Statistics and Experimental Methods

I

Autumn 2018

Lesson 10, 26.11.2018

Martin Heide Jørgensen
Associate Professor, I4.0 programme coordinator
University of Southern Denmark

Anooshmita Das
Ph.D student, Maersk Mc/Kinney Møller Inst.
University of Southern Denmark

Wrap Up from lesson 9

- How is the nature of the null hypothesis (H0) – and the alternative hypothesis (HA)
- How do you explain that the outcome from a test for the mean value from a sample is covered by the 95% Confidence interval.
- *What is the information from a p-value as a part of a "hypothesis test"*
- *What is the information in the "z" value in relation to the "hypothesis test"*
- *What is the definition of a "type 1 error" and a "type 2 error"*

Overview

Paired data

Difference of two means

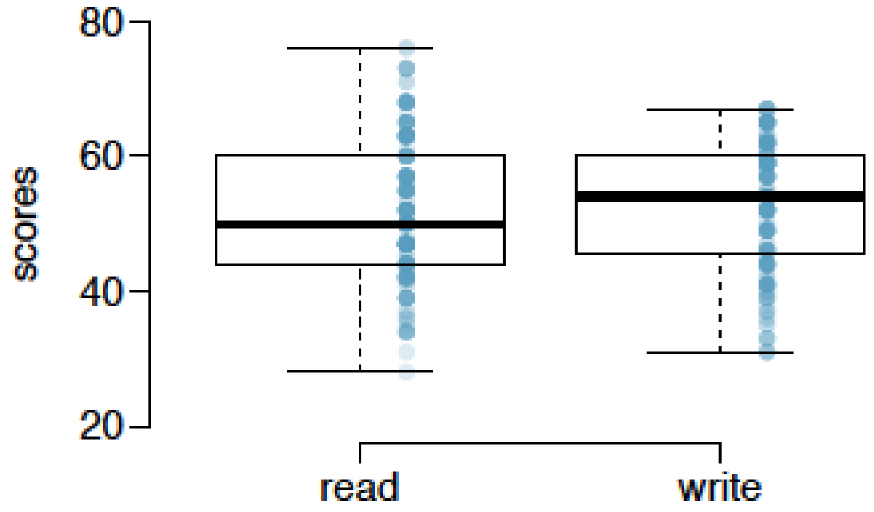
One-sample means with the t distribution

The t distribution for the difference of two means

Comparing means with ANOVA

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

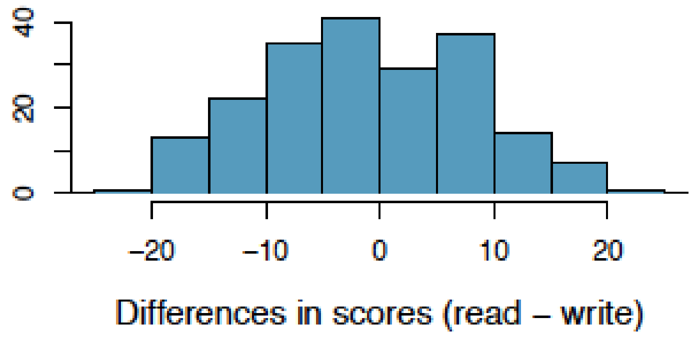
Two sets of observations have special correspondence (not independent): paired.

Observe difference in outcomes of each pair of observations.

diff = read - write

Consistent order of subtraction.

	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 hypothesis

If in fact there was no difference between the scores on the reading and writing exams.

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: no different from what we have done before.

