Week 6 - Midterm Exam

Student: Leonard Armstrong

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 Tuesday, November 12, 2019. Exams submitted after this time will have a late penalty. I will be available to answer questions by email all day on Tuesday, 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.

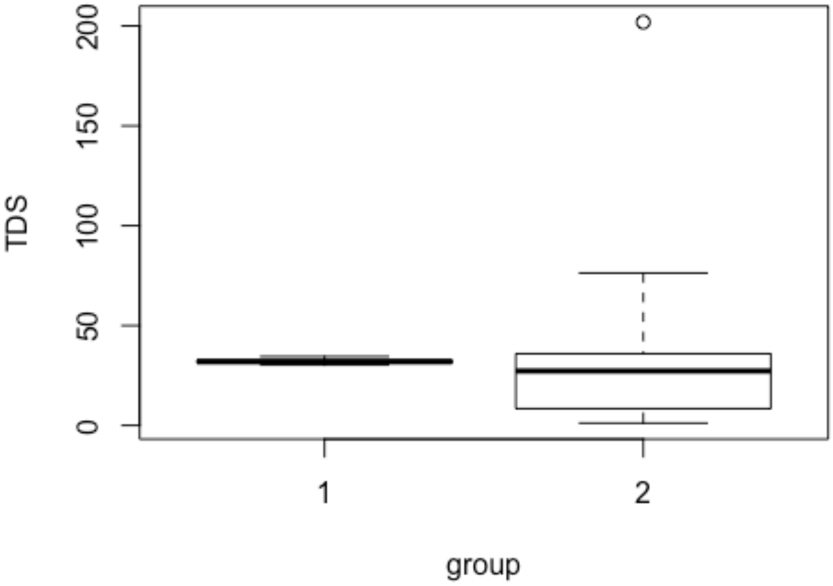
The startup conducts a comparison of batches of dirty water with and without their new treatment. The control group contains batches of water processed using industry standard mechanical filtering methods. The treatment group contains batches of water filtered with the biofilm. The research (alternative) 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's 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:

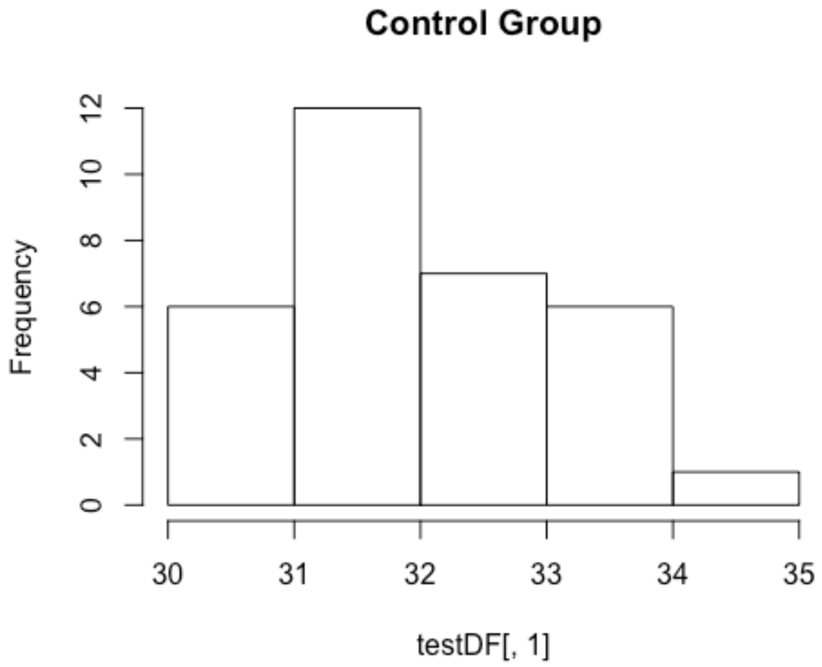
```
> str(testDF)
'data.frame':	32 obs. of  2 variables:
 $ Control  : num  32.8 33.2 30.3 33.2 31.9 ...
 $ Treatment: num  35.1 16.5 31.6 11.8 201.9 ...
