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

CS109 Quiz #2

Take-Home Quiz information

Each quiz will be a 72-hour open-book, open-note exam. We have designed this quiz to approximate about 3 hours of active work (it is about half as long as an in-person exam designed for 3 hours), however as we have learned, some people will chose to spend more time.

- You can submit multiple times; we will only grade the last submission you submit before 2:30pm (Pacific time) on Saturday, February 27th. 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.
- 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 is 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 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): Solution

1 Better Evaluation of Eye Disease [25 points]

When a patient has eye inflammation, eye doctors "grade" the inflammation. When "grading" inflammation they randomly look at a single 1 millimeter by 1 millimeter square in the patient's eye and count how many "cells" they see.

There is uncertainty in these counts. If the true average number of cells for a given patient's eye is 6, the doctor could get a different count (say 4, or 5, or 7) just by chance. As of 2021, modern eye medicine does not have a sense of uncertainty for their inflammation grades! In this problem we are going to change that. At the same time we are going to learn about poisson distributions over space.

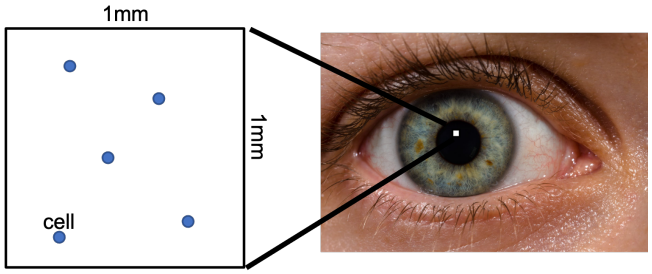

Figure 1: *A 1x1mm sample used for inflammation grading. Inflammation is graded by counting cells in a randomly chosen 1mm by 1mm square. This sample has 5 cells.*

a. (9 points) Explain, as if teaching, why the number of cells observed in a 1x1 square is governed by a poisson process. Make sure to explain how a binomial distribution could approximate the count of cells. Explain what λ means in this context. Note: for a given person's eye, the presence of a cell in a location is independent of the presence of a cell in another location. 100 word limit. Pictures not necessary, but allowed.

Answer. We can approximate a distribution for the count by discretizing the square into a fixed number of equal sized buckets. Each bucket either has a cell or not. Therefore, the count of cells in the 1x1 square is a sum of Bernoulli random variables with equal p, and as such can be modeled as a binomial random variable. This is an approximation because it doesn't allow for two cells in one bucket. Just like with time, if we make the size of each bucket infinitely small, this limitation goes away and we converge on the true distribution of counts. The binomial in the limit, i.e. a binomial as $n \to \infty$, is truly represented by a Poisson random variable. In this context, λ represents the average number of cells per 1×1 sample. See Figure 2.

b. (8 points) For a given patient the true average rate of cells is 5 cells per 1x1 sample. What is the probability that in a single 1x1 sample the doctor counts 4 cells?

Answer. Let X denote the number of cells in the 1x1 sample. We note that $X \sim Poi(5)$. We want to find $P(X = 4)$.

$$P(X = 4) = \frac{5^4 e^{-5}}{4!} \approx 0.175$$

In addition to providing an expression above, please compute a numeric answer: 0.175

