Data 102, Fall 2021 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). 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 of the exam and one near the end. 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:

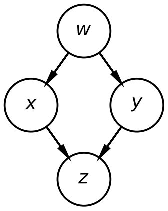
- All of the work submitted here is my original work.
- I did not collaborate with anyone else on this exam.

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, circle either **TRUE** or **FALSE**.

   (a) (1 point) (TRUE / FALSE) In the following graphical model, $x \perp\!\!\!\perp y \mid w$ (that is, x and y are conditionally independent given w):



   Solution: True: w is a common parent of both x and y.

   (b) (1 point) (TRUE / FALSE) When using a GLM, the negative binomial likelihood model is always a better choice than the Poisson likelihood model.

   Solution: False: the Poisson model is simpler and has fewer parameters to fit, and could be a good fit if the data is not overdispersed.

   (c) (1 point) (TRUE / FALSE) When using a binary test with a true positive rate (TPR) of 0.8 and a false positive rate (FPR) of 0.2, the false discovery proportion (FDP) will always be less than 0.2.

   Solution: FALSE: the FDP depends on the prevalence: it could be greater than, equal to, or less than 0.2 depending on it.

   (d) (1 point) (TRUE / FALSE) Gibbs sampling rejects more samples for higher-dimensional problems than for lower-dimensional problems.

   Solution: FALSE: Gibbs sampling never rejects samples, and zero is not more than 0.

   (e) (1 point) (TRUE / FALSE) At each step of Gibbs sampling, we randomly sample one hidden variable conditioned on all the other hidden variables and the data.

   Solution: TRUE: This is the definition of Gibbs sampling.

2. (10 points) Ilin is in charge of collecting results for a large research lab. She works with researchers who provide her with p-values, and her job is to determine which ones correspond to real scientific discoveries. Over the next several years, the researchers will provide her with some number of p-values.

   For parts (a) and (b), assume that she analyzes the p-values online. In other words, after seeing each p-value, she must make a decision, and she cannot ever change that decision in the future.

   (a) (2 points) For this part only, assume she knows in advance that exactly 500 total tests will be performed. Of the techniques below, which of the following can she use to produce decisions online (in other words, which ones can provide a decision after each p-value without ever changing that decision in the future) while guaranteeing that the false discovery rate will be less than or equal to 0.05? Select all that apply.

   (A) Bonferroni

   (B) Benjamini-Hochberg

   (C) LORD

   (D) None of the above

   Solution: A, C

   Bonferroni (A) controls FWER, and we can use it online as long as we know the number of tests in advance. Since FDR≤FWER, it controls FDR too.

   B-H (B) can't be used online without changing decisions: we need to re-run the algorithm after seeing a new p-value in order to control FDR, but this could change old results.

   LORD (C) is designed for online FDR control: we proved it in class.

   (b) (3 points) Suppose she uses LORD to control the false discovery rate at $\alpha = 0.1$. Which of the following must be true? Select all that apply.

   (A) The family-wise error rate (FWER) is also guaranteed to be less than or equal to 0.1.

   (B) The false negative rate is also guaranteed to be less than or equal to 0.1.

   (C) For any given test where the null is true, the probability of rejecting the null increases with the number of discoveries made in the 10 previous tests.

   (D) None of the above

   Solution: C

For parts (c)-(d), suppose that she waits until all the p-values are available before analyzing them all together. She receives a total of five p-values: 0.06, 0.02, 0.01, 0.005, and 0.035.

(c) (2 points) If she wants to control the family-wise error rate at 0.06 using what she learned in Data 102, how many discoveries will she make?

Solution: Two
$\frac{\alpha}{n} = \frac{0.06}{5} = 0.012$, so we reject the third and fourth p-values

(d) (3 points) This time, she instead uses Benjamini-Hochberg to control the false discovery rate α.

Fill in the blanks to make the below statement correct. You do not need to simplify any algebraic or arithmetic expressions in the blanks, but you must show your work.

To receive full credit, you must provide the smallest correct answer for the first blank and the largest correct answer for the second blank.