```

```
> 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  
```

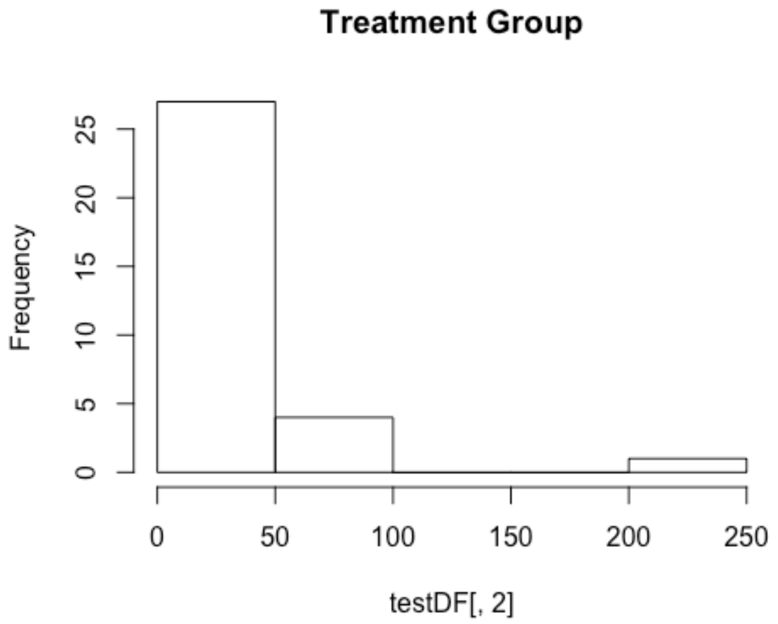
```
> boxplot(list(testDF[,1], testDF[,2]),ylab="TDS",xlab="group")
```



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

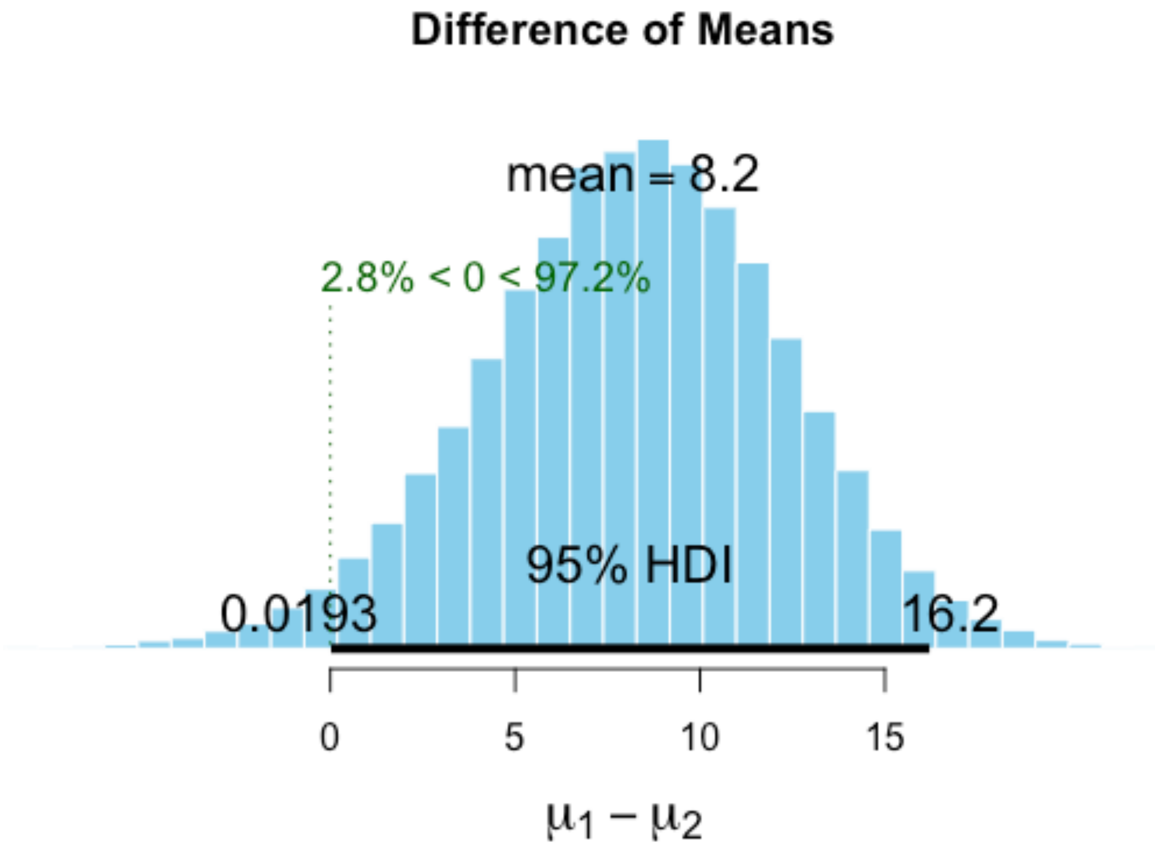


```
> hist(testDF[,2], main="Treatment Group")
```



```
> 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.

