DS102 - Homework 2

If you are submitting a LaTeXversion please use the provided template. If you are submitting a handwritten version please make sure your answers are legible as you may lose points otherwise.
Data science is a collaborative activity. While you may talk with other about the homeworks, we ask that you write your solutions individually. If you do discuss the homework with others please include their name in your submission.
Due by: 1:59pm, Tuesday 24th September, 2019

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 greatly help in the long run!
For sampling from a Gaussian, consider using `numpy.random.normal` and for CDF of Gaussians, consider using `scipy.stats.norm`. Using pandas and matplotlib may be helpful as well.

1. (20 points) In multiple testing, weighted procedures are used for various reasons. In this exercise, we investigate how weights might affect the set of discoveries. For now, we make no assumption about how many true nulls or alternatives there are.

   Suppose we want to test N hypotheses. For example, we might be testing N different treatments for a disease, or we might be running N industrial A/B tests. Assign each hypothesis a weight $w_i \geq 0$ so that the family of weights obeys $\frac{1}{N}\sum_{i=1}^{N} w_i = 1$. Suppose that for every hypothesis we have computed a p-value P_i. As usual, assume that the p-values corresponding to true nulls are uniformly distributed:

   $$\text{If } P_i \text{ corresponds to a null, } \mathbb{P}(P_i \leq u) = u, \text{ for all } u \in [0,1].$$

   Consider the procedure that makes a discovery if $P_i/w_i \leq \alpha/N$, and makes no discovery if $w_i = 0$.

   (a) (5 points) Show that this procedure controls the probability of at least one false discovery (i.e. the FWER) under level α.

   (b) (3 points) What do weights all equal to 1 represent? What does w_i greater or less than 1 represent? What are some scenarios in which weights are useful?

   (c) (12 points) Consider $Z \sim N(\mu, I) \in \mathbb{R}^{500}$, where

   $$\mu_i = \begin{cases} 0 & 1 \leq i \leq 450, \\ \frac{i-450}{5} & \text{otherwise.} \end{cases}$$

   For each i, we want to test whether the null is true - $\mu_i = 0$ or the alternative - $\mu_i > 0$. To do so, we compute the i-th p-value as

$P_i = \Phi(-Z_i)$, where Φ is the standard Gaussian $N(0,1)$ CDF. You will implement the following testing procedures.

(i) "Uncorrected" testing. This just means that you reject a hypothesis if $P_i \leq \alpha$, for a fixed α.

(ii) Bonferroni procedure.

(iii) Benjamini-Hochberg procedure.

(iv) Weighted procedure from part (a) with $w_i = \frac{2i}{501}$.

(v) Weighted procedure from part (a) with $w_i = \frac{2(501-i)}{501}$.

(vi) Weighted procedure from part (a) with $w_i = 0.5$ for $1 \leq i \leq 450$ and $w_i = 5.5$ for $i > 450$.

For all six methods, repeat the experiment 100 times. Over these 100 trials, compute and plot the average number of rejections, average number of true rejections, and achieved false discovery rate, against $\alpha \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Compare the different approaches in your own words. How do the three weightings differ? Which procedure (out of {i, ii, iii} above) is the general weighted procedure most similar to?

2. (20 points) Consider testing $N = 1000$ hypotheses $H_1, \dots, H_N$, and let $\mathcal{H}_0 \subseteq \{1, \dots, N\}$ 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 $P_1, \dots, P_N$ 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 $[0,1]$. The target FDR or FWER level is $\alpha = 0.05$.

In lecture, we argued that 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, again without having 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 $= \mathbb{E}$[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).

3. (20 points) 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 $i \in \{1, \dots, N\}$, 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 $i = 1, \dots, \pi_0 N$, set $\theta_i = 0$, meaning the hypothesis is truly null, and let $P_i \sim \text{Unif}[0, 1]$. For $i = \pi_0 N + 1, \dots, N$, $\theta_i = 1$, 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 $i = 1, \dots, N - \pi_0 N$, 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 $i = N - \pi_0 N + 1, \dots, N$, $\theta_i = 0$, and the hypothesis is truly null; let $P_i \sim \text{Unif}[0, 1]$.

