Chris, Jerry, Katie, Derek, Freya, and Doris
CS 109

Quiz #3
March 17-19, 2021

CS109 Quiz #3 Solution

Take-Home Quiz information

This quiz will is an open-book, open-note exam that should be completed and submitted by Friday, March 19th at 11:59 AOE. We have designed this quiz to approximate about 3 hours of active work, 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 11:59pm (AOE) on Friday, March 19th. 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): _____

1 Fairness in GaussHouse [25 points]

You have recently taken over the management of admissions to The GaussHouse, a mansion for content creators. The house specializes in producing educational short videos about CS109 content. Prospective residents have to apply and be selected to live in GaussHouse. Admission should be based on ability to make CS109 educational content, but is it?

The previous manager left behind data from thousands of applications from last year. You are curious about whether the selection process was biased towards Generation Z (people under 24). In addition to the age category, you also have access to whether or not an applicant had previously produced and uploaded a probability related short video, which we will call Content for short.

Gen-Z	Content	Accepted	Probability
0	0	1	.002
0	1	1	.0075
1	0	1	.0005
1	1	1	.011
0	0	0	.023
0	1	0	.2675
1	0	0	.0995
1	1	0	.589

a. (3 points) What is the probability that an applicant is Gen-Z?

Answer. $(.0005 + .011 + .0995 + .589) = 0.7$

b. (3 points) Given that an applicant was accepted, what is the probability that the applicant was Gen-Z?

Answer. $(.011 + .0005)/(.011 + .0005 + .0075 + .002) = 0.54761904761$

c. (3 points) Given that an applicant was accepted, what is the probability that the applicant was not Gen-Z?

Answer. $((.0075 + .002)/(.011 + .0005 + .0075 + .002)) = 0.45238095238$

d. (8 points) Next let us calculate a few fairness tests. Recall that in our class on fairness we defined random variables A for a demographic variable, and R for result. $R = 1$ is a positive result, in this case acceptance.

   i. **Independence Fairness** means that acceptance rates should be the same for all groups. It is defined as:

$$P(R = 1|A = a) = P(R = 1|A = b) \tag{1}$$

   Did the previous admissions to GaussHouse satisfy independence with respect to age?

Answer. No – $0.54761904761 \neq 0.45238095238$

ii. **Relaxed Independence Condition** is defined as:

$$\frac{P(R = 1|A = a)}{P(R = 1|A = b)} \geq 1 - \varepsilon \tag{2}$$

Where $A = a$ is the event that an applicant is from the group which is less represented in the applicant pool and $A = b$ is the event that an applicant is from the more represented group. Did previous admissions satisfy relaxed independence with respect to age for $\varepsilon = .2$?

Answer. Yes – 0.82608695652

e. (4 points) Your house wants to update their admission system, and one of their goals is to satisfy "fairness through unawareness". One option is to hide an applicant's age from people who review applications. You notice that applications still have people's names. US census data makes it very clear that names change substantially in popularity over time. How could the inclusion of an applicant's name impact fairness through unawareness even if age is removed?

Answer. *Answers varied and were graded without any one specific solution in mind.*

f. (4 points) You want to ensure that future admissions to GaussHouse are fair. Which of the types of fairness that we talked about in class (fairness through unawareness; independence; relaxed independence; separation) would you choose and why? In your answer you should explain in words what each of the four types of fairness means in this context. Base your answer on what you think would be the best fairness standard for this context, not on what current US discrimination law requires. A precise, well justified opinion will receive full credit.

Answer. *Answers varied and were graded without any one specific solution in mind.*

2 Learning to Play Darts [25 points]

The Game of Darts requires a single player to throw a physical dart toward a circular board with hopes of striking the board as close to its center as possible. For our purposes, we'll assume that the center of the circular board is the origin—that is, the coordinate (0, 0)—and that the radius of the dart board is 9 inches. Someone new to the game throws darts that rarely hit anywhere near the center. Rather, the novice throws darts that strike the board according to independent random variables X and Y.

Now, even novices have a sense for left-to-right centering, so $X \sim \mathcal{N}(0, 4)$. But they struggle to understand gravity's pull on the dart, so Y isn't distributed as a Gaussian, because rather as something more complicated: $Y \sim 10 - \text{Exp}(\frac{1}{12})$. This problem takes interest in computing various statistics about dart throws and ultimately predicting how far a dart lands from the center of the board when guided by these two independent distributions.

a. (6 points) Compute $P(X \leq 0, 9 \leq Y \leq 10)$. This is the probability that the dart lands above and to the left of the top of the entire dart board.

