STAT 1221/2220
Introductory Statistics Review Name: Stephen Sladco

The following problems are designed to review the most important content from your introductory statistics course. As a result, this assignment will only be graded for completion. It is extremely important that you understand each question below because these concepts will be used extensively, especially in the first half of the semester. If you get stuck on any part, the R code is provided in a separate document on Canvas. The solutions have also been posted so you can check your work as you go. If you have any questions, please don't hesitate to reach out either via email or during office hours.

An R markdown file titled **intro_stats_review.Rmd** is posted on Canvas with a code chunk for each part you should submit for Problem 3. (You don't need to submit any code for Problems 1 and 2.) Problem 3 will walk you through the process of setting up your workspace in R, writing the code, and knitting it to be submitted. When you are done with the assignment, upload a PDF of the problem set and the knitted HTML file with your R code.

1. Many hypothesis tests we will run in this course will not provide a p-value that directly corresponds to the test. Instead, we will need to calculate the p-value from the test statistic using R and use that value to come to a decision. Use R to calculate the following p-values to four decimal places given the test statistic. Then decide if the null hypothesis would be rejected or not according to the level of significance.

Alternative	Level of Sig.	Test Stat.	DF	P-Value	Decision
Two-sided	5%	$Z = -1.87$		0.0615	fail to reject H_0 (0.0615 > 0.05)
Lower one-sided	5%	$t = -2.15$	33	0.0195	Reject H_0 (0.0195 < 0.05)
Upper one-sided	1%	$t = 1.94$	21	0.0330	Fail to reject H_0 (0.0330 > 0.01)
Two-sided	10%	$t = 1.79$	14	0.0951	Reject H_0 (0.0951 < 0.10)
Upper one-sided	5%	$F = 3.38$	1 and 17	0.0835	fail to reject H_0 (0.0835 > 0.005)
Upper one-sided	1%	$F = 9.45$	2 and 28	0.0007	reject H_0 (0.0007 < 0.01)

2. It is also important to understand how to work with critical values for both confidence intervals and hypothesis tests. A majority of the problems in this course will require the use of critical values from the standard normal distribution, t-distribution, and F-distribution. Use R to report the following critical values to three decimal places based on the level of significance and form of the alternative hypothesis. Then decide if the null hypothesis would be rejected or not according to the test statistic.

Alternative	Level of Sig.	Test Stat.	DF	Critical Value(s)	Decision
Two-sided	10%	$Z = -1.42$		± 1.645	fail to Reject H_0 (1.42 not extreme enough)
Upper one-sided	5%	$t = 1.56$	37	1.687	fail to Reject H_0 (1.56 < 1.687)
Lower one-sided	1%	$t = 2.84$	24	-2.492	Reject H_0 ($-2.84 < -2.492$)
Two-sided	5%	$t = -3.19$	15	± 2.131	Reject H_0 ($-3.19 < -2.131$)
Upper one-sided	5%	$F = 8.06$	2 and 34	3.276	Reject H_0 (8.06 > 3.276)
Upper one-sided	1%	$F = 4.23$	3 and 29	4.538	fail to Reject H_0 (4.23 < 4.358)

3. Last year, the mean monthly energy cost for families nationally in the United States was $200. An economist takes a random sample of families in Pittsburgh and records their energy costs for the previous month. The data is contained in the file **energy_cost.csv**.

(a) If you have not done so, I recommend you create a directory that can house all of your assignments for this course. In the **Documents** folder on your computer, right click and create a new folder. Title this folder **STAT 1221 Homework** (or something similar and recognizable). This will be the folder where you save all of your assignments, code, and datasets to keep everything organized.

(b) Log into Canvas and download the two files under **Introductory Statistics Review**. The first is the R markdown file that contains the framework for writing your code, titled **intro_stats_review.Rmd**. The second is the dataset for the assignment, titled **energy_cost.csv**. When you download these files, they will appear in the **Downloads** folder on your computer. Copy and paste both of these files into the **STAT 1221 Homework** folder you created in part (a).

(c) Double click on the R markdown file **intro_stats_review.Rmd** that now resides in the **STAT 1221 Homework** folder. This will open up the code in RStudio for you to edit. (Do not open the one in your **Downloads** folder. You should always work from the file in the working directory, which is your **STAT 1221 Homework** folder.

(d) In RStudio, move your cursor to the top of the screen. Click on **Session** > **Set Working Directory** > **Choose Directory**. This will bring up a new window. You need to tell RStudio where the files are that you want to use. These files are stored in the same **STAT 1221 Homework** folder created in part (a). Select this folder to set it as your working directory if RStudio has not already identified it.