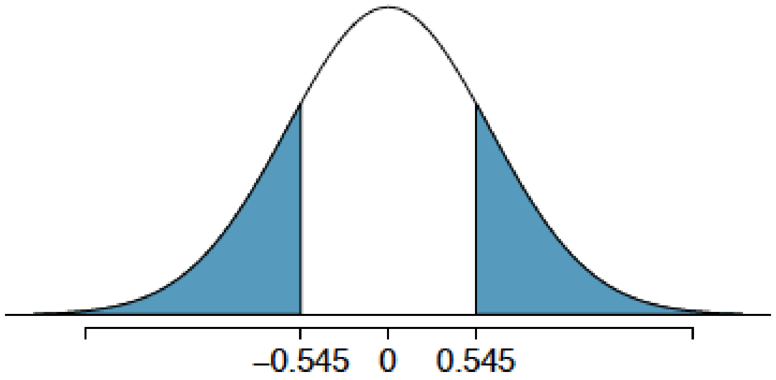
Data from one sample: differences.

Test to see if the average difference is different than 0.

What about conditions?

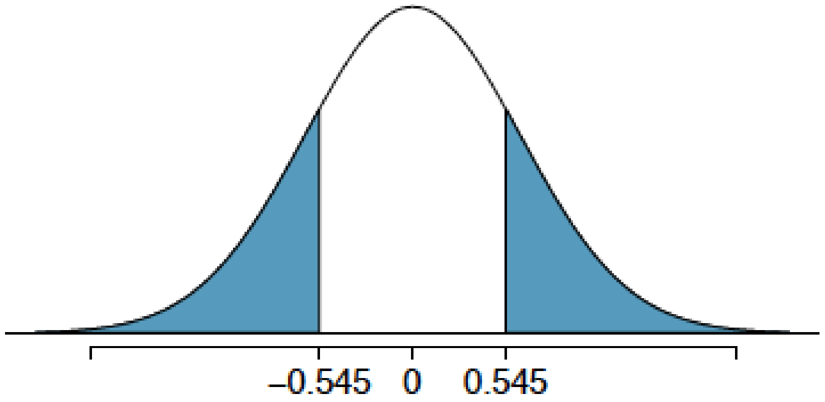
Calculating the test-statistic and the p-value

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 α = 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 α = 0.05.



$$
\begin{aligned}
Z &= \frac{-0.545 - 0}{\frac{8.887}{\sqrt{200}}} \\
&= \frac{-0.545}{0.628} = -0.87 \\
p - value &= 0.1949 \times 2 = 0.3898
\end{aligned}
$$

Since p-value > 0.05, fail to reject

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.

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

If 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

HT <–> CI

If 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

$$
\begin{aligned}
-0.545 \pm 1.96\frac{8.887}{\sqrt{200}} &= -0.545 \pm 1.96 \times 0.628 \\
&= -0.545 \pm 1.23 \\
&= (-1.775, 0.685)
\end{aligned}
$$

