Unit 3: Inference for Categorical and Numerical Data

3. The *t*-distribution

(Chapter 4.1-4.2)

11/7/2018

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. We use the *t* distribution to make up for uncertainty in our sample statistics
1. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)
3. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values

Enrollment in PSYC201

PSYC201 has an enrollment cap of 52 students a year.
But sometimes we go over, and sometimes we go under.

year	enrollment
2010	30
2011	59
2012	44
2013	55
2014	46
2015	25
2016	43
2017	46
2018	44

Does it look like the average number of enrollees is 52?

How to test average enrollment

We want to investigate if the number of people who enroll in PSYC201 is more or less than 52 on average.

H_0: The number of people who enroll on average is 52

H_A: The number of people who enroll on average is *different* from 52

Checking conditions

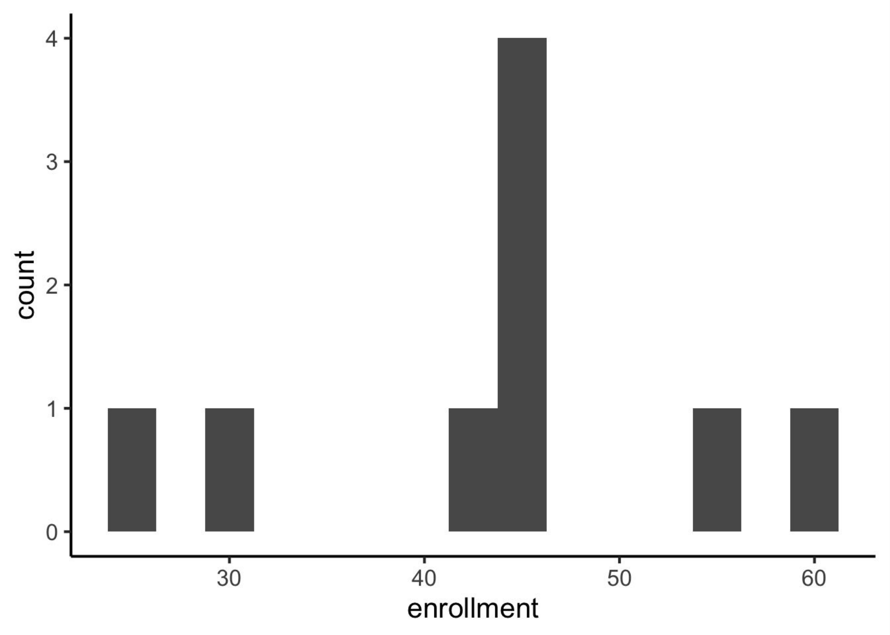
Independence

People might not enroll independently. But maybe this is less problematic across years?

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. Do we?

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…

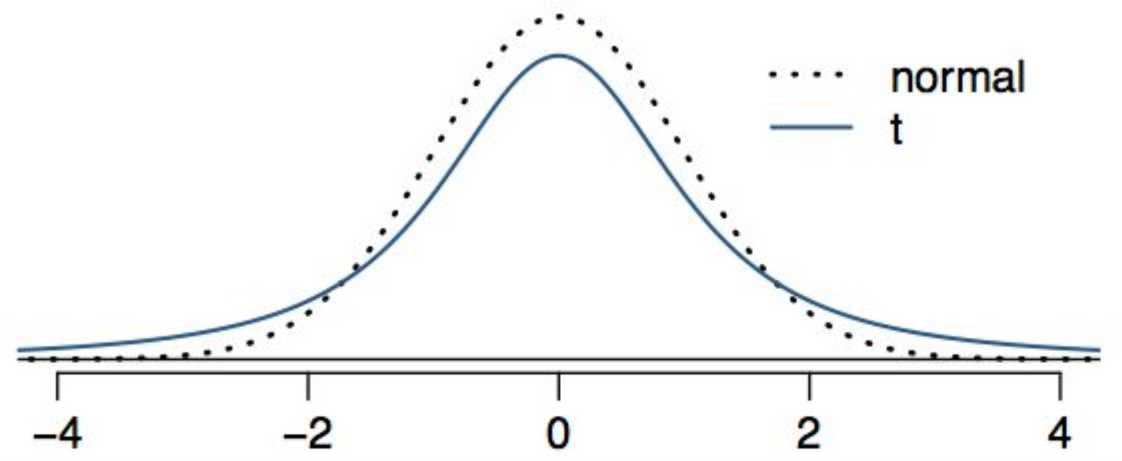
- the sampling distribution of the mean is nearly normal
- $\frac{s}{\sqrt{n}}$ is a reliable estimate of the standard error

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**.

