Review Session 6

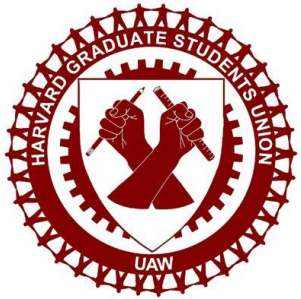
API-201, 10.22.21
Sophie Hill

What are you finding tricky with this week's material?

Plan for today

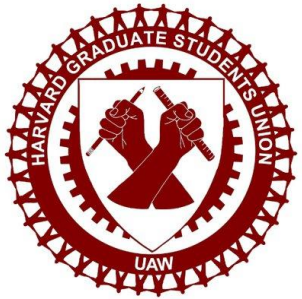
- Review new material (t-distribution, testing difference in means/proportions)
- Practice Problem
- Tips for PSet 6
- Heads up: Sophie is on strike with her union next week Weds-Fri!

Why I am going on strike



- I love teaching you all!
- I also think that my labor is valuable and that Harvard could not function without its student workers

Why I am going on strike



- I love teaching you all!
- I also think that my labor is valuable and that Harvard could not function without its student workers
- The Harvard Graduate Student Union, HGSU, made a series of proposals to the University on August 31st. It has been 7 weeks now and **the University has not bothered to respond.**

Why I am going on strike

- I love teaching you all!
- I also think that my labor is valuable and that Harvard could not function without its student workers
- The Harvard Graduate Student Union, HGSU, made a series of proposals to the University on August 31st. It has been 7 weeks now and **the University has not bothered to respond.**
- We have two main asks:
  - A living wage
  - Real recourse against sexual harassment

What does this mean for you?

The strike has been called for Weds 27th - Fri 29th October.

During that period, I will not be responding to any class-related emails or Slack messages.

I will not be holding Office Hours on Thurs 28th, nor a Review Session on Fri 29th.

I will return to work when the strike is over on Monday Nov 1st.

The first HGSU strike in Dec 2019. (Bloody freezing!!)

What does this mean for you?

Most importantly: I care deeply about your learning progress, and I want all of you to have the resources you need to succeed.

But your TFs and CAs cannot do their best work when they are worried about whether they can make rent this month.

I hope you will understand why this is important to me, and know that I will be ready to continue supporting you when the strike is over!

The first HGSU strike in Dec 2019. (Bloody freezing!!)

t-distribution

Which distribution to use?

Note: σ = population standard deviation

For proportions / difference-in-proportions:

→ normal distribution

For means / difference-in-means:

If σ is known → normal distribution

If σ is unknown → t-distribution

Which distribution to use?

Note: σ = population standard deviation

For proportions / difference-in-proportions:

→ normal distribution

For means / difference-in-means:

If σ is known → normal distribution

If σ is unknown → t-distribution

Why?

T-distribution for Two Types of Uncertainty...

	Sampling distribution	
	Mean	SE
Proportion	$q \approx \hat{q}$	$\sqrt{\frac{q(1-q)}{n}} \approx \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$
Mean	$\mu \approx \bar{x}$	$\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

T-distribution for Two Types of Uncertainty...

If we knew the true population proportion, q, then we would automatically know the mean and SE of the sampling distribution.

	Sampling distribution	
	Mean	SE
Proportion	$q \approx \hat{q}$	$\sqrt{\frac{q(1-q)}{n}} \approx \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$
Mean	$\mu \approx \bar{x}$	$\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

T-distribution for Two Types of Uncertainty...

In general, we don't know q, so we use our sample estimate q-hat to approximate it.

	Sampling distribution	
	Mean	SE
Proportion	$q \approx \hat{q}$	$\sqrt{\frac{q(1-q)}{n}} \approx \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$
Mean	$\mu \approx \bar{x}$	$\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

T-distribution for Two Types of Uncertainty...

For means, we use two different quantities: the sample mean approximates the population mean,

	Sampling distribution	
	Mean	SE
Proportion	$q \approx \hat{q}$	$\sqrt{\frac{q(1-q)}{n}} \approx \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$
Mean	$\mu \approx \bar{x}$	$\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

T-distribution for Two Types of Uncertainty...

For means, we use two different quantities: the sample mean approximates the population mean, and the sample standard deviation approximates the population standard deviation.

