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.

1. Math Stats

Work through the following exercises, and explain your reasoning in your answer.

(a) Suppose a particular drug test is 99% sensitive and 98% specific. 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) the probability of testing positive given H_0.

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

(iii) 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 n independent null p-values $P_1, \ldots, P_n$ that are uniformly distributed on $[0, 1]$. In other words, these are all p-values from tests where the null hypothesis was true. Recall that the level-α Bonferroni procedure rejects all p-values P_i such that $P_i \leq \alpha/n$. What is the probability that the level-α Bonferroni procedure rejects any of the P_i? Show that your answer is less than or equal to α.

Hint : Let E_i be the event P_i is rejected.

(c) Suppose we have a waiting time $T \sim$ Exponential(λ) and wish to test $H_0 : \lambda = c$ versus $H_1 : \lambda = 2c$ for some $c > 0$. The likelihood ratio test considers the ratio of the two density functions, $\text{LR}(T) := \frac{f_1(T)}{f_0(T)}$, and rejects H_0 when $\text{LR}(T) > \eta$ for some threshold η to be determined.

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

(ii) Let $0 < \alpha < 1$. 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.

(iii) What is the TPR of this test? This is also known as the test's *power*.

Remark: The Neyman-Pearson lemma says that the likelihood ratio test is the most powerful test 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.

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

(d) Consider the null hypothesis H_0 that random variable X has tail cdf $\bar{F}_0$, and the alternative hypothesis H_1 that X has tail cdf $\bar{F}_1$. Assume that $\bar{F}_1(x) \geq \bar{F}_0(x)$ for all x and that $\bar{F}_0$ is invertible. Show that, under the *alternative* H_1, the p-value $P = \bar{F}_0(X)$ is *sub-uniform*, i.e. $\mathbb{P}(P \leq p) \geq p$ for all $p \in [0, 1]$. What does this mean for p-values under the alternate hypothesis H_1?

Hint: P-values under the null distribution are uniform.

Hint: Under the condition $F_1(x) \geq F_0(x)$ for all x, then there is a realization of (X, Y) on the same probability space such that $X \sim F_1$, $Y \sim F_0$, and $X \geq Y$.

2. Online Experiments

In some applications of multiple testing, it is not possible to collect all p-values before making decisions about which hypotheses should be proclaimed discoveries. For example, in A/B testing in tech, p-values arrive in a continual stream, so decisions have to be made in an online fashion, without knowing the p-values of future hypotheses. In this question, we compare an online algorithm for FDR control called LORD with the classical Benjamini-Hochberg (BH) procedure. We will provide an implementation of the LORD algorithm, however, for completeness, we also state the steps of the LORD algorithm below. Don't worry if you don't have intuition for the α_t update; the important thing is that such an update ensures that FDR is controlled at any given time t.

Algorithm 1 The LORD Procedure

input: FDR level α, non-increasing sequence $\{\gamma_t\}_{t=1}^{\infty}$ such that $\sum_{t=1}^{\infty} \gamma_t = 1$, initial wealth $W_0 \leq \alpha$

Set $\alpha_1 = \gamma_1 W_0$

for $t = 1, 2, \ldots$ **do**

  p-value P_t arrives

  if $P_t \leq \alpha_t$, reject P_t

  $\alpha_{t+1} = \gamma_{t+1} W_0 + \gamma_{t+1-\tau_1}(\alpha - W_0)\mathbf{1}\{\tau_1 < t\} + \alpha \sum_{j=2}^{\infty} \gamma_{t+1-\tau_j} \mathbf{1}\{\tau_j < t\}$,

  where τ_j is time of j-th rejection $\tau_j = \min\{k : \sum_{l=1}^{k} \mathbf{1}\{P_l \leq \alpha_l\} = j\}$

end

While offline algorithms like Benjamini-Hochberg take as input a *set* of p-values, online algorithms take in an *ordered sequence* of p-values. This makes their performance sensitive to p-value ordering. In this exercise we analyze this phenomenon. Generate $N = 1000$ p-values in three different ways:

(i) For every $i \in \{1, \ldots, N\}$, generate $\theta_i \sim \text{Bern}(1 - \pi_0)$. If $\theta_i = 0$, the p-value P_i is null, and should be generated from Unif$[0, 1]$. If $\theta_i = 1$, the p-value P_i is an alternative. Then, generate $Z_i \sim \mathcal{N}(3, 1)$, and let $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian $\mathcal{N}(0, 1)$ CDF.

(ii) For $i = 1, \ldots, \pi_0 N$, set $\theta_i = 0$, meaning the hypothesis is truly null, and let $P_i \sim$ Unif$[0, 1]$. For $i = \pi_0 N + 1, \ldots, N$, $\theta_i = 1$, and the hypothesis is truly alternative. Then, generate $Z_i \sim \mathcal{N}(3, 1)$, and let $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian $\mathcal{N}(0, 1)$ CDF.

(iii) For $i = 1, \ldots, N - \pi_0 N$, set $\theta_i = 1$, meaning the hypothesis is alternative, generate $Z_i \sim \mathcal{N}(3, 1)$, and let $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian $\mathcal{N}(0, 1)$ CDF. For $i = N - \pi_0 N + 1, \ldots, N$, $\theta_i = 0$, and the hypothesis is truly null; let $P_i \sim \text{Unif}[0, 1]$.

(a) Run the LORD algorithm with $\alpha = 0.05$ on three p-value sequences, given as in (i), (ii) and (iii), respectively. Compute the false discovery proportion (FDP) and sensitivity. Repeat this experiment 100 times to estimate FDR as the average FDP over 100 trials, as well as the average sensitivity. Do this for all $\pi_0 \in \Pi_0 := \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Make the following plots:

- FDR estimated over 100 trials on the y-axis against $\pi_0 \in \Pi_0$ on the x-axis, for the three different scenarios (i), (ii) and (iii).
- Expected sensitivity estimated over 100 trials on the y-axis against $\pi_0 \in \Pi_0$ on the x-axis, for the three different scenarios (i), (ii) and (iii).

For which of the three scenarios (i), (ii), (iii) does LORD achieve highest average sensitivity? Can you give an intuitive explanation for this?

(b) Now also run the Benjamini-Hochberg procedure with $\alpha = 0.05$ for settings (i), (ii), (iii) on the whole batch; generate all of N p-values, and then apply BH. Make the same plots as in part (a). How does the sensitivity of BH compare to the sensitivity of LORD? How does the sensitivity of BH compare in settings (ii) and (iii)?

3. Bias in Police Stops

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`.

(a) 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) Compute p-values $P_i = \Phi(-z_i)$ 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?

(c) A better fit to the z-scores is given by $\mathcal{N}(0.10, 1.40^2)$, called the empirical null. Repeat steps **(a)** and **(b)** treating the empirical null as the null distribution.

(d) 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 bias in this example? What other limitations do you see to this approach to modeling bias?

References

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