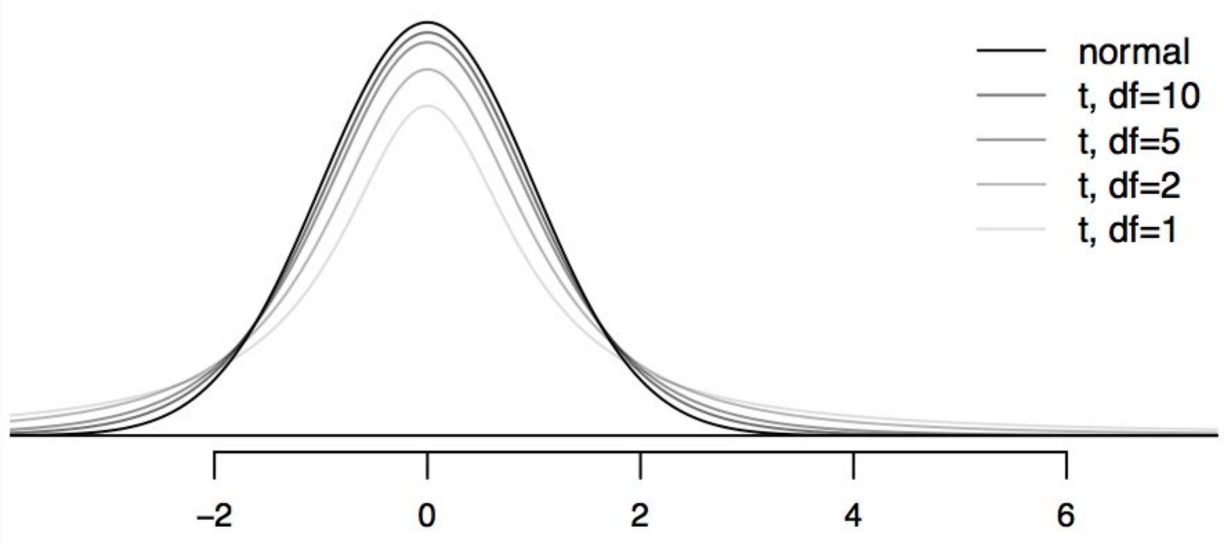
t has a similar shape, but fatter tails (i.e. extreme values are more likely).



The many different *ts*

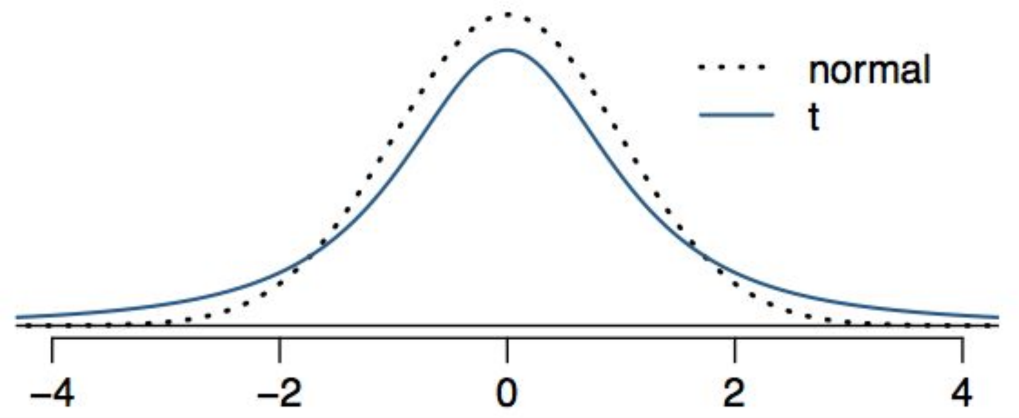
Centered at zero like the standard Normal (*z*-distribution).

