Week 6 - Midterm Exam PRIMROSE

General Instructions: This is an honor system exam that is open book and open notes. You may consult any of the feedback I have provided to you on homework or practice exams. You may not confer or collaborate with any human besides me. Please submit the exam to the LMS by 7:30 PM on Monday, August 10, 2020. Exams submitted after this time will have a late penalty. I will be available to answer questions by email all day on Monday, roughly from 9 AM to 5 PM.

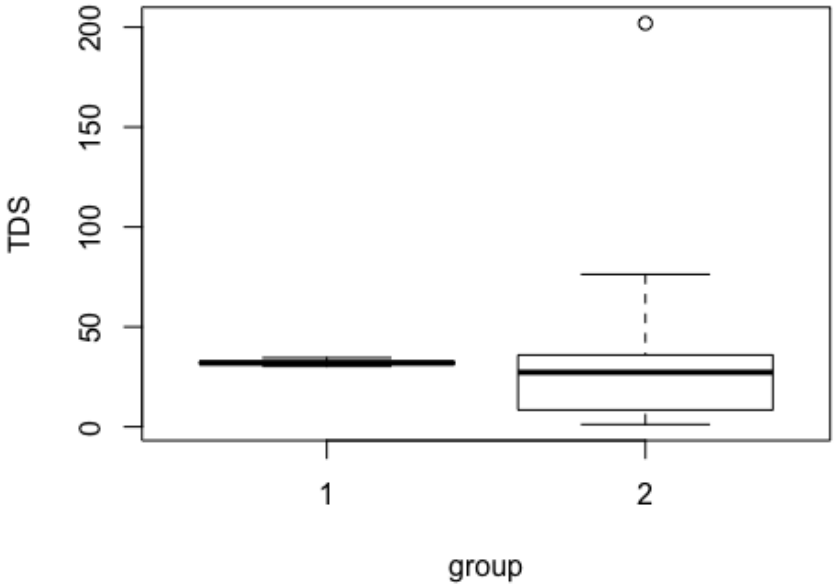
Problem Scenario: A startup company has developed an inexpensive and environmentally friendly biofilm to remove dissolved solids in water treatment plants. TDS is an abbreviation that refers to "total dissolved solids" and is measured in parts per million (PPM). Lower TDS is better – it means the water is cleaner.

The startup conducts a comparison of batches of dirty water with and without their new treatment. The control group contains batches of water processed using industry standard mechanical filtering methods. The treatment group contains batches of water filtered with the biofilm. The research (alternative) hypothesis is that the mean TDS in the treatment group will be lower than the mean TDS in the control group. Specially calibrated, highly sensitive devices are used to measure TDS, so each control and treatment batch costs a lot of money to run.

