# Chapter 7
Inference for numerical data[1]

Department of Mathematics & Statistics
North Carolina A&T State University

---

[1]These notes use content from OpenIntro Statistics Slides by Mine
Cetinkaya-Rundel.

Paired data

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?
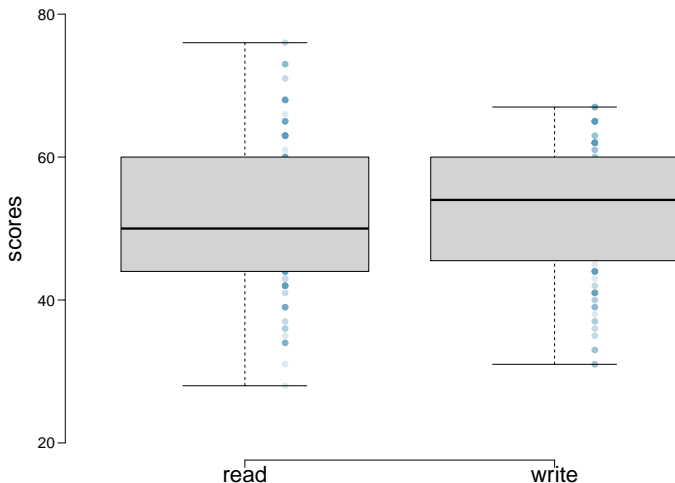
Practice

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

Practice

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

Analyzed paired data

▶ When two sets of observations have this special correspondence (not independent), they are said to be **paired**.

Analyzed 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.

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

Analyzed 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
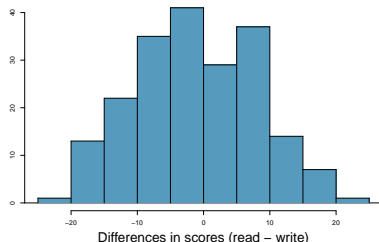
Differences in scores (read – write)

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 : 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 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.
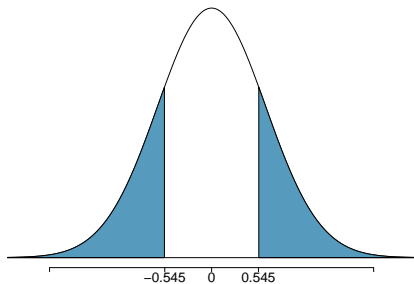
Checking assumptions & conditions

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.
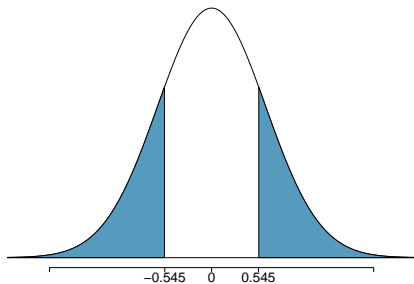
Calculating the test-statistic and the p-value

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 $\alpha = 0.05$.
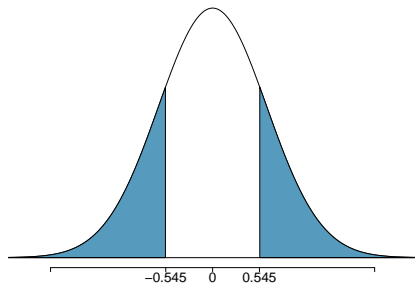
Calculating the test-statistic and the p-value

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 $\alpha = 0.05$.

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

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

$$df = 200 - 1 = 199$$

Calculating the test-statistic and the p-value

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

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 $\alpha = 0.05$.
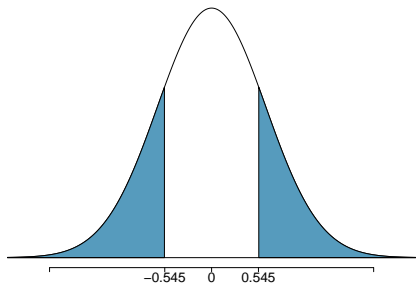

$$T = \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}}$$

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

$$df = 200 - 1 = 199$$

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

Since p-value > 0.05, fail to reject, the data do <u>not</u> 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 it 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 it 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.

Suppose we were to construct 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 ↔ CI

Suppose we were to construct 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.

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

Friday the 13th

Between 1990 - 1992 researchers in the UK collected data on traffic flow, accidents, hospital admissions on Friday 13th and the previous Friday, Friday 6th. Below is an excerpt from this data set on traffic flow. We can assume that traffic flow on given day at locations 1 and 2 are independent.

	type	date	6th	13th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

Friday the 13th

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ One approach is to compare the traffic flow on these two days.

Friday the 13th

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ One approach is to compare the traffic flow on these two days.

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

Friday the 13th

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ One approach is to compare the traffic flow on these two days.

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ H_0 : Average traffic flow on Friday 6th and 13th are equal.
H_A : Average traffic flow on Friday 6th and 13th are different.

Friday the 13th

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ One approach is to compare the traffic flow on these two days.

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ H_0 : Average traffic flow on Friday 6th and 13th are equal.
H_A : Average traffic flow on Friday 6th and 13th are different.

Each case in the data set represents traffic flow recorded at the same location in the same month of the same year: one count from Friday 6th and the other Friday 13th. Are these two counts independent?

Friday the 13th

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ One approach is to compare the traffic flow on these two days.

▶ We want to investigate if people's behavior is different on Friday 13th compared to Friday 6th.

▶ H_0 : Average traffic flow on Friday 6th and 13th are equal.
H_A : Average traffic flow on Friday 6th and 13th are different.

Each case in the data set represents traffic flow recorded at the same location in the same month of the same year: one count from Friday 6th and the other Friday 13th. Are these two counts independent?

