Data 102, Fall 2024
Midterm 2

- You have **110 minutes** to complete this exam. There are **7 questions**, totaling **54 points**.
- You may use **two** 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 page of scratch paper in the **middle** 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: _____

1 True/False and Multiple Choice (6 pts)

Choose the single best answer for each question **by filling in the circle next to it.**

(a) [1 Pt] Given a prediction problem, selecting the model with the best log-likelihood is equivalent to picking the most likely model given the observed data.

◯ True ◯ False

(b) [1 Pt] If W is a valid instrumental variable for estimating the causal effect of Z on Y, then we can compute an unbiased estimate of the ATE by estimating coefficients from two separate specific linear regression models and subtracting the results.

◯ True ◯ False

(c) [1 Pt] Suppose we compute the gradients of $L_1(a, b) = \sin(a^2 + b^2 - \log(a/b))$ and $L_2(a, b) = \frac{a}{7b^4}$, with and without backpropagation. Then the improvement from backpropagation (i.e., number of fewer redundant operations) will be better for L_2 than L_1.

◯ True ◯ False

(d) [1 Pt] To apply the Chernoff bound to a random variable X, we need to know the moment generating function $M_X[\lambda] = \mathbb{E}[e^{\lambda X}]$ for some subset of λ where $M_X[\lambda] < \infty$.

◯ True ◯ False

(e) [1 Pt] Suppose we want to estimate the difference between a population's mean and median. If we gather a very large dataset ($n > 5000$) and estimate this with the sample mean and sample median, then using the bootstrap to produce a confidence interval is a good choice for quantifying uncertainty in our estimate.

◯ True ◯ False

(f) [1 Pt] Consider a sequence of i.i.d. random variables $x_1, \ldots, x_n$, all of which have finite variance. If we use Chebyshev's inequality to produce a confidence interval for the sample mean $y = \frac{1}{n}\sum_{i=1}^{n} x_i$, then the width of that interval will be proportional to:

◯ $\frac{1}{n}$ ◯ $\frac{1}{\sqrt{n}}$ ◯ $\frac{1}{n^2}$ ◯ $\sqrt{\log\left(\frac{1}{n}\right)}$ ◯ None of these

2 GLM or no GLM? (5 pts)

(a) [5 Pts] Consider a prediction model with predictor variables $\{x_1, x_2\}$ and outcome y.

Fill in the missing entries in the table below.

- Mark the "Is a GLM?" column with "Yes" if the model defines a GLM, and "No" if not. *If a model does not define a GLM, then don't fill in the remaining columns of the row associated with the model.*
- Fill in the "Link" column with the corresponding link function name (for example, "logit").
- For the column labeled "Likelihood: $p(y|\beta, x)$" write in the conditional distribution of y given the predictors x, regression coefficients β, and/or intermediate values such as $\hat{y}(x)$. Include any other parameters specified in the model. For example, you may write Bernoulli$(\hat{y}(x))$.

Model	Is a GLM?	Link	Likelihood $p(y\|\beta, x)$
$\hat{y}(x) = \beta_1 x_1 + \beta_2 \sin(x_2)$ $\epsilon \sim \mathcal{N}(0, \sigma^2)$ $y = \hat{y}(x) + \epsilon$			
$\hat{y}(x) = \exp(\beta_1 x_1) + \exp(\beta_2 x_2)$ $y \sim$ Poisson$(\hat{y}(x))$			
$\hat{y}(x) = \tan^{-1}(\beta_1 x_1 - \beta_2 x_2)$ $\epsilon \sim$ Uniform$([-\delta, \delta]), \quad \delta > 0$ $y = \hat{y}(x) + \epsilon$			

3 Heart Disease Prediction (7 pts)

In this question, we'll use the following dataset to build a model that predicts whether patients have heart disease:

	HeartDisease	Age	Sex	RestingBP	Cholesterol	MaxHR	ExerciseAngina
0	0	40	M	140	289	172	N
1	1	49	F	160	180	156	N
2	0	37	M	130	283	98	N
3	1	48	F	138	214	108	Y
4	0	54	M	150	195	122	N

We evaluate three frequentist logistic regression models for predicting `HeartDisease`, each using a different subset of only the `Age`, `Cholesterol`, `RestingBP` and `MaxHR` features.

