Assignment 1 — Due September 29 at 4pm

Homework Assignment Policy and Guidelines

1. Homework assignments should be well organized and reasonably neat. It is required that you show your work in order to receive credit.
2. You may discuss most homework problems with others including your peers, your TA, and instructor, but you must write up your homework solutions by yourself in order to receive credit. Similarly, you must write your own computer code and obtain your computer output independently.
3. Homework assignments will be submitted online. Credit will not be given for homework turned in late.
4. You are expected to complete some homework problems independently (or in a group) without discussing with other students (or other groups) before they are due. You may seek clarification of these problems from your TA and instructor. Such homework problems will be marked `independent work`.
5. For problems that involve data analysis, always interpret the results in the context of the study. This may include comments on whether the results are meaningful or not.
6. For problems that involve technical writing, strive to be clear, concise, and cogent. Organize figures and tables in an efficient manner while maintaining clarity.
7. Unless otherwise stated in a problem, keep the R code and/or R output for all the relevant problems in a **well organized appendix**. Please streamline and briefly document the R code to be clear and concise.
8. Refrain from copying and pasting R code and/or R output directly into your answers. Think of your answers as a mini technical report of your findings and follow the guidelines on technical writing.
9. For hypothesis testing problems, provide all the ingredients of hypothesis testing. That is, define the population parameter(s) of interest in symbols and in words, state the null and alternative hypotheses, give the test statistic and the null distribution, computer the observed test statistic and p-value (with helpful drawing), draw a conclusion, and interpret the results in the context of the study.
10. For confidence interval problems, define the population parameter(s) of interest in symbols and in words, give the confidence interval formula, provide the individual terms in the formula (e.g., confidence level, point estimate, critical values), and the confidence interval. Interpret the confidence intervals in the context of the study.

Homework Problems:

1. There are different ways to estimate the distributions parameters. For example, the three common methods of i- Method of Moments, ii- Least Squares and iii- Maximum Likelihood. Briefly discuss the three methods mentioning pros and cons of each method. Finally, find the MLE of the mean and the variance of the univariate Gaussian distribution.

2. Computer simulations provide a powerful set of tools for studying various statistical ideas. In particular, simulations can be used to study various statistical methods, especially when mathematical or theoretical approaches are not available. We will follow a similar format to the lake clarity example in class and start by assuming that we know the population. We then take a random sample from that population, and calculate something for that **sample**, like the mean, sample variance. Of course, the result of this is random, because it is based on a random sample. To study things in greater generality, we repeat the sampling many times. Write and submit your own `R` code for this problem.

(a) Sample $n = 10$ independent observations from $N(2, 4)$; that is $\mu = 2$ and $\sigma = 2$. Repeat this sampling $S = 100$ times, so at the end we will have 100 samples, each with 10 observations, and for each sample, compute the sample mean and sample variance. Provide the histograms of sample mean and sample variance from the 100 samples.

(b) Repeat (a) but this time let $n = 40$.

(c) Repeat (a) but this time let $n = 160$.

(d) What observations can you make about the histograms in (a)–(c)?

(e) Repeat (a)–(d), but this time, replace $N(2, 4)$ with $B(10, 0.2)$; that is $m = 10$ and $\pi = 0.2$.

(f) Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ denote the sample average and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ be the sample variance, where Y_i's are i.i.d. draws from $N(\mu, \sigma)$. Find $\mathbb{E}(\bar{Y})$, $\mathrm{Var}(\bar{Y})$, and $\mathbb{E}(S^2)$. Compare them with your simulations. What conclusion can you draw? In particular, explain the difference between $\mathrm{Var}(\bar{Y})$ and $\mathrm{Var}(Y)$.

(g) Repeat (f), but this time, replace $N(\mu, \sigma)$ with $B(m, \pi)$.

3. The file 'Spanish.csv' gives the pretest and posttest scores on the MLA listening test in Spanish for 20 executives who received intensive training in Spanish.

(a) We hope to show that the training improves listening skills. State an appropriate H_0 and H_a. Describe in words the parameter(s) that appear in your hypothesis.

(b) Make a graphical check for outliers or strong skewness in the data that you will use in your statistical test, and report your conclusions on the validity of the test.

(c) Carry out a test. Draw your conclusion in the context of the study.

(d) Give a 90% confidence interval for the mean increase in listening score due to the intensive training.

