Course Project Part 1

Sean Leggett - BDA201 Winter 2020
March 7, 2020

```
## load required libraries
library(car)
```

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

Questions

1. The ammonia concentration in your wastewater treatment plant is measured every 6 hours. The data for one year is available (attached as ammonia.csv)

2. Use a visualization plot to determine if the data are from a normal distribution. And then confirm your answer using a suitable plot.

3. Estimate location and spread statistics assuming the data are from a normal distribution.

4. What if I told you that these measured values are not independent. How does it affect your answer from Question 3?

5. Among normal distribution, and student's t distribution briefly explain which would be more suitable for this dataset?

6. What is the probability of having an ammonia concentration greater than 40 mg/L. (use an appropriate distribution)

7. There is 64.7% probability that a random reading will be less than what concentration of ammonia?

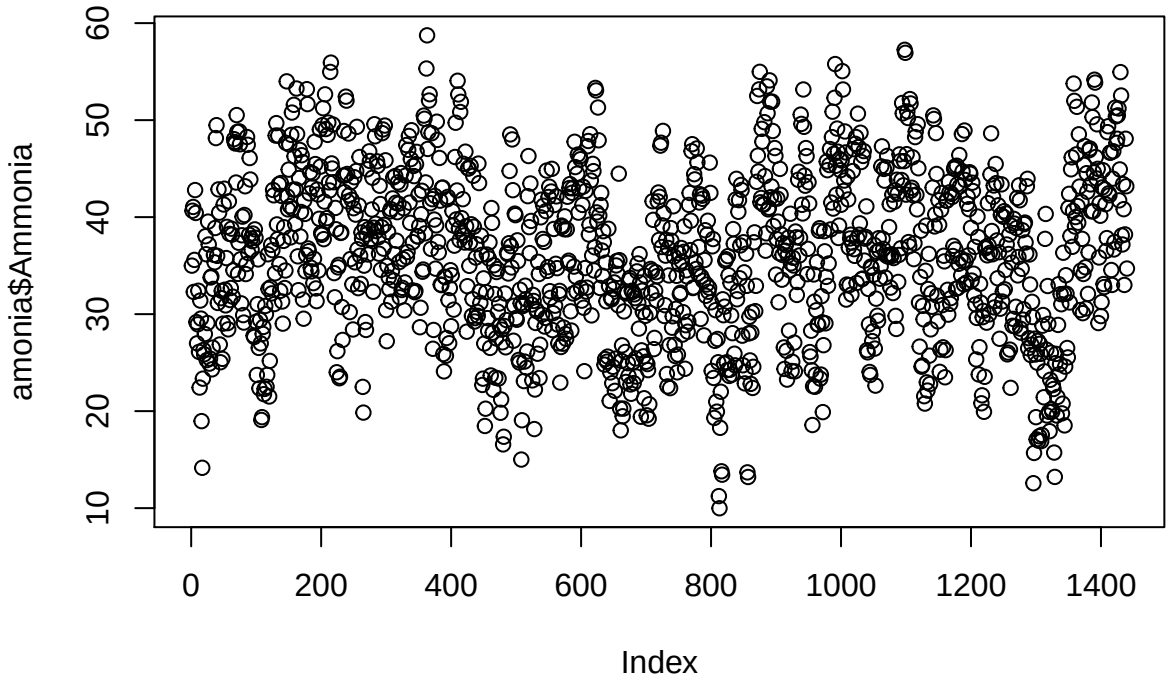
Answers

```
## read csv file into dataframe
amonia <- read.csv("ammonia.csv")
```

Question 2 - visualization of distribution.

First, a histogram to view distribution of variable.