If α is greater than or equal to ____________ *and less than* ____________*, then she will make exactly three discoveries.*

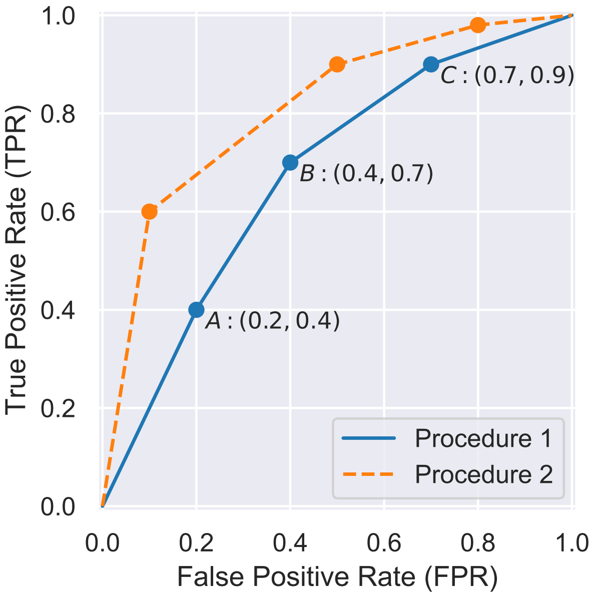
Solution: $\frac{5}{3}(0.02) = \frac{1}{30}$, $\frac{5}{4}(0.035) = \frac{7}{160}$ The p-values in sorted order are 0.005, 0.01, 0.02, 0.035, 0.06.

To make exactly three discoveries, the B-H line must fall **below** the fourth point and **at or above** the third point. So, we must have:

$$0.035 < \frac{4\alpha}{5}$$
$$0.02 \geq \frac{3\alpha}{5}$$

Solving for alpha, we obtain the values above.

3. (9 points) Mazi works for a medical device company. He is trying to predict whether manufactured devices will be defective based on photographs of the assembly line. He sets up a binary decision problem similar to what he learned in Data 102, where $D = 1$ corresponds to predicting that a device is defective. He designs two different procedures to predict defects. He tests the procedures on a dataset of N devices, where m are defective and $N - m$ are not defective. He generates the following ROC curve using scikit-learn, similarly to what was done in lecture.



Three points on the ROC curve for procedure 1 have been labeled as A, B, and C, along with their exact FPR and TPR values.

(a) (2 points) What is the false negative rate (FNR) for point A? You do not need to simplify any algebraic or arithmetic expressions.

Solution: The false negative rate (FNR) is $1 - \text{TPR} = 1 - 0.4 = 0.6$

(b) (3 points) What is the false discovery proportion (FDP) for point B? You do not need to simplify any algebraic or arithmetic expressions.

Solution: We know that the prevalence is $P(R = 1) = m/N$. So:

$$\begin{aligned} FDP = P(R = 0|D = 1) &= \frac{P(D = 1|R = 0)P(R = 0)}{P(D = 1|R = 1)P(R = 1) + P(D = 1|R = 0)P(R = 0)} \\ &= \frac{FPR \times (N - m)/N}{TPR \times m/N + FPR \times (N - m)/N} \\ &= \frac{0.4 \times (N - m)/N}{0.7 \times m/N + 0.4 \times (N - m)/N} \end{aligned}$$

(c) (2 points) Suppose that it's much worse to miss a defective device than to mistakenly predict that a good device is defective. In this case, of the three points on the ROC curve for Procedure 1, which should Mazi prefer? Choose the one best answer.

(A) Point A

(B) Point B

(C) Point C

Solution: C. Point C

A false negative (missing a defective device) is much worse than a false positive (predicting that a good device is defective). So, we should maximize TPR ($= 1 - FNR$), even if it means accepting a high FPR.

(d) (2 points) Which procedure is better (more accurate)? Select only one answer. If you choose option (D), explain your answer (you do not need to explain your answer if you choose any other choice).

(A) Procedure 1 (solid line)

(B) Procedure 2 (dashed line)

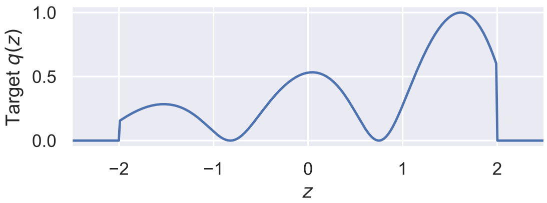
(C) They are exactly the same

(D) Procedure 1 is better in some cases and procedure 2 is better in some cases (if you select this option, write one sentence describing when each procedure is better)

Solution: B. Procedure 2

The ROC curve for procedure 2 is strictly above and to the left: for any given TPR we wish to achieve, Procedure 2 will have a lower FPR than Procedure 1. Similarly, for any given FPR, Procedure 2 will have a higher TPR.

