MP2

Timothy Stubblefield Erika Shults

9/16/2021

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

1 - Read in the Senic Data

```
 #senic <- read.csv("C:\\Users\\timst\\Documents\\SMU\\Fall_2021\\STAT_6336\\Data\\Senic_Data.txt", hea
#Yes_Med <- dplyr::filter(senic, senic$V2 == 1)
#sum(Yes_Med$V2)
 senic <- read.csv("C:\\Users\\timst\\Documents\\SMU\\Fall_2021\\STAT_6336\\Data\\Senic.csv", header = 
print(senic)
```

```
##     ï..Infection.Risk Med.School
## 1                 4.1          2
## 2                 1.6          2
## 3                 2.7          2
## 4                 5.6          2
## 5                 5.7          2
## 6                 5.1          2
## 7                 4.6          2
## 8                 5.4          1
## 9                 4.3          2
## 10                6.3          2
## 11                4.9          1
## 12                4.3          2
## 13                7.7          1
## 14                3.7          2
## 15                4.2          2
```

```
## 16                5.5          2
## 17                4.5          2
## 18                6.4          2
## 19                4.2          2
## 20                4.1          2
## 21                4.2          2
## 22                4.8          2
## 23                5.0          1
## 24                4.8          2
## 25                4.0          1
## 26                3.9          1
## 27                4.5          2
## 28                3.2          2
## 29                4.4          2
## 30                4.9          2
## 31                5.0          2
## 32                5.2          2
## 33                5.3          2
## 34                6.1          2
## 35                6.3          2
## 36                5.0          2
## 37                2.8          2
## 38                4.6          2
## 39                4.1          2
## 40                1.3          2
## 41                3.7          2
## 42                4.7          2
## 43                3.0          2
## 44                5.6          1
## 45                5.5          2
## 46                4.6          1
## 47                6.5          2
## 48                5.5          2
## 49                1.8          2
## 50                4.2          2
## 51                5.6          2
## 52                4.3          2
## 53                7.6          2
## 54                7.8          2
## 55                3.1          2
## 56                3.9          2
## 57                3.7          2
## 58                4.3          2
## 59                3.9          1
## 60                4.5          2
## 61                3.4          2
## 62                5.7          1
## 63                5.4          2
## 64                4.4          2
## 65                5.0          2
## 66                4.3          2
## 67                4.4          2
## 68                3.7          2
## 69                4.5          2
```

