OpenIntro Statistics

CH 05: Inference for Numerical Data

1

One-sample mean with the *t*-distribution

2

Friday the 13th

Between 1990 - 1992 researchers in the UK collected data on traffic flow, accidents, and 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	loc1
2	traffic	1990, July	134012	132908	1104	loc2
3	traffic	1991, September	137055	136018	1037	loc1
4	traffic	1991, September	133732	131843	1889	loc2
5	traffic	1991, December	123552	121641	1911	loc1
6	traffic	1991, December	121139	118723	2416	loc2
7	traffic	1992, March	128293	125532	2761	loc1
8	traffic	1992, March	124631	120249	4382	loc2
9	traffic	1992, November	124609	122770	1839	loc1
10	traffic	1992, November	117584	117263	321	loc2

3

Friday the 13th

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

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

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!

4

Hypotheses

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

A. $H_0 : \boldsymbol{\mu}_{6th} = \boldsymbol{\mu}_{13th}$ $H_A : \boldsymbol{\mu}_{6th} \neq \boldsymbol{\mu}_{13th}$
B. $H_0 : \boldsymbol{p}_{6th} = \boldsymbol{p}_{13th}$ $H_A : \boldsymbol{p}_{6th} \neq \boldsymbol{p}_{13th}$
C. $H_0 : \boldsymbol{\mu}_{diff} = 0$ $H_A : \boldsymbol{\mu}_{diff} \neq 0$
D. $H_0 : \bar{x}_{diff} : 0$ $H_A : \bar{x}_{diff} \neq 0$

5

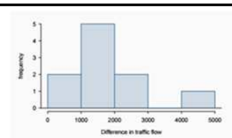
Hypotheses

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

A. $H_0 : \boldsymbol{\mu}_{6th} = \boldsymbol{\mu}_{13th}$ $H_A : \boldsymbol{\mu}_{6th} \neq \boldsymbol{\mu}_{13th}$
B. $H_0 : \boldsymbol{p}_{6th} = \boldsymbol{p}_{13th}$ $H_A : \boldsymbol{p}_{6th} \neq \boldsymbol{p}_{13th}$
C. $H_0 : \boldsymbol{\mu}_{diff} = 0$ $H_A : \boldsymbol{\mu}_{diff} \neq 0$
D. $H_0 : \bar{x}_{diff} : 0$ $H_A : \bar{x}_{diff} \neq 0$

6

Conditions



Independence: We are told to assume that cases (rows) are independent

Sample size / skew:

The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not; probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.

We do not know $\boldsymbol{\sigma}$ and n is too small to assume s is reliable estimate for $\boldsymbol{\sigma}$

So what do we do when the sample size is small?

7

Review: what purpose does a large sample serve?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, $\frac{s}{\sqrt{n}}$, is reliable

8

The normality condition

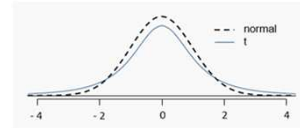
The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal

While this is a helpful special case, it’s inherently difficult to verify normality in small data sets

We should exercise caution when verifying the normality condition for small samples. It is important to not only examine
the data but also think about where the data come from

For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?

The *t* distribution



When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the ***t*** **distribution**.

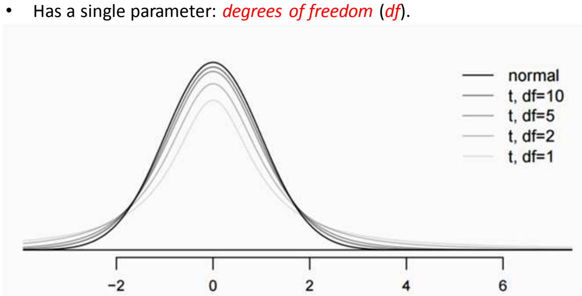
This distribution also has a bell shape, but its tails are thicker than the tail for a normal model

Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution

These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)

The *t* distribution (cont.)

- 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

Back to 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

$\bar{x}_{diff} = 1836$

$s_{diff} = 1176$

Find the test statistic

Test statistic for inference on a small sample mean

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

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

in context...