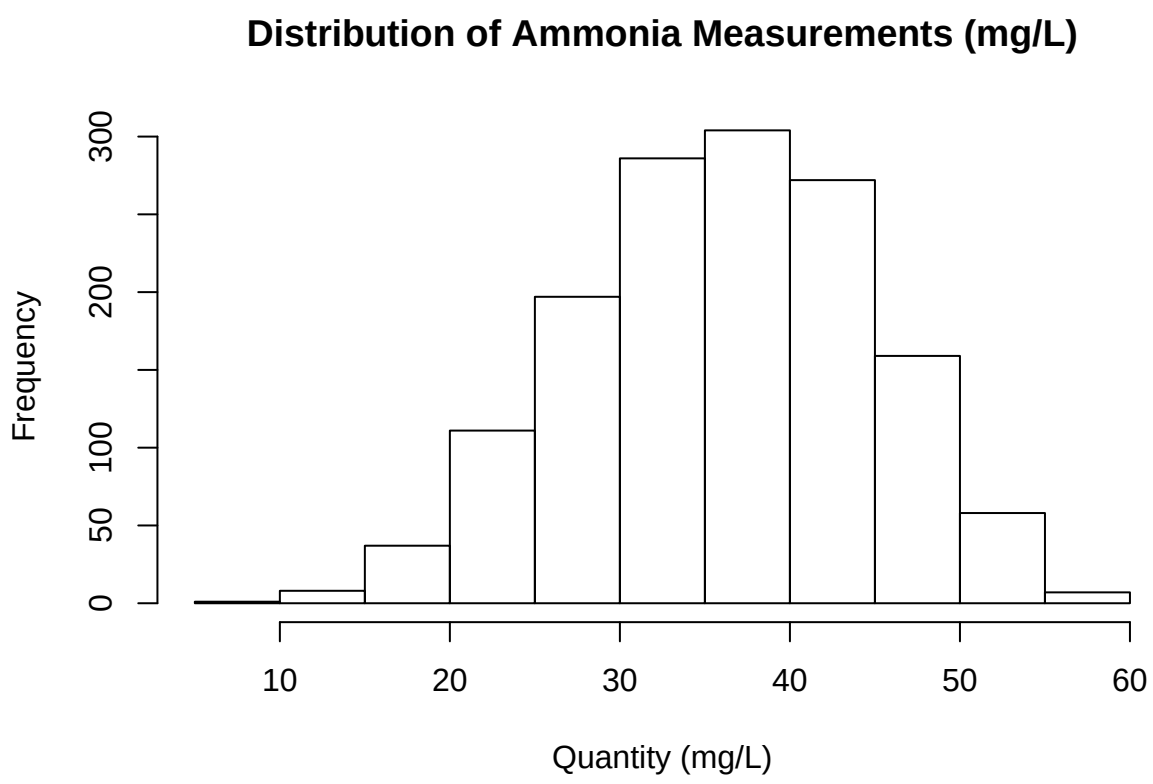
```
scatteramonia <- plot(amonia$Ammonia)
```




```
scatteramonia
```

```
## NULL
```

```
amoniahist <- hist(amonia$Ammonia,
                   main = "Distribution of Ammonia Measurements (mg/L)",
                   xlab = "Quantity (mg/L)")
```



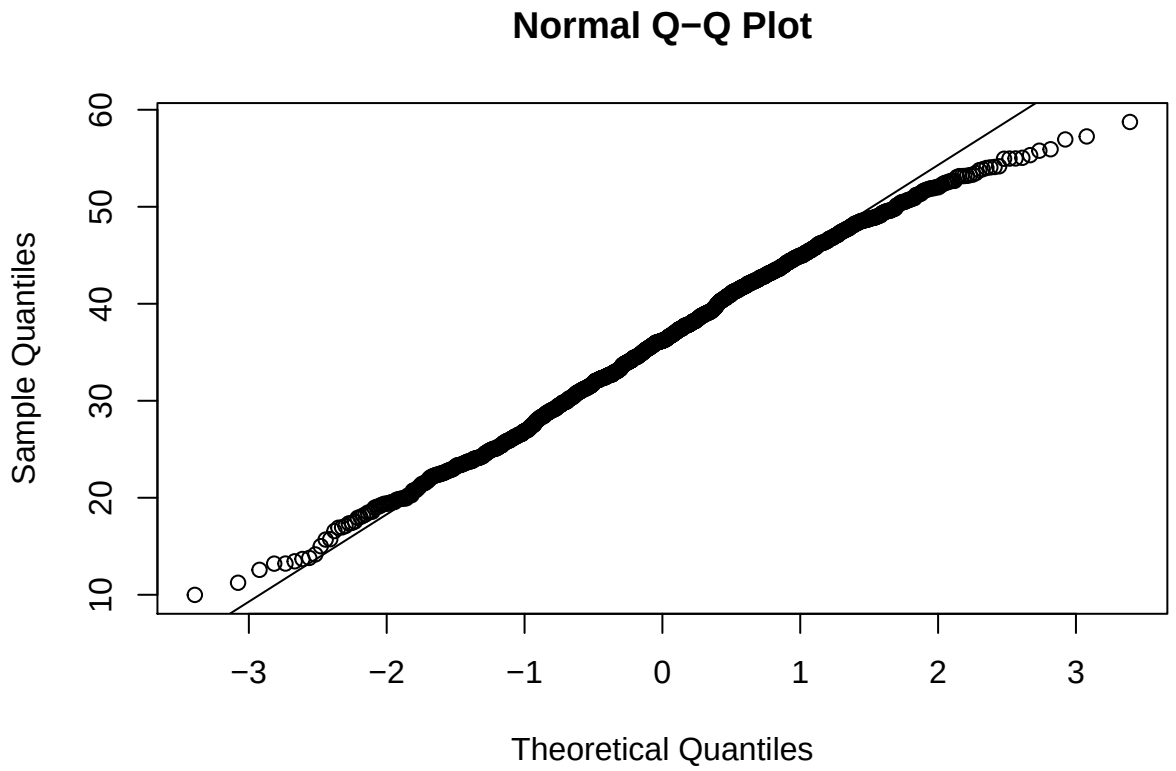

```
amoniahist
```

```
## $breaks
##  [1]  5 10 15 20 25 30 35 40 45 50 55 60
##
## $counts
##  [1]   1   8  37 111 197 286 304 272 159  58   7
##
## $density
##  [1] 0.0001388889 0.0011111111 0.0051388889 0.0154166667 0.0273611111
##  [6] 0.0397222222 0.0422222222 0.0377777778 0.0220833333 0.0080555556
## [11] 0.0009722222
##
## $mids
##  [1]  7.5 12.5 17.5 22.5 27.5 32.5 37.5 42.5 47.5 52.5 57.5
##
## $xname
## [1] "amonia$Ammonia"
##
## $equidist
## [1] TRUE
##
## attr(,"class")
## [1] "histogram"
```

