Homework Set #6 – Due 11:59pm, Friday, March 7.
MATH 445 Winter 2025

Note:

- This homework assignment is based on the Chapters 7 material.
- Your answers must be presented in the same order as the problems given in this assignment, including all graphs and `R` outputs, if applicable (i.e., the graphs and outputs must be in the same order and must not be relegated to the back of your assignment).
- Use your judgment to include the minimal amount of `R` code in your answers as necessary.
- All the `R` code must be well documented and submitted as a separate .R file to Canvas.

The grading scheme is as follows:

- This homework assignment will be graded out of 30 points.
- Every single problem will be checked carefully. Each of the three problems will be graded out of 10 points.
- If the R code is missing or incomplete, at most 5 points will be deducted.
- If the answers are not presented in the right order, at most 5 points will be deducted.

Background: Storey and Tibshirani (2003) analyzed the hereditary breast cancer microarray gene expression data of Hedenfalk et al. (2001). The purpose of their study is to identify potentially differentially expressed (DE) (i.e., 'interesting') genes which may potentially be contributing to some differences between two types of mutations named BRCA1 and BRCA2. A typical (traditional) way of analyzing gene expression data is to compare the means (or distributions) of the log base 2 intensity values of the genes for two groups of subjects. In this dataset, we have $n_1 = 7$ subjects with the BRCA1 mutation and $n_2 = 8$ subjects with the BRCA2 mutation.

The goal of this homework is to understand how microarray data analysis may be performed to identify potentially DE genes. What is unique about microarray data

analysis in general is that it typically consists of thousands (or even tens or hundreds of thousands) of genes. That implies that you need to do at least thousands of two-sample t-tests to identify potentially DE genes. For example, the dataset analyzed in Storey and Tibshirani (2003) consists of $m = 3170$ genes.

Let $\mu_{i,1}$ and $\mu_{i,2}$ denote the population mean of the log base 2 intensity value for the i-th gene for BRCA1 and BRCA2, respectively. Recalling that the p-values of the test are uniformly distributed under $H_{0,i}\colon \mu_{i,1} = \mu_{i,2}$, $i = 1, \ldots, 3170$, these 3170 p-values should form a standard uniform distribution if none of the genes are DE (i.e., 'uninteresting'). In other words, the histogram (with density on the x-axis) should look very much like the pdf of Uniform$(0, 1)$.

However, that cannot be expected because it is likely that at least some of the genes are DE. Thus, it is natural to assume that we have a mixture of distributions for the p-values, where most of them are from Uniform$(0, 1)$ ('uninteresting'), but some are from another distribution (say, F) ('interesting'). That is, if we let X be a randomly chosen p-value from these tests, then

$$X \sim \begin{cases} \text{Uniform}(0,1) & \text{with probability } \pi_0; \\ F & \text{with probability } 1 - \pi_0, \end{cases}$$

for some $\pi_0 \in (0, 1)$, and that F is an unknown distribution for the 'interesting' genes. Here, π_0 means the proportion of non-DE ('uninteresting') genes.

There are several typical research goals for (bio)statisticians, which are listed below:

1. Choosing a suitable test statistic.
2. Choosing an appropriate method (resampling (if so, what resampling?) or non-resampling (if so, what distribution do we assume under $H_{0,i}$?)).
3. Estimating π_0.
4. Determining F.
5. Finding potentially DE genes with an appropriate error rate.

You will be asked to address most of these questions except (4) in this homework. These potentially DE genes are then typically analyzed by biologists.

1. Monte Carlo Simulation: Although the distribution of the log base 2 intensity value is unknown, several papers assume a t-type distribution or a Laplace distribution. It is also very common that this type of study has only several subjects in each group. Thus, we investigate the robustness of different types of t-test under various sample sizes and distributions. Come up with four `R` functions (`ttest1(x,y)`, `ttest2(x,y)`, `ttest3(x,y)`, `ttest4(x,y)`) that return the p-value for the following four types of two-sample t-test.

   (i) The traditional two-sample t-test that assumes equal variance (`ttest1(x,y)`);

   (ii) The traditional two-sample t-test that does not assume equal variance (`ttest2(x,y)`);

   (iii) The permutation-based two-sample t-test that assumes equal variance (`ttest3(x,y)`);

   (iv) The permutation-based two-sample t-test that does not assume equal variance (`ttest4(x,y)`). Note that this might not make sense as permutations are typically used under the assumption of equality of distributions. However, it can be justified *asymptotically* (i.e., large-sample size cases) that the permutation method works even when the assumption of equality of distributions is dropped.

   To specify the equal variance assumption, set `var.equal = TRUE` in the `t.test()` function.

