Data 102, Fall 2024
Homework 1

Due: **5:00 PM** Friday, September 20, 2024

Submission Instructions

Homework assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

Written Portion:

- Every answer should contain a calculation or reasoning.
- You may write the written portions on paper or in LaTeX.
- If you type your written responses, please make sure to put it in a markdown cell instead of writing it as a comment in a code cell.
- Please start each question on a new page.
- It is your responsibility to check that all the work on all the scanned pages is legible.

Code Portion:

- You should append any code you wrote in the PDF you submit. You can either do so by copy and paste the code into a text file or convert your Jupyter Notebook to PDF.
- Run your notebook and make sure you print out your outputs from running the code.
- It is your responsibility to check that your code and answers show up in the PDF file.

Submitting:

You will submit a PDF file to Gradescope containing all the work you want graded (including your math and code).

- When downloading your Jupyter Notebook, make sure you go to File → Save and Export Notebook As → PDF; do not just print page from your web browser because your code and written responses will be cut off.
- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so. As a Berkeley student, you get free access to Adobe Acrobat, which you can use to merge as many PDFs as you want.
- Please see this guide for how to submit your PDF on Gradescope. In particular, for each question on the assignment, please make sure you understand how to select the corresponding page(s) that contain your solution (see item 2 on the last page).

Late assignments will count towards your slip days; it is your responsibility to ensure you have enough time to submit your work.

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

Math Stats

1. (8 points) Work through the following exercises, and explain your reasoning.

   (a) Suppose a particular drug test is 99% sensitive and 98% specific (Here is a Wikipedia link for a refresher on the terminology). The null hypothesis H_0 is that the subject is not using the drug. Assume a prevalence of $\pi_1 = 0.5\%$, i.e. only 0.5% of people use the drug. Consider a randomly selected individual undergoing testing. Rounding to the nearest three significant figures, find

      i. (1 point) the probability of testing positive given H_0.

      ii. (1 point) the probability that they are not using the drug given they test positive.

      iii. (0 points) (Optional) the probability of testing positive a second time given they test positive once. You may assume the two tests are statistically independent given drug user status.

   (b) Suppose we have a waiting time $T \sim \text{Exponential}(\lambda)$ and wish to test

   $$H_0 : \lambda = c \quad \text{vs} \quad H_1 : \lambda = 2c$$

   for some $c > 0$. In this question, you'll use the *likelihood ratio test* (LRT) to compare these two hypotheses. The LRT considers the ratio of the two density functions f_1 and f_0 under the alternative and null respectively:

   $$\text{LR}(T) = \frac{f_1(T)}{f_0(T)},$$

   and rejects H_0 when $\text{LR}(T)$ is greater than some threshold η.

   We use this test because of the *Neyman-Pearson lemma*, which states that the likelihood ratio test is the most powerful test (in other words, it has the highest power, or TPR) of significance level α. That is, out of all possible tests of H_0 vs H_1 with $\text{FPR} = \alpha$, the likelihood ratio test has the highest TPR.

   Hint: For this question, you may find it helpful to brush up on computing probabilities involving continuous random variables. Prob 140 textbook, Chapter 15 *provides a helpful refresher.*

      i. (1 point) Compute $\text{LR}(T)$ explicitly in terms of c.

ii. (3 points) Let α be our false positive rate ($0 < \alpha < 1$). Compute the value of the threshold η so that the FPR of the test is equal to α. We say that such a test has *significance level* α. Your answer should be expressed in terms of any or all of α and c.

Hint: start by expressing the FPR as a conditional probability, then connect it to the LRT decision rule and the densities f_0 and f_1.

iii. (2 points) What is the TPR of this test? This is also known as the test's *power*. Your answer should be expressed in terms of α and c.

Fairness in Binary Prediction

2. (10 points) *Fairness* is a hotly debated topic and active area of study in machine decision-making. In this problem, you will show that two reasonable notions of fairness are mutually incompatible with accuracy, and will consider the consequences of that conclusion. The example is inspired by the paper "Inherent trade-offs in the fair determination of risk scores"[1]. **Please read the full set-up before attempting the problems. You will not be able to answer the problems without reading first.**

   Suppose that a data scientist is asked to design a prediction tool that takes in a job candidate's resume, and predicts whether or not the candidate will succeed in an interview. The data scientist aims to produce a tool that is both accurate and fair.

