Assessment01

STAT414

2024-12-08

```
library(EnvStats)
```

```
##
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

```
# 1. Arsenic concentrations (ppb) collected quarterly at two groundwater monitoring wells.
# These data are stored in the data frame EPA.92c.arsenic3.df.
data <- EPA.92c.arsenic3.df

background <- data[data$Well.type=="Background",]
compliance <- data[data$Well.type=="Compliance",]

background
```

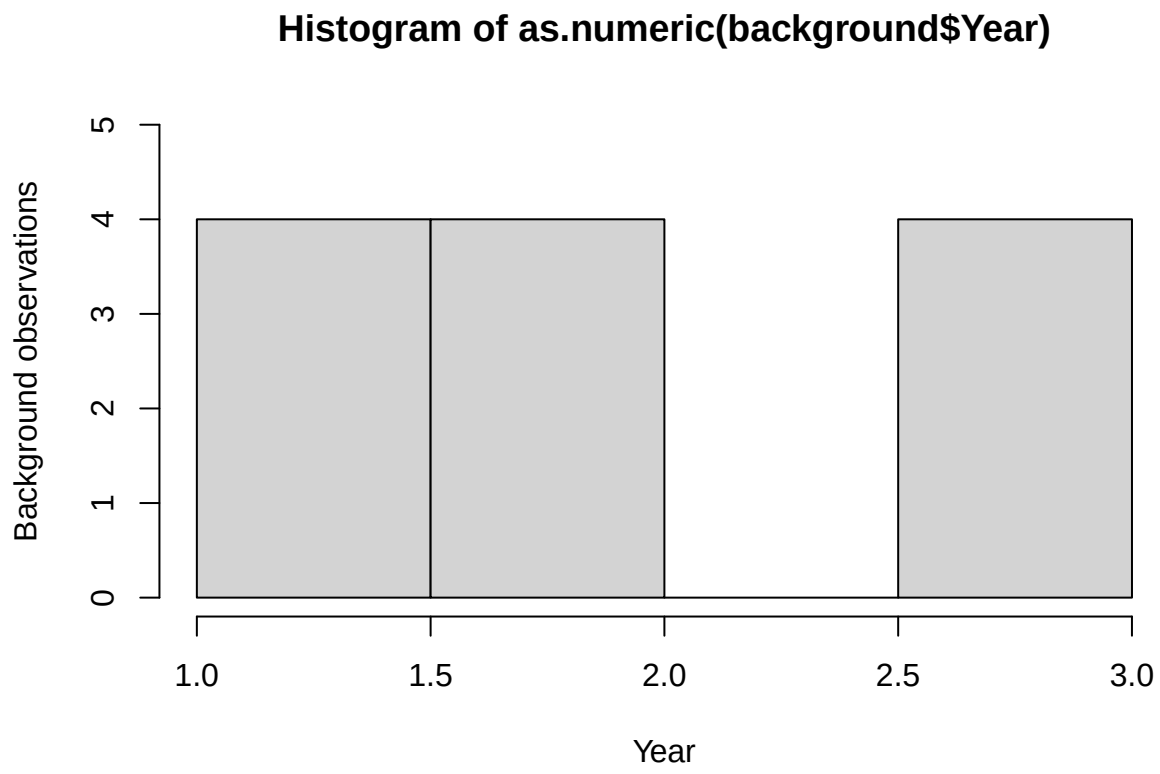
```
##    Arsenic Year  Well.type
## 1     12.6    1 Background
## 2     30.8    1 Background
## 3     52.0    1 Background
## 4     28.1    1 Background
## 5     33.3    2 Background
## 6     44.0    2 Background
## 7      3.0    2 Background
## 8     12.8    2 Background
## 9     58.1    3 Background
## 10    12.6    3 Background
## 11    17.6    3 Background
## 12    25.3    3 Background
```

