Jerry Cain Spring 2023
CS109

CS109 Week 4 Exam Solution

This is a closed calculator/computer exam. You are, however, permitted to consult the two double-sided sheets of notes you've prepared ahead of time. You're otherwise not permitted to refer to any other notes.

In the event of an incorrect answer, any explanation you provide of how you obtained your answer can potentially allow us to give you partial credit for a problem. It is fine for your answers to include summations, products, factorials, exponents, and combinations unless stated otherwise.

I acknowledge and accept the letter and spirit of the honor code.

Signature: ______________________________

Last Name [print]: ______________________________

First Name [print]: ______________________________

SunetID [i.e., your @stanford.edu email]: ______________________________

SUID [i.e., your seven or eight-digit student ID number]: ______________________________

1 Summer Study in London [15 points]

Kathleen is planning to spend the summer at Imperial College London (on scholarship!) to partake in an intense three months of one-on-one tutorials with some of the world's leading experts in machine learning. Kathleen's scholarship allows for her to take a total of 7 tutorials, each of which meets just once per week for precisely one hour.

There are so many tutorials to choose from, though! Imperial offers a total of 25 different tutorials—five on Mondays, five on Tuesdays, five on Wednesday, five on Thursday, and five on Fridays. Fortunately, all tutorials are offered at different times, so Kathleen can choose any 7 of the 25 she wants without introducing any conflicts.

a. [2 points] How many different ways can Kathleen choose her seven tutorials if she can choose the seven without restrictions.

> **Solution.** We have $\binom{25}{7}$ ways of choosing seven tutorials from twenty-five.

b. [5 points] How many different ways can Kathleen choose her seven tutorials so that she has at least one tutorial every weekday?

> **Solution.** In order to choose seven tutorials such that we have at least one tutorial on each weekday, we can consider the following two cases:
>
> - We have two days with two tutorials and three with one.
> - We have one day with three tutorials and four with one.
>
> Since these cases are mutually exclusive, we can sum the counts of each to get the total number. So, we have
>
> $$\binom{5}{2} \cdot \binom{5}{2}^2 \binom{5}{1}^3 + \binom{5}{1} \cdot \binom{5}{3} \binom{5}{1}^4$$
>
> different ways. To explain a bit further, in each summand we first count the number of ways of picking which days have only one tutorial and which have more, then we multiply by the number of ways of choosing from the five available tutorials for each day.

In order to be sure of your answer to part b, you've decided to count the same number a different way: by using the inclusion-exclusion method to count the number of ways Kathleen can have one or more days without any tutorials. Once you count that, you can subtract your answer here from your answer to part a to replicate the same number you derived for part b.

Let A_i be the set of all schedules that include at least one tutorial on the i^{th} weekday, where A_1 counts the number of schedules that include Monday tutorials, A_2 counts the number of schedules that include Tuesday tutorials, and so forth. To count the number of schedules that give Kathleen off one or more weekdays per week, we need to compute the following:

$$\left|A_1^C \cup A_2^C \cup A_3^C \cup A_4^C \cup A_5^C\right|$$

By applying the inclusion-exclusion principle to this particular problem, we arrive at:

$$\left|A_1^C \cup A_2^C \cup A_3^C \cup A_4^C \cup A_5^C\right| = \sum_i \left|A_i^C\right| - \sum_{i<j} \left|A_i^C \cap A_j^C\right| + \sum_{i<j<k} \left|A_i^C \cap A_j^C \cap A_k^C\right|$$

c. [2 points] Clearly but briefly explain why we don't need to include the four-way or five-way intersections of complements for this particular problem statement?

> **Solution.** The cardinalities of the four- and five-way intersections are 0 because we cannot allocate all seven tutorials to fewer than two days of the week. So, not including them doesn't affect the quantity calculated above.

d. [6 points] Present an expression that counts the number of ways that Kathleen can have one or more weekdays off per week. Your expression that include terms and products that are consistent with the formula for $\left|A_1^C \cup A_2^C \cup A_3^C \cup A_4^C \cup A_5^C\right|$ above. Restated, you can't write down your answer to part a minus your answer to part b. Instead, your approach should present combinatorial expressions for $\left|A_1^C\right|$, $\left|A_1^C \cap A_2^C\right|$, and $\left|A_1^C \cap A_2^C \cap A_3^C\right|$ so you can use them to build up your final answer for the full union of all five complements.

