Formulas

The Three Probability Axioms

PA1 **Nonnegativity**: For any statement A, $\Pr(A) \geq 0$.

PA2 **Certainty**: If you're certain that a statement A is true, then $\Pr(A) = 1$ (and if you're certain that A is false, then $\Pr(A) = 0$).

PA3 **Additivity** If A and B are never true at the same time, then $\Pr(A \vee B) = \Pr(A) + \Pr(B)$.

The following is a consequence of the probability axioms:

1. **Partitionality**: If $H_1, H_2, \ldots H_n$ are statements and you know that *exactly one* of them is true, then $\Pr(H_1) + \Pr(H_2) + \cdots + \Pr(H_n) = 1$.

The Ratio Formula and Theorems based on it

1. **The Ratio Formula for Conditional Probability**: If A and B are statements and $\Pr(B) \neq 0$, then $\Pr(A \mid B) = \frac{\Pr(A\&B)}{\Pr(B)}$.
2. **Bayes' Theorem**: If E and H are statements, then

$$\Pr(H \mid E) = \Pr(H)\frac{\Pr(E \mid H)}{\Pr(E)}.$$

3. **Total Probability**: If $H_1, H_2, \ldots H_n$ are statements and you know that *exactly one* of them is true, then for any statement E:

$$\Pr(E) = \Pr(E \mid H_1)\Pr(H_1) + \Pr(E \mid H_2)\Pr(H_2) \cdots + \Pr(E \mid H_n)\Pr(H_n).$$

Conditionalization

The Conditionalization Norm: If you have confidence $\Pr(H)$ in a hypothesis H, and then you learn E, then your new confidence in H upon learning E, $Pr_E(H)$, should equal your conditional probability in H given E:

$$Pr_E(H) \overset{\text{should}}{=} Pr(H \mid E).$$

Evidence For and Independence

1. **Definition**:
   (a) E is **evidence for** H if $\Pr(H \mid E) > \Pr(H)$.
   (b) E is **evidence against** H if $\Pr(H \mid E) < \Pr(H)$.
   (c) H is **independent of** E if $\Pr(H \mid E) = \Pr(H)$.
2. **The Evidence-For Lemma**: The following four are equivalent:

$$\Pr(H \mid E) > \Pr(H) \qquad \Pr(E \mid H) > \Pr(E)$$
$$\Pr(H \mid E) > \Pr(H \mid \neg E) \qquad \Pr(E \mid H) > \Pr(E \mid \neg H).$$

They are still equivalent if the ">" are replaced with "<" or "="

Logic, Reasoning, and Persuasion 07; Deductive Reasoning Problem Set

- Due Date: **Wednesday, November 26**. Points: **100** + 5 extra credit.
- Please write or type your answers *on a separate document/paper*
- Show your calculations! Otherwise, wrong answers will get no partial credit.

1 | Popu lar (15 points + 5 points bonus)

You'd like to figure out whether you're popular in your social group. You decide to ask some trusted friends to go out and gather some data about this and report back to you. But as they are gathering data, you begin to think about your friends' reliability:

1. You think Alice probably wouldn't want to hurt your feelings, so if you were popular, she would definitely tell you, but if you were *unpopular*, there is a 75% chance she would lie to you and tell you that you are popular anyway.
2. Bob loves messing with you, so if you're popular, there's a 75% chance he'll tell you you're *unpopular*, and if you're unpopular, there's a 75% chance he'll tell you you're *popular*.

As you ponder, Cam texts you and says he doesn't like this data gathering thing and he's just gonna flip a coin to decide whether to tell you you're popular or not.

Question 1.1 (15 points): For each statement, is it *evidence for, evidence against, or independent of* the statement "you are popular"?

1. A = Alice tells you you are popular.
2. B = Bob tells you you are popular.
3. C = Cam tells you you are popular.

Use the definition of *evidence for* and/or the evidence-for lemma.

Question 1.bonus (5 points bonus): suppose Alice, Bob, and Cam don't coordinate, so their answers to you are independent of each other. Is the statement "Alice, Bob, and Cam all tell you you are popular" evidence *for* or *against* the statement "you are popular"? Why?

2 | Gotham Policing (30 points)

Residents of Gotham either live downtown or uptown. Recently, the police and police reform groups have disagreed about whether or not Gotham policing is biased against residents of downtown.