```
compliance
```

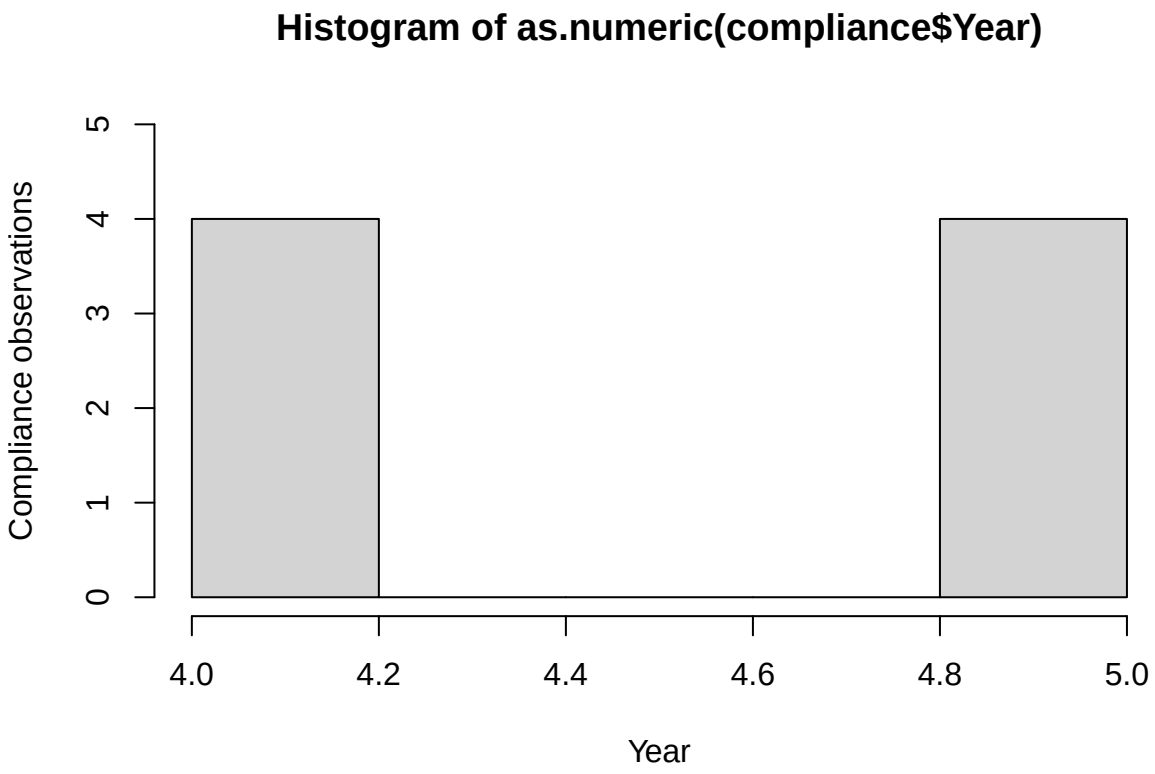
```
##    Arsenic Year  Well.type
## 13    48.0    4 Compliance
## 14    30.3    4 Compliance
## 15    42.5    4 Compliance
## 16    15.0    4 Compliance
```

```
## 17    47.6    5 Compliance
## 18     3.8    5 Compliance
## 19     2.6    5 Compliance
## 20    51.9    5 Compliance
```

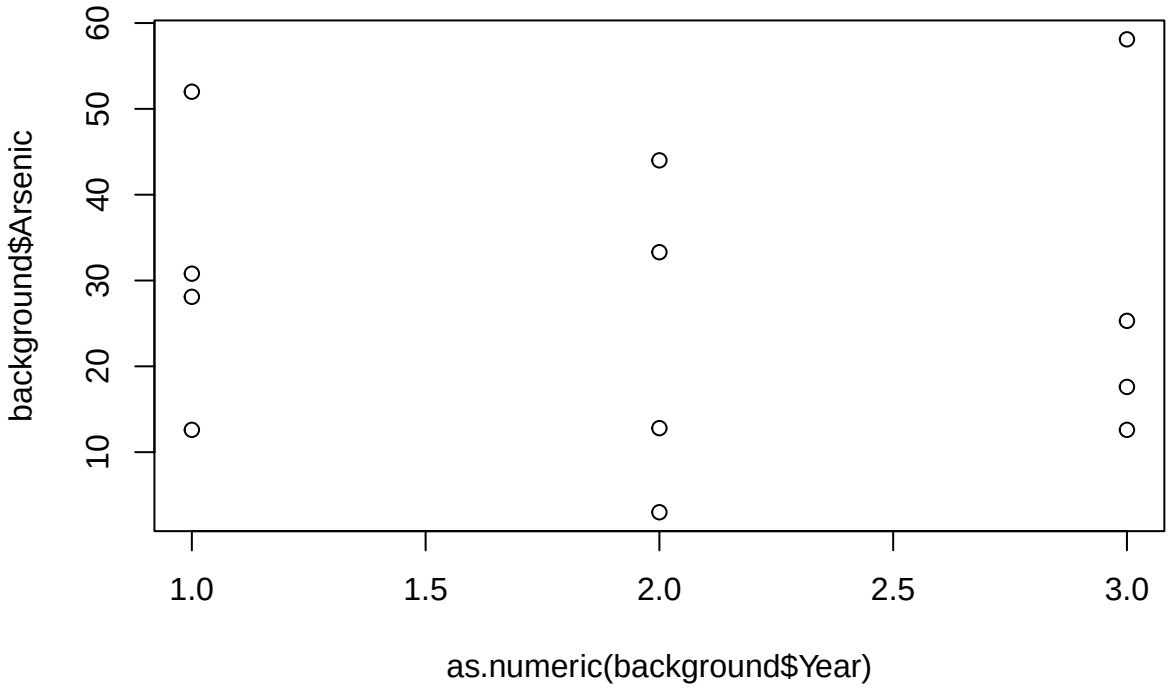
```r
# (a) For each well, plot the observations by year. Do you see any major differences
# between years?

