Eindhoven University of Technology, Department of Mathematics and Computer Science

2DMT00 - Applied Statistics
Assignment 1

Niels Hellinga
Student number: 0977787
n.hellinga@student.tue.nl

Mingpeiyu Zhang
Student number: 1018903
m.zhang.4@student.tue.nl

Vishal Chouksey
Student number: 1034346
v.k.chouksey@student.tue.nl

January 10, 2017

Contents

1 Introduction

In a chemical factory, regular inspections where conducted (hourly) on different sites for safety reasons. The substance is measured on parts per million (PPM) and values below 0.1 mg/unit cannot be detected. After the occurrence of an accident in the factory, concerns have been risen about the safety of the different sites. Each site could be affected by a toxic substance. If this is the case, the level of PPM values has risen to an alarming amount. Measurements where conducted using 2 sensors for each sites. There are 7 sites in total that could be contaminated. A statistical approach can provide an effective method to determine the affected sites.

In this report, a study has been conducted to test which of the 7 sites are contaminated based on the values that where measured before the accident. The measurements before the accident are considered to be normal values and can be used as a reference point. Hence, if the PPM value of a site, after the accident is significantly higher than the before value, the site is considered contaminated and therefore not safe.

The main research question to be answered is:

Have any of the sites been contaminated after the accident?

The main question can be split up in two sub-questions. "Are the measured PPM values at the different sites worst than the normal values?" and "if so, which sites are affected?"

In order to answer these questions. We will first determine a hypothesis H_0 and from that we can determine H_1. We will do some exploratory data analysis to get to know the data set and make our first few assumptions about outliers and normality. After the exploration of the data set, we will use mature and sufficient statistical methods to determine outliers and check for normality. Once we know if the data is normally distributed, we will use appropriate (non-)parametric tests to see if the variance and locations of each site show differences to the Site Before which will then answer our main research question.

1.1 Hypothesis

In this report it is important to prove which of the site is affected and which are not affected. To get a clear understanding we have formulated mathematical expression for the hypothesis:

$$\begin{aligned} H_0 &: \textit{Site is not affected} \\ H_1 &: \textit{Site is affected} \end{aligned} \tag{1}$$

As explained before, we first need to check if our data is normally distributed to determine if we can use parametric or non-parametric tests. These parametric tests are then used to see if a sites variance and location are different to that of the Site Before and hence we can see for which sites we have to reject the H_0.

2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is typically used as a first approach to get to know a certain data set. EDA will provide us with informal, descriptive and graphical representation to explore and understanding the data set. Tools we will use in this section are numeric summary statistics (e.g. mean and median), measures of the spread (e.g. std, var, IQR) and shape (e.g. skewness and kurtosis), as well as graphics (e.g. box plots, histogram and probability plots). These methods will provide us the chance to check the quality of the data set and choose good statistical methods. It will also allow us to confirm if the assumption we made, used by the methods are met.

2.1 Summary Statistics

Table 1 and Figure 1 give an overview of the summary statistics grouped by Site. In the table, the column named 'Site' is the reference of what site we are looking at. We can see that 60 observations are missing in Site 7 compared to the other sites. When we look into the dataset, we found the last hour of Sensor 1 in Site 7 is missing.

Site Before represents the overall measurements of site before the accident, the normal values. Hence, these are the values we need to use as a reference case to test our hypothesis.

Table 1: Summary Statistics by Site

Site	**n**	**min**	**max**	**mean**	**median**	**sd**	**var**
All	3378	0,000	25,681	3,748	2,241	4,063	16,506
Before	120	0,036	10,164	2,132	1,420	2,128	4,526
Site1	480	0,000	10,045	2,786	2,098	1,989	3,958
Site2	480	0,000	4,482	1,154	1,019	0,873	0,761
Site3	459	0,140	19,567	5,790	3,691	5,079	25,799
Site4	459	0,000	25,681	7,917	7,076	6,265	39,250
Site5	480	0,496	8,452	2,619	2,234	1,486	2,209
Site6	480	0,000	16,482	3,019	1,940	3,286	10,801
Site7	420	0,289	12,873	3,608	3,105	2,272	5,164

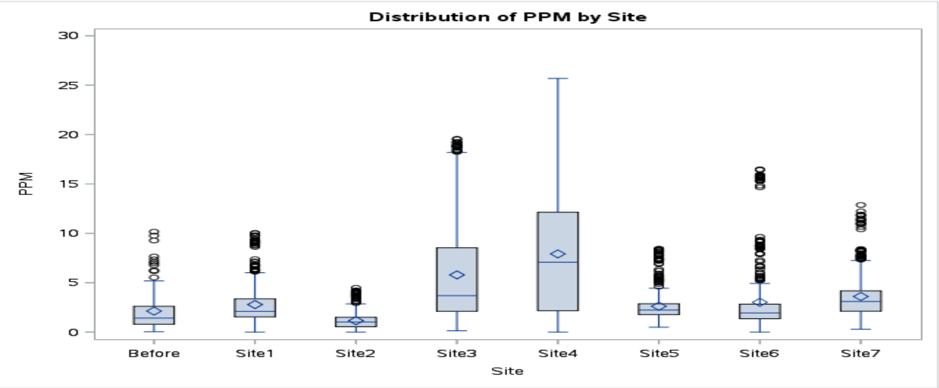

Figure 1: Box Plot: PPM by Site

2.2 Mean and Standard Deviation

Reference: The mean PPM value of the site Before is 2.132 mg/unit with a standard deviation of 2.128 mg/unit.

Looking at the box plot of the summary statistics in Figure 1, we can quickly compare the site Before values to the other sites 1...7. Notice that Site 3 and 4 pop out from the group. They have a bigger spread and we can look at Table 1 to see these exact values. Site 3 has a mean of 5.790 with a standard deviation of 5.080 and Site 4 has a mean of 7.917 with a standard deviation of 6.265. Notice that the means are much higher than the normal values from site Before. However, due to the relatively high standard deviation we cannot say for certain that these two sites are negatively affected by the accident. We need further analysis to see and proof out findings so far about Site 3 and 4.

Taking a look at Site 2, it has a mean of 1.154 and a standard deviation of 0.872. Comparing this to the site Before it could suggests that Site 2 is not affected after the accident since the mean value is lower and also the standard deviation is relatively small. Again, we need futher analysis to test these findings.

2.3 Mean vs. Median

Another observation that we can make is that the mean and median for all sites are relatively close together except for Site 3. Site 3 has a mean of 5.790 and median of 3.691. Notice that the mean is bigger than the median and that the difference is more than 2 mg/units. This difference suggests that there is a 'blob' at the right side of the distribution that increases the mean value. Is this an outlier or do we have valid data that suggest something different? Further investigation needs to be conducted to test this.

Note that the mean is for all sites higher than the median.

2.4 Minimum and Maximum values

Reference: The minimum (min) PPM value of the site Before is 0.036 and the maximum (max) is 10.164.

If we take a look at the min and max values of the sites, we notice that Site 1,2,4 and 6 have a min value of 0. Since the sensors used to measure the PPM values cannot detect values below 0.1, we can either assume that the PPM values have been less than 0.1 for these sites or that the sensors missed some readings. It is also possible that it is a human error. Since it does appear that the sensors did not work at these stages, we will exclude all off the values that are lower than 0.1 from the data set. We also do this since these values might interfere with other tests and if we use want to use a log transformation, we cannot take log of the 0 values.
As already discussed in the mean and standard deviation section above, site 3 and 4 have a bigger spread than the rest. This is also shown by the high max values of the sites. Site 3 has a max value of 19.6 (min of 0.14) and site 4 a max value of 25.7 (min of 0).