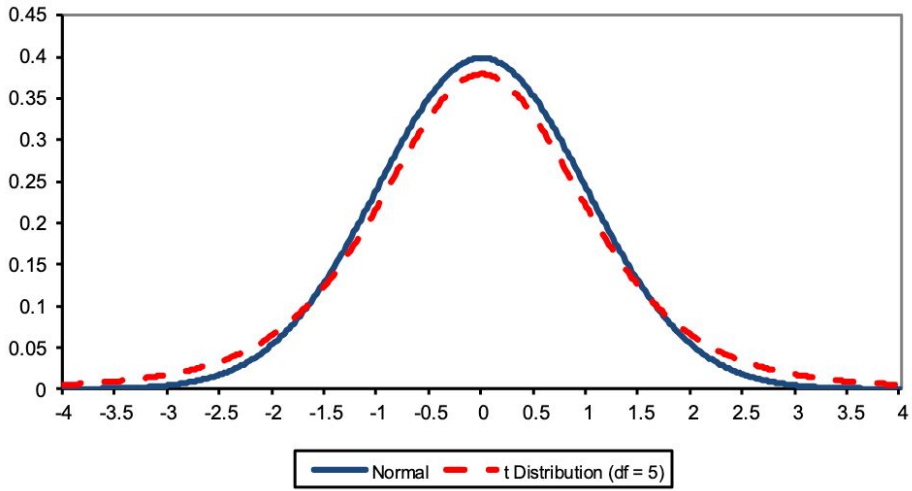
	Sampling distribution	
	Mean	SE
Proportion	$q \approx \hat{q}$	$\sqrt{\frac{q(1-q)}{n}} \approx \sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$
Mean	$\mu \approx \bar{x}$	$\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$

T-distribution for Two Types of Uncertainty…

7. APPENDIX #1: T-TESTS

- *t* distributions have slightly more variance than a normal distribution, to reflect the additional uncertainty from the estimation of not just $\bar{x}$ but also *s*

From Handout 15:



T-distribution vs Normal distribution

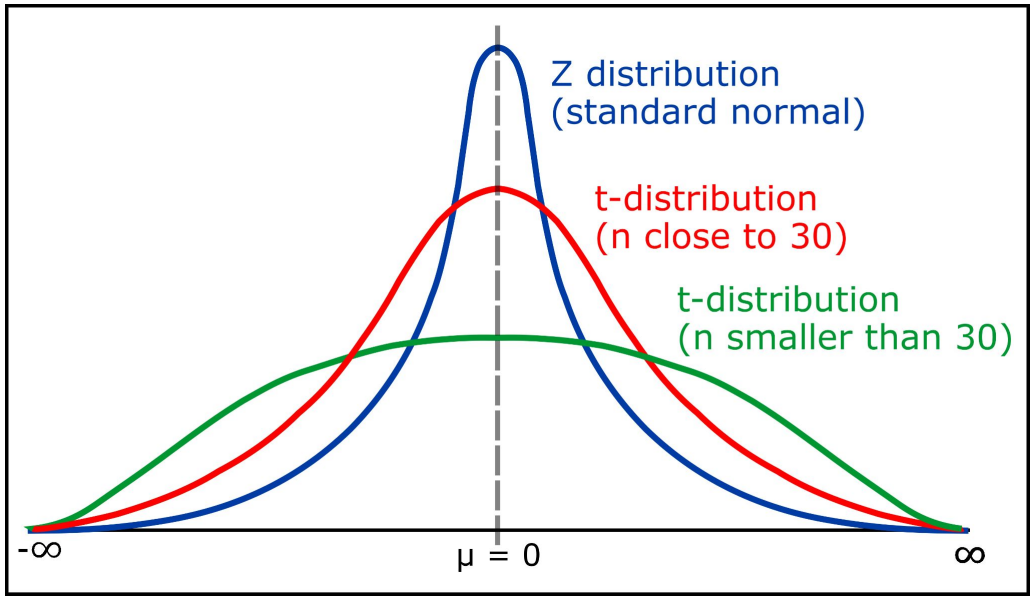
Compared to the Normal distribution, the t-distribution has:

-a similar shape (symmetric, unimodal)

-“Fatter tails”, aka higher spread

-an additional parameter: the degrees of freedom.

For a single mean, df = n-1
For a difference in means we use the smaller of the 2 sample sizes,
df = $n_{smaller}$ - 1



T-distribution vs Normal distribution

When n > 30, the t-distribution becomes v close to normal!

