DS 102 Fall 2020 Midterm

- You have 140 minutes to complete this exam. There are 5 questions, totaling 40 points.
- The exam is open notes and open internet.
- You can choose to write your solutions inside the exam sheet or write them on a separate sheet. You can use use either pen and paper, use a tablet or typeset 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.
- Make sure to write clearly. We can't give you credit if we can't read your solutions.
- If you have clarification questions first check the **Clarification Page** on Piazza, where we will add in real time clarification based on student questions. If you still have a question, make a private Piazza post.
- After 140 minutes have passed, please submit your solutions to Gradescope. You will have 10 minutes to do so, after which the Gradescope submission will close; these 10 minutes are for uploading solutions and not for continuing to work on the exam. *If you have gradescope issues you can email the solutions to us at data102.exam@gmail.com*

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: _____

1. (8 points) For each of the following, answer true or false. **Circle T for true and F for false. You don't need to justify your answer.**

   (a) (1 point) (T / F) In each iteration of the Metropolis-Hastings algorithm, the algorithm transitions from the current state to the proposed state with probability one.

   Solution: False. In each iteration of the Metropolis-Hastings algorithm, the algorithm transitions from the current state to the proposed state with probability $\min\{1, \frac{p(x^{new})}{p(x^{old})} \frac{q(x^{old)|x^{new}}}{q(x^{new}|x^{old})}\}$

   (b) (1 point) (T / F) The family-wise error rate (FWER), i.e., the probability of making at least one false discovery, is at least as big as the false discovery rate (FDR): $FWER \geq FDR$.

   Solution: True. This statement is proven in discussion 3.

   (c) (1 point) (T / F) The function $f(x, \theta) = e^x \ln\theta + e^{3x}\theta^2 + e^x sin(\theta)$ is a linear function of $y = e^x$.

   Solution: False. Rewrite f in terms of y and θ, we get:

   $$f(x, \theta) = f(y, \theta) = y \ln\theta + y^3\theta^2 + y \sin(\theta)$$

   This is not linear in y because of the term $y^3\theta^2$.

   (d) (1 point) (T / F) Conditional on θ, the credible interval of a parameter θ must contain the true value θ^*.

   Solution: False. Conditional on θ, the credible interval of a parameter θ must contain the true value θ^* with some probability p.

   (e) (1 point) (T / F) Suppose we have 10 data points whose true labels are all negatives (all 0s). If our model predicts 8 positives (1s) and 2 negatives (0s), then the False Discovery Proportion (FDP) of the 10 predictions is given by $8/10 = 0.8$.

   Solution: False. FDP is a column-wise ratio as defined in lecture 2. The proportion computed in the problem is a row-wise ratio.

   (f) (1 point) (T / F) We can always use bootstrap to find an unbiased estimate of the median of a random variable X.

Solution: False. When the distribution of X is highly skewed, the estimate may have very high bias.

(g) (1 point) (T / F) Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ for some unknown parameter μ and we observe only one sample X_1. Then the MLE estimate of μ is equal to X_1.

Solution: True.

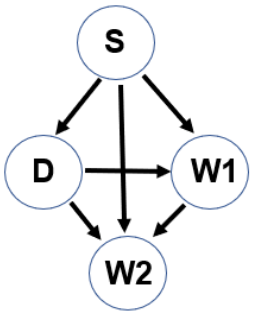
$$\hat{\mu}_{mle} = \arg\max_{\mu} \exp\left(-\frac{1}{2}\frac{(X_1-\mu)^2}{\sigma^2}\right) = \arg\min_{\mu} \frac{1}{2}\frac{(X_1-\mu)^2}{\sigma^2} = X_1$$

(h) (1 point) (T / F) The unconfoundedness assumption, $Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1))|X_i$ is testable.

Solution: False. The unconfoundedness assumption in Causal Inference is not testable.

2. (4 points) Suppose that a scientist conducts an experiment on rats to model the effect of certain drugs on weight loss. Assume that 24 rats (12 female, 12 male) are randomly selected to be administered one of three weight-loss drugs. We observe their weight loss at week one and week two, separately. We model the dependency between the variables in the following way: the weight loss at week one is affected by both the sex of rats and the choice of drugs, and the weight loss at week two is affected by the sex of rats, the choice of drugs, and the weight loss at week one.

   (a) (2 points) Let the random variables S, D, W_1, W_2 denote the sex of rats, choice of drugs, weight loss at week one, and weight loss at week two, respectively. We provide below an unfinished graphical model of the four random variables. Fill in the rest of the arrows to complete it.



