Pre-Midterm Practice Exam ()

Name: \_Abhijith Anil Vamadev\_\_\_

Instructions: This is a “low stakes” (i.e., not graded) learning assessment of your comprehension of the first five weeks of this course*.* Compose brief answers to each of the following questions, typing your response in *italics* below each question. If you are confident in your answers, there is no need to submit this test. If you would like instructor feedback on your responses, please submit to the LMS.

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=98 high traffic pages during a one hour test period:   
     
   A banner: mean of 13.73 clicks (per 1000 impressions) across n=98 pages   
   B banner: mean of 13.94 clicks (per 1000 impressions) across n=98 pages.  
     
   The 95% confidence interval is as follows:   
    -0.59 < (mean difference between A and B) < 0.17.   
     
   Answer the following questions about that confidence interval:   
   1. Does **this particular confidence interval** contain the population mean difference?  
      *It is not possible to tell.*
   2. How confident are you in your answer to a?  
      *Very confident*
   3. Which banner ad do you prefer (A or B) and why?   
      *B did have a higher average, it could be just due to chance. The confidence interval includes 0, there is no data to say either is more preferable.*
   4. 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 each from new data samples: What can you say about this collection of confidence intervals?   
      *If we do the experiment 100 times, then 95 times we expect it to contain the population mean.*
   5. Which command(s) in R would you use to produce the confidence interval for each of the 100 that you constructed?

t.test(a, b)

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 df=30, 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.   
   ***The first line creates a seet of x-values from 3 to -3 by 0.1. Second line plots the density function of x with degrees of freedom of 30. Abline draws two lines at 2.04 and -2.04 and the third abline draws line at 2.5. The first two lines contain the 95% of the observations. The green line might represent a data that is in the extreme value, it is very unlikely to find after the 2.04 mark.***

***Plotting a t-distribution, the 2.5% and 97.5% quantiles are drawn. The green line represents a value or sample that is very unlikely to be observed.***

1. The output below shows 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**

*The confidence interval is 15.39 and 35.81. We can reject the null hypothesis as p-value is less than 0.05. The degrees of freedom is 29.509 and the t-test statistic is 5.1253. If the process was to repeat 100 times, then 95 times it would contain the true population.*

*A t-test was run for the UK driver fatalaties before and after thee seat belt law. . We can reject the null hypothesis as p-value is less than 0.05. The degrees of freedom is 29.509 and the t-test statistic is 5.1253.*

1. The output below shows a Bayesian 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:  
     
   **Waiting for parallel processing to complete...done.**

**MCMC fit results for BEST analysis:**

**100002 simulations saved.**

**mean sd median HDIlo HDIup Rhat n.eff**

**mu1 125.33 1.922 125.34 121.571 129.13 1 55035**

**mu2 99.83 4.993 99.80 89.844 109.57 1 60582**

**nu 43.17 30.717 34.78 5.269 104.78 1 21586**

**sigma1 23.67 1.463 23.64 20.783 26.53 1 44384**

**sigma2 22.90 3.902 22.43 15.904 30.75 1 44349**

**'HDIlo' and 'HDIup' are the limits of a 95% HDI credible interval.**

**'Rhat' is the potential scale reduction factor (at convergence, Rhat=1).**

**'n.eff' is a crude measure of effective sample size.**  


*A bayseian t-test was done to compare UK seat belt law before and after. 95% HDI is between 14.9 and 35.9. The mean of the difference of means is 25.5. All of the data points observed were above 0, which shows consistency and reliability of the data, which shows that the fatalaity was reduced after introduction of the seat belt.*

1. 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. Don’t use R or R-Studio. There are no “trick” questions. All commands below run without error.

|  |  |  |  |
| --- | --- | --- | --- |
| length(X1)  -length of the variable  -7 | length(df$mpg)  -length of the variable mpg  -7 | median(X1)   * Meandian of the x1 variable | max(X2) |
| length(df$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.  
   No, cause the values could have different units and it that makes it inconsistent and incomparable. They aren’t measuring the same thing so it dosen’t make sense.