Review Session 3

API-201, 9.24.21
Sophie Hill


Stats
midterm
MPP1s

7. MIDTERM EXAM - LOGISTICS

- **Friday, Oct 1, 8:00 - 10:00 am**
- Classroom: TBA
- Bring to the exam:
  - A calculator
  - One double-sided cheat sheet
  - Your Harvard ID

- Resources:
  - Practice problems and solutions posted
  - Past midterms and solutions posted
  - Problem set solutions posted
  - Pre-class exercise solutions posted
  - Extra Office Hours

- Two-Stage Exam
  - Stage 1: Individual; 80 minutes
  - Stage 2: Group; 35 minutes

Cheat sheet

- You can bring 1 double-sided cheat sheet into the midterm

- Remember: the best cheat sheet is **the one you make yourself**
  - Focus on the topics you feel *least* comfortable with
  - You can include formulas... but also diagrams, tables, key takeaways, examples from class, and more!

Goals for today

Today we're going to do a BUNCH of practice problems...

- to **review** some recent material
- to practice our **exam strategies**
- and to build our **confidence**!

Practice Problem 1

(a) Use these raw numbers to fill out the table with the PMF (probability mass function) and CDF (cumulative distribution function).

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740		
2 people	44,038		
3 people	19,333		
4 people	16,468		
5 people	7,442		
6 people	2,851		
7+ people	1,714		
Total	127,586		

WAIT!

Before answering, search for clues.

Practice Problem 1

(a) Use these raw numbers to fill out the table with the PMF (probability mass function) and CDF (cumulative distribution function).

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

...s data (2018)

Ho...		P(X=x)	P(X ≤ x)
	...0		
	...8		
	...33		
4 p...	...6,468		
5 people	...42		
6 people	...1		
7+ people			
Total	12...		

Aha! This question must be about **probability distributions**.

Practice Problem 1

(a) Use these raw numbers to fill out the table with the PMF (probability mass function) and CDF (cumulative distribution function).

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

...us data (2018)

Ho...		P(X=x)	P(X ≤ x)
	...0		
	...8		
	...33		
4 p...	...6,468		
5 people	...42		
6 people	...1		
7+ people			
Total	12...		

For part (a), all we have to do is fill out the table. **Phew!**

Practice Problem 1

(a) Use these raw numbers to fill out the table with the PMF (probability mass function) and CDF (cumulative distribution function).

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

...s data (2018)

Ho...		P(X=x)	P(X ≤ x)
	...0		
	...8		
	...33		
4 p...	...6,468		
5 people	...42		
6 people	...1		
7+ people			
Total	12...		

For part (b), we need to write **2-3 sentences**.

WORK SMARTER, NOT HARDER

Practice Problem 1

(a) Use these raw numbers to fill out the table with the PMF (probability mass function) and CDF (cumulative distribution function).

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740		
2 people	44,038		
3 people	19,333		
4 people	16,468		
5 people	7,442		
6 people	2,851		
7+ people	1,714		
Total	127,586		

Practice Problem 1

(a) Use these raw numbers to fill out the table with the PMF (probability mass function) and CDF (cumulative distribution function).

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740	0.28	
2 people	44,038	0.35	
3 people	19,333	0.15	
4 people	16,468	0.13	
5 people	7,442	0.06	
6 people	2,851	0.02	
7+ people	1,714	0.01	
Total	127,586	1.00	

Practice Problem 1

(a) Use these raw numbers to fill out the table with the PMF (probability mass function) and CDF (cumulative distribution function).