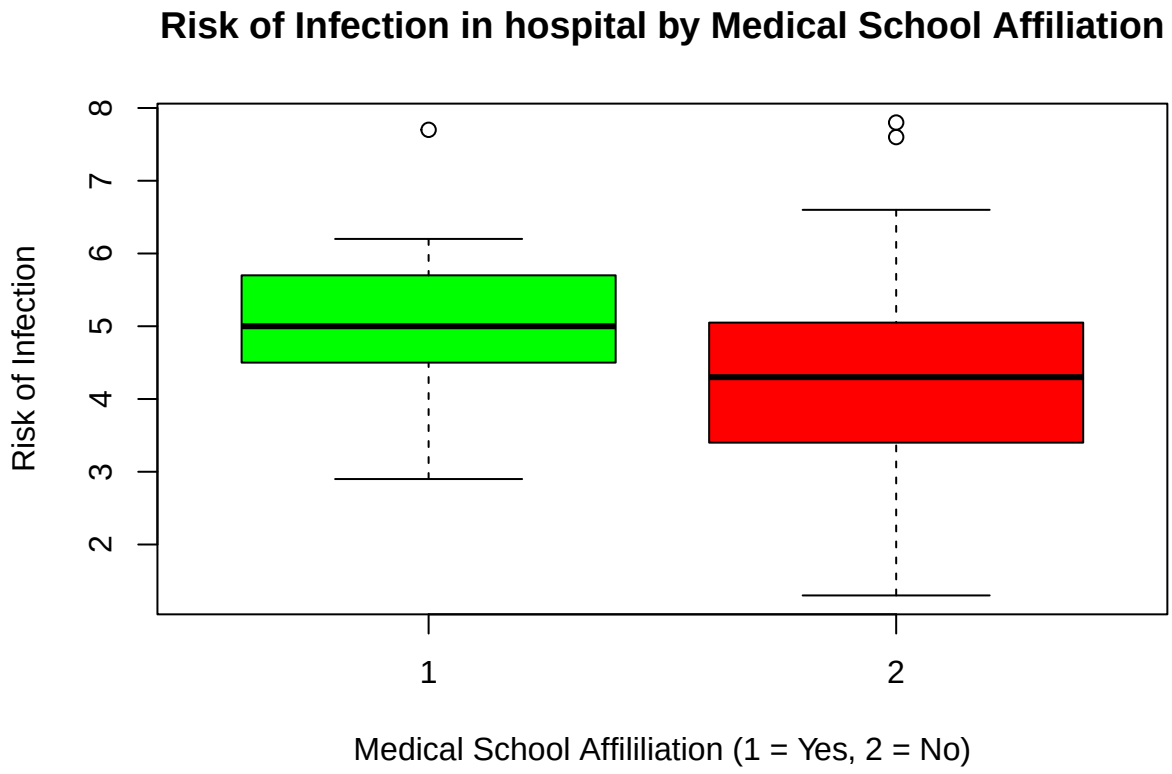
```
## 70                3.5         2
## 71                4.2         2
## 72                2.0         2
## 73                5.2         2
## 74                4.5         1
## 75                3.4         2
## 76                4.5         2
## 77                2.9         2
## 78                4.9         1
## 79                4.4         2
## 80                5.1         2
## 81                2.9         1
## 82                3.5         2
## 83                5.5         2
## 84                4.7         2
## 85                1.7         2
## 86                4.1         2
## 87                2.9         2
## 88                4.3         2
## 89                4.8         2
## 90                5.8         1
## 91                2.9         2
## 92                2.0         2
## 93                1.3         2
## 94                5.3         2
## 95                5.3         2
## 96                2.5         2
## 97                3.8         2
## 98                4.8         2
## 99                2.3         2
## 100               6.2         1
## 101               2.6         2
## 102               4.3         2
## 103               2.7         2
## 104               6.6         2
## 105               4.5         2
## 106               2.9         2
## 107               1.4         2
## 108               2.1         2
## 109               5.7         1
## 110               5.8         2
## 111               4.4         2
## 112               5.9         1
## 113               3.1         2
```

2 - Test for Normality of of One Sample Data

2.1 - Use a Boxplot

```
boxplot(senic[[1]] ~ senic[[2]], data = senic,
        xlab = "Medical School Affililiation (1 = Yes, 2 = No)",
```

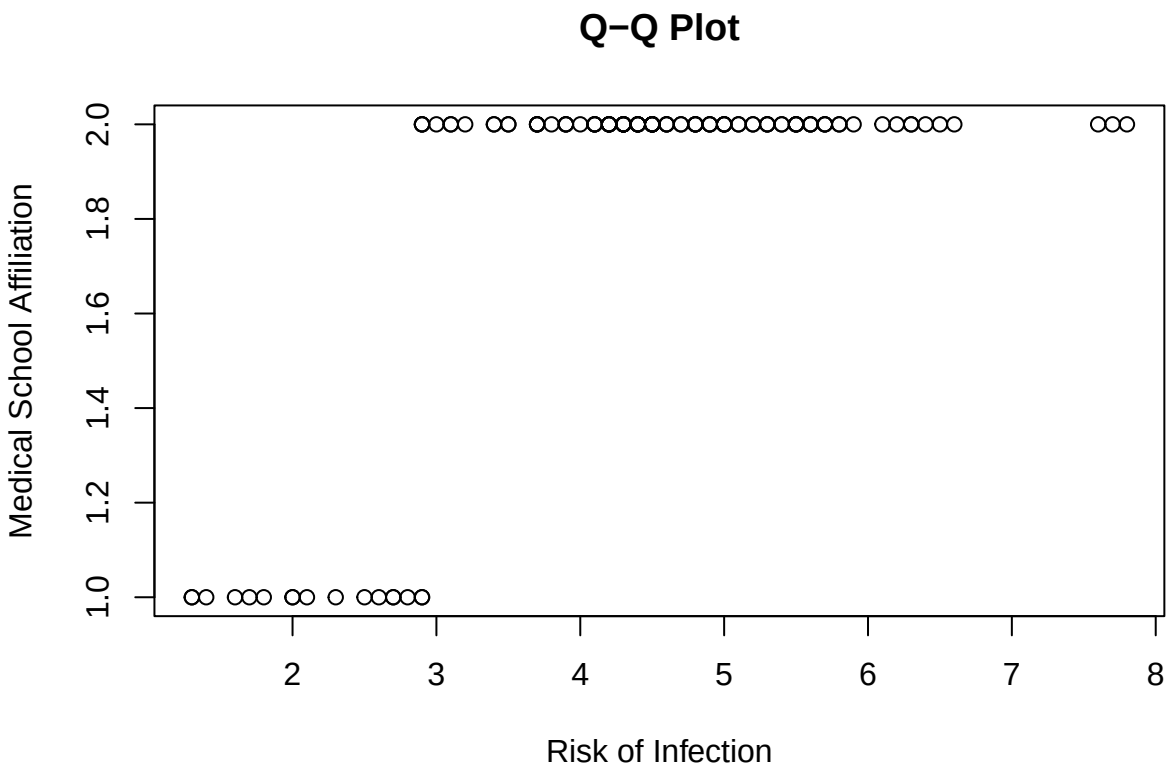
```
        ylab = "Risk of Infection",
        main = "Risk of Infection in hospital by Medical School Affiliation",
        col = c("green","red")
        )
```