Answers

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

   CI lower bound = -12.35187
   CI upper bound = 14.30250

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

   The point estimate mean difference from the t.test is 0.9753. This can be computed both by (a) mean of x – mean of y, and also by (b) taking the mid-point between the CI upper and lower bounds since the t.test CI is computed symmetrically around the estimated mean.

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 for this problem states that "there is no significant difference between the means of the control group and the treatment group." The p-value of 0.8823 resulting from the test is not smaller than an alpha of 0.5 and thus we would not reject the null hypothesis.

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

   HDI Lower: 0.0193
   HDI 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.

   2.8% of the posterior distribution of means is below 0.
   97.2% of the posterior distribution is above 0.

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

Introduction

The purpose of the experiment was to determine if the mean total dissolved solids (TDS) in the treatment group were lower than those in the control group using industry standard filtering practices. While data suggests there *may* be a difference between the two groups, a number of anomalies show cause for concern.

Data Review

An exploratory review of the data found the following.

First, the range and distribution of the treatment group are much wider than those of the control group. While the control group spans a range of approximately 4.26 TDS, the treatment group spans a range of over 200 TDS. While the box plot shown in Figure 1 reveals that one outlier accounts for the span to 200, it can also be seen that the non-outlier maximum mark is around 75 TDS which is still more than double the maximum TDS value for the control group.

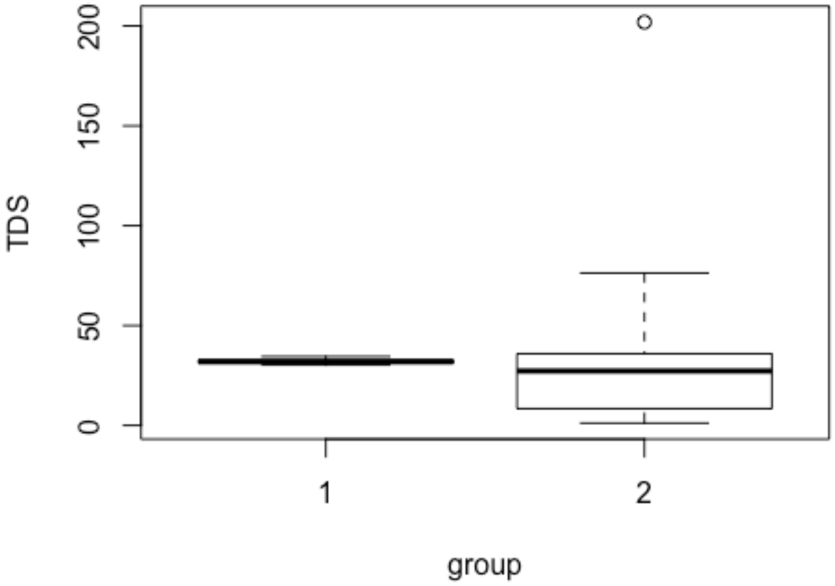


Figure 1- Box plot of control group and treatment group TDS

Looking at the quartile numbers for the two groups directly (see Table 1) shows that at least 25% of the treatment group TDS exceeds the maximum value encountered for the control group. It should also be noted that at least half of the treatment group contained less TDS than any of the control group.