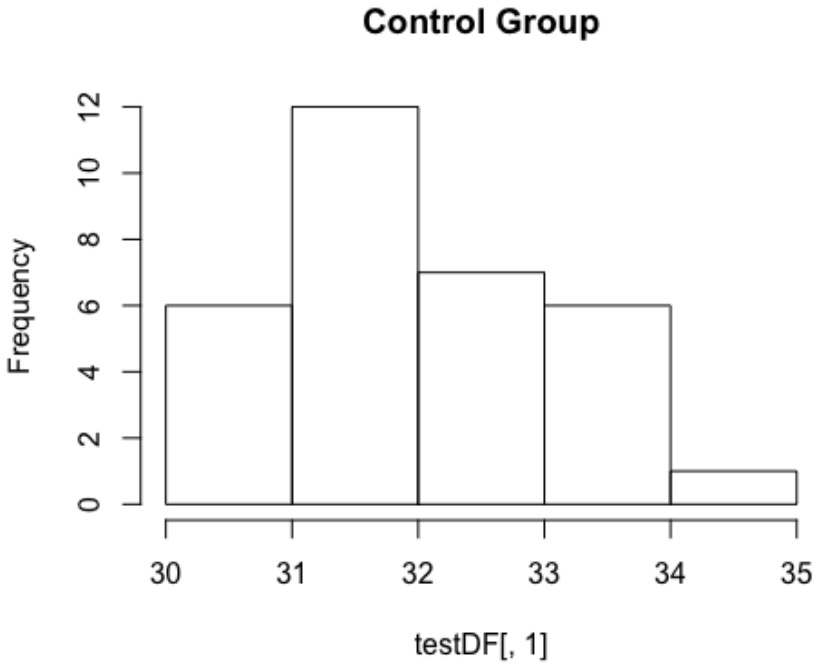
The company will not release the raw data because they consider it a trade secret, but they have provided the following statistical outputs for you. Your job is to produce a report that will guide their biologists and investors on the next steps for this project. As such, the company wants you to evaluate the research hypothesis and write an interpretation of it that can be understood by non-statisticians. Here is the output that they provided to you. You can feel free to cut and paste any of the graphics that appear below into your report, as appropriate for the audience:

```
> str(testDF)
'data.frame':  32 obs. of  2 variables:
 $ Control  : num  32.8 33.2 30.3 33.2 31.9 ...
 $ Treatment: num  35.1 16.5 31.6 11.8 201.9 ...
> summary(testDF)
    Control        Treatment
 Min.   :30.27   Min.   :  1.178
 1st Qu.:31.25   1st Qu.:  8.501
 Median :31.90   Median : 27.259
 Mean   :32.05   Mean   : 31.079
 3rd Qu.:32.81   3rd Qu.: 35.533
 Max.   :34.53   Max.   :201.908
```

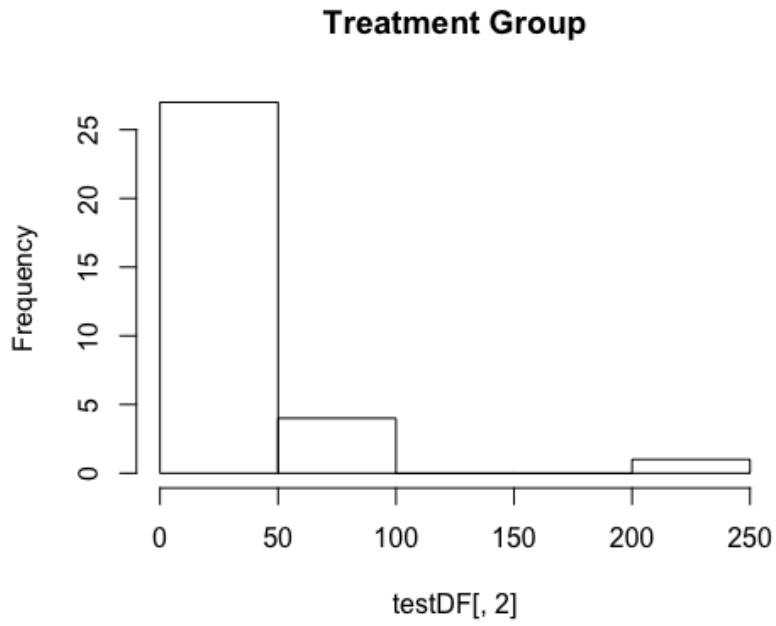
```
> boxplot(list(testDF[,1],
testDF[,2]),ylab="TDS",xlab="group")
```




```
> hist(testDF[,1], main="Control Group")
```