4. (5 points) We want to use sampling to approximate the unnormalized distribution $M \cdot q(z)$ over random variable z, for some constant M:



(a) (2 points) Now suppose we use rejection sampling. Which of the following proposal distributions generate valid proposals for rejection sampling with this distribution? Assume that for each one, we choose the best possible value of M. Choose all that apply.

(A) Uniform$(-1, 2)$

(B) Uniform$(0, 4)$

(C) Uniform$(-2, 2)$

(D) Uniform$(-2.5, 2.5)$

(E) Normal$(0, 1)$

Solution: C, D E

A and B do not cover the full support of the distribution: with A we'll never generate proposals less than -1, and with B we'll never generate proposals less than 0. The others can all generate proposals between -2 and 2.

TODO explain rubric

(b) (3 points) Of the correct choices from the previous section, which one will lead to the highest proportion of accepted samples, and why? Assume that for each one, we choose the best possible value of M.

You must justify your answer to receive full credit.

Solution: C: Uniform$(-2, 2)$

For choice D, 20% of the samples will be rejected no matter what (-2.5 to -2 and 2 to 2.5). For choice E, we must choose an extremely small value of M so that the peak around $z = 1.5$ fits below the standard normal curve. This means that most of the samples will be rejected, since the ratio $Mq(z)/p(z)$ will usually be very small.

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

5. (11 points) Fabiola works at a store and wants to understand shopper behavior. When shoppers enter the store, they take either a basket or a cart. She wants to estimate the probability that they'll take a cart. She gathers data from n randomly selected shoppers. She calls these $x_1, \ldots, x_n$.

For parts (a) - (c), assume that shoppers always take either a cart or a basket (that is, they never enter the store without one or the other, and they never take both).

Let θ be the probability that a shopper takes a cart (and $1-\theta$ be the probability that they take a basket). Fabiola decides to model each shopper's decision as a Bernoulli random variable ($x_i = 1$ if shopper i takes a cart and 0 if they take a basket). She also decides to use a Beta prior for θ:

$$\begin{aligned} x_i|\theta &\sim \text{Bernoulli}(\theta) \\ \theta &\sim \text{Beta}(a, b) \end{aligned}$$

(a) (3 points) Suppose she observes data from 6 shoppers as follows: basket, basket, basket, cart, basket, cart. Give numeric values for a and b such that the maximum a posteriori (MAP) estimate for θ will be 1/4. You must show your work to receive credit.

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 is Beta$(a+2, b+4)$ by conjugacy. The MAP estimate is the mode of the posterior, which must be 1/4. So:

$$\begin{aligned} \frac{a+2-1}{a+2+b+4-2} &= \frac{1}{4} \\ \frac{a+1}{a+b+4} &= \frac{1}{4} \\ 4a+4 &= a+b+4 \\ 3a &= b \end{aligned}$$

Any pair (a, b) of positive numbers that satisfies this equation received full credit. We also gave full credit for plugging in values in the second equation to obtain a ratio of 1/4.

The harder way to solve this problem was to find the maximum of the posterior directly, instead of using the provided formula. We gave full credit for people who applied this approach correctly too. We know that $p(\theta|x_1, \ldots, x_6) \propto \theta^{a-1+2}(1-\theta)^{b-1+4}$. So, we can take the derivative of the log-likelihood and set it equal to 0, then solve for θ:

$$\begin{aligned} \frac{d}{d\theta} \log p(\theta|x_1, \ldots, x_6) &= \frac{d}{d\theta}\left[(a+1)\log\theta + (b+3)\log(1-\theta)\right] \\ 0 &= \frac{a+1}{\theta} - \frac{b+3}{1-\theta} \\ \theta_{MAP} &= \frac{a+1}{a+b+4} \end{aligned}$$

From here, the rest of the solution is the same as above.

(b) (2 points) If Fabiola believes that customers are much more likely to choose carts than baskets, which of the following is the **best** choice for values of a and b?

