Analysis of silicone contamination of batches

Thom Hurks	*0828691*
Puck Mulders	*0737709*
Marijn van Knippenberg	*0676548*
Rik Coenders	*0777004*

May 26, 2016

Contents

1 Introduction

The health inspection has detected a silicone leakage in a machine used in a chemical production process. The leakage was detected in close vicinity of packed product, so the question is whether the product has been affected by the leakage. The machines are regularly inspected, and data on a batch produced before the previous inspection is available and is denoted as batch B0. Batches denoted B1 to B7 have been produced between the previous inspection and the current inspection and may have been affected by the leakage. The data consists of measurements of samples of units from the batches. The silicone that leaked is already part of the product itself, so it will always be present in the measured units to some extent. The goal of this report is to possibly identify the batches that were affected by the silicone leakage. If a batch is affected, we expect the amount of silicone to be higher compared to batch B0 and other unaffected batches. We therefore hypothesize that the affected batches are location shifted such that the mean of the outcome is higher than that of unaffected batches.

In the first section we describe which methods we use to find the affected batches. In the next section we will describe what the results of these methods are. Lastly, we will conclude which of the batches are affected. In this section we will also describe what the weaknesses and risks are of our investigation.

1.1 Reproducible Research

We value the idea of reproducible research, which is publishing data analyses together with the software code, so others can reproduce and verify our findings and possibly extend our research. For this reason this report has been created using Knitr, which is a dynamic report generation tool for the statistical programming language R. The Knitr sourcecode for this report is available from the authors if others are interested in verifying our results.

2 Methods

To decide if any of the batches are affected, we perform both parametric and non-parametric tests. The parametric tests assume that the data is normally distributed, which means we must test for normality and, if necessary, transform the data. Non-parametric tests are generally less powerful than parametric tests, but can be applied regardless of the measure of normality. With parametric tests one must be careful to draw conclusions about the original non-transformed data if the data needs to be transformed in order to meet the normality assumption.

2.1 Exploratory data analysis

Before performing any parametric or non-parametric tests, we will first perform some standard statistical tests to gain insight into the data. As a start, we will plot the data and calculate the mean, standard deviation, variance and median of each batch. Additionally, we will perform a Shapiro-Wilk test and an Anderson-Darling test to see whether the non-transformed data is normal. We need both of these tests, since Shapiro-Wilk is more powerful than Anderson-Darling, but cannot handle ties as well as Anderson-Darling. We have formulated the null hypothesis of both tests as can be found in equation 1:

$$\begin{cases} H_0 : B_i \sim \mathcal{N} & \text{for } 1 \leq i \leq 7 \\ H_1 : B_i \nsim \mathcal{N} \end{cases} \tag{1}$$

We will reject the null hypothesis if the p-value < 0.01. After that, we will perform a Wald-Wolfowitz Runs Test which has as a null hypothesis that each element in the sequence is independently drawn from the same random distribution which is expected if none of the batches are affected by silicones. We reject the null hypothesis if $p < 0.01$.

2.2 Critical value

Most of our hypotheses include testing all seven possibly affected batches. Because of this reason, we have to adjust the commonly used critical value of 0.05 to make sure we do not reject the null hypothesis due to chance. Applying Bonferroni correction gives us a critical value of $\alpha = \frac{0{,}05}{7} = 0,00714$, but since Bonferroni correction is quite conservative we may fail to reject the null hypotheses in cases where we actually want to do so; our goal is to identify the affected batches, so we can be a bit more flexible and raise the alpha value slightly. As such, we will set the critical value to 0.01.
Some tests, however, test data from all batches in one go and for these tests the Bonferroni correction is not necessary. For those tests we will use the original critical value of $\alpha = 0.05$.

