Analysis of silicone contamination of batches

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

May 23, 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.

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.

2 Methods

To decide which of the batches is infected, we are performing both non-parametric tests and parametric tests which assume that the data is normally distributed. Non-parametric tests are generally less powerful than parametric tests, while the parametric tests are not always as reliable if the data is not perfectly normally distributed. As a start, we will plot the data and perform a Shapiro-Wilk test to see whether the non-transformed data is normal. After that, we will perform a Wald-Wolfowitz Runs Test which has as a null hypothesis that each element in the sequence in drawn from the same distribution. If this is not the case, it could be an indicator that some of the batches are affected by silicones. We reject the null hypothesis if $p < 0.05$.

2.1 Normal distribution

To decide whether the data is normally distributed, we first perform a log normal transformation on the data. After the transformation, we will plot the data to get an insight how the data is distributed. We will then perform the Shapiro-Wilk tests on each of the batches to see whether we can reject that the data is from a normal distribution. We will perform a skewness and kurtosis test to see whether batch 0 is in the appropriate range of skewness (between -0,3 and 0,3) and in the appropriate range of kurtosis (between -0,5 and 1,5). $\alpha = 0,05$ is chosen as the critical value. If we can indeed conclude that the batches are normally distributed, we can compare each batch with batch 0. Because batch 0 is produced before the leakage was discovered, we know that the mean of the amount of silicones found in the batch should be similar to the amount of silicone in batch 0. Therefore our hypothesis is the following:

$$H_0 : \mu_0 = \mu_i \text{for} 1 \leq i \leq 7 H_1 \mu_0 \neq \mu_i \text{for} 1 \leq i \leq 7 \tag{1}$$

If the result of the t-test is below 0,05, we will reject the null hypothesis.

Besides a t-test, we will also perform some homogeneity tests, to see whether the data also share the same characteristics. To check if this is the case, we will perform both a Barlett's test, which checks whether all the variances of the batches are equal to each other. The null hypothesis is that all variances are equal (and thus that the sets are homogeneous). We will also perform a F-test to see which batches have similar properties with batch 0.

2.2 Non-parametric tests

Kormogov We will perform a serial correlation test on the means of each batches. The serial correlation test is Von Neumann with Lag-1 Autocorrelation. As a null hypothesis we will assume that the means are random. If they are random, than there will be no clear pattern. This is expected if no of the batches will be affected. We will reject the null hypothesis is $p < 0.05$.

3 Results

3.1 QQ Plot and Shapiro-Wilk

First, the data has been transformed to The first step is to generate a normal QQ plot as an easy visual confirmation.