Answer.

Because X and Y are independent random variables,

$$P(X \leq 0, 9 \leq Y \leq 10) = P(X \leq 0)\, P(9 \leq Y \leq 10).$$

$P(X \leq 0)$ is easily recognized as 0.5, but computing $P(9 \leq Y \leq 10)$ requires some work. To make it easier, let's define a new random variable Z to be equal to $10 - Y$, so that

$$Z = 10 - Y = \text{Exp}(\tfrac{1}{12}).$$

Therefore,

$$\begin{aligned} P(9 \leq Y \leq 10) &= P(9 \leq 10 - Z \leq 10) \\ &= P(-1 \leq -Z \leq 0) \\ &= P(0 \leq Z \leq 1) \\ &= P(Z \leq 1) \end{aligned}$$

This final line is just a simplification of the one above it, since $P(Z)$ is 0 for all $Z < 0$. That means we can tap the CDF of the Exponential distribution with $\lambda = \frac{1}{12}$.

$$P(Z \leq 1) = 1 - e^{-\frac{1}{12} \cdot 1} = 1 - e^{-\frac{1}{12}}$$

Therefore:

$$P(X \leq 0, 9 \leq Y \leq 10) = \frac{1}{2}(1 - e^{-\frac{1}{12}}).$$

b. (5 points) Without relying on any calculus at all, derive values for $E[X^2]$, $E[Y]$, and $E[Y^2]$.

Answer.

$E[Y]$ is the easiest one to compute, since $E[Z]$—where Z is from part a—is 12, so that $E[Y] = 10 - E[Z] = -2$.

$E[X^2]$ is easily computed if we remember that $Var(X) = E[X^2] - E[X]^2$. Since $E[X]$ is 0 and $Var(X) = 4$, it must be the case that $E[X^2] = 4$ as well.

Our approach to computing $E[Y^2]$ is the same, except that we need to compute $Var(Y)$, since that's not explicitly given. But:

$$\begin{aligned} Var(Y) &= Var(10 - \text{Exp}(\frac{1}{12})) \\ &= Var(-\text{Exp}(\frac{1}{12})) \\ &= (-1)^2 Var(\text{Exp}(\frac{1}{12})) \\ &= Var(\text{Exp}(\frac{1}{12})) \\ &= 144 \end{aligned}$$

Since $Var(Y) = E[Y^2] - E[Y]^2$, $E[Y^2] = Var(Y) + E[Y]^2 = 144 + (-2)^2 = 148$. Ta da!

c. (6 points) If you plan to throw 10000 darts over the course of, say, a week of intense practice, you would expect the average y coordinate of all throws to itself conform to a probability distribution. What is that probability distribution?

You should assume all throws are independent. Specify the model and all necessary parameters.

This screams Central Limit Theorem, which tells us that the sample mean $\bar{Y}$ is a random variable distributed as a Normal with parameters $\mu = E[Y]$ and $\sigma^2 = Var(Y)/n$, where $n = 10000$. That means that:

$$\bar{Y} \sim \mathcal{N}(-2, 0.0144)$$

(If your parameters are incorrect, but consistent with the values you generated in part b, we gave you full credit.)

d. (8 points) Consider the random variable $X_{\text{sq}} = X^2$. First, compute the cumulative distribution function $F_{X_{\text{sq}}}(a)$, which is the probability that $X_{\text{sq}} \leq a$ for some value of a in the support of X_{sq}. From that, compute $f_{X_{\text{sq}}}(a)$, which is the PDF of X_{sq}. Your CDF can be defined in terms of Φ, but your PDF should be written as a continuous function on a.*

*We thought about asking you to derive $Y_{\text{sq}} = Y^2$, buy that's even **more** difficult, so we're excluding it.

Answer.

