Chris Piech, Jerry Cain, Freya, and Doris
CS 109

CS109 Quiz #1

Take-Home Quiz information

Each quiz will be a 46.5-hour open-book, open-note exam. We have designed this quiz to approximate about 1-2 hours of active work (*before* typesetting).

- You can submit multiple times; we will only grade the last submission you submit before 1:00pm (Pacific time) on Friday, February 5th. No late submissions can be accepted. When uploading, please assign pages to each question.
- You should upload your submission as a PDF to Gradescope. We provide a LaTeX template if you find it useful, but we will accept any legible submission. You may also find the CS109 Probability LaTeX reference useful: `https://www.overleaf.com/project/5f650a577489e90001f065be`
- Course staff assistance will be limited to clarifying questions of the kind that might be allowed on a traditional, in-person exam. If you have questions during the exam, please ask them as private posts via our discussion forum. We will not have any office hours for answering quiz questions during the quiz, and we can't answer any questions about course material while the quiz it out.
- **For each problem, briefly explain/justify how you obtained your answer at a level such that a future CS109 student would be able to understand how to solve the problem. If it's not fully clear how you arrived at your answer, you will not receive full credit.** It is fine for your answers to be a well-defined mathematical expression including summations, products, factorials, exponents, and combinations, unless the question *specifically* asks for a numeric quantity or closed form. Where numeric answers are required, fractions are fine.

Honor Code Guidelines for Take-Home Quizzes

This exam must be completed individually. It is a violation of the Stanford Honor Code to communicate with any other humans about this exam (other than CS109 course staff), to solicit solutions to this exam, or to share your solutions with others.

The take-home exams are open-book: open lecture notes, handouts, textbooks, course lecture videos, and internet searches for conceptual information (e.g., Wikipedia). Consultation of other humans in any form or medium (e.g., communicating with classmates, asking questions on sites like Chegg or Stack Overflow) is prohibited. All work done with the assistance of any external material in any way (other than provided CS109 course materials) must include citation (e.g., "Referred to Wikipedia page on X for Question 2."). Copying solutions is unacceptable, even with citation. If by chance you encounter solutions to the problem, navigate away from that page before you feel tempted to copy.

If you become aware of any Honor Code violations by any student in the class, your commitments under the Stanford Honor Code obligate you to inform course staff. ***Please remember that there is no reason to violate your conscience to complete a take-home exam in CS109.***

I acknowledge and accept the letter and spirit of the Honor Code:

Name (typed or written): _____

1 The Freya Checkers Invitational [30 points]

Don't tell Stanford, but Chris and Jerry each have side jobs as professional Checkers players, and they're set to compete against one another and play a total of 9 games in the Freya Checkers Invitational. Each game results in either a win for Chris, a win for Jerry, or a tie. A win is worth 2 points, a tie is worth 1 point, and a loss is worth 0.

a. (7 points) How many different ways can Chris arrive at 2 wins, 5 draws, and 2 losses if all 9 games are played?

Answer.

Any 2 of the 9 games can be wins for Chris, and any 2 of the remaining 7 can be wins for Jerry. That means the total number of ways Chris can compile 2 wins, 5 draws, and 2 losses is $\binom{9}{2}\binom{7}{2}\binom{5}{5} = 756$. Another approach computes the number of ways one can permute 2 indistinguishable W's, 2 indistinguishable L's, and 5 indistinguishable T's to form a string of length 9. That computation is just a trinomial, which in this case would be $\binom{9}{2,5,2}$. That, not surprisingly, evaluates to 756 as well.

In addition to providing an expression above,
please compute an integer answer: 756

b. (8 points) What is the probability that Chris gets exactly 10 points when all 9 games are played? Assume that each game results in a win for Chris, a win for Jerry, or a tie with equal likelihood, and that all games are independent.

Answer.

The total number of possible outcomes is $3^9 = 19683$, since each of the nine games results (independently) in a win, tie, or loss for Chris. The number of ways Chris can get 10 points equals the number of ways Chris can get 5 wins, plus the number of ways Chris can get 4 wins and 2 ties, plus the number of ways Chris can get 3 wins and 4 ties. More generally, the number is given as the sum of the ways Chris can win k games and tie 10 - 2k others, for k = 1, 2, 3, 4, and 5. (O wins for Chris isn't an option, since you can't tie 10 games when you only play 9.)

