Nora Lin
IST 772 Fall 2021
Midterm

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

   The lower bound of the 95% frequentist CI is -12.35187
   The upper bound of the 95% frequentist CI is 14.30250

2) ***What is the point estimate of the mean difference?***

   About 0.975 TDS

3) ***Report the outcome of the null hypothesis significance test on the difference of means. Make sure of state the null hypothesis***

   The null hypothesis is that there is no difference between the mean TDS of the treatment group and the mean TDS of the control group. It can also be stated as the null hypothesis is that the true difference in means is equal to zero.

   The t-test evaluated whether there was a mean difference in the control group in terms of TDS values compared to the treatment group's TDS values. The results showed a p-value of 0.8823. Assuming our alpha values is 0.05, we would fail to reject the null hypothesis since our p-value, 0.8823, is greater than our assumed alpha value, 0.05.

4) ***Report the lower and upper bound of the 95% highest density interval for the difference of means.***

   The lower bound of the 95% HDI is 0.0193
   The upper bound of the 95% HDI is 16.2

5) ***Report the percentages of values in the posterior distribution of mean differences that are above zero and below zero***

   2.8% of the values in the posterior distribution of mean differences are below zero
   97.2% of the values in the posterior distribution of mean differences are above zero

6) ***Technical report for other statisticians***

The boxplot, historgram, and Bayesian t-test all show that there is an outlier in the treatment group. The outlier falls outside the 3rd quantile of the treatment group data. It's effects are seen in the histogram, as a treatment group is heavily right skewed. The variance of the control group seen in the Bayesian test is very small, 0.9308, compared to sigma2, 19.5777. The importance of this outlier should be emphasized when discussing the analysis with biologists or investors.

Our null hypothesis is that there is no difference between the mean TDS of the control group and that of the treatment group. From our Welch Two Sample t-test, we are 95% confident that the population parament is between -12.35187 and 14.30250. Our t-value was 0.14925 with 31.048 degrees of freedom and a p-value of 0.8823. Assuming that our alpha value is 0.05, we would fail to reject the null hypothesis. This means that if replicate our whole study 100 times, on average, in 95 of those replications the calculated confidence interval would contain the actual population mean difference. The interval estimate of the population mean difference ranged from -12.35 to 14.30, a span of about 26.8 TDS. This is a wide interval indicating we have high uncertainty.

From our Bayesian test we can state that 95% of the likely values of the population mean difference lie in the bell-shaped area between 0.0193 and 16.2 TDS. The population mean difference is somewhere near 8.2 TDS with the 95% highest density interval ranging form 0.0193 and 16.2 TDS. 97.2% of the mean differences in the distribution were positive and 2.8% were negative.

*7) **Results of your analysis for the presentation to the company's biologists and investors.***

The aim of startup's experiment was to investigate whether the biofilm shows promise as an alternative to traditional filtering techniques. The research hypothesis was that the average TDS of our biofilm group will be lower than the average TDS of the traditional filtered group. We **cannot prove** anything from samples or from statistical inferences. Our 95% confidence interval is a long-run prediction about what would happen if we replicated the study over and over again, giving us a single interval estimate. Our BESTmcmc simulation gives us a distributional model of population parameter. This test is helpful since each TDS reading is expensive.

Our confidence interval contains zero, therefore we cannot determine if there is a difference between the mean of the traditional filtering techniques and that of our biofilm techniques. We have high uncertainty with our results. This means **cannot** say with 100% certainty that the biofilm technique produces lower TDS levels than the traditional filtering techniques. Our BESTmcmc showed that the likelihood of a population mean difference between the traditional filtering and biofilm being the same or that the mean of the traditional filtering is greater than the biofilm is 97.2%. These results are left for the objective interpretation of the biologists and investors. As statisticians, we again can only report the results with high uncertainty.

We want our biofilm results to show a lower amount of TDS than the traditional filtering methods. We caution that the experiment did contain an extreme measure of high TDS for our biofilm filtering method. If budget permits, we suggest that the startup reruns its sampling of TDS levels from the traditional filtering methods and TDS levels with our biofilm.

Cut outs:

his is cause for concern for which our analytical interpretation of any difference in mean should be carefully stated.

There is an outlier in the treatment group that significantly deviates from the mean of the treatment group data. From the boxplot comparing TDS raw values of 32 observations of the control and treatment group, we can observe a tight clustering for the control group but large spread in the treatment group. This is also backed in our histograms, as the control group presents as a normal bell curve while the treatment group is heavily right skewed.

##In our raw data summary output, we see that the mean of the treatment group is smaller than the mean of our control group despite the inclusion of our outlier observation in the treatment group. Mathematically, this signals that our treatment group mean sin outlier should also be smaller than the mean of the control group. This is also backed by our median difference between the two groups, which is unaffected by outlier values. As statisticians, we want to investigate the outlier to see if it was result of mis-entry, experimental error, or truly reflects the observation of the treatment group. Since our BESTmcmc simulation is also affected by the outlier in the raw data, without concerns for money spent, we would recommend repeating the original experiment to obtain another set of 32 observations.

- Bayesian evidence
- Frequentist confidence interval
- Results of the null hypothesis test
- Whether or not the biofilm shows promise as an alternative to traditional filtering techniques
- Provide them with guidance