2.3 Parametric tests

2.3.1 Investigating normality on transformed data

Some statistical tests assume normality of the data, or require the data to at least approach normality. For that reason we will investigate if the data is normally distributed by inspecting normal QQ-plots and by using two tests on normality: the Shapiro-Wilk test and the Anderson-Darling test. If necessary we will transform the data in such a way that the transformed data is normally distributed or nearly normal. For non-parametric tests that do not assume any specific distribution we will use the original non-transformed data. The null hypothesis of both Shapiro-Wilk and Anderson-Darling can be found in equation 1. We will reject the null hypothesis if the p-value < 0.01. Besides the Anderson-Darling and Shapiro-Wilk tests, we will perform skewness and kurtosis tests to determine whether the normality hypothesis can be rejected. The critical value of a two-sided skewness test is 1.157, while the critical value of the kurtosis is 1.49. We will perform a Jarque-Bera test to see if it is appropriate to reject a normal distribution based on skewness and kurtosis. The null hypothesis is the same as with the Shapiro-Wilk and Anderson-Darling test which can be found in equation 1. We will reject normality if the p-value < 0.01.

2.3.2 T-test

Because batch 0 is produced before the previous inspection where no leakage was discovered, we know that the distribution of the amount of silicones found in each batch should be similar to the amount of silicone in batch 0. To check whether they are similar, we will perform a one-sided t-test. We have chosen a one-sided test, since a leakage in silicones will cause the amount to be higher in the batches and thus we have to test only if the test is violated on one side. To perform a proper t-test, we have to know if the variances of the batches are equal. We check this by applying Barlett's test, which checks whether all the variances of the batches are equal to each other. This can be formulated as can be found in 2:

$$\begin{cases} H_0 : \sigma_0 = \sigma_1 = \sigma_2 ... = \sigma_7 \\ H_1 : \sigma_0 \neq \sigma_1 \neq \sigma_2 ... \neq \sigma_7 \end{cases} \tag{2}$$

We will reject the null hypothesis if p-value < 0.05. Based on these results, we will apply either a equal or unequal t-test. The hypothesis of the t-test is the following:

$$\begin{cases} H_0 : \mu_0 \geq \mu_i & \text{for } 1 \leq i \leq 7 \\ H_1 : \mu_0 < \mu_i \end{cases} \tag{3}$$

If the result of the t-test is less than 0.01, we will reject the null hypothesis.

2.3.3 Variance testing

Besides performing a t-test, we will also perform some homogeneity tests to see whether the data also share the same characteristics. We will perform an F-test to see which batches have a variance similiar to batch 0. This can be formulated as can be found in equation 4:

$$\begin{cases} H_0 : \sigma_0 = \sigma_i & \text{for } 1 \leq i \leq 7 \\ H_1 : \sigma_0 \neq \sigma_i \end{cases} \tag{4}$$

We will reject the null hypothesis if the p-value < 0.01.

2.3.4 Outliers

We will perform a Grubbs outlier test on the log-transformed to check whether there are any indications that one of the measurements has some incorrect value due to measurement errors, for example. We will do this for both outliers on the left and right tail. Our null hypothesis is can be found in equation 5.We will reject the null hypothesis if the p-value < 0.01.

$$\begin{cases} H_0 : \text{Batch}_i \text{ contains no outliers} & \text{for } 1 \leq i \leq 7 \\ H_1 : \text{Batch}_i \text{ contains outliers} \end{cases} \tag{5}$$

We will reject the null hypothesis if the p-value< 0.01.

2.4 Non-parametric tests

2.4.1 Kolmogorov-Smirnov Test

If the non-transformed data does not follow a normal distribution, we will also perform some non-parametric tests that do not assume a normal distribution. Such a parametric test is the Kolgomorov-Smirnov test. For the two-sample version of this test, we need to make the following assumptions:

- The two samples are independent.
- The outcomes are ordinal or numerical.