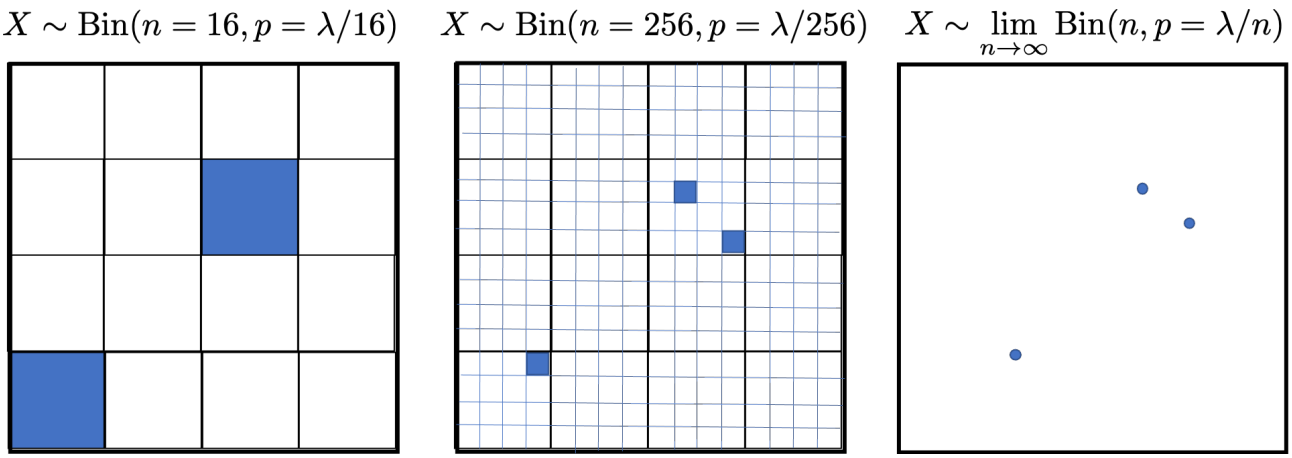

Figure 2: *X is counts of events in discrete buckets. In the limit, as n (number of buckets)* $\to \infty$*, X becomes a Poisson.*

c. (8 points) For a given patient the true average rate of cells is 5 cells per 1mm by 1mm sample. In an attempt to be more precise, the doctor counts cells in **two** different, larger **2mm by 2mm** samples. Assume that the occurrences of cells in one 2mm by 2mm samples are independent of the occurrences in any other 2mm by 2mm samples. What is the probability that she counts 20 cells in the first samples and 20 cells in the second?

Answer. Let Y_1 and Y_2 denote the number of cells in each of the 2x2 samples. Since there are 5 cells in a 1x1 sample, there are 20 samples in a 2x2 sample since the area quadrupled, so we have that $Y_1 \sim Poi(20)$ and $Y_2 \sim Poi(20)$. We want to find $P(Y_1 = 20 \wedge Y_2 = 20)$. Since the number of cells in the two samples are independent, this is equivalent to finding $P(Y_1 = 20)P(Y_2 = 20)$.

$$P(Y_1 = 20 \wedge Y_2 = 20) = P(Y_1 = 20)P(Y_2 = 20) = \left(\frac{20^{20}e^{-20}}{20!}\right)^2 = 0.00789$$

In addition to providing an expression above,
please compute a numeric answer: 0.00789

2 Stirring the Pot [25 points]

You are cooking a big pot of vegetables and the instructions say to “stir the pot". If you don’t stir the pot some vegetables will be under-cooked and some will be overcooked.

How much heat a vegetable gets depends on if it is on the bottom of the pot or not. If it is on the bottom of the pot it will get 5 units of heat per minute. If it is not, it will get 1 unit of heat per minute.

Each time you stir, each vegetable, independently, has a 1/5 chance of ending up at the bottom of the pot. For each stir, the position of a vegetable is independent of previous position and of the position of other vegetables.*

a. (8 points) You stir the pot a single time and then let the vegetables cook for 10 minutes. What is the variance of heat on a vegetable?

Answer.

In addition to providing an expression above, please compute a numeric answer: 256

.

Let X represent the units of heat a vegetable gets over the course of 10 minutes. Then,

$$E[X] = \frac{1}{5} * (50) + \frac{4}{5}(10) = 18$$

and it follows that

$$Var(X) = \frac{1}{5} * (50 - 18)^2 + \frac{4}{5}(10 - 18)^2 = 256$$

.

b. (9 points) You stir the pot **once every minute** for 10 minutes. What is the variance of heat on a vegetable?