That means the number of ways Chris can get 10 points is:

$$\sum_{w=1}^{5} \binom{9}{w,\, 10-2w,\, w-1} = \sum_{w=1}^{5} \binom{9}{w}\binom{9-w}{w-1} = 2907$$

So, the probability Chris gets precisely 10 points after 9 games is $\frac{2907}{19683}$, or 0.1477. (I wrote a short Python program to compute both the numerator, the denominator, and their ratio.)

In addition to providing an expression above,
please compute a numeric answer: 0.1477

c. (7 points) Because Freya is in a hurry to get Chris home, Chris and arch-nemesis Jerry agree to play a best-of-9 tournament, where play stops when one accumulates 10 points, or they've played all 9 games, whichever comes first. How many different ways can Chris win the tournament by a score of 10 to 8?

Answer.

If the score ended up being 10 to 8, then Chris and Jerry played all 9 games. The score going in to the final score must have been either 8-8 (in which case Chris must win the final game) or 9-7 in Chris's favor (in which case Chris and Jerry must tie in the final game). That's two mutually exclusive events, but each event size can be computed via the same approach I took in part b.

- The number of ways Chris can accumulate 8 points in 8 games equals the number of ways Chris can get 4 wins and 0 ties, plus the number of ways Chris can get 3 wins and 2 ties, plus the number of ways Chris can get 2 wins and 4 ties, plus the number of ways Chris can get 1 win and 6 ties, plus the number of ways Chris can get 0 wins and 8 ties. How's that for exhaustive? That sum is given by given by:

$$\sum_{w=0}^{4} \binom{8}{w,8-2w,w} = \sum_{w=0}^{4} \binom{8}{w}\binom{8-w}{w} = 1107.$$

- The number of ways Chris can accumulate 9 points in 8 games equals the number of ways Chris can get 4 wins and 1 tie, plus the number of ways Chris can get 3 wins and 3 ties, plus the number of ways Chris can get 2 wins and 5 ties, but the number of ways Chris can get 1 win and 7 ties. I'm being similarly exhaustive, just to be careful. That sum is given by given by:

$$\sum_{w=1}^{4} \binom{8}{w,9-2w,w-1} = \sum_{w=1}^{4} \binom{8}{w}\binom{8-w}{w-1} = 1016.$$

Lo and behold, there are 1107 + 1016 = 2123 ways for Chris to win 10 to 8 in 9 games.

In addition to providing an expression above,
please compute an integer answer: 2123

d. (8 points) Eager to get Chris home to Freya even more quickly, Jerry proposes they keep playing until one player's point total exceeds the other's by 3 or more points. Assume that Chris and Jerry each win a single game with probability of 0.25 and 0.15, respectively, and they tie with probability 0.6. What is the probability that Chris eventually wins?

Answer.

Let's let $p_c = 0.25$, $p_j = 0.15$, and $p_t = 0.60$. From those three values, we need to compute the probability that Chris wins before Jerry does after a sequence of zero or more ties. If we let p_{cwbj} represent that probability **C**hris **W**ins **B**efore **J**erry, it can be computed using as:

$$p_{cwbj} = p_c + p_t p_c + p_t^2 p_c + \ldots = p_c \sum_{k=0}^{\infty} p_t^k = p_c \frac{1}{1 - p_t} = 0.25 \frac{1}{1 - 0.60} = 0.625$$

Intuitively, this shouldn't surprise you, since that's syncs well with the 5 to 3 ratio between Chris's and Jerry's win probabilities: $\frac{p_c}{p_c + p_j} = \frac{0.25}{0.25+0.15} = \frac{0.25}{0.40} = 0.625$.

If we define p_{cw} to be the probability that Chris eventually wins the tournament, then p_{cw} can be recursively defined to be:

$$p_{cw} = p_{cwbj}^2 + 2 p_{cwbj} (1 - p_{cwbj}) p_{cw}$$

Basically, Chris either wins the tournament by winning two games outright—that's the p_{cwbj}^2—or he wins one, loses one (or loses one, wins one), and continues to win the tournament from what is effectively the starting line—that's the $2 p_{cwbj} (1 - p_{cwbj}) p_{cw}$ part.