Just looking at the Boxplots, both variables seem to follow a normal distribution. We will do further tests to see if we get corroborating results.

2.2 - Use a QQPlot

```
x <- senic[[1]]
y <- senic[[2]]
qqplot(x, y, xlab = "Risk of Infection", ylab = "Medical School Affiliation", main = "Q-Q Plot")
```



For this first QQPlot, it isn't correct since it is Risk of Infection vs Med School Affiliation. The latter variable is a binary variable so I need to plot against a theoretical normal distribution for each variable.

2.3 - QQ Plot for Risk of Infection

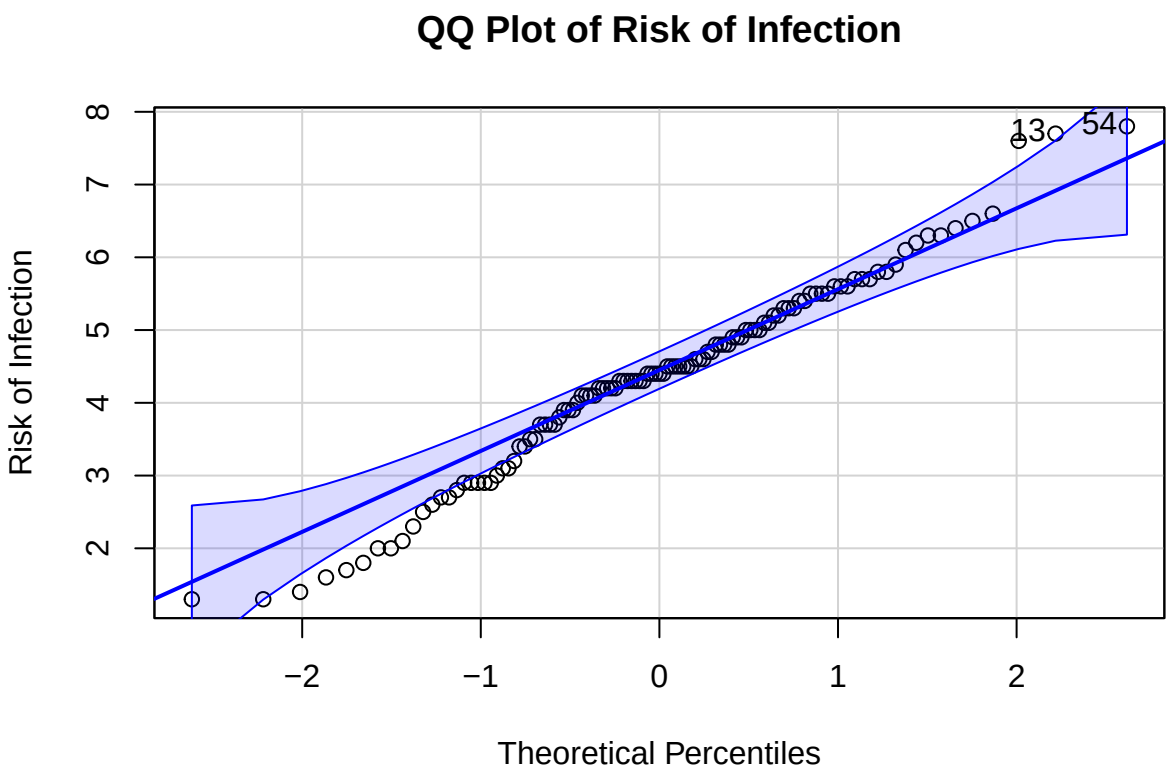
```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
qqPlot(x, distribution = "norm",
       xlab = "Theoretical Percentiles",
       ylab = "Risk of Infection",
       main = "QQ Plot of Risk of Infection")
```




```
## [1] 54 13
```

So, it looks like it is close to a normal distribution, but is not quite a normal distribution. A normal distribution would have roughly all of the points roughly within that blue band. Since some of the points are not within the band, that could indicate some skewness, but the T-Tools are generally robust to skewness anyway.

