UCLA Department of Statistics
STATS M254 Homework 1

Instructor: Jingyi Jessica Li
Due date: Wednesday, Feb 5, 2025 at 11:59 pm on BruinLearn

You are recommended to use R to answer numerical and graphical questions.
Please submit your codes along with your answers.

Problem 1: Multiple Testing (20 pts)

We will use the `NCI60` cancer cell line microarray data, which consist of $6,830$ gene expression measurements on 64 cancer cell lines. You need to install the `ISLR` package in `R` by

```
> install.packages("ISLR")
```

Then you load the package into `R` using

```
> library(ISLR)
```

Then you have the object `NCI60` in your `R` workspace. The object is a list of two elements `NCI60$data` and `NCI60$labs`, where `NCI60$data` is a numeric matrix of 64 rows (i.e., cancer cell lines) and $6,830$ columns (i.e., genes) and `NCI60$labs` is a 64-element character vector containing the cancer types of the cell lines. For example, you can explore the data using the following commands.

```
> dim(NCI60$data)
> head(NCI60$labs)
> table(NCI60$labs)
```

For every gene in the data set, perform a two-sample t-test with a two-sided alternative hypothesis to check if the gene is differentially expressed between the `NSCLC` cell lines and the `RENAL` cell lines. You can use the `R` function `t.test()`. This will result in $6,830$ tests in total.

1. Order the $6,830$ p-values from the smallest to the largest. Plot the ordered p-values as the y-axis and their corresponding ranks as the x-axis. If all the $6,830$ null hypotheses are true, the p-values should be uniformly distributed. Are the observed p-values likely from a uniform distribution?

2. What will be the p-value cutoff if you would like to control the Family Wise Error Rate under 0.05 using the Bonferroni correction? How many genes will be identified as differentially expressed using this cutoff?

3. What will be the p-value cutoff if you would like to control the False Discovery Rate under 0.05 using the Benjamini-Hochberg procedure? How many genes will be identified as differentially expressed using this cutoff?

4. What will be the p-value cutoff if you would like to control the Per-Comparison Error Rate (i.e., significance level) under 0.05? How many genes will be identified as differentially expressed using this cutoff?

Problem 2 (15 pts)

Use the `NCI60` data, give an example to each of the following data analysis tasks (please refer to the Leek and Peng, Science (2015) article we studied in our first lecture). Note: this is an open question; different people may give difference examples to the same task.

1. Descriptive study
2. Exploratory study
3. Inferential study
4. Predictive study
5. Causal study
6. Mechanistic study

Problem 3 (20 pts)

We will use a data set `babies.txt` to demonstrate the use of the permutation test. You may use the following R commands to read the data set into R:

```
> babies <- read.table("babies.txt", header=TRUE)
> bwt.nonsmoke <- subset(babies, smoke==0)$bwt
> bwt.smoke <- subset(babies, smoke==1)$bwt
```

The two R objects `bwt.nonsmoke` and `bwt.smoke` contain the rows that correspond to the weights of the babies with nonsmoking and smoking mothers respectively.

1. We will generate the following statistics based on a sample size of 10 and observe the following difference:

   ```
   > n <- 10
   > set.seed(1)
   > nonsmokers <- sample(bwt.nonsmoke , n)
   > smokers <- sample(bwt.smoke , n)
   > diff <- mean(smokers) - mean(nonsmokers)
   ```

   The question is whether this observed difference is statistically significant. We do not want to rely on the assumptions needed for the t test, so instead we will use permutations. We will reshuffle the data and recompute the mean. We can create one permuted sample with the following code:

   ```
   > dat <- c(smokers, nonsmokers)
   > shuffle <- sample(dat)
   > smokers_star <- shuffle[1:n]
   > nonsmokers_star <- shuffle[(n+1):(2*n)]
   > diff_star <- mean(smokers_star)-mean(nonsmokers_star)
   ```

The last value is one observation from the null distribution we will conduct. Set the seed at 1, and then repeat the permutation for 1,000 times to create a null distribution. What is the permutation derived p-value for our observation `diff`?

2. Repeat the above exercise, but instead of the differences in mean, consider the differences in median.

   ```
   > diff_med <- median(smokers) - median(nonsmokers)
   ```

   What is the permutation based p-value?

Problem 4 (10 pts)

We can use the same data set `babies.txt` to demonstrate the use of t test. For a sample with sample size $n = 10$ like in Problem 3, the Central Limit Theorem does not apply, and we need to check whether the data are approximately Gaussian under each condition. The following R codes can be used to check the Gaussian assumption:

```
> qqnorm(bwt.nonsmoke)
> qqline(bwt.nonsmoke,col=2)

> qqnorm(bwt.smoke)
> qqline(bwt.smoke,col=2)
```

1. Please display the Q-Q plots and argue whether the Gaussian assumption is reasonable for this data set.

2. Perform the t test using the following R codes. Compare the resulting p-value with what you obtain from the permutation test in Problem 3. Do you reach the same conclusion?

   ```
   > result <- t.test(nonsmokers, smokers)
   > result$p.value
   ```

Problem 5 (10 pts)

A test for cystic fibrosis has an accuracy of 99%. Specifically, we mean that:

$$P(+|D) = 0.99$$

and

$$P(-|ND) = 0.99,$$

where + and − stand for positive and negative test results respectively, and D and ND represent the existence and nonexistence of the disease respectively.

The cystic fibrosis rate in the general population is 1 in 3,900, that is

$$P(D) = 0.00025.$$

If we select a random person and they test positive, what is probability that the person has the disease?

Hint: Use the Bayes Theorem.

Note: The posterior mean you derive in this problem is often called the "maximum a posteriori" (MAP) estimator of μ.

Hint: Please follow my derivation for the Bayesian estimator of σ^2 in the t-test setting.

Problem 6 (10 pts)

This problem is to help you understand what we mean by a random sample, and why we say that the sample mean is a random variable.

Use the following R code to simulate a sample with sample size $n = 10$ and calculate the sample mean:

```
n <- 10
x <- rnorm(n)
mean(x)
```

Repeat the simulation for one million times, plot the distribution of the one million sample means.

Now increase n to 100. Repeat the above simulation, how does the distribution of the one million sample means change?

Now further increase n to 1000. Repeat the above simulation, how does the distribution of the one million sample means change?

What do we conclude about the distribution of the sample mean?