2.5 Outliers

The box plot, Figure 1 shows us a lot of (theoretical) outliers. So many that we need to do in depth analysis of these 'outliers' using mature and sophisticated tools. These tests will be conducted later on in this report. For now we already have a assumption that these are not outliers and that the distribution is not normal.

2.6 Spread and Shape

The spread of the data set has already been discussed. To summaries the spread, site 3 and 4 have the biggest spread with a variance of 25.8 and 39.2 respectively. The shape of the data set can be seen in probability plots, Figure 2. Looking at these plots, it appears that the data set for each Site is not normally distributed. To proof this hypothesis, we will use adequate statistical measure to determine the distribution of the sites later in this reports.

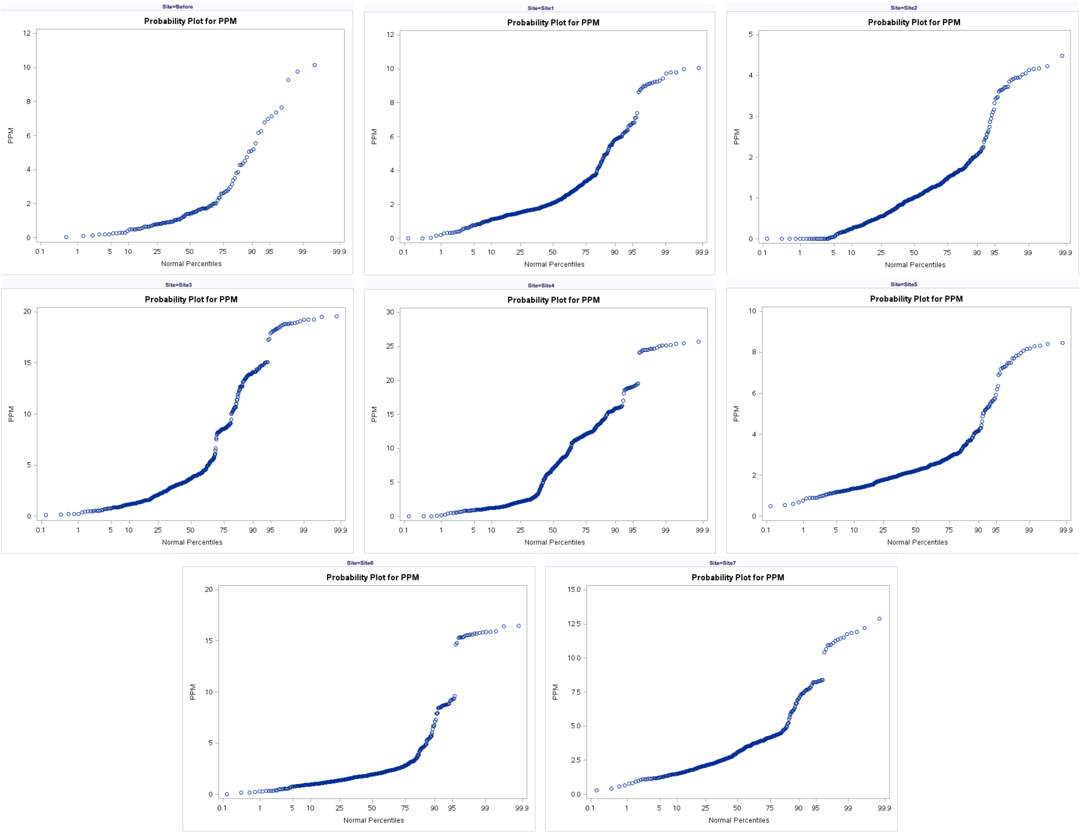

Figure 2: Probability Plot: PPM by Sensor

3 Paired sample test

Before further tests we have to check if the paired samples (Sensor 1 and Sensor 2) have the same distribution. We need to test this in order to make a decision whether to group by site only or that we need to take sensors into account. Hence, if the sensors do not have the same distribution, we need to group by Site by Sensor.

3.1 Preprocess the data

In the summary statistics section we found that there are many zero values and many observations are below 0.1. Due to the fact that values below 0.1 mg/unit cannot be detected, we deleted all the observations whose PPM values is smaller than 0.1. We use the processed data for the rest of our tests.

3.2 Testing for the paired samples

Data is considered paired whenever two measurements are taken on the same unit. In our data set, we have two measurements (Sensor 1 and Sensor 2) of the same variable (ppm value). Besides, the ppm value in the data set is continuous, not binary. There are several tests to test on this scenario: Paired t-test, Sign test and Wilcoxon signed rank test. These three tests are implemented in SAS in the same procedure, we can therefor run them at the same time.

The data for each site is $(y_{11}, y_{21}), (y_{12}, y_{22}), \ldots, (y_{1n}, y_{2n})$. We take the difference $z_j = y_{2j} - y_{1j}$ or log difference $z_j = \log(y_{2j}) - \log(y_{1j})$ or ratio $z_j = (y_{2j}/y_{1j}) - 1$. Then we use μ_0 to denote the mean of z_j. Here we have formalized the following hypothesis.

$$\begin{aligned} H_0 :& F_1 = F_2 \rightarrow \mu_0 = 0 \\ H_1 :& F_1 \neq F_2 \rightarrow \mu_0 \neq 0 \end{aligned} \tag{2}$$

We use the univariate procedure in SAS and the results of the tests are shown in the appendix in Table 11, Table 12 and Table 13. Here we do not take the normality assumption since the summary analysis shows that most of the samples may not be normally distributed. Hence, we do not look at the result of Student's t-test. For the Sign test and Wilcoxon signed rank test, if the p-value is smaller than 0.05 (chosen value for alpha), we reject the null hypothesis.

3.3 Conclusion

The result of variable 'diff' shows that we can not reject the null hypothesis of Site Before, Site 1 and Site 3, which means that the two sensors in these sites have the same distribution.

The result of variable 'zl' shows that we cannot reject the null hypothesis of Site Before, Site 1 and Site 3. And in the result of variable 'zr', the results of Sign test and Wilcoxon signed rank test are different in many Sites. For this reason we choose not take the variable 'zr' into consideration. The result of variable 'zl' is the same as the variable 'diff'.

Based on the results of the test, we conclude that the the two sensors in Site Before, Site 1 and Site 3 have the same distribution, while the distribution of the two sensors in the other sites are significantly different. Since the two sensors in Site Before, Site 1 and Site 3 have no difference, we merge them together in the further tests, which means that there will be only one sensor in Site Before, one in Site 1 and one in Site 3.

4 Outliers Test

An outlier is an observation that appears to deviate markedly from other observations in the sample. We have to exclude the outliers in the dataset for further tests. There are many tests to find possible outliers in the data. In order to test for outliers: Dixon test, Hampel test, Tukey test, Grubbs test and Doornbos test. Dixon test uses the differences between ordered values instead of standardized/ studentized values. Since there are hundreds of observations in each sample set, we do not have sufficient critical values. Hence, it is not possible to use this to find outliers using this test. All the other mentioned tests can be used.

