Overview

Submit your writeup, including any code, as a PDF via gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to reuse code will help in the long run!

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 in your answer.
   (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.

      Solution: 0.02, or 1 - specificity.

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

      Solution: $\pi_0 + \pi_1 = 1$. Applying Bayes rule,

      $$\mathbb{P}(H_0 \mid +) = \frac{\mathbb{P}(+ \mid H_0)\pi_0}{\mathbb{P}(+ \mid H_0)\pi_0 + \mathbb{P}(+ \mid H_1)(1-\pi_0)}.$$

      From part (a), $\mathbb{P}(+ \mid H_0) = .02$, and similarly $\mathbb{P}(+ \mid H_1) = 0.99$ is the given sensitivity, so

      $$\mathbb{P}(H_0 \mid +) = \frac{0.02 \cdot 0.995}{0.02 \cdot 0.995 + 0.99 \cdot 0.005} = 0.801.$$

      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.

[1] In Jupyter, you can download as PDF or print to save as PDF

Solution: Let $+_1$ denote a positive test result on the first test, and let $+_2$ denote a positive test result on the second test. By the law of total probability,

$$\mathbb{P}(+_2 \mid +_1) = \mathbb{P}(+_2 \mid +_1, H_0)\mathbb{P}(H_0 \mid +_1) + \mathbb{P}(+_2 \mid +_1, H_1)\mathbb{P}(H_1 \mid +_1).$$

Since the tests are independent given drug user status, $\mathbb{P}(+_2 \mid +_1, H_i) = \mathbb{P}(+_2 \mid H_i)$ for $i = 0, 1$. Hence,

$$\begin{aligned}\mathbb{P}(+_2 \mid +_1) &= \mathbb{P}(+_2 \mid H_0)\mathbb{P}(H_0 \mid +_1) + \mathbb{P}(+_2 \mid H_1)\mathbb{P}(H_1 \mid +_1) \\ &= .02 \cdot 0.801 + 0.99 \cdot (1 - .801) = 0.213.\end{aligned}$$

The positive test result will only be reproducible about 21.3% of the time.

(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 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.

Solution: We know that $f_1(t) = 2ce^{-2ct}$ and $f_0(t) = ce^{-ct}$. Therefore, $\text{LR}(t) = \frac{f_1(t)}{f_0(t)} = \frac{2ce^{-2ct}}{ce^{-ct}} = 2e^{-ct}$.

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.

Solution: FPR $= \mathbb{P}(LR(T) > \eta | H_0)$ and we want FPR $= \alpha$. For ease of calculation, we want to isolate T and put our equation in the form of $\mathbb{P}(T < \eta' | H_0)$.

$$\begin{aligned} & LR(T) > \eta \\ \implies & 2e^{-cT} > \eta \\ \implies & T < -\frac{1}{c}\log(\frac{\eta}{2}) \end{aligned}$$

Hence, we have $\mathbb{P}(T < \eta' | H_0)$ by defining $\eta' := -\frac{1}{c}\log(\frac{\eta}{2})$ (and equivalently, $\eta = 2e^{-c\eta'}$). $\mathbb{P}(T < \eta' | H_0) = \int_0^{\eta'} f_0(t)dt = \int_0^{\eta'} ce^{-ct}dt = 1 - e^{-c\eta'}$, so we have $\alpha = 1 - e^{-c\eta'}$. Rearranging, $\eta' = -\frac{1}{c}\log(1-\alpha)$, so $\eta = 2e^{-c\cdot(-\frac{1}{c}\log(1-\alpha))} = 2(1-\alpha)$.

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.

Solution: TPR or Power can be written as $\mathbb{P}(LR(T) > \eta | H_1)$. Following similar calculations as above, we have $\mathbb{P}(LR(T) > \eta | H_1) = \mathbb{P}(T > \eta' | H_1) = \int_0^{\eta'} f_1(t)dt = \int_0^{\eta'} 2ce^{-2ct}dt = 1 - e^{-2c\eta'}$. Substituting in η' from part (b), we get TPR $= 1 - e^{-2c\eta'} = 1 - (1-\alpha)^2$.

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.

We provide the data in a file `policez.csv`. Assume that under the null hypothesis (that the officers do not show racial bias), the z-scores should follow a standard normal distribution.

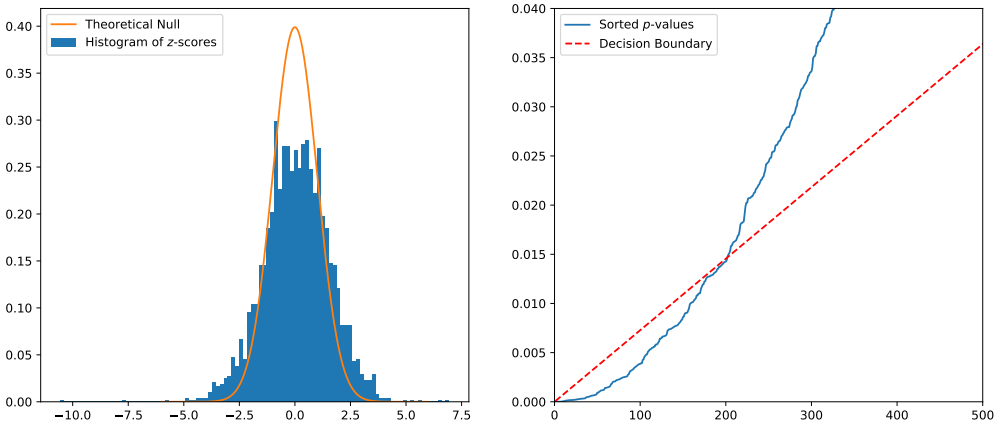
Note that throughout this question, the word "bias" refers to police officers' racial bias, rather than the statistical term.

