Due on Friday, 30 January 2015 by 17:00 if submitting electronically (follow the instructions carefully) or submit in class on Friday, if handing in on paper.

---

1. The Reid vapour pressure is a quality property measured on a distillation column. Data from 30 samples has a median of 67 and a MAD of 5 units. Show, if possible, how you might estimate the probability of observing a value that is 75 units or higher. Show all assumptions you make in calculating your answer.

   Solution, based on that from Mason-Hopkin-Reiche (thank you!)

   We can use the median as an estimate for the mean, and the MAD as an estimate for standard deviation, in the absence of other information, and assuming no major outliers to skew the data (the mean and median will centre around the same value).
   Then, assuming a normal distribution:
   z = xi - mean / standard deviation
   z = (75 - 67)/5 = 1.6
   find area under normal distribution when z >= 1.6
   `pnorm(1.6)` = area when z is less than or equal to 1.6, or about 94.5%
   So z is >= 1.6 for 100 - 94.5% = 5.5%

   Therefore the probability is about 5.5% of observing a value of 75 units of higher.

2. We received the 95% confidence interval from a potential supplier of sulphuric acid. The impurity levels, they stated, is 429 ppm at the lower bound, and 673 ppm at the upper bound. Their interval was based on 16 samples.
   a. What was the standard deviation of their process?
   b. We need the 99% confidence interval though. What is that confidence interval?
   c. Please explain why the values of the revised interval changed relative to the originally provided interval.

   Solution:

   a. Back calculate the standard deviation: $c_t \frac{s}{\sqrt{n}} = \frac{UB-LB}{2} = 122 \Rightarrow s = 228.9$ since the critical value is $c_t = 2.13$using the **`qt(0.025, df=15)`** in R, or from tables.
   b. At the 99% level, that leaves 0.5% in each tail, so **`qt(0.005, df=15) = -2.95`** and we know that the $\bar{x}$ is midway in the interval, so $\bar{x} = 551$from which we get the new lower bound is 382.3 and the upper board in 719.7.

c. As expected, the wider range from a higher confidence level makes sense, because a wider range has a higher probability of containing the true population value.

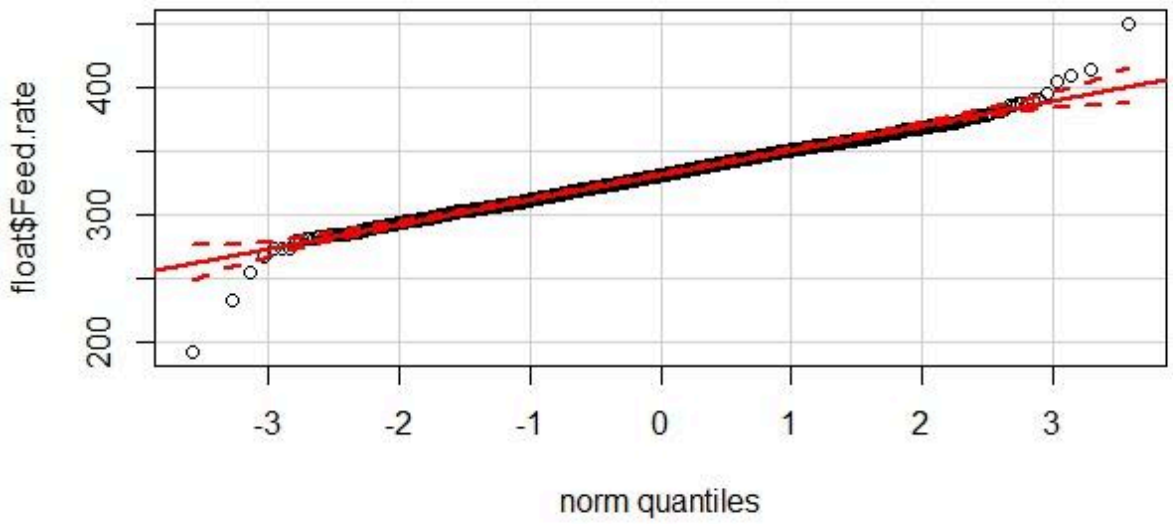
3. A student in the course had asked why a cut-off value of 2.5 standard deviations was suggested by Rousseeuw to identify outliers in a data set.
   a. If a data set was truly normally distributed with no actual outliers, how many data points out of 2922 would be falsely rejected as outliers?
   b. Load the Flotation cell data set and investigate the `Feed.rate` variable. Would you say it is normally distributed? Support your answer with proof.
   c. In relation to the prior answer, calculate and compare the median and mean. Also calculate and compare the standard deviation and MAD. Do you still believe it is normally distributed?
   d. Now attempt to answer the question: how many values in this variable exceed 2.5 standard deviations above the mean? and how many values are more than 2.5 standard deviations below the mean? Does this mostly agree with your value from part "a"?
   e. Draw a time-series plot of the data sequence. Superimpose the mean on the plot, and superimpose two other horizontal lines: one 2.5 standard deviations above, and the other 2.5 standard deviations below the mean. Use the `abline(...)` function in R to do this, if you are using R.
   f. Based on your plot in part "e", would you deem the 2.5 cut-off to be a reasonable way to detect outliers for normally distributed data?