Solving for p_{cw}, we get:

$$\begin{aligned} p_{cw} &= p_{cwbj}^2 + 2 p_{cwbj} (1 - p_{cwbj}) p_{cw} \\ &= 0.625^2 + 2 \cdot 0.625 \cdot 0.375 p_{cw} \\ &= 0.390625 + 0.46875 p_{cw} \\ 0.53125 p_{cw} &= 0.390625 \\ p_{cw} &= \frac{0.390625}{0.53125} = 0.7353 \end{aligned}$$

In addition to providing an expression above,
please compute a numeric answer: 0.7353

2 Malware Detection [30 points]

Jerry's nephew, Andrew, just started high school this past September, and because all course instruction is online, his high school provided him with an old laptop so he could Zoom in for class. To better protect its own equipment from malware and viruses, the school installed two browser extensions to scrutinize all web downloads. Each download is either safe (event S) or unsafe (event S^C), and all downloads are examined by both extensions. The first extension marks the download as either safe (event A) or unsafe (event A^C), and the second extension marks the download as safe (event B) or unsafe (event B^C).

Assume that 94% of Andrew's web downloads are safe, and that:

- the first browser extension accurately marks unsafe downloads as unsafe with probability 0.93, but improperly marks safe downloads as unsafe with probability 0.04.
- the second browser extension accurately marks unsafe downloads as unsafe with probability 0.85, but improperly marks safe downloads as unsafe with probability 0.02.

Assume that given a download is safe, the two browser extensions independently mark that download as safe. Similarly, the two browser extensions independently mark unsafe downloads as unsafe.

a. (6 points) Assuming a downloaded document is safe, what is the probability that at least one of the two extensions marks the download as unsafe?

Answer.

Let's jot down everything we've been given:

$$P(S) = 0.94, \text{ so } P(S^C) = 1 - P(S) = 0.06$$

$$P(A^C|S^C) = 0.93, \text{ so } P(A|S^C) = 1 - P(A^C|S^C) = 0.07$$
$$P(A^C|S) = 0.04, \text{ so } P(A|S) = 1 - P(A^C|S) = 0.96$$

$$P(B^C|S^C) = 0.85, \text{ so } P(B|S^C) = 1 - P(B^C|S^C) = 0.15$$
$$P(B^C|S) = 0.02, \text{ so } P(B|S) = 1 - P(B^C|S) = 0.98$$

$$P(AB|S) = P(A|S)P(B|S)$$
$$P(A^C B^C|S^C) = P(A^C|S^C)P(B^C|S^C)$$

The probability that one or both extensions mark a safe download as unsafe is more easily computed as 1 minus the probability that **both** mark a safe download as safe.

$$\begin{aligned} P((AB)^C|S) &= 1 - P(AB|S) \\ &= 1 - P(A|S)P(B|S) \\ &= 1 - 0.96 \cdot 0.98 \\ &= 0.0592 \end{aligned}$$

The second line above follows from the first because A and B are conditionally independent given S.

In addition to providing an expression above, please compute an numeric answer: 0.0592

b. (6 points) Assuming a downloaded document is safe, what is the probability that exactly one of the two extensions marks the download as unsafe?

Answer.

We're still in the world of safe downloads, so:

$$\begin{aligned} P(A^C B|S) + P(AB^C|S) &= P(A^C|S)P(B|S) + P(A|S)P(B^C|S) \\ &= 0.04 \cdot 0.98 + 0.96 \cdot 0.02 \\ &= 0.0584 \end{aligned}$$

Note that if A and B are conditionally independent given S, then so are A and B^C, as are A^C and B.

In addition to providing an expression above, please compute an numeric answer: 0.0584

c. (10 points) Given that both extensions mark the download as safe, what is the probability that the download is unsafe?

Answer.