(e) At this point, you should be ready to read in the dataset. In the gray box under the heading "Read in data", type **energy = read.table("energy_cost.csv", header = TRUE, sep = ",")**. Click the "Run" button (the sideways triangle) on the right side of this box. If everything worked correctly, you should now see a new row on the far right side of your screen that says **energy** under data. This is the data frame from which you will work. (If you get an error message that says "cannot open the connection", then either your working directory in part (d) is not set correctly or the files are not in the working directory. Understanding how to get these lined up is the most challenging part so please ask for help if need be.)

(f) Describe the population and parameter.

Population: All the energy cost for families nationally in the USA ; There are two observations outside the bounds $64 and $274

Parameter: The mean Monthly energy cost of all families nationally in the USA.

(g) In the gray box under "Problem 3g", create a histogram of the data in R. Put a label on the x-axis with "Monthly Energy Bill" as the label.

(h) In the gray box under "Problem 3h", create a boxplot of the data in R. Put a label on the x-axis with "Monthly Energy Bill" as the label.

(i) In the gray box under "Problem 3i", obtain the following summary statistics using R: sample mean, sample standard deviation, sample size, and five number summary. Report the values in the table below. Round decimals to two decimal places.

Statistic	Mean	SD	Sample Size	Five Number Summary
Energy Cost	174.07	49.50	30	64, 147.50, 181, 203, 274

(j) Describe the symmetry and modality of the histogram.

The histogram is unimodal with a peak between $150 and $200, and is symmetric. This results in the histogram being approx. normal.

(k) Describe the data using the boxplot, including if there are any outliers.

The Boxplot is closely symmetric at both whiskers with about the same height, but the lower half of the boxplot is a little wider compared to the upper half. There are no outliers.

(l) Observations that are more than two standard deviations away from the mean are noteworthy and often deserve closer attention. Calculate the values that are two standard deviations above and below the mean.

$$\text{Stat} \pm CV \cdot SE$$

$$174.07 \pm 2 \cdot 49.50 = (75.07, 273.07)$$

(m) Open the data frame in R by clicking on **energy** in the top, right section of RStudio. The data should show up in the top, left section of the screen. Click on the column titled **cost** to sort the data from smallest to largest. Report any values in the data that go outside the bounds from part (l).

Two are outside bounds: $64 and $274

(n) In the gray box under "Problem 3n", obtain the critical value for finding a 95% confidence interval for the mean energy cost for families in Pittsburgh.

(o) Calculate a 95% confidence interval for the mean monthly energy cost for families in Pittsburgh. Show the calculation.

$$174.07 \pm 2.045\left(\frac{49.50}{\sqrt{30}}\right) = 174.07 \pm 18.48$$

$$= (155.59, 192.55)$$

(p) Interpret the interval from part (o) above.

We are 95% confident that the true population mean monthly energy cost for families in Pittsburgh is between $155.59 and 192.55.

(q) Test if the mean monthly energy cost for families in Pittsburgh is significantly less than the $200 national average using a 5% level of significance. Type the code for the critical value and p-value in the gray box under "Problem 3q" in the R markdown file.

i. State the hypotheses.

H_0: $\mu = 200$ vs. H_A: $\mu < 200$

ii. Calculate the test statistic. Show the calculation.

$$\frac{174.07 - 200}{\frac{49.50}{\sqrt{30}}} = -2.87$$

iii. Use R to obtain and report the critical value. 1.699

iv. Use R to obtain and report the p-value. 0.038

v. Write a conclusion in the context of the problem.

Reject H_0 and will conclude that the true energy cost for families each month is significantly less than $200

(r) Are the confidence interval and hypothesis test from above consistent? (i.e. Do you come to the same conclusion regarding the mean energy cost being $200 using the confidence interval and hypothesis test?) Justify your answer.

Yes. The hypothesis test concludes that the true mean is less than $200, and the confidence interval is below $200, meaning that the true mean is lower.

(s) Once you are done with the assignment, the final step is to knit your R file to HTML. Knitting is the process of taking your code and turning it into a document. To do this, find the button that says either **Preview** or **Knit** in the toolbar near the top of the window. (The one you see may depend on your computer, but they both do the same thing.) Click the arrow next to this button to reveal a dropdown menu. Finally, click **Knit to HTML**. A pop-up window may appear with RStudio telling you that several packages must be installed. Click **Yes** and allow RStudio to install these packages. Once that is complete, RStudio will knit your code to an HTML file called **intro_stats_review.html**, which you can find in your **STAT 1221 Homework** folder created in part (a).