Difference of two means

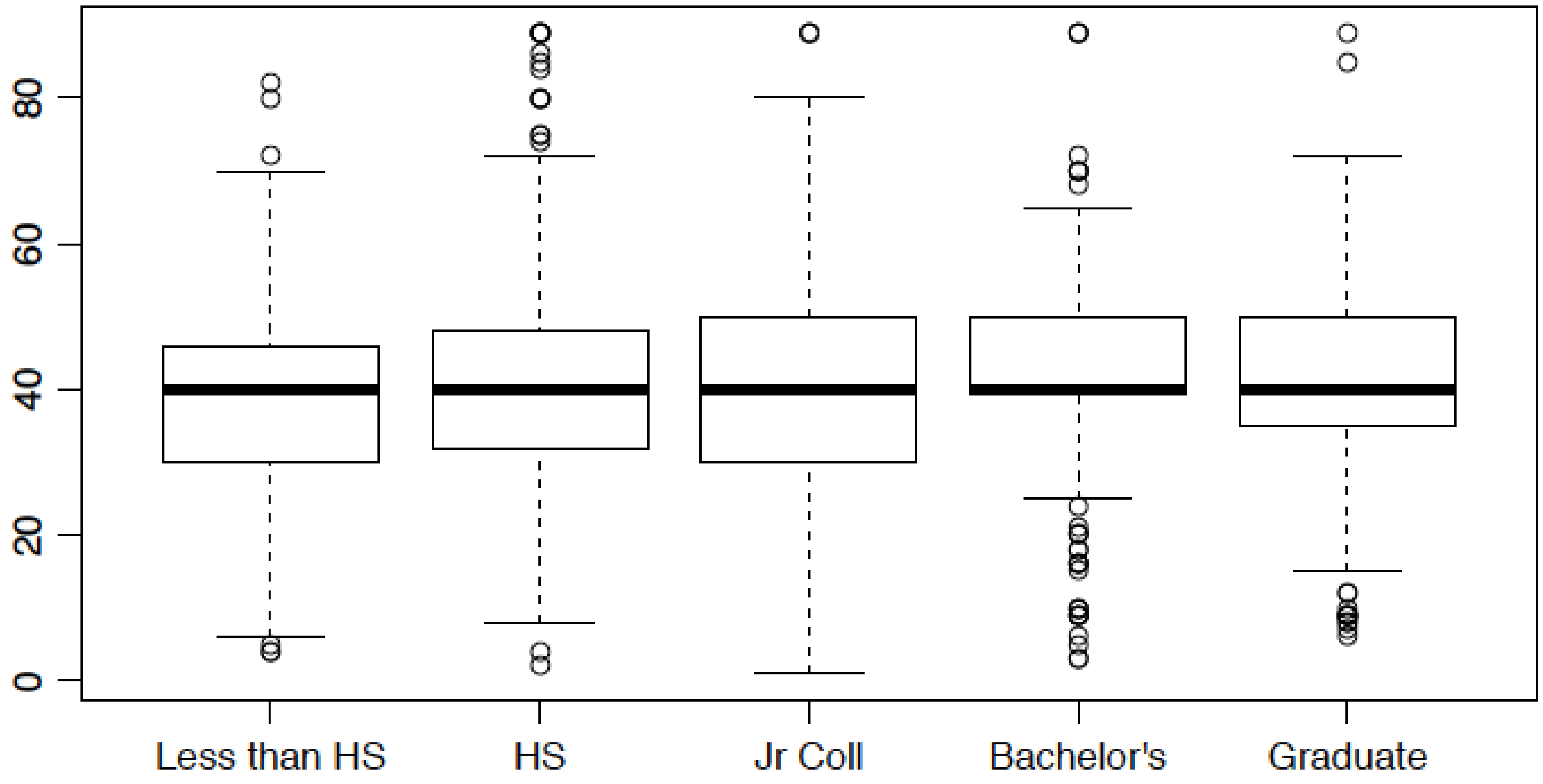
Difference of two means

The General Social Survey (GSS) conducted by the Census Bureau contains a standard ‘core’ of demographic, behavioral, and attitudinal questions, plus topics of special interest. Many of the core questions have remained unchanged since 1972 to facilitate time-trend studies as well as replication of earlier findings. Below is an excerpt from the 2010 data set. The variables are number of hours worked per week and highest educational attainment.

	degree	hrs1
1	BACHELOR	55
2	BACHELOR	45
3	JUNIOR COLLEGE	45
⋮		
1172	HIGH SCHOOL	40

Exploratory analysis

What can you say about the relationship between educational attainment and hours worked per week?