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740	0.28	0.28
2 people	44,038	0.35	0.63
3 people	19,333	0.15	0.78
4 people	16,468	0.13	0.91
5 people	7,442	0.06	0.96
6 people	2,851	0.02	0.99
7+ people	1,714	0.01	1
Total	127,586	1.00	

Practice Problem 1

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740	0.28	0.28
2 people	44,038	0.35	0.63
3 people	19,333	0.15	0.78
4 people	16,468	0.13	0.91
5 people	7,442	0.06	0.96
6 people	2,851	0.02	0.99
7+ people	1,714	0.01	1
Total	127,586	1.00	

Practice Problem 1

(b) Explain the meaning of the PMF and the CDF in this context, in terms that a non-statistician would understand. [2-3 sentences]

The PMF tells us the probabilities of picking a household with size *x*: e.g., there is a 28% chance of picking a single-person household.

The CDF tells us the probability of picking a household with size *x* or less: e.g., there is a 96% chance of picking a household of up to 5 people.

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740	0.28	0.28
2 people	44,038	0.35	0.63
3 people	19,333	0.15	0.78
4 people	16,468	0.13	0.91
5 people	7,442	0.06	0.96
6 people	2,851	0.02	0.99
7+ people	1,714	0.01	1
Total	127,586	1.00	

Practice Problem 1

(c) What is the probability of picking a household with 4 or more people at random from this sample?

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740	0.28	0.28
2 people	44,038	0.35	0.63
3 people	19,333	0.15	0.78
4 people	16,468	0.13	0.91
5 people	7,442	0.06	0.96
6 people	2,851	0.02	0.99
7+ people	1,714	0.01	1
Total	127,586	1.00	

Practice Problem 1

(c) What is the probability of picking a household with 4 or more people at random from this sample?

Using the PDF:

P(4 ppl or more)
= P(4 ppl) + P(5 ppl)
+ P(6 ppl) + P(7 ppl or more)
= 0.13 + 0.06 + 0.02 + 0.01
= 0.22

Number of households by size
US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740	0.28	0.28
2 people	44,038	0.35	0.63
3 people	19,333	0.15	0.78
4 people	16,468	0.13	0.91
5 people	7,442	0.06	0.96
6 people	2,851	0.02	0.99
7+ people	1,714	0.01	1
Total	127,586	1.00	

Practice Problem 1

(c) What is the probability of picking a household with 4 or more people at random from this sample?

Using the CDF:

P(4 ppl or more)
= 1 - P(3 ppl or less)
= 1 - 0.78
= 0.22

Number of households by size

US Census data (2018)

Household Size (x)	n	P(X=x)	P(X ≤ x)
1 person	35,740	0.28	0.28
2 people	44,038	0.35	0.63
3 people	19,333	0.15	0.78
4 people	16,468	0.13	0.91
5 people	7,442	0.06	0.96
6 people	2,851	0.02	0.99
7+ people	1,714	0.01	1
Total	127,586	1.00	

Practice Problem 1

(c) What is the probabili
of picking a hous
with 4 or more p
random from thi
sample?

Using the CDF:

P(4 ppl or more)
= 1 - P(3 ppl or less)
= 1 - 0.78
= 0.22

Number of households by size
US Census data (2018)

Household Size (x)	p	P(X=x)	P(X ≤ x)
			0.28
			0.63
			0.78
			0.91
			0.96
			0.99
			1

If you just wrote the number 0.22, you *might* get full credit. But if you made a typo or an error in your calculation, you'd get none!

By laying out what I'm doing (using the PDF/CDF) and writing out the equation in notation before plugging in the numbers, I am ensuring that I can get *some* credit even if the final answer is wrong.

You gotta show your work (work work work work)

Practice Problem 2

Your friend at a different school just got back their midterm results. They got 83/100 on the stats midterm and 94/100 on the econ midterm.

(a) Your friend tells you: “I scored 11 points higher on the econ midterm than the stats midterm. I never knew I was so much better at econ than at stats!” Do you agree with their conclusion? [1-2 sentences]

Practice Problem 2

This question is about **comparing distributions**. But note that part (a) only requires 1-2 sentences.

Your friend at a different sch… …lts. They got 83/100 on the stats … on the econ midterm.