Since the results of the samples do not depend on each other, the first assumption can be made. Second, the outcomes are all numerical so the second assumption can be made too. We perform a two-sided test to get a general view of whether non-parametric testing picks up on any differences between the distributions of batch 0 and the other batches. Our hypothesis is the following: We will reject the null hypothesis when $p < 0.01$. The Kolmogorov-Smirnov test will not give the exact p-value due to ties in the data, so it could be the case that the p-value resulting from the test somewhat deviates from the real p-value. Because we have picked a somewhat high critical value, some bias of p-values can be handled.

2.4.2 Wilcoxon Rank-Sum Test

A second non-parametric test we will perform is the Wilcoxon Rank-Sum Test, which needs the additional assumption that both distributions are from an ordinal distribution. This implies that the data cannot contain any ties. Since we have some ties in the data, we will not get the exact p-value. Since our value does not contain a substantial amount of ties, the bias of the p-value will not be extremely high. As with the Kolmogorov-Smirnov test, we think that our critical value can handle some bias of the p-values calculated by the Wilcoxon Rank-Sum Test. The hypotheses of the Wilcoxon Rank-Sum test can be found in equation 6:

$$\begin{cases} H_0 : \text{Batch}_0 \text{ and batch}_i \text{ come from the same population} & \text{for } 1 \leq i \leq 7 \\ H_1 : \text{Batch}_0 \text{ and batch}_i \text{ come from a different population} \end{cases} \tag{6}$$

This null hypothesis indicates that the change that the median of batch_0 is larger than the median of batch_1 is the same as that the median of batch_0 is smaller than the median of batch_1. We will reject the null hypothesis if p-value < 0.01.

3 Results

3.1 Exploratory data analysis

The results of the standard data analysis can be found in the table below.

	mean	median	variance	st.dev
B0	2.65	1.08	9.16	3.03
B1	2.53	1.72	2.55	1.60
B2	8.05	4.76	63.77	7.99
B3	7.17	4.88	29.80	5.46
B4	1.07	0.96	0.22	0.47
B5	3.05	2.32	3.12	1.77
B6	2.58	2.04	4.59	2.14
B7	4.49	3.92	9.41	3.07

We also discovered that batch seven has a missing value ("NA"). Some of the statistical tests that we plan to perform do not require that the number of data points between batches is the same, and since there is only a single missing value in the whole dataset, we can simply remove the missing value in those cases. The next step is to generate a normal QQ plot for batch 0 as an easy visual confirmation of normality.