Collapsing levels into two

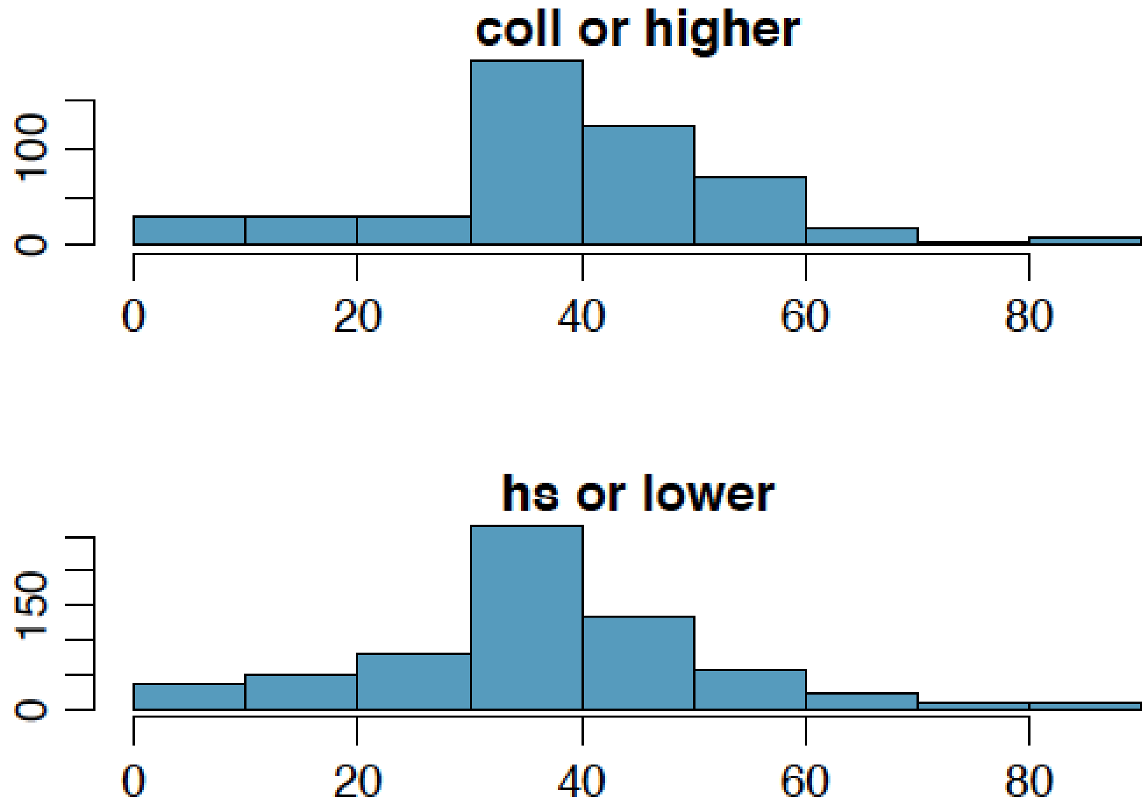
We need the difference between college and non-college graduates.

Combine the levels of education:

- `hs or lower` 🡐 less than high school or high school
- `coll or higher` 🡐 junior college, bachelor's, and graduate

Exploratory analysis - another look

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667



Parameter and point estimate

We want to construct a 95% confidence interval for the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower. What are the parameter of interest and the point estimate?

Parameter and point estimate

We want to construct a 95% confidence interval for the average difference between the number of hours worked per week by Americans with a college degree and those with a high school degree or lower. What are the parameter of interest and the point estimate?

Parameter of interest: Average difference in the number of hours worked per week by all Americans with a college degree and those with a high school degree or lower.

$$\mu_{coll} - \mu_{hs}$$

Point estimate: Average difference in the number of hours worked per week by sampled Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} - \bar{x}_{hs}$$

+ Checking assumptions & conditions

1. Independence within groups:

Both the college graduates and those with HS degree or lower are sampled randomly.

505 < 10% of all college graduates and 667 < 10% of all students with a high school degree or lower.

2. Independence between groups: 🡐 new!

Since the sample is random, the college graduates in the sample are independent of those with a HS degree or lower.

3. Sample size / skew:

Both distributions look reasonably symmetric, and the sample sizes are at least 30.

Hence, sampling distributions are nearly normal.

Confidence interval for difference between two means

Same old form:

point estimate ± *critical value × SE of point estimate*

Point estimate is $\bar{x}_1 - \bar{x}_2$

Critical value is z*

Standard error of the difference between two sample means

$$SE_{(\bar{x}_1 - \bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

In context…

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

In context…

Calculate the standard error of the average difference between the number of hours worked per week by college graduates and those with a HS degree or lower.

