Jake Sauter
Environmental Statistics
HW 3

An investigator is concerned about the condition of the environment around an old industrial site.

1) They are first interested in determining if the level of pollutant (mg/l) in a water source nearby is below an acceptable threshold of 250 (mg.l). A sample of 25 observations was taken resulting in the data set GWhw3. Analyze the resulting data set for the investigator and report your findings.

Possible tests for this analysis:

One Sample t-Test

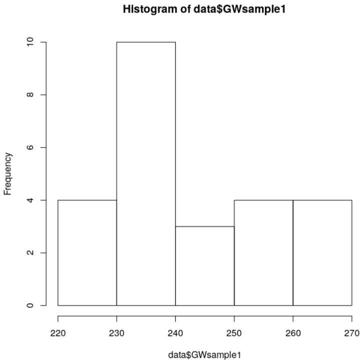
One Sample Wilcoxn Signed Rank Test

One Sample Sign Test

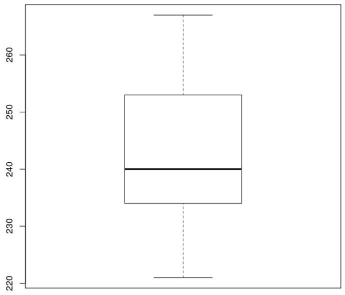
The possible statistical tests for this analysis are listed in order of their power, and also in the order in the amount of assumptions we have to make. To choose the best method, we will work our way down the list from top to bottom, and stop when the acquired assumptions for the test are fulfilled, as it will be the better than the tests below it, and any test above it have unfulfilled assumptions and cannot be performed.

One Sample t-Test: In order to perform a one sample t-test, we must make the assumption that the data is fairly normally distributed. To test this, we can first take a look at the data through graphical methods to assess normality, then implement a Shapiro-Wilkes test for normality. We can take all of these results into consideration to determine if our data is normal, and therefore if we can implement the one sample t-test to analyze this data

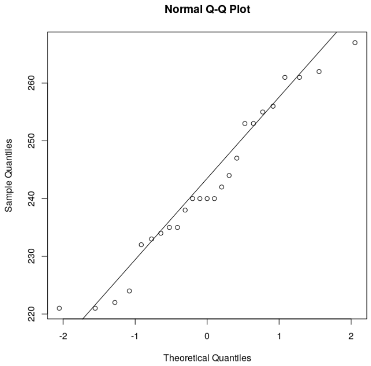
```
>> data = read.table(”data/GWhw3.txt”)
>> hist(data$GWsample1)
```



```
>> boxplot(data$GWsample1)
```



```
>> qqplot(data$GWsample1)
>> qqline(data$GWsample1)
```



```
>> shapiro.test(data$GWsample1)
Shapiro-Wilk normality test
data: data$GWsample1
W = 0.9533, p-value = 0.2971
```

The graphical analysis methods shown on the left indicate that the the data may not be normally distributed. This can be seen from the histogram not resembling a normal distribution in being symmetric about the mean, as well as not having a similar density distribution. The boxplot shows a right skew with the median being closer to the first quartile. Finally on the QQ plot, most of the points do not lie on the main diagonal, which would provide evidence that the quantiles of the data and the quantiles of the normal distribution match well.

When we use the Shapiro-Wilkes hypothesis test to assess normality, we see a low p-value, which does indicate that the data could be taken as normal, though is far from a p-value that would strongly indicate normality. Therefore the necessary assumptions for the one sample t-test are not met well enough for me to condisder the assumption valid.

One Sample Wilcoxn Signed Rank Test: The one sample Wilcoxn signed rank test requires the assumption that our data is symmetric. From the histogram and boxplot generated above, we see that this is not a valid assumption as the data is left skewed.

