Jake Sauter
Environmental Statistics
HW 2

For all of these solutions, first the data needs to be entered

>> pollution_data = c(5.67, 1.98, 2.07, 5.53, 1.52, 1.88, 3.35, 1.30, 17.99, 16.44, 2.50, 4.40, 16.11, 0.58, 1.8, 0.16, 3.97, 1.71, 4.54, 2.54)

1) Computer arithmetic mean, geometric mean, and median for the data set. Which of these appears to be the best estimate of central location? Produce a plot containing each of these measures of entral location plotted with the kernel density of the data, and include a legend.

>> arith_mean = mean(pollution_data)

>> geo_mean = prod(pollution_data)$\hat{}$(1/length(pollution_data))

>> median = median(pollution_data)

arithmetic mean: 4.802, geometric mean: 2.831343, median: 2.52

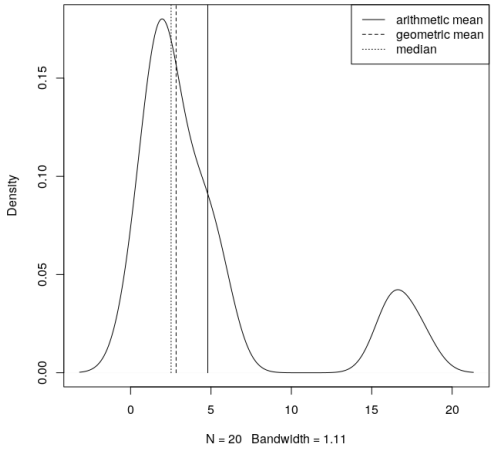
>> plot(density(pollution_data))

>> segments(arith_mean, 0, arith_mean, 1, lty=1)

>> segments(geo_mean, 0, geo_mean, 1, lty=2)

>> segments(median, 0, median, 1, lty=3)

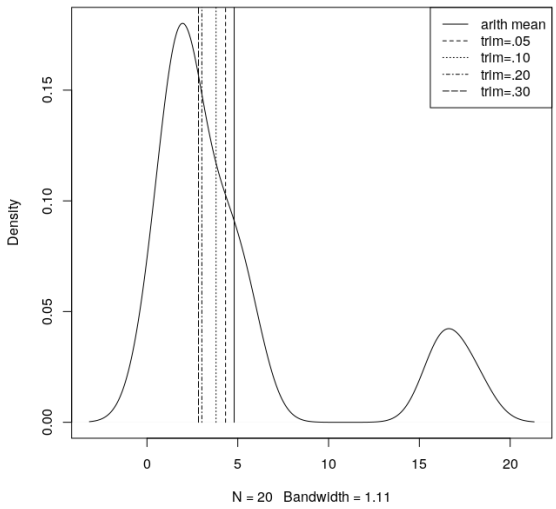
>> legend("topright", c("arithmetic mean", "geometric mean", "median"), lty=c(1,2,3))



Median appears to be the best measure of central location for the most prominent data in the data set, though if the second smaller distribution is not perceived to be outliers, one may wish to not disregard them as much as median does and use either geometric or arithmetic mean.

2) Compute the trimmed mean removing 5%, 10%, 20% and 30% of the data. What do you notice occurring with the estimate of the trimmed mean as you increase the percent of data removed. Produce a graph containing each of the trimmed means you have computed and the arithmetic mean plotted with the kernel density of the data.

```
>> trim_mean_05 = mean(pollution_data, trim=.05)
>> trim_mean_10 = mean(pollution_data, trim=.10)
>> trim_mean_20 = mean(pollution_data, trim=.20)
>> trim_mean_30 = mean(pollution_data, trim=.30)
>> plot(density(pollution_data))
>> segments(arith_mean, 0, arith_mean, 1, lty=1)
>> segments(trim_mean_05, 0, trim_mean_05, 1, lty=2)
>> segments(trim_mean_10, 0, trim_mean_10, 1, lty=3)
>> segments(trim_mean_20, 0, trim_mean_20, 1, lty=4)
>> segments(trim_mean_30, 0, trim_mean_30, 1, lty=5)
>> legend("topright", c("arith mean", "trim=.05", "trim=.10", "trim=.20", "trim=.30"),
lty=c(1,2,3,4,5))
```



As the trim parameter is increased for trimmed mean, less of the data is being used from both ends. This is causing the mean to be drawn closer to the strongest mode, however due do the closeness of the weaker mode, the strongest mode might actually be eliminated from the data set before the mean is pulled all the way to it.

3) We know the mathematical formula for the geometric mean is $GM = (x_1 X_2 \cdots x_n)^{1/n} = (\prod_{i=1}^{n} x_i)^{1/n}$, show that the geometric mean equal to the exponentiated mean of the log of the observations.

$$
\begin{aligned}
e^{\frac{\ln(x_1 x_2 \cdots x_n)}{n}} &= e^{\frac{1}{n}(log(x_1) log(x_2) \cdots (log(x_n)} \\
&= \left(e^{log(x_1)} e^{log(x_2)} \ldots e^{log(x_n)}\right)^{\frac{1}{n}} \\
&= \left(x_1 x_2 \cdots x_n\right)^{\frac{1}{n}} \\
&= (\prod_{i=1}^{n} x_i)^{1/n} \\
&= GM
\end{aligned}
$$

4) Compute the standard deviation, interquartile range, and coefficient of variation for this data set, and compare those values.

```
>> SD = sd(pollution_data)
>> quartiles = quantile(pollution_data)
>> IQR = quartiles[4] - quartiles[2]
>> CV = SD/arith_mean
```