hist(as.numeric(background$Year), xlab="Year",
     ylab="Background observations", ylim=c(0,5))
```



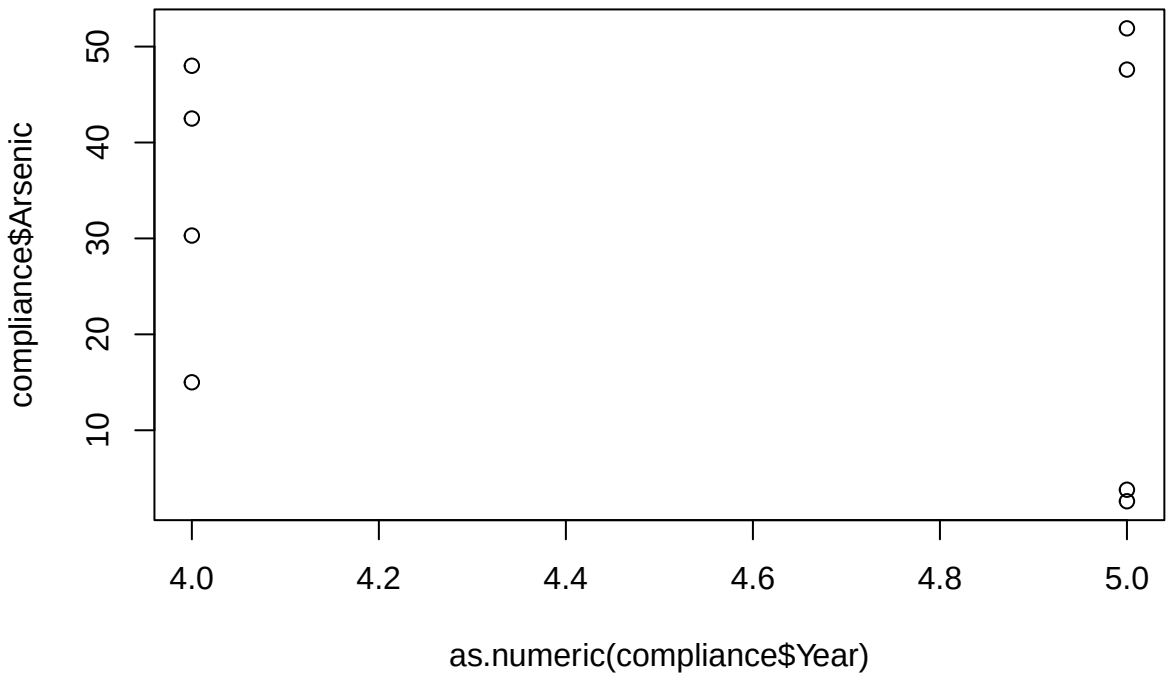

```r
hist(as.numeric(compliance$Year), xlab="Year",
     ylab="Compliance observations", ylim=c(0,5))
```