One Sample Sign Test: In order to use this test, we must meet the assumption of not too many ties in the data, where a tie is indicated by a data value having the same value as the threshold value. We can assess this in the the following way

```
>> data = data$GWsample1 ; threshold = 250
>> length(data[data == threshold])
       0
```

Here we see that no values are at the threshold value, so our assumptions for this test are met and this test can be implemented to determine if the level of pollutant in the water source is below the acceptable threshold of 250 (mg/l)

```
>> librray(BSDA)
>> SIGN.test(data, md=250, alternative="less")
```

This test produces a p-value of 0.05388, which is greater than our prescribed $\alpha = 0.05$ significance level, so we fail to reject the null hypothesis that the level of the pollutant is greater than or equal to the acceptable level.

2) Next the investigator is interested in examining the level of a certain heavy metal (mg/kg) that was a byproduct of something being produced in this factory. They are again interested in determining if the levels of this heavy metal are below the level 2.5 (mg/kg). Again a sample of 25 observations was collected resulting the data set SShw3, where any measured value that was below .8 was recorded as the reporting limit of .8. Analyze the resulting data set for the investigator and report your findings

This problem is very similar to Problem 1, though NDs being present in the data

Possible tests for this analysis:

One Sample t-Test

One Sample Wilcoxn Signed Rank Test

One Sample Sign Test

One Sample t-Test: This test cannot be used as NDs are be present and cannot be to perform the test.

One Sample Wilcoxn Signed Rank Test: This test cannot be used due to the presence of non-detects.

One Sample Sign Test: In order to use this test, we must meet the assumption of not too many ties in the data, where a tie is indicated by a data value having the same value as the threshold value. We can assess this in the the following way

```
>> length(data[data == 2.5])
       0
```

Here we see that no values are at the threshold value, so our assumptions for this test are met and this test can be implemented to determine if the level of pollutant is below the acceptable threshold of 2.5 (mg/kg)

```
>> SIGN.test(data$x, md=2.5, alternative=”less”)
```

This test produces a p-value of 0.007317, well below our significance threshold of α = 0.05. This means that we reject our null hypothesis that the pollutant level is equal to or above the threshold, and accept our alternative hypothesis that the pollutant is below the threshold level.

3) The investigator is now interested in the levels of a certain toxin (mg/kg) in the soil around the site. There is no standard level for this toxin around the site to the naturally occurring levels in the soil. They are interested to see if the levels of toxin are higher at the industrial site than the levels that naturally occur. Samples are collected from the site (variable labeled site) and compared to the levels in a control sample (variable labeled control) collected far away from the industrial site and recorded in the dataset toxHW3. Analyze the resulting data set for the investigator and report your findings.

Possible tests for this analysis:

Two Sample t-Test

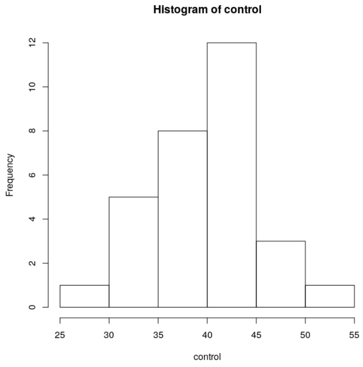
Wilcoxn Rank Sum Test

Two Sampe t-Test: The two sample t-test requires both data sets to be normal, and for the data sets to have equal variance, so we must first assess the normality of both of the data sets.

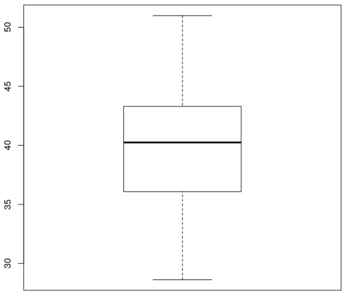
The output of the analysis steps are shown below. From these results, we can see that the control data is fairly likely to be normally distributed, though there is a left skew and less evidence for normality for the site data. In situations like this it might be be best to perform both analyses anyway (including the next analysis which does not require normality) to see if the results are still valid if the assumption does not hold.

```
>> data = read.table(”toxHW3.txt”)
```

>> hist(data$GWsample1)

