Midterm exam

Name: Maya Mileva

**REPORT**

Environment friendly biofilm that removes dissolved solids in water treatment is developed. TDS is an abbreviation that refers to “total dissolved solids” and measures in parts per million (PPM). Lower TDS gives better results, meaning the water is cleaner.

Comparison of batches of dirty water with and without the new treatment are conducted. 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 target of the research is to prove that the mean TDS in the treatment group will be lower than the mean TDS in the control group.

The raw data will not be released by the company, they consider it a trade secret. It consisted of 32 observations with 2 variables – control and treatment group.

**> 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**

The range of the control grouped varies from 30. 27ppm (as minimum) to 34.53ppm (as maximum value). We are observing very tight range. 1st Qu. refers to the dividing line at the top of the ﬁrst quartile. If we look at all the control values and line them up in order, we can divide up the whole into 4 groups, where each group had the same number of 2 observations (smallest on the left and largest on the right). The 1st Qu(31.25ppm) is the value of the control group that divides the ﬁrst quarter of the cases from the other three quarters. 3rd Qu is the third quartile(32.81ppm). It represents the third ﬁnal dividing line that splits all the cases into 4 equal parts. Quartiles give a sense of the shape of the distribution, can be used for comparisons too (if we want to know if a sample is drawn by speciﬁc data set). Median, 31.90ppm, refers to the value that splits the whole Control group in half (half having higher values and half - lower). The average DTS for the Control group is 32.05ppm. The mean and the median have really close value, we can have normal distribution (evenly spread data).

With the treatment group we are observing wider range, from minimum value of 1.18ppm to maximum value of 201.98ppm. That might be due to extreme values(outliers) in the data used. When we look at the distribution, represented by the quartiles, we noticed that 50% of the data is spread out between 8.5ppm and 35.53 ppm (27.03 interquartile range). We are observing that median in the Treatment group has a value of 27.29 which mean that, half of the values are higher with max of 201.9ppm and half are lower with min of 1.17ppm. We must acknowledge that the medium is not sensitive to outliers presenting in the data and that the distribution here is not normal. The average DTS for the Treatment group is 31.07ppm.



Visualizing the 2 groups of water batches with a box plots is giving us a better view of their distributions. Box plots pack a lot of information into a small piece. In each case upper and lower boundaries of the box represent the ﬁrst and third quartiles respectively (25% of the cases are above the box and 25% are below the box). The dark band in the middle represents the median for each variable. Whiskers represents the position of the max and min values, respectively.

For the Control group we can confirm the small range, no outlies. The mean is slightly higher than the mean of the Treatment group, and the boxplots overlap, which indicates similarity between these groups. The whiskers show that very lowest value for Control falls at the 3rd quartile for the Treatment reafﬁrming their diﬀerence. We are working with samples so we cannot be certain that the diﬀerence between these three groups can be trusted. The treatment group has an outlier, which can skew the data distribution. Has higher TDS values in general and like we notice from the histograms less variance. We can see that the mean is toward the upper end so that means bigger variance from it.

The mean values of the two groups are really close to each other, indicationg that they are similar. Control group has higher values in general, which makes is the worse. We need to remove the outlier in order to get better understanding of the distribution and decide which one is better.

Histogram are used to display the data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data.

* Control group histogram



The control group contains batches of water processed using industry standard mechanical filtering methods. The x-axis are the diﬀerent values of TDS, and the y-axis are frequency of that value. Histogram is left skewed, with most values between 31 and 32ppm. The mean is at 32.02ppm. Small range, with a peak around the median.

* Treatment group histogram



The treatment group contains batches of water filtered with the biofilm. We can see more variance from the mean. With smaller sample sizes tails are higher than the normal curve , representing greater uncertainty. Most of the values are between 1.17and 50ppm. There is obvious outlier, that can be skewing the data and giving it wider range and spread meaning higher band of uncertainty. In general treatment group has lower values, but they overlap with the control group.

The get better understanding of the difference between these 2 groups t-test is conducted between them.

**> 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**

The t-test examined whether there was a mean difference in batches of dirty water with and without the new treatment. Mean values of TDS without treatment were 32.05, while mean values after the treatment was conducted were 31.07. The observed p-value for this test was 0.8823. We also constructed 95% confidence interval around the mean difference of .98ppm. That confidence interval ranged from -12.35 up to 14.30. Note that it is impossible to know whether this particular confidence interval actually contains the population mean difference. The width of the confidence interval, about 26.65ppm, gives us sense of uncertainty in these estimates. To reduce this uncertainty, we would have to increase sample sizes, reduce variability in weight within groups or both.