```
plot(as.numeric(background$Year), background$Arsenic)
```




```
plot(as.numeric(compliance$Year), compliance$Arsenic)
```




```
cat("The background well data was collected years 1-3, and the compliance well
    was collected years 4-5. All data collected quarterly each year. No major
    differences except that the background well has 1 extra year of quarterly data
    collection, and the wells were not tested at the same time for any given year.")
```

```
## The background well data was collected years 1-3, and the compliance well
##     was collected years 4-5. All data collected quarterly each year. No major
##     differences except that the background well has 1 extra year of quarterly data
##     collection, and the wells were not tested at the same time for any given year.
```

```
# (b) Compute summary statistics for each well (combine years).
summary(background)
```

```
##     Arsenic      Year       Well.type
##  Min.   : 3.00   1:4   Background:12
##  1st Qu.:12.75   2:4   Compliance: 0
##  Median :26.70   3:4
##  Mean   :27.52   4:0
##  3rd Qu.:35.98   5:0
##  Max.   :58.10
```

```
summary(compliance)
```

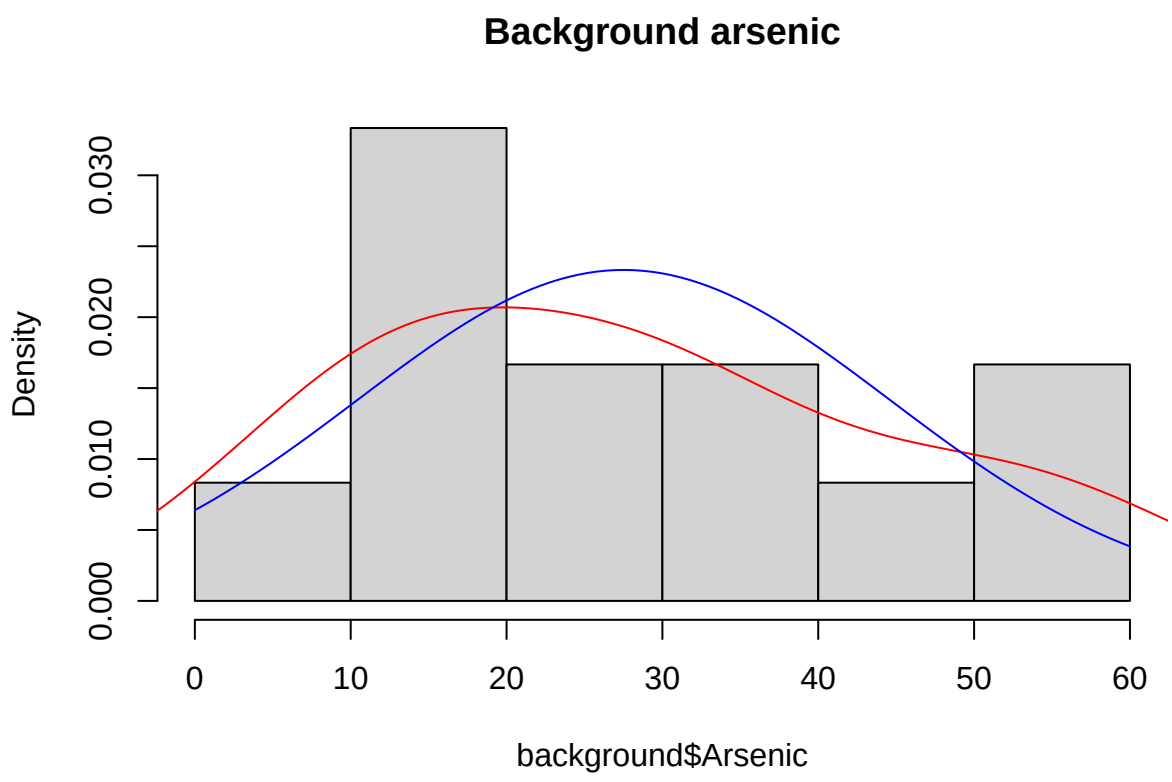
```
##     Arsenic      Year       Well.type
```

```
##  Min.   : 2.60   1:0   Background:0
##  1st Qu.:12.20   2:0   Compliance:8
##  Median :36.40   3:0
##  Mean   :30.21   4:4
##  3rd Qu.:47.70   5:4
##  Max.   :51.90
```

```
# (c) Compare the observed distribution of arsenic at each well. Use whatever types
# of plots you wish. Does the compliance well appear to show any evidence of
# contamination? Why or why not?