2.4 - Shapiro Wilk Test for Normality

For the SW Test, we have the Hypotheses:

$$H_0 : \text{Distribution is Normal}$$

$$H_a : \text{Distribution is NOT Normal}$$

```
# x is senic$'Risk of Infection'
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.98204, p-value = 0.1339
```

Since the p-value is not very small, for $\alpha = 0.05$ and $\alpha = 0.1$, we would fail to reject H_0.
With that in mind, there is no substantial evidence to show that the distribution is significantly different from the normal distribution. Thus, when we do the One Sample T-Test, the normality assumption should be satisfied.

3 - Test for Normality of Two Sample Data

3.1 - Separate the Risk of Infection data by Medical School Affiliation

```
sample1 <- dplyr::filter(senic, senic[[2]] == 1)
sample2 <- dplyr::filter(senic, senic[[2]] == 2)
sample1
```

```
##    ï..Infection.Risk Med.School
## 1                5.4          1
## 2                4.9          1
## 3                7.7          1
## 4                5.0          1
## 5                4.0          1
## 6                3.9          1
## 7                5.6          1
## 8                4.6          1
## 9                3.9          1
## 10               5.7          1
## 11               4.5          1
## 12               4.9          1
## 13               2.9          1
## 14               5.8          1
## 15               6.2          1
## 16               5.7          1
## 17               5.9          1
```

```
sample2
```

```
##    ï..Infection.Risk Med.School
## 1                4.1          2
## 2                1.6          2
## 3                2.7          2
## 4                5.6          2
## 5                5.7          2
## 6                5.1          2
## 7                4.6          2
## 8                4.3          2
## 9                6.3          2
## 10               4.3          2
## 11               3.7          2
## 12               4.2          2
## 13               5.5          2
## 14               4.5          2
## 15               6.4          2
```

```
## 16              4.2         2
## 17              4.1         2
## 18              4.2         2
## 19              4.8         2
## 20              4.8         2
## 21              4.5         2
## 22              3.2         2
## 23              4.4         2
## 24              4.9         2
## 25              5.0         2
## 26              5.2         2
## 27              5.3         2
## 28              6.1         2
## 29              6.3         2
## 30              5.0         2
## 31              2.8         2
## 32              4.6         2
## 33              4.1         2
## 34              1.3         2
## 35              3.7         2
## 36              4.7         2
## 37              3.0         2
## 38              5.5         2
## 39              6.5         2
## 40              5.5         2
## 41              1.8         2
## 42              4.2         2
## 43              5.6         2
## 44              4.3         2
## 45              7.6         2
## 46              7.8         2
## 47              3.1         2
## 48              3.9         2
## 49              3.7         2
## 50              4.3         2
## 51              4.5         2
## 52              3.4         2
## 53              5.4         2
## 54              4.4         2
## 55              5.0         2
## 56              4.3         2
## 57              4.4         2
## 58              3.7         2
## 59              4.5         2
## 60              3.5         2
## 61              4.2         2
## 62              2.0         2
## 63              5.2         2
## 64              3.4         2
## 65              4.5         2
## 66              2.9         2
## 67              4.4         2
## 68              5.1         2
## 69              3.5         2
```

```
## 70              5.5         2
## 71              4.7         2
## 72              1.7         2
## 73              4.1         2
## 74              2.9         2
## 75              4.3         2
## 76              4.8         2
## 77              2.9         2
## 78              2.0         2
## 79              1.3         2
## 80              5.3         2
## 81              5.3         2
## 82              2.5         2
## 83              3.8         2
## 84              4.8         2
## 85              2.3         2
## 86              2.6         2
## 87              4.3         2
## 88              2.7         2
## 89              6.6         2
## 90              4.5         2
## 91              2.9         2
## 92              1.4         2
## 93              2.1         2
## 94              5.8         2
## 95              4.4         2
## 96              3.1         2
```

So sample1 contains all the hospitals with a medical school affiliation. There are 17 of them ($n_1 = 17$). Likewise, sample2 contains all the hospitals without a medical school affiliation. There are 95 of them ($n_2 = 96$). Generally, we want the sizes of the Two Samples to have a similar number of observations. Since these two samples differ quite a bit, we may want to be careful.