> **Solution.** For each day that we cannot schedule tutorials, we have five fewer tutorials from which we can choose. We thus have that
>
> - $\sum_i \left|A_i^C\right| = \sum_i \binom{20}{7} = \binom{5}{1}\binom{20}{7}$
> - $\sum_{i<j} \left|A_i^C \cap A_j^C\right| = \sum_{i<j} \binom{15}{7} = \binom{5}{2}\binom{15}{7}$
> - $\sum_{i<j<k} \left|A_i^C \cap A_j^C \cap A_k^C\right| = \sum_{i<j<k} \binom{10}{7} = \binom{5}{3}\binom{10}{7}$,
>
> so we can write
>
> $$\left|A_1^C \cup A_2^C \cup A_3^C \cup A_4^C \cup A_5^C\right| = \binom{5}{1}\binom{20}{7} - \binom{5}{2}\binom{15}{7} + \binom{5}{3}\binom{10}{7}.$$

2 Combinatorial Proofs [5 points]

Consider the following combinatorial identity for all integers $n \geq 2$:

$$\sum_{k=2}^{n} \binom{k}{2}\binom{n}{k}^2 = \binom{n}{2}\binom{2n-2}{n-2}$$

Present a **combinatorial** proof of the above identity, without relying on any algebra. As a hint, consider the selection of n members for a Stanford committee, where the n members are selected from a group of n professors and n students, where the co-chairs of the committee must both be professors.

> **Solution.** The right-hand side consists of two parts. The first term, $\binom{n}{2}$, is the number of ways for picking the 2 professors who will be co-chairs out of the n professors we can select from. After we have selected the two professors who will be co-chairs, the second term, $\binom{2n-2}{n-2}$ is the number of ways we can select the remaining $n-2$ people who will be on our committee from the total $2n-2$ remaining candidates. There are $2n-2$ remaining candidates since we have n professors and n students (so $2n$ people to choose from total), but we already selected 2 professors, so we remove them from our remaining candidate pool.
>
> On the left hand side, we can have k represent the possible number of professors we choose for our committee. We know that we have a minimum of 2 professors (since the committee has two professors as the co-chairs), but we could have up to n professors total (in the case of a committee of only professors), hence we have a sum from 2 to n.
>
> For each possible option of k professors, we will choose 2 of the k professors on the committee to be the co-chairs, and there are $\binom{k}{2}$ ways to do so. We have to select the k professors that we will place on the committee, so we have our first $\binom{n}{k}$ term representing the number of ways to do so. Finally, with our n students, all but k of them can be part of the committee (since k of the n slots are taken up by professors). Thus, we have $\binom{n}{k}$ ways of selecting the students who will not be part of the committee. In conclusion, for every possible option of k professors on the committee, we have $\binom{k}{2}\binom{n}{k}^2$ ways of selecting the committee.
>
> **Note:** While our solution here is fairly long, we did not expect student answers to be this verbose. We have provided extra explanations for the sake of student understanding. We gave full credit to every answer that had a reasonable explanation for each term on both the left and right hand sides. For example, "The second term, $\binom{2n-2}{n-2}$ is the number of ways we can select the remaining $n-2$ people who will be on our committee from the total $2n-2$ remaining candidates." is all we would need as explanation for the $\binom{2n-2}{n-2}$ term.

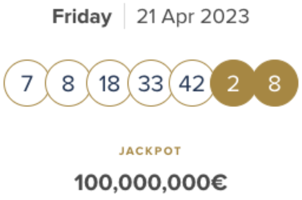
3 The EuroMillions Lottery [15 points]

EuroMillions is a lottery where residents from participating European countries can play with hopes of winning the jackpot, which is generally tens (and often hundreds) of millions of euros. When purchasing a single lottery ticket, the player must

- choose five distinct integers between 1 and 50, inclusive, as your **main** numbers, and independently...
- choose two distinct integers between 1 and 12, inclusive (though the numbers chosen here may repeat one or two numbers chosen as part of the main five). These additional two numbers are known as the **Lucky Stars**.

Unsurprisingly, you win the jackpot (or at least share it in the event of multiple winners) when your five numbers match those drawn in the lottery and your two Lucky Stars match those drawn in the lottery as well. The order of the numbers are drawn doesn't matter.