   Using the idea presented in the Chapter 5 notes for the traditional versions and in the Chapter 7 notes for the permutation-based versions, report the empirical Type I error rates of the four tests in each of the following cases at $\alpha = 0.05$ with at least 1000 Monte Carlo simulations. Then, decide which test seems to be best in terms of robustness. For the permutation-based t-tests, use 1000 permutations. Also, if any of the test(s) did not perform well, provide a reason why.

   (a) $X_j \sim t(6)$, $j = 1, \ldots, 7$, and $Y_j \sim t(7)$, $j = 1, \ldots, 8$.

   (b) $X_j \sim t(4)$, $j = 1, \ldots, 7$, and $Y_j \sim t(20)$, $j = 1, \ldots, 8$.

   (c) $X_j \sim \text{Laplace}(0, 0.5)$, $j = 1, \ldots, 7$, and $Y_j \sim \text{Laplace}(0, 2)$, $j = 1, \ldots, 8$.

   (d) $X_j \sim t(6)$, $j = 1, \ldots, 7$, and $Y_j \sim \text{Laplace}(0, 0.5)$, $j = 1, \ldots, 8$.

   Note: Running simulations may take at least a couple of hours, so it is best to start early, and do a preliminary run with a smaller number of simulations first to ensure that there is no error in your code. Also, even though the problem says "at least 1000", it is suggested to do at least 10000 Monte Carlo simulations.

2. Calculation of p-values: The dataset `brcadata2.txt` (which has a header) contains $m = 3170$ rows and 15 columns. The first 7 columns are for the BRCA1 subjects and the other 8 columns are for the BRCA2 subjects.

   (a) Calculate the p-value for each of the $m = 3170$ rows using the best t-test statistic in terms of robustness. Store these p-values in a vector called `pvec`.

   (b) Plot the 3170 p-values using the `hist()` function. Make sure to set `freq=FALSE` and `breaks=seq(0,1,1/14)`.

   (c) Referring to (b), estimate (by eye) the minimum value of the x-axis so that the density becomes flat after that value.

   (d) Let λ (`lambda` in the `R` code below) denote the value you found in (c). Check whether or not your choice in (c) is reasonable by reporting the result of the (one-sample) Kolmogorov-Smirnov test. Specifically, use
   ```
   p.null <- sort(pvec)[(sum(pvec < lambda)+1):m]
   ks.test(p.null, "punif", lambda, 1)
   ```
   which tests to see if `p.null`, which is the vector of p-values beyond λ, is Uniform$(\lambda, 1)$ distributed. See the help file of `ks.test()` for more details if necessary.
   Note: The Kolmogorov-Smirnov test is sensitive to an outlier. Thus, a small p-value of the Kolmogorov-Smirnov may be somewhat deceiving sometimes. For this type of analysis, what is more important is to rely on visual tools such as histogram. That said, a large p-value from the Kolmogorov-Smirnov test would imply no statistically significant evidence against the distribution specified under the null hypothesis (i.e., Uniform$(\lambda, 1)$).

   (e) Estimate π_0 by using the formula $\hat{\pi}_0 = \sum_{i=1}^{m} I(p_i > \lambda)/[m(1-\lambda)]$, where p_i is the p-value for the i-th gene, $i = 1, 2, \ldots, m$. Using that, state the estimated proportion of genes that are potentially DE.

3. Even though we have successfully estimated the proportion of potentially DE genes in 2(e), we may still have a large number of potentially DE genes. Having too many candidate genes would be problematic for biologists. In addition, statisticians have realized that the p-values are not necessarily appropriate for identifying potentially DE genes.

   Therefore, they have come up with q-values, which can be thought of as the "opposite" of the p-values. In particular, while the p-value is used to express the minimum Type I error rate (i.e., the probability of rejecting H_0 when H_0 is true,

and is also known as the false positive rate) of rejecting H_0, the q-value calculates the minimum false discovery rate (FDR) for calling the result statistically significant. Here, the FDR is the expected proportion of false positives (FPs) among the ones called statistically significant (True positives (TPs) + False positives (FPs)). Just like p-values, smaller q-values indicate higher statistical significance, although its interpretation is different than that of p-values.

(a) Watch the video `https://www.youtube.com/watch?v=4AytJuNkeSM` to understand the basic idea behind FDR. (No need to report.)

(b) One way of converting p-values into q-values is to use the method of Benjamini and Hochberg (1995), often abbreviated as BH. Using the `p.adjust()` function, convert the vector of p-values (`pvec`) into a vector of q-values. Make sure to set `method="fdr"`. (No need to report.)

(c) Using 0.1 as the significance level for q-values, calculate the number of genes that are potentially DE.