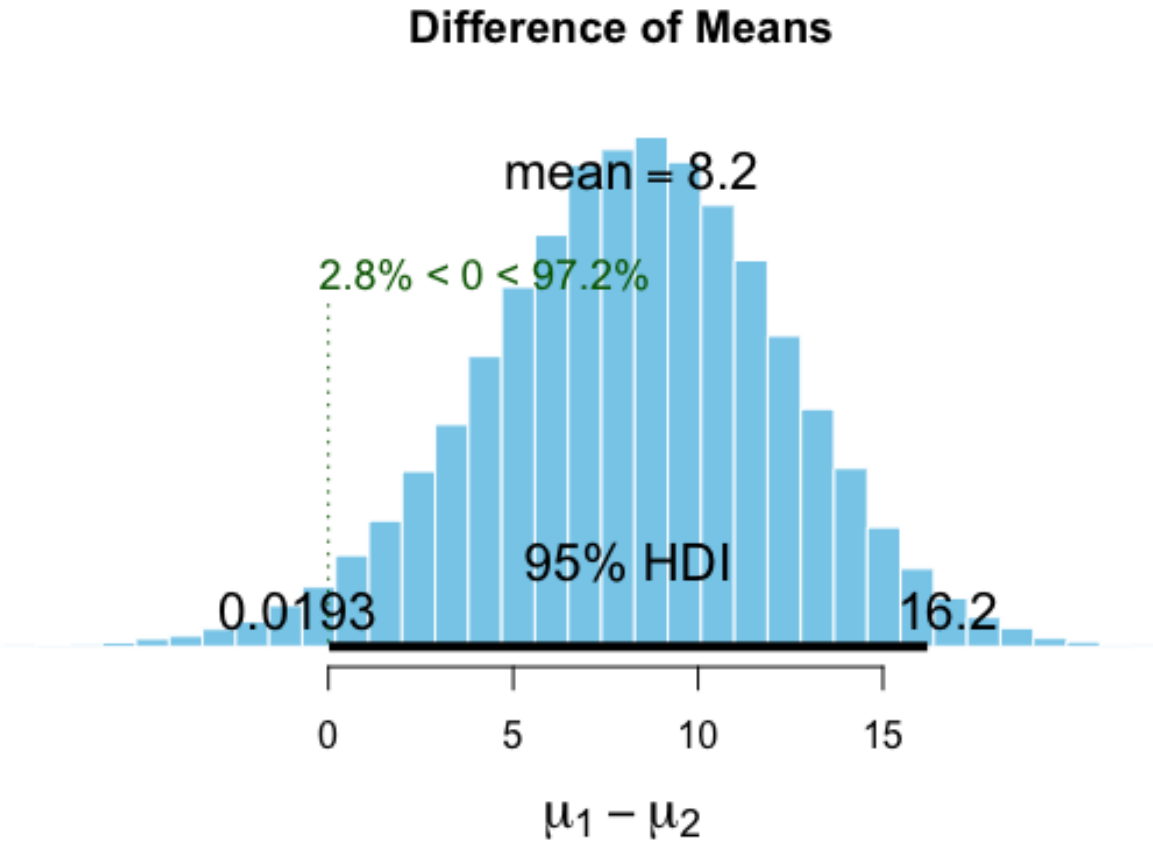


```
> hist(testDF[,2], main="Treatment Group")
```




```
> t.test(x=testDF[1],y=testDF[2])

     Welch Two Sample t-test

data:  testDF[1] and testDF[2]
t = 0.14925, df = 31.048, p-value = 0.8823
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -12.35187  14.30250
sample estimates:
mean of x mean of y
 32.05410  31.07879

> bestOut <- BESTmcmc(y1=testDF[,1],y2=testDF[,2])
Waiting for parallel processing to complete...done.
> print(bestOut)
MCMC fit results for BEST analysis:
100002 simulations saved.
         mean     sd  median   HDIlo  HDIup  Rhat n.eff
mu1    32.0202 0.1969 32.0183 31.6371 32.409 1.000 57403
mu2    23.8196 4.1250 23.7048 15.7920 31.971 1.000 47821
nu      4.9958 2.9804  4.2830  1.5269 10.157 1.012  9307
sigma1  0.9308 0.1582  0.9182  0.6327  1.246 1.000 39628
```

```
sigma2 19.5777 4.1796 19.1006 12.2217 28.019 1.001 20224
'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.


> plot(bestOut)
```



Report Components: Make sure your report includes all of the following elements.

1. (1 point) What are the lower bound and upper bounds of the (frequentist) 95% confidence interval of the mean difference?

