Data 102, Spring 2022 Midterm 1

- You have 110 minutes to complete this exam. There are 6 questions, totaling 40 points.
- You may use one 8.5×11 sheet of handwritten notes (front and back), and the provided reference sheet. No other notes or resources are allowed.
- You should write your solutions inside this exam sheet.
- You should write your name and Student ID on every sheet (in the provided blanks).
- Make sure to write clearly. We can't give you credit if we can't read your solutions.
- Even if you are unsure about your answer, it is better to write down partial solutions so we can give you partial credit.
- We have provided one blank page of scratch paper at the end of the exam. No work on this page will be graded.
- You may, without proof, use theorems and facts that were given in the discussions or lectures, **but please cite them**.
- There will be no questions allowed during the exam: if you believe something is unclear, clearly state your assumptions and complete the question.
- Unless otherwise stated, no work or explanations will be graded for multiple-choice questions.
- Unless otherwise stated, you must show your work for free-response questions in order to receive credit.

Last name	
First name	
Student ID (SID) number	
Calcentral email (`@berkeley.edu`)	
Name of person to your left	
Name of person to your right	

Honor Code

I will respect my classmates and the integrity of this exam by following this honor code.

I affirm:

- All of the work submitted here is my original work.
- I did not collaborate with anyone else on this exam.

Signature: _____

1. (5 points) For each of the following, determine whether the statement is true or false. For this question, no work will be graded and no partial credit will be assigned.

   (a) (1 point) Gibbs sampling is always a better choice than rejection sampling because it never rejects any samples.

   ◯ A. TRUE ◯ B. FALSE

   (b) (1 point) Consider two medical tests for two different diseases, A and B. Disease/test A has TPR = TNR = 0.99. Disease/test B has TPR = TNR = 0.7. Then, the FDR for disease/test A must be higher than the FDR for disease/test B.

   ◯ A. TRUE ◯ B. FALSE

   (c) (1 point) Markov Chain Monte Carlo methods can only be applied to Bayesian models if we choose a conjugate prior for our likelihood.

   ◯ A. TRUE ◯ B. FALSE

   (d) (1 point) Gibbs sampling is based on conditioning on only the first observed data point (e.g., x_1), then sampling all hidden variables, then repeating for the second data point (e.g., x_2), and so on.

   ◯ A. TRUE ◯ B. FALSE

   (e) (1 point) When using the LORD algorithm with desired FDR α on a sequence of 100 p-values, the false discovery rate (FDR) after the first 10 p-values is guaranteed to be less than or equal to α.

   ◯ A. TRUE ◯ B. FALSE

2. (4 points) Your friend Gabriel has two data sets that he needs help analyzing. He knows that you have recently learned about generalized linear models (GLMs) in Data 102, so he turns to you for help.

   (a) (2 points) Gabriel's first data set is about Cal dining. Gabriel wants to predict the total number of students who will go to a Cal dining hall on a given day using information about the day (day of week, weather, etc.).

   Specify an inverse link function and two different likelihood models that Gabriel could use for this dataset.

   Solution: The two likelihood models Gabriel could use are **Poisson** and **negative binomial**, which are both probability distributions over non-negative integers and are thus appropriate for modeling counts.

   The ideal inverse link function to use with these would be the **exponential link function**.

   (b) (2 points) In Gabriel's second dataset, he wants to use information from patients' medical records to predict whether they will get diabetes. He proposes using an *identity link function* along with a *Bernoulli likelihood model*.

   Briefly explain why this is not a good idea.

