Homework 4

Due May 6,2023

Problem 1

In this homework we will write our own code to implement the BH procedure and the adaptive z-value procedure (the optimal procedure based on posterior probability). We will also see the importance of estimating the null distribution. Here is the model:

$$X_i \sim (1-\pi)f_0 + \pi f_1.$$

The theoretical null is $f_0 \sim N(0,1)$, but in reality the null distribution follows $N(\mu_0, \sigma_0^2)$ with μ_0 and σ_0^2 unknown. The p-value under the theoretical null is computed as

$$p_i = 2\Phi(-|x_i|),$$

where $\Phi(\cdot)$ is the CDF of $N(0,1)$. Under the actual null, the p-value is computed as

$$p_i = 2\Phi\left(-\left|\frac{x_i-\mu_0}{\sigma_0}\right|\right).$$

Define $z_i = \dfrac{x_i-\mu_0}{\sigma_0}$, note that under the theoretical null $z_i = x_i$.

1.1

Write a function called "bh.func" that implements the BH procedure.
Input: a vector of p-values $(p_1, p_2, ..., p_m)$, a number α represents the desired FDR level.
Output: A vector of 0 and 1's that represents your decision. 0 = not reject the ith null, 1 = rejects the ith null.

1.2

Write a function called "az.func" that implements the adaptive z-value procedure.
Input: a vector of z-values $(z_1, z_2, ..., z_m)$, a number α represents the desired FDR level.
Output: A list contains the following: a vector de of 0 and 1's that represents your decision. 0 = not reject the ith null, 1 = rejects the ith null. A number pi represents the estimated alternative proportion.

Hint: You can use the following code for density estimation, here zv is the vector of z-values

```
den=density(zv, from=min(zv)-10, to=max(zv)+10, n=2000)
```

Note that the above code only gives you estimated density at $den\$x$. To estimate the density at points that are not in $den\$x$, you can connect the estimated density at the two adjacent points using a straight line.
Alternatively you can use use $den\$bw$ as bandwidth and calculate your own kernel estimate.

1.3

Write a function called "EstNull.func" that estimates the null distribution:
Input: the observation vector $(x_1, x_2, ..., x_m)$
Output: a vector $(\hat{\mu}_0, \hat{\sigma}_0)$ represents the estimated null distribution.

1.4

Import the data:

```
d <- read.csv("hw4training")
```

d is a 10000×2 dataframe. The first column $d\$x$ is the observation, the second column $d\$theta$ is the ground truth, 0 means the observation is generated from the null distribution, 1 means the observation is generated from the alternative distribution. Assume the theoretical null, apply the BH and the adaptive z-value procedure on $d\$x$ at $\alpha = 0.1$. What are the FDPs? How many alternative hypotheses are correctly rejected by each procedure? Is the result expected?
Hint: The FDP for the adaptive z-value procedure should be very close to α, and the FDP for BH should be slightly less than α.

1.5

Import the data:

```
d <- read.csv("hw4data")
```

d is a 10000×2 dataframe. The first column $d\$x$ is the observation, the second column $d\$theta$ is the ground truth, 0 means the observation is generated from the null distribution, 1 means the observation is generated from the alternative distribution. Assume the theoretical null, apply the BH and the adaptive z-value procedure on $d\$x$ at $\alpha = 0.1$. What are the FDPs? How many alternative hypotheses are correctly rejected by each procedure?
Hint: The FDPs should be higher than α.

1.6

The fact that the observed FDPs are higher than α in problem 1.5 indicates that the use of theoretical null is not appropriate. Now use "EstNull.func" to

estimate the null hypothesis. What is null estimated null? Recompute the p-values and z-values then apply the BH and the adaptive z-value procedure on $d\$x$. What are the FDPs? Which procedure is more powerful?
Hint: Now the FDPs should be around or less than α .