Data 102, Fall 2025
Midterm 2

- You have **110 minutes** to complete this exam. There are **6 questions**, totaling **50 points**.
- You may use **two** 8.5×11 sheets 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 **not** provided any scratch paper, but you may use the back of the reference sheet for this purpose.
- 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. **For questions where a box for work/answers is given, any work or answers outside the provided box will not be graded.**

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 or False and Multiple Choice [7 Pts]

For (a) - (d), determine whether the statement is true or false. For all parts of this question, no work will be graded and no partial credit will be assigned.

(a) [1 Pt] True or False: While formulating a GLM, Mortimer looks at the observed y-values for each x-value, and calculates the average $\bar{y}(x)$ and standard deviation $\sigma_y(x)$. If the values follow the relationship $\sigma_y(x) \approx \sqrt{\bar{y}(x)}$ for all values of x, then Poisson regression will be overdispersed.

◯ True ◯ False

(b) [1 Pt] True or False: For a multi-armed bandit problem, when deciding which arm to pull with Thomson sampling, we always choose the arm with the highest posterior mean.

◯ True ◯ False

(c) [1 Pt] True or False: Backpropagation is an algorithm for efficiently computing gradients.

◯ True ◯ False

(d) [1 Pt] True or False: When training a random forest on a training set with n data points, each tree is trained by sampling n data points with replacement and choosing a random subset of the features.

◯ True ◯ False

(e) [1 Pt] Consider a sequence of i.i.d. random variables $r_1, \ldots, r_m$ that only take on values between 0 and 1. If we use Hoeffding's inequality to construct a confidence interval for the sample mean $y = \frac{1}{m}\sum_{i=1}^{m} r_i$, then the width of the interval will be proportional to which of the following? Choose the single best answer **by filling in the circle next to it.**

◯ $\sqrt{\log(m^{-1})}$ ◯ m^{-2} ◯ m^{-1} ◯ $m^{-1/2}$ ◯ Can't use Hoeffding ◯ None of these

(f) [2 Pts] Consider a treatment S and outcome R, with a candidate instrumental variable T. For each statement below, determine whether it makes T an invalid or useless instrument. Select all answers that apply **by filling in the square next to each correct answer**.

□ $\text{cov}(T, R) = 0$

□ $\text{cov}(T, S) = 0$

□ Conditioned on S, T is **not** independent of at least one confounder.

2 Elaine's Boba, Again! [3 pts]

Elaine owns many Boba shops in the Bay Area. For each shop, she gathers dozens of features about the shop and surrounding neighborhood (parking, distance to transit, neighborhood demographics, etc.) and wants to predict the number of orders at that shop. She finds a large national database with similar information for over $5{,}000$ shops. She trains two different models on the national data, and then tests each one on data from her stores over the past two years. The two models are:

- Model A: a frequentist negative binomial GLM
- Model B: a random forest with 100 trees

(a) [3 Pts] For each of the scenarios described below, decide which model Elaine should use. If both models meet her requirements, or if neither is a good fit, select the corresponding option. Choose the single best answer **by filling in the circle next to it.**

(i) Elaine's top priority is explainability, and both models have comparable performance on the test set.

◯ Model A ◯ Model B ◯ Both ◯ Neither

(ii) Elaine discovers that there are complex nonlinear relationships between the features and outcome, and wants the most accurate model.

◯ Model A ◯ Model B ◯ Both ◯ Neither

(iii) Elaine wants a model where she can quantify the uncertainty in her model's predictions.

◯ Model A ◯ Model B ◯ Both ◯ Neither

3 The One with Beetle Predictions (10 pts)

On some randomly chosen days, Cindy goes out to count the number of beetles in her favorite park, calling this b. She also records the temperature f (in $^\circ F$), amount of rainfall in the last two days r (in inches), and the number of people in the park p. She records her data in a table. Here are the first two rows:

date	f (temperature)	r (rainfall)	p (people)	b (beetles)
2025/03/08	57	0.0	30	22
2025/05/01	50	2.3	3	17

For parts (a) and (b), Cindy uses a frequentist Poisson GLM to predict b (beetle count) using f (temperature) and p (number of people), with $\beta_f = 0.01$ and $\beta_p = -0.05$ as their respective coefficients, and $\beta_0 = 3$ as an intercept.