Here are the results of EuroMillions from last Friday:



Friday | 21 Apr 2023

7 8 18 33 42 2 8

JACKPOT

100,000,000€

Note that last Friday's lottery just happened to draw an 8 two times. That's totally legitimate, since one of the 8's was drawn as part of the first five numbers and the second 8 was drawn separately as one of the two Lucky Stars.

a. [2 points] Assuming all size-5 subsets of $\{1, 2, \ldots, 50\}$ are equally likely to be drawn and all size-2 subsets of $\{1, 2, \ldots, 12\}$ are equally likely to be drawn (independently of the size-5 subsets), what is the probability that you match all seven numbers?

> **Solution.** This is 1 out of all ways to choose the main numbers and the lucky stars, so:
>
> $$\frac{1}{\binom{50}{5} \cdot \binom{12}{2}}$$
>
> because there are $\binom{50}{5}$ to choose the main numbers and $\binom{12}{2}$ ways to choose the lucky stars. Notice the denominator need has both co

b. [4 points] What is the probability that you match precisely three of the five main numbers given that you match at least one?

> **Solution.** With conditional probability,
>
> a. Let G be the binary event of **one or more** of the 5 main numbers.
>
> b. Let M_3 be the binary event of matching **exactly three** of the 5 main numbers.

> First,
>
> $$P(G) = 1 - \frac{\binom{45}{5}}{\binom{50}{5}}$$
>
> $$P(M_3) = \frac{\binom{5}{3} \cdot \binom{50-5}{5-3}}{\binom{50}{5}}$$
>
> Then,
>
> $$P(M_3|G) = \frac{P(M_3, G)}{P(G)} = \frac{P(M_3)}{P(G)} = \frac{\frac{\binom{5}{3} \cdot \binom{50-5}{5-3}}{\binom{50}{5}}}{1 - \frac{\binom{45}{5}}{\binom{50}{5}}} \approx 0.011$$

c. [5 points] What is the probability that all seven numbers are different? Note that last Friday's drawing is legal, but because the 8 is repeated twice (once among the five and a second time as one of the two Lucky Stars), that particular outcome wouldn't contribute to the probability you're computing here.

> **Solution.** **Solution 1:** Fix the last 2 numbers. We get,
>
> $$P = \frac{\binom{48}{5}\binom{12}{2}}{\binom{50}{5}\binom{12}{2}} = \frac{\binom{48}{5}}{\binom{50}{5}}$$
>
> because we need all ways the main numbers could avoid the two lucky stars that we have fixed.
>
> Common errors
>
> a. (Minor Error):
>
> $$P = \frac{\binom{38}{5}}{\binom{50}{5}}$$
>
> This solution assumes the 5 main numbers cant be 1-12
>
> b. (Minor Error):
>
> $$\frac{\binom{50}{5}\binom{7}{2}}{\binom{50}{5}\binom{12}{2}}$$
>
> This solution assumes the 5 main numbers are all from 1-12.
>
> **Solution 2:** Sum over all cases where the main numbers could have 1-5 of the lucky number candidates.
>
> $$P = \frac{\sum_{i=1}^{5}\binom{12}{i}\binom{50-12}{5-i}\binom{12-i}{2}}{\binom{50}{5}\binom{12}{2}}$$
>
> Common mistakes
>
> a. (Numerical error) Using $\binom{50-i}{5-i}$ instead of $\binom{50-12}{5-i}$
>
> b. (Minor error) Summing only for $i = 0, 1, 2$.
>
> **Solution 3:** (Grader: my solution but it wasn't popular :P).
>
> Approach: fully ordered event and sample space. Pretend the two lucky numbers are already drawn.
>
> $$P = \prod_{i=0}^{4} \frac{48-i}{50-i}$$
>
> i.e, just "miss" the two lucky numbers 5 times.
>
> **Solution 4**:
>
> $$\frac{\binom{50}{5} * \binom{12}{2} - 2 * \binom{49}{2} * \binom{12}{2} + \binom{48}{3} * \binom{12}{2}}{\binom{50}{5} * \binom{12}{2}}$$
>
> Common errors

a. (Minor Error): Forgot the (12 choose 2) in the numerator

General Major Errors

a. Ordered-ness

$$\frac{48 \cdot 47 \cdot 46 \cdot 45 \cdot 44 \cdot 12 \cdot 11}{\binom{50}{5}\binom{12}{2}}$$

