Data 102, Fall 2022 Midterm 1

- You have 110 minutes to complete this exam. There are 5 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 two blank pages of scratch paper, one at the beginning and one at the end of the exam. No work on these pages 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: ______________________________

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

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) If our null hypothesis is "the means of groups A and B are equal" and our alternative hypothesis is "the mean of group A is larger than group B", then we have enough information to use the likelihood ratio test.

   ○ A. TRUE [✓ ○ B. FALSE]

   (b) (1 point) When using a frequentist GLM, the estimated coefficient values that we compute from the training data are random (not fixed) quantities.

   [✓ ○ A. TRUE] ○ B. FALSE

   (c) (1 point) If using a GLM to fit continuous data with high variance, negative binomial regression is usually the best choice.

   ○ A. TRUE [✓ ○ B. FALSE]

   (d) (1 point) The choice of prior distribution has no effect on the width of the credible interval.

   ○ A. TRUE [✓ ○ B. FALSE]

   (e) (1 point) For any Bayesian model, we can use a posterior predictive check to evaluate how well the model represents the training data.

   [✓ ○ A. TRUE] ○ B. FALSE

2. (9 points) **The one with even more beetles**

Cindy decides to classify pictures of beetles that she took in her yard. She determines two species that she is most interested in, A and B, and wants to estimate which one is more common in her yard. She defines the following:

- A parameter q, which represent the probability of observing beetles belonging to species A in her yard.
- Data $x_1, \ldots, x_n$, indicating whether each picture she took is of species A (i.e., if x_i is 1, the picture is of species A, and if $x_i = 0$, the picture is of species B)

Cindy assumes that the x_i are i.i.d., and each $x_i \sim \text{Bernoulli}(q)$.

(a) (1 point) She observes 10 beetles, and correctly computes the maximum likelihood estimate for q as 0.8. How many beetles of species A did she observe?

Solution: The MLE estimate is (number of A beetles) / (total number of beetles). The denominator is 10, so the numerator must be 8.

(b) (3 points) Cindy's neighbor Mindy decides to try a Bayesian approach, and arbitrarily decides to choose the following prior for q_A:

$$q_A \sim \text{Beta}(6, c)$$

Mindy observes 18 beeetles of species A and 2 beetles of species B. Find a numeric value for c such that Mindy's LMSE/MMSE estimate will equal 0.8.
Hint: For a Beta(α, β) distribution, the expectation is $\frac{\alpha}{\alpha+\beta}$ and the mode (most likely value) is $\frac{\alpha-1}{\alpha+\beta-2}$.

Solution: The posterior distribution will be $\text{Beta}(6 + 18, 2 + c)$, and its mean must be 0.8:

$$\begin{aligned}
\frac{6+18}{6+18+2+c} &= 0.8 \\
24 &= 0.8(26 + c) \\
c &= \frac{24 - (0.8)(26)}{0.8} \\
c &= 30 - 26 = \boxed{4}
\end{aligned}$$

(c) (2 points) Cindy looks online and finds the following information on how common different species of beetles are in her county.

Species	Frequency in county
Species A	60%
Species B	20%
Species C	10%
Species D	5%

She decides to use this information to take a Bayesian approach to estimating q_A. Based on this information, which of the following are appropriate choices for a prior distribution for q? Select all answers that apply.

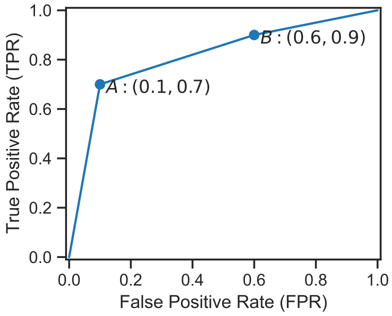
- □ A. $\text{Beta}(6, 2)$
- □ B. $\text{Beta}(2, 6)$ $\text{Beta}(12, 4)$
- □ C. $\text{Beta}(4, 12)$
- □ D. $\text{Normal}(6, 2)$
- □ E. $\text{Normal}(2, 6)$
- □ F. $\text{Normal}(12, 4)$
- □ G. $\text{Normal}(4, 12)$
- □ H. None of the above

(d) (3 points) Cindy's online group has k people in different locations who also have the same two species of beetle in their yards. They decide to use a hierarchical model to model beetle proportions in their yard, with shared parameters a and b, and separate probabilities for each yard, $q_1, \ldots, q_k$. They use the notation x_{ij} to represent the i-th beetle from the j-th person's yard.

Assuming there are two people in the group ($k = 2$) and each person observes three beetles ($n = 3$), draw a graphical model that represents the relationships between all the qs and all the xs.

Note: When shading in random variables that are observed, please only shade the edge of the circle, so that we can read your answer after scanning.

