Data 102, Fall 2023
Midterm 1

- You have **110 minutes** to complete this exam. There are **6 questions**, totaling **50 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 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 something so we can give you partial credit.
- We have provided a blank pages of scratch paper at the beginning of the exam. No work on this page will be graded.
- You may, without proof, use theorems and facts given in the discussions or lectures, **but please cite them**.
- We don't answer questions individually. If you believe something is unclear, bring your question to us and if we find your question valid we will make a note to the whole class.
- 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	
Berkeley email	
Name of person to your left	
Name of person to your right	

Honor Code [1 pt]:
As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. I am the person whose name is on the exam, and I completed this exam in accordance with the Honor Code.

Signature: _____

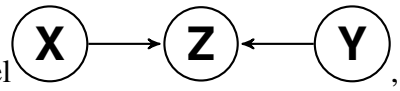
This page has been intentionally left blank.

1 True or False [5 Pts]

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.

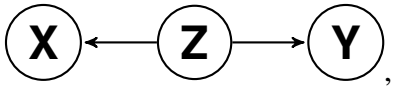
(a) [1 Pt] Define the Positive Predictive Value (PPV) as $\frac{\text{TP}}{\text{TP+FP}}$. This is a column-wise rate.

◯ True ◯ False

(b) [1 Pt] In the graphical model 

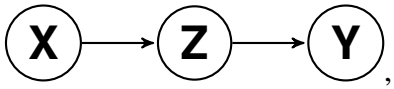
, the random variables X and Y are conditionally independent given Z.

◯ True ◯ False

(c) [1 Pt] In the graphical model 

, the random variables X and Y are conditionally independent given Z.

◯ True ◯ False

(d) [1 Pt] In the graphical model 

, the random variables X and Y are conditionally independent given Z.

◯ True ◯ False

(e) [1 Pt] For rejection sampling to generate samples from the Beta$(4, 1)$ density, it is more efficient to use Uniform$[0, 1]$ as the proposal density compared to Beta$(2, 1)$.

◯ True ◯ False

2 Project Cybersyn [6 Pts]

Dr. Allende builds a model called *Cybersyn* to classify patients' tumor scans as benign (0) or malignant (1).

(a) [2 Pts] The first iteration of the model has an overall accuracy of 65% and a False Discovery Proportion (FDP) of 20%. However, Dr. Allende lost a sheet of paper containing the confusion matrix. He can only remember the two quantities shown below.

Help him complete the confusion matrix. You may use the box below for scratch work, but no work will be graded.

		Decision	
		0	**1**
Reality	**0**	5	
	1		8

(b) [2 Pts] To better adjust the model between false negatives and false positives, Dr. Allende defines the following loss function:

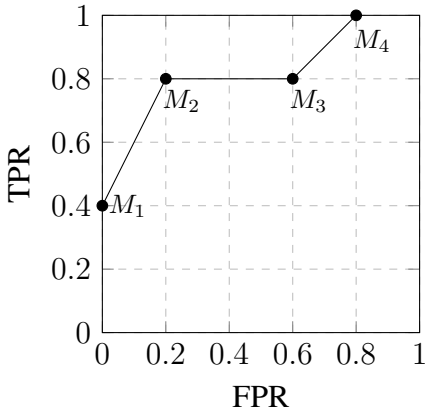
$$\begin{cases} \ell(D = 0 \mid R = 1) = 1 \\ \ell(D = 1 \mid R = 0) = k \\ \ell(D = 0 \mid R = 0) = \ell(D = 1 \mid R = 1) = 0, \end{cases}$$

where D represents decision and R represents reality.

Which of the following statements is/are true? Select all that apply.

- ☐ A. If $k > 1$, Dr. Allende thinks classifying a malignant tumor as benign is worse than classifying a benign tumor as malignant.
- ☐ B. If $k = 0$, the minimum loss is achieved by classifying every tumor as malignant.
- ☐ C. If $k = 0$, the classifier that minimizes the loss will have an FDP of $1 - \pi_1$, where π_1 is the prevalence of malignant tumors in the dataset.
- ☐ D. If $0 < k < 1$, the minimum loss is achieved by classifying every tumor as benign.