4.1 Hampel and Tukey Test

Both Hampel Test and Tukey Test assume that the sample data is normally distributed. However, as we have seen in Figure 1, if we assume the samples have normal distribution, then we will have too many (theoretical) outliers when applying the Tukey Test. This can be explained by a the skewness to the right, which means that the data is not centered around the median but has some ‘blob’ at the right hand side (higher levels). Since there are so many of them, we cannot conclude that they are outliers and for this reason we should not delete them but try different methods.

Table 2: The outliers from Hampel test

	Site	**Hour**	**Sensor**	**PPM**	**Z**
1	Before	-1	1	6.246611915	3.7332968387
2	Before	-1	1	6.773781709	4.2275373116
3	Before	-1	2	6.178422734	3.6693670509
4	Before	-1	2	6.962093956	4.4040867556
...	...	...	...	...	...
578	Site7	4	2	6.098761949	3.5946822254
579	Site7	4	2	6.881808523	4.3288163003
580	Site7	4	2	7.040019105	4.4771443604

Table 3: The outliers from Turkey test

	Site	**Hour**	**Sensor**	**PPM**
1	Before	-1	1	6.246611915
2	Before	-1	1	6.773781709
3	Before	-1	1	5.531590706
4	Before	-1	2	6.178422734
...	...	...	...	...
246	Site7	2	2	7.499478835
247	Site7	2	2	8.248157024
248	Site7	2	2	7.252841178

4.2 Grubbs Test

Grubbs test, also known as the maximum normed residual test or extreme studentized deviate test, is a statistical test used to detect outliers in a univariate data set assumed to come from a normally distributed population. Table 4 shows all the outliers found for this test. However, since this test assumes a normally distributed data set, we cannot take these potential outliers out of the dataset because as we have see from the analysis so far, the data does not appear to have a normal distribution. Hence, we also applied the Doornbos Test.

Table 4: Outliers using Grubbs Test

Site	**Hour**	**Sensor**	**PPM**	**ID**	**mean**	**sd**	**var**	**n**	**G_value**	**cutoff**
Before	-1	1	9,767	14	2,063	1,862	3,465	60	4,138	3,027
Before	-1	2	10,164	69	2,200	2,378	5,655	60	3,349	3,027
Site7	2	1	12,873	3136	3,251	2,716	7,376	180	3,543	3,400

4.3 Doornbos Test

The Doornbos Test is in this case the best option to find outliers. Doornbos used the externally studentized values to investigate the existence of a single outlier.

We have calculated the critical values for each site based on its sample size (n). For each calculated Doornbos value (D_value), we check if it is bigger than the critical value. Table 5 shows us all the detected outliers based on the Doornbos test. Compairing these findings with the once in Table 4, we see that for Site Before, we get the same outliers (see 'ID'). These values will be excluded from the data set as they might interfere with the normality tests.

Table 5: Doornbos Test Outlier

Site	Hour	Sensor	PPM	ID	mean	sd	var	n	D-value	D-criteria
Before	-1	1	9.77	14	2.13	1.86	3.44	58	4.93	3.52
Before	-1	2	10.16	69	2.20	2.38	5.65	60	3.73	3.53

4.4 Conclusion

Based on above discussion, we choose only to take the the testing result of Doornbos test. Since all the other tests assume that the samples come from a normally distributed population which we do not know so far. The results of other tests can not be trusted. We then remove the detected outliers listed in Table 5 and use the processed dataset for further tests.

5 Normality Test

In order to answer the question “Are the data normally distributed?”, we do the normality test on the previously processed dataset.

5.1 Test for normality

In the probability plots discussed above we saw that the data does not appear to be normally distributed, grouped by site. Hence, we will check the normality by sensor. We do this also based on the findings of the paired sample section discussed above where we noticed that not all sensors (per site) have the same distribution. For the sites that have sensors with the same distribution we will combine them. Having set our strategy, we have formalized our hypothesis. Equation 3 displays the hypothesis.

$$\begin{aligned} &H_0 \text{ :Sensor is normally distributed} \\ &H_1 \text{ :Sensor is not normally distributed} \end{aligned} \tag{3}$$

H_0 is defined such that the sensor (1 and/or 2) is normally distributed. We will use several methods (Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling) to test for normality. If the p-value (for each test) is less than our chosen alpha of 0.05 (95% confidence), then H_0 is rejected and there is evidence that the sensor is not normally distributed. Vise verse, if the p-value is greater than 0.05 (alpha) we will not reject the H_0 and the sensor will be normally distributed with a confidence level of 95%.

Table 6: Normality test result (Site before,1,3)

Site Before (Sensor 1, Sensor 2)		
	P-Value	P-value(Log)
Shapiro-Wilk	<0.0001	0.2689
Kolmogorov-Smirnov	<0.0100	>0.1500
Cramer-von Mises	<0.0050	>0.2500
Anderson-Darling	<0.0050	>0.2500

Site 1 (Sensor 1, Sensor 2)		
	P-Value	P-value(Log)
Shapiro-Wilk	<0.0001	<0.0001
Kolmogorov-Smirnov	<0.0100	<0.0100
Cramer-von Mises	<0.0050	<0.0050
Anderson-Darling	<0.0050	<0.0050

Site 3 (Sensor 1, Sensor 2)		
	P-Value	P-value(Log)
Shapiro-Wilk	<0.0001	<0.0001
Kolmogorov-Smirnov	<0.0100	<0.0100
Cramer-von Mises	<0.0050	<0.0050
Anderson-Darling	<0.0050	<0.0050

Table 7: Paired test result for Site 2, 4, 5, 6, 7

Site 2				
	Sensor 1		Sensor 2	
	P-Value	P-value(Log)	P-Value	P-value(Log)
Shapiro-Wilk	<0.0001	0.0017	0.0003	<0.0001
Kolmogorov-Smirnov	<0.0100	0.0606	<0.0100	<0.0100
Cramer-von Mises	<0.0050	0.0673	0.0155	<0.0050
Anderson-Darling	<0.0050	0.0213	<0.0050	<0.0050

Site 4				
	Sensor 1		Sensor 2	
	P-Value	P-value(Log)	P-Value	P-value(Log)
Shapiro-Wilk	<0.0001	<0.0001	<0.0001	<0.0001
Kolmogorov-Smirnov	<0.0100	<0.0100	<0.0100	<0.0100
Cramer-von Mises	<0.0050	<0.0050	<0.0050	<0.0050
Anderson-Darling	<0.0050	<0.0050	<0.0050	<0.0050

Site 5				
	Sensor 1		Sensor 2	
	P-Value	P-value(Log)	P-Value	P-value(Log)
Shapiro-Wilk	0.5163	<0.0001	<0.0001	0.0113
Kolmogorov-Smirnov	>0.1500	<0.0100	<0.0100	0.0177
Cramer-von Mises	>0.2500	<0.0050	<0.0050	0.0244
Anderson-Darling	>0.2500	<0.0050	<0.0050	0.0130

Site 6				
	Sensor 1		Sensor 2	
	P-Value	P-value(Log)	P-Value	P-value(Log)
Shapiro-Wilk	<0.0001	<0.0001	<0.0001	<0.0001
Kolmogorov-Smirnov	<0.0100	<0.0100	<0.0100	<0.0100
Cramer-von Mises	<0.0050	<0.0050	<0.0050	<0.0050
Anderson-Darling	<0.0050	<0.0050	<0.0050	<0.0050

Site 7				
	Sensor 1		Sensor 2	
	P-Value	P-value(Log)	P-Value	P-value(Log)
Shapiro-Wilk	<0.0001	<0.0001	<0.0001	0.0004
Kolmogorov-Smirnov	<0.0100	0.0877	<0.0100	<0.0100
Cramer-von Mises	<0.0050	0.0062	<0.0050	<0.0050
Anderson-Darling	<0.0050	<0.0050	<0.0050	<0.0050

Looking at the Table 6 and Table 7 and only taking the ‘normal’ PPM values into account for now (not the logarithmic values), we notice that for Site 5 sensor 1 we cannot reject the H_0. Hence, this sensor is normally distributed. All the other sites do not appear to be normally distributed. Hence, we can reject the H_0 for them.