Solution [12 points]: courtesy of: Rachel, Melissa and Sean

a. z = 2.5 for a probability of 2.5 standard deviations. In R, pnorm(2.5) = 0.994, which represents the percentage of data points within 2.5 standard deviations. To determine the number of outliers, it would be 1- 0.99 = 6.2097×10^{-3}. We have to take into consideration both tails of the normal distribution, therefore the calculated percentage is be multiplied by two. Out of 2922 data points, 1.24% would be outliers, which results in 36.2 outliers, therefore around 37 data points would be falsely rejected as outliers.

b. Normally Distributed, as less than 5 percent of the data is outside of the 95% confidence bounds.

```
qqPlot(float$Feed.rate)
```



c.

Median=331.7

Mean=331.8

SD=19.49

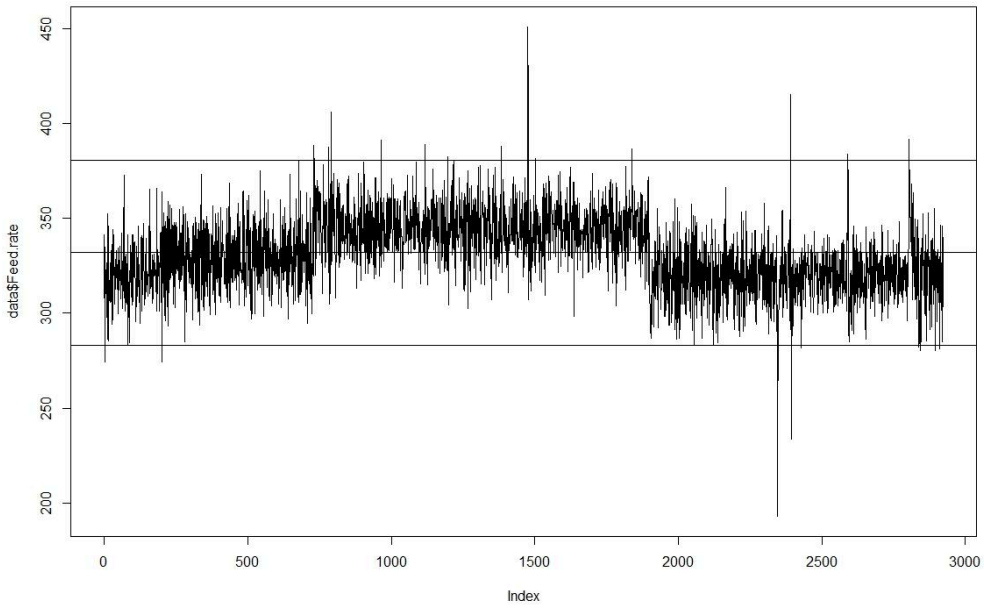
MAD=19.44

Since the mean and the median, as well as the SD and the MAD are very similar, it is determined that the data is in fact normal. If the values were significantly different it could suggest that the distribution was left or right skewed.