For this question, you should assume that for small values of x (i.e., $|x| < 0.1$), $\exp(x) = 1 + x$.

(a) [2 Pts] Write an expression for the probability, according to Cindy's fitted Poisson GLM, that she will observe 20 beetles on a day where the temperature is $62^\circ F$ and she observes 8 people. You do not need to simplify your expression, but it should contain only numbers and no variables (i.e., someone should be able to put it into a calculator to compute the result).

(b) [2 Pts] Which of the following are valid interpretations of the coefficients of Cindy's model? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ On a day with an average temperature and average number of people, the expected number of beetles observed is $\exp(3)$.
- ☐ Cindy's model predicts that for every additional person she observes, she will observe (on average) 5% fewer beetles.
- ☐ Cindy's model predicts that if she observes 20 fewer people, she will observe (on average) 1 fewer beetle.

For the remainder of this question, Cindy uses Bayesian linear regression to predict the temperature (f) from rainfall (r, with coefficient β_r), the number of people (p, with coefficient β_p), and an intercept β_0.

(c) [3 Pts] She computes a 90% credible interval for β_r, the coefficient of rainfall, $[-8.3, 0.1]$. Which of the following must be true? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ If the posterior distribution for β_r is unimodal (i.e., has only one mode), then this interval must be unique.
- ☐ Given her observations, there is a 90% probability that the coefficient is between -8.3 and 0.1.
- ☐ If she were to observe many different observed datasets and compute a credible interval for each one, then 90% of them would contain the fixed true value.
- ☐ If she computed the interval by using samples to approximate the posterior, then 10% of her sample values for β_r must have been smaller than -8.3 or larger than 0.1.

(d) [3 Pts] Cindy uses a posterior predictive check (PPC) on a held-out test set to evaluate her model. Let $(r_{train}, p_{train}, f_{train})$ be her training set, and let σ be the standard deviation of the likelihood. Fill in the blanks in the instructions/pseudocode below to help her conduct her PPC. Note that some blanks have hints beneath them.

(1) Use sampling to approximate the posterior distribution $p(\beta_r, \beta_p, \beta_0, \sigma | r_{train}, p_{train}, f_{train})$ with 1000 samples, calling each one $(\beta_r^{(t)}, \beta_p^{(t)}, \beta_0^{(t)}, \sigma^{(t)})$ for $t \in \{1, 2, \ldots, 1000\}$.

(2) For $t = 1, 2, \ldots, 1000$:

For each test set point $(r_j, p_j, f_j), j \in \{1, \ldots, N_{test}\}$:

(a) Compute the average prediction

$\hat{y} =$ _____.

(b) Draw a sample from the _____ distribution and store it.
DistributionName(parameter, parameter,...)

(3) Compare all the stored samples from all iterations of Step 2(b) to the observed

_____-values from the _____ set and see if the distributions are similar.
r, p, or f test or training

4 Jewel Bandits (10 pts)

Marc, a former Data 102 student, is now the leader of a large group of expert jewel thieves. Each week, the group steals a jewel according to the following process:

1. Pick one of four different large neighborhoods in the city
2. Choose a house in that neighborhood uniformly at random
3. Steal the most expensive jewel in the chosen house

For this question, assume that the thieves never get caught.

Marc assumes that each neighborhood has a different distribution of jewel value (how much he resells the stolen jewels for). He assumes that the jewels he steals will all be worth between $\$1,000$ and $\$501{,}000$. He decides to use multi-armed bandits to help choose the neighborhood with the best jewels (highest average resale value). The following table summarizes the group's performance for the first twenty weeks:

Neighborhood	Times chosen	Average stolen jewel price
1	4	$\$20{,}000$
2	8	$\$80{,}000$
3	6	$\$100{,}000$
4	2	$\$4{,}000$

(a) [2 Pts] List one assumption that Marc is making with multi-armed bandits that might **not** be satisfied by the setup above, and explain why not.