   (b) (2 points) With the above graphical model, write down the factorization of their joint density using the multiplication of (conditional) probabilities:

$$p(S, D, W_1, W_2) = \_\_\_\_\_$$

Solution: (a) See the above figure.
(b) $p(S, D, W_1, W_2) = p(S)p(D|S)p(W_1|S, D)p(W_2|S, D, W_1)$.

3. (10 points) **Causal Inference**

Let's first consider a simple causal inference problem that begins with a data scientist announcing that they have discovered a correlation between shoe size and income.

(a) (2 points) What is a possible confounding variable that casts doubt on any assertion that this is a causal relationship? Explain why this confounder accounts for the correlation.

Solution: The confounder that I had in mind was age: children have a small shoe size and they also have a small income. Adults have a larger shoe size and a larger income. But presumably there are other reasonable confounders; indeed, anything that correlates with age, including years of education.

(b) (2 points) In a subsequent study, let's suppose that we observe that confounding variable as data and discretize it into two values. We then condition on those discrete values for a causal analysis. Does it make sense to additionally use the original, non-discretized variable as a covariate in a regression as part of our causal analysis? Why or why not?

Solution: Yes, it still makes sense in general. The discretization provides a partial ability for the observation of the confounder to render the treatment and the outcome conditionally independent. Regressing on the full variable potentially allows more unconfounding.

Now consider a general causal inference problem with a covariate X_i, where X_i takes on only two possible values, x and x'. Suppose that conditional on X_i, the treatment indicator Z_i is independent of the potential outcomes. Suppose that 100 experimental units have been sampled from a population, with 60 of them having covariate $X_i = x$ and 40 having covariate $X_i = x'$. Outcomes Y_i are measured for each of these units.

(c) (2 points) Write down an estimator of the average treatment effect that is based on breaking the overall data into two subsets, one where the covariate is x and the other where the covariate is x', and comparing average outcomes in the two cases. Your expression should use indicators and sums. Explain why your estimator is a reasonable estimator.

Solution: For each value of $X_i \in \{x, x'\}$, since treatment is independent of the potential outcomes, we can estimate the average treatment effect among units with that value of the covariate as a simple difference in means. For example, conditional on $X_i = x$ a reasonable estimate of the average treatment effect is

$$\hat{\tau}_x = \frac{1}{n_{x,1}} \sum_{i:X_i=x} Y_i Z_i - \frac{1}{n_{x,0}} \sum_{i:X_i=x} Y_i(1 - Z_i),$$

where $n_{x,b} = \sum_{i:X_i=x} 1\{Z_i = b\}$. Then the overall estimate of average treatment effect should weight the conditional estimates by their relative likelihood:

$$\hat{\tau} = \frac{60}{100}\hat{\tau}_x + \frac{40}{100}\hat{\tau}_{x'}$$

(d) (2 points) Suppose that for both x and x', we observe that 1/5 of the units are given the treatment and 4/5 are given the control. Propose an estimator of the propensity score.

Solution: Recall that the population propensity score is $e(X) = \mathbb{P}(Z = 1 \mid X)$. Since X only takes on two values, it's reasonable to use the empirical proportion of units receiving treatment within each group as an estimate of propensity score, i.e.

$$\hat{e}(x) = \hat{e}(x') = 1/5.$$

(e) (2 points) Given your estimator of the propensity score, write down the inverse weighted propensity score estimator of the average treatment effect, plugging in numerical values wherever you can.

Solution: The IPW estimator is

$$\begin{aligned}\hat{\tau}^{IPW} &= \frac{1}{100}\sum_{i=1}^{100}\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)} \\ &= \frac{1}{100}\sum_{i=1}^{100}\frac{Z_i Y_i}{1/5} - \frac{(1-Z_i)Y_i}{4/5}\end{aligned}$$

(bonus) How does the estimator in (e) relate to your estimator in (c)?

Solution: Continuing from the last step in (e),

$$\begin{aligned}\hat{\tau}^{IPW} = \frac{60}{100}&\left(\frac{1}{60}\sum_{i:X_i=x}\frac{Z_i Y_i}{1/5} - \frac{(1-Z_i)Y_i}{4/5}\right) \\ &+ \frac{40}{100}\left(\frac{1}{40}\sum_{i:X_i=x'}\frac{Z_i Y_i}{1/5} - \frac{(1-Z_i)Y_i}{4/5}\right) = \frac{60}{100}\hat{\tau}_x + \frac{40}{100}\hat{\tau}_{x'} = \hat{\tau}\end{aligned}$$

so the estimators are the same.