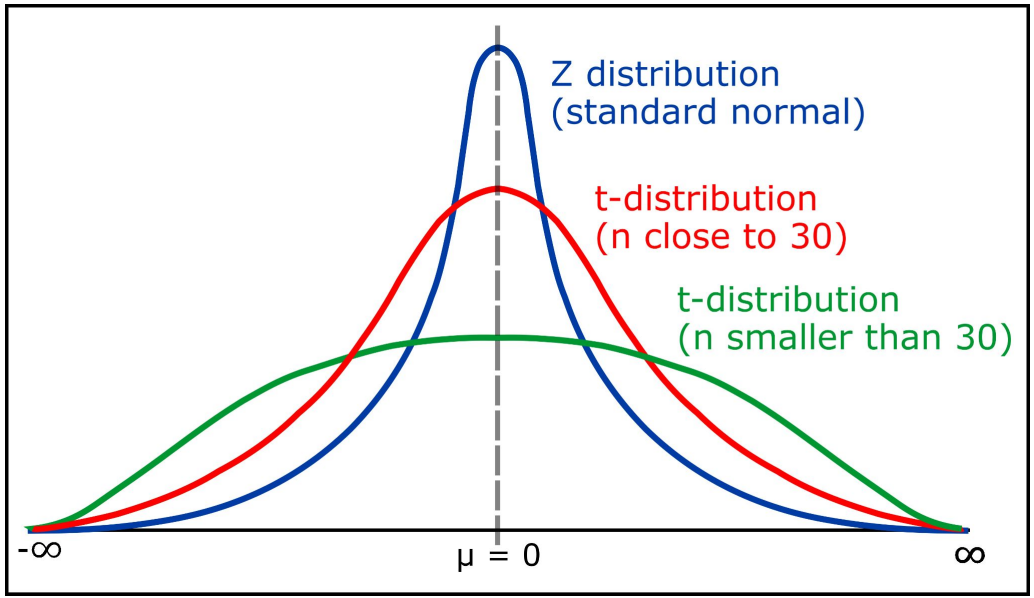
Compared to the Normal distribution, the t-distribution h

-a similar shape (symmetric, unimodal)

-"Fatter tails", aka higher spread

-an additional parameter: the degrees of freedom.

For a single mean, df = n-1
For a difference in means we use the smaller of the 2 sample sizes,
df = $n_{smaller}$ - 1



Which distribution to use?

Note: σ = population standard deviation

For proportions / difference-in-proportions:

→ normal distribution

For means / difference-in-means:

If σ is known → normal distribution

If σ is unknown → t-distribution

Which distribution to use?

Note: σ = population standard deviation

For proportions / difference-in-proportions:

→ normal distribution

For means / difference-in-means:

If σ is known → normal distribution

If σ is unknown → t-distribution

(though if we have a "large sample", say n > 30, this is basically equivalent to the normal anyway!)

Practice Problem

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

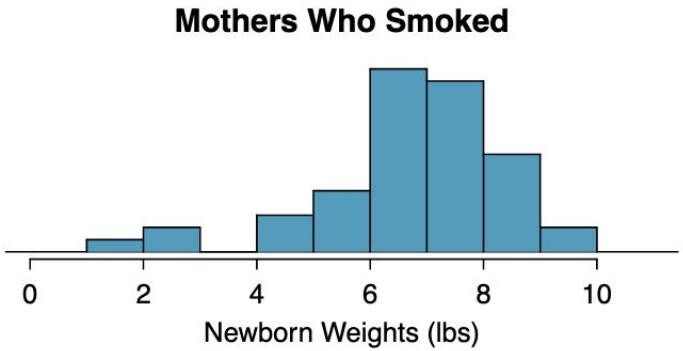
	fage	mage	weeks	weight	sex	smoke
1	NA	13	37	5.00	female	nonsmoker
2	NA	14	36	5.88	female	nonsmoker
3	19	15	41	8.13	male	smoker
⋮	⋮	⋮	⋮	⋮	⋮	
150	45	50	36	9.25	female	nonsmoker

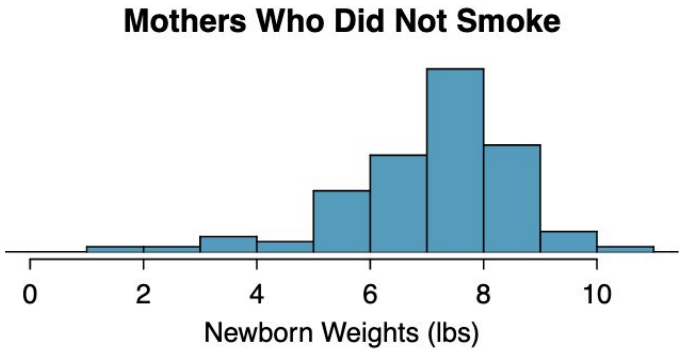
From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.