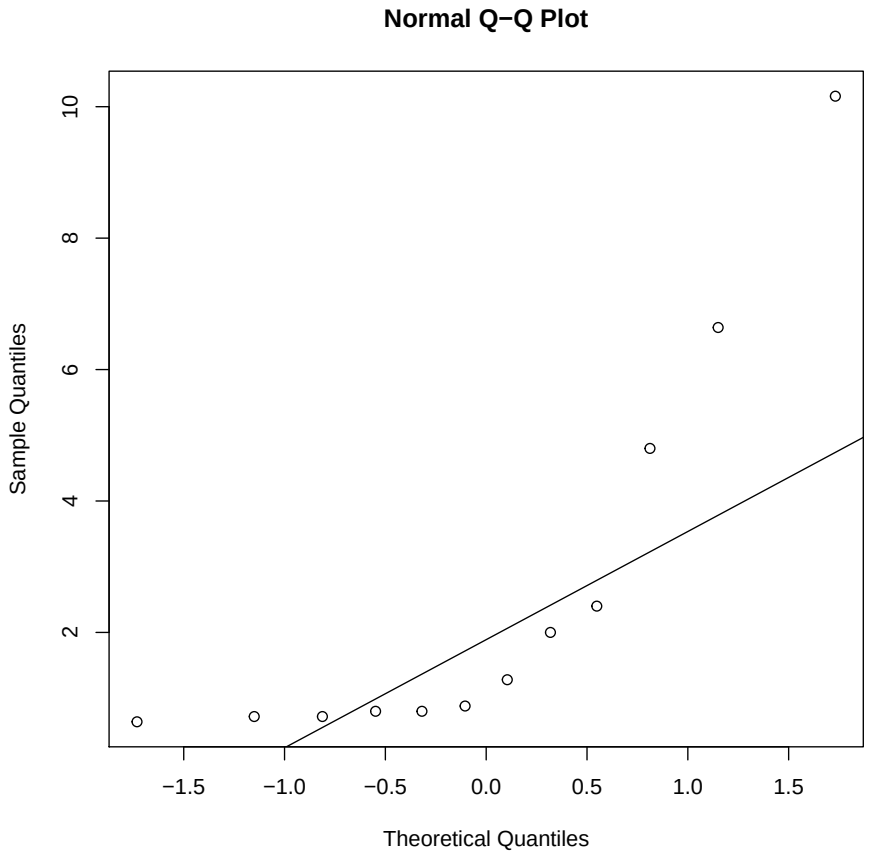

Figure 1: QQ Plot of batch 0

The normal QQ plot already seems to suggest that batch 0 is not normally distributed. When we also perform the Shapiro-Wilk normality test we can see the p-value $= 0.0013049 < 0.01$ so we can reject the null-hypothesis that the outcome is normally distributed.

Next we can generate a normal QQ plot for all batches 0 to 7. Of course generating this plot for multiple batches at once assumes that the units are somewhat consistent across different batches.

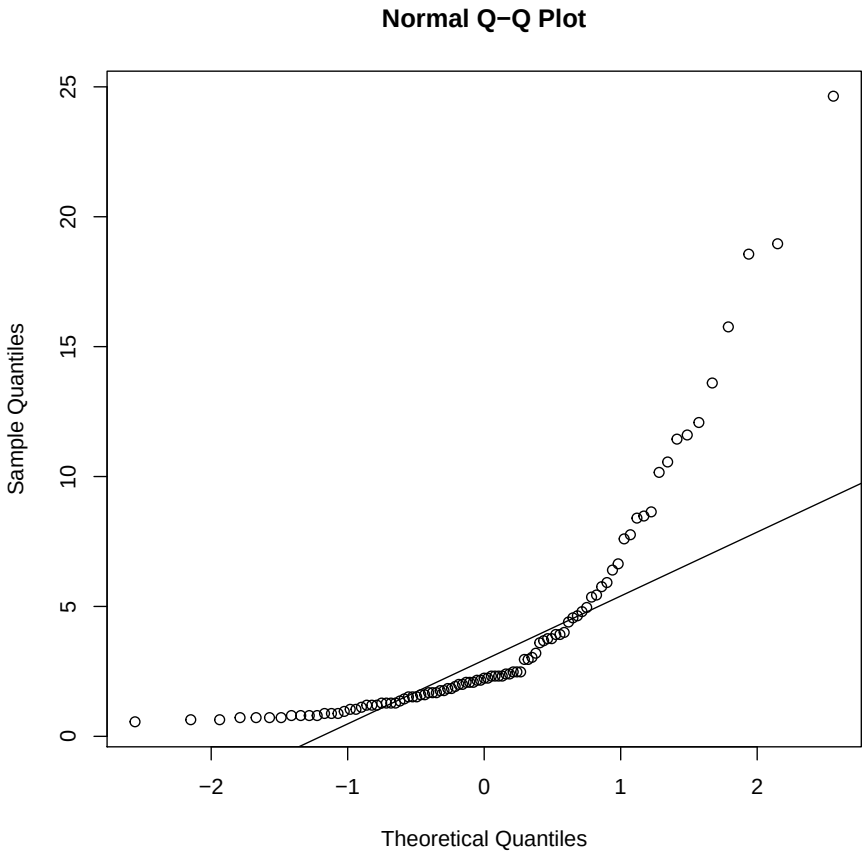

Figure 2: QQ Plot of batches 0 to 7

The normal QQ plot here too seems to suggest that the data is not normally distributed. We can again compute the Shapiro-Wilk test p-value for all batches to gather more information.

	p.value
B0	0.0013
B1	0.0113
B2	0.0442
B3	0.1218
B4	0.0759
B5	0.0020
B6	0.0006
B7	0.0865

We can reject the normality null hypothesis for batches zero, five and six with p-values respectively 0.0013049, 0.0019666 and 6.2737479×10^{-4}. We can also observe that for the other batches the p-values are not particularly high, the exception being batch 3 with a p-value of 0.1218305.