Due to so few normality cases, we decided to transform the data using a logarithmic transformation. Applying the same methods again, we can now see the results in the Table 6 and Table 7 when looking at the LOGPPM values. Here, we do see some cases for which we cannot reject the H_0. For example, Site Before now is normally distributed since we cannot reject the H_0 for any method used. Site 5 sensor 1 however went from a normally to a not normal distribution when applying the transformation. We do have a few new cases where we see the p-value greater than 0.05. However, these p-values are very close to the chosen alpha and since the other methods still do reject H_0, we still have to assume that the data is not normally distributed.

5.2 Conclusion

We have two cases where we can detect normality, namely Site Before (sensor 1+2) with the transformed LOGPPM values and Site 5 for Sensor 1 with the 'normal' PPM values. Since the rest is not normally distributed and we cannot compare Site 5 to Site Before, we will need to use non-parametric tests. We need non-parametric tests for further analysis to answer the main question of 'which site is worst than normal'. We cannot use parametric tests since they all assume normality and in our case we cannot compare two sites with one another since both do not have normal distributions for the same (transformed) PPM values.

6 Location and Variances

We have now arrived at the final stage where we want to answer the question "Do the PPM levels of the sites have different 'locations' and 'variances' compared to the PPM levels before the accident ?". First, we have to do homogeneity tests to test for the variances, then do some other tests for the location shift.

6.1 Homogeneity test

F-test and Bartletts test require normality assumption, the former is the optimal test under normality. Previous results, shown in Table 6 and Table 7 reveal that observations in most of the sites and the sensors are significantly deviate from the normally assumption. Hence, F-test and Bartletts test are not appropriate test for this data set. Krutosis is also not a good choice because it relay's on the visual representation of histograms. Next, the robust version of Bartletts (adjusted for kurtosis) is not a suitable choice either. Therefore, Levene's and BF tests, which do not require normality assumption, are the logical choices. From the lecture, we know that BF is more powerful to detect variance differences.

As the results shown in Table 14 and Table 15 of the appendix, equality of variances are rejected for Site 2, Site 3, Site 4 and Sensor 1 of Site 5 based on BF version of Levenes test because theirs ProbF value are smaller than 0.05. However, we can not reject the null hypothesis of Site 1, Site 5 (Sensor 2), Site 6 and Site 7. Hence, we can conclude that Site 1, Site 5 (Sensor 2), Site 6 and Site 7 do not have different variances compared to the PPM levels before the accident (Site Before).

6.2 Location test

Three tests can be applied to see if the location of two samples are the same: t-test, Wilcoxon rank sum test and Kolmogorov-Smirnov test. However, t-test requires normality assumption. Since the observations in most of the sites and the sensors significantly deviate from the normally assumption, we do not use the t-test. Next, we will show our test procedures for the Wilcoxon rank sum test and Kolmogorov-Smirnov test.

6.2.1 Wilcoxon rank sum test

Assumptions of Wilcoxon rank sum test:

- Each sample has been randomly collected from the population it represents
- The two samples are independent of one another
- The original variable observed is a continuous random variable
- The underlying distributions from which the samples are derived are identical in shape

The last assumption means that the variance of the two samples should be not deviate from each other. From previous homogeneity test we only have the conclusion that the Site 1, Site 5 (Sensor 2), Site 6 and Site 7 do not have different variances compared to the PPM levels before the accident (Site Before). Hence, we can only perform Wilcoxon rank sum test on these sites (sensors).

In order to perform Wilcoxon rank sum test, firstly we have to do randomness test on the dataset. The tests for randomness are (un)conditional runs, autocorrelation and serial correlation. Autocorrelation, in contrast to (un)conditional runs test and serial correlation, is not a rank-based test so it is more powerful. But it should be used with caution because it requires normality assumption. Since all the above results show that most of the samples are not normally distributed, we cannot use this test. While for the (un)conditional runs test, due to the sample size, we cannot run them either. Hence we choose to use serial correlation test on Site Before, Site 1, Site 5 (Sensor 2), Site 6 and Site 7.

Table 8: The result for the serial correlation test

Result for site before						
Obs	ETA	MU	SIG	STAT	C_Oneside	C_Twoside
1	-0.078263	-.00862069	.008485308	-0.75603	1.64485	1.95996

Result for site 1						
Obs	ETA	MU	SIG	STAT	C_Oneside	C_Twoside
1	0.32743	-.001686341	.001681210	8.02665	1.64485	1.95996

Result for site 5 sensor 2						
Obs	ETA	MU	SIG	STAT	C_Oneside	C_Twoside
1	0.55318	-.004166667	.004135231	8.66706	1.64485	1.95996

Result for site 6 sensor 1						
Obs	ETA	MU	SIG	STAT	C_Oneside	C_Twoside
1	0.29758	-.0041841	.004152401	4.68298	1.64485	1.95996

Result for site 6 sensor 2						
Obs	ETA	MU	SIG	STAT	C_Oneside	C_Twoside
1	0.10232	-.004166667	.004135231	1.65592	1.64485	1.95996

Result for site 7 sensor 1						
Obs	ETA	MU	SIG	STAT	C_Oneside	C_Twoside
1	0.48831	-.005555556	.005499563	6.65959	1.64485	1.95996

Result for site 7 sensor 2						
Obs	ETA	MU	SIG	STAT	C_Oneside	C_Twoside
1	0.50772	-.004166667	.004135231	7.96018	1.64485	1.95996

As the result shown in Table 8, the statistics for Site Before and Sensor 2 of Site 6 are in the range of -1.65 and 1.96, which means that there is not enough evidence to reject the null hypothesis (H0: there is no serial correlation (algorithm) among observations) in favour of one-sided or two-sided hypothesis because the statistic is smaller than the critical value for two sided (1.96), and larger than the (left) one-sided critical value (-1.65). However for the other Sites (Sensors), we can reject the null hypothesis. Hence, the sample of Site Before and Sensor

2 of Site 6 are random. Site 1, Sensor 2 of Site 5, Sensor 1 of Site 6 and Site 7 are not random.

Since most of the samples either do not have same variance or are not random, we choose not to use Wilcoxon rank sum test.

6.2.2 Kolmogorov-Smirnov test

For Kolmogorov-Smirnov test, it only requires that: the two samples are independent and the outcomes are ordinal or numerical. Since our dataset satisfies these two conditions, we can apply Kolmogorov-Smirnov test.

SAS outputs are provided as following. It calculates an asymptotic version of statistic and reports it as KS. Due to the relatively large sample size, Monte Carlo estimation is applied for the KS exact test.