The coefficients from the models, along with the log-likelihood for each, are listed in the table below. If a feature is marked "-", then it was not used in the model.

Name	`Age`	`RestingBP`	`Cholesterol`	`MaxHR`	Log-Likelihood
Model 1	4.8	2.62	6	0.6	-1377.5
Model 2	7.4	-	9.2	-	-1378.1
Model 3	-	-	-	9.2	-1623.1

(a) [2 Pts] In the space below, identify which model you would select using AIC. (*Hint: you can compute the AIC from the information provided. If you do not solve for the AIC, then you should explain your answer.*) What does the AIC account for that the log-likelihood alone does not?

For the remainder of the question, we implement a Bayesian logistic regression model with all four numeric predictors and priors chosen based on consulting with medical experts. We fit the model on a training set. You may assume that the posterior over all coefficients is jointly unimodal.

(b) [2 Pts] For each of the following terms, choose whether (A) or (B) best describes its meaning. In each of the sentences below the "unknown" stands for any particular regression coefficient you are attempting to estimate.

(A) The interval is a range that probably contains the unknown given our data.

(B) The interval is chosen so that, if the process was repeated for different independent draws of the data, then the interval would usually contain the unknown.

(i) Confidence Interval: ◯ (A) ◯ (B)

(ii) Credible Interval: ◯ (A) ◯ (B)

(c) [1 Pt] We find that the 95% Highest Density Interval (HDI) for the coefficient of `MaxHR` is $[-0.13, 1.72]$. Which of the following statements must be true? Select all answers that apply **by filling in the square next to each correct answer**.

☐ The posterior samples forming this credible interval are independent from the corresponding posterior sample values for `Cholesterol`

☐ A 99% HDI for the same coefficient under the same model is guaranteed to contain the value 1.71.

☐ The 95% HDI is unique.

(d) [2 Pts] Suppose that you divide your data set into a training set, $(X_{\text{train}}, Y_{\text{train}})$ and a testing set $(X_{\text{test}}, Y_{\text{test}})$. Then, assume that you have defined a Bayesian logistic regression model with unknown regression coefficients β. Assume that you have access to a tool that can sample from the posterior defined by your Bayesian model.

Explain a two-step process for sampling from the Posterior Predictive Distribution (PPD) given X_{train}. At each step, briefly specify the distribution used for sampling and name a relevant sampling technique. If you cannot explain how to sample, define the PPD for partial credit.

test$given the training pairs$(X˙train,Y˙train), $and the test features$X˙test.″

4 Stargazing (6 pts)

Sarah goes camping every night for 30 nights and counts the number of shooting stars she sees. She gathers some data to help her predict the number of shooting stars she can expect to see on a given night. The first two rows of her dataset are shown below:

	Elevation_ft	Light_Pollution_Index	Cloud_Cover_Percent	Number_of_Shooting_Stars
0	7270	0.91	72.7	9
1	860	6.18	32.7	3

(a) [2 Pts] Sarah uses Poisson regression and obtains the following coefficients:

$$\hat{\beta}_{\texttt{Elevation\_ft}} = 0.002, \quad \hat{\beta}_{\texttt{Light\_Pollution\_Index}} = -0.08, \quad \hat{\beta}_{\texttt{Cloud\_Cover\_Percent}} = -0.09$$

For a particular night with `Cloud_Cover_Percent`$= 43$, let s be the number of shooting stars she observes that night. The model's average prediction for s is 3.7.

If the cloud cover percent were to increase from 43 to 53, what distribution for s would the model predict? Write your answer as a random variable. For example, you might write $s \sim \text{Bernoulli}\left(\frac{\log(9.01)}{e^2}\right)$.

For the remainder of the question, Sara compares her Poisson regression model to a decision tree and a random forest. All three are trained on the same dataset, and all three have good prediction accuracy on the training set.

(b) [2 Pts] Sara evaluates all three models on another similar dataset gathered by one of her friends at a nearby location. Select *the single best* option in the sentences below by completely filling in the circle next to the correct answer.