Since our data contains some ties and that is a weakness of the Shapiro-Wilk test, we also perform the Anderson-Darling normality test.

	p.value
B0	0.0005
B1	0.0054
B2	0.0512
B3	0.1337
B4	0.1357
B5	0.0009
B6	0.0004
B7	0.1193

With Anderson-Darling we see that we can reject the normality hypothesis for batches zero, one, five and six. The hypothesis for batch two is not rejected, but it must be noted it has a low p-value of 0.0511721

3.2 Parametric tests

3.2.1 Shapiro-Wilk on log-transformed data

In the previous section we observed that we could reject normality for batches zero, five and six on both normality tests and also for batch one using Anderson-Darling. This includes batch zero, which is important because we intend to compare it with other batches. In order to perform parametric tests that assume normality, we can transform the data such that all batches are normally distributed or close to a normal distribution. We transform the data using the well-known log transformation and then check for each batch what the p-values of the Shapiro-Wilk and Anderson-Darling tests are in order to confirm that the transformation has the desired effect.

	p.value
B0	0.0388
B1	0.1414
B2	0.6190
B3	0.6635
B4	0.8979
B5	0.0620
B6	0.2064
B7	0.9643

We cannot reject the normality hypothesis for any of the transformed batches. It must be noted the p-value for batch zero is still somewhat low with 0.0387691

3.2.2 Anderson-Darling on log-transformed data

We also perform the Anderson-Darling test to investigate if the transformation causes the data to be more normally shaped.

	p.value
B0	0.0380
B1	0.0805
B2	0.7584
B3	0.7127
B4	0.8509
B5	0.0438
B6	0.2118
B7	0.9179

Again, we cannot reject the normality hypothesis for any batch, but the p-value for batch zero is again somewhat low with 0.0379542. We can generate a normal QQ-plot to get more insight into the log transformed batch 0.