Let's formally rely on $F_{X_{\text{sq}}}(a)$ to denote the cumulative distribution function. By definition, $F_{X_{\text{sq}}}(a) = P(X_{\text{sq}} \leq a)$. Because $X_{\text{sq}} = X^2$ and $X \sim \mathcal{N}(0, 4)$, we have that

$$\begin{aligned}
P(X_{\text{sq}} \leq a) &= P(X^2 \leq a) \\
&= P(-\sqrt{a} \leq X \leq \sqrt{a}) \\
&= \Phi(\frac{\sqrt{a}}{2}) - \Phi(-\frac{\sqrt{a}}{2}) \\
&= \Phi(\frac{\sqrt{a}}{2}) - (1 - \Phi(\frac{\sqrt{a}}{2})) \\
&= 2\Phi(\frac{\sqrt{a}}{2}) - 1.
\end{aligned}$$

In order to compute $f_{X_{\text{sq}}}(a)$, we need to take the derivative of $F_{X_{\text{sq}}}(a)$ with respect to a. That means that:

$$\begin{aligned}
f_{X_{\text{sq}}}(a) &= 2\frac{d}{da}(\Phi(\frac{\sqrt{a}}{2}) - 1) \\
&= 2\frac{d}{da}\Phi(\frac{\sqrt{a}}{2}) \\
&= 2\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{\sqrt{a}}{2})^2}\frac{d}{da}\frac{\sqrt{a}}{2} \\
&= 2\frac{1}{\sqrt{2\pi}}e^{-\frac{a}{8}}\frac{d}{da}\frac{\sqrt{a}}{2} \\
&= 2\frac{1}{\sqrt{2\pi}}e^{-\frac{a}{8}}\frac{1}{4\sqrt{a}} \\
&= \frac{1}{2\sqrt{2\pi a}}e^{-\frac{a}{8}}
\end{aligned}$$

3 Lambda Estimation for Eye Evaluation[25 points]

This quarter in CS109 we are helping our eye doctor friends think about probability for evaluation of inflammation in the eye. In our last quiz we learned that when a patient has eye inflammation, eye doctors "grade" the inflammation. When "grading" inflammation they look at a single, randomly-chosen 1 millimeter by 1 millimeter square in the patient's eye and count how many "cells" they see. The number of cells that a doctor counts, X, is governed by a **Poisson** process where the parameter λ is the true average rate of cells in a 1mm by 1mm square, $X \sim \text{Poi}(\lambda)$. For example, a patient could have a true inflammation rate of $\lambda = 4$, then in a particular 1x1 sample, the chance that a doctor observes $X = 3$ would be the probability of a Poisson with rate $\lambda = 4$ producing a 3.

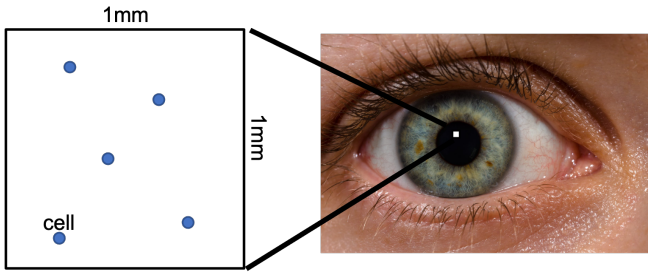

Figure 1: *Recall: A 1x1mm sample used for inflammation grading. Inflammation is graded by counting cells in a randomly chosen 1mm by 1mm square. This leads to a Poisson process.*

In Quiz 2 we estimated the probability of different counts given the true rate λ. In this quiz we are going to think about our belief regarding the true rate λ based on an observed count. This is going to require us to think of the true rate as a random variable.

Inflammation prior: Based on millions of historical patients, doctors have learned that the prior probability density function of true rate of cells is:

$$f(\lambda) = K \cdot \lambda \cdot e^{-\frac{\lambda}{2}}$$

Where K is a normalization constant and λ must be greater than 0.

a. (8 points) A doctor takes a single sample and counts 4 cells. Give an equation for the updated probability density of λ. Use the "Inflammation prior" as the prior probability density over values of λ. Your probability density may have a constant term.

Answer. Let θ be the random variable for true rate. Let X be the random variable for the count

$$\begin{aligned} f(\theta = \lambda | X = 4) &= \frac{P(X = 4 | \theta = \lambda) f(\theta = \lambda)}{P(X = 4)} \\ &= K \cdot \lambda^4 e^{-\lambda} \lambda e^{-\frac{\lambda}{2}} \\ &= K \cdot \lambda^5 e^{-\frac{3}{2}\lambda} \end{aligned}$$

b. (7 points) A doctor takes a single sample and counts 4 cells. What is the Maximum A-Posteriori estimate of λ? Use the "Inflammation prior".