(b) [1 Pt] For the next week, which neighborhood should Marc pick if he is prioritizing **exploration**? Choose the single best answer **by filling in the circle next to it.**

◯ 1 ◯ 2 ◯ 3 ◯ 4

(c) [1 Pt] For the next week, which neighborhood should Marc pick if he is prioritizing **exploitation**? Choose the single best answer **by filling in the circle next to it.**

◯ 1 ◯ 2 ◯ 3 ◯ 4

(d) [2 Pts] Given the data for the observed pattern above, which of the following are possible algorithms that Marc could have been using for the first twenty-one weeks? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ Explore-then-commit (ETC)
- ☐ Upper confidence bound (UCB)
- ☐ Thomson sampling (TS)
- ☐ None of these

(e) [2 Pts] For this part only, Marc learns that some of the houses may contain extremely valuable jewels worth well over $\$1{,}000{,}000$, and he can no longer assume an upper bound of $\$501{,}000$ on the sale price of stolen jewels. Which of the following must be true, assuming he uses only what he learned in Data 102? Select all answers that apply **by filling in the square next to each correct answer**.

- ☐ If he uses an unbounded likelihood distribution (e.g., log-normal), then he can still use Thomson sampling.
- ☐ He can no longer use Hoeffding's inequality to construct upper bounds for the UCB algorithm.

(f) [2 Pts] The police interview every household in the city about the most expensive jewel in their home. For each neighborhood j, they compute v_j, the average value of the most expensive jewels in each house in that neighborhood. They find that $\max_j v_j = v_2$. Write an expression for Marc's pseudo-regret, using only the v_j variables and the numbers provided in the table above. If this cannot be computed from the information given, explain why.

Hint: for this part, you should assume that the neighborhoods are very large, and that there are no outliers in jewel prices within each one.

5 Causal Slop (12 Pts)

David runs a large video-sharing social media platform. The platform shows ads before some videos, and those ads can be skipped. He wants to know whether watching a higher proportion of AI-generated videos causes users to skip ads less often. For every user i ($i \in \{1, \ldots, n\}$, where n is very large), he defines:

- S_i: the ad skip rate (proportion of ads shown to that user that were skipped)
- A_i: the proportion of videos watched by the user that are AI-generated
- D_i: demographic data: age, gender, etc.
- P_i: user's political preferences

(a) [3 Pts] For this part only, assume that:

- David randomly assigns exactly half of all users to see exactly 80% AI-generated videos for a week, and the other half to see exactly 20% AI-generated videos for the same week.
- Users can only watch videos suggested by the platform, and they always watch every video suggested.

David computes y as described below. Which of the following must be true about y? Select all answers that apply **by filling in the square next to each correct answer**.

$$y = \left[\frac{1}{n/2}\sum_i S_i \cdot \mathbb{1}(A_i = 0.8)\right] - \left[\frac{1}{n/2}\sum_i S_i \cdot \mathbb{1}(A_i = 0.2)\right]$$

☐ If $y = 0.3$, then David should conclude that watching 80% AI-generated videos causes ad skip rates to increase by 0.3 (on average) compared to watching 20% AI-generated videos.

☐ If $y < 0$, then David should conclude that **any increase** in the proportion of AI-generated content shown to users will (on average) cause a decrease in ad skip rates.

☐ The group assignment for each user (80% or 20% AI-generated videos) is independent of the potential outcomes for each user.

For the remainder of the question, assume that users choose which videos to watch based on a combination of algorithmic recommendations and personal preference, and that A_i can now take on values other than 0.8 and 0.2.

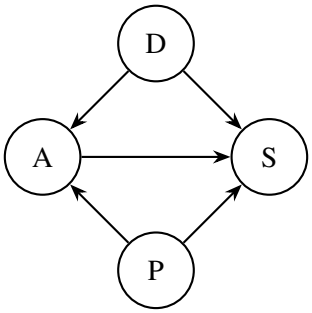
(b) [2 Pts] David asks an intern who took Data 102 to use exact matching to estimate the causal effect of watching AI-generated videos on ad skip rates. Assuming the intern uses only methods learned in this class and uses them correctly, which of the following must be true? Select all answers that apply **by filling in the square next to each correct answer**.

☐ The unconfoundedness assumption must be satisfied.