	$\bar{x}$	s	n
coll or higher	41.8	15.14	505
hs or lower	39.4	15.12	667

$$
\begin{aligned}
SE_{(\bar{x}_{coll}-\bar{x}_{hs})} &= \sqrt{\frac{s_{coll}^2}{n_{coll}} + \frac{s_{hs}^2}{n_{hs}}} \\
&= \sqrt{\frac{15.14^2}{505} + \frac{15.12^2}{667}} \\
&= 0.89
\end{aligned}
$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$

Confidence interval for the difference (cont.)

Estimate (using a 95% confidence interval) the average difference between Americans with a college degree and those with a high school degree or lower.

$$\bar{x}_{coll} = 41.8 \qquad \bar{x}_{hs} = 39.4 \qquad SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$

$$\begin{aligned}
(\bar{x}_{coll} - \bar{x}_{hs}) \pm z^{\star} \times SE_{(\bar{x}_{coll} - \bar{x}_{hs})} &= (41.8 - 39.4) \pm 1.96 \times 0.89 \\
&= 2.4 \pm 1.74 \\
&= (0.66, 4.14)
\end{aligned}$$

Setting the hypotheses

What are the hypotheses for testing if there is a difference between college graduates and those with a HS degree or lower?

Setting the hypotheses

What are the hypotheses for testing if there is a difference between college graduates and those with a HS degree or lower?

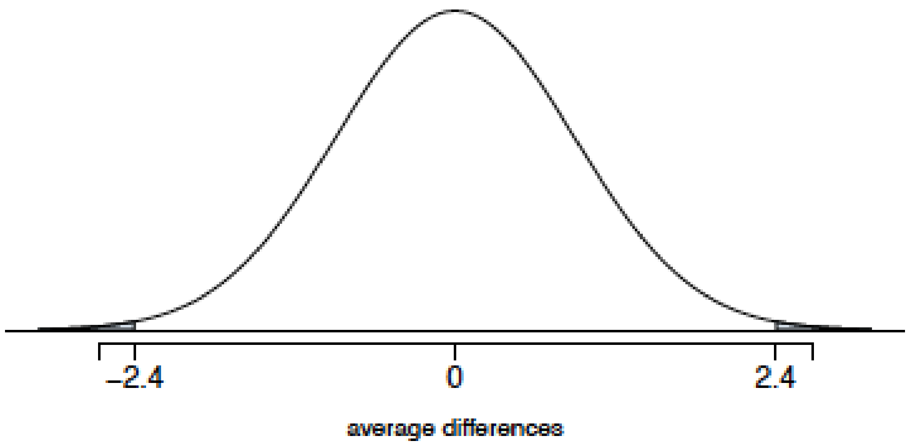
$H_0: \mu_{coll} = \mu_{hs}$

$H_A: \mu_{coll} \neq \mu_{hs}$

Calculating the test-statistic and the p-value

$$H_0: \mu_{coll} = \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} = 0$$
$$H_A: \mu_{coll} \neq \mu_{hs} \rightarrow \mu_{coll} - \mu_{hs} \neq 0$$

$$\bar{x}_{coll} - \bar{x}_{hs} = 2.4,\ SE_{(\bar{x}_{coll} - \bar{x}_{hs})} = 0.89$$



$$\begin{aligned} Z &= \frac{(\bar{x}_{coll} - \bar{x}_{hs}) - 0}{SE_{(\bar{x}_{coll} - \bar{x}_{hs})}} \\ &= \frac{2.4}{0.89} = 2.70 \\ upper\ tail &= 1 - 0.9965 = 0.0035 \\ p - value &= 2 \times 0.0035 = 0.007 \end{aligned}$$

Conclusion of the test

Which of the following is correct based on the results of the hypothesis test we just conducted?