d. There are 28 values outside 2.5 standard deviations of the mean. 16 data points are more than 2.5 standard deviations above the mean and 12 data points are less than 2.5 standard deviations below the mean. Therefore there are less outliers than expected from a normal distribution (37 outliers, as calculated in part a). This means that we can once again confirm that our data is normal.

```
> sum(data$Feed.rate>380.525)
[1] 16
> sum(data$Feed.rate<283.075)
[1] 12
```

e.




```
plot(data$Feed.rate, type="l")
abline(h=4)
abline(h=331.8)
abline(h=380.525)
abline(h=283.075)
```

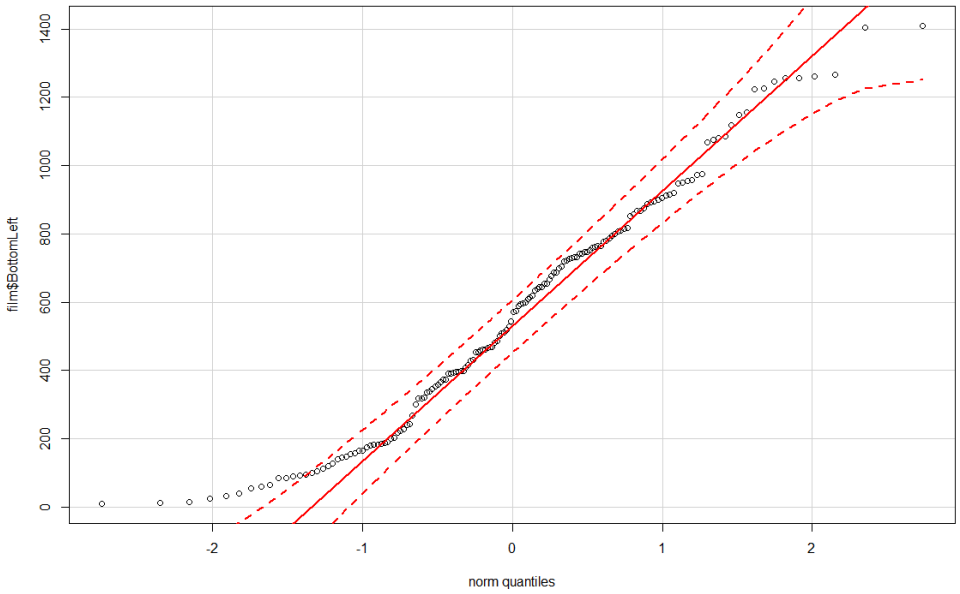
f. Yes, we feel like 2.5 standard deviations is a valid cut-off to determine normally distributed data. As we can see from the plot in part e), the majority of the data is within the bounds outlined.

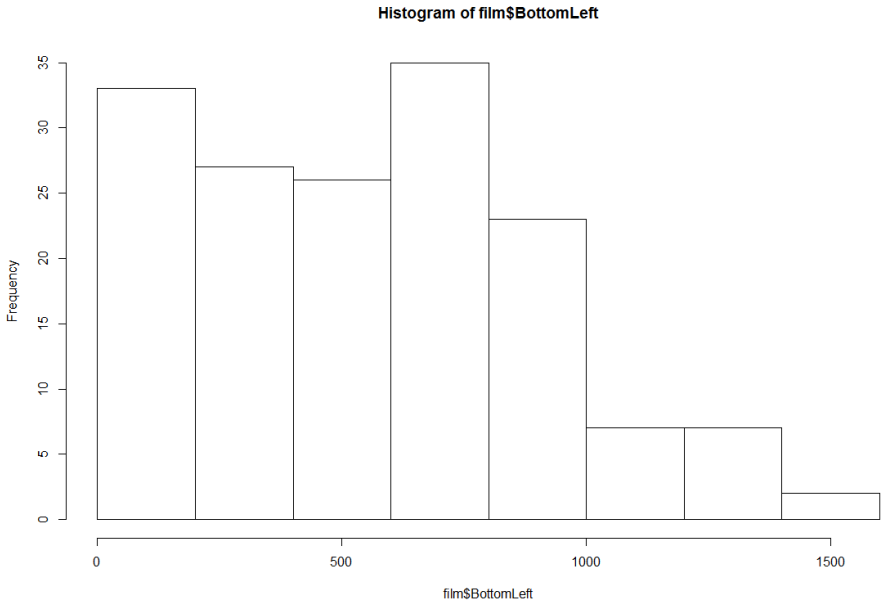
4. Load the Film thickness data set, containing the thickness of a plastic films, measured at 4 positions. It is a data set used for quality control.
   a. Which of the 4 positions on the film is normally distributed?
   b. Plot the q-q plot for the **`BottomLeft`** position. Also plot the histogram. Can you confirm what you see in the histogram in the q-q plot? In other words, describe how the histogram and the q-q plot are related to each other in this example.
   c. Now plot the time series plot for the **`BottomLeft`** position. Make sure you connect the points with a line. What trends do you notice? Are these points independent?
   d. Repeat the histogram and q-q plot for the **`TopRight`** position. What do you notice?