Answer. Maximize the likelihood of the parameter given the data:

$$\begin{aligned}\arg\max_{\lambda} K \cdot \lambda^5 e^{-\frac{3}{2}\lambda} &= \arg\max_{\lambda} \log K \cdot \lambda^5 e^{-\frac{3}{2}\lambda} \\ &= \arg\max_{\lambda} \log K + 5\log\lambda - \frac{3}{2}\lambda\end{aligned}$$

Calculate the derivative

$$\begin{aligned}\frac{\partial}{\partial\lambda} f(\theta = \lambda | X = 4) &= \frac{\partial}{\partial\lambda} \log K + 5\log\lambda - \frac{3}{2}\lambda \\ &= \frac{5}{\lambda} - \frac{3}{2}\end{aligned}$$

Set it equal to 0

$$\begin{aligned}\frac{5}{\lambda} - \frac{3}{2} &= 0 \\ \frac{5}{\lambda} &= \frac{3}{2} \\ \lambda &= \frac{10}{3}\end{aligned}$$

c. (3 points) Explain, in words, the difference between the two estimates of lambda in part (a) and part (b).

Answer. The estimate in the first part is a “distribution" (also called a soft estimate) whereas the estimate in the second part is a single value (also called a point estimate). The former contains information about confidence.

d. (8 points) A patient comes on two separate days. The first day the doctor counts 5 cells, the second day the doctor counts 4 cells. Based only on this observation, and treating the true rates on the two days as independent, what is the probability that the patient’s inflammation has gotten better (in other words, that their λ has decreased)? For full credit provide an expression to calculate the probability exactly. For most of the credit provide an explanation for how you could approximate the probability.

Answer. Let θ_1 be the random variable for lambda on the first day and θ_2 be the random variable for lambda on the second day.

$$\begin{aligned}f(\theta_1 = \lambda | X = 5) &= K_1 \cdot \lambda^6 e^{-\frac{3}{2}\lambda} \\ f(\theta_2 = \lambda | X = 4) &= K_2 \cdot \lambda^5 e^{-\frac{3}{2}\lambda}\end{aligned}$$

The question is asking what is $P(\theta_1 > \theta_2)$? There are a few ways to calculate this exactly:

$$\begin{aligned}
&\int_{\lambda_1=0}^{\infty} \int_{\lambda_2=0}^{\lambda_1} f(\theta_1 = \lambda_1, \theta_2 = \lambda_2) \\
&= \int_{\lambda_1=0}^{\infty} \int_{\lambda_2=0}^{\lambda_1} f(\theta_1 = \lambda_1) \cdot f(\theta_2 = \lambda_2) \\
&= \int_{\lambda_1=0}^{\infty} f(\theta_1 = \lambda_1) \int_{\lambda_2=0}^{\lambda_1} f(\theta_2 = \lambda_2) \\
&= \int_{\lambda_1=0}^{\infty} K_1 \cdot \lambda^6 e^{-\frac{3}{2}\lambda} \int_{\lambda_2=0}^{\lambda_1} K_2 \cdot \lambda^5 e^{-\frac{3}{2}\lambda}
\end{aligned}$$

Another approach is to use convolution to solve for the probability that $P(\theta_1 - \theta_2 > 0)$. To do so we can first solve for $f(\theta_1 - \theta_2 = c)$ and then you can find the probability that that convolution is greater than 0.
In both approaches students should explain how to calculate K_1 and K_2, for a small number of points.

$$\begin{aligned}
K_1 &= 1 / \int_{\lambda=0}^{\infty} \lambda^6 e^{-\frac{3}{2}\lambda} \\
K_2 &= 1 / \int_{\lambda=0}^{\infty} \lambda^5 e^{-\frac{3}{2}\lambda}
\end{aligned}$$

4 Predicting Drug Effectiveness [25 points]

You have been given a grant to work with the FDA to build a machine learning model that can predict how effective new drugs will be at curing diseases. A drug's effectiveness is the probability that it cures a patient.

The FDA gives you a mountain of historical data of drugs. The data contains n drugs all of which are in the exact same format: $x^{(i)}, p^{(i)}$ where $x^{(i)}$ is an m dimensional vector of 0s and 1s describing features of the drug and $p^{(i)}$ is the observed probability that the drug cures the disease when given to a patient. Note that $p^{(i)}$ must be in the range 0 to 1 as it is a probability.