(c) [2 Pts] Dr. Allende trains four models M_1, M_2, M_3, M_4 and plots their FPR and TPR using the following ROC curve.



Let n_0 and n_1 be the number of benign tumors and malignant tumors in the dataset. Which of the following is/are true? Select all that apply.

- ☐ A. If $n_0 = n_1$, M_2 is the model with the highest accuracy.
- ☐ B. If $n_0 = n_1$, all four classifiers have a higher accuracy than any random classifier.
- ☐ C. There exists some n_0 and n_1 such that M_1 is the model with the highest accuracy.
- ☐ D. There exists some n_0 and n_1 such that M_3 is the model with the highest accuracy.

3 Making Bad Decisions [13 Pts]

Juliet is a big fan of the band *The Strokes*. She collects vinyl records of the band's album *The New Abnormal*. However, vinyl records worn out as you play it. Juliet decides to see how long it lasts.

Let the lifetime of the record be T (in hours). She has two hypotheses on T:

- Null Hypothesis (H_0): $T \sim$ Exponential($\frac{1}{60}$)
- Alternative Hypothesis (H_1): $T \sim$ Exponential($\frac{1}{120}$)

(a) [2 Pts] Which of the following is/are true about Juliet's hypothesis testing scheme? Select all that apply.

- ☐ A. The null hypothesis is a simple hypothesis.
- ☐ B. The alternative hypothesis is a composite hypothesis.
- ☐ C. Out of all possible tests with significance level α, the Likelihood Ratio Test maximizes the True Positive Rate (TPR).
- ☐ D. Out of all possible tests with significance level α, the Likelihood Ratio Test maximizes the True Negative Rate (TNR).

(b) [2 Pts] Juliet wants to design a Likelihood Ratio Test. Which of the following is the Likelihood Ratio (LR)? Select the only correct option and **show all your work in the provided box**.

- ◯ A. LR$(T) = \frac{1}{2}e^{T/120}$
- ◯ B. LR$(T) = 2e^{T/40}$
- ◯ C. LR$(T) = \frac{1}{2}e^{-T/40}$
- ◯ D. LR$(T) = 2e^{-T/120}$

(c) [5 Pts] Recall that the Likelihood Ratio Test rejects the null hypothesis for a data point T if $\text{LR}(T) \geq \eta$, for some threshold value η.

Juliet has **10** copies of the same record and decides to test the hypothesis for each record. To control the Family-wise Error Rate (FWER) of the 10 tests at 0.1, she uses Bonferroni correction.

Derive the threshold η. Simplify your answer as much as possible. Show all your work and fill in the blank.

Hint: You don't need to compute any integral. The CDF of $X \sim \text{Exponential}(\lambda)$ is $1 - e^{-\lambda x}$.

$\eta =$ _____

(d) [2 Pts] Instead of FWER, Juliet now wants to control the False Discovery Rate (FDR) of the 10 tests at 0.1. She decides to use the Benjamini-Hochberg procedure. In particular, she calculates the p-values for the 10 data points and sorts them in non-decreasing order: $P_{(1)}, P_{(2)}, \ldots, P_{(10)}$, where $P_{(1)}$ is the smallest and $P_{(10)}$ is the largest.

Which of the following is/are true? Select all that all apply.

- ☐ A. If $P_{(10)} \leq 0.1$, she rejects H_0 for all data points.
- ☐ B. If $P_{(1)} = P_{(2)} \cdots = P_{(10)} > 0.1$, she fails to reject H_0 for all data points.
- ☐ C. If $P_{(1)}, \ldots, P_{(10)} < 0.01$, she makes less rejections compared to part (c).
- ☐ D. If $0.01 < P_{(3)} < 0.03$, she makes more rejections compared to part (c).