Table 9: Kolmogorov-Smirnov Two-SampleTest result

Monte Carlo Estimate			
Site (sensor)	Exact Pr >= D	Exact Pr >= D+	Exact Pr >= D-
site 1	<.0001	1	<.0001
site 2 sensor1	0.0168	0.0093	0.8873
site 2 sensor2	<.0001	<.0001	1
site 3	<.0001	0.9978	<.0001
site 4 sensor1	<.0001	1	<.0001
site 4 sensor2	<.0001	0.9927	<.0001
site 5 sensor1	<.0001	0.0135	<.0001
site 5 sensor2	<.0001	0.9803	<.0001
site 6 sensor1	<.0001	0.962	<.0001
site 6 sensor2	0.0016	0.9804	0.0008
site 7 sensor1	<.0001	1	<.0001
site 7 sensor2	<.0001	0.9839	<.0001

Table 9 shows the result of Kolmogorov-Smirnov test. We can see that all the p-values of Exact Pr >= D are pretty small. Hence, from the result we we reject the null hypothesis for each of the sites (sensors) which means that all the sites (sensors) are different from site before. Hence, all sites (sensors) have different locations from the site before.

However, we can also see from Table 9 that site 1, site 3, site 4, site 6 and site 7 shift to a larger PPM value, while site 2 shift to a smaller PPM value. The result of two sensors in site 5 result in a contradiction.

6.3 Conclusion

Having done these test, we can see that the location has changed for all the sites when compared to the Site Before. However, only Site 2 has its location shifted to the left. This indicates that the values are not worst since only higher values will be toxic. The variance of the sites 2, 3, 4 and Sensor 1 of Site 5 are all different to Site Before. Site 1, 5 (Sensor 2), 6 and Site 7 have the same variance as Site Before.

7 Conclusion and Answer to Research Question

After having done all the tests finding results, we are now able to answer the research questions:
1) Are the measured PPM values at the different sites worst than normal level?
2) If so, which sites are really affected?

To answer these questions, we will discuss them for each site vs. the site before. The summary Table 10 gives a quicker overview of our finding.

Site 1 vs Site Before
The measured PPM values for Site 1 do differ from the normal levels (Site Before). We can conclude this based on the findings of the location and variance testing. Even though the variance is not significantly different to the Site Before its location has shifted to the right. Since higher values of PPM make the findings worst. As said before, the variance is the same and in order to see if the site is really affected, we created a box plots by hour to see is there are big changes in the values. As we can see in Figure 3, there are not big changes by hour. Hence, it is not really affected.

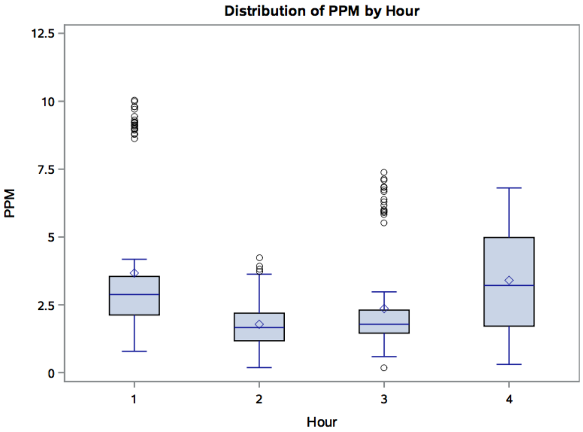

Figure 3: Box Plot: PPM by Site

Site 2 vs Site Before
The measured PPM values for Site 2 do differ to the normal levels. However, they have shifted to the left which implies that we have better PPM values than Site Before. Hence, Site 2 is not worst than before and therefore not affected by the accident.

Site 3 vs Site Before
The measures PPM values for Site 3 do differ from the normal levels. We have also seen that the variance is different than to the Site Before and the location shifted to the right hand side which indicates that the site is worst than Site Before. Looking also at the summary statistics table, we can see that the mean and median are far apart which suggests that Site 3 is really affected by the accident.

Site 4 vs Site Before

The measured PPM values for Site 4 do also differ from the normal levels. We have seen that the variance of the data for Site 4 is different to Site Before and that the location has shifted to the right. Looking again at the median and mean of this Site, we can say that it has really been affected by the accident. The maximum value is also 10 mg/unit higher than the highest value of Site Before.

Site 5 vs Site Before

The measured PPM values for Site 5 do also differ from the normal levels. The variance of sensor 2 is the same as Site Before but the variance of sensor 1 is different. Looking at the results of the location part, we noticed that sensor 1 does something very strange. It indicates that the location has shifted, but it has not shifted to the right or to the left. This seems very strange! For sensor 2 its location has shifted to the right. Taking these findings into account, we can conclude that the site is worst than the Site Before (normal level) but we cannot say for sure if it is really affected by the accident. If we consider sensor 1 to be broken, we can conclude that the site is really affected.

Site 6 vs Site Before

The measured PPM values for Site 6 do differ from the normal levels. Even though the variance is same the location has shifted to the right. Hence, it is worst than Site Before but is it really affected by the accident? Looking at the box plot figure in the summary statistics section, we see a 'blob' at the high end. Also the maximum value is 6mg/units higher than the maximum value of the Site Before. We would still like to create a boxplot, like with site 1, to see if there are changes per hour. A change by hour would indicate that the site is really affected by the accident. As Figure 4 (left sensor 1, right sensor 2) show us, there is a big change in sensor 1 for hour 2 and a big change for sensor 2 in hour 3. Hence, this site is really affected.

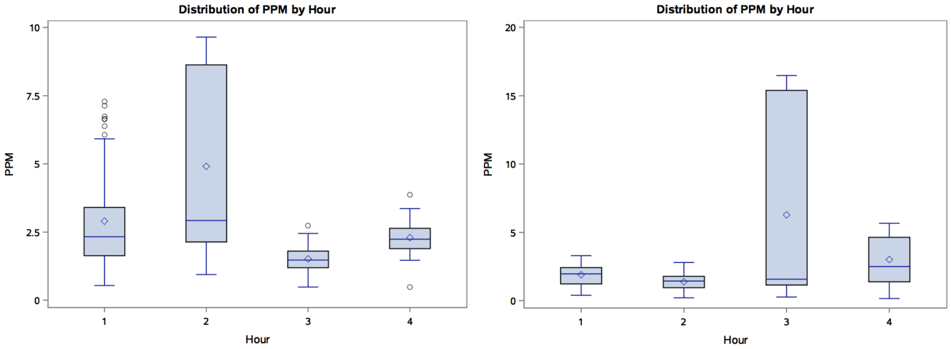

Figure 4: Box Plot: PPM by Site

Site 7 vs Site Before

The measured PPM value for Site 7 have shown some difficulty. Even though we have notices that they do differ from the normal levels and that they are worst. We cannot conclude that Site 7 is affected by the accident since we miss a lot of data for the measured PPM values. The summary statistics table show that we miss 60 measurements.

Overall

Question 1: Is/are there outlier/s in the data set?
Yes, we have found 2 outliers for Site Before. These outliers can be found in Table 5.

Question 2: Are the data normally distributed? The data is only normally distributed for Site Before with logarithmic PPM transformed values. Site 5 sensor 1 is also normally distributed with 'normal' PPM values. However, sensor 1 of Site 5 has given us throughout the whole analysis strange results so something is going on with this sensor.

Question 3: Do the PPM levels of the sites have different locations and variances compared to the PPM levels before the accident? Provide arguments for your choice of test.
As we have seen in the 'Location and Variance' section, all of the sites do not have the same location. However, only Site 2 has its location shifted to the left. The variance of the sites 2, 3, 4 and Sensor 1 of Site 5 are all different to Site Before. Site 1, 5 (Sensor 2), 6 and Site 7 have the same variance as Site Before. Arguments for the choices of our tests can be found through out this report and are included in each section.