	Control	**Treatment**
Min	30.27	1.178
1st Quartile	31.25	8.501
Median	31.90	27.259
Mean	32.05	31.079
3rd Quartile	32.81	35.533
Max	34.53	201.908

Table 1- Control and Treatment Group Statistics Comparison

To gain a better sense of the two distributions distribution it is recommended to recreate the treatment group histogram using bins every 5 TDS. This would provide better visual comparison between the small range of the control group and the approximate overlapping range of the treatment group as well as provide the distribution shape for the treatment group on either side of the control group range.

Frequentist t.test

A review of the results of the frequentist t.test indicate the null hypothesis (i.e., that there is *no significant difference* between the two groups) *not* be rejected as the resulting p value of ~0.88 far exceeds an alpha of 0.05. Additionally, the confidence interval for the mean difference (-12.35187 to 14.30250) spans 0. However, as should be noted with any t.test, we do not know if this is the case with this sample as an outlier, or with the majority of samples. Additional samples would be needed to see how the sample means (control group: 32.05410; treatment group: 31.07879) established by this test compare with additional tests in the long run.

Bayesian Markov Chain Monte Carlo Simulations

Since the cost of obtaining additional samples may be prohibitive, some of the limitations of t.test with respect to making claims of a likely population mean from one sample of data have been mitigated by the fact that a Bayesian-based Markov chain Monte Carlo (MCMC) simulation was also run on the samples. As a result of the MCMC model, an estimated mean difference of 8.2 TDS between the groups was calculated with the control group lowering the mean by that amount. As an additional point of concern, this is over 8 times higher than the point estimate mean computed in the t.test of 0.9753.

Additionally, a broader review of the results suggests this result is on the borderline of meeting its statistical significance and, given a relatively small sample size of 32 observations per group, a few data points could change the viability of this result in either direction. While the HDI does not span 0 it is very close at 0.0193 TDS for an HDI spanning almost 16.2 TDS in total. Additionally, 2.8% of the left tail is to the left-side of 0. Assuming a 2-tailed test with 2.5% on the left and 2.5% on the right of the HDI, this means that more than half of the 2.5% on the left-hand side is less than 0. This is possible since the computed distribution is not a perfect normal distribution.

Given the sample size of 32 observations, the addition of the one large outlier at 201 TDS accounting for approximately 3% of observations make a difference in your results. Assuming this point is truly an outlier, this could change the results of your two test significantly.

Summary

In short, the conflicting data of these two tests along with the fact that, while the results of the two MCMC test may indicate an overall decrease in mean TDS, there are still questions about the addition of larger TDS values for ¼ of the sample (and potentially population) to handle. There is also a remaining question of how to handle the 201 TDS outlier.

Recommendations for Next Steps

The following recommendations are herein suggested next steps deriving from the provided data:

1. Regenerate the treatment group histogram with bins of 5 TDS for better comparison between the control group and the treatment group.
2. Run additional statistical tests such as ANOVA to see if more information can be gained. Although running ANOVA on only a pair of groups may not

produce results very similar to t.test so additional statistical tests should be considered.

3. Given a control group resulting in a fairly localized TDS (between 30 and 35) determine what percentage of treatment observations can be in excess of 35 TDS even if the overall mean is lowered. Are the 25% of cases in excess of 35 TDS in the treatment group acceptable given an overall lower mean, or will those results cause damage to the company?
4. Determine if a span of 200 TDS (or 75 TDS, ignoring outliers) is acceptable. Was such a large range known by the treatment research and development team and, if not, was any specific range targeted?
5. Rerun the test with the treatment group outlier removed. If the results change drastically additional studies must be performed to determine if this data point is truly an outlier (given a relatively small sample).
6. If at all possible, run a second with new (preferably larger) samples.

Gratefully Yours,
The Technical Staff at Armstrong Data Sciences

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.

Executive Report

Introduction

After reviewing the water treatment experiment data that was provided to us, we have determined that there may be cause for optimism with respect to the viability of your new treatment, but there are also questions that must first be answered and issues that must be resolved before considering a move to market. While one analysis suggests that your treatment *may* lower the overall TDS of treated water, there are also a number of issues that show cause for concern.

The Good and the Bad

Provided data showed that water treated with industry standard practices (hereafter, *the control group*) resulted in between 30 and 35 TDS. On the positive side, over ½ of the data points in the sample treated with your treatment (hereafter, *the treatment group*) resulted in lower than 30 TDS. That is, your treatment resulted in lower TDS than even the lowest control group TDS, half of the time.