(e) [2 Pts] Now instead of knowing the lifetime of all 10 records at the same time and conducting hypothesis testing, Juliet plays the records one by one until each one is worn out. She wants to make a decision immediately after seeing each lifetime. Assume she knows in advance that there are 10 records. Which of the following multiple testing strategies can she use? Select all that apply.

- ☐ A. Naive Thresholding
- ☐ B. Bonferroni Correction
- ☐ C. Benjamini-Hochberg Procedure
- ☐ D. LORD

4 Surgical Survival Rates [7 Pts]

A new hospital just opened up in our neighborhood and we are interested in the survival rate θ for a high-risk operation at this new hospital. The historical survival rates for this procedure at $N = 10$ nearby hospitals are given by

$$0.90, 0.99, 0.87, 0.75, 0.81, 0.90, 0.99, 0.96, 0.74, 0.93$$

(a) [2 Pts] We would like to convert this data from nearby hospitals into a prior density for the survival rate θ for this operation in this new hospital. Which of the following densities presents a suitable prior? Select the best option.

◯ A. Beta$(9, 1)$. ◯ B. Beta$(1, 9)$.
◯ C. Beta$(90, 10)$. ◯ D. Beta$(10, 90)$.

Suppose your selected prior from the previous part is Beta(a, b). Now consider the additional information: n **patients are operated in this new hospital and all of them survived**. Assume that the n surgeries are **independent** of each other, given the survival rate θ. Answer the following questions in terms of a, b and n.

(b) [2 Pts] What is the posterior distribution of the survival rate θ at the new hospital? If it is a known distribution, provide the name and parameters; if not, provide the density function.

(c) [3 Pts] What is the probability that the next patient at this new hospital will survive the operation? State clearly the random variables you are using and the assumptions you are making to answer this question. Show all your work.

5 Height-based Penguin Data Correction [10 Pts]

Dr. Okoro is an biologist specializing in penguins. She maintains detailed records of the morphological features of various penguin species. However, an error led to a mix-up in her datasets for chinstrap and emperor penguins. To rectify this, she plans to use the `"height"` variable (stored in an array named `combined_penguin_heights`) to categorize the mixed records back into separate "chinstrap" and "emperor" groups. She's aware that emperor penguins, typically 120 cm tall, are taller than chinstrap penguins, which have an average height of 70 cm.

For the entire question, assume `numpy` as been imported as `np`, and `PyMC` as `pm`.

(a) [2 Pts] Consider the following PyMC model:

```
model_ONE = pm.Model()
with model_ONE:
    theta = pm.Uniform("theta", lower = -200, upper = 200)
    log_sigma = pm.Uniform("log_sigma",
                           lower = -50, upper = 50)
    sigma = pm.Deterministic("sigma",
                           pm.math.exp(log_sigma))
    Y = pm.Normal("Y", mu = theta, sigma = sigma,
                  observed = combined_penguin_heights)
    idata = pm.sample(2000, chains = 2)
```

Draw a graphical model for `model_ONE` in terms of θ (`theta`), σ (`sigma`), and Y_1, Y_2, Y_3, assuming the number of records in `combined_penguin_heights` is 3.

(b) [2 Pts] Would `model_ONE` help carry out Dr. Okoro's task?

- If yes, select the variable whose posterior samples returned by `PyMC` will help us separate the penguins and describe how you would do it.
- If not, select option E and explain why not.

Your answer should be in two sentences or less.

◯ A. `theta` ◯ B. `log_sigma` ◯ C. `sigma` ◯ D. `Y` ◯ E. N/A

(c) [3 Pts] Consider the PyMC model:

```
N = len(combined_penguin_heights)
model_TWO = pm.Model()
with model_TWO:
    w = pm.Uniform("w", lower = 0, upper = 1)
    thetas = pm.Normal("thetas", mu = np.array([70, 120]),
                       sigma = 20, shape = 2)
    z = pm.Bernoulli("z", p = w, shape = N)
    Y = pm.Normal("Y", mu = thetas[z],
                      sigma = 15,
                      observed = combined_penguin_heights)
    idata = pm.sample(1000, chains = 2)
```

Draw a graphical model for `model_TWO` in terms of w, `thetas` (θ_0, θ_1), z_1, z_2, z_3, and Y_1, Y_2, Y_3, assuming the number of records in `combined_penguin_heights` is 3.