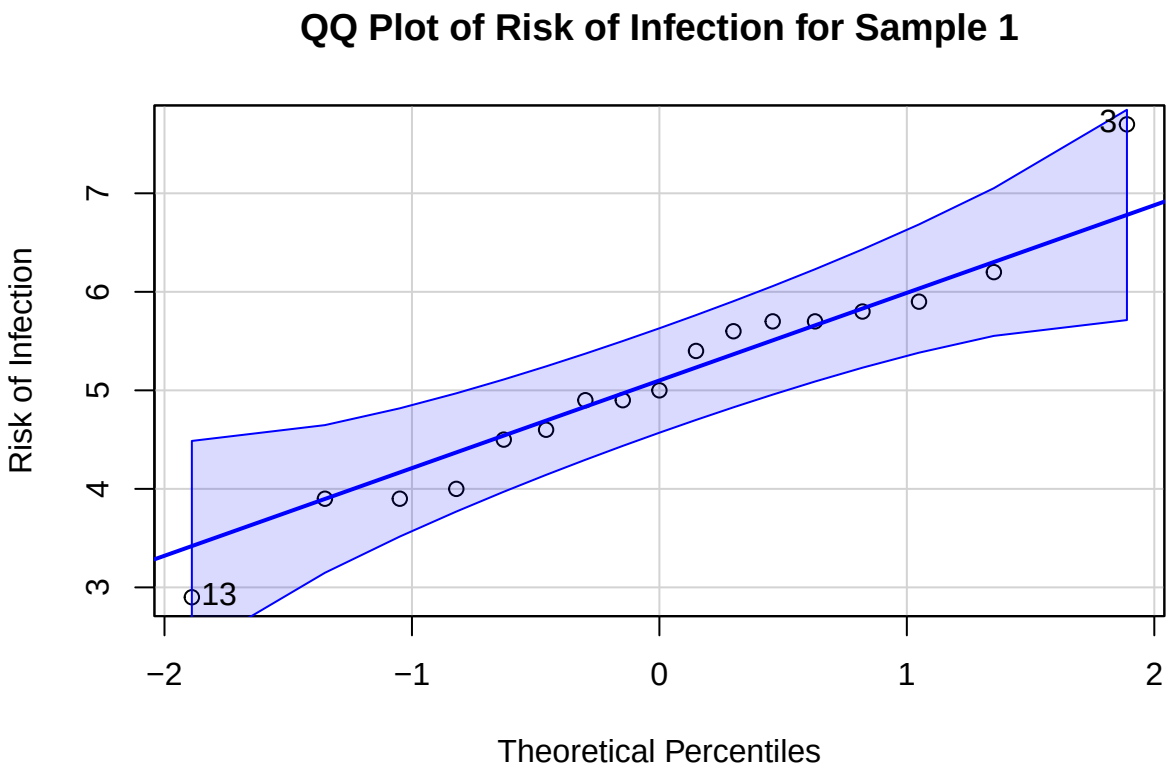
3.2 - Use the Boxplot

We already did this in (2.1)

3.3 - QQ Plots

3.3.1 - Sample 1 (Has a Med School)

```
library(car)
qqPlot(sample1[[1]], distribution = "norm",
       xlab = "Theoretical Percentiles",
       ylab = "Risk of Infection",
       main = "QQ Plot of Risk of Infection for Sample 1")
```