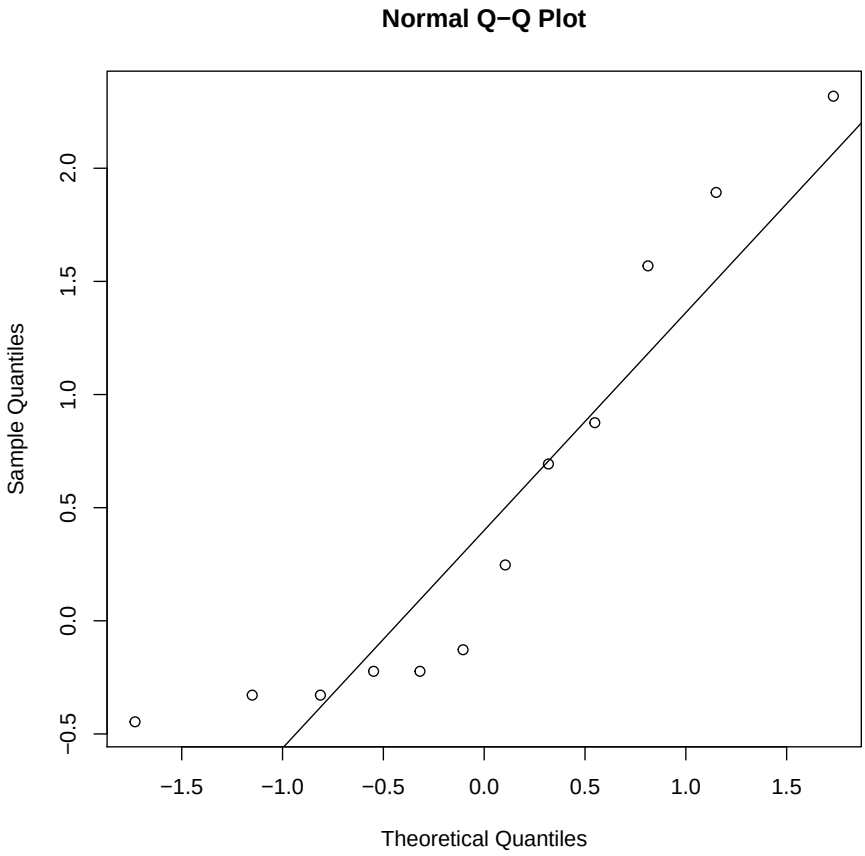

Figure 3: QQ Plot of the log of batch 0

We can visually observe that the normal QQ-plot matches the results of the normality tests; the data seems to be near normality, but deviates enough to justify more testing.

3.3 Skewness and Kurtosis

After performing the Shapiro-Wilk and Anderson-Darling tests, we perform skewness and kurtosis tests to gain more insight in the way the log-transformed data follows normality. In most batches the skewness seems a bit too far away from zero, while in some batches the kurtosis is away from three.

	kurtosis	skewness
B0	-0.703	0.862
B1	-0.773	0.720
B2	-1.306	-0.131
B3	-1.196	-0.062
B4	0.086	0.540
B5	1.123	1.282
B6	1.426	1.094
B7	-0.553	0.130

Kurtosis and skewness tests are also formalized in the Jarque-Bera test, which tests if the skewness is close to zero and the kurtosis close to three.

	Jarque.Bera
B0	0.4640
B1	0.5395
B2	0.6538
B3	0.6898
B4	0.7693
B5	0.2834
B6	0.3850
B7	0.8437

The p-value of the Jarque-Bera test is above 0.01 with a big margin for all batches, so we cannot reject the normality hypothesis based on skewness and kurtosis. Not only do the Shapiro-Wilk and Anderson-Darling tests indicate that assuming normality on the log-transformed data is not rejected, but also Jarque-Bera indicates that the batches having a normal shape cannot be rejected. Therefore the normality assumption for the transformed data is reasonable for the following tests. It must also be noted that the t-test is robust against data that deviates somewhat from normality, so our assumption at least for that test seems to be very reasonable.

3.3.1 T-Test

Performing the Barlett's tests gives us a p-value of $0.004178 < 0.05$. This rejects the null hypothesis that all the batches have equal variance. This implies that we have to perform a t-test that assumes unequal variances. Batch seven has one missing value, which should not be a problem for a t-test.
The results of the t-tests can be found in the following table:

	p.value
B1	0.1986
B2	0.0159
B3	0.0020
B4	0.9395
B5	0.0596
B6	0.2336
B7	0.0145

The results indicate that batch three has a mean that is significanlty different from the mean of the unaffected batch, since its p-value is less than 0.01. This could indicate that this batch is affected. It must also be noted that the hypothesis for batches two and seven is very near being rejected as well.

3.3.2 Variance testing

After performing an F-test that compares all the batches with batch 0, we have found the following results:

	F.test
B1	0.083
B2	0.486
B3	0.605
B4	0.006
B5	0.020
B6	0.152
B7	0.239

These results indicate that batch four has a significantly different variance than the other batches.

3.3.3 Outliers

	min	p.value
B1	1.04	1.00
B2	0.72	0.91
B3	1.52	0.72
B4	0.56	0.80
B5	1.60	1.00
B6	0.96	1.00
B7	1.28	0.51

	max	p.value
B1	5.92	0.34
B2	24.64	0.67
B3	18.56	0.69
B4	2.24	0.14
B5	7.60	0.07
B6	8.64	0.05
B7	11.60	0.35

Both the extreme on the left and the right tail are not significantly big to reject them as an outlier. This indicates that no values can be indicated as errors due mistakes that people maked while measuring the amount of μg in the batches.