3. (7 points) A phone manufacturer develops an algorithm to flag each incoming text message as spam or not. A decision of 1 corresponds to spam, and a decision of 0 corresponds to not spam. If a message is flagged as spam, it is not shown to the user. They test their algorithm on a large database of text messages, and obtain the following ROC curve.



After interviews with users, they decide that they will choose a version of the algorithm that corresponds to one of the two labeled points, A or B.

(a) (2 points) For each of the following, fill in the blank next to the sentence with which version of the algorithm that person would prefer. For example, if the blank said "Person who prefers version B", then you would write "B" in the blank.

__________ Salesperson who receives many text messages from new clients and needs to see all of them to do their job

__________ Person who never uses their phone to send or receive text messages with any of their friends or other contacts

Solution: A, B

For the rest of the question, assume that the true prevalence of spam messages is q.

(b) (2 points) Compute the **precision** for version B. You may leave your answer as a numeric or algebraic expression involving q: you do not need to simplify. If you do not have enough information to solve the question, explain in plain English what additional information you would need.

Hint: precision is equal to $1 - FDP$.

Solution:

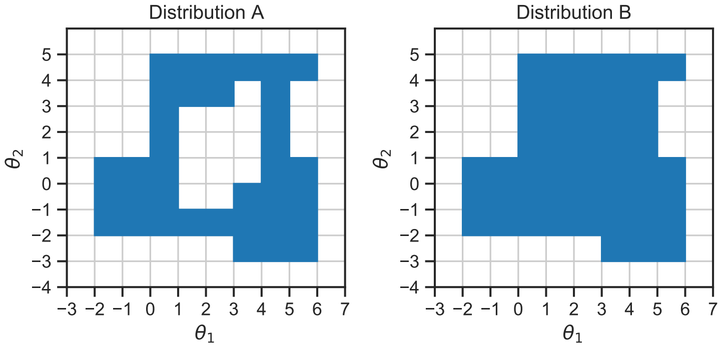
$$\begin{aligned}
\text{Precision} &= P(R=1|D=1) \\
&= \frac{P(D=1|R=1)P(R=1)}{P(D=1)} \\
&= \frac{P(D=1|R=1)P(R=1)}{P(D=1|R=1)P(R=1)+P(D=1|R=0)P(R=0)} \\
&= \frac{0.9q}{0.9q+0.6(1-q)}
\end{aligned}$$

(c) (3 points) The manufacturer determines that the loss of incorrectly flagging each non-spam message as spam is \$1, and the loss of incorrectly flagging each spam message is \$0.05. The loss for a correct decision is \$0. Compute the expected loss of version B. You may leave your answer as a numeric or algebraic expression involving q: you do not need to simplify. If you do not have enough information to solve the question, explain in plain English what additional information you would need.

Solution:

$$\begin{aligned}\text{Expectedloss} &= \text{lossforFP} \times P(D=1, R=0) + \text{lossforFN} \times P(D=0, R=1) + 0 + 0 \\ &= 1 \times P(D=1|R=0)P(R=0) + 0.05 \times P(D=0|R=1)P(R=1) \\ &= 0.6(1-q) + (0.05)(0.1)q\end{aligned}$$

4. (7 points) Consider two distributions over θ_1 and θ_2 that are uniform over the shaded regions in the graphs below. We will approximate these distributions with samples.



(a) (2 points) Suppose we use rejection sampling to approximate distribution A (left). Which of the following are valid proposal distributions? Select all answers that apply.

□ A. $\theta_1 \sim \text{Uniform}(-3, 7)$ and $\theta_2 \sim \text{Uniform}(-3, 7)$

□ B. $\theta_1 \sim \mathcal{N}(0, 2)$ and $\theta_2 \sim \mathcal{N}(0, 2)$ $\theta_1 \sim \text{Uniform}(-2, 6)$ and $\theta_2 \sim \text{Uniform}(-3, 5)$

□ C. $\theta_1 \sim \text{Uniform}(-2, 6)$ and $\theta_2 \sim \text{Uniform}(-2, 5)$

(b) (2 points) Suppose we use rejection sampling to approximate each distribution (A and B), choosing a correct proposal distribution from the choices above, and any other parameters optimally to get the largest possible proportion of accepted samples for each.