(a) Your friend tells you: "I scored 11 points higher on the econ midterm than the stats midterm. I never knew I was so much better at econ than at stats!" Do you agree with their conclusion? [1-2 sentences]

Practice Problem 2

Your friend at a different school just got back their midterm results. They got 83/100 on the stats midterm and 94/100 on the econ midterm.

(a) Your friend tells you: “I scored 11 points higher on the econ midterm than the stats midterm. I never knew I was so much better at econ than at stats!” Do you agree with their conclusion? [1-2 sentences]

No. In order to compare our friend’s performance on both midterms, we would want to know how they performed *relative to others in the class*. To do this, we would need some more information about the distribution of both sets of test scores.

Practice Problem 2

Your friend at a different school just got back their midterm results. They got 83/100 on the stats midterm and 94/100 on the econ midterm.

(b) Your friend takes your advice and asks the professors for more information. The stats professor says their scores were normally distributed with a mean of 77 and a standard deviation of 3. The econ professors says their scores were normally distributed with a mean of 88 and a standard deviation of 6.

"Aha!," says your friend, "I scored 6 points above the class average on both tests. That means I am equally good at stats and econ!" Do you agree with this conclusion? [1 paragraph]

Practice Problem 2

Your friend at a different school just got bac
They got 83/100 on the stats midterm and 9
midterm.

(b) Your friend takes your advice and asks the
information. The stats professor says their s
distributed with a mean of 77 and a standard de
econ professors says their scores were normally distribu
mean of 88 and a standard deviation of 6.

"Aha!," says your friend, "I scored 6 points above the class
average on both tests. That means I am equally good at stats
econ!" Do you agree with this conclusion? [1 paragraph]

We are given some explicit numbers here so we should expect to do some calculations. "1 paragraph" means we need to explain and interpret these calculations.

Practice Problem 2

Stats ~ N(77, 3), our friend scored 83

Econ ~ N(88, 6), our friend scored 94

Practice Problem 2

Stats ~ N(77, 3), our friend scored 83

Econ ~ N(88, 6), our friend scored 94

Our friend is correct that they scored 6 points above the class average on both midterms. However, in order to interpret this, we need to take into account the *spread* of the test scores, which is captured by the standard deviation. We can do this by calculating the Z-scores:

$$Z = (\text{student's score} - \text{mean}) / (\text{standard deviation})$$

$Z_{stats} = (83 - 77)/3 = 6/3 = 2$ $Z_{econ} = (94 - 88)/6 = 6/6 = 1$

Our friend scored 2 standard deviations above the mean on stats, but only 1 standard deviation above the mean on econ. So our friend actually did better (relative to the class) on stats than on econ!

Practice Problem 2

Summarizing information from the question

Stats ~ N(77, 3), our friend scored 83

Econ ~ N(88, 6), our friend scored 94

Our friend is correct that they scored 6 points above the class average on both midterms. However, in order to interpret this, we need to take into account the *spread* of the test scores, which is captured by the standard deviation. We can do this by calculating the Z-scores:

Z = (student's score - mean) / (standard deviation)

$Z_{stats} = (83 - 77)/3 = 6/3 = 2$

$Z_{econ} = (94 - 88)/6 = 6/6 = 1$

Our friend scored 2 standard deviations above the mean on stats, but only 1 standard deviation above the mean on econ. So our friend actually did better (relative to the class) on stats than on econ!

Practice Problem 2

Showing conceptual understanding

Stats ~ N(77, 3), our friend scored 83

Econ ~ N(88, 6), our friend scored 94

Our friend is correct that they scored 6 points above the class average on both midterms. However, in order to interpret this, we need to take into account the *spread* of the test scores, which is captured by the standard deviation. We can do this by calculating the Z-scores:

$$Z = (\text{student's score} - \text{mean}) / (\text{standard deviation})$$

$Z_{stats} = (83 - 77)/3 = 6/3 = 2$ $Z_{econ} = (94 - 88)/6 = 6/6 = 1$

Our friend scored 2 standard deviations above the mean on stats, but only 1 standard deviation above the mean on econ. So our friend actually did better (relative to the class) on stats than on econ!

