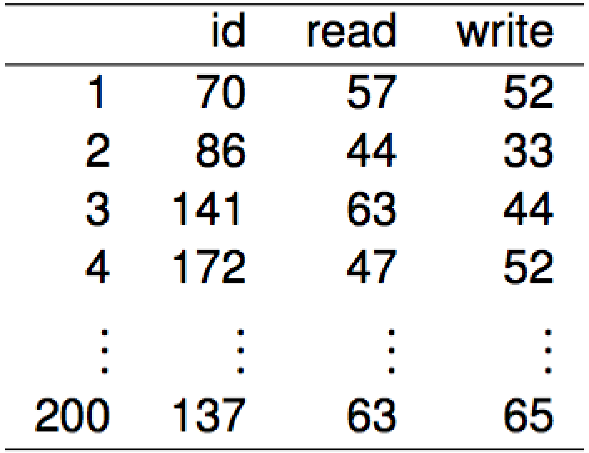
Paired observations

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?



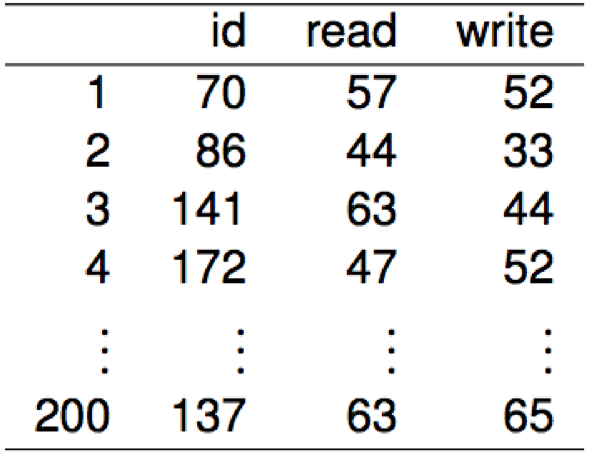
	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

(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

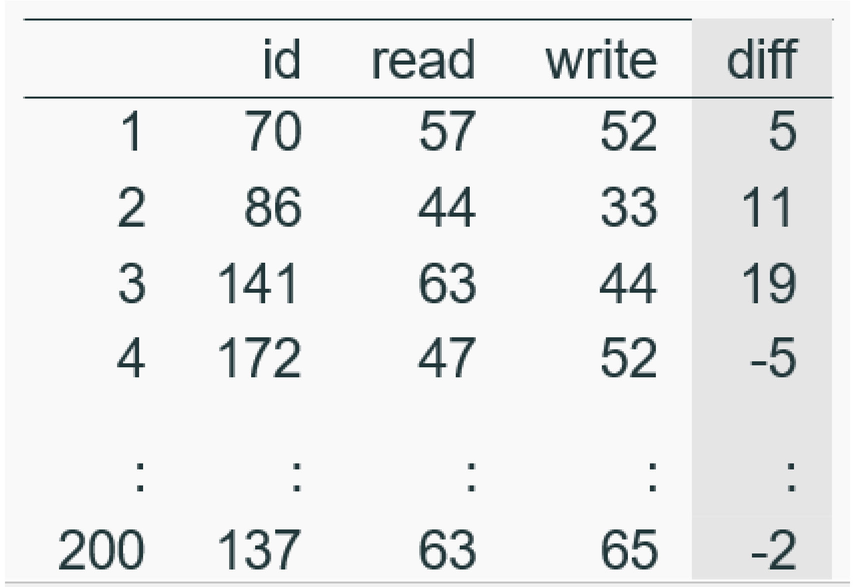
diff = read - write

Analyzing paired data

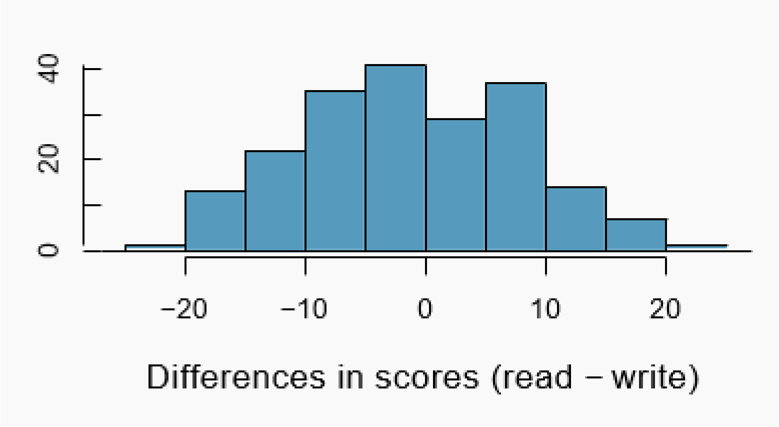
- When two sets of observations have this special correspondence (not independent), they are said to be *paired*
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations

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

$$\mu_{diff}$$

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 scores for reading and writing are equal.

$$\mu_{diff} = 0$$

H_A : Average scores for reading and writing 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*

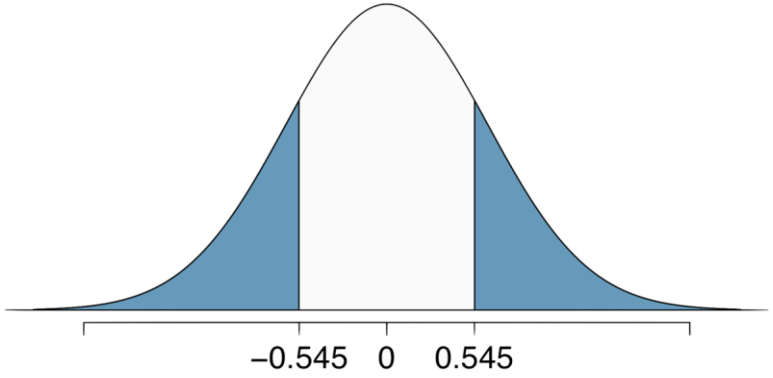
B. The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test

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

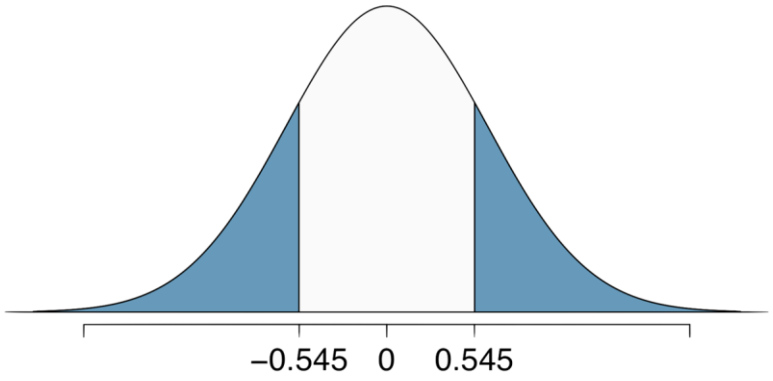
Calculating the test-statistics and the p-value

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 α = 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



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

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

$$df = 200 - 1 = 199$$

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

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

Sources

- openintro.org/os (Chapter 7, Section 7.2)
- https://www.jmp.com/en_us/statistics-knowledge-portal/t-test/t-distribution.html
- https://www.jmp.com/en_nl/statistics-knowledge-portal/t-test/paired-t-test.html
- https://www.statstutor.ac.uk/resources/uploaded/paired-t-test.pdf

Helpful Links:

- https://www.youtube.com/watch?v=JiQR0lHLe74
- https://www.youtube.com/watch?v=Q0V7WpzICI8