Week 5 Practice Exam (#2)

Name: \_\_\_\_Quinn Knudsen\_\_\_\_\_\_\_\_\_

Instructions: This is a “low stakes” (i.e., not graded) learning assessment of your comprehension of the first four weeks of this course*.* Compose brief answers to each of the following questions, typing your response in *italics* below each question.

1. Your boss at the social media marketing company asks you to conduct another A/B test on two different social media ad configurations. Each of the two ads is displayed on n=96 high traffic social media pages:   
     
   The A banner gets an average of 1373 clicks per hour.   
   The B banner gets an average of 1394 clicks per hour.   
     
   The 95% confidence interval is as follows:   
    -23 < (mean difference between A and B) < 17.   
     
   Answer the following questions about that confidence interval:   
   1. Does this particular confidence interval contain the population mean difference?  
       *We can estimate with a 95% confidence interval the population mean difference based upon these sample statistics (unfortunately without certainty on whether it is positive or negative… A or B larger) but not necessarily contained in this sample.*
   2. Which banner ad do you prefer (A or B) and why?   
       *I cannot be certain because the 95% confidence interval crosses the zero plane. While the means are different, we are not going to have a p value <.05 to reject the null hypothesis that the values are equal. The standard deviations (variance) impact this conclusion.*
   3. Your boss tells you to run the same experiment 99 more times, calculating a new confidence interval each time. Now you have a collection of 100 confidence intervals, each of which was constructed in the same way, but from new data samples: What can you say about this collection of confidence intervals?

*These confidence intervals will be more accurate than if I had just one confidence interval to base my results off of. With greater sample size I will minimize the ill effects of outliers and sampling deficiencies. Sample size in this instance is our friend. However, if the 95% confidence interval still ranges, on average, from a negative value to a positive value, then we cannot reject the null. Means of means will form a normal distribution converging in theory around the true population mean, but that does not help us reject the null. 95 of these should contain the population mean.*

* 1. Which command in R would you use to produce the confidence interval for each of the 100 that you constructed?  
     *I would use t.test() to get a value for the confidence interval for each of the 100 I constructed. If we calculate the standard deviation then we can run a sample like below using the frequentist approach.*

*Toy example:*

*t.test(rnorm(100,mean=1373,sd=3.8),rnorm(100000,mean=1394,sd=33.8))*

1. Some output appears from a t-test that compared annual U.K. driver fatalities for several years before and after a seat belt law was enacted. Interpret these results in a brief paragraph, making sure to explain as much of the statistical output as you can:  
     
   **Welch Two Sample t-test**

**data: FatalitiesPreLaw and FatalitiesPostLaw**

**t = 5.1253, df = 29.609, p-value = 1.693e-05**

**alternative hypothesis: true difference in means is not equal to 0**

**95 percent confidence interval:**

**15.39892 35.81899**

**sample estimates:**

**mean of x mean of y**

**125.8698 100.2609** *Based upon these results, we can reject the null hypothesis that fatalities pre law (125.9) and fatalities post law (100.3) are equal (p<.001). We see here that the mean of x 126 is greater than the mean of y 100. The 95% confidence interval for the difference in means ranges from 15 to 36. We have 29.6 degrees of freedom and a t statistic of 5.13 (corresponding with the associated p value provided). This provides support for the alternative hypothesis that the two groups are not equal.*

1. Explain the following diagram, which was created from these five lines of code:  
     
   **x <- seq(from=-3,to=3,by=.1)  
   plot(x, dt(x,df=30))  
   abline(v=-2.04)  
   abline(v=2.04)  
   abline(v=2.5,col="green")**  
     
   Hint: dt() is the probability density function of the t-distribution, so the total area under the curve equals 1. The -0.025 quantile for t, with 30 degrees of freedom, is -2.04. Make sure to explain what the green line might represent and the consequences of its position on the extreme right of the diagram.

*X is a vector of values ranging from -3 to 3 with an increasing increment of .1. Based upon the green line in this diagram we would likely reject the null as it falls beyond the 97.5 quantile located here at +2.5 sd from the mean of this normal distribution. In this instance, the green value would be grounds to reject the null (p<.05). Using the -0.025 quantile for t, with 30 degrees of freedom, is -2.04. and the distribution, the .975 quantile is .025 0r 2.04. With 30 degrees of freedom the 99% would fall at 2.457. This adds support to our claim and means this line may fall at p<.01.*

*Source:* [*http://homepage.cs.uiowa.edu/~jblang/probability.calculators/t.table.htm*](http://homepage.cs.uiowa.edu/~jblang/probability.calculators/t.table.htm)

*4.* Imagine that you just ran the following R code:

X1 <- c(32, 48, 23, 23, 23, 21, 28)

X2 <- c(51, 32, 33, 50, 26, 66, 27)

df <- data.frame(mpg=X1, wt=X2)

1. Fill in the data table below so that it resembles what you would see as a result of running the R-Studio command “View(df)”. Make sure to fill in the column labels!

|  |  |  |
| --- | --- | --- |
| *Observation* | *mpg* | *wt* |
| *1* | *32* | *51* |
| *2* | *48* | *32* |
| *3* | *23* | *33* |
| *4* | *23* | *50* |
| *5* | *23* | *26* |
| *6* | *21* | *66* |
| *7* | *28* | *27* |

1. Next, review the R code in each of these boxes and write in the box what you would see if you ran that code at the console after creating df with the code above. There is no need for a calculator for any of these items. There are no “trick” questions. All commands below run without error.

|  |  |  |  |
| --- | --- | --- | --- |
| length(X1)  7 | length(df$mpg)  7 | median(X1)  23 | max(X2)  66 |
| length(df$mpg)== length(df$wt)  TRUE | min(df$wt)  26 | X2[1]  51 | df$wt[1]  51 |

1. In the data set for the previous question, would it or would it not make sense to run a t-test comparing df$mpg and df$wt? Briefly explain why.

*This would not make sense because they are separate variables.*