Here, the numerator is ordered, yet the denominator is unordered. This overcounts.

b. Ignore dependency

$$\frac{48 \cdot 47 \cdot 46 \cdot 45 \cdot 44 \cdot 12 \cdot 11}{50^5 12^2}$$

Because the main/lucky numbers are distinct so each drawing of a number is dependent on the previous one.

d. [4 points] Present a pseudo-Python implementation of a program that estimates the expected number of **consecutive** lotteries needed before all 50 main numbers appear at least once **and** all 12 Lucky Star numbers appear once as well. Your implementation should compute this estimation by running the same experiment 100000 times, where each experiment simulates the lottery process as many times as needed until all numbers show up. Your implementation doesn't need to be all that efficient, but it needs to be correct. You needn't use `scipy` or `numpy` or anything fancy unless you really want to. We expect you to rely on simple Python data structures like Python lists and/or dictionaries and the `random.choice` function, as illustrated below.

Again, don't worry about syntax. We're perfectly happy with pseudo-code as long as there's a clear Python equivalent for each line of your implementation.

```
from random import choice
# choice(range(1, 13)) returns a random integer between 1 and 12 inclusive,
# all being equally likely
def estimate_expectation():
    num_lotteries_needed = 0
    for count in range(100000):
        num_lotteries_needed += run_one_simulation()
    return num_lotteries_needed/100000

def run_one_simulation():
    # place your implementation in the space below
```

Solution.

```
def run_one_simulation():
    # Setup sets.
    seen_main = set()
    seen_lucky = set()
    count = 0

    # While not all main numbers seen and not all lucky numbers seen.
    while len(seen_main) != 50 or len(seen_lucky) != 12:

        # Containers to simulate a lottery.
```

```
            main = []
            lucky = []

            # Choose the main and lucky numbers distinctly.
            while len(main) != 5:
                cand = choice(range(1, 51))
                if cand not in main:
                    main.append(cand)
            while len(lucky) != 2:
                cand = choice(range(1, 13))
                if cand not in lucky:
                    lucky.append(cand)

            # Mark numbers as seen, remember set add only adds unique values.
            for num_main in main:
                seen_main.add(num_main)
            for num_lucky in lucky:
                seen_lucky.add(num_lucky)

            # Increment the number of sample we have drawn.
            count += 1
        return count
```

Common Errors.

a. Simulating only 1 lottery and returning any variation of counting numbers on that.

b. While condition checks `len(seen_main) != 50 and len(seen_lucky) != 12`. This is wrong because we could have played enough lotteries to see all 12 lucky numbers, but not enough to see all 50 main numbers.

c. Not checking `if cand not in main`. This check allows us to sample without replacement which is needed to simulate the "distinct" property of the 5 main numbers and 2 lucky stars.

4 Spam Detection [10 points]

In an attempt to reduce the amount of spam reaching your inbox, you've installed two separate anti-spam browser extensions. Any single email is either spam (S) or not (S^C), and each of the two programs either marks an email as spam (M_k) or legitimate (M_k^C), for $k = 1, 2$. Assume that 75% of all email is spam—i.e., $P(S) = \frac{3}{4}$, that the first browser extension correctly classifies as spam or legitimate with probability p_1–i.e., $P(M_1|S) = P(M_1^C|S^C) = p_1$, and that the second browser extension is accurate with probability p_2—i.e., $P(M_2|S) = P(M_2^C|S^C) = p_2$. For simplicity, assume the two extensions are conditionally independent of each other, regardless of whether the email is spam or not. Also assume that $0 < p_1 < p_2 < 1$, which among other things, says that the second browser extension correctly classifies emails more often than the first one does.

a. [2 points] Does the order in which the two browser extensions are applied matter? Restated, might our belief that an email is spam be different depending on whether or not the first extension is applied before the second versus the second being applied before the first? Explain your answer.

> **Solution.** The order in which the extensions are applied does not matter. This is because M_1 and M_2 are conditionally independent, so M_2's distribution is not influenced by M_1's result at all (i.e. $P(M_2|S) = P(M_2|S, M_1) = P(M_2|S, M_1^c)$ and likewise for M_2^c).

b. [4 points] What is the probability that a single email is spam even though the first browser extension identifies it as legitimate?