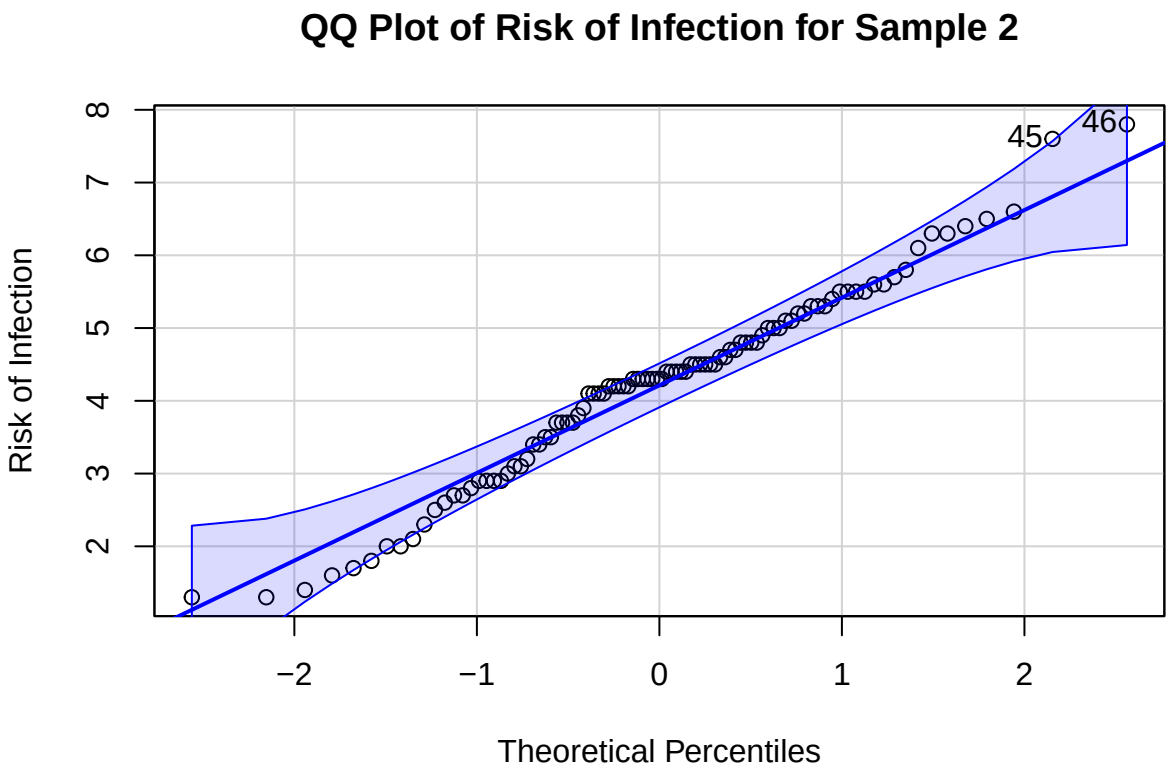



```
## [1]  3 13
```

As we can see, the QQ Plot of Sample 1 seems to be normally distributed as it tends to form a line when plotted against theoretical percentiles.

3.3.2 - Sample 2 (Does NOT have a Med School)

```
qqPlot(sample2[[1]], distribution = "norm",
       xlab = "Theoretical Percentiles",
       ylab = "Risk of Infection",
       main = "QQ Plot of Risk of Infection for Sample 2")
```




```
## [1] 46 45
```

While not quite as normally distributed as Sample 1, Sample 2 also seems to be mostly normal. There could be a little skew but T-Tools tend to be robust to that anyway.

3.4 - Shapiro Wilk Tests For Normality

3.4.1 - For Sample 1 (Has a Med School)

The Hypotheses are:

$$H_0 : \text{Sample 1 is Normally Distributed}$$

$$H_a : \text{Sample 1 is NOT Normally Distributed}$$

```
shapiro.test(sample1[[1]])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample1[[1]]
## W = 0.96792, p-value = 0.7811
```

The p-value of the SW Test is 0.7636 suggests strong evidence not to reject H_0. Therefore, it seems that Sample 1 is likely normally distributed.

3.4.2 - For Sample 2 (Does NOT have a Med School)

The Hypotheses are:

$$H_0 : \text{Sample 2 is Normally Distributed}$$

$$H_a : \text{Sample 1 is NOT Normally Distributed}$$

```
shapiro.test(sample2[[1]])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sample2[[1]]
## W = 0.98061, p-value = 0.1669
```

With a low p-value of 0.1404, we fail to reject H_0. Thus, there is no evidence to support that Sample 2 is not normally distributed, so we can conclude that Sample 2 is normally distributed. Just as the QQ Plots indicated in (3.3), we Sample 2 does not seem to be "as normal" as Sample 1 (as evidenced by the lower p-value). However, both samples were still determined to be normal in the SW Test.