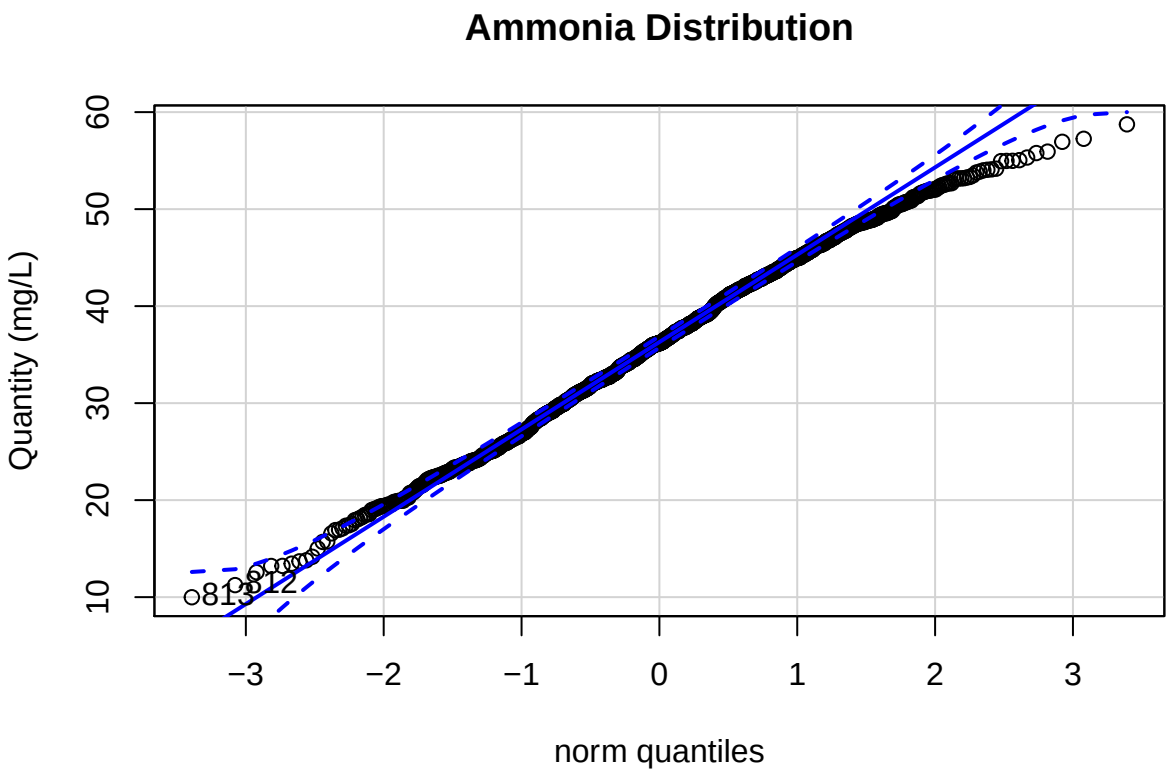
The scatter plot does not yield a visual assistance to determining distribution. However, the histogram