The lower and upper bounds come from the performing a Confidence Interval t-test as provided below. A particular outcome provided by this test is an interval estimate which provides a range of lower and upper bounds of the population value. The highlighted values below how the lower bound of `-12.35187` and an upper bound of `14.30250`.

```
> t.test(x=testDF[1],y=testDF[2])

     Welch Two Sample t-test

data:  testDF[1] and testDF[2]
t = 0.14925, df = 31.048, p-value = 0.8823
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -12.35187  14.30250
sample estimates:
mean of x mean of y
 32.05410  31.07879
```

2. (1 point) What is the point estimate of the mean difference?

The point estimate of the mean difference is the center of the estimate range. In this case the amount is at the center between `-12.35187 and 14.30250.` Adding the two we get an absolute width of 26.65437. When divided by two we get a span of 13.327185 on either side. Subtracting this amount from one of the range estimates or 14.30250 – 13.327185 = .975315 which is the point estimate of the mean difference. In other words the control may be more efficient than the treatment by .975315

3. (1 point) Report the outcome of the null hypothesis significance test on the difference of means. Make sure to state the null hypothesis.

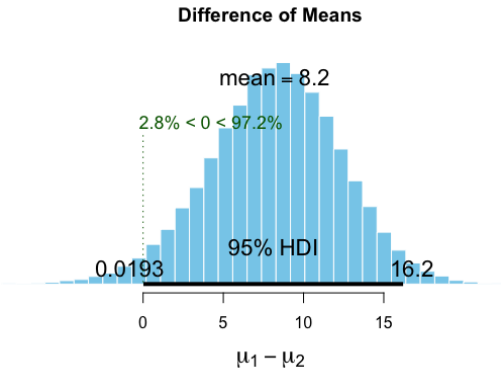
The null hypothesis is there is no difference between the control and treatment groups. As stated, the alternative hypothesis is that the treatment will be lower than the control group. As evidenced in the results (repeated below) the results include the results of the null hypothesis significance test. Given the standard alpha level of .05 a significant value is calculated and shown as the p-value highlighted below. Since the p-value is above the alpha threshold we DO NOT REJECT the null hypothesis. This does not mean we accept the null hypothesis. It means we do not have enough information either way.

```
> t.test(x=testDF[1],y=testDF[2])

     Welch Two Sample t-test
data:  testDF[1] and testDF[2]
t = 0.14925, df = 31.048, p-value = 0.8823
alternative hypothesis: true difference in means is not
equal to 0
```

```
95 percent confidence interval:
 -12.35187  14.30250
sample estimates:
mean of x mean of y
 32.05410  31.07879
```

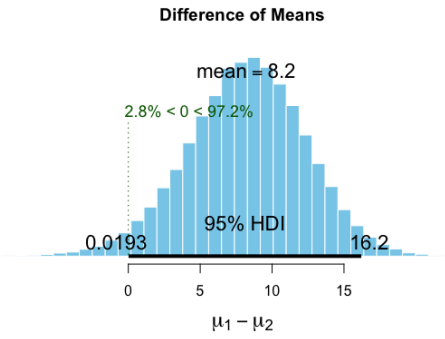
4. (1 point) Report the lower and upper bounds of the 95% Highest Density Interval for the difference of means.

As shown below the lower and upper bounds of the HDI is as follows: lower = 0.0193 and the upper = 16.2.



5. (1 point) Report the percentages of values in the posterior distribution of mean differences that are above zero and below zero.

As shown in green below the percentage of value in the posterior distribution of mean differences above zero is 97.2% and the amount below zero is 2.8%