No.

Hypotheses

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6th and 13th?

A) $H_0 : \mu_{6th} = \mu_{13th}$
   $H_A : \mu_{6th} \neq \mu_{13th}$

B) $H_0 : p_{6th} = p_{13th}$
   $H_A : p_{6th} \neq p_{13th}$

C) $H_0 : \mu_{diff} = 0$
   $H_A : \mu_{diff} \neq 0$

D) $H_0 : \bar{x}_{diff} = 0$
   $H_A : \bar{x}_{diff} = 0$

Hypotheses

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6th and 13th?

A) $H_0 : \mu_{6th} = \mu_{13th}$
   $H_A : \mu_{6th} \neq \mu_{13th}$

B) $H_0 : p_{6th} = p_{13th}$
   $H_A : p_{6th} \neq p_{13th}$

C) $H_0 : \mu_{diff} = 0$
   $H_A : \mu_{diff} \neq 0$

D) $H_0 : \bar{x}_{diff} = 0$
   $H_A : \bar{x}_{diff} = 0$

Friday the 13th

	type	date	6th	13th	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

$$\bar{x}_{diff} = 1836$$
$$s_{diff} = 1176$$
$$n = 10$$

Finding the test statistic

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

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

Finding the test statistic

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...

$$\text{point estimate} = \bar{x}_{diff} = 1836$$

Finding the test statistic

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...

$$\text{point estimate} = \bar{x}_{diff} = 1836$$

$$SE = \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372$$

Finding the test statistic

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...

$$\text{point estimate} = \bar{x}_{diff} = 1836$$

$$SE = \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372$$

$$T = \frac{1836 - 0}{372} = 4.94$$

Finding the test statistic

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...

$$\text{point estimate} = \bar{x}_{diff} = 1836$$

$$SE = \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372$$

$$T = \frac{1836 - 0}{372} = 4.94$$

$$df = 10 - 1 = 9$$

_____

Note: Null value is 0 because in the null hypothesis we set $\mu_{diff} = 0$.

Finding the p-value

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

Finding the p-value

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

▶ Using a web app:
https://gallery.shinyapps.io/dist_calc/

Finding the p-value

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

▶ Using a web app:
https://gallery.shinyapps.io/dist_calc/

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

Conclusion of the test

What is the conclusion of this hypothese test?

Conclusion of the test

What is the conclusion of this hypothese test?

Since the p-value is quite low, we conclude that the data provide strong evidence of a difference between traffic flow on Friday 6th and 13th.

What is the difference?

▶ We concluded that there is a difference in the traffic flow between Friday 6th and 13th.

What is the difference?

▶ We concluded that there is a difference in the traffic flow between Friday 6th and 13th.

▶ But it would be more interesting to find out what exactly this difference is.

What is the difference?

▶ We concluded that there is a difference in the traffic flow between Friday 6th and 13th.

▶ But it would be more interesting to find out what exactly this difference is.

▶ We can use a confidence interval to estimate this difference.

Confidence interval for a small sample mean

▶ Confidence intervals are always of the form

point estimate $\pm\ ME$

Confidence interval for a small sample mean

▶ Confidence intervals are always of the form

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

▶ ME 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

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

▶ ME 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^*).

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

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the difference between the traffic flow between Friday 6th and 13th?

$$\bar{x}_{diff} = 1836 \quad s_{diff} = 1176 \quad n = 10 \quad SE = 372$$

A) $1836 \pm 1.96 \times 372$

B) $1836 \pm 2.26 \times 372$

C) $1836 \pm -2.26 \times 372$

D) $1836 \pm 2.26 \times 1176$

Constructing a CI for a small sample mean

Which of the following is the correct calculation of a 95% confidence interval for the difference between the traffic flow between Friday 6th and 13th?

$$\bar{x}_{diff} = 1836 \quad s_{diff} = 1176 \quad n = 10 \quad SE = 372$$

A) $1836 \pm 1.96 \times 372$

B) $1836 \pm 2.26 \times 372 \qquad \rightarrow (995, 2677)$

C) $1836 \pm -2.26 \times 372$

D) $1836 \pm 2.26 \times 1176$

Interpreting the CI

Which of the following is the **best** interpretation for the confidence interval we just calculated?

$$\mu_{diff:\ 6th-13th} = (995, 2677)$$

We are 95% confident that...

A) The difference between the average number of cars on the road on Friday 6th and 13th is between 995 and 2677.

B) On Friday 6th there are 995 to 2677 fewer cars on the road than on the Friday 13th, on average.

C) On Friday 6th there are 995 to 2677 more cars on the road than on the Friday 13th, on average.

D) On Friday 13th there are 995 to 2677 fewer cars on the road than on the Friday 6th, on average.

Interpreting the CI

Which of the following is the **best** interpretation for the confidence interval we just calculated?

$$\mu_{diff: 6th-13th} = (995, 2677)$$

We are 95% confident that...

A) The difference between the average number of cars on the road on Friday 6th and 13th is between 995 and 2677.

B) On Friday 6th there are 995 to 2677 fewer cars on the road than on the Friday 13th, on average.

C) On Friday 6th there are 995 to 2677 more cars on the road than on the Friday 13th, on average.

D) On Friday 13th there are 995 to 2677 fewer cars on the road than on the Friday 6th, on average.

Synthesis

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

Synthesis

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

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

Synthesis

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

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

Do you think the findings of this study suggests that people believe Friday 13th is a day of bad luck?

Synthesis

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

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

Do you think the findings of this study suggests that people believe Friday 13th is a day of bad luck?

No, this is an observational study. We have just observed a significant difference between the number of cars on the road on these two days. We have not tested for people's beliefs.