implies a nice, neat normal distribution. Let's confirm using qqplots...

```
qqnorm(amonia$Ammonia)
qqline(amonia$Ammonia)
```




```
qqPlot(amonia$Ammonia,
       main = "Ammonia Distribution",
       ylab = "Quantity (mg/L)")
```



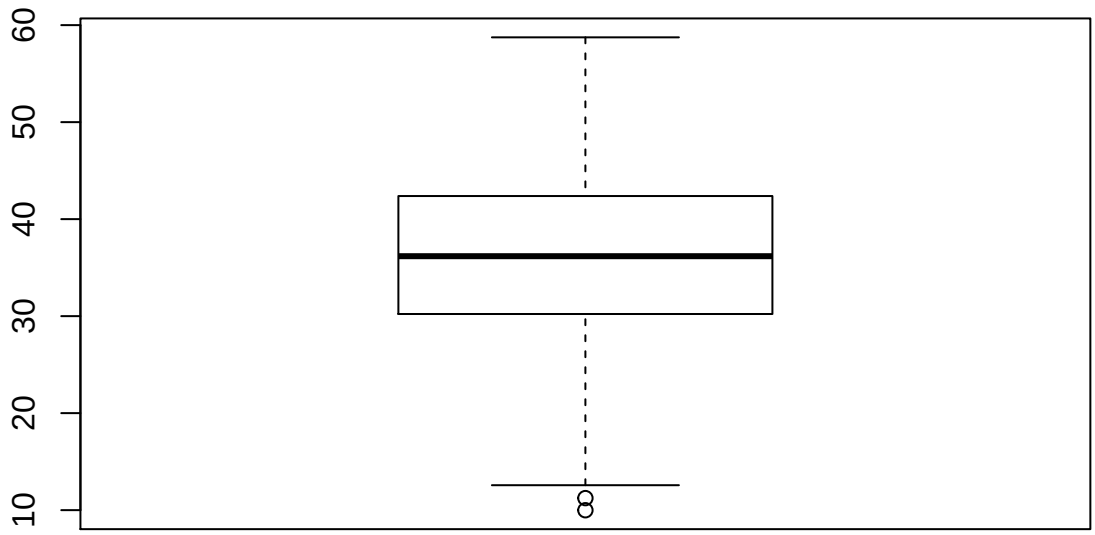

```
## [1] 813 812
```

Interestingly, there is a trail off outside the 95% confidence intervals. Outliers may exist but at least visually there seems to be a drift towards the top end of the distribution.

Question 3 - Statistics

A box plot should provide most of the statistics for us in terms of location and spread. We establish this first but will repeat the specific calculations below for confirmation. Summary() function would also yield some of the statistics for us but we infer that manual calculation is preferred in this exercise.

```
amoniabox <- boxplot(amonia$Ammonia)
```




```
amoniamean <- mean(amonia$Ammonia)
amoniamedian <- median(amonia$Ammonia)
amoniaspread <- var(amonia$Ammonia)
amoniasd <- sd(amonia$Ammonia)
amoniamean
```

```
## [1] 36.09499
```

```
amoniamedian
```

```
## [1] 36.18
```

```
amoniaspread
```

```
## [1] 72.57213
```

```
amoniasd
```

