STAT 231 Spring 2021 Assignment 5
Crowdmark Portion

Assignment 5 is due on **Wednesday July 21st at 2:00pm** EDT. Your assignment submission **must be typed** and submitted as a pdf. There are no exceptions. Any submitted answer which is not typed will not be marked and given a mark of zero, unless the question specifically indicates that part can be written by hand.

Upload your assignment to **Crowdmark as a pdf file for marking.** You can upload your assignment as one document or individually for each problem. If you upload one document then you must drag and drop the pages for each problem to the appropriate question as indicated in Crowdmark. You can resubmit your assignment any number of times before the due time. Therefore, to ensure that there are no issues with uploading we advise you to upload your assignment well in advance of the due time. **Assignments which are left as a single document and not uploaded to the appropriate places in Crowdmark will be assigned a 10% penalty.**

In addition to submitting in Crowdmark for marking, you must submit your pdf to the **Assignment 5 LEARN dropbox** to facilitate the running of your assignment through plagiarism detection software. You will also submit your R file to this dropbox.

A penalty of 10% per hour is applied for late assignments.

Checklist to complete for this assignment:

- ☐ Work on the crowdmark portion of the assignment throughout the week. Review the R Tutorial for Assignment 5 for assistance with R code.
- ☐ Upload the PDF of your assignment solutions to Crowdmark by the deadline.
- ☐ Upload **a single** PDF of your assignment solutions to the appropriate learn dropbox by the deadline.
- ☐ Upload the R file of your assignment to the appropriate learn dropbox by the deadline.
- ☐ Complete the Mobius portion of the assignment on Wednesday July 21st before the deadline.

Marks for Assignment 5:

Assignment Section	Total marks	% of Assignment 4 Grade
Crowdmark	60	60%
Learn Dropbox of solutions + R file	2	
Mobius	15	40%

Assignment 5 Crowdmark Intended Learning Outcomes

Below we have outlined the list of intended learning outcomes for this assignment. This assignment includes a reflective question at the end addressing these outcomes. Keep them in mind as you solve the various questions, and we recommend you try and keep track of how these learning outcomes are achieved by each of the questions to follow.
Enjoy 😊

- Fit and analyse a simple linear regression model.
- Interpret diagnostic plots and make recommendations on how model issues can be fixed.
- Perform two-sample testing for independent samples and make conclusions based on the observed data.

(30 marks) In **Question 1**, we will analyse the **Cost of Product** and **Weight in grams variates** from the E-commerce shipping dataset. The full dataset of 500 observations can be used in your analysis. The purpose of this question is to investigate the possible linear association that may exist between these two variates.

(a) (2 marks) In Assignment 1 Question 6 you were asked to create a scatter plot and solve for the correlation coefficient between Cost of Product and Weight in grams. Insert the scatter plot and correlation coefficient previously found in your assignment. In the context of regression analysis comment on whether the scatter plot and correlation coefficient suggest a linear association. Explain.

(b) (5 marks) Assume that a linear model was appropriate with Weight in grams as the explanatory variate x and Cost of Product as the response variate y. Use R to fit the simple linear regression model

$$Y_i \sim G(\alpha + \beta x_i, \sigma), i = 1,2, \dots, n \text{ independently}$$

where the $x_i's$ are assumed to be known constants to your data.
Use your fitted model to complete the following table based on the output from the R command `lm()`, where to get the actual model output we make use of the command `summary()`:

number of complete observations	
$\bar{x},\ \bar{y}$	
sample correlation coefficient (in a simple linear regression the correlation coefficient is the square root of the multiple R-squared value in the model output. The sign is determined by the sign of the slope parameter estimate)	
maximum likelihood estimate of the intercept α	

maximum likelihood estimate of the slope β	
equation of the fitted line: $y = \hat{\alpha} + \hat{\beta}x$	
unbiased estimate of σ	
estimate of the standard deviation of $\tilde{\beta}$, the estimator of the slope	

(c) (3 marks) The parameters β, $\mu(x) = \alpha + \beta x$, and σ correspond to what attributes of interest in the study population?

