

# Data 102 Spring 2021

## Midterm 2

- Please write your solutions using either pen/pencil and paper, or a tablet. Each question should start on a new page. At the end of the exam period (or earlier), please upload your exam to the “Midterm 2” assignment on Gradescope. **It is your responsibility to make sure your work will be legible!**
- We will not answer any questions during the exam. If you think a question is unclear, state your assumptions and answer accordingly.
- You have 80 minutes to work on the exam: you must stop working at 11:00AM PT.
- This exam has 6 questions, for a total of 40 points. **You must complete all 6 questions to receive full credit.** There are multiple versions of this exam.
- Unless otherwise stated, you must show your work to receive full credit.
- You may, without proof, use theorems and facts that were given in the lectures, homework, lab, or discussions.
- **You must complete this honor pledge in order to receive credit on the exam:** We ask that you act in accordance with the honor code. Please copy the following statement by hand and sign your name, and include this in your submission.

<p><b>As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others. These answers are my own work.</b></p>
---

0. Make sure you complete the honor pledge on the previous page.

1. (5 points) **Nonparametric methods.**

- (a) (1 point) (True/False) When training a random forest, each tree is trained with the same features, but in a different order.
- (b) (1 point) (True/False) Techniques like LIME use a simple, interpretable model to approximate a more complex model.
- (c) (1 point) (True/False) If there are nonlinear interactions between the input variables and a binary output label, then there is no way to use logistic regression to model the relationship between them.
- (d) (1 point) (True/False) Backpropagation is an algorithm that is only used for training neural networks.
- (e) (1 point) (True/False) When thinking about the bias-variance tradeoff, logistic regression has higher bias than a decision tree (with no depth limit).

2. (15 points) **Causal inference.** You are working with a developmental economist to understand the effect of free school lunches on school attendance. To study this, the economist conducted a completely randomized experiment that randomized Grade 5 students into receiving a free lunch ( $T = 1$ ) or not ( $T = 0$ ), and then observed whether they attended school ( $Y = 1$ ) or not ( $Y = 0$ ). The results are reported in the following table.

	$T = 0$	$T = 1$
$Y = 0$	100	50
$Y = 1$	700	450

Table 1: Grade 5 students

- (a) (2 points) Compute the Neyman (difference-in-means) estimate for the average treatment effect (ATE) of school lunches on school attendance for Grade 5 students.
- (b) (1 point) Write one sentence of plain English interpreting the ATE. Your answer should be understandable to a general audience, and should make the strongest valid conclusion that you can. *Hint: What is the effect of receiving a free school lunch?*
- (c) (1 point) We compute a 95% confidence interval for the true ATE using the Neyman variance. If the interval does not contain 0, which of the following null hypotheses can we reject (at the 95% level)? **Select all that apply (or write “none”).**
- A. Fisher’s strong null hypothesis
- B. Neyman’s weak null hypothesis
- (d) (2 points) The economist simultaneously did a completely randomized experiment on Grade 6 students, with the results reported in Table 2.

	$T = 0$	$T = 1$
$Y = 0$	200	200
$Y = 1$	300	300

Table 2: Grade 6 students

For the rest of this question, we investigate the results using a super-population framework. We introduce a covariate  $X$  such that  $X = 1$  for students in Grade 6, and  $X = 0$  for students in Grade 5. Compute the estimated propensity score function  $\hat{e}(x)$  for  $x = 0, 1$ .

- (e) (2 points) Is  $X$  (which grade a student is in) a confounding variable? In one sentence, explain why or why not.
- (f) (2 points) Does the unconfoundedness assumption hold? In one sentence, explain why or why not.

- (g) (2 points) The next two parts are about the inverse-propensity weighting (IPW) estimate for the average treatment effect (ATE) of school lunches on school attendance for the **combined population** of Grade 5 and 6 students. The estimate has the form.

$$\hat{\tau}_{IPW} = \frac{1}{n} \left( \frac{A}{\hat{e}(0)} + \frac{B}{\hat{e}(1)} - \frac{C}{1 - \hat{e}(0)} - \frac{D}{1 - \hat{e}(1)} \right). \quad (1)$$

What are the values of  $A$ ,  $B$ ,  $C$  and  $D$ ?

