Data 102, Spring 2024
Homework 1

Due: **5:00 PM** Friday, February 2nd, 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. (10 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. (2 points) 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 $LR(T)$ 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 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.

Bias in Police Stops

2. The following example is taken from [1, 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.

(b) (2 points) 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. How many discoveries did you make (i.e. for how many police officers in the dataset did you reject the null hypothesis)?

(c) (2 points) 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 steps **(a)** and **(b)**, treating the empirical null as the null distribution.

(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

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