(a) There is a 0.7% chance that there is no difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(b) Since the p-value is low, we reject H_0. The data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(c) Since we rejected H_0, we may have made a Type 2 error.

(d) Since the p-value is low, we fail to reject H_0. The data do not provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

Conclusion of the test

Which of the following is correct based on the results of the hypothesis test we just conducted?

(a) There is a 0.7% chance that there is no difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(b) Since the p-value is low, we reject H_0. The data provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

(c) Since we rejected H_0, we may have made a Type 2 error.

(d) Since the p-value is low, we fail to reject H_0. The data do not provide convincing evidence of a difference between the average number of hours worked per week by college graduates and those with a HS degree or lower.

One-sample means with the *t* distribution
example - 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. Assume that traffic flow on given day at locations 1 and 2 are independent.

	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

Scanlon, T.J., Luben, R.N., Scanlon, F.L., Singleton, N. (1993), "Is Friday the 13th Bad For Your Health?," BMJ, 307, 1584-1586.

Friday the 13th

Is people's behavior different on Friday 13th compared to Friday 6th?

One approach: compare the traffic flow on these two days.

H0 : Average traffic flow on Friday 6th and 13th are equal.

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

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

How do you specify hypotheses?

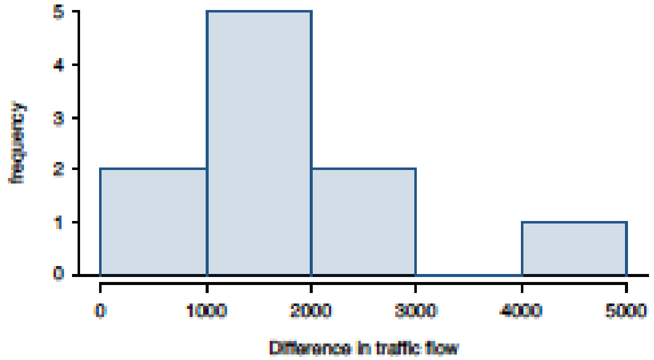
Conditions

Independence: Stated

Sample size / skew: seems equally likely to have days with lower than average traffic and higher than average traffic

n < 30!

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



Review: why large sample?

If observations are independent, and the population distribution is not extremely skewed…

the sampling distribution of the mean is nearly normal

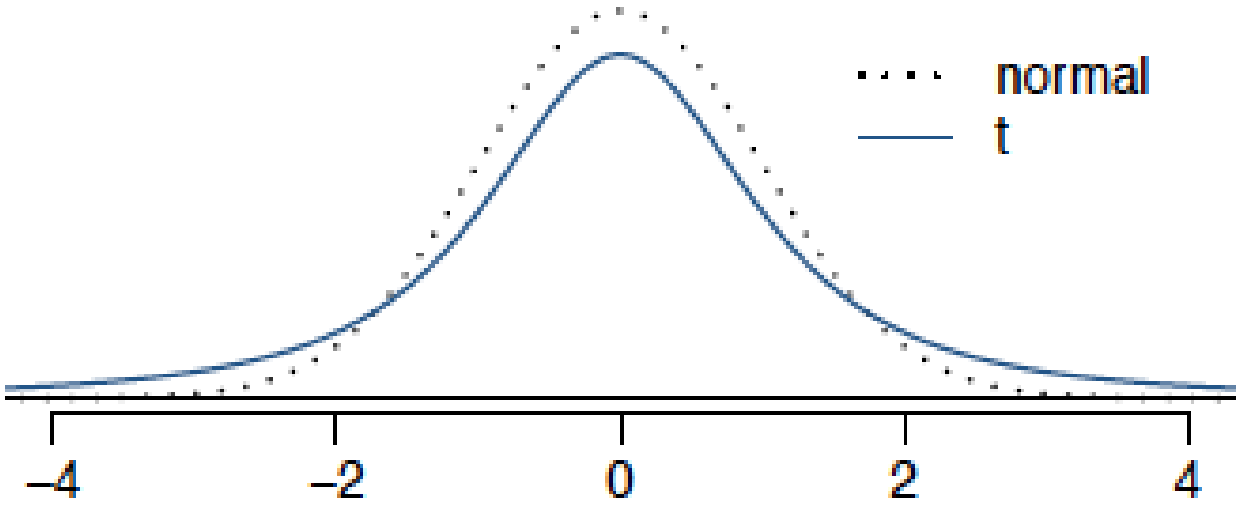
the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is more reliable