(d) [3 Pts] Would `model_TWO` help carry out Dr. Okoro's task?

- If yes, select the variable whose posterior samples returned by `PyMC` will help us separate the penguins and describe how you would do it.
- If not, select option E and explain why not.

Your answer should be in two sentences or less.

◯ A. `w` ◯ B. `theta` ◯ C. `z` ◯ D. `Y` ◯ E. N/A

6 A Crime Dataset [8 Pts]

This problem concerns a dataset named `crime` on arrests from the Introductory Econometrics book by Wooldridge. This dataset contains information for $n = 2725$ adult men on the following variables:

- `narr86` (y): Number of arrests in the year 1986 (this variable equals zero for 1970 of the 2725 men in the dataset)
- `pcnv` (x_1): Proportion of previous arrests that led to a conviction
- `tottime` (x_2): Total time (in months) in prison since turning 18
- `inc86` (x_3): Legal income in 1986 (in hundreds of dollars)
- `qemp86` (x_4): Number of quarters employed in 1986
- `black` (x_5): Binary variable which equals 1 if the individual is black and 0 otherwise

For this entire question, assume `statsmodels.api` has been imported as `sm`.

(a) [2 Pts] We use the following code to fit a model to this dataset:

```
Y = crime['narr86']
X = crime[['pcnv', 'tottime', 'inc86',
           'qemp86', 'black']].copy()
X = sm.add_constant(X)
model_ONE = sm.OLS(Y, X).fit()
model_ONE.summary()
```

Write down the math equation(s) for the model fitted by the above code in terms of $y, x_1, \ldots, x_5$ and weights $\beta_0, \ldots, \beta_5$.

(b) [2 Pts] The summary for `model_ONE` is given in Table 1.

Table 1: `model_ONE` summary

	coef	std err	t	P$>|t|$	[0.025	0.975]
const	0.5825	0.034	17.071	0.000	0.516	0.649
`pcnv`	-0.1441	0.041	-3.549	0.000	-0.224	-0.064
`tottime`	0.0005	0.004	0.143	0.886	-0.006	0.007
`inc86`	-0.0016	0.000	-4.775	0.000	-0.002	-0.001
`qemp86`	-0.0349	0.014	-2.456	0.014	-0.063	-0.007
`black`	0.2700	0.044	6.078	0.000	0.183	0.357

In this table, the "coef" for the variable `black` is given as 0.27. How would you interpret this coefficient?

(c) [2 Pts] We fit another model to this dataset using the code below:

```
Y = crime['narr86']
X = crime[['pcnv', 'tottime', 'inc86',
        'qemp86', 'black']].copy()
X = sm.add_constant(X)
model_TWO = sm.GLM(Y, X, family=sm.families.Poisson()).fit()
model_TWO.summary()
```

Write down the math equation(s) for the model fitted by the above code in terms of $y, x_1, \ldots, x_5$ and weights $\beta_0, \ldots, \beta_5$.

(d) [2 Pts] The summary for `model_TWO` is given in Table 2.

Table 2: `model_TWO summary`

	coef	std err	z	P$>|z|$	[0.025	0.975]
const	-0.5271	0.058	-9.053	0.000	-0.641	-0.413
`pcnv`	-0.5000	0.084	-4.987	0.000	-0.586	-0.255
`tottime`	0.0004	0.006	0.074	0.941	-0.010	0.011
`inc86`	-0.0086	0.001	-8.288	0.000	-0.011	-0.007
`qemp86`	-0.0030	0.029	-0.105	0.916	-0.059	0.053
`black`	0.5000	0.069	7.189	0.000	0.360	0.629

Which of the following interpretations of the results is/are true? Select all that apply.