Practice Problem 2

Knowing the correct formula to use

Stats ~ N(77, 3), our friend scored 83

Econ ~ N(88, 6), our friend scored 94

Our friend is correct that they scored 6 points above the class average on both midterms. However, in order to interpret this, we need to take into account the *spread* of the test scores, which is captured by the standard deviation. We can do this by calculating the Z-scores:

Z = (student's score - mean) / (standard deviation)

$Z_{stats} = (83 - 77)/3 = 6/3 = 2$

$Z_{econ} = (94 - 88)/6 = 6/6 = 1$

Our friend scored 2 standard deviations above the mean on stats, but only 1 standard deviation above the mean on econ. So our friend actually did better (relative to the class) on stats than on econ!

Practice Problem 2

Applying the formula

Stats ~ N(77, 3), our friend scored 83

Econ ~ N(88, 6), our friend scored 94

Our friend is correct that they scored 6 points above the class average on both midterms. However, in order to interpret this, we need to take into account the *spread* of the test scores, which is captured by the standard deviation. We can do this by calculating the Z-scores:

$$Z = (\text{student's score} - \text{mean}) / (\text{standard deviation})$$

$Z_{stats} = (83 - 77)/3 = 6/3 = 2$

$Z_{econ} = (94 - 88)/6 = 6/6 = 1$

Our friend scored 2 standard deviations above the mean on stats, but only 1 standard deviation above the mean on econ. So our friend actually did better (relative to the class) on stats than on econ!

Practice Problem 2

Interpreting the results

Stats ~ N(77, 3), our friend scored 83

Econ ~ N(88, 6), our friend scored 94

Our friend is correct that they scored 6 points above the class average on both midterms. However, in order to interpret this, we need to take into account the *spread* of the test scores, which is captured by the standard deviation. We can do this by calculating the Z-scores:

$$Z = (\text{student's score} - \text{mean}) / (\text{standard deviation})$$

$Z_{stats} = (83 - 77)/3 = 6/3 = 2$ $Z_{econ} = (94 - 88)/6 = 6/6 = 1$

Our friend scored 2 standard deviations above the mean on stats, but only 1 standard deviation above the mean on econ. So our friend actually did better (relative to the class) on stats than on econ!

Practice Problem 3

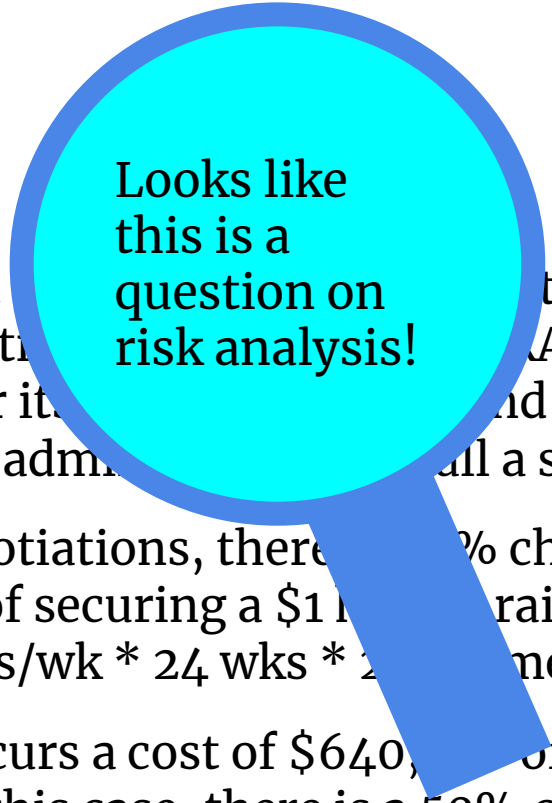
The Harvard Graduate Student Union (HGSU) is currently bargaining with the administration over a new contract for TFs, CAs, and RAs. The union wants to achieve the biggest pay rise for its 2,000 members, and it has two options: continue to negotiate with the administration or call a strike.