4 - Perform some Tests on the Data

4.1 - First, get a summary of the Data

```
summary(senic)
```

```
##  ï..Infection.Risk   Med.School
##  Min.   :1.300     Min.   :1.00
##  1st Qu.:3.700     1st Qu.:2.00
##  Median :4.400     Median :2.00
##  Mean   :4.355     Mean   :1.85
##  3rd Qu.:5.200     3rd Qu.:2.00
##  Max.   :7.800     Max.   :2.00
```

4.2 - Perform One Sample T-Test on Risk of Infection for μ since we might have Normality

4.2.1 - Assumptions

The usual Assumptions for performing a T-Test are:

1. The data are normally distributed.
2. Constant Variability
3. Independence

In considering these assumptions, the SW Test from (2.4) shows strong evidence that the Risk of Infection follows a normal distribution. As for the independence assumption, that could best be answered by knowing how the data was collected. We can just assume that the samples are independent. Finally, since the default T-Test in R is the Welch T-Test, that test does NOT require constant variability, so we don't really need that assumption.

4.2.2 - The Design

This is a One Sample T-Test to see if $\mu = 4$. Since this is a One Sample T-Test, the degrees of freedom are:

$$df = n - 1 = 113 - 1 = 112$$

$$df = 112$$

4.2.2 - The Hypotheses and Test So, we have the Hypotheses:

$$H_0 : \mu = 4$$

$$H_a : \mu \neq 4$$

```
t.test(x, mu = 4)
```

```
##
##  One Sample t-test
##
## data:  x
## t = 2.8132, df = 112, p-value = 0.005794
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  4.104933 4.604801
## sample estimates:
## mean of x
##  4.354867
```

Since the p-value = 0.005794 is very small, we do not have evidence to show that $\mu = 4.355$. Therefore, it is likely that $\mu \neq 4.355$.

4.3 - Two Sample T-Test for Risk of Infection by Hospital Affiilation

4.3.1 - Assumptions

1. Both samples are normally distributed.
2. Samples have equal variance
3. Independence

In the SW Tests from (3.4), both Sample 1 and Sample 2 were determined to be normally distributed. Therefore, that assumption is met. Also, since R uses the Welch's T-Test, the assumption of equal variances does NOT need to be satisfied. Finally, we will just assume that the samples are independent.

4.3.2 - The Design

Essentially, the data will be split up by medical school affiliation. One Sample will have all of the observations where the hospital was affiliated with a medical school and the other sample will have all of the observations where the hospital was NOT affiliated with a medical school. We will conduct a T-Test to determine if the means of these two groups are different. If the means are different, that could indicate a correlation between medical school affiliation and the risk of becoming infected. The degrees of freedom are:

$$df = n_1 + n_2 - 2$$

$$df = 25.252$$

The summaries of the Samples are:

```
summary(sample1)
```

```
##  ï..Infection.Risk   Med.School
##  Min.   :2.900     Min.   :1
##  1st Qu.:4.500     1st Qu.:1
##  Median :5.000     Median :1
##  Mean   :5.094     Mean   :1
##  3rd Qu.:5.700     3rd Qu.:1
##  Max.   :7.700     Max.   :1
```

```
summary(sample2)
```

```
##  ï..Infection.Risk   Med.School
##  Min.   :1.300     Min.   :2
##  1st Qu.:3.400     1st Qu.:2
##  Median :4.300     Median :2
##  Mean   :4.224     Mean   :2
##  3rd Qu.:5.025     3rd Qu.:2
##  Max.   :7.800     Max.   :2
```

4.3.3 - The Hypotheses and Test

So, we have the Hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 \neq 0$$

. In laymans terms, the Hypotheses are:

H_0 : Medical School affiliation does NOT affect Risk of Infection

H_a : Medical School affiliation does affect Risk of Infection

```
#Remember:
#x = senic$'Risk of Infection'
#y = senic$Medical School'
t.test(x~y)
```

```
##
##  Welch Two Sample t-test
##
## data:  x by y
## t = 2.8772, df = 25.01, p-value = 0.008093
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal to 0
## 95 percent confidence interval:
##  0.2472939 1.4930247
## sample estimates:
## mean in group 1 mean in group 2
##        5.094118        4.223958
```

Notice that that p-value = 0.0345, which is relatively low. Also, 0 is NOT within the 95% Confidence Interval of the difference between the means. That is quite indicative that the means are not likely the same. Therefore, we can say that the Two Sample T-Test shows there is evidence that the average risk of infection at hospitals with a medical school differs from the average risk of infection at hospitals with no medical school.