(d) (2 marks) Use the output from the R function `summary()` to test the hypothesis H_0: β = 0 (the hypothesis of no relationship). Your answer should be displayed in the following four step structure:

Step 1: State the hypothesis that you are testing (both null and alternate)
Step 2: Write out the formula of the test statistic being calculated and include the appropriate value from the R output.
Step 3: State the formula of the p-value being solved for and include the appropriate value from the R output
Step 4: Use your findings to make a conclusion statement (use Table 5.1 from the course notes)

(e) (8 marks) Insert the following plots in your assignment:

- scatterplot of your data with the fitted line superimposed.
- plot of the standardized residuals versus the explanatory variate.
- plot of the standardized residuals versus the fitted values.
- qqplot of the standardized residuals.

For each plot indicate which assumptions of the simple linear regression model are being checked, what you expect to see for each plot if the model assumptions hold, and what you observe for your data.

What is your conclusion regarding the fit of the simple linear regression model to your data? If the linear regression model is not a good fit, then what do these plots suggest is wrong with the model?

(f) (2 marks) Discuss what the results of your model fit tell you about the relationship between these two variates. Your answer should be phrased in 'real-world' terms, that is, discussing the actual variate names and not simply referring to algebraic notation. Your discussion should include an interpretation of $\hat{\beta}$. Are you surprised by your results? Why or why not?

For parts g) to j) refer to the R coding provided in Chapter 6 Exercise 12 on page 273.

(g) (2 marks) Use the output from the R command `confint()` to give a 95% confidence interval for the slope β. Only provide the output and give an interpretation of your confidence interval.

(h) (2 marks) Use the output from the R command `predict()` to give an estimate and a 90% confidence interval for the mean response $\mu(x) = \alpha + \beta x$ for x = 3000. Only provide the output and give an interpretation of your confidence interval.

(i) (2 marks) Use the output from the R command `predict()` to give an estimate and a 99% prediction interval for the predicted response for x = 3000. Give reasons for why this interval is wider than the interval found in part (h).

(j) (2 marks) Use R to obtain a 95% confidence interval for σ. Provide your R code and output, include an interpretation of your confidence interval.

(19 marks) Question 2 focuses on the **Discount offered** and **Gender** variates from the E-commerce shipping dataset. In this question you can recode blank entries as "NA" and then ask R to omit these when creating the qqplots and boxplots. This can also be done in the t.test commands. Some useful code is included below:

```
dataset[dataset==""]<-NA #code any blank spots as NA
boxplot(y~x, na.omit=TRUE) #plot boxplot ignoring NA entries
```

```
t.test(y~x, na.omit=TRUE) #perform t-test ignoring NA entries

# code to subset the data based on Gender ignoring the NA
entries

DiscountbyMale<-subset(datasetDD, Gender=="M", na.omit=TRUE)
```

The purpose of this question is to perform two sample testing for independent samples investigating whether there is a statistical difference in the average discount offered by customers' gender.

In our previous assignments and Midterm, we have best modeled the variate Discount offered as coming from an Exponential distribution. However, in performing a two sample test an underlying assumption is Normality. In the Chapter 6 course notes (pg 247-248) it mentions how often transformations on the variate can help adjust for heteroscedasticity issues. These transformations can also be used to adjust for Normality issues. As such for this question we will work with the **natural log of the variate Discount offered, i.e ln(Discount_Offered).**

(a) (4 marks) Use R to create a qqplot for the Discount Offered variate and the Log Discount Offered variate. Insert both qqplots in your assignment and comment on any differences and similarities you may notice between the two plots. Based on the qqplot can the variate Log Discount offered be more correctly modeled by the Gaussian distribution? Explain.

Let Y_{mi} be the log discount for the ith male customer, i = 1,2,…,n_m. Assume a $G(\mu_m, \sigma_m)$ model for the Y_{mi}'s. In addition let Y_{fi} be the log discount for the ith female customer, i = 1,2,…,n_f. Assume a $G(\mu_m, \sigma_m)$ model for the Y_{fi}'s.