hist(background$Arsenic, probability="TRUE",main="Background arsenic")
lines(density(background$Arsenic), add=TRUE, col="red")
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "add" is not a graphical
## parameter
```

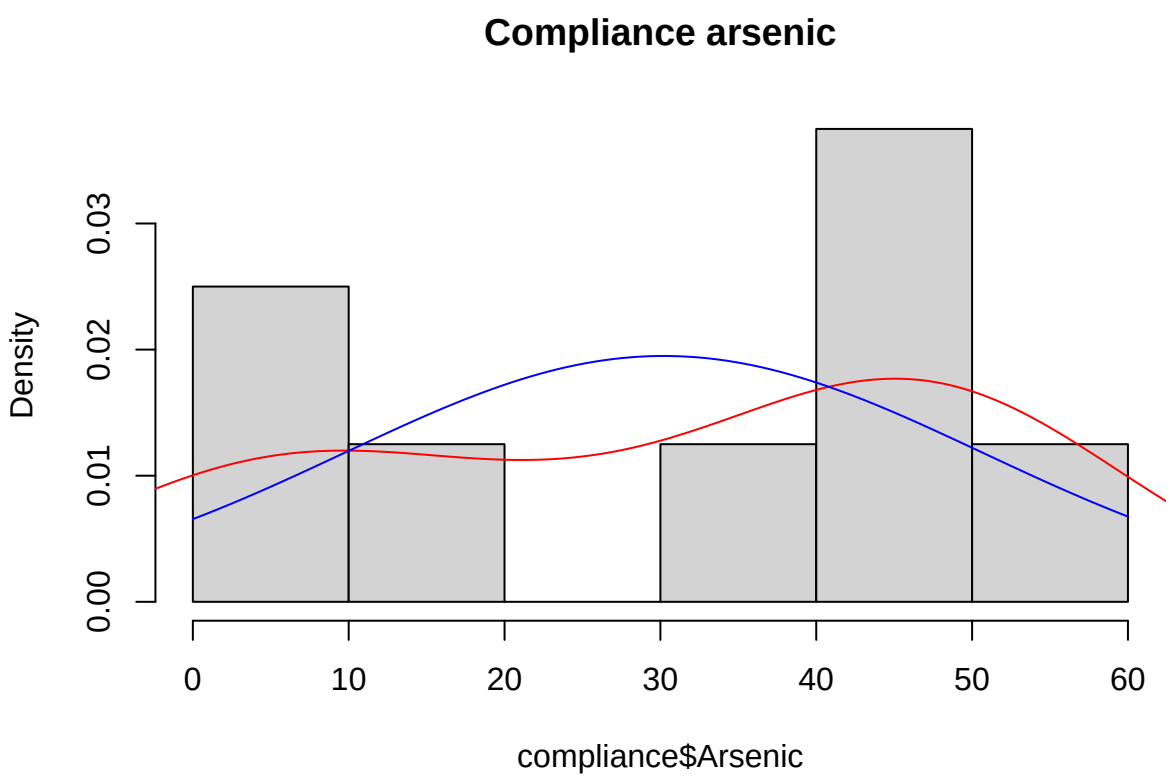
```
#superimpose normal curve
curve(dnorm(x,mean=mean(background$Arsenic), sd=sd(background$Arsenic)), add=TRUE, col="blue")
```




```
hist(compliance$Arsenic, probability="TRUE",main="Compliance arsenic")
lines(density(compliance$Arsenic), add=TRUE, col="red")
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "add" is not a graphical
## parameter
```

```r
#superimpose normal curve
curve(dnorm(x,mean=mean(compliance$Arsenic), sd=sd(compliance$Arsenic)), add=TRUE, col="blue")
```




```r
cat("The compliance well does show evidence of aresenic contamination because its
    histogram distribution is slightly left-skewed, showing that it commonly has
    high ppb concentration of arsenic.")
```

