Chapter 7: Inference for numerical data

OpenIntro Statistics, 4th Edition

Slides developed by Mine Çetinkaya-Rundel of OpenIntro.

One-sample means with the t distribution

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	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.
- 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

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) $H_0 : \mu_{diff} = 0$

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

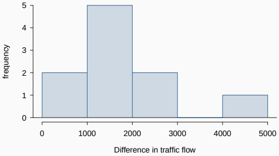
(d) $H_0 : \bar{x}_{diff} = 0$

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

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 σ and n is too small to assume s is a reliable estimate for σ.



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

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...

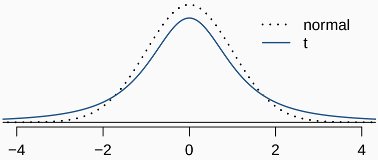
- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, holds 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?

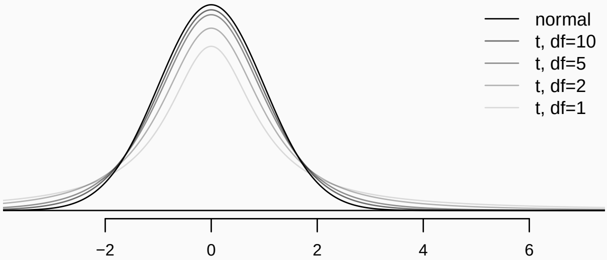
The t distribution

- 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 normal model's.
- 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 shape of the t distribution as df increases?

Approaches normal.

Back to Friday the 13th

	type	date	6^{th}	13^{th}	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

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

$$s_{diff} = 1176$$

Finding 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 $df = n - 1$.

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

in context...

$$\begin{aligned} point\ estimate &= \bar{x}_{diff} = 1836 \\ SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \\ T &= \frac{1836 - 0}{372} = 4.94 \\ df &= 10 - 1 = 9 \end{aligned}$$

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.
- Using R:

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

[1] 0.0008022394
```

- 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 hypothesis 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 6^{th} and 13^{th}.
- 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

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

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

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

Finding the critical t ($t^\star$)

Using R:

```
> qt(p = 0.975, df = 9)

[1] 2.262157
```

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?

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

(a) $1836 \pm 1.96 \times 372$

(b) $1836 \pm 2.26 \times 372$ $\rightarrow$ *(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 6^{th} and 13^{th} is between 995 and 2,677.

(b) on Friday 6^{th} there are 995 to 2,677 fewer cars on the road than on the Friday 13^{th}, on average.

(c) on Friday 6^{th} there are 995 fewer to 2,677 more cars on the road than on the Friday 13^{th}, on average.

(d) *on Friday 13^{th} there are 995 to 2,677 fewer cars on the road than on the Friday 6^{th}, on average.*

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.

Recap: Inference using the t-distribution

- If σ is unknown, use the 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:

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

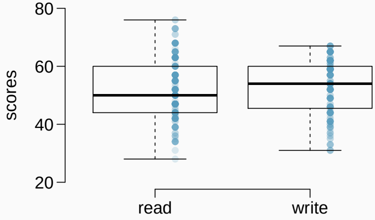
- Confidence interval:

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

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, therefore the mechanics are

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?



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*

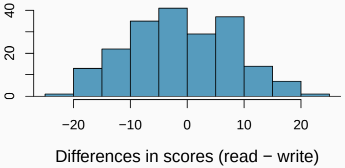
Analyzing paired data

- 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.

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

- 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.

$$\mu_{diff}$$

- *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: 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.*

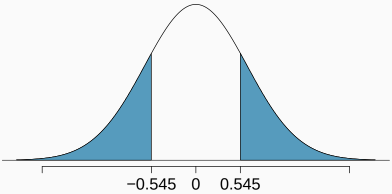
(b) The distribution of differences is bimodal, therefore we cannot continue with the hypothesis test.

(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.

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 $\alpha = 0.05$.



$$
\begin{aligned}
T &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} \\
&= \frac{-0.545}{0.628} = -0.87 \\
df &= 200 - 1 = 199 \\
p - value &= 0.1927 \times 2 = 0.3854
\end{aligned}
$$

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.

HT ↔ 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

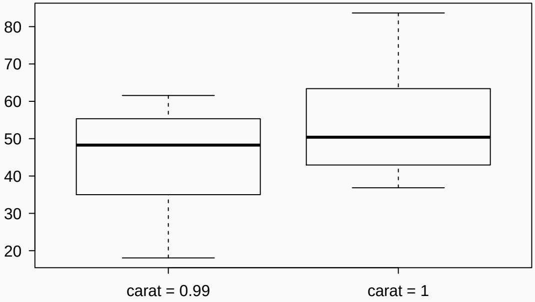
$$
\begin{aligned}
-0.545 \pm 1.97 \frac{8.887}{\sqrt{200}} &= -0.545 \pm 1.97 \times 0.628 \\
&= -0.545 \pm 1.24 \\
&= (-1.785, 0.695)
\end{aligned}
$$

Difference of 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

These data are a random sample from the `diamonds` data set in `ggplot2` R package.

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.

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

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.*

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	*1 carat*
	pt99	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

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

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

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

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.*

Equivalent confidence level

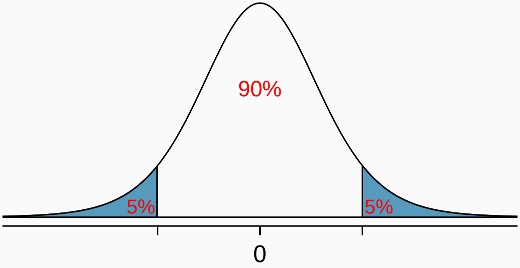
What is the equivalent confidence level for a one-sided hypothesis test at $\alpha = 0.05$?

(a) *90%*

(b) 92.5%

(c) 95%

(d) 97.5%



Critical value

What is the appropriate $t^\star$ for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

(a) 1.32

(b) *1.72*

(c) 2.07

(d) 2.82

```
> qt(p = 0.95, df = 22)

[1] 1.717144
```

Confidence interval

Calculate the interval, and interpret it in context.

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

$$\begin{aligned}
(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^\star_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\
&= -8.93 \pm 6.12 \\
&= (-15.05, -2.81)
\end{aligned}$$

We are 90% confident that the average point price of a 0.99 carat diamond is \$15.05 to \$2.81 lower than the average point price of a 1 carat diamond.

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a t-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.
- 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

Recap: Inference using difference of two small sample means

- If σ_1 or σ_2 is unknown, difference between the sample means follow a t-distribution with $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_1}}$.
- 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:

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