Here are the arrest numbers in the year 2025 for Gotham:

	Downtown	Uptown
Population	750000	250000
Arrests	15000	1250

Let Pr_D be the probabilities for people living downtown, and Pr_U be the probabilities for people living uptown. So $\Pr_D(\text{arrested})$ is the probability that someone living downtown gets arrested.

Question 2.1 (5 points): What's the probability of being arrested if one lives downtown? How about if one lives uptown?

1. $Pr_D(\text{arrested}) =$
2. $Pr_U(\text{arrested}) =$

The Gotham court has an oracle that can determine with perfect accuracy whether someone who has been arrested for a crime is guilty of that crime. Here are the numbers of arrested people the oracle determined are guilty in the year 2000:

	Downtown	Uptown
Arrested and Guilty	3000	250

Question 2.2 (5 points): What's the probability of being arrested and guilty if one lives downtown? How about uptown?

1. $Pr_D(\text{arrested } \& \text{ guilty}) =$
2. $Pr_U(\text{arrested } \& \text{ guilty}) =$

The city concludes based on this information that the crime rate is much higher downtown than uptown. Officials say that, given that the crime rate is almost quadruple in downtown what it is in uptown, more policing is needed in the higher-crime areas.

Question 2.3 (5 points): What's the probability of being found guilty, given that one is arrested, if one lives downtown or uptown?

1. $Pr_D(\text{guilty} \mid \text{arrested}) =$
2. $Pr_U(\text{guilty} \mid \text{arrested}) =$

In response to objections that this focus on downtown policing is biased over-policing against downtown residents, the police department argues based on the numbers above that downtown residents are just as likely to be found guilty when they are arrested, and so the police are not making mistakes in arrests. There is simply more crime downtown than uptown.

However, another report from the oracle reveals the *total* number of people guilty of committing crimes in 2000, whether or not they were arrested by the police:

	Downtown	Uptown
People committing crimes	15000	5000

Question 2.4 (5 points): What's the probability that someone has committed a crime if one lives downtown or uptown?

1. $Pr_D(\text{guilty}) =$
2. $Pr_U(\text{guilty}) =$

§2 Gotham Policing (30 points)

Policing reform groups point out that the data in Question 2.4 implies that people living downtown are *not* more likely to be guilty of a crime than people living uptown.

So we have three points of data:

1. People are more likely to be arrested and found guilty if they live downtown:

$$Pr_D(\text{arrested \& guilty}) > Pr_U(\text{arrested \& guilty}). \tag{1}$$

2. People are *not* more or less likely to be found guilty after being arrested depending on where they live:

$$Pr_D(\text{guilty} \mid \text{arrested}) = Pr_U(\text{guilty} \mid \text{arrested}). \tag{2}$$

3. People are *not* more likely to be guilty of a crime (whether or not they are arrested) depending on where they live.

$$Pr_D(\text{guilty}) = Pr_U(\text{guilty}). \tag{3}$$

Based on (1), the police say that there is more crime downtown. Based on (2), the police say they are not mistakenly arresting people at a greater rate downtown than uptown. Based on (3), policing reform groups say that more intensive policing downtown is not justified. We want to know whose argument is better.

Let's ask: are the police more accurate for one group of people than for another?

1. The **false positive** arrest rate is the probability of getting arrested when you are not guilty.
2. The **false negative** arrest rate is the probability of *not* getting arrested when you *are* guilty.

If the false positive rate or the false negative rate are higher for downtown residents than for uptown residents, this would be evidence that there is bias in the policing.

Question 2.5 (5 points):
False Negatives: What's the probability of not getting arrested when you've committed a crime, if you live downtown/uptown?

1. $Pr_D(\text{not arrested} \mid \text{guilty}) =$
2. $Pr_U(\text{not arrested} \mid \text{guilty}) =$

False Positives: What's the probability of getting arrested when you have *not* committed a crime, if you live downtown/uptown?

1. $Pr_D(\text{arrested} \mid \text{not guilty}) =$
2. $Pr_U(\text{arrested} \mid \text{not guilty}) =$

Question 2.6 (5 points): Based on these numbers, do you think that that policing in Gotham in 2025 is biased against residents of downtown? Why or why not? (There is no uniquely correct answer here.)

