IST772 Midterm

Kyle Walter

5/23/2021

1. The lower bound of the confidence interval is -12.35. While the upper bound is 14.3.

2.
```
PEM <- mean(c(-12.35187,14.30250))
PEM
```

```
## [1] 0.975315
```

   The Point Estimate Mean Difference is .975

3. The null hypothesis is that mean difference between the control method and biofilm is the same or that the biofilm shows a higher level of Total Dissolved Solids than that of the control treatment.

   The alternative hypothesis is that the biofilm shows lower levels of Total Dissovled Solids in the water.

   Given a P-value of .88, and the confidence interval both passes through zero and contains several values less than zero meaning in the oberservations there are several values point to the Control Treatment out performing the biofilm. As such, we fail to reject the null hypothesis as there isn't support in the data validate the alternative hypothesis

4. The lower bound of the HDI is mean difference of 0.0193 and the upper bound of the HDI mean difference 16.2.

5. The percentage of values in the posterior mean distribtuion below zero is 2.8% while 97.2% of the values are above zero.

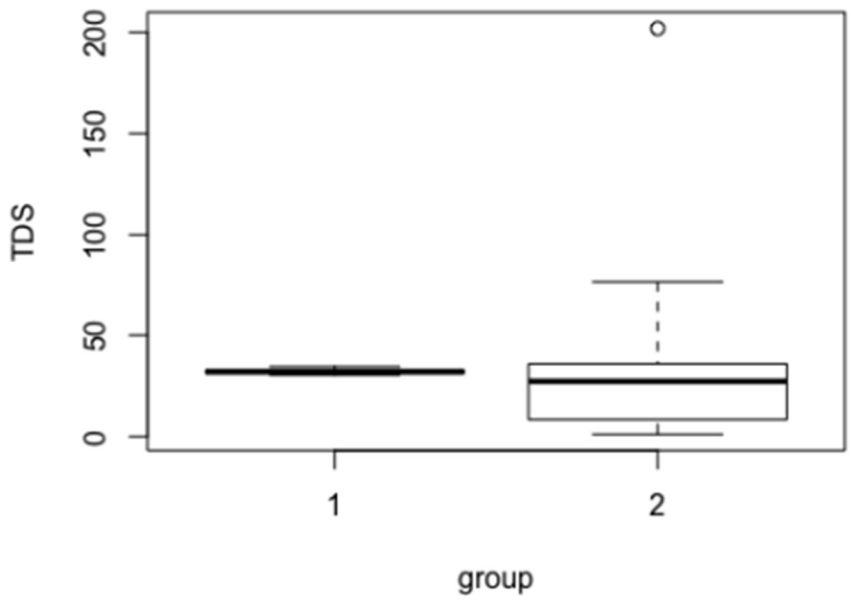
6. Technical Report: In reviewing the Biofilm test product against the control Treatment. I have reviewed by the frequenties an bayesian results. Starting with the measure of dispersion.

```
dispersion <- data.frame(control=c(30.27,31.25,31.90,32.05,32.81,34.53),
Treatment = c(1.78, 8.501, 27.259, 31.079, 35.533, 201.908), row.names =
c("Min.", "1st Quantile", "Median","Mean","3rd Quantile","max."))
dispersion
```

```
##              control Treatment
## Min.           30.27     1.780
## 1st Quantile   31.25     8.501
## Median         31.90    27.259
## Mean           32.05    31.079
```