4. (10 points) **Hypothesis Testing and FDR control.** Consider a random variable X that comes from either a $\mathcal{N}(0,1)$ distribution or a $\mathcal{N}(1,1)$ distribution. To decide which of the two possibilities is correct, we frame this problem in the context of hypothesis testing, with the null being $H_0 : X \sim \mathcal{N}(0,1)$ and the alternative being $H_A : X \sim \mathcal{N}(1,1)$.

   (a) (2 points) Suppose we have one observation, $X_1 = 1$. What is the p-value associated with this observation? Write your expression in terms of the cumulative distribution function for the standard normal distribution, $\Phi(x)$.

   Solution: By definition, p-value of an observation x is equal to the probability of observing x or more extreme values under the null hypothesis. Therefore,

   $$\text{p-value} = \mathbb{P}(X \geq 1 | X \sim \mathcal{N}(0,1)) = 1 - \Phi(1)$$

   (b) (3 points) Suppose we use the decision rule:

   $$\delta(p, \alpha) = \begin{cases} \text{reject null} & p \leq \alpha \\ \text{accept null} & p > \alpha, \end{cases} \tag{1}$$

   where $\alpha \in (0,1)$ is the significance level of the test. We have n independent observations $X_1, X_2, ..., X_n$ and the corresponding n independent p-values $p_1, ..., p_n$. If in reality, all X_i's are sampled from $\mathcal{N}(0,1)$, what is the probability that the FDP (False Discovery Proportion) is equal to 1? In other words, give an expression for $\mathbb{P}(FDP = 1)$ in terms of α.

   Solution: From lecture and discussion, we know that p-value $\sim Uniform(0,1)$ under null distribution. By definition:

   $$FDP = \frac{\text{Number of False Discoveries}}{\text{Number of False Discoveries} + \text{Number of True Discoveries}}$$

   Since in reality, all $X_i's$ are sampled from the null distribution, all discoveries, if any, are False Discoveries. In other words, Number of True Discoveries should always be 0. Therefore,

   $$FDP = \frac{\text{Number of False Discoveries}}{\text{Number of False Discoveries}} = 1$$

   as long as we have at least one discovery. As a result:

   $$\begin{aligned} \mathbb{P}(FDP = 1) &= \mathbb{P}(\text{at least 1 discovery}) \\ &= 1 - \mathbb{P}(\text{no discoveries}) \\ &= 1 - (1-\alpha)^n \end{aligned} \tag{2}$$

(c) (2 points) If we observe five sorted p-values, 0.005, 0.009, 0.012, 0.02, 0.03 and set the significance level to be $\alpha = 0.05$, (i) How many tests are rejected with the Bonferroni correction? (ii) How many tests are rejected with the Benjamin-Hochberg procedure?

Solution:

1. Bonferroni correction: A discovery is made if p-value $p < \frac{\alpha}{n}$. In the problem, $\frac{\alpha}{n} = \frac{0.05}{5} = 0.01$. Since 0.005 and 0.009 are less than 0.01, **2 tests are rejected** with Bonferroni correction.

2. Benjamin-Hochberg: We find the k such that $k = \max_i \{P_{(i)} \leq \frac{i}{n}\alpha\}$ where $P_{(i)}$ is the i-th smallest p-value and reject all tests whose sorted index is less than or equal to k. In the problem, we have that $k = 5$. Therefore, **5 tests are rejected** with the Benjamin-Hochberg procedure.

(d) (3 points) Suppose we observe n independent random p-values, $p_1, p_2, \ldots, p_n$. At significance level α, what is the probability that the Benjamini-Hochberg procedure makes **at least one** discovery? Write your final expression in terms of n and α.