The p-value (0.8823) represent all of the area in the tails of the distribution, beyond the observed t-value – the probability of obtaining a value of t at least as high as what was actually observed. If we assume alpha is equal to .05, the p-value from the t-test is larger than alpha. The bigger the diﬀerence, the more convincing the results are. Still we fail to reject the null hypothesis, stating that there is no mean diﬀerence between the mean of the two groups (control and treatment). The target of the research is to prove that the mean TDS in the treatment group will be lower than the mean TDS in the control group proving difference in those means).

Failing to reject null hypothesis does not mean that we accept the null hypothesis, rather than we have no good evidence either way. Likewise, p-value does not inform the question of how likely the null hypothesis is.

The frequentist methods suggest no significant difference.

**> bestOut <- BESTmcmc(y1=testDF[,1],y2=testDF[,2])**

Waiting for parallel processing to complete...done.

The BESTmcmc() function computes a probability distribution for the mean diﬀerence between the two groups, using the full information available in the two samples of data.

**> 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.

The beginning of the output shows that 100,002 steps occurred in the MCMC. Each of those steps generated one set of credible estimates of the population parameter. There is a table of statistics describing each distribution of the population parameter., represented by column headings: mean, standard deviation, median, the lower bound of HDI and upper bound of HDI. For instance 32.02ppm is a point estimate of the mean population value of the control group. The next values on that line are as labeled. Rhat is close to 1, so that shows that MCMC process converged properly on sensible results.

> plot(bestOut)



This Bayesian t-test examined whether there was a mean difference in batches of dirty water with and without the new treatment. Mean values of TDS without treatment were 32.05, while mean values after the treatment was conducted were 31.07. The Bayesian t-test produced a distribution of estimates for the mean difference. The center of this distribution was a difference of 8.2ppm (point estimate). The population mean difference probably lies somewhere close to this value: The 95% highest density interval for this distribution of estimates ranged from 0.0193 to 16.2 (HDI is explained as the 95% probability that the population mean diﬀerence between the two groups falls within this range). The “mu” symbol was used to indicate the population mean.

2.8% < 0 < 97.2% - show the proportion of the mean differences in the MCMC run that were negative versus the proportion that were positive. We can see an evidence here: 2.8% were negative and 97.2% were positive, meaning that treatment group was more effective.

Putting everything all together into a single sentence we might say that the population mean difference is somewhere near 8.2ppm, with 95% HDI ranging from 0.19ppm to 16.2ppm. The likelihood of the population mean difference of 0 or larger is 97.2%. The HDI doesn`t overlap with 0, so we can state that we have credible results and significant difference.

We must mention the statistical issues here too. Models are not perfect - they are modeling what the data shows in relation to a set of priors. There is nothing perfect about the BEST procedure but hey will often compensate for distributional anomalies in a sample. If you have an outlier, for instance, that would drag a mean in in a direction that would cause the T test to become significant. That might be the reason for the contradiction between the two performed test here, there is an outlier in the treatment group. The Bayesian model in this case can be showing a better picture of the actual results. The common assumptions made when doing a t-test include normality of the data. The input data in this case does not fit this assumption, which affects the credibility of the result too.

Boxplot showed that treatment group values are coming a little lower in general. To have more exact result removing the outliers and replotting the data is suggested.

t-test was not statistically significant (P-value >.05), we fail to reject the null hypothesis that there is no mean diﬀerence between the mean of the two groups. The Confidence interval overlap with 0 (-12.35 up to 14.30). If the 95% CI of the difference contains 0, then there is no difference in TDS between groups (If it doesn't contain 0, then there is a statistically significant difference between groups). T-test is showing us a wide confidence interval - about 26.65ppm, giving us a large span of uncertainty, and only .95ppm difference between the means. That support the evidence of uncertain result. From that test we can`t say if the treatment group is better than the control group.

HDI showed that 95% of the posterior estimates that were generated for the main difference ended up in range from 0.0193 to 16.2 (most likely value for the main difference). The mean difference here was bigger: 8.2ppm on average. That HDI doesn’t overlap with 0 so that is an evidence for credible difference which contradict with the results from the t-test. 97.2% of the cases falls in the right tail above 16.2 and only 2.8% of the cases are lower than 0.0193. That lower are just 0.3% over the 2.5% (considering two tail test) which is another evidence for difference (we accept the positive direction, mean of control group is higher than the mean of the treatment).

Although the different between control and treatment groups is not statistically significant, the addition of the Bayesian evidence makes it clear that this is a weak result that should only be interpreted cautiously. T-test results might be affected but the existence of extreme values, which lead us to the idea that treatment group is better. Anybody who's going to use this to promote the importance or value of treatment group to has to be very cautious about that and because the pieces of evidence that we have are contradictory and that suggests a very weak result. Using more test to prove the alternative hypothesis that the mean TDS in the treatment group will be lower than the mean TDS in the control group.