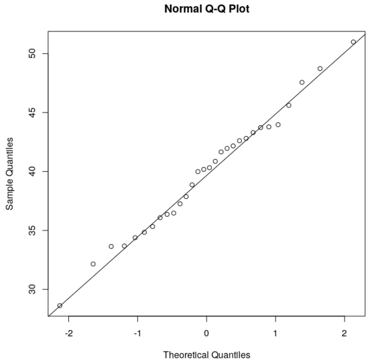


>> boxplot(data$control)



>> qqplot(data$control)

>> qqline(data$control)



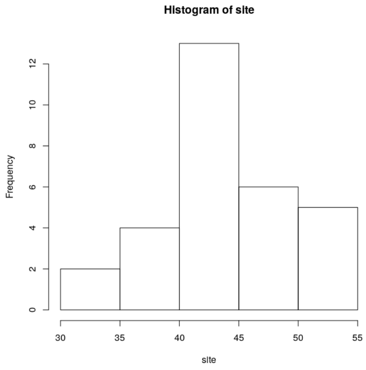
>> shapiro.test(data$control)

Shapiro-Wilk normality test

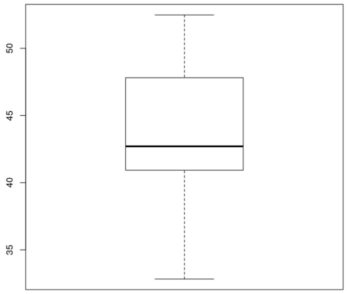
data: data$control W = 0.98872,

p-value = 0.983

>> hist(data$site)

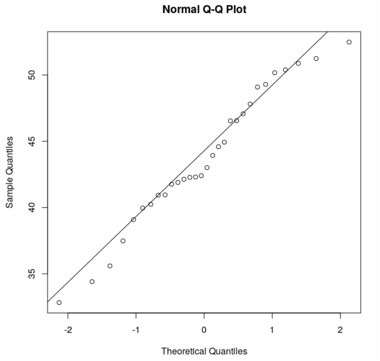


>> boxplot(data$site)



>> qqplot(data$site)

>> qqline(data$site)



>> shapiro.test(data$site)

Shapiro-Wilk normality test

data: data$site W = 0.9704,

p-value = 0.5501

We have to check one last thing before performing the t-test, being equal variances between the two groups

```
>> var(data$control)
          26.64761
>> var(data$site)
          26.41671
```

These variances are certainly close enough to fulfill the equal variance assumption, and now we can perform our t-test

```
>> t.test(data$site, data$control, alternative=”greater”)
```

This test produced a p-value of 0.002504, less than our prescribed significance level of $\alpha = 0.05$, permitting us to reject the null hypothesis, and accept our alternative hypothesis that toxin levels are higher at the industrial site under question than in a control site.

Wilcox Rank Sum Test: This test does not require the assumption of normality of both samples, though does require both samples to be roughly symmetric, as well as having equal variances. We have seen previously that these samples do meet the assumption for equal variances. Symmetry is also questionable from the histogram and boxplots shown above, though it is a weaker assumption to make than normality, which seems like a possible assumption.

```
>> wilcox.test(data$site, data$control, alternative=”greater”)
```

This hypothesis tests produced a p-value of .004, also being well below our prescribed significance level of $\alpha = 0.05$. This result indicates that we must reject our null hypothesis and accept our alternative hypothesis that the level of toxicity in the industrial site is at a higher level than that of the control site.

4) On a separate project that you have been assigned to, an environmentalist is interested in the effect of urban environments have seen on stream temperatures. The goal of the study is to determine if the temperature of a stream running through an urban environment is warmer than that of a river that does not run through an urban area. Is is assumed that there will be a trend in the data due to a time factor (temperatures change throughout the year) a paired test (variable labeled urban)

Possible tests for this analysis:

Paired Two Sample t-Test

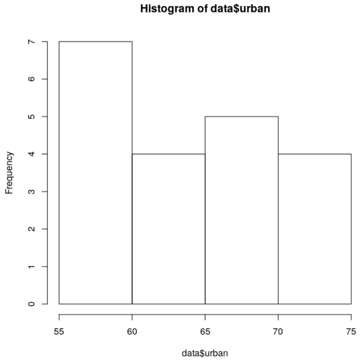
Paired Wilcox Signed Rank Test

Two Sample Paired Sign Test

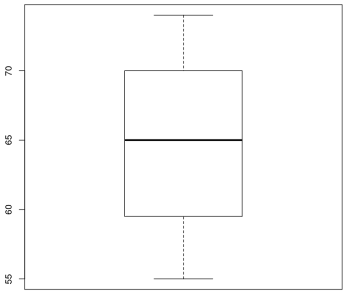
Paired Two Sample t-Test: The paired two sample t test requires both of our data samples to be normal, and for the variances of the samples to be the same. On the page below you can see the results of the normality analysis for both samples, and conclude neither sample to be normal for some of the same reasons stated in Problem 1. The boxplot for the urban data does seem to indicate normality, but if we look back at the histogram, we can see that this analysis does not hold.

Paired Wilcoxn Signed Rank Test: This test requires that the residuals between the pairs be symmetric. Through the same results shown below, the residuals cannot be assumed to be symmetric. This is more visible in the boxplot for the rural data, as we can see a clear right skew, indicated by the median being much closer to Q3 than Q1.

>> hist(data$urban)

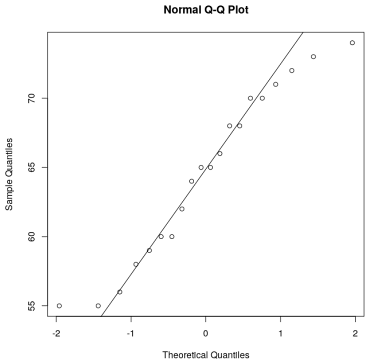


>> boxplot(data$urban)



>> qqplot(data$urban)

>> qqline(data$urban)



>> shapiro.test(data$urban)

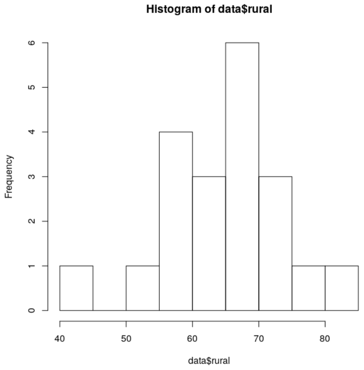
Shapiro-Wilk normality test

data: data$urban W = 0.9448,

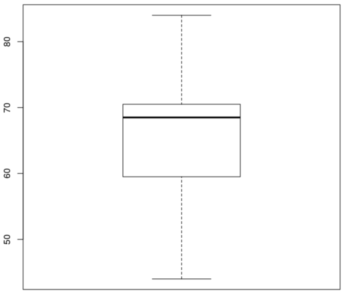
p-value = 0.2949

p-value = 0.983

>> hist(data$rural)

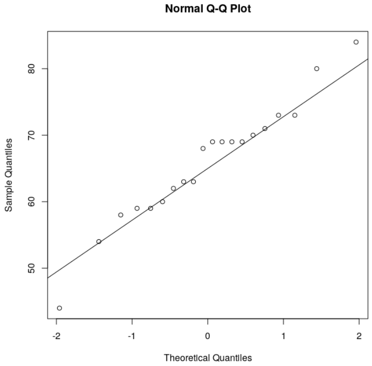


>> boxplot(data$rural)



>> qqplot(data$rural)

>> qqline(data$rural)



>> shapiro.test(data$rural)

Shapiro-Wilk normality test

data: data$rural W = 0.96676,

p-value = 0.6855

Two Sample Paired Sign Test: This test requires that there are not many ties. Ties here would be indicated by paired data values having the same value. This is assessed through R below

```
>> length(data$urban[(data$urban - data$rural) == 0])
        2
>> length(data$urban[(data$rural - data$urban) == 0])
        2
```

This amount of ties should not affect our analysis, so we can now apply the two sample paired sign test

```
>> SIGN.test(data$urban, data$rural, alternative=”greater”)
```

This test results in a p-value of .8811, being much larger than our prescribed significance level of $\alpha = 0.05$, failing to reject the null hypothesis that the streams have no difference in temperature.