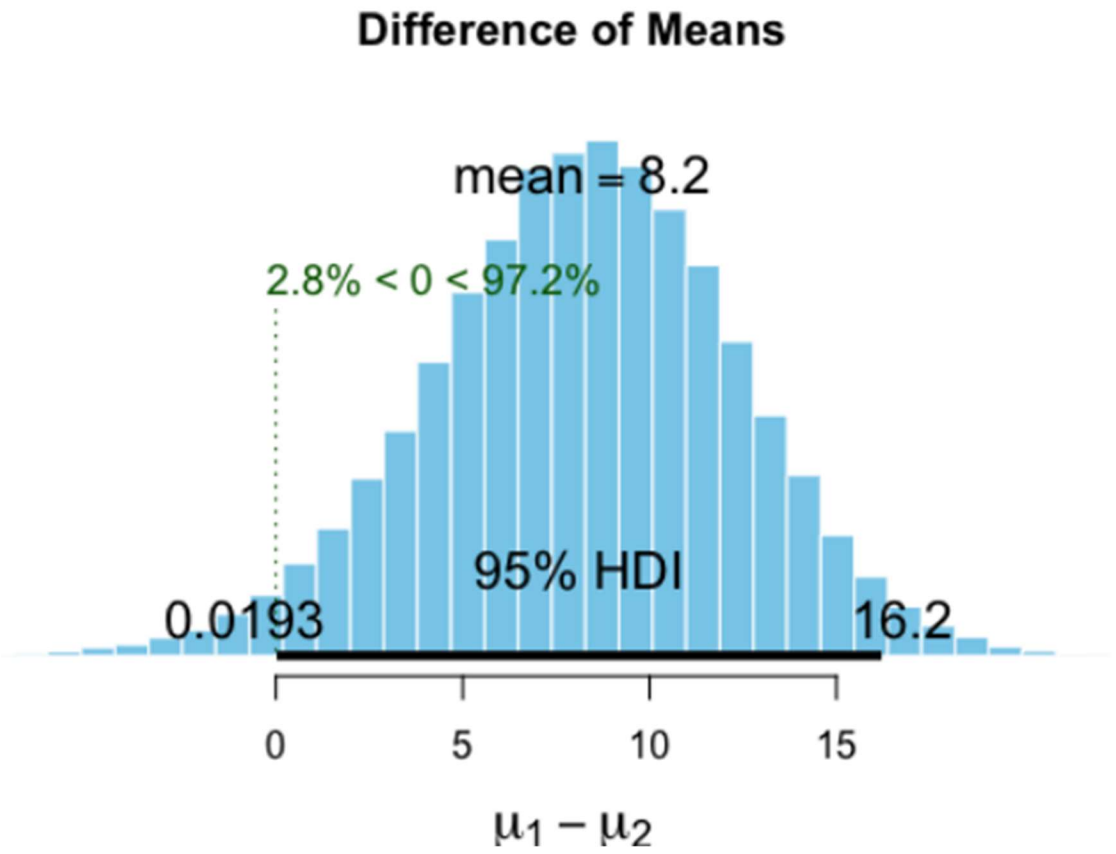
```
## 3rd Quantile   32.81    35.533
## max.           34.53   201.908
```

The control shows a very narrow range from 30.27 to 34.53 and average of 32.05. While the biofilm covers quite a large range total dissolved solids parts per million from 1.78 to 201.91. The average is slightly lower at 31.08. Looking at the boxplots provided, we can see the majority of measurements are below treatment cluster, however there are outliers that are much higher such as max measurement. It could be that there are problems in the data collection, application of the product, or training of how to read the results. I have included the chart below for quick reference.



Client provided boxplot of Control vs Biofilm water treatment

In addition to looking at quick measures of dispersion, I have reviewed the results of the frequentist T-test and baysian MCMC High Density Model. The T-test shows us a really wide confidence interval ranging from -12.35 up to 14.3. The interval passes through zero meaning that the population difference of means could very well be zero. In addition the high p-value of .88 also lets us know that the observation is not a rare event, thus failing to reject the null hypothesis. HDI does hold a bit of better news. 95% of the value realized by the montecarlo model are between .0193 to 16.2 with model predicting the population mean of 8.2, which suggests that the biofilm is out preforming the control test. I do however caution since 2.8% of the values are zero or less that would assert no difference or worse performance is possible. I have included the chart for quick reference below. I would suggest given the conflict results of the two different tests and the outliers in measurements that sampling from the dataset to grown the measures of center or if budget exists resting with stronger controls around measurement of the solids and application of the product.



7. Business Summary: Dear Management Team,

   I have reviewed the statistical results provided by your technical team. Two different systems of were applied in this review to test the certainty and the results are inconclusive about the ability of biofilm to preform relative to the current market standard for purifying water. The first test, the t-test, produces two measure. The first a probability value that indicates how rare it would be to see an event as a percentage, this was 88%, and using common standards a passing test would be less than 5%. The other result is a confidence interval, or a band of numbers where if we could measure the entire population of samples we would find the true measurement. The larger this interval the less certain we become of the results. In this case the interval was quite large and represented many values where no improvement and less efficacy of the product were reasonable. Using these two measure we could conclude that the data has failed to reject the null hypothesis test, meaning the data does not support that the biofilm preforms better than that of the control system.

   The other measure applied, provided significantly better results. An algorithm was applied where the attempt to find the true population mean difference through trials where the most commonly observed value. This method indicated that the data from the biofilm likely pointed to the 8 parts per million improvement when using the biofilm over the current market standard. The model still indicated a 2.8%

chance of the value being no difference or worse performance. The other promising trend, when looking at the how the dataset appear next each other, almost 3/4th of the data shows better results than the current standard. I do have concerns around the how widely dispersed the data is, especially when looking at the higher values. There may be data collection issues, that if controlled for would enough certainty.

In conclusion, the data shows some promising trends with an algorithm showing a large improvement over the control, and the dispersion views showing the majority of measurements less than the tightly grouped cluster of the control. However; as it failed to gain support for rejecting the null hypothesis, there is not enough evidence to argue the biofilm preforms better than the current market standard.