(e) Let D_i denote the difference between the pretest score and posttest score for i-th individual, where $D_i \sim_{i.i.d.} N(\mu, 1)$ for all $i = 1, \ldots, n$. Suppose the true effect size is $\mu = 2$. Now we want to perform the test $H_0: \mu = 0$ vs. $H_a: \mu \neq 0$, with significance level $\alpha = 0.05$ and power $1 - \beta = 0.80$. What is the minimum sample size n needed?
Hint: Let $z_{\alpha/2}$ be the critical value for the test statistics $T = T(n)$ with significance level $\alpha = 0.05$. Find minimum n such that

$$\mathbb{P}(T(n) \in [-z_{\alpha/2}, z_{\alpha/2}] | D_i \sim N(2, 1), \text{ i.i.d. for } i = 1, \ldots, n) \leq \beta.$$

4. Below are measurements on stem volume (in cubic centimeters) of 2-year-old seedlings. One group was propagated from virus-infected buds whereas the other was propagated from healthy buds.

Healthy			Infected				
1870	1324	1446	1121	408	184	16	741
1325	1759	1652	170	991	711	734	202
1364	1515	1065	893	742	335	444	

(a) Present appropriate summary plots of these data and also summarize them using the sample median, mean, and standard deviation. What tentative conclusions might be drawn about the effect of the virus?

(b) (Hand calculation with calculator) Perform an appropriate T test to determine whether there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is smaller than those propagated from healthy buds. Identify the ingredients of hypothesis testing and draw conclusions in the context of the study.

(c) (Hand calculation with calculator) Construct a 95% confidence interval for the difference of the mean stem volume of 2-year-old seedlings between the two groups.

(d) Perform an appropriate T test to determine whether there is evidence that the mean stem volume of 2-year-old seedlings propagated from virus-infected buds is *different from* those propagated from healthy buds.

(e) What assumptions are made for parts (b), (c), and (d)? Assess the assumptions by using suitable histograms, box plots, QQ plots, and Levene's test. How reasonable are the assumptions? What remedial measures are desirable, if any?

(f) Perform a Welch's T test and construct the corresponding 95% confidence interval for the difference of the mean stem volume of 2-year-old seedlings between the two groups.

(g) Perform a suitable randomization test.

(h) Perform a suitable nonparametric test.

(i) Compare the results obtained from the tests in parts (d), (f), (g), and (h).

(j) Suppose this dataset is to be used to inform the design of a new study in the future. What recommendation would you make about the sample sizes n_1 and n_2? Provide reasoning. Assume $\alpha = 0.05$ and power $= 0.80$ although the researcher is not sure what the smallest scientifically significant mean difference between the two groups should be.

5. An experiment was conducted to examine the effectiveness of a biological control for reducing damage of corn (maize) by the European corn borer. In the experiment, 8 plots were identified in a large field of alfalfa. The plots were planted to corn, and each plot was divided into two equal subplots. Within each plot, one of the two subplots was randomly assigned to be treated with the biological control; the other subplot was assigned a standard chemical treatment. The data (# of moths captured in a peak week) are shown below:

Plot	1	2	3	4	5	6	7	8
Biological control	37	42	12	32	97	68	14	29
Chemical control	41	45	18	106	107	171	12	30

(a) Present a useful display (or displays) of these data and also summarize them by useful summary statistics. What tentative conclusions might be drawn about the biological control versus the standard chemical treatment?

(b) Perform an appropriate T test to determine whether there is evidence that the means are different between the two types of control. Identify the ingredients of hypothesis testing and draw conclusions in the context of the study.

(c) Construct a 95% confidence interval for the difference of the means between the two types of control.

(d) What assumptions are made for parts (b) and (c)? Assess the assumptions by using suitable histograms, box plots, and/or QQ plots. How reasonable are the assumptions? What remedial measures are desirable, if any?

(e) Perform an appropriate transformation and then an appropriate T test.

(f) Perform a suitable randomization test.

(g) Perform a suitable nonparametric test.

(h) Compare the results obtained from all the tests above.

6. Develop a "cool" product that *effectively* explains

   (a) hypothesis testing

   (b) confidence interval

   (c) central limit theorem.

   to a lay audience using words and/or visuals. The format of the product can be of an article or a report. Be creative. Describe the types of data that are available and relevant, as well as possible statistical techniques that can be employed to analyze such data and draw conclusions. You are also encouraged to pair up with another student (preferably with a different background) to work on this problem together but please document briefly the role each individual has played in the creation of this product.

   Pick a dataset on your own and explain the above concepts in the context of your data. Possible data source:

   - UCI Machine learning repository: https://archive.ics.uci.edu/ml/index.php
   - Kaggle datasets: https://www.kaggle.com/datasets

   Of course, feel free to use other dataset that you are interested in and/or passionate about. The maximum length of your write-up for this problem is **1 page** (excluding code and figures). Have fun with your collaboration.

   `independent work`