Sara should expect the (◯ *decision tree*, ◯ *random forest*) to have higher prediction accuracy on the new dataset because bagging reduces the (◯ *bias*, ◯ *variance*) and helps prevent overfitting to the training dataset.

(c) [2 Pts] Now, Sarah not only wants to predict the number of shooting stars, but also wants to understand the most important factors that contribute to nights with many shooting stars.

Out of the three models (Poisson regression, decision tree, random forest), which is best suited to Sarah's task? Explain your answer in two sentences or less.

This page has been intentionally left blank. No work on this page will be graded.

5 Causal Inference (10 Points)

Professor X teaches an advanced course with many prerequisites, where every unit of the course depends on the prerequisite material. She creates an optional 4-hour intensive session designed to help students remember the prerequisite material. She wants to know how much better this session **causes** students to do in her course. She collects the following information about each student:

- P: Average GPA in all prerequisite courses (continuous, 0-4)
- S: Whether the student attended the session (binary)
- M: Student's percentage score on the first midterm (continuous, 0-100)
- F: Student's overall composite grade in the course, including midterms, final, etc. (continuous, 0-100)

She wants to estimate the causal effect of the session S on midterm 1 grades M.

(a) [2 Pts] Draw a causal Directed Acyclic Graph (DAG) representing the relationship between P, S, M, and F.

(b) [2 Pts] **For this part only**, assume students are randomly assigned to either be required to attend the session, or not allowed to attend, and that all students follow their random assignment.

Which of the following must be true? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ The overall composite grade F is a collider for this causal question.
- ☐ The midterm score potential outcomes $M(0)$ and $M(1)$ are (unconditionally) independent of session attendance S.
- ☐ Midterm score M is independent of session attendance S.
- ☐ If some students leave the session halfway through, then SUTVA could be violated.

(c) [3 Pts] She decides to use inverse propensity weighting to measure the causal effect of the session S on **midterm 1 grade** M, due to the presence of confounding variables. She computes the propensity score using logistic regression.

For this part, assume the four variables listed are the only relevant variables: in other words, that there are no other variables with causal effects related to what she is trying to study.

(i) (1pt) Which variable should she use as the target y for logistic regression? Choose the single best answer **by filling in the circle next to it. No work will be graded for this part.**

◯ P ◯ S ◯ M ◯ F

(ii) (2pt) Which variable(s) should she use as the predictor(s) X for her logistic regression? Select all answers that apply **by filling in the square next to each correct answer**. **You must justify your answer to receive credit.**

□ P □ S □ M □ F

Justification:

(d) [3 Pts] Professor X chooses some students at random before the semester starts, and offers them extra credit worth 3% of their overall course grade if they attend the session. Let E be whether each student received this offer. Her TAs propose using the offer E as an instrumental variable (IV) to determine the causal effect of the session S on **midterm 1 grade** M.

For each of the three assumptions needed for IVs, **briefly state the assumption, state whether E satisfies it (yes/no), and then briefly explain why.** Answers outside the provided space will not be graded.

6 Multi-Armed Brat-dits (11 pts)

Charli XCX has released different versions of songs from her album Brat onto a streaming platform. Each features a different artist. To maximize engagement, the platform chooses versions to promote, and tracks how many times a listener likes each promoted version.

The following table summarizes their data part way through their trial:

Featured artist	Times promoted	Times liked	Fraction liked
Billie Eilish	1500	350	0.23
Troye Sivan	2500	450	0.18
Lorde	2000	700	0.35

The platform wants to promote the most popular featured artist: the artist whose version is liked the most often when it is played.

(a) [2 Pts] Assuming no prior belief about each artist's popularity, identify **(i)** which artist they should promote if they prioritize exploration, and **(ii)** which they should promote if they prioritize exploitation. **To earn full credit, you must explain your answers without algebra.**

(b) [3 Pts] The platform plans to use the UCB algorithm. They design their confidence bounds to fail with probability at most δ. Select the true statements below regarding the intervals and the UCB criteria **by filling in the square next to each correct answer**.

- ☐ Decreasing δ will make the intervals narrower.
- ☐ To prioritize exploration over exploitation, the platform should decrease δ.
- ☐ For a fixed δ, observing more trials will make an interval narrower.