Standard Deviation (SD): 5.413989

Interquartile range (IQR): 3.01

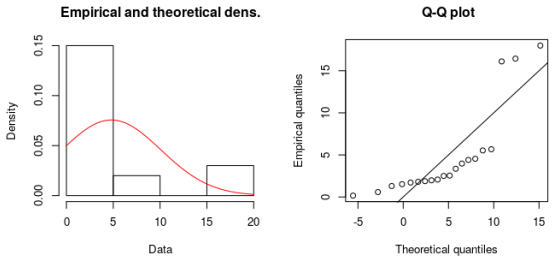
Coefficient of Variation (CV): 1.127445

All of the values seen above are measures of dispersion. Coefficient of variation (CV) and standard deviation (SD) are very similar, the only difference being that CV is divided by the sample mean, to produce a unit-less measure, while standard deviation has the same units as the data set. Both of these values are representative but not resistant. Interquartile range (IQR) however is a resistant measure, but is not very representative. We see that IQR is less than SD here, as it is more resistant to the outliers. As discussed CV is less than both measures, though it possesses no unit.

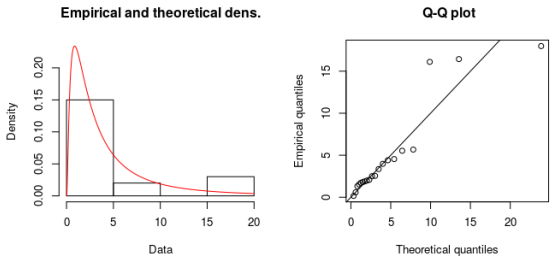
5) Using the R package *fitdistrplus*, assess the data set to determine if the observations appear to come from one of the distributions discussed in class. Explain the reasoning behind your conclusion.

>> library(fitdistrplus)

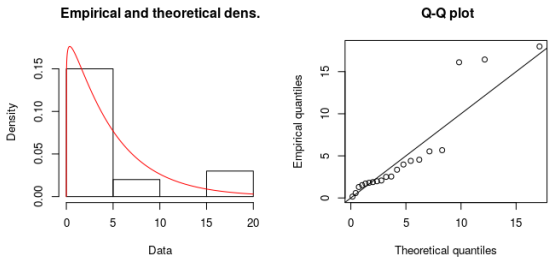
>> plot(fitdist(pollution_data, distr="norm"))



>> plot(fitdist(pollution_data, distr="lnorm"))



>> plot(fitdist(pollution_data, distr="gamma"))



From the graphs above it appears that the data comes from a log-normal distribution. This can be seen from the close match of the distribution on the left plot of the log-normal distribution. Notice that is matches better than the gamma distribution as more points lay on the best fit line on the Q-Q plot for the log-normal distribution.

6) Regardless of your answer in question 3, compute a t-distribution based 95% confidence interval for the mean of the data set. Explain why this confidence interval is, or is not appropriate for the data set.

>> t.test(pollution_data, conf.level=95)

Confidence Interval: (2.268175 7.335825)

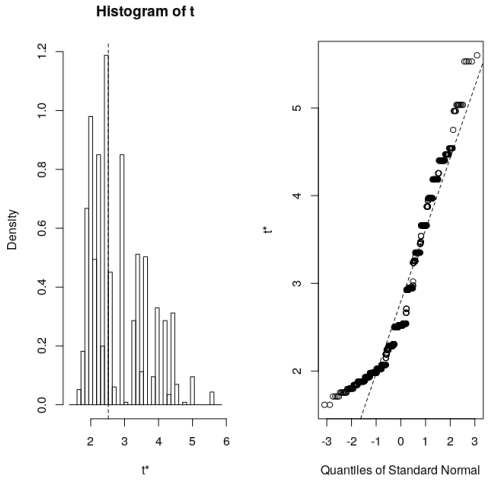
This approach is not appropriate for the data set because in using the t-distribtuion to form the confidence interval, we are making the assumption that the data follows a normal distribution with enough samples. As we saw in the previous problem it clearly does not follow the normal distribution.

7) Regardless of your answer in question 3, computer a 95% boostrap confidence interval for the median of the data set with 1000 bootstraps; state which bootstrap confidence interval you would choose and explain your reasoning. Explain why this approach is or is not appropriate.

>> boot.function = function(x,i) quantile(x[i],probs=.5)

>> medianF.boot = boot(pollution_data, boot.function, R=1000)

>> plot(medianF.boot)



>> boot.ci(medianF.boot,conf=.95)

Normal: (0.643, 3.852) Basic: (0.502, 3.240)

Percentile: (1.800, 4.538) BCa: (1.795, 4.400)

I would choose the BCa confidence interval. This would be my choice because the data does not meet the normally distributed assumption for normal confidence interval, and it does not meet the assumption of the distribution of the medians being symmetric for the percentile method. Both the basic and BCa confidence intervals make the least amount of assumptions, however the BCa method includes bias and acceleration corrections.

This approach is appropriate for the data set if one was truly interested in the median, as there are no other methods that we have seen to perform this task with the few assumption we have here. I might note that if one was just interested in a measure of central tendency, that a geometric mean (fairly resistant and explanatory) can be calculated by performing a t-test on the log of the data (which will possess the normal assumption if the data is log-normal), then exponentiating the bounds that are found for the confidence interval.