The *t* distribution

When working with small samples, and the standard deviation is unknown (almost always):

t distribution

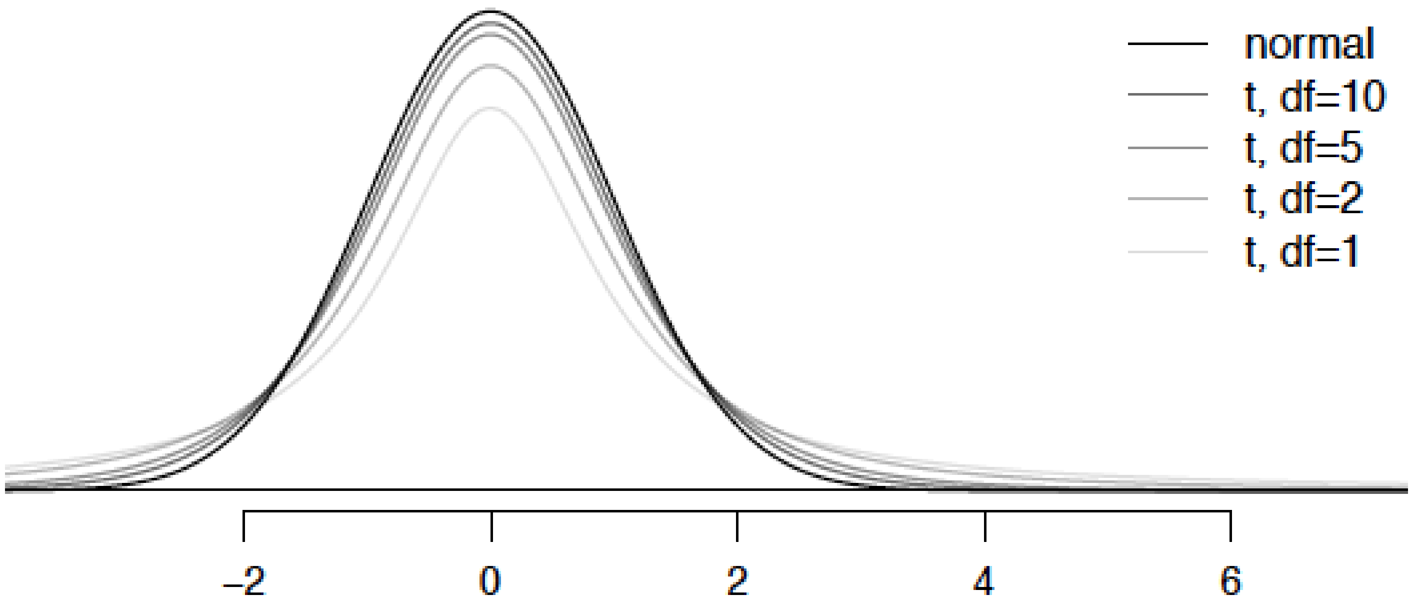
- also bell shaped, thicker tails
- observations - more likely to fall beyond two SDs



The *t* distribution (cont.)

Always centered at zero

Single parameter: degrees of freedom (df).



What happens to shape of the t distribution as *df* increases?