Answer.

In addition to providing an expression above, please compute a numeric answer: 25.6

.

Let X_i represent the heat a vegetable receives at minute i where $1 \leq i \leq 10$. Then

$$E[X_i] = \frac{1}{5} * (5) + \frac{4}{5}(1) = 1.8$$

and we find that

$$Var(X_i) = \frac{1}{5} * (5 - 1.8)^2 + \frac{4}{5}(1 - 1.8)^2 = 2.56$$

*This independence assumption doesn’t perfectly match the real world – for example in the real world it wouldn’t be possible for all vegetables to end up at the bottom of the pot. While the independence assumption is wrong, it has a very small impact on the final answer, and it makes for a more straightforward quiz.

. To find the variance of the heat on a vegetable we must sum over the entire time period. Then, we want to find

$$Var(\sum_{i=1}^{10} X_i) = \sum_{i=1}^{10} Var(X_i) = \sum_{i=1}^{10} 2.56 = 25.6$$

.

c. (8 points) Let's explore what this problem would look like with continuous numbers. In the continuous version, when you stir the pot, each vegetable has a distance from the bottom which is equally likely to be any real valued number between 0 and 1. The heat received by a vegetable, H, with distance d is: $H(d) = 1 - d^2$. You stir the pot a single time and then let the vegetables cook for 10 minutes. Show an expression that you would need to solve in order to calculate the expectation of heat. Your expression should include an integral. Then, solve the expression to get a numeric answer.

Answer.

In addition to providing an expression above, please compute a numeric answer: $\frac{20}{3}$

.

H is a rate measured in heat-units per minute. D is the distance measured in distance-units.

$$D \sim Uni(0, 1); p_D(d) = 1 \text{ for } 0 \leq d \leq 1, \text{ else } 0$$

$$H = g(D) = 1 - D^2$$

We want to compute $E[10H]$.

$$E[10H] = E[10g(D)] = 10E[g(D)]$$

Using the Law of the Unconscious Statistician:

$$10E[g(D)] = 10 \int_0^1 g(d)p_D(d)dd$$

$$= 10 \int_0^1 (1 - d^2) * (1)dd$$

$$= 10 \int_0^1 (1 - d^2)dd$$

$$= 10[d - \frac{d^3}{3}]_0^1$$

$$= 10(\frac{2}{3})$$

$$= \frac{20}{3}$$

3 Gender Composition of Sections [20 points]

A massive online Stanford class has sections with 10 students each. Each student in our population has a 50% chance of identifying as female, 47% chance of identifying as male and 3% chance of identifying as non-binary. Even though students are assigned randomly to sections, a few sections end up having a very uneven gender distribution just by chance. You should assume that the population of students is so large that the percentages of students who identify as male / female / non-binary are unchanged, even if you select students without replacement.

a. (6 points) Write an expression for the exact probability that a section has 0, 1, 9 or 10 people who identify as female.

Answer. Let X be the number of students in a section who identify as female, where we have $X \sim Bin(n = 10, p = 0.5)$. Here, X is binomial random variable with 50% probability of selecting a female student, or $P(Female) = 0.5$.

We want to find an expression for the exact probability that a section has 0, 1, 9, or 10 people who identify as female, which you can achieve by summing up the 4 mutually exclusive event probabilities as such:

$$P(X = 0, 1, 9, 10) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 9) + P(X = 10)$$

$$= \binom{10}{0}(0.5)^0(0.5)^{10} + \binom{10}{1}(0.5)^1(0.5)^9 + \binom{10}{9}(0.5)^9(0.5)^1 + \binom{10}{10}(0.5)^{10}(0.5)^0$$

$$= 0.5^{10} + 10 \cdot 0.5^{10} + 10 \cdot 0.5^{10} + 0.5^{10} \approx 22 \cdot 0.5^{10} = \mathbf{0.02148}$$

Alternatively, this can equivalently be found by doing $1 - P(2 \leq X \leq 8)$:

$$P(X = 0, 1, 9, 10) = 1 - P(2 \leq X \leq 8)$$

$$= 1 - \sum_{i=2}^{8} \binom{10}{i}(0.5)^i(0.5)^{10-i}$$

$$= 1 - \sum_{i=2}^{8} \binom{10}{i}(0.5)^{10} \approx \mathbf{0.02148}$$

b. (8 points) Forty students are randomly selected to be in a single review session. Use an approximation to efficiently estimate the probability of there being more than 10 and less than 30 people who identify as female in the review session.