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



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

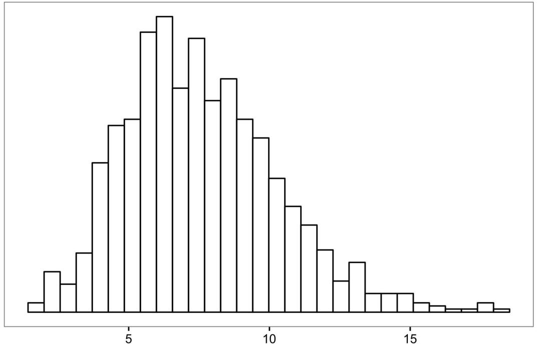
Why do we want fatter tails?



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

$$SE = \frac{s}{\sqrt{n}}$$

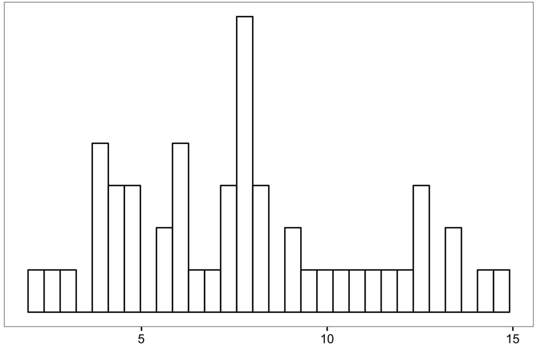
A reminder about the Central Limit Theorem



When I draw **independent samples** from the population, as sample size **approaches infinity,** the distribution of means approaches normality

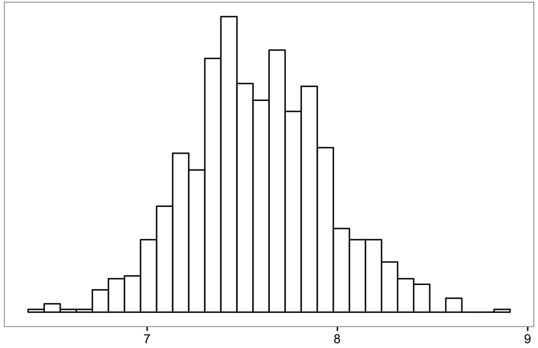
But what is it's SD?

The Sample's Standard Error!

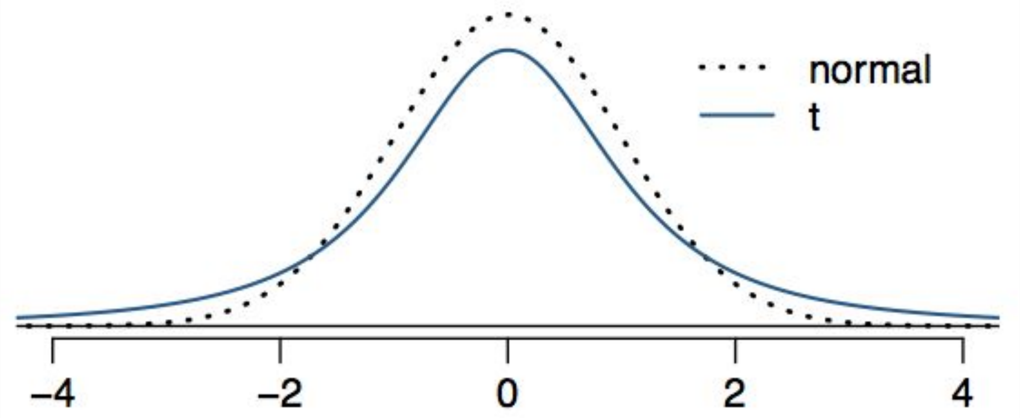


Take the mean,

Repeat many times...



Small samples have more variable standard deviations



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

$$SE = \frac{\boxed{s}}{\sqrt{n}}$$

Computing the test-statistic

year	enrollment
2010	30
2011	59
2012	44
2013	55
2014	46
2015	25
2016	43
2017	46
2018	44