   We'll use the following notation:

   1. $i \in \{1, 2, \dots N\}$ indexes each candidate. I will denote a uniform random individual: $I \sim \text{Uniform}(1, 2, \dots, N)$.
   2. $y_i \in \{0, 1\}$ is an indicator for whether candidate i, if interviewed, will succeed, i.e. pass the interview.
   3. **Protected Classes:** $\mathcal{C} = \{C_1, C_2, \dots C_n\}$ is a collection of protected classes (e.g. age, gender, race, etc.), where each $C_j \subset \{1, 2, \dots, N\}$ contains a subset of the candidates all belonging to the same class. For example, the first protected class C_1 might consist of all individuals who are female, Black, and 40-50 years old; the second protected class C_2 might contain all candidates who are male, American Indian/Alaska Native, and 50-60 years old; and so on. You may assume that the classes are disjoint.
   4. p_j is the probability that a random individual from class C_j would pass the interview: $p_j = \mathbb{P}(y_I = 1 \mid I \in C_j)$
   5. **Predictive types:** $\mathcal{T} = \{T_1, T_2, \dots, T_m\}$ is a collection of disjoint candidate types used for prediction. Each $T_k \subset \{1, 2, \dots, N\}$ contains a subset of candidates who the prediction tool will treat as interchangeable. It will assign all individuals of a given type the same probability of passing their interview. For example, suppose the tool tracks education level achieved. Then the types could be:
      - T_1 contains all individuals who did not graduate high school;
      - T_2 contains all high school graduates (who did not earn any further degrees);

- T_3 contains all college graduates (who did not earn any further degrees);and
- T_4 contains all applicants with a graduate or professional degree.

6. $t_k = |T_k|$ is the number of individuals of type k. In the example above, t_4 is the number of applicants with graduate or professional degrees
7. $k(i)$ is the index corresponding to the type of individual i. In the example above, $k(77) = 3$ means that applicant 77 has a college degree and no further education.
8. The prediction tool assigns each type T_k a "score", s_k, then recommends individuals of type k with probability equal to their score s_k.
9. $z_i \in \{0, 1\}$ is the decision returned by the tool for individual i. If $z_i = 1$, then the tool recommends the individual. Specifically, $z_i \sim \text{Bernoulli}(s_{k(i)})$.

The data scientist aims to choose the scores, $\{s_1, \ldots, s_m\}$, so that the score assigned to each individual closely matches their chance of succeeding in the interview **without treating protected classes differently**. So, they adopt three objectives:

1. **Calibration:** The conditional probability that the tool *recommends* an individual should equal the conditional probability that the individual *would succeed* in their interview, given their type and class membership, for all types and classes:

$$\mathbb{P}(z_I = 1 \mid I \in C_j, I \in T_k) = \mathbb{P}(y_I = 1 \mid I \in C_j, I \in T_k) \text{ for all } j, k. \tag{1}$$

   This is an **accuracy** objective, since it matches the tool's recommendation probabilities (scores) to the true success probabilities for each predictive type and each protected class.

2. **Matching True Positive Rates:** The true positive rate for a uniformly drawn individual *should not* depend on their class:

$$P(z_I = 1 \mid y_I = 1, I \in C_j) = P(z_I = 1 \mid y_I = 1, I \in C_{j'}) = \text{TPR for all } j, j'. \tag{2}$$

   This is a **fairness** objective: when we compare individuals who would pass the interview, it says that the tool should be equally accurate on all of them, even if they're from different protected classes.

3. **Matching True Negative Rates:** The true negative rate for a uniformly drawn individual does not depend on their class:

$$P(z_I = 0 \mid y_I = 0, I \in C_j) = P(z_I = 0 \mid y_I = 0, I \in C_{j'}) = \text{TNR for all } j, j'. \tag{3}$$

   This is also a **fairness** objective: when we compare individuals who would *not* pass the interview, it says that the tool should be equally accurate on all of them, even if they're from different protected classes.

In this question, you'll show that, except in very special cases, no prediction tool can satisfy all three objectives at once.

(a) (1 point) The setup above describes a binary decision-making problem. Thinking back to the way we set up such problems in lecture and the textbook, which variable corresponds to reality and which corresponds to the decision?

(b) (3 points) The data scientist interviewed various experts who described what they want out of the tool. For each description below, state which of the three objectives (calibration, matching TPR, matching TNR) it most closely matches, and very briefly explain why.