If the union continues the negotiations, there is 20% chance of a deadlock (no pay raise), and a 80% chance of securing a $1 hourly raise. (That equates to an aggregate gain of $1/hr * 20hrs/wk * 24 wks * 2,000 members = $960,000.)

If the union calls a strike, it incurs a cost of $640,000 of strike pay for workers whose wages are withheld. In this case, there is a 50% chance of a deadlock and a 50% of a $3 hourly raise.

Assuming that the union is risk neutral, what should it do?

Practice Problem 3



The Harvard Graduate Student tly bargaining with the administration over a new conti As. The union wants to achieve the biggest pay rise for it nd it has two options: continue to negotiate with the adm ll a strike.

If the union continues the negotiations, there % chance of a deadlock (no pay raise), and a 80% chance of securing a $1 raise. (That equates to an aggregate gain of $1/hr * 20hrs/wk * 24 wks * members = $960,000.)

If the union calls a strike, it incurs a cost of $640, of strike pay for workers whose wages are withheld. In this case, there is a 50% chance of a deadlock and a 50% of a $3 hourly raise.

Assuming that the union is risk neutral, what should it do?

Practice Problem 3

$EV_{\text{no strike}}$

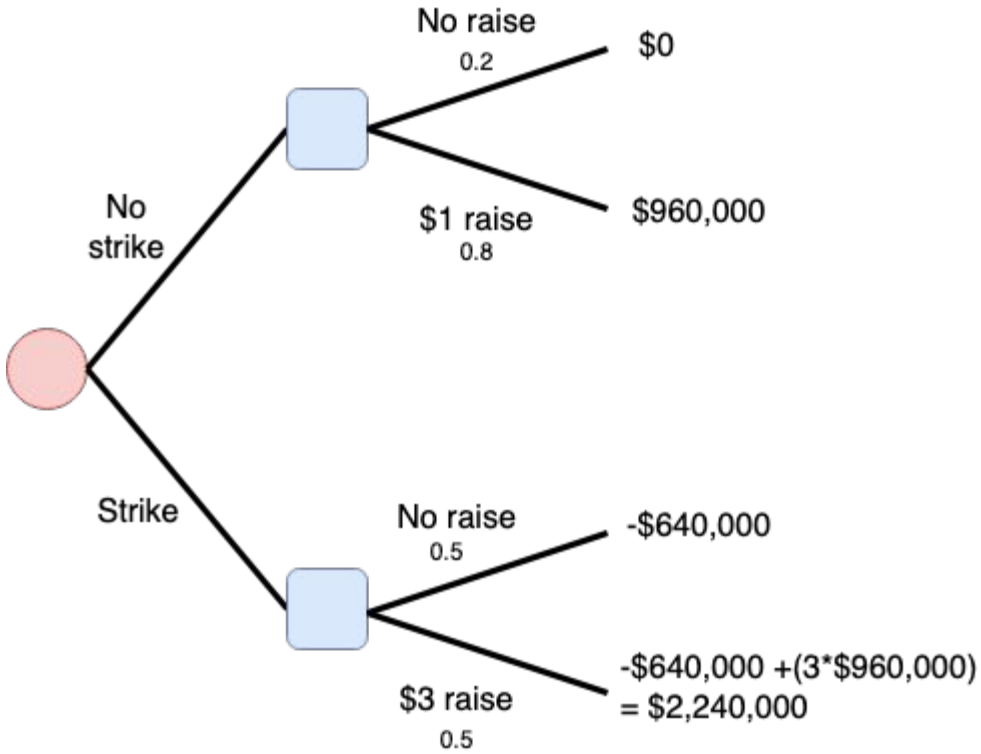
= 0.2*$0 + 0.8*$960,000

= $768,000

EV_{strike}

= 0.5*(-$640,000) +
0.5(-$640,000 + 3*$960,000)