Answer. The exact way to model this would be to have $X \sim Bin(n = 40, p = 0.5)$. However, since we are asked to use an approximation (choosing between Poisson and Normal approximations so far learned in class), we first look at the distinguishing criteria:

- $n = 40 > 20 \rightarrow$ sufficient for both approximations
- $p = 0.5 \rightarrow$ moderate, so leaning towards normal

- $np(1-p) = 40(0.5)(0.5) = 10 \geq 10 \rightarrow$ good to use for a normal approximation
- $\lambda = np = 40(0.5) = 20 \rightarrow$ too large to use a Poisson approximation
- Students are independently selected $\rightarrow$ also great for a normal distribution

Therefore, we should proceed using a normal approximation, so we let $X \sim N(\mu = 20, \sigma^2 = 10)$, where $\mu = np = 40 \cdot 0.5 = 20$ and $\sigma^2 = np(1-p) = 40 \cdot 0.5 \cdot (0.5) = 10$. Since we are using a continuous RV to approximate a discrete one, we apply a continuity correction and thus solve for $F_x(29.5) - F_x(10.5)$:

$$P(10 < X < 30) = F_x(29.5) - F_x(10.5)$$

$$= \Phi\left(\frac{29.5 - 20}{\sqrt{10}}\right) - \Phi\left(\frac{10.5 - 20}{\sqrt{10}}\right) \approx \mathbf{0.9974}$$

In addition to providing an expression above,
please compute a numeric answer: 0.9974

c. (6 points) A given section has 5 people who identify as female, 3 who identify as non-binary and 2 who identify as male. What is the probability of this exact composition?

Answer. We can calculate the probability of this exact composition by modeling the distribution of gender identities in a given section by using a multinomial random variable. We know the probabilities of someone identifying as female ($X_F = 0.50$), male ($X_M = 0.47$), and non-binary ($X_{NB} = 0.03$).

Therefore, we can calculate this exact composition probability by the following:

$$\binom{10}{5, 3, 2}(0.5)^5(0.03)^3(0.47)^2$$

$$\frac{10!}{5!3!2!}(0.5)^5(0.03)^3(0.47)^2 \approx \mathbf{0.00047}$$

In addition to providing an expression above,
please compute a numeric answer: 0.00047

4 Section Assignment Algorithm [20 points]

At the start of the quarter we assign students to sections based on their time preferences. Let's consider a "greedy" section assignment algorithm and compute the expectation of the number of people who we are unable to fit in a section ("unassigned").

A class has $k = 12$ students and $r = 4$ sections. Each section can have $m = 3$ people max (though they can have less, even zero students). In this question assume each person gives exactly one preference and that a student's preference is equally likely to be any of the r sections.

The greedy algorithm we are going to analyze goes through students one-by-one (arbitrarily). When considering a student, it checks if there is space in their preferred section. If so they are assigned to that section. Otherwise they are "unassigned". Note: this problem is meant to be a challenge.

Algorithm: Greedy Section Time Assignment

```
# keep track of how big each section is
section_counts = [0 for i in range(r)]
# count how many students go unassigned
n_unassigned = 0
# loop over students in an arbitrary order
for student in students:
  # each student has a single section preference
  section_pref = student['pref']
  # does their preferred section have space?
  if section_counts[section_pref] < m:
     section_counts[section_pref] += 1
  else:
     n_unassigned += 1
# this is what we are interested in.
print(n_unassigned)
```

a. (10 points) What is the probability that the very last person considered by the algorithm is unassigned?