(b) (3 marks) Insert side-by-side boxplots comparing the female log discount offered with the male log discount offered in your assignment. Based on the plot describe all the differences and similarities between these two data sets.

(c) (3 marks) Assuming that the Gaussian models are reasonable, determine a 95% confidence interval for σ_f and a 95% confidence interval for σ_m. Based on these intervals, is it reasonable to assume $\sigma_f = \sigma_m$? Give reasons for your answers.

(d) (2 marks) Explain what the null hypothesis H_o: $\mu_f = \mu_m$ means in real-world terms, that is, explain what this null hypothesis means in words that a non-statistician would understand.

(e) (3 marks) Assuming the variances of the two samples are not equal, i.e. $\sigma_f \neq \sigma_m$, use the output from the R command t.test to test the hypothesis H_o: $\mu_f = \mu_m$. Be sure to include the observed value of the test statistic,

$$D = \frac{\left|\bar{Y}_f - \bar{Y}_m\right|}{\sqrt{\frac{S_f^2}{n_f} + \frac{S_m^2}{n_m}}}$$

the degrees of freedom associated with the test statistic, the p-value, and the conclusion. Does your conclusion agree with what you observed in (b)?

Notes:

1. **R does not report the absolute value of the test statistic. If your test statistic is negative, be sure to report the absolute value.**
2. **As stated in the tutorial R uses the t-distribution to perform this test, where the exact degrees of freedom are based on a fair involved formula and are most often not integer valued. The course notes describe this method for large values of the sample size so that pivotal quantities are obtained from the standard Normal distribution. For purposes of this assignment please quote the values as given in the R output. You do not need to recalculate the p-value.**

(f) (2 marks) Give a 99% confidence interval for the difference $\mu_f - \mu_m$. Include a conclusion statement.

(g) (2 marks) Based on your findings what does the data suggest about the average discount offered to females and males? Explain.

(6 marks) Question 3 focuses on the **Weight in grams** and **mode of shipment** variates from the E-commerce shipping dataset. The full dataset of 500 observations can be used in your analysis. The purpose of this question is to perform two sample testing for independent samples investigating whether there is a statistical difference in average weight in grams by mode of shipment.

In our previous assignments and Midterm we have best modeled the variate Weight in grams as coming from a Gaussian distribution.

Let Y_{si} be the weight in grams for the ith product transported by ship, i = 1,2,…,n_s. Assume a G(μ_s, σ_s) model for the Y_{si}'s. Let Y_{ri} be the weight in grams for the ith product transported by road, i = 1,2,…,n_r. Assume a G(μ_r, σ_r) model for the Y_{ri}'s. In addition let Y_{fi} be the weight in grams for the ith product transported by flight, i = 1,2,…,n_f. Assume a G(μ_f, σ_f) model for the Y_{fi}'s.

(a) (3 marks) Assuming that the Gaussian models are reasonable, determine a 95% confidence interval for σ_s and a 95% confidence interval for σ_r. Based on these intervals, is it reasonable to assume $\sigma_s = \sigma_r$? Give reasons for your answers.

(b) (3 marks) Assuming the variances of the samples for different modes of transportation are equal, i.e. $\sigma = \sigma_s = \sigma_r$, use the output from the R command t.test to test the hypothesis H_o: $\mu_s = \mu_r$. Be sure to include the observed value of the test statistic,

$$D = \frac{|\bar{Y}_s - \bar{Y}_r|}{S_p\sqrt{\frac{1}{n_s} + \frac{1}{n_r}}}$$

the degrees of freedom associated with the test statistic, the p-value, and the conclusion.

Note: R does not report the absolute value of the test statistic. If your test statistic is negative, be sure to report the absolute value.

(5 marks) Question 4 is intended to help you check that you have identified and internalized the important things to learn from this assignment.

Write an approximately 1-2 paragraph reflection (in full sentences) that summarizes how you achieved (or not if you're still not confident with them) the intended learning outcomes by completing this assignment.