Table 10: Summary Result of Findings

	Worst PPM value than before	Really Affected?
Site 1	Yes	No
Site 2	No	No
Site 3	Yes	Yes
Site 4	Yes	Yes
Site 5	Yes	Cannot say
Site 6	Yes	Yes
Site 7	Yes, but lack some values	Cannot say

8 Appendix

8.1 Tables

Table 11: Paired test result for difference

Before,Tests for Location: Mu0=0, diff				
Test	Statistic		p Value	
Student's t	t	0.288488	Pr > $\lvert t \rvert$	0.7740
Sign	M	-2	Pr >= $\lvert M \rvert$	0.6989
Signed Rank	S	-31	Pr >= $\lvert S \rvert$	0.8217

Site1,Tests for Location: Mu0=0,diff				
Test	Statistic		p Value	
Student's t	t	-0.27594	Pr > $\lvert t \rvert$	0.7828
Sign	M	12	Pr >= $\lvert M \rvert$	0.1375
Signed Rank	S	1145	Pr >= $\lvert S \rvert$	0.2885

Site2,Tests for Location: Mu0=0,diff				
Test	Statistic		p Value	
Student's t	t	-5.72001	Pr > $\lvert t \rvert$	<.0001
Sign	M	-22	Pr >= $\lvert M \rvert$	0.0047
Signed Rank	S	-4366	Pr >= $\lvert S \rvert$	<.0001

Site3,Tests for Location: Mu0=0,diff				
Test	Statistic		p Value	
Student's t	t	-0.00299	Pr > $\lvert t \rvert$	0.9976
Sign	M	4	Pr >= $\lvert M \rvert$	0.6515
Signed Rank	S	-42	Pr >= $\lvert S \rvert$	0.9690

Site4,Tests for Location: Mu0=0,diff				
Test	Statistic		p Value	
Student's t	t	-8.22578	Pr > $\lvert t \rvert$	<.0001
Sign	M	-31	Pr >= $\lvert M \rvert$	<.0001
Signed Rank	S	-7563	Pr >= $\lvert S \rvert$	<.0001

Site5, Tests for Location: Mu0=0,diff				
Test	Statistic		p Value	
Student's t	t	8.879892	Pr > $\lvert t \rvert$	<.0001
Sign	M	47	Pr >= $\lvert M \rvert$	<.0001
Signed Rank	S	8344	Pr >= $\lvert S \rvert$	<.0001

Site6,Tests for Location: Mu0=0,diff				
Test	Statistic		p Value	
Student's t	t	0.695044	Pr > $\lvert t \rvert$	0.4877
Sign	M	-32	Pr >= $\lvert M \rvert$	<.0001

Signed Rank	S	-2321	Pr >= $\lvert S\rvert$	0.0308

Site7,Tests for Location: Mu0=0,diff				
Test	Statistic		p Value	
Student's t	t	3.26841	Pr > $\lvert t\rvert$	0.0013
Sign	M	41	Pr >= $\lvert M\rvert$	<.0001
Signed Rank	S	3711	Pr >= $\lvert S\rvert$	<.0001

Table 12: Paired test result for log difference

Before,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	-0.54354	Pr > $\lvert t\rvert$	0.5888
Sign	M	-2	Pr >= $\lvert M\rvert$	0.6989
Signed Rank	S	-92	Pr >= $\lvert S\rvert$	0.5028

Site1,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	1.434436	Pr > $\lvert t\rvert$	0.1528
Sign	M	12	Pr >= $\lvert M\rvert$	0.1375
Signed Rank	S	1716	Pr >= $\lvert S\rvert$	0.1111

Site2,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	-4.52509	Pr > $\lvert t\rvert$	<.0001
Sign	M	-22	Pr >= $\lvert M\rvert$	0.0047
Signed Rank	S	-3977	Pr >= $\lvert S\rvert$	<.0001

Site3,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	2.104379	Pr > $\lvert t\rvert$	0.0364
Sign	M	4	Pr >= $\lvert M\rvert$	0.6515
Signed Rank	S	1841	Pr >= $\lvert S\rvert$	0.0873

Site4,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	-5.68439	Pr > $\lvert t\rvert$	<.0001
Sign	M	-31	Pr >= $\lvert M\rvert$	<.0001
Signed Rank	S	-5673	Pr >= $\lvert S\rvert$	<.0001

Site5,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	7.981056	Pr > $\lvert t\rvert$	<.0001
Sign	M	47	Pr >= $\lvert M\rvert$	<.0001

Signed Rank	S	7832	Pr >= $\|S\|$	<.0001

Site6,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	-2.29693	Pr > $\|t\|$	0.0225
Sign	M	-32	Pr >= $\|M\|$	<.0001
Signed Rank	S	-3168	Pr >= $\|S\|$	0.0031

Site7,Tests for Location: Mu0=0,ZL				
Test	Statistic		p Value	
Student's t	t	6.846991	Pr > $\|t\|$	<.0001
Sign	M	41	Pr >= $\|M\|$	<.0001
Signed Rank	S	4461	Pr >= $\|S\|$	<.0001

Table 13: Paired test result for ratio

Before,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student's t	t	1.969744	Pr > $\|t\|$	0.0536
Sign	M	-2	Pr >= $\|M\|$	0.6989
Signed Rank	S	177	Pr >= $\|S\|$	0.1950

Site1,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student's t	t	5.749321	Pr > $\|t\|$	<.0001
Sign	M	12	Pr >= $\|M\|$	0.1375
Signed Rank	S	5368	Pr >= $\|S\|$	<.0001

Site2,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student's t	t	2.183062	Pr > $\|t\|$	0.0300
Sign	M	-22	Pr >= $\|M\|$	0.0047
Signed Rank	S	-876	Pr >= $\|S\|$	0.3932

Site3,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student's t	t	5.067299	Pr > $\|t\|$	<.0001
Sign	M	4	Pr >= $\|M\|$	0.6515
Signed Rank	S	5108	Pr >= $\|S\|$	<.0001

Site4,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student's t	t	2.455597	Pr > $\|t\|$	0.0148
Sign	M	-31	Pr >= $\|M\|$	<.0001

Signed Rank	S	-2640	Pr >= $\lvert S\rvert$	0.0139

Site5,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student’s t	t	8.382088	Pr > $\lvert t\rvert$	<.0001
Sign	M	47	Pr >= $\lvert M\rvert$	<.0001
Signed Rank	S	8820	Pr >= $\lvert S\rvert$	<.0001

Site6,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student’s t	t	4.073359	Pr > $\lvert t\rvert$	<.0001
Sign	M	-32	Pr >= $\lvert M\rvert$	<.0001
Signed Rank	S	-384	Pr >= $\lvert S\rvert$	0.7221

Site7,Tests for Location: Mu0=0,ZR				
Test	Statistic		p Value	
Student’s t	t	7.288574	Pr > $\lvert t\rvert$	<.0001
Sign	M	41	Pr >= $\lvert M\rvert$	<.0001
Signed Rank	S	5121	Pr >= $\lvert S\rvert$	<.0001

Table 14: Test heterogeneity of variances using Levene’s test

Levene’s test result for Site 1									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	16.8067	16.8067	0.24	0.6213
2	Sensor	PPM	LV	Error	591	40663.3	68.8043		