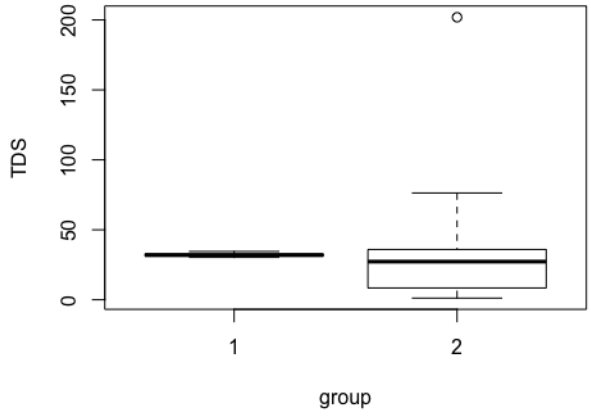
6. (5 points) Write a 1-2 paragraph technical report. The technical report should contain the detailed information that it would be *important for other statisticians to know* about the data, about the analytical results, about any anomalies you observed, and about how any such anomalies may have affected the reported results. You can cut and paste any of the graphics included above, as long as you provide a 2-3 sentence explanation of what the graphic means.

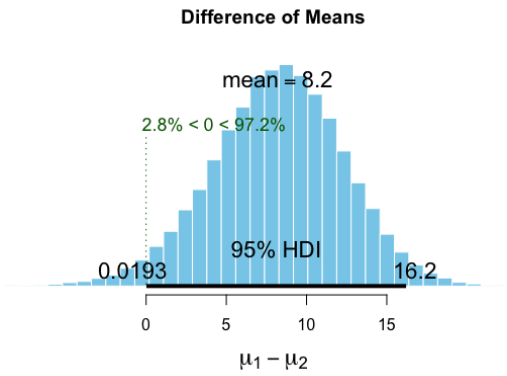
Technical Report: The request presented is to evaluate the research hypothesis. The research hypothesis is the mean TDS (total dissolved solids) of the treatment group is lower than the mean TDS of the control group. The outcome of our evaluation concerns decisions about next steps for the biofilm project. Outputs reviewed includes the following: (str() dataframe object data, dataframe summary data, boxplot of control and treatment, histogram of both control and treatment , t.test Confidence Interval, Bayesian BEST procedure output, and BEST HDI plot graph.

Analysis of the data and results provided converged on the idea that the control group provided better TPD than the treatment group. The BEST HDI data gave us probability distribution of population mean centered on a positive 8.2 in favor of the control group. Regarding the BEST HDI output, 97.2% of results are positive indicating that the control is performing better than the control Additionally, the t-test output reveals a p-value of .8823 which indicated that we do not reject the null hypothesis. Thus we are not able to accept the research hypothesis. However there are data issues to be recognized: the confidence interval spans 1.0 which causes concern about the quality of data reviewed. The confidence interval breadth is very wide. The interval width of 26.65437 between `-12.35187 and 14.30250` indicates a high degree of uncertainty. The point estimate of the mean difference as shown within the confidence interval is .975315 while the BEST difference of probability population means is 8.2. The min of the treatment group is 1.178 vs 30.27 for the control showing significant improvement in TDS efficacy. This continues through the 1st Quartile until the 3rd Quartile. However by the 3rd Quartile this improvement reverses at the 3rd Quartile and beyond. An outlier max of TPD of 201.908 for the treatment group is a suggested area for investigation. While we cannot reject the null hypothesis the improvement shown early in the results suggest additional research and testing is necessary.

```
> summary(testDF)
    Control         Treatment
 Min.   :30.27   Min.   :  1.178
 1st Qu.:31.25   1st Qu.:  8.501
 Median :31.90   Median : 27.259
 Mean   :32.05   Mean   : 31.079
 3rd Qu.:32.81   3rd Qu.: 35.533
 Max.   :34.53   Max.   :201.908
```

```
> boxplot(list(testDF[,1],
testDF[,2]),ylab="TDS",xlab="group")
```





7. (5 points) Write a 1-2 paragraph report of the results of your analysis for presentation to the company's biologists and investors. *This report should be in plain language, interpretable by non-statisticians*. Make sure to integrate the Bayesian evidence, the frequentist confidence interval, and the results of the null hypothesis significance test. The biologists and investors need to decide what the startup should do next: The essential question they want to answer is whether or not the biofilm shows promise as an alternative to traditional filtering techniques. Use the results of these statistical analysis to provide them with guidance.