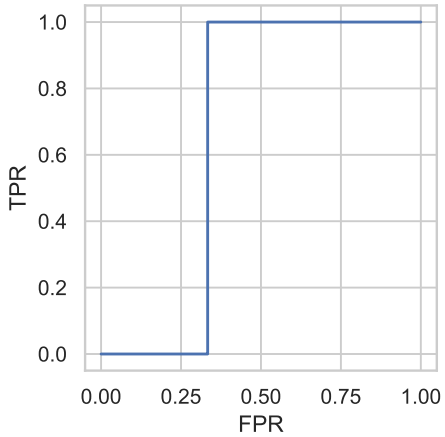
   Solution: This is **not** a good choice because the Bernoulli likelihood function expects a number between 0 and 1 as input, but there is no guarantee that the identity link function returns such a number. When working with a Bernoulli likelihood, you should use a sigmoid inverse link function, which guarantees that the input to the Bernoulli likelihood will be between 0 and 1. Note that a GLM with a sigmoid inverse link function and Bernoulli likelihood corresponds to logistic regression.

3. (6 points) After graduating, Alyona starts a tea cafe next to campus. To attract more business from students, she will try a targeted marketing campaign that offers discounts and promotions to students based on their favorite drink. To do so, she first needs a classifier which will tell her whether a student likes tea. Assume that $y = 0$ corresponds to not liking tea, and $y = 1$ corresponds to liking tea.

   Alyona uses social media to mine data about students, and trains a logistic regression model to predict tea preferences. She tests her model on four students, obtaining the following results along with an ROC curve.

   She wants to choose a threshold z and make a decision of 1 (student likes tea) when $f(X) > z$.

Y	f(X)
0	0.7
0	0.1
1	0.4
0	0.3



   (a) (1 point) Suppose Alyona sets a threshold $z = 0.2$, and makes a decision 1 (student likes tea) when $f(X) > z$. What is the resulting false discovery proportion (FDP)?

   Solution: Alyona makes 3 discoveries, but only one of them is a true discovery, so the FDP is $\frac{2}{3}$.

   (b) (2 points) Provide a threshold value such that the resulting false discovery proportion is *lower* than the FDP you computed in part a. If there is no such threshold value, state "no such threshold exists."

   Solution: Any $.3 \leq z < .4$ would result in one true discovery and one false discovery, leading to an FDP of $\frac{1}{2}$.

(c) (1 point) Provide a threshold value that corresponds to the point $(0.3, 1.0)$ on the ROC curve. In other words, determine a threshold z such that Alyona makes a decision 1 when $f(X) > z$, and her decisions have the given FPR and TPR.

Solution: Any value for z between $0.3 \leq z < 0.4$ would lead to 2 discoveries, leading to FPR of $\frac{1}{3} \approx 0.3$, and a TPR of $\frac{1}{1} = 1$.

Note: With only 3 values in the dataset for which $Y = 0$, technically an FPR of .3 is impossible—the point of interest was really $\left(\frac{1}{3}, 1.0\right)$. Answers stating "there is no such threshold value," providing above reason were awarded full credit.

(d) (2 points) Alyona's friend thinks that this is all too much work, and suggests instead making predictions completely at random: in other words, her friend predicts that each student likes tea with probability 0.5, independent of all other students.

For at least one of the points on the ROC curve above, the expected FPR and TPR for her friend's approach would be strictly better than the classifier. Which point(s) are those? Select all answers that apply.

☐ A. $(0, 0)$ ☐ B. $(0.3, 0)$ ☐ C. $(0.3, 1)$ ☐ D. $(1, 1)$

Solution: For a random classifier, the expected TPR and FPR will both be 0.5, since all points are classified randomly. A random classifier would, on average, correctly label 2 of the 4 data points in our training set. Point A correctly classifies the 3 tea-drinkers by making no discoveries. Point B makes 1 discovery but still correctly classifies 2 students. Point C makes 2 discoveries and correctly classifies the one coffee-drinker and 2 of the tea-drinkers. Point D makes 3 discoveries and only correctly classifies the coffee-drinker and one of the tea-drinkers. And point E only correctly classifies the tea-drinker. Therefore, points A, B, and C get more than 2 correct labels, so those thresholds are better than random.

4. (7 points) Rinzen works with researchers testing different drugs. The researchers will provide him with p-values for each drug trial's hypothesis test, and he must decide for which tests to reject the null hypothesis.