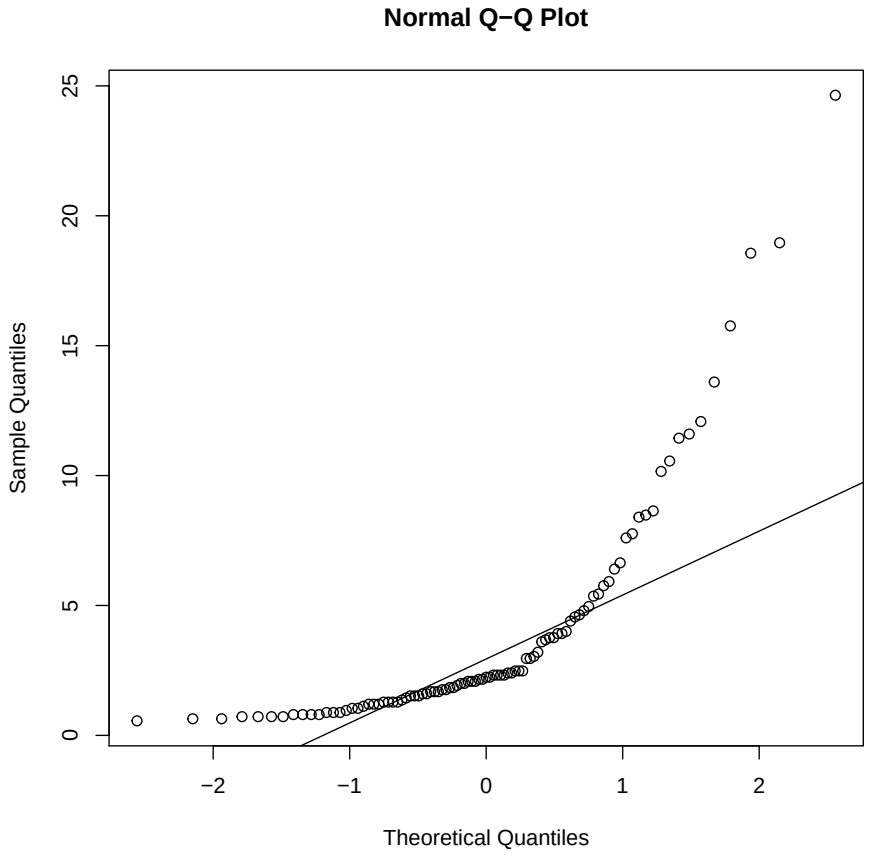

Figure 1: QQ Plot

The normal QQ plot already seems to suggest that the data is not normally distributed. When we also perform the Shapiro-Wilk normality test we can see the p-value $= 1.1536883 \times 10^{-12} < 0.05$ so we can reject the null-hypothesis that the outcome is normally distributed.

3.2 Normal distribution

3.2.1 Shapiro-Wilk, skewness and kurtosis

After performing a log normal transformation, the results of the Shapiro-Wilk test can be found in the table below.

Batch	P-value
B0	0.03876912
B1	0.1414101
B2	0.6190074
B3	0.6635468
B4	0.8978743
B5	0.06197854
B6	0.2063684
B7	0.9643288

Only the p-value of batch 0 is lower than 0,05, which means that only in that batch normality is rejected according to the Shapiro-Wilk test, but normality is not rejected with a very low p-value. When we perform the skewness and kurtosis test, we get values of 1,36 and 0,51 respectively. This means that although the kurtosis is in the appropriate range to be normally distributed, the skewness of the data is a bit too high. The data is skewed right, which probably is the reason that the Shapiro-Wilk test rejects a normal distribution. The p-value of the Shapiro-Wilk test is not very low and the results of the skewness and kurtosis tests are not very extreme, in combination with the fact that all the other batches are normally distributed, we will assume a normal distribution for the following tests.

3.2.2 T-Test

To see which t-test we have to perform, we have analysed the variance of all the batches. The variance differs from each batch, as can be see in the table below. Because of that, we will perform a t-test that assumes unequal variances.

Batch	P-value
B0	0.9399675
B1	0.3146983
B2	1.446813
B3	0.6830243
B4	0.1557867
B5	0.2116048
B6	0.3831803
B7	0.4372429

The results of the t-test can be found in the following table:

Batch	P-value
B1	0.3972427
B2	0.03181353
B3	0.004011126
B4	0.1210506
B5	0.1191315
B6	0.4672437
B7	0.02896952

The results indicate that batches 2, 3 and 7 have a mean that is different from the mean of the unaffected batch. This could indicate that these batches are affected.

3.2.3 Homogeneity tests

Performing the Barlett's tests gives us a p-value of $4,523 * 10^{-15}$. This rejects the null hypothesis that all the batches have the same variance. After performing an F-test that compares all the batches with batch 0, we have found the following results:

Batch	P-value
B1	0.04429337
B2	0.003227914
B3	0.06260923
B4	$4.570405 * 10^{-7}$
B5	0.08812147
B6	0.266533
B7	0.9583088

These results indicate that batch 1, 2, and 4 have different properties than batch 0.

3.3 Non-parametric tests

3.3.1 Serial correlation tests with means

We have performed a Rank von Neumann Test for lag-1. The p-value is $0.4895337 < 0.05$ and we therefore have to reject the null hypothesis. This means that the serial correlation test detects some pattern in the means of the batches.

3.3.2 Wald-Wolfowitz Runs Test

We have performed a Wald-Wolfowitz Runs test. 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, but part of the samples may or may not be tainted, and this result may point to that being true.

3.3.3 Rank Serial Correlation Test

Here we perform the rank serial autocorrelation test at lag 1 using the rank von Neumann ratio. This tests the null hypothesis that the lag-k autocorrelation is 0 for all values of k greater than 0 (i.e., the time series is purely random). The procedure that we use emits some warnings because the dataset contains ties, and with this specific test that can make the p-value less accurate. With that warning in mind, we present the table containing the rank serial autocorrelation test p-values of each batch:

	batch	p.value
1	B0	0.99
2	B1	0.13
3	B2	0.26
4	B3	0.36
5	B4	0.68
6	B5	0.21
7	B6	0.49
8	B7	0.37

As one can observe, the p-values are all large, which means that we cannot reject the null hypothesis that the batches are random using this test.

4 Results