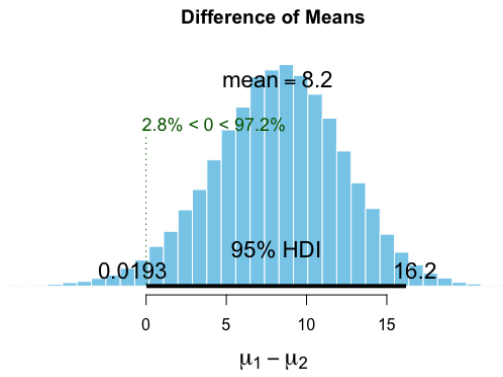
Report on Research Response: Thank you for the opportunity to assist you in your review of your biofilm project. We understand our report to you will be used to guide you in determining next steps for this project. As part of our evaluation we reviewed the following outputs: a summation of existing data (raw data being proprietary), graphic boxplots of summary control and treatment data, histograms of control and treatment, along with three specific outputs: a Welch Two Sample t-test report plus specific Bayesian BEST, and BEST HDI plot graph output.

Our evaluation of the data summaries provided suggests the control remains superior to the biofilm. We reference three results to substantiate our findings: First we evaluated the Welch Two Sample t-test results provided which includes in its output the null hypothesis significance test. In this evaluation the hypothesis provided us is the biofilm will perform superior to the existing mechanical process. We evaluate this hypothesis by performing a null hypothesis significance test. To perform this test a standard assumption is made which assumes there will be no difference between the current mechanical testing and the biofilm testing in terms of TDS results. This assumption of no difference is termed the null hypothesis. If the results of this test are below 0.05 the result is deemed significant and the null assumption of no difference is rejected in favor of there being a difference. If the result is greater than 0.05 the null is not rejected meaning there is not enough evidence to support or reject the assumption of no difference. The result is 0.8823. So we are unable to reject the null assumption. In the second evaluation we considered the provided confidence interval of a single estimate results. The result of the confidence interval is shown below. The results suggests we can be 95% confident that over the long run, assuming 100 tests of this type that the true population mean difference between the mechanical and the biofilm process would lie within the range of `-12.35 to 14.30` with a center point estimate of the mean difference being 0.975315 plus/minus of 13.33. This positive center point supports the earlier trend of mechanical being superior to biofilm. Lastly, we evaluate the Bayesian BEST HDI and Bayesian MCMC detailed results shown below based on 100,002 actual data simulations. The BEST HDI data gave us a probability distribution of population mean centered on a positive 8.2 in favor of the control group of the mechanical process with 95% plus or minus 8.1. Additionally, the BEST HDI output reveals 97.2% of results are positive indicating that the mechanical process is performing better than biofilm. However there are data issues to be recognized: the confidence interval spans 1.0 which causes concern about the quality of data reviewed. The confidence interval breadth is very wide. The interval width of 26.65437 between `-12.35187 and 14.30250` indicates a high degree of uncertainty. The min of the biofilm is 1.178 vs 30.27 for the mechanical showing significant improvement in TDS efficacy for biofilm. This continues through the 1st Quartile until the 3rd Quartile. Here biofilm show significant promise. However by the 3rd Quartile this improvement reverses at the 3rd Quartile and beyond. An outlier max of TPD of 201.908 for the treatment group is a suggested area for investigation. A possible next step is to understand why the superior TPS performance weakens by the 3rd Quartile and perform additional research which may increase the biofilm outcomes in future experiments.

```
Welch Two Sample t-test

data:  testDF[1] and testDF[2]
t = 0.14925, df = 31.048, p-value = 0.8823
```

```
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -12.35187  14.30250
sample estimates:
mean of x mean of y
 32.05410  31.07879
```

BEST HDI




```
> summary(testDF)
    Control         Treatment
 Min.   :30.27   Min.   :  1.178
 1st Qu.:31.25   1st Qu.:  8.501
 Median :31.90   Median : 27.259
 Mean   :32.05   Mean   : 31.079
 3rd Qu.:32.81   3rd Qu.: 35.533
 Max.   :34.53   Max.   :201.908
```