3.4 Non-parametric tests

3.4.1 Wald-Wolfowitz Runs Test

We have performed a Wald-Wolfowitz Runs test on the whole original data for which the null hypothesis is that the data is randomly drawn from the same distribution. We see that the p-value $= 7.1225335 \times 10^{-5} < 0.05$ so we can reject that null hypothesis. This is an interesting result, because we know the data is obtained from the same source.

3.4.2 Kolmogorov-Smirnov Test

The results of the Kolmogorov-Smirnov test can be found in the following table.

	batch	p.value
1	B1	0.0996
2	B2	0.2485
3	B3	0.0337
4	B4	0.5176
5	B5	0.0337
6	B6	0.0996
7	B7	0.1134

None of the batches violate the null-hypothesis that the samples are independently and randomly drawn from the same distribution, since all p-values are larger than 0.01. However, we can observe that the p-values for batch three and five are low. Note that the Kolmogovor-Smirnov test is a stronger test than the Wald-Wolfowitz test to detect differences between distributions that differ in location.

3.4.3 Wilcoxon Rank-Sum Test

	batch	p.value
1	B1	0.1181
2	B2	0.0162
3	B3	0.0033
4	B4	0.8306
5	B5	0.0415
6	B6	0.1070
7	B7	0.0227

We can observe that the p-value for batch three $= 0.0033165 < 0.01$ and so for batch three we reject the null hypothesis that the population mean ranks are similar. The other batches do not deviate significantly from batch zero.

4 Conclusion

If we consider the results, we see that batch three fails both the two-sided t-test and the Wilcoxon Rank-Sum Test. Batch three also reports a low p-value on the Kolmogorov-Smirnov test. We have hypothesized that we expect some kind of location shift of the distribution towards a higher outcome. Both the Wilcoxon Rank-Sum Test and t-test are tests that investigate whether such a shift has taken place, and by rejecting the null hypothesis they indicate that batch three indeed has shifted towards a higher amount of silicones in its batch. The Kolmogorov-Smirnov test can also be used to detect location-shifts between distributions, so a low p-value for batch three on that test may also be indicative. Because of these reasons, we advise not to allow batch three to be sold, since we strongly suspect that batch three is affected by the silicones.

Batch two and batch seven are also highly suspicious. Both batches pass the Wilcoxon Rank-Sum Test and t-test with a p-value very close to the critical value. It would be wise to do some further investigation on these two batches to get some more insight into whether the batches are contaminated.

If we consider the other tests, we see that batch four does not pass the homogeneity tests. This implies that the data from batch three has a different variance than batch 0. This is indeed the case, since the variance of batch three is quite low. Because batch four easily passes both the t-test and the Wilcoxon Rank-Sum Test we do not have any concerns that batch four is affected by silicones.

Since all the rest of the batches pass all the tests that could indicate an affection of silicons, we see no reason to state that batch one, two, four, five, six and seven are affected. Those batches should be suitable for sale.

4.1 Weaknesses of our approach

One of the weaknesses of our data analysis is the amount of data. Each batch has very little data, which increases the probability that the data is biased. This could mean that the comparison we have made with batch zero has been made with a highly biased data set. This would result in wrongly accepting or rejecting null hypotheses.

The other weakness of our investigation is that we have only performed some conservative tests. Because we adjusted the critical value, we increase the probability that a type II error occurs. Especially batch two and batch seven could be the victim of a type II error, so it may be useful to obtain more data of batch zero, two and seven to investigate further whether they are acceptable or affected. As an extension of the research we have already performed, it could also be useful to perform an ANOVA test to see if this yields a similar result. We log-transformed the data in order to perform tests that assume normality.

This changes the meaning of the tests with respect to the original data, which is why we also performed non-parametric tests. If more data on the suspected batches is made available, our research can be easily repeated because of our use of the dynamic report generation tool Knitr.