Solution [7 points]: courtesy of: Rachel, Melissa and Sean

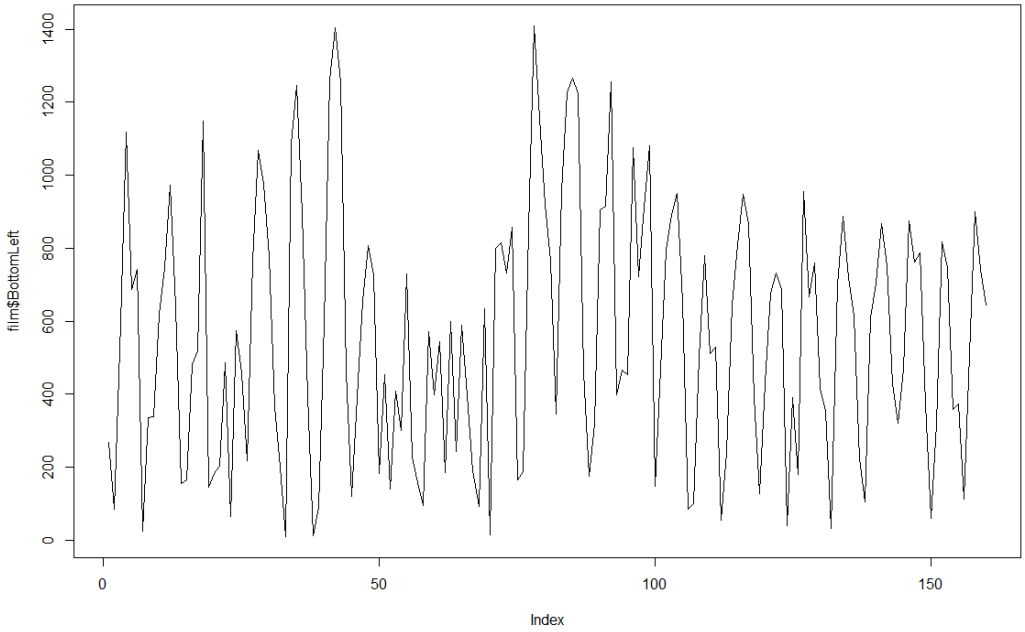
a. BottomRight (from qqPlot)

b. Yes, the data is skewed to the left. There are many outliers on the left side of the qqPlot which relates to the histogram because there are large bars on the left side.