= $800,000

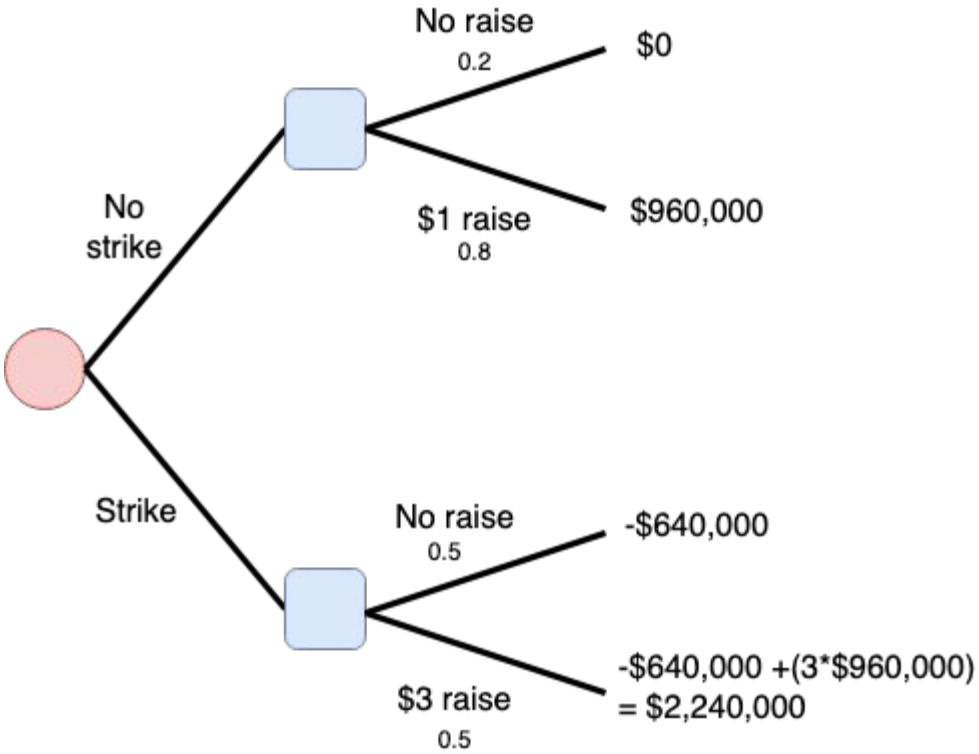


Practice Problem 3

$EV_{\text{no strike}}$ = **$768,000**

EV_{strike} = **$800,000**

The question tells us that the union is risk neutral. Therefore, the union should choose to call a strike, since option provides a higher expected value.



Practice Problem 3

(b) Suppose the union is uncertain about the total cost of strike pay (since it is not clear *ex ante* how many workers will have their wages withheld).

What is the cost of strike pay that would make the union indifferent between the two options?

Practice Problem 3

(b) Suppose the union is uncertain about the total cost of strike pay (since it is not clear *ex ante* how many workers will have their wages withheld).

What is the cost of strike pay that would make the union indifferent between the two options?

The union is indifferent when:

$$EV_{\text{no strike}} = EV_{\text{strike}}$$

In this sensitivity analysis, we are going to let the cost of strike pay vary, but everything else is fixed. So $EV_{\text{no strike}}$ is still \$768,000.