$\bar{x} = 43.56$

$s = 10.67$

$n = 9$

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

$$point\ estimate = \bar{x} = 43.56$$

$$SE = \frac{s}{\sqrt{n}} = \frac{10.67}{\sqrt{9}} = 3.56$$

$$T = \frac{43.56 - 52}{3.56} = -2.37$$

$$df = 9 - 1 = 8$$

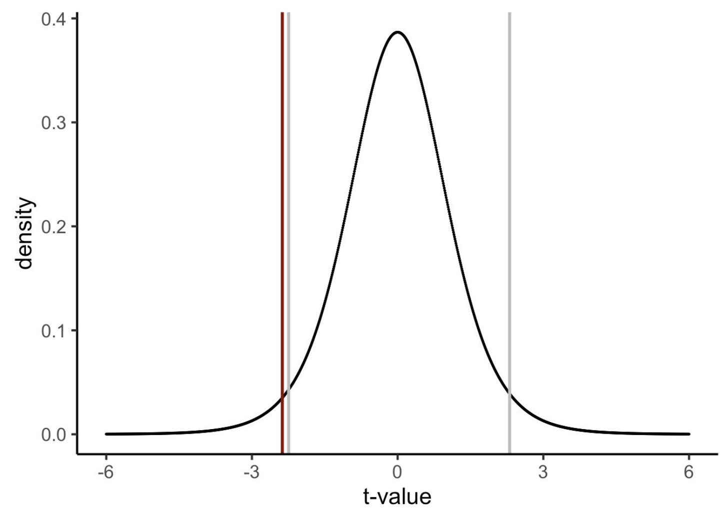
Finding the p-value

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

So for *t*(8), Using R:

```
> 2 * pt(-2.37, df = 8,
       lower.tail = TRUE)
[1] 0.045
```

Fewer than 52 people enroll on average



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 Enrollment.

Which of the following is the correct calculation of a 95% confidence interval for the number of people who enroll on average in PSYC201?

t*: `qt(p = .975, df = 8)`

```
2.306
```

$\bar{x} = 43.6$ $s = 10.67$ $n = 9$ $SE = 3.56$

(a) $43.6 \pm 1.96 \times 3.56$
(b) $43.6 \pm 2.31 \times 3.56$
(c) $43.6 \pm 2.31 \times 10.67$

Practice Question 2: Confidence interval for Enrollment.

Which of the following is the correct calculation of a 95% confidence interval for the number of people who enroll on average in PSYC201?

t*: `qt(p = .975, df = 8)`

`2.306`

$\bar{x} = 43.6$ $s = 10.67$ $n = 9$ $SE = 3.56$

(a) $43.6 \pm 1.96 \times 3.56$

(b) $43.6 \pm 2.31 \times 3.56$ → (35.3, 51.8)

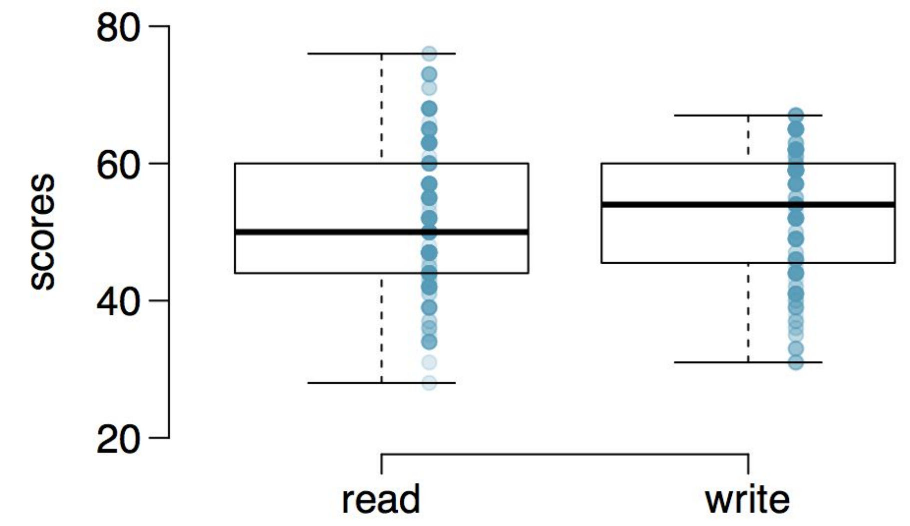
(c) $43.6 \pm 2.31 \times 10.67$

What does this mean?

An example of paired data

200 observations were randomly sampled from the HS&B survey. The same students took a reading and writing test, here are their scores.

Does there appear to be a difference between the average reading and writing test score?



An example of paired data

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

Are the reading and writing scores of each student independent of each other?

(a) Yes (b) No

An example of paired data

	id	read	write
1	70	57	52
2	86	44	33
3	141	63	44
4	172	47	52
⋮	⋮	⋮	⋮
200	137	63	65

Are the reading and writing scores of each student independent of each other?