(A) $a = 30, b = 1$

(B) $a = 3, b = 1$

(C) $a = 3, b = 3$

(D) $a = 1, b = 3$

(E) $a = 1, b = 30$

Solution: A

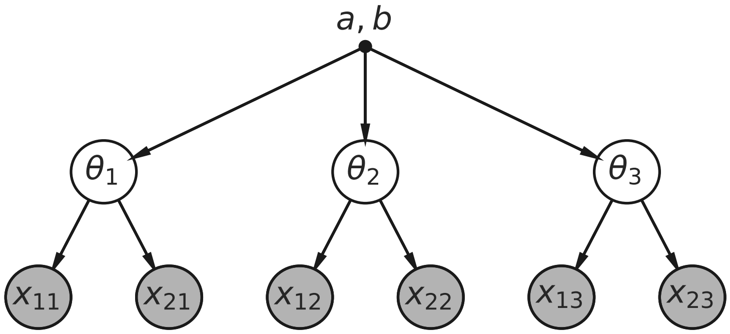
Choices A and B will lead to a prior distribution where large values of θ are more likely, which corresponds to shoppers being more likely to choose carts. A, the stronger prior, is a better choice because she believes they are **much** more likely to choose carts.

(c) (3 points) **For this part only**, Fabiola decides to apply her approach to multiple stores. Since the stores are all in the same geographical area and are of similar size and purpose, she believes that the cart/basket proportion for each of the m stores will be only slightly different. She decides to use a hierarchical model, with shared parameters α and β, and separate cart probabilities for each store, $\theta_1, \ldots, \theta_m$. For the data, she uses the notation x_{ij} to represent the i-th shopper from the j-th store (for example, x_{23} represents the second shopper from store 3).

Assuming that there are three stores ($m = 3$) and two people sampled for every store ($n = 2$), draw a graphical model that describes the relationships between all the θs 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.

Solution:



For a hierarchical model with shared parameters, this model is the best choice. Note that a and b are not random variables, so they should be marked with a dot, square, or diamond, rather than a circle, but we did not take off points for this.

(d) (3 points) **For this part, Fabiola goes back to modeling data for only one store.**

She decides to change her model to consider shoppers who don't take a cart or a basket (in other words, they enter the store with neither). She uses the following probability model for x_i:

$$p(x_i|\theta_c, \theta_b, \theta_n) = \begin{cases} \theta_c & \text{if } x_i = \text{cart} \\ \theta_b & \text{if } x_i = \text{basket} \\ \theta_n & \text{if } x_i = \text{neither} \end{cases},$$

with the constraint that $\theta_c + \theta_b + \theta_n = 1$. For her prior over these three parameters, she uses a (three-dimensional) **Dirichlet distribution** with parameters α_c, α_b, and α_n:

$$p(\theta_c, \theta_b, \theta_n) \propto \theta_c^{\alpha_c} \theta_b^{\alpha_b} \theta_n^{\alpha_n} \tag{1}$$

Suppose she observes N_b shoppers who use baskets, N_c shoppers who use carts, and N_n shoppers who use neither. Show that the posterior distribution for θ_c, θ_b, and θ_n is also a Dirichlet distribution, and express its parameters in terms of the α and N variables.

Hint: For convenience, you may abbreviate c for cart, b for basket, and n for neither. It may be helpful to express the likelihood in terms of indicator functions $\mathbb{1}(x_i = c)$, $\mathbb{1}(x_i = b)$, and $\mathbb{1}(x_i = n)$.

Solution: We begin by rewriting the likelihood using the hint:

$$p(x_i|\theta_c, \theta_b, \theta_n) = \theta_c^{\mathbb{1}(x_i=c)} \theta_b^{\mathbb{1}(x_i=b)} \theta_n^{\mathbb{1}(x_i=n)}$$

Now, we can write the posterior:

$$\begin{aligned} p(\theta_c, \theta_b, \theta_n | x_1, \dots, x_n) &\propto p(\theta_c, \theta_b, \theta_n) \left[\prod_i p(x_i|\theta_c, \theta_b, \theta_n) \right] \\ &\propto \theta_c^{\alpha_c} \theta_b^{\alpha_b} \theta_n^{\alpha_n} \left[\theta_c^{N_c} \theta_b^{N_b} \theta_n^{N_n} \right] \\ &\propto \theta_c^{\alpha_c + N_c} \theta_b^{\alpha_b + N_b} \theta_n^{\alpha_n + N_n} \\ &= \text{Dirichlet}(\alpha_c + N_c, \alpha_b + N_b, \alpha_n + N_n) \end{aligned}$$

Note that the question has a very small mistake: the Dirichlet distribution is usually specified as

$$p(\theta_c, \theta_b, \theta_n) \propto \theta_c^{\alpha_c - 1} \theta_b^{\alpha_b - 1} \theta_n^{\alpha_n - 1},$$

which would introduce -1s throughout the solution but not change the final result.