However, on the downside, at least ¼ of the data points in the treatment group resulted in greater than the control group maximum 34.53 TDS. These data points spanned to twice the TDS max with one outlier resulting in over 200 TDS or over 5 ½ times the maximum control group TDS. In fact, this outlier may have had an impact on your overall results. If it is truly an outlier, then we recommend rerunning the statistical tests with this data point removed. However, while this data point is a statistical outlier for this relatively small data set it is unknown how truly rare such a value will be encountered with a larger set of observations.

TDS Ranges

First, the range and distribution of the treatment group are much wider than those of the control group. While the control group spans a range of approximately 4.26 TDS, the treatment group spans a range of over 200 TDS when one includes the outlier data point. Without that outlier, the treatment group maximum mark is around 75 TDS which is still more than double the maximum TDS value for the control group. Given the 500mg/L EPA standard for total dissolved solids[1], this may be an acceptable value, or your company standards and/or customer standards may require a stronger threshold.

The second concern with respect to the data range is that the treatment group spanned a 200 TDS range of results (or 75 TDS if the outlier is excluded) compared to a less-than 5 TDS range for the control group. How important is consistency of treatment for your organization especially given that 25% of the times your treatment resulted in TDS values that were higher than the maximum achieved from current standards?

Was Overall Mean (Average) TDS Reduced?

We were provided with the results of two separate statistical test to determine if the mean TDS was lowered, on average, with your treatment. Unfortunately, the results of the tests are somewhat contradictory.

The first test (called a *t.test*) used a model that emphasizes frequency and proportions of data to draw inferences. Applying this test to your water treatment problem, one hypothesizes that *there is no statistically significant difference in the control group vs. treatment group* and then looks for significant evidence to reject

[1] Water Research Center. "Water Testing Total Dissolved Solids Drinking Water Quality". Oram, Brian. https://water-research.net/index.php/water-treatment/tools/total-dissolved-solids.

this hypothesis. Execution of this test on your data set did not provide evidence that the hypothesis could be to be rejected. While we also cannot conclude that the hypothesis is true from the test, we also know that this test provides more meaningful information when run against multiple data sets (samples) and the data provided was for only one result.

The second test (called an *MCMC* for *Markov chain Monte Carlo simulation*) is based on a statistical philosophy that used new data to make inferences on existing data. This test showed your treatment lowered the mean by 8.2 TDS when compared to the control group. However, we also observed that this result teeters on the 95% confidence level, satisfying the criteria to assume a lower mean. Given the relatively small sample size of 32 data points, a few additional data points (such as the aforementioned outlier) could have swayed this result more toward or away from the 95% confidence level.

In short, the conflicting data of these two tests along with the borderline findings on the MCMC test lead us to recommend you conduct additional statistical tests and, ideally additional experiments before moving to market. Your research and development staff have been provided with more details in our technical report.

Recommendations

The following recommendations are herein suggested next steps deriving from the provided data:

1. Have your research staff rerun the two previously executed statistical tests with the 201 TDS outlier removed. If these results clarify the reduction in mean TDS then additional research much be performed to determine if this data point was truly an outlier in the long run.
2. Have your research staff run additional alternate statistical tests to see if more information can be gained since the two tests executed provided conflicting results.
3. Determine what percentage of treatment observations can be in excess of 35 TDS even if the overall mean is lowered. Are the 25% of cases in excess of 35 TDS in the treatment group acceptable given an overall lower mean, or will those results cause damage to the company's reputation and/or revenue?
4. Determine if a span of 200 TDS (or 75 TDS, ignoring outliers) is acceptable. Will such a large range be accepted by your customers, even if your treatment results in lower TDS on (mean) average? We recommend further research directed to lower the upper bound of this range to come in line with current processing standards of 35 TDS.
5. If at all possible, run a second with new (preferably larger) samples. It is understood that such an experiment will be costly but without such research you will be going to market with conflicting evidence as to the overall benefit of your test.

It has been our pleasure to serve you,

The Executive Staff at Armstrong Data Sciences