Your decide to build a model, inspired by Logistic Regression, called Logistic-Beta. Logistic-Beta uses the features of a drug x and predicts how likely it will be that a drug has a given effectiveness value p. To do so, our model first generates two values, a and b. These values are then interpreted as the parameters of a Beta distribution:

$$a = 1 + \sum_{j=0}^{m} x_j \cdot \theta_j$$

$$b = 1 + \sum_{j=0}^{m} x_j \cdot \phi_j$$

$$Y \sim \text{Beta}(a, b)$$

What is likelihood of a single drug with features x and effectiveness p? It is the probability density that the Beta random variable Y, which is computed using x, takes on the value p. The model has $2m + 2$ learnable parameters across two equal sized vectors: θ and ϕ.

a. (8 points) Imagine you have $m = 2$ features per drug and have finished training your model. You have learned parameters $\phi = [-1, 0, 3]$ and $\theta = [2, 1, 2]$. For a new drug with features $x = [1, 1]$ what is your model's belief that the effectiveness is greater than 0.8? You may use a computer, but must still show your work.

Answer.

$$\begin{aligned} a &= 1 + \theta_0 + x_1 \cdot \theta_1 + x_2 \cdot \theta_2 \\ &= 1 + 2 + 1 \cdot 1 + 1 \cdot 2 = 6 \\ b &= 1 + \phi_0 + x_1 \cdot \phi_1 + x_2 \cdot \phi_2 \\ &= 1 - 1 + 1 \cdot 0 + 1 \cdot 3 = 3 \end{aligned}$$

Thus $Y \sim \text{Beta}(a = 6, b = 3)$. We use python to solve for $1 - F_Y(0.8)$

```
>>> 1- stats.beta.cdf(0.8, 6, 3)
0.20308224
```

b. (8 points) Write an expression for the log likelihood of the n drugs in the dataset.

Answer. Let $\alpha^{(i)} = \theta^T x^{(i)}$ and $\beta^{(i)} = \phi^T x^{(i)}$

$$\begin{aligned} L(\theta, \phi) &= \prod_i f(Y^{(i)} = p^{(i)}) \\ &= \prod_i \frac{1}{\mathrm{B}(\alpha^{(i)} + 1, \beta^{(i)} + 1)} \cdot {p^{(i)}}^{\alpha^{(i)}} (1 - p^{(i)})^{\beta^{(i)}} \end{aligned}$$

$$\begin{aligned} LL(\theta, \phi) &= \log L(\theta, \phi) \\ &= \sum_i -\log \mathrm{B}(\alpha^{(i)} + 1, \beta^{(i)} + 1) + \alpha^{(i)} \log p^{(i)} + \beta^{(i)} \log(1 - p^{(i)}) \end{aligned}$$

c. (9 points) Compute all derivatives that you would need in order to learn the parameters θ and ϕ via gradient descent.

Answer. This last part of the exam had one especially tricky part. Does the student realize that they need to also optimize through the $\log \mathrm{B}(\alpha^{(i)+1}, \beta^{(i)+1})$ term? Realizing that is worth some points. Being able to solve it completely would be extra credit.

$$\frac{\partial LL}{\partial \theta_i} = \sum_i -\frac{\partial}{\partial \theta_j} \log \mathrm{B}(\alpha^{(i)} + 1, \beta^{(i)} + 1) + \frac{\partial}{\partial \theta_j} \left[\alpha^{(i)} \log p^{(i)} + \beta^{(i)} \log(1 - p^{(i)}) \right]$$

We can solve this in parts, first the more straight forward part:

$$\begin{aligned} \frac{\partial}{\partial \theta_j} \left[\alpha^{(i)} \log p^{(i)} + \beta^{(i)} \log(1 - p^{(i)}) \right] &= \frac{\partial}{\partial \theta_j} \alpha^{(i)} \log p^{(i)} \\ &= x_j^{(i)} \log p^{(i)} \end{aligned}$$

Then the truly complex part. To solve this part you can either use the gamma definition of B or you can use the definition from the reader where it is the constant that makes the PDF integrate to 1.

$$\frac{\partial}{\partial \theta_j} \log \mathrm{B}(\alpha^{(i)+1}, \beta^{(i)+1}) = \frac{\partial}{\partial \theta_j} \log \int_{x=0}^{1} {p^{(i)}}^{\alpha^{(i)}} (1 - p^{(i)})^{\beta^{(i)}}$$

Getting to this point is worth full credit. If students are able to get substantially further, that would be extra credit! Mabrouk.
The students also need recognize that they must calculate $\frac{\partial LL}{\partial \phi_i}$. Since it is symmetric to $\frac{\partial LL}{\partial \theta_i}$ they don't have to show as much work, and we shouldn't double ding them for making the same mistakes twice. The derivative will have a $1 - p^{(i)}$ in place of $p^{(i)}$ because ϕ shows up in the β term.