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? Think about the "wealth" interpretation given in lecture.

(b) (5 points) Now also run the Benjamini-Hochberg procedure 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)?

4. (20 points) In this exercise, we prove that the Benjamini-Hochberg (BH) procedure controls FDR. Recall the steps of the procedure:

Algorithm 2 The Benjamini-Hochberg Procedure

```
input: FDR level $\alpha$, set of $N$ p-values $P_1, \dots, P_N$
Sort the p-values $P_1, \dots, P_N$ in non-decreasing order $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(N)}$
  Find $R = \max\{i \in \{1, \dots, N\} : P_{(i)} \leq \frac{\alpha}{N} i\}$
  Reject the hypotheses corresponding to $P_{(1)}, \dots, P_{(R)}$
```

Assume that the p-values $P_1, \dots, P_N$ are independent. Let R denote the number of rejections made by the BH algorithm. Now imagine that we erase a single P_i from the input p-value sequence, for some $i \in \{1, \dots, N\}$. Denote the ordered p-values in this "modified" set by $P_{(1)}^{(-i)}, \dots, P_{(N-1)}^{(-i)}$. Also let $R^{(-i)} = \max\{j \in \{1, \dots, N-1\} : P_{(j)}^{(-i)} \leq \frac{\alpha}{N}(j+1)\}$. The variable $R^{(-i)}$ is the number of rejections in $P_{(1)}^{(-i)}, \dots, P_{(N-1)}^{(-i)}$, if we "lifted" the usual linear threshold by α/N. Recall the visual illustration of the BH procedure and try to illustrate $R^{(-i)}$ for yourself.

(a) (5 points) Argue that the event $\{P_i \leq \frac{\alpha}{N}r, R = r\}$ is equal to the event $\{P_i \leq \frac{\alpha}{N}r, R^{(-i)} = r - 1\}$, for every $r \in \mathbb{N}$.
(Hint: The visual illustration of the BH method could be very useful. Draw the ordered set of p-values before and after P_i is erased.)

(b) (5 points) Let $\mathcal{H}_0$ denote the indices in $\{1, \dots, N\}$ which correspond to null p-values/hypotheses. The false discovery proportion is equal to

$$\text{FDP} = \frac{1}{R} \sum_{i \in \mathcal{H}_0} \mathbf{1}\{P_i \leq \frac{\alpha}{N} R, R > 0\}.$$

Using part (a), prove that the FDP is equal to

$$\sum_{i \in \mathcal{H}_0} \sum_{r=1}^{N} \frac{1}{r} \mathbf{1}\{P_i \leq \frac{\alpha}{N} r, R^{(-i)} = r - 1\}.$$

(Hint: Notice that you can write $\mathbf{1}\{R > 0\} = \sum_{r=1}^{N} \mathbf{1}\{R = r\}$, and recall that $\mathbf{1}\{E_1, E_2\} = \mathbf{1}\{E_1\}\mathbf{1}\{E_2\}$, for any two events E_1 and E_2.)

(c) (2 points) The FDR is equal to the expectation of the FDP, FDR = $\mathbb{E}[\text{FDP}]$. Using part (b), prove that

$$\text{FDR} = \sum_{i \in \mathcal{H}_0} \sum_{r=1}^{N} \frac{1}{r} \mathbb{P}\left(P_i \leq \frac{\alpha}{N} r, R^{(-i)} = r - 1\right).$$

(d) (3 points) Use the fact that null p-values are uniformly distributed to show

$$\text{FDR} = \sum_{i \in \mathcal{H}_0} \sum_{r=1}^{N} \frac{\alpha}{N} \mathbb{P}\left(R^{(-i)} = r - 1 | P_i \leq \frac{\alpha}{N} r\right).$$

(e) (5 points) Use independence between p-values to show FDR $\leq \alpha$.
(Hint: Are P_i and $R^{(-i)}$ dependent?)