We are interested in $P(S^C|AB)$, which according to Bayes' Theorem is $\frac{P(AB|S^C)P(S_C)}{P(AB)}$. Some of these probabilities are given, but others have to be computed from scratch.

$$\begin{aligned}
P(AB|S^C) &= P(A|S^C) \cdot P(B|S^C) \\
&= 0.07 \cdot 0.15 \\
&= 0.0105 \\
P(S^C) &= 0.06 \\
P(AB) &= P(AB|S)P(S) + P(AB|S^C)P(S^C) \\
&= P(A|S)P(B|S)P(S) + 0.0105 \cdot 0.06 \\
&= 0.96 \cdot 0.98 \cdot 0.94 + 0.0105 \cdot 0.06 \\
&= 0.884982
\end{aligned}$$

Now we can compute $P(S^C|AB)$ as $\frac{0.0105 \cdot 0.06}{0.884982}$, or 0.000712. That's less than $\frac{1}{10}^{th}$ of a percent, which means the probability an unsafe download goes undetected is super small.

In addition to providing an expression above, please compute an numeric answer: 0.000712

d. (8 points) Are the unconditioned events where the two extensions mark a download as safe independent? Why or why not?

Answer.

Nope. Intuitively, you would expect the second extension is more likely to flag a download as safe when the first extension does, and vice versa. Mathematically, we examine $P(AB)$, which we've already computed to be 0.88498, and compare that to $P(A)P(B)$. We haven't computed $P(A)$ or $P(B)$ yet, so let's do that now:

$$
\begin{aligned}
P(A) &= P(A|S)P(S) + P(A|S^C)P(S^C) \\
&= 0.96 \cdot 0.94 + 0.07 \cdot 0.06 \\
&= 0.9066 \\
P(B) &= P(B|S)P(S) + P(B|S^C)P(S^C) \\
&= 0.98 \cdot 0.94 + 0.15 \cdot 0.06 \\
&= 0.9302 \\
P(A)P(B) &= 0.9066 \cdot 0.9302 = 0.84332 \neq 0.884982 = P(AB)
\end{aligned}
$$

That's a solid mathematical defense that A and B are **not independent**.

3 Doris and Inventory Audits [30 points]

Jerry's Boston Terrier, Doris, is a San Francisco-based accountant, and much of her work has her auditing retail companies—literally showing up unannounced–to confirm their in-house inventory matches what their inventory software claims. If she finds evidence to suggest there are widespread discrepancies between physical in-store inventory and what's shared with the Internal Revenue Service (that's the U.S. Federal Tax Bureau) via inventory reports, she informs the IRS that a more thorough audit is warranted.

Assume Doris is auditing a company that maintains separate records for 4500 different items, 4400 of which are accurate and 100 of which are inaccurate. Rather than examining all 4500 records, Doris arbitrarily samples 100 of them (without replacement, 25 per paw). Let X be the random variable counting the number of inaccurate record found among all 100 samples, assuming all samples are equally likely.

a. (10 points) Compute $p_X(x)$, which is the probability mass function of X as defined above.

Answer.

There are a total of $\binom{4500}{100}$ samples, and:

- there are $\binom{4400}{100}\binom{100}{0}$ ways to sample 100 records and get $X = 0$ inaccuracies.
- there are $\binom{4400}{99}\binom{100}{1}$ ways to sample 100 records to get just $X = 1$ inaccuracy.
- there are $\binom{4400}{98}\binom{100}{2}$ ways to sample 100 records to get 2 inaccuracies.

  ...
- there are $\binom{4400}{2}\binom{100}{98}$ ways to sample 100 records to get 98 inaccuracies.
- there are $\binom{4400}{1}\binom{100}{99}$ ways to sample 100 records to get 99 inaccuracies.
- there are $\binom{4400}{0}\binom{100}{100}$ ways to sample 100 records to get all the inaccurate ones.

I think we can safely generalize::

$$p_X(x) = \frac{\binom{4400}{100-x}\binom{100}{x}}{\binom{4500}{100}}, 0 \leq x \leq 100.$$

b. (6 points) Compute the expected value of X, i.e. E[X]. You should express your answer as a sum of a finite number of terms (which you share as part of your answer), but then write a short Python program (which need not be included) to compute the summation and present a single numeric value.

Answer.