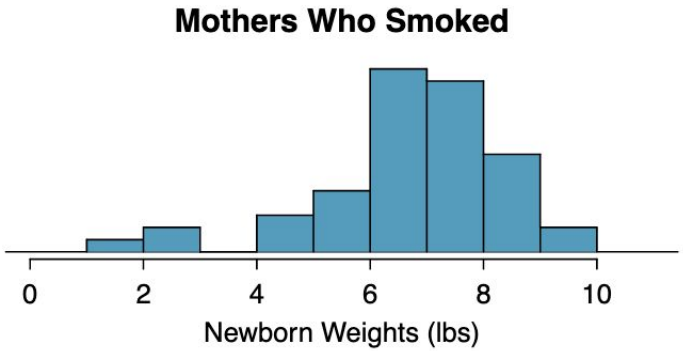


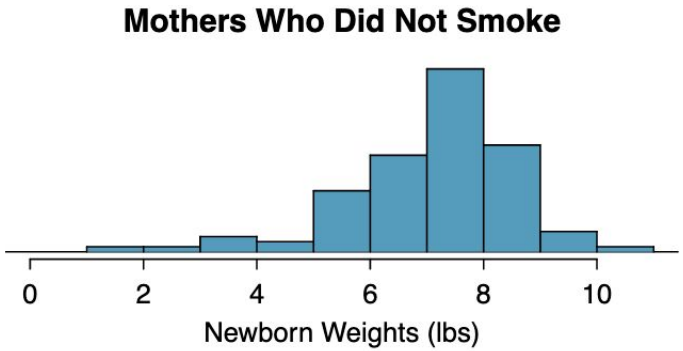
From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.





From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

1. **What is the null hypothesis?**

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

1. **What is the null hypothesis?**

H_0: There is no difference in average birth weight for newborns from mothers who did and did not smoke.

In notation: $\mu_n - \mu_s = 0$

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

2. What is our estimate of the difference? $\mu_n - \mu_s$

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

2. What is our estimate of the difference? $\mu_n - \mu_s$

$\bar{x}_n - \bar{x}_s$ = 7.18 - 6.78 = 0.4

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

3. What is the standard error of this estimate?

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

3. What is the standard error of this estimate?

$$SE = \sqrt{\frac{\sigma_n^2}{n_n} + \frac{\sigma_s^2}{n_s}} \approx \sqrt{\frac{s_n^2}{n_n} + \frac{s_s^2}{n_s}}$$

$$= \sqrt{\frac{1.60^2}{100} + \frac{1.43^2}{50}} = 0.26$$

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

4. What is the test statistic?

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

4. What is the test statistic?

$$T = \frac{0.40 - 0}{0.26} = 1.54$$

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

4. What is the p-value corresponding to this test statistic?

(a) First decide what the degrees of freedom should be.

(b) Then use this Excel formula to get the p-value:
=TDIST(abs(T), df , 2)

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

4. What is the p-value corresponding to this test statistic?

(a) $df = n_{smaller} - 1 = 50 - 1 = 49$

(b) =TDIST(abs(1.54), 49 , 2) = 0.13

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

5. With a significance level of 5%, what is the conclusion of our hypothesis test?

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Practice Problem: Baby weights

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

Let's test this using a dataset from North Carolina including 50 mothers who smoke and 100 mothers who don't.

5. With a significance level of 5%, what is the conclusion of our hypothesis test?

0.13 > 0.05

Fail to reject the null that there is no difference in average birth weight based on mother's smoking status.

	smoker	nonsmoker
mean	6.78	7.18
st. dev.	1.43	1.60
samp. size	50	100

From OIS Exercise 7.3.2

Tips for PSet 6

Tips for PSet 6

- For Q1, you might find it helpful to look at the glossary on Handout 15, Appendix #2

Tips for PSet 6

- For Q1, you might find it helpful to look at the glossary on Handout 15, Appendix #2

- For the second part: one way to “relate” two concepts is to ask “what happens if this one gets bigger/smaller?”

Tips for PSet 6

- For Q1, you might find it helpful to look at the glossary on Handout 15, Appendix #2

- For the second part: one way to “relate” two concepts is to ask “what happens if this one gets bigger/smaller?”

- On Q2, remember that the margin of error = multiplier x SE, so for a 95% CI the MoE = 1.95 x SE

Tips for PSet 6

- For Q1, you might find it helpful to look at the glossary on Handout 15, Appendix #2
- For the second part: one way to "relate" two concepts is to ask "what happens if this one gets bigger/smaller?"
- On Q2, remember that the margin of error = multiplier x SE, so for a 95% CI the MoE = 1.95 x SE
- For Q3 (3), see the footnote for guidance on how to draw random numbers in Excel. To prevent these from being recalculated when you sort the table, you can highlight the column, Copy, and then click "Paste Special → Values only".