```
## The compliance well does show evidence of aresenic contamination because its
##     histogram distribution is slightly left-skewed, showing that it commonly has
##     high ppb concentration of arsenic.
```

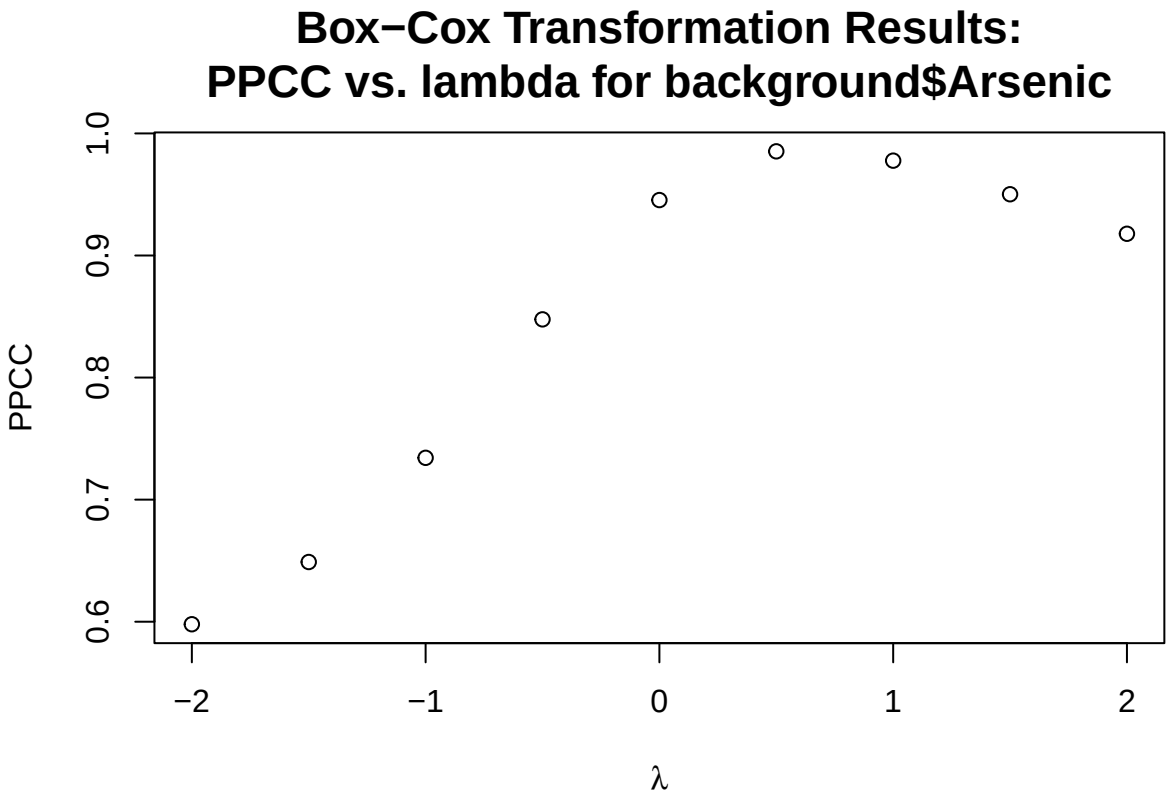
```r
# 2. Consider only the data from the Background wells for this part.


# (a) Does it appear that the background well data may be modeled as coming from a
# normal distribution? Support your conclusions with plots and tests of normality
# statistics.

# qqplot(background$Arsenic)
# qqline(background$Arsenic)

cat("From the histogram we see that the shape of the distribution and its density
    line somehawt resembles the shape of the superimposed normal curve, therefore
    conclusions cannot be drawn from the distribution alone, there are noticable
    differences and errors in the shape while still remaining similar.")
```

```
## From the histogram we see that the shape of the distribution and its density
##     line somehawt resembles the shape of the superimposed normal curve, therefore
##     conclusions cannot be drawn from the distribution alone, there are noticable
##     differences and errors in the shape while still remaining similar.
```

```
kurt <- kurtosis(background$Arsenic)
skew <- skewness(background$Arsenic)

cat("From the kurtosis and skewness values, we know that for a normal distribution
    their values should be 0 or close to 0. The value of skewness for the arsenic
    concentration for the background well is", skew, "and the kurtosis is", kurt,
    "
    neither of which are 0 or close to 0. Therefore we can say that from these values
    and the histogram distribution that the background well data is most likely not
    modeled from a normal distribution.")
```

```
## From the kurtosis and skewness values, we know that for a normal distribution
##     their values should be 0 or close to 0. The value of skewness for the arsenic
##     concentration for the background well is 0.4895164 and the kurtosis is -0.6688492
##     neither of which are 0 or close to 0. Therefore we can say that from these values
##     and the histogram distribution that the background well data is most likely not
##     modeled from a normal distribution.
```