> **Solution.** The answer is $P(S|M_1^c)$, which can be computed by
>
> $$\begin{aligned} P(S|M_1^c) &= \frac{P(M_1^c|S)P(S)}{P(M_1^c|S)P(S) + P(M_1^c|S^c)P(S^c)} \\ &= \frac{\frac{3}{4}(1-p_1)}{\frac{3}{4}(1-p_1) + \frac{1}{4}p_1} \\ &= \boxed{\frac{3(1-p_1)}{3-2p_1}}. \end{aligned}$$
>
> *Commentary.* This problem was fairly well done overall. Some common errors:
>
> - Flipping something, e.g. having p_1 and $1 - p_1$ switched around, computing $P(S|M_1)$ instead of $P(S|M_1^c)$.
> - Forgetting prior terms in the denominator, i.e. computing
>
> $$P(M_1^c) = P(M_1^c|S) + P(M_1^c|S^c) = 1$$
>
> - Computed $P(S, M_1^c)$ instead of $P(S|M_1^c)$.
>
> Some stylistic errors:
>
> - If you use any new notation (e.g. if you rewrite the conditional probability as $P(E|F)$), you must define it! (e.g. "Let E denote the event that ...")
>
>   In this case, this is more work than it is worth, since we defined all the events for you already in the problem.
> - Forgetting to simplify probabilities given in the problem (e.g. $P(S)$, $P(M_1^c|S^c)$).
> - Overloading notation for probabilities and event (e.g. writing $P(p_1|S)$).

c. [4 points] What is the probability that a single email is spam when the two extensions disagree on whether or not the email is spam?

Solution. For simplicity, we use the notation for set difference $E \triangle F = (E \cap F^c) \cup (E^c \cap F)$. The probability is

$$\begin{aligned} P(S|M_1 \triangle M_2) &= \frac{P(S, M_1 \triangle M_2)}{P(M_1 \triangle M_2)} \\ &= \frac{P(M_1 \triangle M_2|S)P(S)}{P(M_1 \triangle M_2|S)P(S) + P(M_1 \triangle M_2|S^c)P(S^c)} \\ &= \frac{\frac{3}{4}(p_1(1-p_2) + p_2(1-p_1))}{\frac{3}{4}(p_1(1-p_2) + p_2(1-p_1)) + \frac{1}{4}(p_1(1-p_2) + p_2(1-p_1))} = \boxed{\frac{3}{4}}. \end{aligned}$$

Notably, the answer does not depend on p_1 or p_2 at all!

Conditional independence was used in the evaluation of the denominator. For instance,

$$\begin{aligned} P(M_1 \triangle M_2|S) &= P(M_1, M_2^c|S) + P(M_1^c, M_2|S) \\ &= P(M_1|S)P(M_2^c|S) + P(M_1^c|S)P(M_2^c|S) \\ &= p_1(1-p_2) + p_2(1-p_1) \end{aligned}$$

Alternative Solution. One can compute the denominator differently: the correctness of each extension is independent, so $P(\text{disagree}) = p_1(1-p_2) + p_2(1-p_1)$. Proceed as per the previous solution.

Commentary. The complexity of the problem led to many conceptual errors. When in doubt over a conditional probability, always fall back to understanding it in terms of the joint distributions.

- (By far the most common error.) When conditioning on a disjoint union, we cannot add it over the two disjoint parts, i.e.

  $$(\text{answer}) = P(S|M_1 \triangle M_2) = P(S|M_1, M_2^c) + P(S|M_2, M_1^c)$$

  is wrong.

- Conditional independence doesn't mean that we can factor probabilities over the *conditioned event*, e.g.

  $$P(S|M_1, M_2^c) = P(S|M_1)P(S|M_2^c)$$

  is wrong.

- M_1 and M_2 being conditionally independent of S doesn't mean that M_1 and M_2 are independent, e.g. one should not do

  $$P(M_1 \triangle M_2) = P(M_1)P(M_2^c) + P(M_1^c)P(M_2)$$

Stylistic errors:

- $P(E|(F|G))$ is not a thing. If you meant that both F and G should happen, it's $P(E|F, G)$.