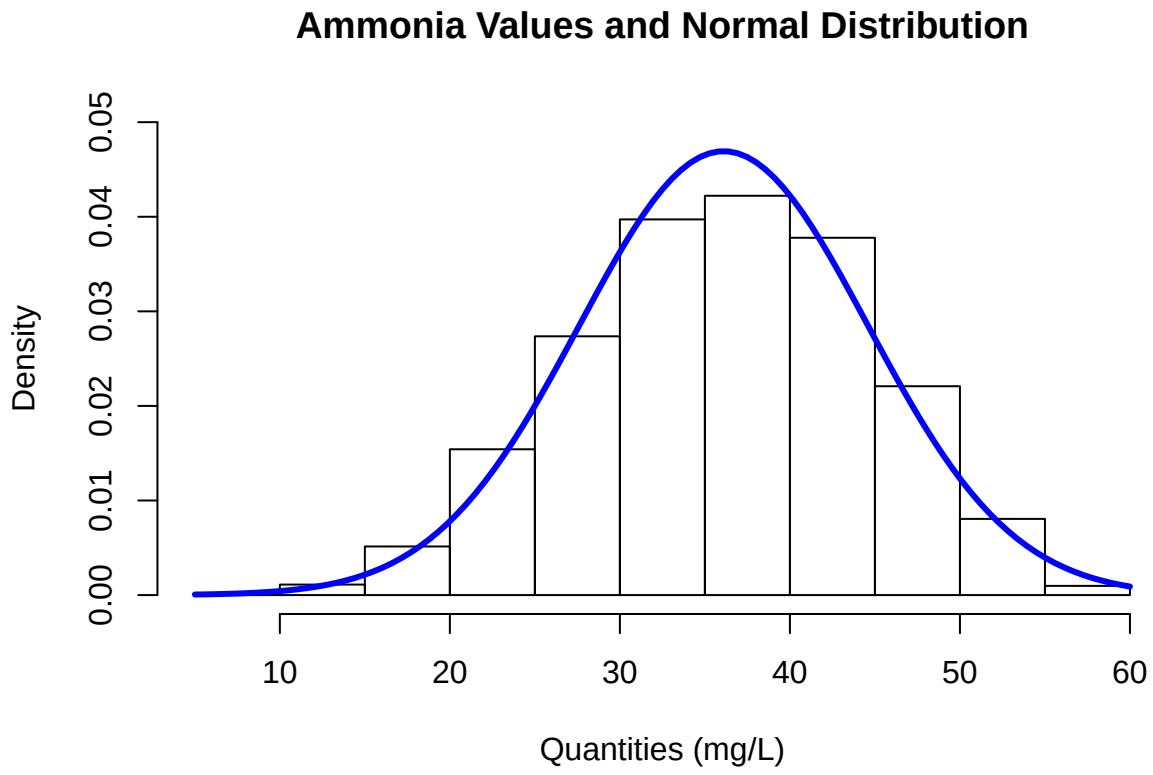
```
## [1] 8.518928
```

Question 4 - Non-Independence
If the ammonia measurements were not actually independent, we would look for the possible correlated variable. Options for this correlated variable are not provided with the exercise. However, rather than

variance, we would look for covariance between ammonia measurements and potentially related variables. P-Value and Correlation Coefficient would be important quantities to determine and establish a statistical relationship in this regard. We could refine an understanding of this relationship using methods such as Pearson or Spearman rank coefficients.

Question 5 - Normal vs T-Student Distributions
To our relatively untrained skillset, the data seem to fit the normal distribution best. To our understanding, the T-Student Distribution may actually be better for this purpose in terms of having a sample rather than a population. However, this sample is quite large and as such, the T-Student Distribution will ultimately begin to look like the normal distribution with such a large sample size. We infer this data to be a population rather than a sample. Ultimately, we believe the normal distribution is most appropriate. From the evidence of the earlier qqplots and the slight misfit shown below, it may be interesting to investigate T-Student distribution in the future. However, we are comfortable with normal distribution.

```
## plot with dnorm
histnorm <- hist(amonia$Ammonia, probability = TRUE,
                 ylim = c(0, .05),
                 main = "Ammonia Values and Normal Distribution",
                 xlab = "Quantities (mg/L)")
curve(dnorm(x, 36.1, 8.5), add=TRUE, col="blue", lwd=3)
```



Question 6 - Probability of amonia > 40mg/l
Answer using normal distribution is 32.33354%

```
great40 <- 1 - pnorm(40, mean = 36.09499, sd = 8.518928)
great40*100
```

```
## [1] 32.33354
```

Question 7 - 64.7% probability that a random reading will be less than what concentration of ammonia?

Answer using normal distribution is 39.30862 mg/l

```
less64 <- qnorm(.647, mean = 36.09499, sd = 8.518928)
less64
```

```
## [1] 39.30862
```

For some sanity check we run summary() function on the amonia data.

```
summary(amonia)
```

```
##     Ammonia
##  Min.   : 9.99
##  1st Qu.:30.23
##  Median :36.18
##  Mean   :36.09
##  3rd Qu.:42.37
##  Max.   :58.74
```