- (h) (1 point) What is the value of  $n$  in equation (1)?
- (i) (2 points) Denote your answer in (a) using  $\hat{\tau}_5$ , and denote the Neyman estimate for the corresponding ATE computed over Table 2 using  $\hat{\tau}_6$ . The economist proposes four estimates for the average treatment effect (ATE) of school lunches on school attendance for the **combined population** of Grade 5 and 6 students. They are as follows.
- (A) The IPW estimate from part (g) (Equation (1)).
- (B) Add up the counts in Tables 1 and 2 and compute the Neyman estimate for the ATE using the resulting table.
- (C)  $\frac{1}{2}\hat{\tau}_5 + \frac{1}{2}\hat{\tau}_6$ .
- (D)  $(1 - w)\hat{\tau}_5 + w\hat{\tau}_6$ , where  $w = \mathbb{P}(X = 1)$ .

Which of these estimates are unbiased for the true ATE? **Select all that apply (or write “none”)**.

3. (3 points) **Instrumental variables.**

Consider the linear structural model

$$Y = \alpha + \tau Z + \beta X + \epsilon,$$

$$Z = \alpha' + \gamma W + \beta' X + \delta.$$

We wish to estimate the treatment effect  $\tau$  of  $Z$  on  $Y$  using  $W$  as an instrumental variable. In order for  $W$  to be valid instrumental variable, we need some assumptions on  $Y$ ,  $Z$  and  $W$  and  $X$ . For each of the quantities below, specify whether it **must be zero** ( $= 0$ ), **must be nonzero** ( $\neq 0$ ), or **does not matter**.

- (i)  $\text{Cov}(W, Y)$
- (ii)  $\text{Cov}(W, X)$
- (iii)  $\text{Cov}(W, Z)$
- (iv)  $\text{Cov}(W, \epsilon)$
- (v)  $\text{Cov}(W, \delta)$

4. (4 points) **Concentration inequalities.** Let  $X_1, X_2, \dots, X_n$  be independent, identically distributed (i.i.d.) random variables, each with mean 0, and having the same distribution as a  $\sigma$ -sub-Gaussian random variable  $X$ . Let  $S_n = \sum_{i=1}^n X_i$ .
- (a) (2 points) Suppose we are told (only for this part of the question) that  $X$  is bounded between  $-a$  and  $a$ . Based on this information, what is a valid value for  $\sigma^2$ ? State the smallest possible valid value.
- (b) (2 points) By Hoeffding's inequality, we have

$$\mathbb{P}(|S_n| > t) \leq \exp\left(-\frac{t^2}{2n\sigma^2}\right). \quad (2)$$

Which of the following changes can we **make** to the assumptions and still guarantee that inequality (2) still hold? **Select all that apply (or write "none")**.

- (A)  $X_1, \dots, X_n$  are not identically distributed.
- (B)  $X_1, \dots, X_n$  are not independent.
- (C) We have  $\mathbb{E}[X_i] = \mu_i$  (not necessarily 0) for each  $i = 1, \dots, n$ , but  $\sum_{i=1}^n \mu_i = 0$ .
- (D) Each  $X_i$  is  $\sigma_i$ -sub-Gaussian (not necessarily all the same), with  $\sum_{i=1}^n \sigma_i^2 = n\sigma^2$ .

5. (8 points) **Bandit Algorithms.** Consider a bandit environment with  $K = 2$  arms, with 1-sub-Gaussian arm reward distributions  $P_a$  and means  $\mu_a$  for  $a = 1, 2$ . Assume that arm 1 is the optimal arm (i.e.,  $\mu_1 > \mu_2$ ), and so we may define the suboptimality gap  $\Delta = \mu_1 - \mu_2$ .

A learner has already played 7 rounds. We are told they pulled arm 1 a total of 2 times, and arm 2 a total of 5 times.

- (a) (2 points) Which of the following are possible policies that the learner was following? **Select all that apply (or write “none”).**
- (A) The upper confidence bound algorithm (UCB).
  - (B) Explore-then-commit (ETC) with 3 rounds of exploration ( $m = 3$ ).
  - (C) Thompson-sampling (TS).
- (b) (2 points) Suppose the learner was following a *deterministic* strategy (i.e. the arm choices were determined before the start of the algorithm. We still assume arms 1 and 2 were pulled 2 and 5 times respectively.) As usual, denote the observed average reward for each arm by  $\hat{\mu}_a = \frac{1}{T_a(7)} \sum_{s=1}^7 X_s \mathbf{1}(A_s = a)$  for  $a = 1, 2$ . Using Hoeffding’s inequality, we compute the following probability bound:

$$\mathbb{P}(\hat{\mu}_1 - \hat{\mu}_2 > t) \leq \exp\left(-\frac{(t + A)^2}{B}\right).$$

What are the values for  $A$  and  $B$ ? Express your answer in terms of  $\mu_1, \mu_2$ , and  $\Delta$ .