(i) "I have a list of 50 candidates who passed their interviews. If I apply the tool to them, it should be just as likely to predict successful Black candidates as successful Asian candidates."

(ii) "I don't want a tool that's going to take these superstar candidates who'd almost certainly pass an interview, and just give them low scores!"

(iii) "Let's look at these older candidates and these younger candidates, and suppose none of them would pass the interview. The tool shouldn't give an unfair advantage to the younger candidates, OK?"

(c) (1 point) . By marginalizing over predictive types, show that, if the prediction tool is well calibrated (i.e., it meets the calibration objective), then:

$$\mathbb{P}(z_I = 1 \mid I \in C_j) = p_j. \tag{4}$$

Hint: you should not make any assumptions about either fairness objective for this part.

(d) (2 points) Let TPR_j be the true positive rate for protected class j (and similarly for TNR_j and FPR_j). Show that if the tool satisfies both of the fairness objectives (matching TPR and matching TNR), then for any protected class C_j, the TPR, TNR, and FPR satisfy the following equation:

$$(2 - (\text{TPR}_j + \text{TNR}_j))p_j = \text{FPR}_j \text{ for all } j. \tag{5}$$

Hint: Expand the left hand side of equation (4) by using the law of total probability, partitioning over the possible outcomes when candidate I is interviewed. Then, use the definitions of TPR, FPR, and FNR.

Hint: you should not make any assumptions about calibration or accuracy for this part.

(e) (2 points) Show that equation (5) can only be solved in one of two cases. Either, the probability that an individual succeeds in their interview does not depend on their class ($p_j = p_{j'}$ for all j, j'), or, the tool is perfect and never makes any mistakes (i.e. perfect prediction). If the tool is well-calibrated, is perfect prediction possible for arbitrary p_j? If not, what are the only situations when perfect prediction is possible?

(f) (1 point) *(Graded on completion)* What do *you* think this result implies about fairness in binary decision-making? Think about the difference between enforcing constraints exactly and approximately, whether the chosen definitions of fairness and

accuracy could be reasonably adapted or replaced, and, what other judgments the data scientist must make if they cannot satisfy all of their goals at once.[1]

Hint: "Fairness is impossible" is not a correct answer, and won't receive any credit!

(g) (0 points) (*Optional*) In the example above, we assumed that the predictive types corresponded to maximal degree achieved. Did that specific choice matter, or are calibration, matching true positive, and matching true negative rates incompatible for any definition of types? Is it possible to align the set of types $\mathcal{T}$ to the protected classes $\mathcal{C}$ so that the three constraints are compatible?

Bias in Police Stops

3. The following example is taken from [2, Ch. 6]:

> A study of possible racial bias in police pedestrian stops was conducted in New York City in 2006. Each of $N = 2749$ officers was assigned a score z_i on the basis of their stop data, with large positive values of z_i being possible evidence of bias. In computing z_i, an ingenious two-stage logistic regression analysis was used to compensate for differences in the time, place, and context of the individual stops.

To see how these z-scores are computed, check the original paper. We provide the data in a file `policez.csv` on DataHub.

We often assume each police officer acts **independently** of each other and therefore z_i's are **independent and identically distributed** (i.i.d.). Does this assumption hold? We will see more later in the question.

For now, if we assume that this is true, we can use hypothesis testing on each police officer to determine if they showed racial bias when stopping pedestrians. Formally, for the ith officer in the dataset, we have

- **Null Hypothesis**: $z_i \sim \mathcal{N}(0, 1)$; i.e. the ith officer does not show racial bias
- **Alternative Hypothesis**: z_i follows some other probability distribution; i.e. the ith officer shows some racial bias

Note that throughout this question, the word "bias" refers to police officers' racial bias, rather than the statistical term.

(a) (1 point) In the paper, the authors show that if the z_i's are i.i.d, they should be distributed according to $z_i \sim \mathcal{N}(0, 1)$. Let's see if this is true: in one plot, make a normalized histogram (e.g. a histogram with total area equal to one) of the z-scores and a line plot of the pdf of the theoretical null $\mathcal{N}(0, 1)$. Does the theoretical null fit the data exactly? If not, describe how the data differ from the pdf of the theoretical null.

[1] This question will be graded for completion, not accuracy.

(b) (0 points) *(Optional)* Compute p-values $P_i = \Phi(-z_i)$ (where Φ is the standard normal CDF) and then apply the BH procedure with $\alpha = 0.2$. Plot the sorted p-values as well as the decision boundary. You should find that you make **201 discoveries**.