(a) (1 point) In one plot, make a normalized histogram of the z-scores and a line plot of the pdf of the theoretical null $\mathcal{N}(0,1)$. Describe how the fit looks.

(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?

Solution: (a) and (b)

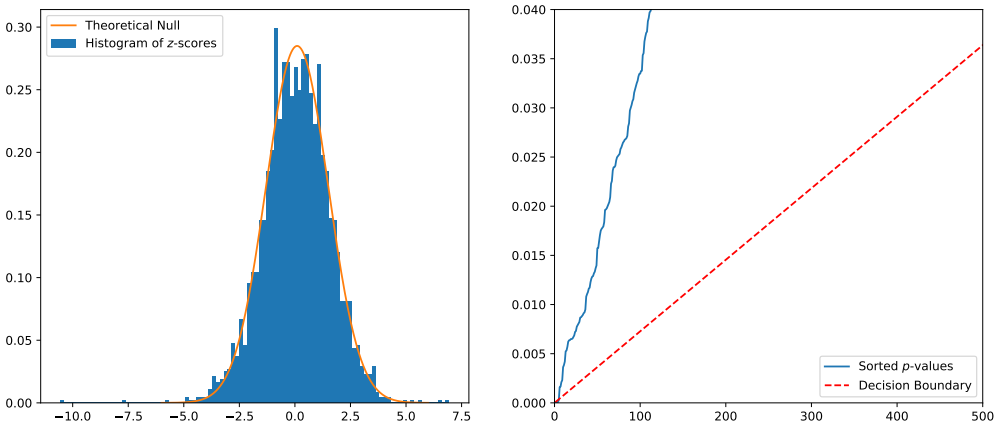
The z-values have heavier tails than the theoretical null. This could either mean a significant fraction of the subjects are non-null, or that the theoretical null $\mathcal{N}(0,1)$ otherwise fails to describe these statistics [1, see section 6.4]. Applying BH with these p-values, we made 201 discoveries.



(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.

Solution:

The empirical null has a much better fit. Set the p-values to $P_i = \Phi(-\frac{z_i - .1}{1.4})$ Applying BH with these p-values, we made 5 discoveries.



(d) (2 points) What assumption(s) are we implicitly making in part **(c)** by replacing the theoretical null $\mathcal{N}(0, 1)$ with one which fits the data well $\mathcal{N}(0.10, 1.40^2)$? What are the limitations of using the theoretical null? Which approach would you take when reporting discoveries of racial bias in this example? What other limitations do you see to this approach to modeling racial bias?

Solution: See [1, section 6.4] for a discussion of issues with the theoretical null. Correlation between the z_i might explain over-dispersion.

By fitting a single normal distribution to the nulls, we are essentially assuming π_0 is large (in fact, $\pi_0 \geq .9$ was assumed in fitting the empirical null). The authors of the original study [**?**] noted this limitation of the empirical null:

> Implicit in the proposed framework, which draws on a multiple-comparison idea relevant to hypothesis testing, is an assumption that numerous officers have the same level of bias, which is either near zero or identically equal to zero. Although the method compares officers to their peers, it is not necessarily the case that their peers are unbiased. If, for example, all of the officers in a precinct act in a racially biased manner then when each is compared with the others, none of the officers in this precinct will be flagged as problematic. Only in the case that most officers are unbiased and only a few are problematic, the setting several police executives have suggested, will the method actually measure race bias among officers.

Studies of racial bias in policing should not consider individual officers as the only unit of analysis, but should also look at deployment and enforcement disparities within and across entire departments. See also [**?**] for a Bayesian regression analysis of stop-and-frisk records.

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('Education', '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. DATA 8 Textbook *provides a helpful refresher.*

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.

Solution: The following are results after 10,000 simulations, note that students' results may be slightly different due to Monte Carlo variance.

	Age	Hours per week
Post HS?	0.70748	0.9288
Ever Married?	0	0.09776
Gender	0.0426	0.01708
50K?	4e-5	0.23516

(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?

Solution: With the p-values above, we make 4 rejections: Ever married and Age; Gender and Age, 50K and Age, Gender and hours per week

(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?

Solution: If we want to guarantee a FWER of .05, we need to use the Bonferroni correction, which reduces the rejection threshold from .05 to $\frac{.05}{m}$, where $m = 8$ is the number of tests. Our new threshold is thus $\frac{.05}{8} = .00625$, and the only two rejections we make are Ever married and Age; 50K and Age.

(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.

Solution: Per the hint, we need to use the Benjamini-Hochberg question to control the FDR.
With this less conservative BH rejection threshold, we also reject Gender and Hours per Week (in addition to the two rejections under Bonferroni). Note that it is possible (though very very unlikely to get different rejection sets due to RNG. confirm students have used at least 2,500 permutations and that their p-values are similar to the 2(b) table.).

(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.

Solution: Bonferroni is controlling for FWER, which is the probability of *at least one* false positive. This is a very tough requirement to meet, so Bonferroni results in a lot stricter threshold. Benjamini-Hochberg, on the other hand, is controlling for FDR, which allows for more false positives, as long as there are more true positives as well. So the threshold with B-H is not as strict.

(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.

Solution: There are many acceptable answers here, including:

- Different threshold could be chosen for income (higher or lower than $50,000)
- Comment about using gender instead of sex, or vice versa
- Marital status could be "currently married" or "in a committed relationship" instead of "ever married."

For example, higher different thresholds for the income could make the correlation with age and hours per week higher. Any sort of general explanation about how changing the discretization and how that effects the correlation between values should constitute a correct answer.

References

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