   (a) (2 points) He has not calculated the P-values yet, but he is deciding between using the Bonferroni or Benjamini-Hochberg procedures, each with parameter α. Fill in the blanks in the sentence below, using "Bonferroni" for one and "Benjamini-Hochberg" for the other:

   The number of discoveries made by the _____ procedure will be greater than or equal to the number of discoveries made by the _____ procedure, if the corresponding error rate for both procedures is set to α.

   Solution: 1) Benjamini-Hochberg, and 2) Bonferroni.
   An error rate of α for Benajmini-Hochberg controls FDR, while an error rate of α for Bonferroni controls FWER. The Bonferroni threshold is α/N. This is the smallest possible threshold for Benjamini-Hochberg (if only the first p-value is below the line $k\alpha/N$.
   So, any null hypothesis that is rejected by Bonferroni must have $p_i \leq \alpha/N < k\alpha/N$ for any k that denotes P_i's index in the sorted list.

   (b) (2 points) Now, Rinzen obtains the p-values for the 6 drug trials from the researchers. They are $P_1 = \frac{1.5\alpha}{6}$, $P_2 = \frac{3.1\alpha}{6}$, $P_3 = \frac{5.2\alpha}{6}$, $P_4 = \frac{5.5\alpha}{6}$, and $P_5 = \frac{2.5\alpha}{6}$ and $P_6 = \frac{4.5\alpha}{6}$, for some α between 0 and 1/2.
   He correctly uses the Benjamini-Hochberg procedure to control the false discovery rate at level α. For which p-values does he make a discovery? Select all answers that apply. **You must show your work to receive credit.**
   Hint: he makes at least one discovery, so "none of the above" is not a correct answer.

   ☐ A. P_1 ☐ B. P_2 ☐ C. P_3 ☐ D. P_4 ☐ E. P_5 ☐ F. P_6

   Solution: All tests hypotheses are rejected because when sorted, the largest value is $5.5\alpha/6 < \alpha$, so all earlier hypothesis are also rejected.

   (c) (3 points) One of the researchers made a mistake while calculating their p-value and sends Rinzen the new, correctly calculated, p-value.
   Rinzen still uses Benjamini-Hochberg to control the FDR at level α. This time, five of the p-values are the same as in part (b), but exactly one has changed. What is the **smallest** number of discoveries he could make? State which p-value would change

and its new value for this to happen (for example, you might say "Three discoveries if P_6 changes to 0.3") and **justify your choice by showing your work.**

Solution: Only P_4, the largest p-value, was within the critical value. Therefore, if this p-value changed to be greater than $6\alpha/6$ (i.e. greater than α), then no discoveries would be made. In this scenario, the smallest number of discoveries he could make would be 0.

Also, if any other p-value changed to be larger than α, that would reduce the number of discoveries to 0 as well. This is because $P_4 = \frac{5.5\alpha}{6} > \frac{5\alpha}{6}$, so it would be above the line.

5. (12 points) Isaac is trying to figure out how many discussion worksheets to print for the first in-person discussion section in Data 102. He decides to use Bayesian modeling and inference to answer this question.

   Isaac wants to estimate the unknown quantity q, which he defines as **the probability that any given student in Data 102 attends in-person discussion section**. He knows there are n students in the course. So, if he knew what q was, he could print nq worksheets.

   (a) (2 points) Isaac guesses that about 40% of students will attend in-person discussion section. However, he is not very confident in this belief: while he still thinks 40% is most likely, he suspects that anywhere between 30% and 50% of students attending section is also reasonably likely, and it's possible that it could be outside of that range. Which of the following prior distributions best reflects Isaac's beliefs? Choose the single best answer by filling in the circle next to it.

   ◯ A. Uniform(0.3, 0.5)
   ◯ B. Uniform(0, 1)
   ◯ C. Uniform(0.3, 1)
   ◯ D. Uniform(0, 0.5)
   ◯ E. Beta(9, 6)
   ◯ F. Beta(6, 9)
   ◯ G. Beta(90, 60)
   ◯ H. Beta(60, 90)

   (b) (3 points) Isaac then collects data from the first discussion section of the semester (which was held remotely), and finds that x students attended. He assumes x follows a Binomial(n, q) distribution, where $n = 300$ is the number of students in the class and on the waitlist.