(c) [3 Pts] The platform considers Explore-then-Commit (ETC), Upper Confidence Bound (UCB), and Thompson Sampling (TS) as methods to maximize engagement. Each method has different theoretical properties. Select the correct choice in each parenthesis by fully shading in the bubble next to it.

1. ETC has (◯ *linear*, ◯ *logarithmic*) regret.
2. UCB has (◯ *linear*, ◯ *logarithmic*) regret and is (◯ *frequentist*, ◯ *Bayesian*).
3. TS has (◯ *linear*, ◯ *logarithmic*) regret and is (◯ *frequentist*, ◯ *Bayesian*).

Assume that the platform uses Thompson Sampling for the rest of this question. Based on past streaming data, they use a Beta(50, 100) prior for Billie Eilish, a Beta(50, 80) prior for Troye Sivan, and a Beta(500, 1000) prior for Lorde.

(d) [1 Pt] Given these priors, which featured artist is the platform most likely to promote first (before collecting the trial data in Table 1)? **Justify your answer. You may give a heuristic answer based on the separate exploitation and exploration objectives.**

(e) [2 Pts] Calculate the posterior distribution for Lorde given the data in Table 1 and the information above. Report your answer as a named distribution and specify its parameters (for example, Normal$(10, 3)$). Then, explain how Thompson Sampling would use the posteriors for all three artists to select an artist to promote.

Lorde posterior distribution:

Explanation of how TS would use all three posteriors to select an artist:

7 Painting Probabilities (8 pts)

Liana works as a freelance painter. She paints n houses in a month, and for each house i she paints, she has some costs x_i and some income y_i (both in dollars). She does some research and finds that $E[x_i] = \$750$ and $E[y_i] = \$1,000$. Let $C_n = \sum_{i=1}^{n} x_i$ be her total costs, and $I_n = \sum_{i=1}^{n} y_i$ be her total income for the month.

At the start of the month, Liana wants to know how much money she'll need to borrow so she can cover her total cost C_n. Let $t(n)$ be the amount of money she borrows.

For parts (a) and (b), assume $n = 6$. In all parts, partial work will receive partial credit. If you cannot simplify, set up each relevant inequality, then show how to solve for the requested value.

(a) [2 Pts] Using only the information above, find the smallest value of t so that she can be 95% sure that her total cost, C_n, will be less than t. Give your answer as an integer.

You must show your work to earn credit.

(b) [3 Pts] Liana can't find anyone to loan her the amount from part (a) each month, so she decides she'll only take on houses where her costs won't exceed $\$1025$. She also determines that for any house, her costs will always be at least $\$525$.

Using this information, find the smallest value of t so that she can be 95% sure that her total cost, C_n, won't exceed t. Give your answer as an integer. You may make the approximation $\log(0.05) = -3$.

You must show your work to earn credit.

(c) [3 Pts] Liana's income per house is between \$775 and \$1,275. As in part (b), her cost per house is between \$525 and \$1025. She does her best to match her price to the expense of the job, but cannot guarantee her costs until she is finished. Let $\bar{P}_n = \frac{1}{n}(I_n - C_n)$ denote her average profit per house if she completes n houses each month.

Find a lower bound on the number of houses, n, that Liana should paint to guarantee that her average profit $\bar{P}_n$ will be positive 95% of the time.

You may leave your answer as an algebraic expression for the lower bound on n. You must isolate n in your answer to receive full credit. For example, an implicit bound $\sqrt{1 + an^2} \geq b$ would not earn full credit, but an explicit bound $n \geq \sqrt{\frac{1}{a}(b^2 - 1)}$ would.

You must show your work to earn credit.

8 Congratulations [0 Pts]

Congratulations! You have completed Midterm 2.

- **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] Draw a picture or cartoon that's related to your favorite thing you've learned in Data 102 so far.

Midterm 2 Reference Sheet

Useful Distributions:

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

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

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

Generalized Linear Models

Regression	Inverse link function	Likelihood
Linear	identity	Gaussian
Logistic	sigmoid	Bernoulli
Poisson	exponential	Poisson
Negative binomial	exponential	Negative binomial

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