We really just get to tap the definition of expectation here and plug in our support and $p_X(x)$, like so:

$$\begin{aligned} E[X] &= \sum_{x=0}^{100} x \cdot p_X(X) \\ &= \sum_{x=0}^{100} x \cdot \frac{\binom{4400}{100-x}\binom{100}{x}}{\binom{4500}{100}} \\ &= \frac{20}{9} = 2.22222 \end{aligned}$$

I didn't analytically arrive at $\frac{20}{9}$ myself. I instead used the Python command line to compute the sum for me.

In addition to providing an expression above, please compute an numeric answer: 2.22222

To guard against some unlucky sampling, Doris stops if the first sample produces either 0 (she takes that as conclusive) or three or more inaccurate records. If she finds either one or two inaccurate records in her initial sample of 100, she draws a new sample of 100 from the original 4500, just as she did originally, and counts the number of inaccuracies the same exact way.

c. (7 points) Let Y be the random variable counting the number of inaccuracies counted between the first and, if needed, second audits. What is $P(Y = 2)$?

Answer.

There are two ways that Doris can arrive at a total of two inaccuracies.

- Doris might find 2 inaccurate records in the first sample, and then find 0 in the second.
- Doris might find 1 inaccuracy in the first sample, and then 1 more in the second.

Those two scenarios above are mutually exclusive, so:

$$P(Y = 2) = p_Y(2) = p_{X_1}(2)p_{X_1}(0) + p_{X_1}(1)p_{X_2}(1)$$

Here, I use X_1 and X_2 to denote independent random variables modeling the two audits, each of which is really just X with a subscript to be clear which audit they're modeling.

$$\begin{aligned} P(Y=2) &= \boxed{p_{X_1}(2)p_{X_2}(0) + p_{X_1}(1)p_{X_2}(1)} \\ &= \frac{\binom{4400}{98}\binom{100}{2}}{\binom{4500}{100}} \cdot \frac{\binom{4400}{100}\binom{100}{0}}{\binom{4500}{100}} + \frac{\binom{4400}{99}\binom{100}{1}}{\binom{4500}{100}} \cdot \frac{\binom{4400}{99}\binom{100}{1}}{\binom{4500}{100}} \\ &= 0.08551 \end{aligned}$$

I arrived at the value of 0.08551 using the Python command line. Even though we expected a numeric result, if you arrived at the boxed expression above or something like it, you'll get the majority of the points.

In addition to providing an expression above,
please compute an numeric answer: 0.08551

d. (7 points) Compute $E[Y]$ in terms of $E[X]$. Your answer should be framed in terms of $E[X]$ instead of the exact value you computed for part b, just in case your answer to part b was incorrect.

Answer.

Let's formally define $Y = X_1 + (p_{X_1}(1) + p_{X_1}(2))X_2$. Note that $(p_{X_1}(1) + p_{X_1}(2))$ is a constant, not a random variable, so:

$$\begin{aligned} Y &= X_1 + (p_{X_1}(1) + p_{X_1}(2))X_2 \\ E[Y] &= E[X_1] + (p_{X_1}(1) + p_{X_1}(2)) \cdot E[X_2] \\ &= E[X] + (p_{X_1}(1) + p_{X_1}(2)) \cdot E[X] \\ &= \boxed{(1 + p_{X_1}(1) + p_{X_1}(2)) \cdot E[X]} \\ &= (1 + \frac{\binom{4400}{99}\binom{100}{1}}{\binom{4500}{100}} + \frac{\binom{4400}{98}\binom{100}{2}}{\binom{4500}{100}}) \cdot E[X] \\ &= (1 + 0.23957 + 0.27289) \cdot E[X] \\ &= 1.51246 \cdot E[X] \\ &= 1.51246 \cdot 2.22222 \\ &= 3.36102 \end{aligned}$$

I computed $p_{X_1}(1)$ and $p_{X_1}(2)$ via Python's command line. The boxed expression above, however, is enough to get full credit, since we asked that $E[Y]$ be framed in terms of $E[X]$, and we aren't concerned—at least for part d—that you numerically evaluated $p_{X_1}(1)$ and $p_{X_1}(2)$. The boxed part is the interesting result and demonstrates an understanding of linearity of expectation.

In addition to providing an expression above,
please compute an numeric answer: 3.36102