(a) Yes **(b) No**

Analyzing paired data

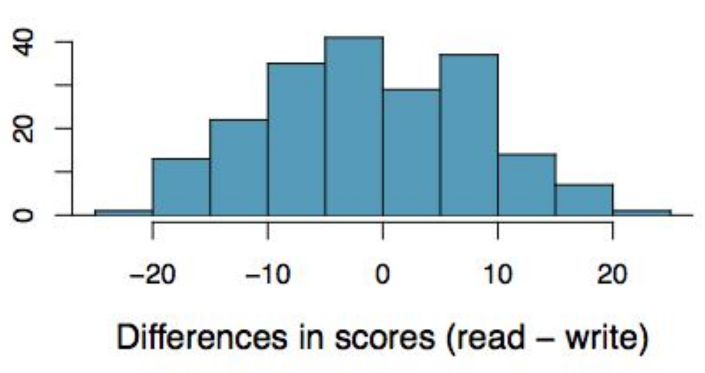
Two sets of data are ***paired*** if each data point in one set depends on a particular point in the other set.

To analyze paired data, we first compute the difference between in outcomes of each pair of observations.

diff = read - write

Note: It's 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



What counts as paired?

1. Verbal SAT and Math SAT from the same person
2. Spouse 1’s height and Spouse 2’s height
3. Parental anxiety score and child’s anxiety score
4. SAT scores at Harvard and Yale
5. “Hot shots” and “not shots” Steph Curry’s games
6. Control group blood pressure and Treatment group blood pressure

Two sets of data are paired if each data point in the first set has one clear “partner” in the second data set.

Parameter and point estimate

Parameter of interest: Average difference between the reading and writing scores of <u>all</u> high school students.

μ_{diff}

Point estimate: Average difference between the reading and writing scores of <u>sampled</u> high school students.

$\bar{x}_{diff}$

Setting up the Hypotheses

If there were no difference between scores on reading and writing exams, what difference would you expect on average?

0

What are the hypotheses for testing if there is a difference between the average reading and writing scores?

H0: There is no difference between the average reading and writing score.

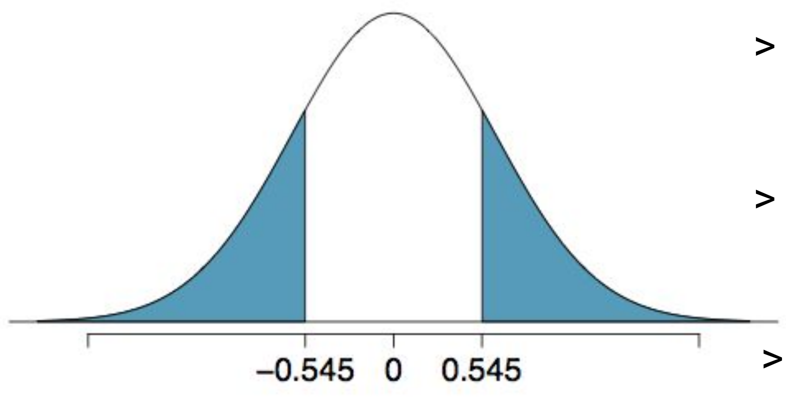
$$\mu_{diff} = 0$$

HA: There is a difference between the average reading and writing score.

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

Calculating the test-statistics and p-values

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 suggest a difference between the average scores on the two exams at $\alpha = 0.05$?



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

```
> t <- (-.545 - 0) / (8.887/ sqrt(200))
    = -.87

> pt(-.87, df = 199)
   = .1927

> p_val <- .1949 * 2
        = .3898
```

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.

Interpreting the p-value

Which of the following is the correct interpretation of the p-value?

(a) Probability that the average scores on the two exams are equal.

(b) Probability that the average scores on the two 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.

Interpreting the p-value

Which of the following is the correct interpretation of the p-value?

(a) Probability that the average scores on the two exams are equal.

(b) Probability that the average scores on the two 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.

Hypothesis testing and Confidence Intervals

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

Hypothesis testing and Confidence Intervals

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

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

Key ideas

1. When our samples are too small, we shouldn't use the Normal distribution. We use the *t* distribution to make up for uncertainty in our sample statistics
1. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)
3. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values