- ☐ A. The "coef" for the variable `black` is given as 0.5, this means that the average number of arrests in 1986 for a black man is roughly 45% higher than a non-black man, if we hold the values of other variables fixed.
- ☐ B. The average number of arrests for each individual goes down by roughly 5% if the probability of conviction is slightly increased by 0.1, if we hold the values of other variables fixed.
- ☐ C. According to the model, there is a strong positive association between an individual's the total time spent in prison and their number of arrests in 1986.
- ☐ D. The two most important factors (among `pcnv`, `tottime`, `inc86`, `qemp86`, `black`) that are associated with `narr86` are `inc86` and `black`.

Hint: Below are some powers of e:

x	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$y = e^x$	1.05	1.11	1.22	1.35	1.49	1.65	1.82	2.01	2.23	2.46	2.72

7 Congratulations [0 Pts]

Congratulations! You have completed Midterm 1.

- **Make sure that you have written your student ID number on *every other page* of the exam.** You may lose points on pages where you have not done so.
- Also ensure that you have **signed the Honor Code** on the cover page of the exam for 1 point.
- If more than 10 minutes remain in the exam period, you may hand in your paper and leave. If $\leq$ 10 minutes remain, please **sit quietly** until the exam concludes.

[Optional, 0 pts] What's your favorite joke?

Midterm 1 Reference Sheet

Algorithm 1 The Benjamini-Hochberg Procedure

Inputinput FDR level α, set of n p-values $P_1, \ldots, P_n$ Sort the p-values $P_1, \ldots, P_n$ in non-decreasing order $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(n)}$
Find $K = \max\{i \in \{1, \ldots, n\} : P_{(i)} \leq \frac{\alpha}{n} i\}$
Reject the null hypotheses (declare discoveries) corresponding to $P_{(1)}, \ldots, P_{(K)}$

Useful Distributions:

Distribution	Support	PDF/PMF	Mean	Variance	Mode
$X \sim \text{Poisson}(\lambda)$	$k = 0, 1, 2, \ldots$	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	$\lfloor \lambda \rfloor$
$X \sim \text{Beta}(\alpha, \beta)$	$0 \leq x \leq 1$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1}$	$\frac{\alpha-1}{\alpha+\beta-2}$
$X \sim \text{Gamma}(\alpha, \beta)$	$x \geq 0$	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
$X \sim \mathcal{N}(\mu, \sigma^2)$	$x \in \mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$	μ	σ^2	μ
$X \sim \text{Exponential}(\lambda)$	$x \geq 0$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	0

Conjugate Priors: For observations x_i, $i = 1, \ldots, n$:

Likelihood	Prior	Posterior
$x_i \vert \theta \sim \text{Bernoulli}(\theta)$	$\theta \sim \text{Beta}(\alpha, \beta)$	$\theta \vert x_{1:n} \sim \text{Beta}\left(\alpha + \sum_i x_i, \beta + \sum_i (1 - x_i)\right)$
$x_i \vert \mu \sim \mathcal{N}(\mu, \sigma^2)$	$\mu \sim \mathcal{N}(\mu_0, 1)$	$\mu \vert x_{1:n} \sim \mathcal{N}\left(\frac{\sigma^2}{\sigma^2+n}\left(\mu_0 + \frac{1}{\sigma^2}\sum_i x_i\right), \frac{\sigma^2}{\sigma^2+n}\right)$
$x_i \vert \lambda \sim \text{Exponential}(\lambda)$	$\lambda \sim \text{Gamma}(\alpha, \beta)$	$\lambda \vert x_{1:n} \sim \text{Gamma}\left(\alpha + n, \beta + \sum_i x_i\right)$

Generalized Linear Models

Regression	Inverse link function	Likelihood
Linear	identity	Gaussian
Poisson	exponential	Poisson

Some powers of e:

x	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$y = e^x$	1.05	1.11	1.22	1.35	1.49	1.65	1.82	2.01	2.23	2.46	2.72