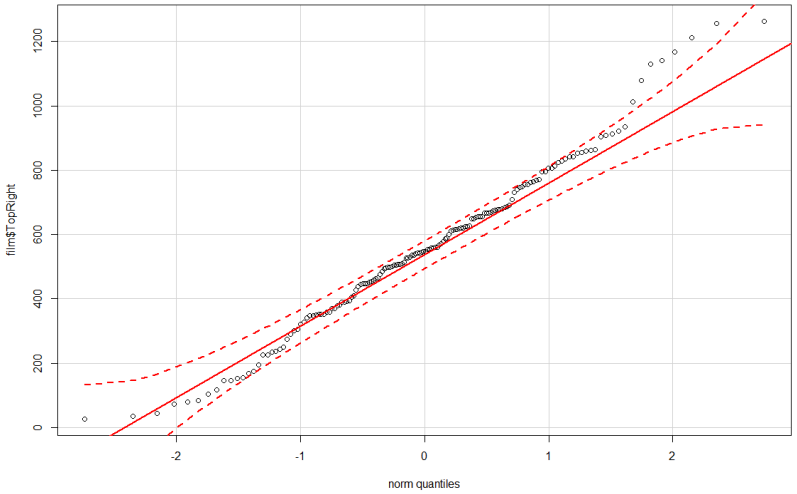


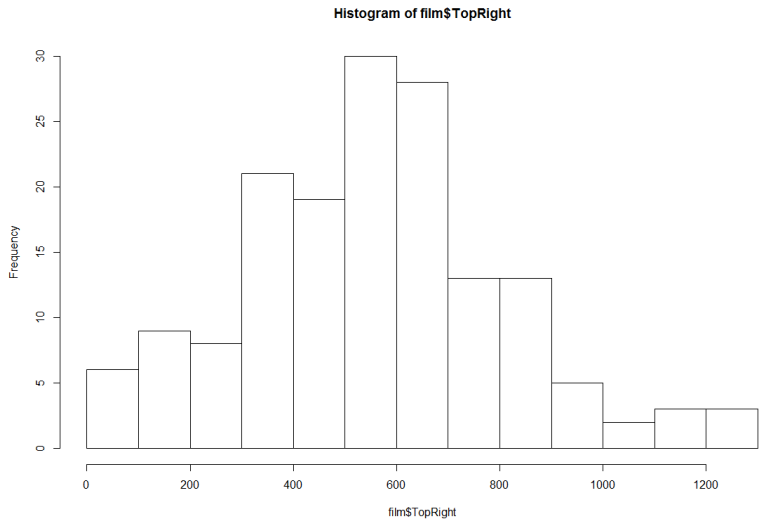


c. Trends: oscillatory, general increase then decrease. The points are not independent.


film$BottomLeft
0
200
400
600
800
1000
1200
1400
0
50
100
150
Index

d.


film$TopRight
0
200
400
600
800
1000
1200
-2
-1
0
1
2
norm quantiles



We noticed that the data looks right skewed.

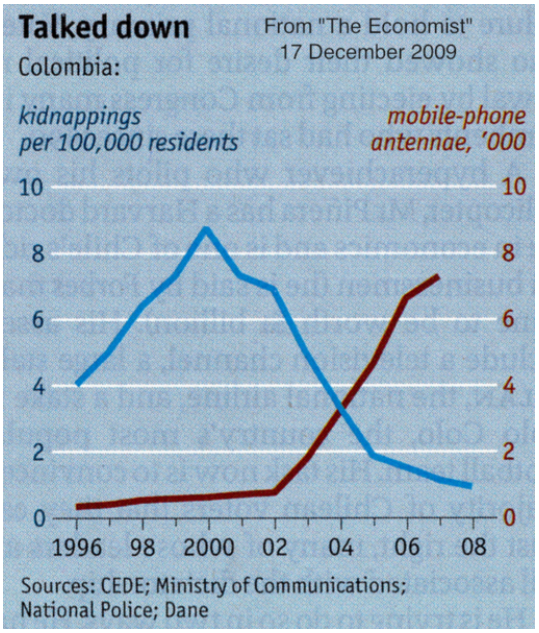
5. A forum posting in the course indicates that the concept of the central limit theorem is confusing. Without copying from the textbook or website resources, try to best explain in your own words what the central limit theorem tells us. Write your answer as if you are explaining idea idea to another final year engineer.

Solution [4 points]: courtesy of: Kima and Shabnam

The Central Limit Theorem states that if we take *n* independent samples from any distribution with finite variance and calculate their mean, the calculated mean will be from a normal distribution. In other words, if we repeat the process of taking independent samples and calculating their mean many times the set of calculated means will form a normal distribution.