```
# (b) Find the value of lambda corresponding Box-Cox transformation of the data to normality.
#Plot a histogram of the transformed data and overlay the appropriate Normal curve to
#illustrate the quality of fit of this model. Repeat the above  exercise with
#the original (non-transformed) data and illustrate how the transformation improves
#the normality fit.
  box <- boxcox(background$Arsenic, lambda=c(-2,-1.5,-1,-.5,0,.5,1,1.5,2), plot="TRUE")
  plot(box)
```



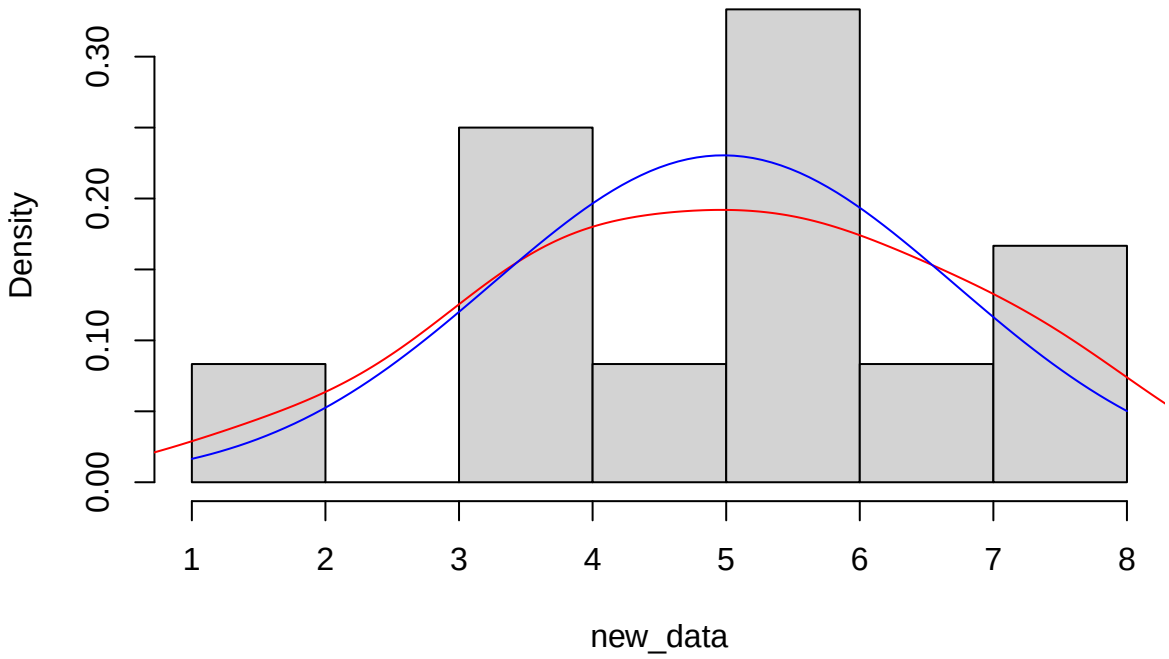

```
  cat("Choose lambda = .5 since the PPCC value is maxmimized at lambda=.5")
```

```
## Choose lambda = .5 since the PPCC value is maxmimized at lambda=.5
```

```
  new_data <- background$Arsenic ^ .5
  hist(new_data, probability="TRUE",main="Background lambda=.5 arsenic")
  lines(density(new_data), add=TRUE, col="red")
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "add" is not a graphical
## parameter
```

```
  #superimpose normal curve
  curve(dnorm(x,mean=mean(new_data), sd=sd(new_data)), add=TRUE, col="blue")
```




```
cat("We can clearly see from the transformed data distribution that is matches
    the expected normal distribution, and that it is a much better fit compared
    to the original histogram distribution as calculated previously. The transformation
    improves the normality fit by making the density line similar in shape and structure
    to the superimposed normal curve.")
```

```
## We can clearly see from the transformed data distribution that is matches
##       the expected normal distribution, and that it is a much better fit compared
##       to the original histogram distribution as calculated previously. The transformation
##       improves the normality fit by making the density line similar in shape and structure
##       to the superimposed normal curve.
```