$\$768,000 = 0.5*(-x) + 0.5(-x + 3*\$960,000)$

$\$768,000 = -x + 0.5*3*\$960,000$

$\$768,000 = -x + \$1,440,000$

$x = \$1,440,000 - \$768,000$

$= \$672,000$

Practice Problem 4

A relative of yours just texted you the following:

“I heard on the news that 75 out of the 100 people in the ICU with COVID had been vaccinated!! I TOLD you, those “vaccines” actually make you *more* likely to get serious COVID! It’s such a shame that only 600 people in our town of 20,000 are still unvaccinated.” **Do you agree with their conclusions?**

Practice Problem 4

A relative of yours just texted you the following:

"I heard on the news that 75 out of the 100 people in the ICU with COVID had been vaccinated!! I TOLD you, those "vaccines" actually make you *more* likely to get serious COVID! It's such a shame that only 600 people in our town of 20,000 are still unvaccinated." **Do you agree with their conclusions?**

P(unvax) = 600/20,000 = 0.03

P(vax) = 1 - 0.03 = 0.97

P(ICU) = 100/20,000 = 0.005

P(not in ICU) = 1 - 0.005 = 0.995

P(vax | ICU) = 0.75

P(unvax | ICU) = 1 - 0.75 = 0.25

Practice Problem 4

P(unvax) = 0.03 P(vax) = 0.97

P(ICU) = 0.005 P(not in ICU) = 0.995

P(vax | ICU) = 0.75P(unvax | ICU) = 0.25

Our relative has got their conditional probabilities mixed up. They are pointing to the fact that P(vax | ICU) > P(unvax | ICU).

But in order to determine whether the vaccinated or unvaccinated have a higher chance of ending up in the ICU with Covid, we actually need to compare **P(ICU | vax)** and **P(ICU | unvax)**.

Practice Problem 4

The multiplication rule tells us that P(A|B) = P(A&B)/P(B).

So we can rewrite these conditional probabilities as follows:

$$P(ICU \mid vax) = P(ICU \,\&\, vax) / P(vax)$$

$$P(ICU \mid unvax) = P(ICU \,\&\, unvax) / P(unvax)$$

Practice Problem 4

We can use the multiplication rule *again* - this time to rewrite the joint probability!

$$P(ICU \mid vax) = P(ICU \,\&\, vax) / P(vax)$$

$$= [P(vax \mid ICU) * P(ICU)] / P(vax)$$

$$P(ICU \mid unvax) = P(ICU \,\&\, unvax) / P(unvax)$$

$$= [P(unvax \mid ICU) * P(ICU)] / P(unvax)$$

Practice Problem 4

P(unvax) = 0.03 P(vax) = 0.97

P(ICU) = 0.005 P(not in ICU) = 0.995

P(vax | ICU) = 0.75P(unvax | ICU) = 0.25

P(ICU | vax) = P(ICU & vax) / P(vax)

= [P(vax | ICU)*P(ICU)] / P(vax) = (0.75*0.005)/0.97 = **0.004**

P(ICU | unvax) = P(ICU & unvax) / P(unvax)

= [P(unvax | ICU)*P(ICU)] / P(unvax) = (0.25*0.005) / 0.03 = **0.04**

Practice Problem 4

The probability of being in the ICU given that you are vaccinated is 0.4%.

The probability of being in the ICU given that you are unvaccinated is 4%.

So, we do not agree with our relative's conclusions. Based on these numbers, the risk of serious Covid is actually 10 times higher for the unvaccinated compared to the vaccinated.

Practice Problem 4

(b) One of your friends, who is also taking API-201, says that your relative's confusion stems from the Base Rate Fallacy. Explain what the Base Rate Fallacy is and how it might apply in this particular context. [1 paragraph]

Practice Problem 4

(b) One of your friends, who is also taking API-201, says that your relative's confusion stems from the Base Rate Fallacy. Explain what the Base Rate Fallacy is and how it might apply in this particular context. [1 paragraph]

The Base Rate Fallacy occurs when we draw conclusions about a characteristic based on new information, without taking the prevalence of that characteristic into account.

In this case, our relative failed to take the prevalence of vaccination into account when interpreting the fact that 75% of those in the ICU were vaccinated. Consider an extreme case, where 100% of the town is vaccinated, and there is just 1 person in the ICU. Then 100% of the ICU patients are vaccinated, but the risk of serious COVID with vaccination is very small!

Great job everyone!!