- (c) (2 points) Suppose we were running Thompson Sampling with Gaussian priors and likelihoods. For concreteness, suppose  $\mu_1 = 5$ ,  $\mu_2 = 3$ ,  $\hat{\mu}_1(7) = 4.1$  and  $\hat{\mu}_2(7) = 2.5$ . Let  $\mathcal{N}(z_a, v_a^2)$  be the prior for arm  $a = 1, 2$ . For each of the following parameters, we have suggested a number of possible values. For each parameter, choose the **one** option that maximizes the probability of pulling arm 1 in the next (eighth) round.

$$\begin{array}{l} z_1 : \quad -10, \quad 4.1, \quad 5, \quad 10 \\ z_2 : \quad -10, \quad 2.5, \quad 3, \quad 10 \\ v_1^2 : \quad 0.05, \quad 10 \\ v_2^2 : \quad 0.05, \quad 10 \end{array}$$

For example, if you chose a value of 10 for all parameters, your solution might look like:  $z_1 = 10$ ,  $z_2 = 10$ ,  $v_1^2 = 10$ ,  $v_2^2 = 10$ . (This is not necessarily the correct answer, just an example.)

- (d) (2 points) A very risk-averse data scientist has proposed the following *lower* confidence bound algorithm. Define the lower confidence bound

$$\text{LCB}_a(t, \delta) = \hat{\mu}_a(t) - \sqrt{\frac{2 \log(1/\delta)}{T_a(t)}}$$

At each round  $t$ , the learner selects:

$$A_t = \begin{cases} t & t \leq K \\ \operatorname{argmax}_{a=1,\dots,K} \operatorname{LCB}_a(t-1, 1/t^3) & t > K. \end{cases}$$

Does this algorithm have logarithmic regret? Explain why or why not. You don't need to provide a full proof, but you must provide a convincing explanation.



6. (5 points) **Uncertainty quantification for GLM.** You are consulting for an ice-cream company that wants to investigate the relationship between mean daily temperature  $X$  (in degrees Celsius) and the number of ice-cream cones sold  $Y$  (a count). You model this using Poisson regression, with

$$\mathbb{E}[Y|X] = e^{\beta_0 + \beta_1 X}. \quad (3)$$

To fit the model, we use a data set  $S$  that contains i.i.d. samples  $(X_1, Y_1), \dots, (X_n, Y_n)$  from our population of interest, obtaining coefficient estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$ .

- (a) (2 points) We wish to use the bootstrap to get a 95% confidence interval for  $\beta_1$ . We have already generated 1000 bootstrap replicates of  $\hat{\beta}_1$ , which are stored in a one-dimensional numpy array `beta_boot`. Write **no more than two lines** of code in Python that gives the left and right end-points of such an interval. You may assume that we have already run the line `import numpy as np`.
- (b) (1 point) After running your code, you discover a bug: each bootstrap replicate  $\hat{\beta}^*$  was obtained by drawing  $2n$  samples at random with replacement from  $S$  (instead of just  $n$  samples). Compared to the correct bootstrap confidence interval, is the width of your confidence interval smaller, larger, or roughly the same?
- (c) (1 point) Suppose we know that the mean temperature tomorrow is going to be  $35^\circ\text{C}$ . Given a model with regression coefficients  $\beta = (\beta_0, \beta_1)$ , what is the probability  $p(Y = 90|X = 35, \beta)$  that 90 ice-cream cones will be sold tomorrow?
- (d) (1 point) Suppose we instead use a Bayesian approach, and fit a Bayesian Poisson GLM. Let  $q(\beta)$  denote the posterior distribution (density) that we compute over  $\beta = (\beta_0, \beta_1)$ . Write a formula for the **posterior predictive probability** that 90 ice-cream cones will be sold tomorrow given that the mean temperature tomorrow is going to be  $35^\circ\text{C}$ . You may leave your answer in terms of  $q$  and  $p(Y = 90|X = 35, \beta)$ .