3 | Gotham Policing 2049 (25 points)

24 years later, there is a new generation of Gotham residents, and a reformed police. Let's assume that the population has stayed the same in both downtown and uptown. The police department announces the following statistics, rounded to 0.01%:

	Downtown	Uptown
$Pr(\text{arrested} \mid \text{guilty})$	7.00%	7.01%
$Pr(\text{arrested} \mid \text{not guilty})$	0.51%	0.51%

The police department thus says that uptown and downtown residents are about equally likely to be arrested if guilty and if not guilty, and concludes that their policing practices are *not* biased against downtown residents.

However, a data journalism group finds that the policing practices from 2025 have affected crime rates in 2049. If a Gotham resident is a child of someone who got arrested in 2025, they are 20% likely to commit a crime in 2049. Otherwise, they are only 1.5% likely to commit a crime. In other words, one is over 13 times more likely to commit a crime if one has a parent who was arrested.

Question 3.1 (10 points): What is the probability that a Gotham resident who lives downtown/uptown is guilty of a crime in 2049?

1. $Pr_{D2049}(\text{guilty}) =$
2. $Pr_{U2049}(\text{guilty}) =$

Assume that the probability that a Gotham resident in 2049 is the child of someone who got arrested in 2025 is the same as the probability that a Gotham resident got arrested in 2025. **Hint**: Use the law of total probability and the statistics from 2025. For instance, for downtown residents:

$$\begin{aligned} &Pr_{D2049}(\text{guilty}) \\ &= Pr_{D2049}(\text{guilty} \mid \text{had a parent arrested}) Pr_{D2049}(\text{had a parent arrested}) \\ &\quad + Pr_{D2049}(\text{guilty} \mid \text{had no parent arrested}) Pr_{D2049}(\text{had no parent arrested}) \end{aligned}$$

Question 3.2 (5 points): What is the probability that a Gotham resident who lives downtown/uptown is arrested for a crime in 2049?

1. $Pr_{D2049}(\text{arrested}) =$
2. $Pr_{U2049}(\text{arrested}) =$

Hint: Use the law of total probability, the statistics from 2049, and the result from **Question 3.1**.

Question 3.3 (5 points): What is the probability that a Gotham resident who lives downtown/uptown who is arrested for a crime is guilty in 2049?

1. $Pr_{D2049}(\text{guilty} \mid \text{arrested}) =$
2. $Pr_{U2049}(\text{guilty} \mid \text{arrested}) =$

Hint: Use Bayes' Theorem.

Question 3.4 (5 points): Is policing in Gotham in 2049 fair? By the numbers, it seems so. The police are not *overpolicing* downtown based on the false positive and false negative rates: their policing rates are based on the slightly *higher* crime rate downtown. However, the higher crime rate downtown is entirely due to the overpolicing in 2025, which created a new generation more likely to commit crime. So in a sense the higher arrest rates downtown are a direct result of the overpolicing in 2025. Is there something *unfair* or *biased* about the higher arrest rate of downtown residents compared to uptown residents in 2049? Why or why not? (No unique answer here).

4 | Small Language Model (30 points)

In this set of problems, we'll build a simple auto-complete algorithm that approximates how large language models generate text. This is how the algorithm will work:

1. We'll get samples of text.
2. We'll generate new sentences, one word at a time, based on the *most likely next words*, given what has already been typed.

Specifically, the algorithm will look like this:

Algorithm

To generate a sentence

word1 word2 word3 word4 word5 word6 word7,

we follow these steps:

1. word1 = most likely first word of the sentence.
2. word2 = most likely second word, given word1.
3. word3 = most likely third word, given word2.

   ...
7. word7 = most likely seventh word, given word6.

But how do we figure out the most likely next words? We pull from a set of *sentence data*, which we gather from a user's other sentences, as well as perhaps the most common sentences. In the case of large language models, the sentence data can correspond to a large swath of the internet. In our example, we'll have just ten sentences:

	1	2	3	4	5	6	7
A	The	dog	jumped	over	our	chair	.
B	The	dog	jumped	over	a	table	.
C	The	dog	jumped	on	our	chair	.
D	The	dog	leaned	on	our	chair	.
E	The	cat	sat	on	our	bed	.
F	The	cat	jumped	on	our	bed	.
G	The	cat	sat	on	a	bed	.
H	The	cat	really	hates	our	chair	.
I	The	cat	sat	on	a	table	.
J	The	cat	leaned	over	a	table	.