For which distribution will we obtain a higher proportion of accepted samples? Choose the single best answer by filling in the circle next to it. Explain your answer in two sentences or less. **You must explain your answer to receive credit.**

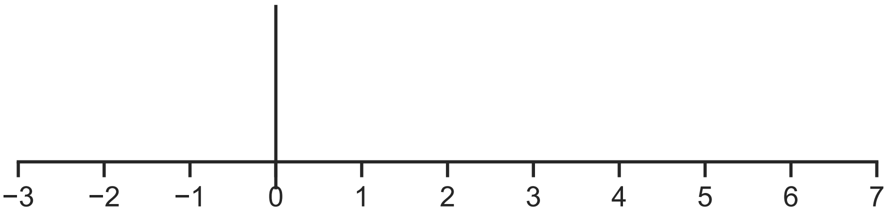
○ A. A

○ B. B

○ C. They will be the same

Solution: Distribution B has more probability mass in the center of the distribution, and will be more closely matched to any of the correct proposal distributions above. With distribution A, we will reject all samples in the center, but with distribution B, we are likely to accept many more of them.

(c) (3 points) Suppose we use Gibbs sampling to approximate distribution A (left). If $\theta_2^{(t)} = 0.5$, draw the conditional distribution that we should use to sample the next value of $\theta_1^{(t+1)}$. To receive full credit, you must label your axes and specify the height of the normalized density.



Solution: Union of Uniform(-2,1) and Uniform(4,6) with a height of 0.2

5. (12 points) Ilham writes a popular newsletter with a large number of non-paying subscribers, and wants to increase the number of subscribers who pay for a premium subscription. He tries an A/B test of k different interventions on k separate groups of subscribers. For each one, he obtains a p-value for the following hypothesis test:

   - H_0: The intervention **has no effect on** the number of people who pay for a premium subscription after reading the newsletter
   - H_1: The intervention **increases** the number of people who pay for a premium subscription after reading the newsletter.

   For parts (a) and (b) only, assume $k = 20$.

   Note: The choices available for parts (a) and (b) are exactly the same.

   (a) (2 points) For this part, he uses the **Bonferroni** procedure to control the family-wise error rate at $\alpha = 0.1$. You may assume that all assumptions necessary for the Bonferroni procedure are satisfied. Which of the following statements are true? Select all answers that apply.

   □ A. The expected number of discoveries he will make is 2.

   □ B. Out of the discoveries he makes, exactly 10% of them (rounded to the nearest integer) will be incorrect.

   □ C. For the interventions where the null hypothesis is true, the expected proportion of correct decisions he will make is 0.1.

   □ D. The probability that at least one of his discoveries is incorrect is 0.1.

   □ E. The expected number of discoveries that will be incorrect is 2.

   (b) (3 points) For this part, he uses the **Benjamini-Hochberg (B-H)** procedure to control the false discovery rate at $\alpha = 0.1$. You may assume that all assumptions necessary for the B-H procedure are satisfied. Which of the following statements are true? Select all answers that apply.

   □ A. The expected number of discoveries he will make is 2.

   □ B. Out of the discoveries he makes, exactly 10% of them (rounded to the nearest integer) will be incorrect.

   □ C. For the interventions where the null hypothesis is true, the expected proportion of correct decisions he will make is 0.1.

   □ D. The probability that at least one of his discoveries is incorrect is 0.1. The expected number of discoveries that will be incorrect is 2.

(c) (2 points) For this part only, Ilham decides to try the interventions sequentially, so he uses the LORD algorithm to make the decisions online. He is deciding whether to try his best ideas (most likely to convince people to pay) first, or his worst ideas first. Under which conditions is LORD more likely to make more discoveries? Choose the single best answer by filling in the circle next to it.

○ A. Trying the best ideas first

○ B. Trying the worst ideas first

For the remainder of the question, let $k = 4$. The p-values for the first 3 interventions are as follows: $0.8\alpha, 0.3\alpha, 0.6\alpha$ for some number α where $0 \leq \alpha \leq 0.5$.

(d) (2 points) Suppose we use **Bonferroni correction** with parameter α. Specify all possible values for the fourth p-value such that there will be no discoveries, or explain why no such values exist.

Solution: For Bonferroni correction, our p-value threshold will be $\alpha/4$. All of the provided p-values are above this threshold, so we require that the last p-value is also above this threshold: $\boxed{p_4 > \alpha/4}$

(e) (3 points) Suppose we use the **Benjamini-Hochberg procedure** with input parameter α. Specify all possible values for the fourth p-value such that there will be no discoveries, or explain why no such values exist.

Solution: In sorted order, the p-values so far are $0.3\alpha, 0.6\alpha, 0.8\alpha$. The critical values will be $0.25\alpha, 0.5\alpha, 0.75\alpha, \alpha$. If the last p-value is less than 0.8α, then the last p-value in sorted order will be 0.8α, which will be below the last critical value α, and we will make four discoveries. If the last p-value is greater than 0.8α, then the first three p-values are above the corresponding critical values. Therefore, the last p-value must be greater than the last critical value: $\boxed{p_4 > \alpha}$

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