As engineers we often use averages since we report properties such as density, viscosity, etc. in bulk. Although our raw data comes from unknown distributions, based on the central limit theorem the average of our raw data is from a normal distribution if the two restrictions of the theorem are met. The restriction of having a finite variance is satisfied for all practical cases but as engineers we often violate the restriction of the samples being independent.

6. Regarding the plot alongside:
   a. What type of plot is this?
   b. Describe the phenomenon displayed.
   c. Which plot type asks you to draw a cause and effect relationship? Is there a cause-and-effect here?
   d. Use rough values from the given plot to construct an approximate example of the plot you proposed in part "c".
   e. What advantage is there to the plot given here, over the type in your answer to part "c".



Solution [9]:

a. A time-series plot.

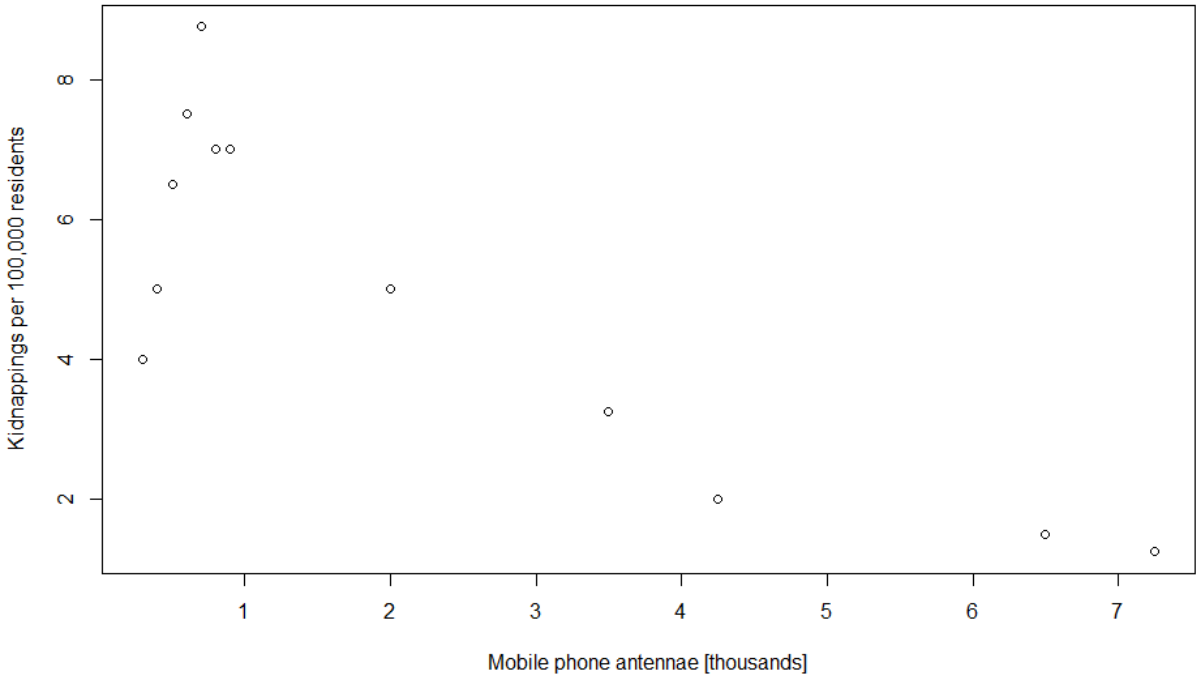
b. The rate of cellphone usage (expected to be proportional to number of mobile phone antennae) has increased in Columbia, especially since 2002. Likely this is this usual case where the price comes down, leading to greater use. Though some other political or economic change may have taken place in 2002 leading to increased phone use.

The rate of kidnappings peaked in 2000, at a rate of 8 per 100,000 residents, and has steadily decreased since that peak.

c. A scatter plot. Yes, by increase in rate of cellphone usage, the rate of kidnapping has decreased.

d. A scatter plot, from approximate values on the plot, is generated by the following code (you may use any software to construct your plot)

```
# Data from 1996 to 2007
kidnap <- c( 4, 5, 6.5, 7.5, 8.75, 7, 7, 5, 3.25, 2, 1.5, 1.25)
mobile <- c(0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 2, 3.5, 4.25, 6.5, 7.25)
plot(mobile, kidnap, type='p', xlab="Mobile phone antennae [thousands]",
          ylab="Kidnappings per 100,000 residents")
```



e. The advantage of the time-series plot is that you are able to clearly see any time-based trends - those are lost in the scatter plot (though you can recover some time-based information when you connect the dots in time order).