   He calculates the posterior distribution for q, $p(q|x)$, and plans to use this to determine how many worksheets to print out.

   Based only on the information provided so far, which of the following assumptions is Isaac making? Select all answers that apply.

   ☐ A. Each student's decision about whether to attend is independent of all other students.
   ☐ B. The probability of attending, q, is a fixed and unknown parameter.
   ☐ C. The probability of attending remote discussion section is the same as the probability of attending in-person discussion section.
   ☐ D. Waitlisted students and enrolled students have the same probability of attending.
   ☐ E. The probability of attending, q, decreases over time.
   ☐ F. The number of students who attended the first discussion section, x, is equal to nq.

Solution: TODO explain here

(c) (2 points) Suppose Isaac chooses a Beta$(4, 4)$ prior for q. He observes $x = 100$ students attended discussion, out of 300 students total. What is the posterior distribution $q|x$? Express your answer in the form of a well-known distribution, along with its parameters (for example, Normal$(3, 2)$).

Solution: Beta(104, 204). Recall the Beta-Binomial conjugate pair. With a q following a Beta(α, β) prior, and x having a Binomial(n, q) likelihood, the posterior distribution for $q|x$ is a Beta($\alpha + x$, $\beta + n - x$) distribution. Plugging in appropriate values for α, β, x, and n, we get a Beta(104, 204) distribution.

(d) (3 points) Isaac wants a point estimate of q so that he can calculate how many worksheets to print. He knows that if he wants to minimize the Bayesian posterior risk with mean squared error loss, he should choose the mean of the posterior distribution, $E_{q|x}[q|x]$ (LMSE estimate). Isaac thinks about the tradeoffs in printing too many or too few worksheets, and decides to use the following loss function for true value q and estimate $\hat{q}$:

$$\ell(q, \hat{q}) = \begin{cases} 2(\hat{q} - q)^2 & \text{if } \hat{q} \geq q \\ (\hat{q} - q)^2 & \text{otherwise} \end{cases}$$

He minimizes the Bayesian posterior risk with this loss function and obtains a value q^*.

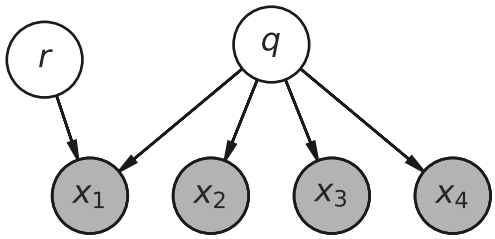
Will q^* be greater than, less than, or equal to the LMSE estimate? Choose the single best answer by filling in the circle next to it. **You must justify your answer to receive credit.**

◯ A. Greater than ◯ B. Equal to ◯ C. Less than

Solution: Isaac's new loss function penalizes large values more: in other words, if he overestimates, the loss will be higher than if he underestimates. So, in order to minimize this loss, we must pick a value below the LMSE estimate (because values above the LMSE estimate will incur double the loss).

(e) (2 points) For this part only, assume that Isaac has data from the first four discussion sections x_1, x_2, x_3, x_4, where x_1 is from the week when sections were held remotely, and all the others were held in person. In particular, x_i is the total number of students that attend discussion section on week i.

He decides to add a new (scalar) variable r to his model, and draws the following graphical model:



He shares his graphical model with the other GSIs, but forgets to describe what r means. Which of the following are reasonable interpretations of r based on the graphical model?

- ☐ A. r represents the effect that holding sections remotely has on attendance.
- ☐ B. r represents the effect that time of day has on attendance.
- ☐ C. r represents the effect that the first week of classes has on attendance.
- ☐ D. r represents the effect that different GSIs have on attendance.