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

$\downarrow$

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

$$s_{diff} = 1176$$

$$n = 10$$

Finding the test statistic

Test statistic for 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…

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

Finding the p-value

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 applet:

http://www.socr.ucla.edu/htmls/SOCR_Distributions.html

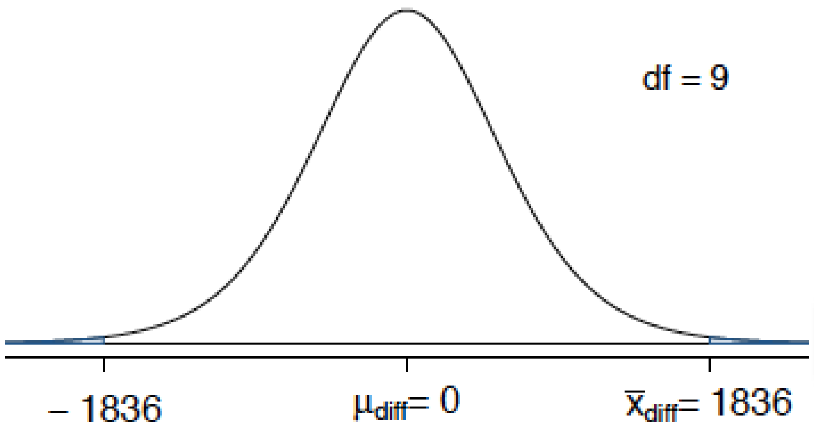
Or when these aren't available, we can use a *t* table.

Finding the p-value

one tail / two tails	0.100 / 0.200	0.050 / 0.100	0.025 / 0.050	0.010 / 0.020	0.005 / 0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
⋮	⋮	⋮	⋮	⋮	
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
⋮	⋮	⋮	⋮	⋮	
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.64	1.96	2.33	2.58

Finding the p-value (cont.)

one tail	0.100	0.050	0.025	0.010	0.005	
two tails	0.200	0.100	0.050	0.020	*0.010*	→
df 6	1.44	1.94	2.45	3.14	3.71	
7	1.41	1.89	2.36	3.00	3.50	
8	1.40	1.86	2.31	2.90	3.36	
9	1.38	1.83	2.26	2.82	*3.25*	→
10	1.37	1.81	2.23	2.76	3.17	



$T = 4.94$

What is the conclusion of the hypothesis test?

The data provide convincing

traffic flow on Friday 6th and 13th.

What is the difference?

There is a difference in the traffic flow between Friday 6^{th} and 13^{th}.

What exactly this difference is?

CI for a mean of small sample

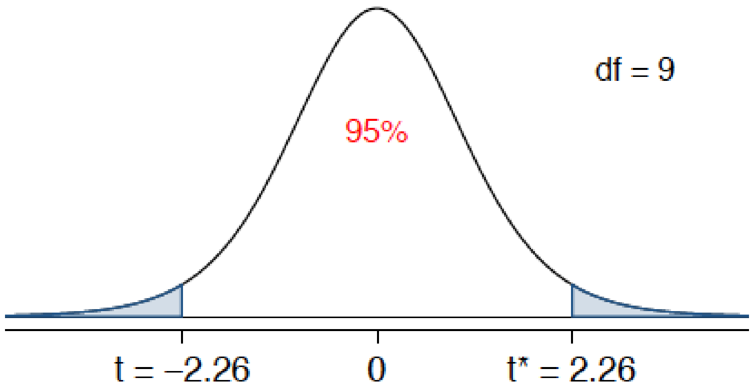
As usual:

point estimate ± *critical value* * *SE*

As we have *t* distribution, the critical value is a t^*

point estimate ± $t^* \times SE$

Finding the critical t (t^*)



$n = 10$, $df = 10 - 1 = 9$, $t^\star$ is at the intersection of row $df = 9$ and two tail probability 0.05.

one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17

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?

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

(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?

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

(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) $1836 \pm 2.26 \times 1176$

Synthesis

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

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

Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a 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)
  - $n < 30$ and 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 the same as when we are working with just one sample.

Next class

Inference for numerical data, continued