7. 600-level students only: answer the same questions as in question 6 above, but for the plot just a little down the page in this article, http://www.motherjones.com/environment/2013/01/lead-crime-link-gasoline, the first plot with the title "The Pb Effect".

Solution [9 points]: courtesy of: Kima and Shabnam

a. Time series

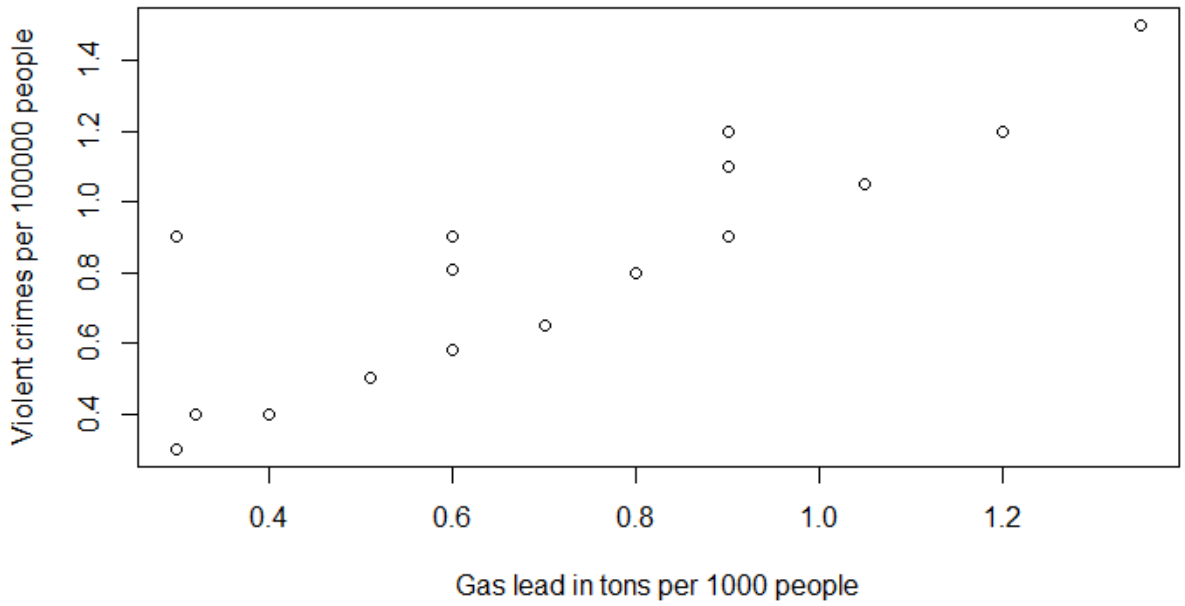
b. The blue graphs shows lead level in tones per people for the years 1937-1986. Lead level was generally increasing from 1937-1972 and after that it started decreasing.

The brown graph shows the number of violent crimes per 100,000 people for the years 1960-2009. Violent crimes were generally increasing from 1960-1995 and they started decreasing.

The two graphs combined aim to show that the increase in lead level exposure has caused an increase in crimes two decades later and also the decrease of lead exposure has caused a decrease in number of crimes.

c. The scatter plot

The cause and effect relationship is between the two graphs : Exposure to lead caused violent crimes two decades later.



d.

e.

The advantage of the given plot is that it depicts two sets of data which are two decades apart and you can compare their trend over time. The scatter plot does not give any information about the sequence of the data and by looking at it we can only understand that generally with an increase in Gas lead exposure there has been an increase in crimes. On the other hand, in the original plot because data are plotted against time, we understand that there has been two periods, a period of increase in lead exposure which has caused an increase in crimes two decades later and a period of decrease in lead exposure which has caused a decrease in crimes of two decades later.