HDS Exercise set 2.

YOUR NAME (STUDENT NUMBER)

Return by **10.15** o'clock on **12.11.2019** to the Moodle area of the course. You can use 'HDS_ex2.Rmd' file as a template for your answers, in which case each question will also be shown there before your own solution. Return the final file in pdf format with name "HDS2_yourname.pdf".

Problem 1.

Let's work with example data from `qvalue` package. Load the package `library(qvalue)` (see first lecture notes for installing `qvalue` from Bioconductor if you haven't done that yet) and call `data(hedenfalk)`.

This Hedenfalk data has 3170 genes tested for differential gene expression between $n_1 = 7$ BRCA1- and $n_2 = 8$ BRCA2-mutation-positive tumors. The analysis is done by t-test statistic of difference in group means and the P-values have been computed by permutation test over the tumor type labels (because standard t-test assumptions may not hold). The data has a component `hedenfalk$p` of P-values.

(a) Plot a histogram of P-values using 50 bins. Draw a horizontal line that corresponds to Uniform(0,1) in this histogram. What do you infer visually about proportion of null and non-null P-values here?

(b) Compute a crude estimate of π_0, the proportion of null P-values in the histogram, using the method from lectures (Q-value part) with parameter $\lambda = 0.5$. Apply `qvalue()` function to these P-values and extract the `pi0` estimate from the output of `qvalue()`. Does it agree with your manual estimate?

(c) For increasing sequence of threshold `t=seq(0.01,0.99,0.01)` compute how many discoveries you would do when you determined the significance threshold by P-value, by Q-value by BH adjusted P-value or by Bonferroni corrected P-value. Don't print these numbers out, but make a plot where t is on x-axis and log10(number of discoveries) is on y-axis and represent the four inference methods by four lines with different colors. Explain whether the picture looks as you had expected?

(d) Run `plot(qval)` where `qval` is the object returned by `qvalue()` and read from the plots: How many discoveries you would make if you allowed 20 false discoveries? What about if you allowed 10% of false discoveries among all discoveries?

Problem 2.

Let's examine the relationship between BH adjusted P-values and Storey's Q-values. Generate $p = 5000$ P-values by command `pval = c(rbeta(m,1,100), runif(p-m,0,1))` where $m = 1000$. (So here first m P-values come from the alternative distribution and the remaining $p - m$ are from the null.)

Apply both BH adjustment (`p.adjust(,method="BH")`) and `qvalue()` to these P-values.

Print out the estimate of $\pi_0 = p_0/p$ given by `qvalue`.

Do linear regression of qvalues on BH-adjusted P-values. Compare the slope to the estimate of π_0 from `qvalue` and explain why you see what you see.

Problem 3.

Let's evaluate how `qvalue` works for different alternative distributions. Let's generate data sets with $p = 5000$ P-values of which $m = 500$ correspond to true effects, whose P-values are generated from Beta(b_1,b_0) distribution (use `rbeta(m, b.1, b.0)` in R), and the remaining $p_0 = p - m = 4500$ P-values come from the null distribution Uniform(0,1).

(a) To familiarize yourself with the Beta distribution, that we use to generate the non-null P-values, draw the density functions of all three beta distributions speficied by varying parameters (b_1, b_0) below. (You can either use `dbeta(x,b.1,b.0)` to get the densities where vector `x` spans the interval (0,1) and use `plot()` once and then `lines()` to add the curves or you can use `curve()` as in the lecture notes by specifying the function and the interval.)

(b) For each set of parameters (b_1, b_0) given below, generate $R = 500$ replications of the above described data simulation. For each replicate, apply `qvalue()` to calculate the false discovery proportion (FDP) and the proportion of true effects that are discovered ("power"), both at the FDR level $\alpha_F = 0.1$, and collect also the estimate of `pi0`. For each set of parameters, draw three histograms: FDP, power and `pi0` across R replications and show the means of the distributions in the titles.

Does `qvalue()` work as promised in terms of FDR control? What explains the differences in power across the settings (i),(ii) and (iii)?

(i) $b_1 = 1, b_0 = 1$

(ii) $b_1 = 1, b_0 = 100$

(iii) $b_1 = 1, b_0 = 500$.

Problem 4.

Let's study how P-values and Q-values behave in a (very) discrete space.

Suppose you are given $p = 1000$ coins and your task is to determine which proportion of them are fair (that is, on average, will result in heads in 50% of tosses and in tails in 50% of tosses). You do an experiment where you toss each coin 2 times and record the number of heads $y_j \in \{0, 1, 2\}$ for each coin $j \leq p$. Your null hypothesis for each coin is that the coin is fair.

(a) What is the null distribution of the outcome values of a single coin tossed 2 times? What is the null distribution of (two-sided) P-values of a single coin tossed 2 times? (In lectures, it was stated that the null distribution of P-values is Uniform(0,1), or equivalently, that $\Pr(P_j \leq t \,|\, \text{NULL}) = t$ for all $t \in [0, 1]$, but now we learn that, in a discrete state space, we need to restrict this formula to exactly those threshold values t that correspond to the P-values attainable in the discrete state space.)

(b) Suppose that, with $p = 1000$ coins, the observed counts of outcomes 0, 1 and 2 heads are 180, 366 and 454, respectively. What is the observed P-value distribution of these p observations? How would you estimate π_0, the proportion of fair coins, by comparing the observed P-value distribution to the null distribution? What is your estimate $\widehat{\pi}_0$? What is your estimate of Q-value for obsevations whose P-value is 0.5? (You may assume that all biased coins are fully biased, that is, can only yield either heads or tails but never both.)

(c) What would be the estimate of $\widehat{\pi}_0(\lambda)$ from the lectures HDS3 for value of $\lambda = 0.4, 0.5, 0.9$? Which of these three values (if any) agrees with what you inferred in part (b)? Apply `qvalue()` to the set of your p P-values and show `plot(qvalue())`. What is the estimated Q-value for obsevations whose P-value is 0.5 using `qvalue()`? Does `qvalue()` seem to work well for these kinds of discrete data?

Problem 5.

Let's see how well lfdr approximates the posterior probability of the null hypothesis. Simulate $p = 1000$ P-values of which $p_0 = 800$ come
from the null distribution (Uniform(0,1)) and $m = 200$ from the non-null distribution Beta(1,100).

(a) For each P-value P_j, compute the posterior probability that the P-value comes from the null hypothesis (H_j) given the knowledge of the true non-null distribution and the true proportion $\pi_0 = p_0/p$. (HINT: Expand $\Pr(H_j \,|\, P_j, \pi_0, b_1 = 1, b_0 = 100)$ by using Bayes formula to switch the roles of H_j and P_j.)

(b) Make a scatter plot of posteriors from part (a) and lfdr values from `qvalue()` function applied to the P-values using different colors according to the known null/non-null status. Do these two quantities look similar?

(c) Compute the average values separately for truly null and non-null hypotheses using (i) the exact posterior probability of null hypothesis and (ii) estimated lfdr.