DS 102 Homework 1

Name:

Discussed with:

If you are handwriting your solution please make sure your answers are legible, as you may lose points otherwise.

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you do discuss the homework with others, please include their names on your submission.

Due on Gradescope by 9:29am, Tuesday 11th February, 2020

This homework involves a fair amount of coding, so 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!

1. (20 points) **Fundamentals of Decision Theory.** For the following two parts, identify the components of the decision theoretic framework: the data X, the parameter of interest θ, the decision procedure $\delta(X)$, and the loss function $\ell(\theta, \delta(X))$.

   (a) (3 points) During World War II, the Allied forces attempted to estimate the total number of German tanks that had been produced, given the serial numbers of the k tanks they had captured. Assume serial numbers corresponded to the order in which the tanks were produced (*e.g.* tank #126 was the 126-th tank produced). A method that turned out to be quite accurate, in terms of squared error, was to estimate the maximum serial number m captured, plus the average gap between captured serial numbers: $m + (m - k)/k$.

   (b) (3 points) Evolutionary biologists are interested in detecting mutations that are adaptive (arose then persisted due to selection) rather than neutral (arose then persisted just by chance). If the same mutation (say, from an `A` to a `C` at certain site) occurred independently in N different species, one way to decide it is adaptive is if the probability that the specific mutation arose N times under neutral mutation rates* is below some threshold η. Because this decision is used for downstream scientific conclusions, there is a cost C_{01} when an adaptive mutation is labeled as neutral, and a cost C_{10} when a neutral mutation is labeled as adaptive, and no cost otherwise.

   * Values of $\mathbb{P}(i \text{ mutates to } j \text{ then does not mutate})$ for $i, j \in \{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ when mutations are known to be neutral.

   For the following two parts, derive the decision procedure δ^* that minimizes the risk given different loss functions. That is, provide an expression for

   $$\delta^* = \min_{\delta} R(\delta) = \min_{\delta} \mathbb{E}_{\theta, X \sim \mathbb{P}}[\ell(\theta, \delta(X))]$$

   when:

(c) (5 points) $\ell(\theta, \delta(X)) = (\theta - \delta(X))^2$

(d) (9 points) $\ell(\theta, \delta(X)) = \mathbf{1}[\theta \neq \delta(X)]$

2. (20 points) **Likelihood Ratio Test and Neyman-Pearson Lemma.**

Let X be a continuous random variable distributed over the closed interval [0,1]. Under the null hypothesis H_0, X is uniform:

$$\mathbb{P}(X|H_0) = \begin{cases} 1 & 0 \leq X \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Under the alternative hypothesis H_1, the conditional pdf of X is as follows:

$$\mathbb{P}(X|H_1) = \begin{cases} 2X & 0 \leq X \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The *a priori* probability that X is uniformly distributed is $p \in (0, 1)$. Associated with the possible decisions are the costs $C_{00} = C_{11} = 0$ and $C_{01}, C_{10} \in [0, \infty]$, where C_{ij} is the cost of deciding $\hat{H}(X) = H_i$ when the correct hypothesis is $H = H_j$.

(a) (10 points) **Find** the decision rule that minimizes the probability of error.
Hint: one approach is to proceed in two steps. First, the rule that minimizes the probability of error corresponds to a likelihood ratio test for some significance level (why?). Next, note that the probability of error can be written as a function of the threshold in the likelihood ratio test.

(b) (5 points) Define P_D and P_F as::

$$P_D = \mathbb{P}(\hat{H} = H_1 | H = H_1)$$
$$P_F = \mathbb{P}(\hat{H} = H_1 | H = H_0)$$

Express P_D as a function of P_F for the likelihood ratio test (LRT).
Note that this closed form expression is called the operating characteristic of LRT.

(c) (5 points) **Determine** whether the following statement is **true or false**, and **justify** your answer:
For at least some value(s) of P_F, there exists some other decision rules that can achieve a greater P_D than that corresponding to the operating characteristic you found in part (b).

3. (20 points) **Offline FDR Control.** Consider testing $N = 1000$ hypotheses $H_1, \ldots, H_N$, and let $\mathcal{H}_0 \subseteq \{1, \ldots, N\}$ denote the indices of the nulls among them (so that $i \in \mathcal{H}_0$ if index i correspond to a null). Denote by π_0 the proportion of true null hypotheses, $\pi_0 = |\mathcal{H}_0|/N$. Denote by $P_1, \ldots, P_N$ the corresponding p-values. Suppose that the alternative p-values $P_i, i \notin \mathcal{H}_0$ are equal to 0.01 with probability one, and that the null p-values $P_i, i \in \mathcal{H}_0$ are as usual independent and uniformly distributed on $[0, 1]$. Our target FDR or FWER level is $\alpha = 0.05$.

FDR is a problem when the proportion of alternatives among the N hypotheses is low. In this exercise, we aim to demonstrate this statement in practice.

(a) (5 points) Suppose that you apply the "classical" uncorrected decision strategy: reject H_i if $P_i \leq \alpha$. Express the resulting FDR in terms of π_0, N and α.
(Hint: The number of null hypotheses H_i which have $P_i \leq \alpha$ is in some sense the number of "successes" in $N\pi_0$ trials, where each trial succeeds with probability $\mathbb{P}(P_i \leq \alpha) = \alpha$. What distribution is this? You don't have to simplify the final expression much.)

(b) (3 points) Assuming the decision rule from part (a), plot the FDR against π_0, for $\pi_0 \in \Pi_0 := \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. Do *not* approximate it; use the formula obtained in part (a). What is the expected sensitivity of this decision rule? Recall that

$$\mathbb{E}[\text{sensitivity}] = \mathbb{E}\left[\frac{\text{number of true discoveries}}{N(1-\pi_0)}\right].$$

(c) (5 points) Now consider the Bonferroni correction. What is the expected sensitivity of this procedure? Express the FDR in terms of π_0, N and α (you don't need to simplify the expression). Plot this expression for the FDR against $\pi_0 \in \Pi_0$, where Π_0 is defined in part (b). Recall that FDR $= \mathbb{E}[\text{FDP}]$, and FDP $= 0$ if there are no discoveries (i.e. we "assume" 0/0=0).

(d) (2 points) Assuming that the alternative p-values are still constant (but not necessarily equal to 0.01), how much would you have to decrease them for the Bonferroni procedure to discover all of them?

(e) (5 points) Now use the Benjamini-Hochberg procedure to find discoveries. In this part, we will approximate the average sensitivity and FDR through simulation. Approximate both by averaging the false discovery proportion and sensitivity over 100 independent simulations. Note that the randomness should only come from the null p-values. Plot the resulting FDR and expected sensitivity against π_0, for $\pi_0 \in \Pi_0$. Compare your observations in part (c) and part (e).

4. (20 points) **Online FDR Control.** 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 the IT industry, p-values arrive in a virtually never-ending 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 $\alpha$, 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 $\tau_j$ 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.

(a) (15 points) You will generate $N = 1000$ p-values in three different ways:

(i) For every $i \in \{1, \ldots, N\}$, generate $\theta_i \sim \text{Bern}(\pi_0)$. 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 $Z_i \sim N(3, 1)$, and let $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian $N(0, 1)$ 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 $Z_i \sim N(3, 1)$, and let $P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian $N(0, 1)$ CDF.

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

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?