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

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

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

$$df = 10 - 1 = 9$$

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

13

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the t distribution
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)
[1]  0.0008022394
```

- Or when these aren't available, we can use a t-table
  - Locate the calculated T statistic on the appropriate *df* row, obtain the p-value from the corresponding column heading (one or two tail, depending on the alternative hypothesis).

14

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

15

Confidence interval for a small sample mean

- Confidence intervals are always of the form

$$point\ estimate \pm ME$$

- ME is always calculated as the product of a critical value and SE

$$point\ estimate \pm t^{*}\ x\ SE$$

- Since small sample means follow a t distribution (and not a z distribution), the critical value is a t^* (as opposed to a z^*?).

16

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?

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

A. 1836 ± 1.96 x 372
B. 1836 ± 2.26 x 372
C. 1836 ± -2.26 x 372
D. 1836 ± 2.26 x 1176

17

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?

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

A. 1836 ± 1.96 x 372
B. 1836 ± 2.26 x 372 → (995, 2677)
C. 1836 ± -2.26 x 372
D. 1836 ± 2.26 x 1176

18

Interpreting the CI

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

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

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 2,677
B. on Friday 6th there are 995 to 2,677 fewer cars on the road than on the Friday 13th, on average
C. on Friday 6th there are 995 fewer to 2,677 more cars on the road than on the Friday 13th, on average
D. on Friday 13th there are 995 to 2,677 fewer cars on the road than on the Friday 6th, on average

19

Interpreting the CI

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

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

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 2,677
B. on Friday 6th there are 995 to 2,677 fewer cars on the road than on the Friday 13th, on average
C. on Friday 6th there are 995 fewer to 2,677 more cars on the road than on the Friday 13th, on average
D. on Friday 13th there are 995 to 2,677 fewer cars on the road than on the Friday 6th, on average

20

Synthesis

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

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

21

Recap: Inference using the *t*-distribution

- If $\boldsymbol{\sigma}$ is unknown, use the $\boldsymbol{t}$-distribution with $SE = \frac{s}{\sqrt{n}}$
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, *n* < 10% of population)
  - no extreme skew
- Hypothesis Testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = n-1$$

- Confidence interval: $point\ estimate \pm t^{*}_{df} \times SE$

Note: The example we used was for paired means (difference between dependent groups). We took the difference between the observations and used only these differences (one sample) in our analysis

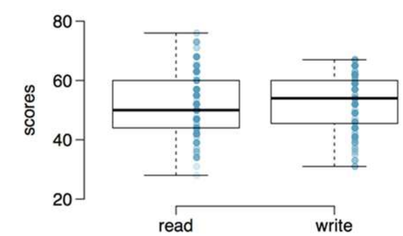
22

Paired Data

23

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?



24

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

25

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

26

Analyzing paired data

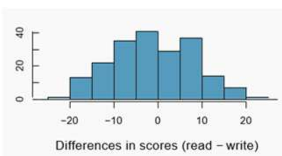
When two sets of observations have this special correspondence (not independent), they are said to be *paired*

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



27

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

$$\bar{x}_{diff}$$

28

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 traffic flow on Friday 6th and 13th are equal.

$$\mu_{diff} = 0$$

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

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

29

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

30

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

31

Checking assumptions & conditions

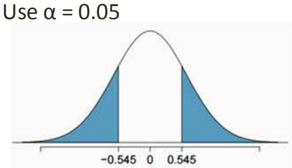
Which of the following is true?

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

32

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 $\alpha = 0.05$



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

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

$$df = 200 - 1 = 199$$

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

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

33

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

34

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

35

Difference in two means

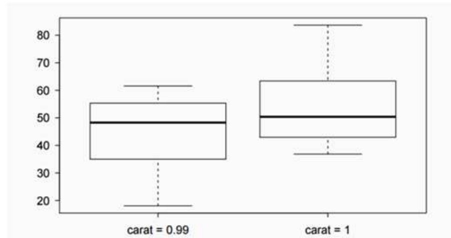
36

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

37

Data



	0.99 carat pt99	*1 carat* 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.

38

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

$$\bar{x}_{pt99} - \bar{x}_{pt100}$$

39

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. $H_0: \mu_{pt99} = \mu_{pt100}$ $\quad H_A: \mu_{pt99} \neq \mu_{pt100}$

B. $H_0: \mu_{pt99} = \mu_{pt100}$ $\quad H_A: \mu_{pt99} > \mu_{pt100}$

C. $H_0: \mu_{pt99} = \mu_{pt100}$ $\quad H_A: \mu_{pt99} < \mu_{pt100}$

D. $H_0: \bar{x}_{pt99} = \bar{x}_{pt100}$ $\quad H_A: \bar{x}_{pt99} < \bar{x}_{pt100}$

40

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

41

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

42

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

43

Test statistics

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.

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

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

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

44

Test statistics (cont.)

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

in context...

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

45

Test statistics (cont.)

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

A. 22
B. 23
C. 30
D. 29
E. 52

46

Test statistics (cont.)

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

A. 22
B. 23
C. 30
D. 29
E. 52

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

47

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. 0.01
C. 0.02
D. between 0.01 and 0.02

48

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. 0.01
C. 0.02
D. between 0.01 and 0.02

49

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

50

Recap: Inference for difference of two 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:
  $$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = \min(n_1 - 1, n_2 - 1)$$
- Confidence interval: $point\ estimate \pm t^{\star}_{df} \times SE$

51

Computing the power for a 2-sample test

Comparing means with ANOVA

52

We are skipping these sections.

53