Levene’s test result for Site 2 Sensor 1									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	450.1	450.1	21.73	<.0001
2	Sensor	PPM	LV	Error	342	7083.6	20.7124		

Levene’s test result for Site 2 Sensor 2									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	804.3	804.3	42.92	<.0001
2	Sensor	PPM	LV	Error	340	6370.5	18.7369		

Levene’s test result for Site 3									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	45809.3	45809.3	37.14	.0001
2	Sensor	PPM	LV	Error	573	706676	1233.3		

Levene’s test result for Site 4 Sensor 1									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF

1	Sensor	PPM	LV	Sensor	1	137726	137726	56.31	<.0001
2	Sensor	PPM	LV	Error	354	865844	2445.9		

Levene's test result for Site 4 Sensor 2

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	33216.5	33216.5	57.86	<.0001
2	Sensor	PPM	LV	Error	330	189441	574.1		

Levene's test result for Site 5 Sensor 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	788.1	788.1	43.59	<.0001
2	Sensor	PPM	LV	Error	354	6401.3	18.0829		

Levene's test result for Site 5 Sensor 2

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	3.8718	3.8718	0.10	0.7550
2	Sensor	PPM	LV	Error	352	13978.8	39.7124		

Levene's test result for Site 6 Sensor 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	218.3	218.3	2.42	0.1206
2	Sensor	PPM	LV	Error	353	31840.9	90.2007		

Levene's test result for Site 6 Sensor 2

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	3596.2	3596.2	4.18	0.0416
2	Sensor	PPM	LV	Error	345	296563	859.6		

Levene's test result for Site 7 Sensor 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	792.0	792.0	3.81	0.0519
2	Sensor	PPM	LV	Error	293	60924.8	207.9		

Levene's test result for Site 7 Sensor 2

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	LV	Sensor	1	1.8629	1.8629	0.05	0.8165
2	Sensor	PPM	LV	Error	354	12226.9	34.5391		

Table 15: Test heterogeneity of variances using BF test

BF test result for Site 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	0.8601	0.8601	0.34	0.5624
2	Sensor	PPM	BF	Error	591	1512.8	2.5597		

BF test result for Site 2 Sensor 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	19.2256	19.2256	15.94	<.000
2	Sensor	PPM	BF	Error	342	412.4	1.2058		

BF test result for Site 2 Sensor 2

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	51.0641	51.0641	60.20	<.0001
2	Sensor	PPM	BF	Error	340	288.4	0.8482		

BF test result for Site 3

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	550.0	550.0	39.94	<.0001
2	Sensor	PPM	BF	Error	573	7890.8	13.7710		

BF test result for Site 4 Sensor 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	1351.1	1351.1	111.94	<.0001
2	Sensor	PPM	BF	Error	354	4272.9	12.0704		

BF test result for Site 4 Sensor 2

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	604.7	604.7	71.80	<.0001
2	Sensor	PPM	BF	Error	330	2778.9	8.4210		

BF test result for Site 5 Sensor 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	48.6007	48.6007	57.50	<.0001
2	Sensor	PPM	BF	Error	354	299.2	0.8452		

BF test result for Site 5 Sensor 2

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	0.2305	0.2305	0.11	0.7383
2	Sensor	PPM	BF	Error	352	725.9	2.0623		

BF test result for Site 6 Sensor 1

OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	1.1556	1.1556	0.34	0.5607

2	Sensor	PPM	BF	Error	353	1202.7	3.4071		

BF test result for Site 6 Sensor 2									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	6.5270	6.5270	0.98	0.3235
2	Sensor	PPM	BF	Error	345	2303.8	6.6776		

BF test result for Site 7 Sensor 1									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	8.0079	8.0079	2.00	0.1587
2	Sensor	PPM	BF	Error	293	1175.0	4.0102		

BF test result for Site 7 Sensor 2									
OBS	Effect	Dependent	Method	Source	DF	SS	MS	Fvalue	ProbF
1	Sensor	PPM	BF	Sensor	1	1.1278	1.1278	0.61	0.4337
2	Sensor	PPM	BF	Error	354	649.9	1.8359		

8.2 CODE

8.2.1 Summary Statistics

```
LIBNAME A1 '/folders/myfolders/Assignment 1/';
DATA dataset;
        SET A1.assignmentdata;
        ID = _n_;
RUN;

proc sort data=dataset;
        by Site Sensor Hour;
run;

/* calc means of dataset */
Title 'Summary Statistics';
proc means data=dataset n min max mean median std var;
        class Site;
/*      class Site Hour Sensor; */
        var PPM;
        output out=meansdetails n=n min=min max=max mean=mean median=median std
run;

Title 'Sensor 1 by Hour';
proc means data=dataset n min max mean median std var;
/*      class Site Sensor; */
        class Site Hour;
        var PPM;
        where sensor = 1;
```

```
run;

Title 'Sensor 2 by Hour';
proc means data=dataset n min max mean median std var;
/*        class Site Sensor; */
          class Site Hour;
          var PPM;
          where sensor = 2;
run;
```

8.2.2 Probability Plots

```
LIBNAME A1 '/folders/myfolders/Assignment 1/';
DATA dataset;
          SET A1.datanooutliers;
          if (PPM < 0.1) then delete;
          LOGPPM = LOG(PPM);
RUN;

proc sort data=dataset;
          by Site Hour;
run;

proc univariate data=dataset normal noprint;
          by Site;
          probplot PPM /normal;
          probplot LOGPPM/normal;
run;
```

8.2.3 Box Plot

```
LIBNAME A1 '/folders/myfolders/Assignment 1/';
DATA dataset;
          SET A1.assignmentdata;
          ID = _n_;
          if (ppm < 0.1) then delete;
RUN;

proc sort data=dataset;
          by Site Sensor;
run;

/* ods graphics off; */
proc boxplot data=dataset;
   plot PPM*Site / boxstyle = schematic;
run;

proc boxplot data=dataset;
```

```
    plot PPM*Sensor / boxstyle = schematic;
    by Site;
run;
```

8.2.4 Paired Sample testing

8.3 Outlier testing

Grubbs Test

```
LIBNAME A1 '/folders/myfolders/Assignment 1/';
DATA dataset;
        SET A1.datasetnooutliers;
        ID = _n_;
        if (ppm < 0.1) then delete;
RUN;

proc sort data=dataset ;
        by Site Sensor;
run;

/* calc means of dataset */
proc means data=dataset mean std var n;
        var PPM;
        class Site Sensor;
        output out=meansdetails mean=mean std=sd var=var n=n;
run;

proc sort data=meansdetails;
        by Site Sensor;
run;

/* merge into one dataset */
data mergeddataset;
        merge dataset meansdetails;
        by Site Sensor;
run;

/* Grubbs */
Data grubbs;
        set mergeddataset;
        u = (PPM - mean)/sd;
        Grubbs_final_value = abs(u);
run;

/* outliers */
data grubbs_outlier;
        set grubbs;
        /* two sided grubbs */
        t = tinv( 0.05 /(n),n-2) ;
        cutoff = sqrt((((n-1)**2)*(t**2))/(n*((t**2)+n-2)));
        if (cutoff < 0) then return;
        if (Grubbs_final_value > cutoff) then output;
run;

proc print data=grubbs_outlier;
run;
```

