Unit 3: Inference for Categorical and Numerical Data

3. One Sample Means using the *t*-distribution

(Chapter 4.1)

11/1/2017

Recap from last time

1. You can use the Normal approximation for the difference of two proportions
2. The margin of error is not just the sum of the margin of errors for each proportion
3. If you think two proportions come from the same population, you can use a pooled estimate

Key ideas

1. When our samples are too small, we shouldn't use the Normal distribution
2. We should use the *t* distribution to make up for uncertainty in our sample statistics
3. All of our statistical theory still holds, we are just plugging in different distributions

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. Let's assume that 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,et al. (1993) BMJ, 307, 1584-1586.

How to test for the effect of 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 represents traffic flow 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

Practice Question 1: Setting up the 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) $H_0: \bar{x}_{diff} = 0$
$H_A: \bar{x}_{diff} \neq 0$

Checking conditions

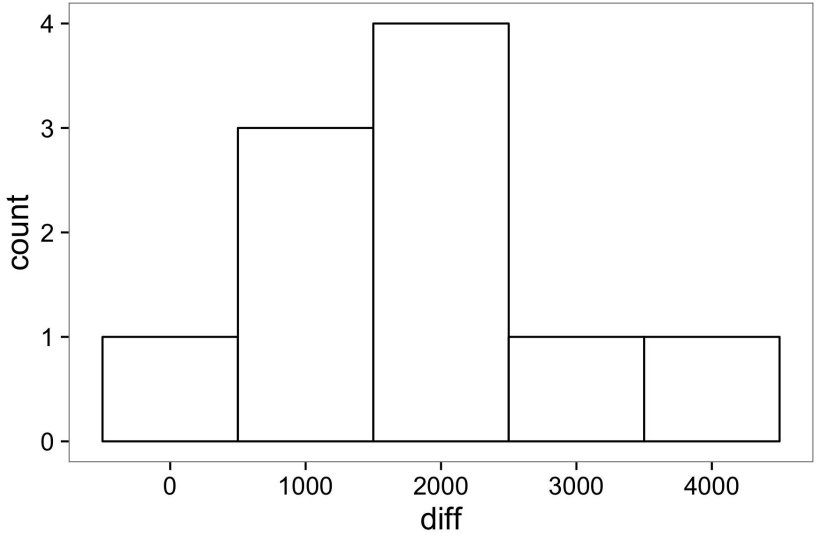
Independence

We're told to assume that cases (rows) are independent

Sample size / skew

Distribution doesn't look very skewed, but hard to assess with small sample. Worth thinking about whether we *expect* it to be skewed. Probably not?

But n < 30! What should we do?



Review: Why do we want a large sample?

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

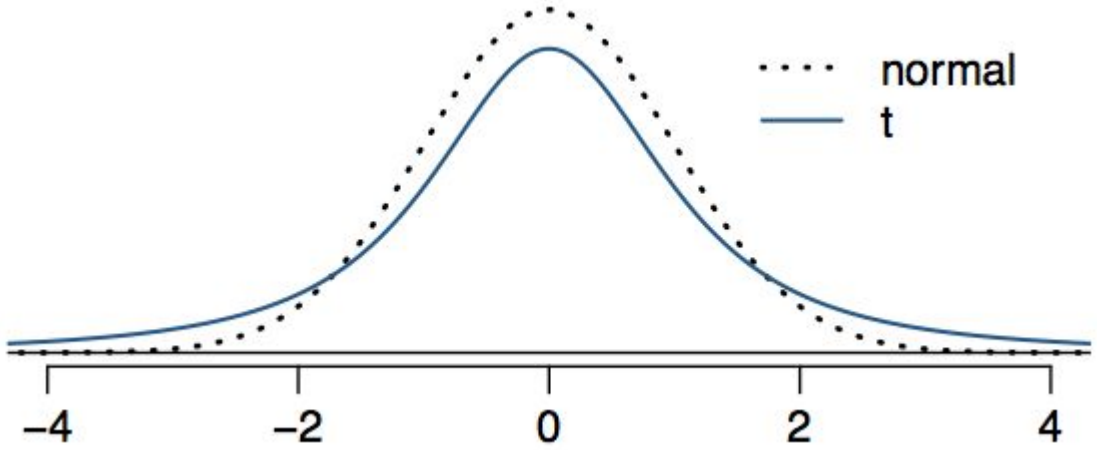
- the sampling distribution of the mean is nearly normal
- $\frac{s}{\sqrt{n}}$ is a reliable estimate of the standard error

What about small samples?

The *t* distribution

When working with small samples, and the population standard deviation is unknown, we hedge for the uncertainty of the standard error estimate by using a new distribution: the ***t*-distribution**.