(c) (1 point) Looking at the data, we can get a better fit to the distribution of z-scores if we use $\mathcal{N}(0.10, 1.40^2)$, called the empirical null (instead of the theoretical null from part (a)). Repeat parts **(a)** and (optionally) **(b)**, treating the empirical null as the null distribution. If you repeat part (b), should find that you make **only 5 discoveries**.

(d) (3 points) In practice, when the assumptions of our hypothesis tests are violated, we can alleviate adverse effects by utilizing the empirical null instead of the theoretical null to generate our p-values. However, this leads to a change in the way we interpret our results. Consider the following questions, and respond in plain english:

i. (1 point) In this study, the researcher argues in favor of using the empirical null to get around violated assumptions of each test (namely, that z_i are i.i.d). What might be a reasonable explanation as to why our z_i's fail to meet this assumption?

ii. (1 point) As we mentioned, using the empirical null comes with a new set of implicit assumptions. In particular, for each test we conduct, the null hypothesis is that the officer acts in an unbiased way. If the null hypothesis stays the same, but we switch from using the theoretical null distribution to using an empirical null distribution that better fits our policing data, what belief are we implicitly encoding about police officer behavior in general?

iii. (1 point) Based on your response to part **(ii)**, is this study design able to uncover department-wide patterns in racial bias? Why or why not?

p-values, FDR and FWER

4. The `adult.csv` file contains data from a random sample of the US adult population. It includes two numerical fields: `Age` and `Hours worked per week`. It also includes four categorical fields (which we have binarized for you): Gender, Education, Marriage status and whether the person's income is greater than $50,000. We will use this dataset to test the hypotheses of whether each of the categorical fields have any effect on the expectation of the numerical fields. For example, one test tests whether married individuals work significantly more or less than unmarried individuals.

(a) (3 points) Write a function `avg_difference_in_means` that takes as input two column names: `binary_col`, the name of a column with binary data, and `numerical_col`, the name of a column with numerical data. The function should compute the p-value for a test of the following hypothesis test:

H_0 : There is no difference in the average value of `numerical_col` between the two groups specified in `binary_col`.

H_1 : The average value of `numerical_col` is different for the two groups specified in `binary_col`.

For example, the result of `avg_difference_in_means('Post HS?', 'Age')` should be a p-value for testing whether there is a significant difference in age between college-educated and non-college-educated adults. You should use a permutation test (i.e., an A/B test from Data 8) to compute your p-values, using at least 25,000 permutations to form your final null distribution. Using such a large number of permutations will stabilize the p-values so that random noise is unlikely to lead to differing results across the class. On Datahub, running the full loop of tests should take a couple minutes.

Hint: It might be useful to recall how to run the simulations to get the necessary p-values. The Data 8 Textbook walks through an example of running a permutation test.

Hint: To shuffle a single column of a dataframe in pandas, you can use code similar to the following line. Make sure you use the correct arguments to the `sample` method!

```
df['my_column'] = df['my_column'].sample(...).values
```

(b) (1 point) Use your function to compute eight p-values, one for each possible combination of categorical and numerical column.

(c) (1 point) Suppose we use a naive p-value threshold of 0.05 to make a decision for each hypothesis test. Given the p-values from above, for which tests do we reject the null hypothesis?

(d) (2 points) Suppose we want to guarantee a Family-wise Error Rate (FWER) of 0.05. Given the p-values from above, for which tests do we reject the null hypothesis?

(e) (2 points) Suppose we want to guarantee a False Discovery Rate (FDR) of 0.05. Given the p-values from above, for which tests do we reject the null hypothesis?

Hint: Use the Benjamini-Hochberg algorithm.

(f) (2 points) How do the results from (d) and (e) compare? Explain how and why these results are different.

Hint: Recall how FWER and FDR are conceptually different.

(g) (2 points) Most variables don't always fit neatly into binary categories. As described earlier, we binarized these columns for you. Look at the original data in `adult_original.csv`. For one categorical column, give an example of how that variable could have been binarized differently, and how that might change the results from the earlier parts.

You aren't required to do any computation for this part: just explain how you might binarize one variable differently, and how that might change the results or your interpretation of them.

References

[1] Kleinberg, Jon and Mullainathan, Sendhil and Raghavan, Manish. *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807, 2016.

[2] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Cambridge University Press, 2012.