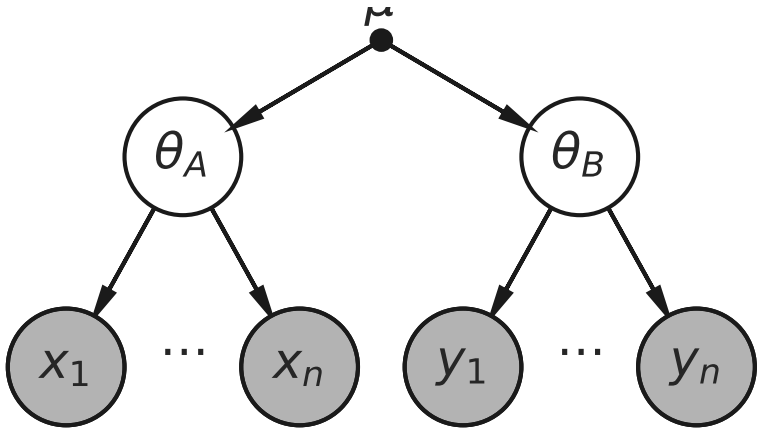
Solution: The first and third options are compatible with r, since they both will only affect attendance in the first week of class.

6. (6 points) As part of a class project, you are simulating the number of students that attend Data 102 office hours. You decide on the following:

   - Let θ_A be the rate of attendance on weeks that homework is due, and θ_B be the rate of attendance on weeks that homework is not due.
   - Let $x_1, \ldots, x_n$ be the number of students who attended on n randomly chosen days from weeks that HW is due.
   - Let $y_1, \ldots, y_m$ be the number of students who attended on m randomly chosen days from weeks that HW is not due.
   - You assume each x_i is drawn from a Poisson(θ_A) distribution, and each y_i is drawn from a Poisson(θ_B) distribution.
   - You assume that θ_A and θ_B are independently drawn from exponential distributions with parameter μ: $\theta_A \sim \text{Exp}(\mu)$ and $\theta_B \sim \text{Exp}(\mu)$.

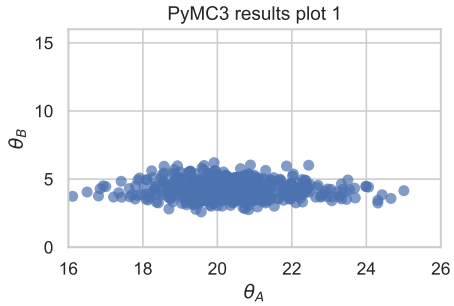
   (a) (3 points) Draw a graphical model for the probability model described above.

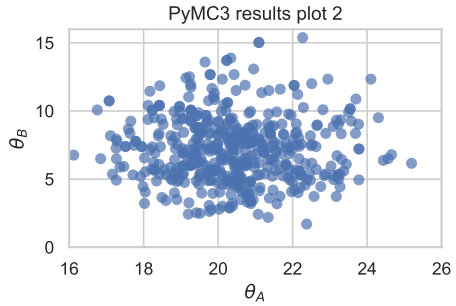
Solution:



(b) (3 points) You collect data from 60 days (30 from weeks that HW is due and 30 from weeks that HW is not due). Your friend implements the model from part (a) in PyMC3, and tries to produce a scatterplot where each point represents one posterior sample, and shows the values of θ_A and θ_B for that sample.

Unfortunately, your friend's Jupyter notebook is very disorganized, and he gives you two versions of the plot. One was computed using only two y-values (instead of 30), and the other was computed correctly using all of them. He doesn't know which is which, and sends you both of them.





Which of the following are valid conclusions from these plots? Select all answers that apply.

Hint: Recall that the mean of an Exp(μ) distribution is $1/\mu$.

- ☐ A. The plot on the right is more likely to have the full data than the plot on the left.
- ☐ B. The rate of student attendance is higher during weeks that HW is due.
- ☐ C. The LMSE estimate for θ_A is higher than the LMSE estimate for θ_B.
- ☐ D. The value of μ used must have been between $1/8$ and $1/2$.

Solution:

This page intentionally left blank for scratch work. No work on this page will be graded.