To figure out the most likely next word, given the current word, we can take all the possible next words and calculate

$$\Pr(\text{next word} \mid \text{current word}).$$

For instance, suppose the current word is 'The'. We see that there are 10 places in the sentences where the word is 'The', all in column 1. Of those, the next word is 'cat' 6 of the times, and is 'dog' 4 of the times. So

$$\Pr(\text{word2 is 'cat'} \mid \text{word1 is 'The'}) = 6/10.$$
$$\Pr(\text{word2 is 'dog'} \mid \text{word1 is 'The'}) = 4/10.$$

Question 4.1 (5 points). The most likely first word of the sentence is 'The' (that's the *only* possible first word of the sentence, given the data). We saw above that the most likely second word of the sentence is 'cat', given that the first word is 'The'. Calculate the most likely third word of the sentence. Do this by calculating each of the following:

$$\Pr(\text{word3 is 'jumped'} \mid \text{word2 is 'cat'}) =$$
$$\Pr(\text{word3 is 'sat'} \mid \text{word2 is 'cat'}) =$$
$$\Pr(\text{word3 is 'leaned'} \mid \text{word2 is 'cat'}) =$$
$$\Pr(\text{word3 is 'really'} \mid \text{word2 is 'cat'}) =$$

Then pick the word with the highest probability. **Note:** in the case of *ties* for the most probable word, you can choose either word.

Question 4.2 (10 points). Suppose you use this algorithm to generate every word of a sentence from the data above. Write down the sentence generated:

1	2	3	4	5	6	7

Question 4.3 (5 points). Autocomplete *suggests* words to the user, but the user can always write different words.

1. Suppose that the user writes 'dog' instead of 'cat' as the second word, so that the sentence begins "The dog…". Complete the rest of the sentence:
   The dog

2. Suppose that the user writes 'leaned' as the third word and 'over' as the fourth word, so that the sentence begins "The cat leaned over…". Complete the rest of the sentence.
   The cat leaned over

§4 Small Language Model (30 points)

Probabilistic Auto-Complete

Let's inch closer to something that looks like a large language model. Suppose we chose the next word not just by picking the most likely next word, but *randomly* based on the probabilities. For instance, instead of simply choosing 'cat' for the second word, we choose 'cat' 60% of the time and 'dog' 40% of the time. Then the probability of generating a sentence

$$\texttt{word1} \quad \texttt{word2} \quad \texttt{word3} \quad \texttt{word4} \quad \texttt{word5} \quad \texttt{word6} \quad \texttt{word7}$$

can be found by multiplying all of the probabilities of getting the next word in the sentence at each step:

$$\Pr(\texttt{word 1})\Pr(\texttt{word 2} \mid \texttt{word 1})\Pr(\texttt{word 3} \mid \texttt{word 2}) \cdots \Pr(\texttt{word 7} \mid \texttt{word 6}).$$

Question 4.4 (5 points). Suppose we choose the next word randomly based on the probabilities. What is the probability of generating each of the following sentences?

1. The dog sat on our table.
2. The cat leaned over our chair.
3. The dog leaned over a table.

NOTE: sentences 2 and 3 corrected on Nov 16.

Discussion. In fact, the cat hates our chair and would never do anything on or over it. Sentence H says this explicitly, and furthermore none of the sentences have the cat doing anything with our chair. However, as you saw in the first problem, the most likely generated sentence is one about a cat sitting on our chair.

Question 4.5 (5 points). How might you modify the algorithm to try to avoid cases where the "most likely" generated sentences might *contradict* the sentences in the sentence data, like in the algorithm we have been considering? (No uniquely correct answer here, but please do this based on your own reasoning, not by searching the web for the (massive body of) existing ways to improve on such models).

Please specify your use of outside assistance on this problem set. For instance, if you used AI to check your answers, write which questions, and indicate if AI was able to correct your answers. If you did the problem set together with another student, please indicate who. If you used the internet, please indicate which websites and how. If you *did not* use AI, the internet, or collaboration, please also indicate this with a statement saying "I did not use AI, the internet, or discussion with other students to complete this assignment" (appropriately modified to reflect what you did and did not use).