Answer. The desired probability is $\sum_{i=3}^{11} \binom{11}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{11-i}$. Since the previous 11 students each independently want the same section as the last student with probability 1/4, the number of previous students with the same preference is a binomial random variable with $n = 11$ and $p = \frac{1}{4}$. The last student is unassigned whenever 3 or more previous students share the last student's preference.

In addition to providing an expression above, please compute a numeric answer: 0.545

b. (10 points) What is the expected number of people who will be unassigned? You do not need to provide an exact answer, instead you can leave your answer as an expression.

Answer. The answer is $\sum_{i=3}^{11} \sum_{j=3}^{i} \binom{i}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{i-j}$ (or any equivalent expression).
Students from student 4 onwards could possibly be unassigned, so for $3 \leq i \leq 11$, we can define X_i to be an indicator variable such that $X_i = 1$ whenever student $i + 1$ is unassigned. The expected number of unassigned students is then $\sum_{i=3}^{11} X_i$.
Similar to part (a), for $3 \leq i \leq 11$, X_i depends on a binomial distribution with $n = i$ and $p = \frac{1}{4}$. We have that $E[X_i]$ is the probability that at least 3 students prior to student $i + 1$ had the same preference. This happens with probability $\sum_{j=3}^{i} \binom{i}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{i-j}$. Thus, the expected number of unassigned students is $E[\sum_{i=3}^{11} X_i] = \sum_{i=3}^{11} E[X_i] = \sum_{i=3}^{11} \sum_{j=3}^{i} \binom{i}{j} \left(\frac{1}{4}\right)^j \left(\frac{3}{4}\right)^{i-j}$.
Alternate solution (inspired by a student solution):
The answer can be equivalently found as $4 * \sum_{i=4}^{12} (i-3) \binom{12}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{12-i}$. For $1 \leq i \leq 4$, let Y_i equal the number of students who preferred section i but were not able to be assigned. Then the expected number of unassigned students is $E[\sum_{i=1}^{4} Y_i] = \sum_{i=1}^{4} E[Y_i]$. If there are i students who prefer a section for $i > 3$, then $i - 3$ of those students are unassigned. We also recall that the distribution of students preferring a particular section is distributed as Bin(12, $\frac{1}{4}$). Therefore, $E[Y_i]$ always equals $\sum_{i=4}^{12} (i-3) \binom{12}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{12-i}$, giving us our final answer of $4 * \sum_{i=4}^{12} (i-3) \binom{12}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{12-i}$, which is equal to the alternate answer.

5 Optional Meme [0 points]

That's all folks! Optionally include a meme about the CS109 material. Of course the meme should be in line with the Stanford Fundamental Standard. This question is worth zero points and is just for fun.

As an aside each of these problems are drawn from real world phenomena. Problem 1: in 2004 a group of international uveitis specialists created a standardization of grading eye inflammation. Until this quiz it has been used without thinking about probability. The next step in this line of thought is to think of λ as a random variable, allowing you to be explicit about your uncertainty after observation, and to do important things, like know the probability that a patient has actually gotten worse over a period of time. Later in CS109 we will talk about thinking of parameters like λ as random variables. Problem 2: Chris and Jerry spent a lot of time cooking this last year. And thinking about probability at the same time. Problem 3: Code in Place was a massive section-based class to teach the first half of Cs106A that Stanford offered as a community service project in the time of Covid-19. It was so large that, even though gender was split very evenly, a large number of sections ended up being all female or all male just by chance. Everyone was surprised. Problem 4: The algorithm used to assign CS109 students to section is constantly being improved. We don't use this particular one, but we do use something similar. In practice our expectation of unassigned students is very close to 0. Can you think of an algorithm for CS109 section assignments?