8.3.1 Normality testing

```
LIBNAME A1 '/folders/myfolders/Assignment 1/';
DATA dataset;
  SET A1.exclude;
RUN;

DATA dataset;
  set dataset;
  if (site = 'Before') then sensor = 1;
  if (site = 'Site1') then sensor=1;
  if (site = 'Site3') then sensor=1;
run;

DATA dataset;
  set dataset;
  ID = _n_;
  LOGPPM = LOG(PPM);
run;

proc sort data=dataset;
  by site sensor;
run;


Title 'Normality Testing';
proc univariate data=dataset normal outtable=normality_table;
  var PPM LOGPPM;
  by Site sensor;
run;
```

8.3.2 Variance testing

```
/* Test for Variance */
/* LEVENE */
/* Test for Variance */
Title 'site_1';
PROC GLM DATA=site_1 ;
```

```
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;


 Title 'site_2 sensor 1';
PROC GLM DATA=site_2_1  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_2 sensor 2';
PROC GLM DATA=site_2_2  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_3 ';
PROC GLM DATA=site_3  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_4 sensor 1';
PROC GLM DATA=site_4_1  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_4 sensor 2';
PROC GLM DATA=site_4_2  ;
```

```
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title ’site_5 sensor 1’;
PROC GLM DATA=site_5_1 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title ’site_5 sensor 2’;
PROC GLM DATA=site_5_2 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title ’site_6 sensor 1’;
PROC GLM DATA=site_6_1 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title ’site_6 sensor 2’;
PROC GLM DATA=site_6_2 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title ’site_7 sensor 1’;
PROC GLM DATA=site_7_1 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_7 sensor 2';
PROC GLM DATA=site_7_2 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=LEVENE;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;

/* bf */

 Title 'site_1 ';
PROC GLM DATA=site_1 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;


 Title 'site_2 sensor 1';
PROC GLM DATA=site_2_1 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_2 sensor 2';
PROC GLM DATA=site_2_2 ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_3 ';
```

```
PROC GLM DATA=site_3  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
Title 'site_4 sensor 1';
PROC GLM DATA=site_4_1  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
Title 'site_4 sensor 2';
PROC GLM DATA=site_4_2  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
Title 'site_5 sensor 1';
PROC GLM DATA=site_5_1  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
Title 'site_5 sensor 2';
PROC GLM DATA=site_5_2  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
Title 'site_6 sensor 1';
PROC GLM DATA=site_6_1  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_6 sensor 2';
PROC GLM DATA=site_6_2  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_7 sensor 1';
PROC GLM DATA=site_7_1  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
 Title 'site_7 sensor 2';
PROC GLM DATA=site_7_2  ;
        CLASS SENSOR;
        MODEL ppm = sensor;
        MEANS sensor/ HOVTEST=BF;
        ODS OUTPUT HOVFTest=Hete_of_Var;
RUN;QUIT;
PROC PRINT DATA=Hete_of_Var;
RUN;
```

8.3.3 Location Testing

```
LIBNAME A1 '/folders/myfolders/Assignment 1/';
DATA dataset;
        SET A1.exclude;
        keep site hour sensor ppm ;

RUN;

DATA dataset;
        set dataset;
        if (site = 'Before') then sensor = 1;
        if (site = 'Site1') then sensor=1;
        if (site = 'Site3') then sensor=1;
/*      i = 1; */
run;

data site_before site_1 site_2 site_3 site_4 site_5 site_6 site_7;
        set dataset;
        if (Site = 'Before') then output site_before ;
        if (Site = 'Site1') then output site_1 ;
        if (Site = 'Site2') then output site_2 ;
        if (Site = 'Site3') then output site_3 ;
        if (Site = 'Site4') then output site_4 ;
        if (Site = 'Site5') then output site_5 ;
        if (Site = 'Site6') then output site_6 ;
        if (Site = 'Site7') then output site_7 ;
run;
data site_before;
        set site_before;
        sensor = 3;
run;


data site_1;
        set site_1 site_before;
run;
data site_3;
        set site_3 site_before;
run;
data site_2_1;
        set site_2;
        if (sensor = 2) then delete;
run;
data site_2_2;
        set site_2;
        if (sensor = 1) then delete;
run;

data site_4_1;
        set site_4;
        if (sensor = 2) then delete;
run;
data site_4_2;
        set site_4;
        if (sensor = 1) then delete;
run;

data site_5_1;
        set site_5;
        if (sensor = 2) then delete;
run;
data site_5_2;
        set site_5;
        if (sensor = 1) then delete;
run;

data site_6_1;
        set site_6;
        if (sensor = 2) then delete;
run;
data site_6_2;
        set site_6;
        if (sensor = 1) then delete;
run;

data site_7_1;
        set site_7;
        if (sensor = 2) then delete;
run;
data site_7_2;
        set site_7;
        if (sensor = 1) then delete;
run;

data site_2_1;
        set site_2_1 site_before;
run;
data site_2_2;
        set site_2_2 site_before;
run;

data site_4_1;
        set site_4_1 site_before;
run;
data site_4_2;
        set site_4_2 site_before;
run;

data site_5_1;
        set site_5_1 site_before;
run;
data site_5_2;
        set site_5_2 site_before;
run;

data site_6_1;
```

```
        set site_6_1 site_before;
run;
data site_6_2;
        set site_6_2 site_before;
run;

data site_7_1;
        set site_7_1 site_before;
run;
data site_7_2;
        set site_7_2 site_before;
run;


Title 'site_1 ';
PROC NPAR1WAY DATA=site_1;
        CLASS sensor;
        VAR PPM;
        EXACT KS / MC;
RUN;
Title 'site_2_1 ';
PROC NPAR1WAY DATA=site_2_1;
        CLASS sensor;
        VAR PPM;
        EXACT KS / MC;
RUN;
Title 'site_2_2 ';
PROC NPAR1WAY DATA=site_2_2;
        CLASS sensor;
        VAR PPM;
        EXACT KS / MC;
RUN;
Title 'site_3 ';
PROC NPAR1WAY DATA=site_3;
        CLASS sensor;
        VAR PPM;
        EXACT KS / MC;
RUN;
Title 'site_4_1 ';
PROC NPAR1WAY DATA=site_4_1;
        CLASS sensor;
        VAR PPM;
        EXACT KS / MC;
RUN;
Title 'site_4_2 ';
PROC NPAR1WAY DATA=site_4_2;
        CLASS sensor;
```

```
          VAR PPM;
          EXACT KS / MC;
RUN;
Title 'site_5_1';
PROC NPAR1WAY DATA=site_5_1;
          CLASS sensor;
          VAR PPM;
          EXACT KS / MC;
RUN;
Title 'site_5_2';
PROC NPAR1WAY DATA=site_5_2;
          CLASS sensor;
          VAR PPM;
          EXACT KS / MC;
RUN;
Title 'site_6_1';
PROC NPAR1WAY DATA=site_6_1;
          CLASS sensor;
          VAR PPM;
          EXACT KS / MC;
RUN;
Title 'site_6_2';
PROC NPAR1WAY DATA=site_6_2;
          CLASS sensor;
          VAR PPM;
          EXACT KS / MC;
RUN;
Title 'site_7_1';
PROC NPAR1WAY DATA=site_7_1;
          CLASS sensor;
          VAR PPM;
          EXACT KS / MC;
RUN;
Title 'site_7_2';
PROC NPAR1WAY DATA=site_7_2;
          CLASS sensor;
          VAR PPM;
          EXACT KS / MC;
RUN;
```