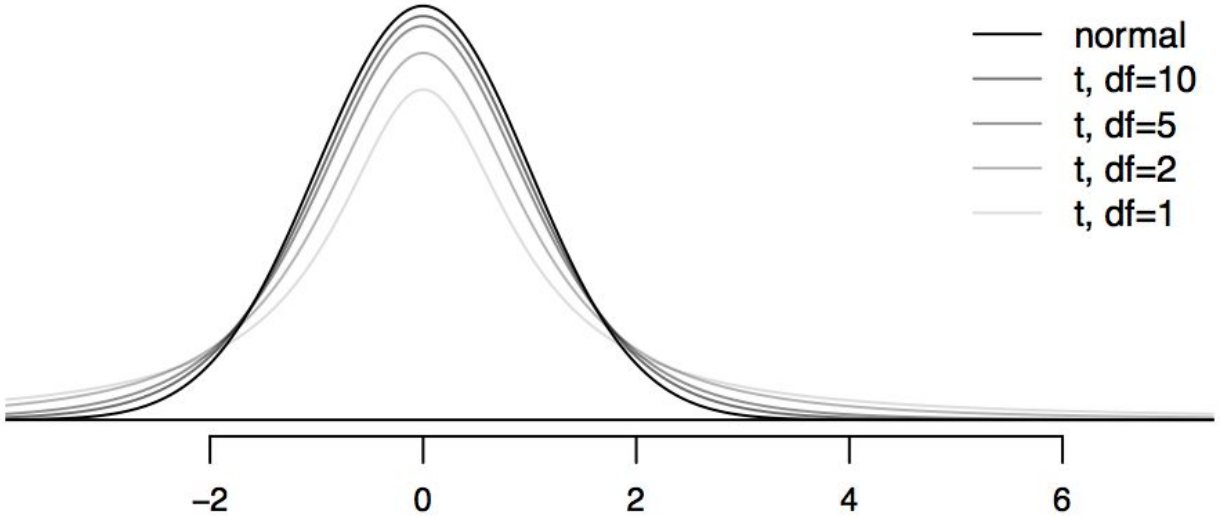
t has a similar shape, but thicker tails (i.e. extreme values are more likely).



The many different *t*s

Centered at zero like the standard Normal (*z*-distribution).

Has only one parameter: **degrees of freedom (df)**



What happens as df increases? **Approaches the Normal (z)**

Testing the effect of 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

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

$s_{diff} = 1176$

$n = 10$

Let's compute the test statistic

The test statistic for inference of the mean is the *T*-statistic with *df=n-1*.

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

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

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

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

$$df = 10 - 1 = 9$$

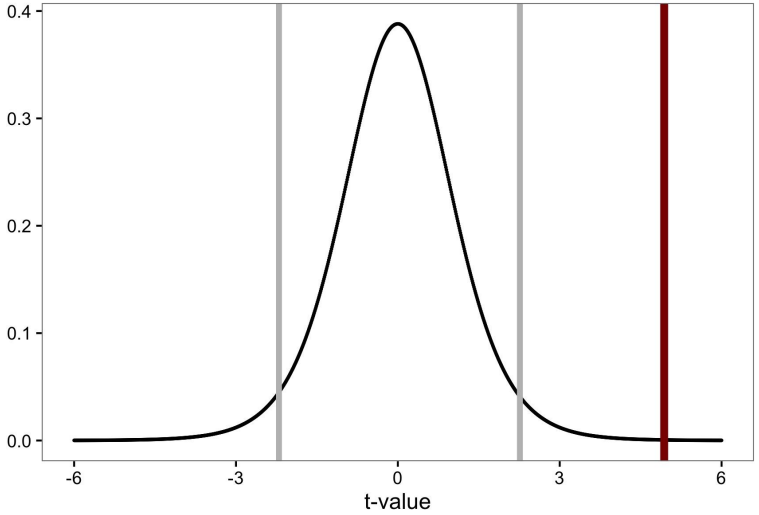
Finding the p-value

As always, the p-value is probability of getting a value *at least* this extreme given our null distribution.

So for *t*(9), Using R:

```
> 2 * pt(4.94, df = 9,
       lower.tail = FALSE)
[1] 0.000802239
```

Something is freaky about Friday the 13th!



Why 2 times? **We want to consider extreme data in the other tail as well**

Confidence intervals for the t-distribution

Confidence intervals are always of the form

point estimate ± Margin of Error

and Margin of error is always

critical value * SE

But since small sample means follow a t-distribution (and not a z distribution), the critical value is a t*.

point estimate ± t* x SE

Practice Question 2: Confidence interval for Friday the 13th.

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?

t*: `qt(p = .975, df = 9)`

`2.262157`

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

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

(b) $1836 \pm 2.26 \times 372$

(c) $1836 \pm 2.26 \times 1176$

Practice Question 2: Confidence interval for Friday the 13th.

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?

t*: `qt(p = .975, df = 9)`

`2.262157`

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

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

(b) $1836 \pm 2.26 \times 372$ → (995, 2677)

(c) $1836 \pm 2.26 \times 1176$

Practice Question 3: Interpreting the CI

Which of the following is the best interpretation for the confidence interval we just calculated? $\mu_{\text{diff: 6th - 13th}} = (995, 2677)$

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 2,677.

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

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

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

Practice Question 3: Interpreting the CI

Which of the following is the best interpretation for the confidence interval we just calculated? $\mu_{\text{diff: 6th - 13th}} = (995, 2677)$

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 2,677.

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

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

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

Key ideas

1. When our samples are too small, we shouldn't use the Normal distribution
2. We should use the *t* distribution to make up for uncertainty in our sample statistics
3. All of our statistical theory still holds, we are just plugging in different distributions