Solution:

$$\mathbb{P}(\text{at least one discovery}) = 1 - \mathbb{P}(\text{No Discovery})$$

In order to have 0 discoveries with the Benjamini-Hochberg procedure, we must have that $k = 0$ where $k = \max_i \{P_{(i)} \leq \frac{i}{n}\alpha\}$ where $P_{(1)}, P_{(2)}, ..., P_{(n)}$ is the sequence of sorted p-values in ascending order. Therefore, by symmetry, we have:

$$\begin{aligned}
\mathbb{P}(\text{No Discovery}) &= \mathbb{P}(k = 0) \\
&= \sum_{\text{permutations}} \mathbb{P}(k = 0, p_1 \leq p_2 \leq ... \leq p_n) \\
&= \sum_{\text{permutations}} \mathbb{P}(k = 0 | p_1 \leq p_2 \leq ... \leq p_n)\mathbb{P}(p_1 \leq p_2 \leq ... \leq p_n) \\
&= n!\mathbb{P}(p_1 > \frac{\alpha \cdot 1}{n}, p_2 > \frac{\alpha \cdot 2}{n}, ..., p_n > \frac{\alpha \cdot n}{n} | p_1 \leq p_2 \leq ... \leq p_n)\frac{1}{n!} \\
&= \mathbb{P}(p_1 > \frac{\alpha \cdot 1}{n}, p_2 > \frac{\alpha \cdot 2}{n}, ..., p_n > \frac{\alpha \cdot n}{n} | p_1 \leq p_2 \leq ... \leq p_n) \qquad (3)
\end{aligned}$$

Finding the exact probability of the expression above requires order statistics which is beyond the scope of this class. This is a mistake on our part (we should've realized this while making the exam). As a result, we are grading this subpart generously. One common solution student have is

$$\mathbb{P}(\text{at least one discovery}) = 1 - \mathbb{P}(\text{No Discovery}) = 1 - \prod_{i=1}^{n} \left(1 - \frac{\alpha \cdot i}{n}\right)$$

This solution is correct if there is no condition in the final expression of (3).

5. (8 points) **Rejection Sampling and Gibbs Sampling.** We would like to sample from a two-dimensional distribution with joint probability density

$$p(x,y) = \begin{cases} \frac{x+2y+1}{6} & x \in [0,2], y \in [0,1], \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

(a) (3 points) Assume we are going to sample from $p(x,y)$ using Rejection Sampling, with the proposal distribution $q(x,y)$ being uniform and supported on the range $x \in [0,2], y \in [0,1]$:

$$q(x,y) = \begin{cases} \frac{1}{2} & x \in [0,2], y \in [0,1], \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Find the constant c such that $c \cdot p(x,y) \leq q(x,y)$ on the support.

(b) (3 points) Describe the procedure of obtaining samples from $p(x,y)$ via Rejection Sampling using the constant c that you obtained in part (a).

(c) (2 points) We would like to estimate the expectation $\mathbb{E}\left[\sqrt{X^2+Y^7}\right]$. Assume that we are given a subroutine that can sample directly from $p(x|y)$ and $p(y|x)$. Fill in the blank (i), (ii) below to estimate $\mathbb{E}\left[\sqrt{X^2+Y^7}\right]$ using Gibbs Sampling.

Estimation of expectation: Initialize $X^{(0)}$ and $Y^{(0)}$.

For $t = 1, \ldots, T$, where T is some large stopping time:

1. Start with $(X^{(t-1)}, Y^{(t-1)})$ from the previous iteration.
2. Sample $X^{(t)}$ from the distribution $\mathbb{P}(X|Y = Y^{(t-1)})$.
3. (i) _____ (Hint: Step for sampling $Y^{(t)}$)

End For.

Estimate the expectation of $\sqrt{X^2+Y^7}$ using the samples $\{(X^{(i)}, Y^i)\}_{i=1}^n$ as:

(ii)_____

Solution: (a) We want that for any $x \in [0,2], y \in [0,1]$,

$$c \leq \frac{q(x,y)}{p(x,y)} = \frac{3}{x+2y+1}. \tag{6}$$

The right hand side is minimized when $x = 2, y = 1$, thus we have

$$c \leq \frac{3}{5}. \tag{7}$$

Thus any positive constant that is smaller or equal to 3/5 satisfies the requirement.

(b)

For $t = 1, \ldots, T$ for some large stopping time T:

1. Sample $(X^{(t)}, Y^{(t)})$ according to $q(x, y)$, which is uniform distribution on $x \in [0, 2], y \in [0, 1]$.

2. Generate a sample r from uniform distribution $U[0, 1]$. If $r \leq \frac{c \cdot p(x,y)}{q(x,y)}$, accept the sample $(X^{(t)}, Y^{(t)})$.

(c) (i) Sample $Y^{(t)}$ from the distribution $\mathbb{P}(Y|X = X^{(t)})$.
(ii) $\frac{1}{n} \sum_{i=1}^{n} \sqrt{(X^{(i)})^2 + (Y^{(i)})^7}$