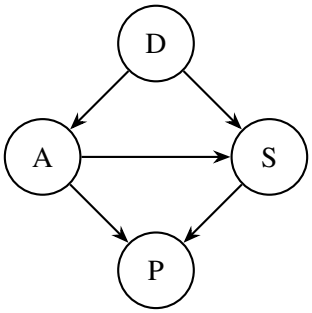
☐ The intern must have binarized A_i (e.g., by thresholding).

☐ Every user has at least one other user who has the same values of all confounding variables.

For parts (c) and (d), David is considering the following two DAGs for the relationships between S, A, P, and D. He assumes that there are no variables of interest other than these four.



DAG 1



DAG 2

(c) [4 Pts] For each statement below, determine whether it is consistent with DAG 1, DAG 2, both, or neither.

(i) Watching more AI-generated videos can cause someone to take more politically conservative views.

◯ DAG 1 ◯ DAG 2 ◯ Both ◯ Neither

(ii) When investigating the causal effect of AI-generated videos on ad skip rates, political views are a collider.

◯ DAG 1 ◯ DAG 2 ◯ Both ◯ Neither

(iii) When investigating the causal effect of AI-generated videos on ad skip rates, the set $\{D, P\}$ satisfies the backdoor criterion.

◯ DAG 1 ◯ DAG 2 ◯ Both ◯ Neither

(iv) When investigating the causal effect of AI-generated videos on ad skip rates, demographic information is a confounding variable.

◯ DAG 1 ◯ DAG 2 ◯ Both ◯ Neither

(d) [3 Pts] For this part, David assumes that DAG 1 is correct. Based on this, he decides to use inverse propensity weighting, using logistic regression to compute propensity scores.

(i) Which one of the following variables should he use as the **target** or y-variable when fitting his logistic regression model? Choose the single best answer **by filling in the circle next to it.**

◯ A ◯ P ◯ D ◯ S

(ii) Which of the following variable or variables should he use as the **predictors** or x-variables when fitting his logistic regression model? Select all answers that apply **by filling in the square next to each correct answer**.

□ A □ P □ D □ S

6 Gambling [7 Pts]

Agatha buys a ticket from GamblingCo that lets her play a game of chance 200 times. Each play, she receives an i.i.d. random monetary reward, where the average reward per play is $\$0.50$.

Let T be Agatha's total reward from all 200 plays of the game. The ticket costs $\$150$, so her net winnings are $T - 150$.

(a) [1 Pt] Given only the information so far, find the smallest number q such that the probability of her net winnings being nonnegative is q.

For the remainder of the question, assume the reward is always between $\$0.01$ and $\$5.01$.

(b) [3 Pts] With this new information, find the smallest number q such that the probability of her net winnings being nonnegative is q.

(c) [3 Pts] GamblingCo designs a different kind of ticket. Here's how it works:

1. The player first chooses a number n, and then plays the game n times, keeping all the rewards.
2. If the player's total reward in n plays is at least $3n$ dollars, then **the ticket is free**.
3. Otherwise, **the ticket costs** $3n$ **dollars**.

As before, assume the reward is always between $\$0.01$ and $\$5.01$, and that the average reward is $\$0.50$.

Find the smallest integer n such that the company can be at least 90% confident that the player will **not** win a free ticket, or prove why no such number exists. **For full credit, you must express your answer as an integer or show a clear proof why no such number exists.**

Where necessary, you should use the approximations $\log(0.9) \approx -0.1$ and/or $\log(0.1) \approx -2.5$.

7 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	$\lfloor \lambda \rfloor$
$X \sim \text{Binomial}(n, p)$	$x \in \{0, 1, \ldots, n\}$	$\binom{n}{x} p^x (1-p)^{1-x}$	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

Hoeffding's Inequality: If $X_1, \ldots, X_n$ are independent random variables bounded between a and b, then

$$P\left(\frac{1}{n}\sum_{i=1}^n (X_i - E[X_i]) \ge t\right) \le \exp\left(-\frac{2nt^2}{(a-b)^2}\right)$$

$$P\left(\frac{1}{n}\sum_{i=1}^n (X_i - E[X_i]) \le -t\right) \le \exp\left(-\frac{2nt^2}{(a-b)^2}\right)$$

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - E[X_i])\right| \ge t\right) \le 2\exp\left(-\frac{2nt^2}{(a-b)^2}\right)$$