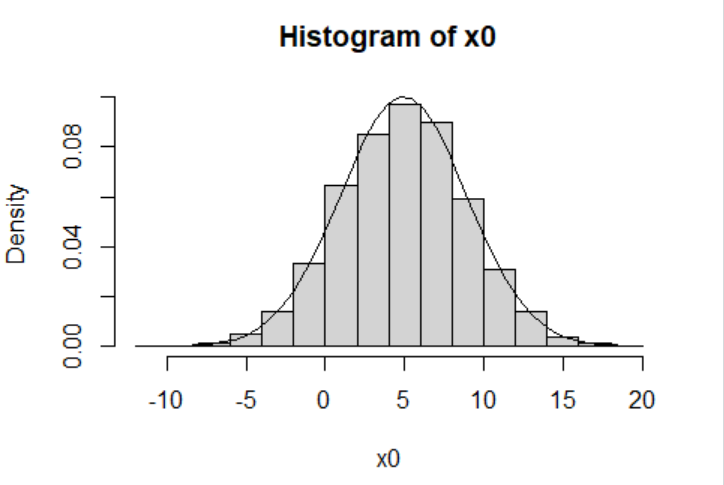
Q5.

My friend may have produce a histogram(with normal density curve), a boxplot, a scatterplot, a QQ plot and a density plot.

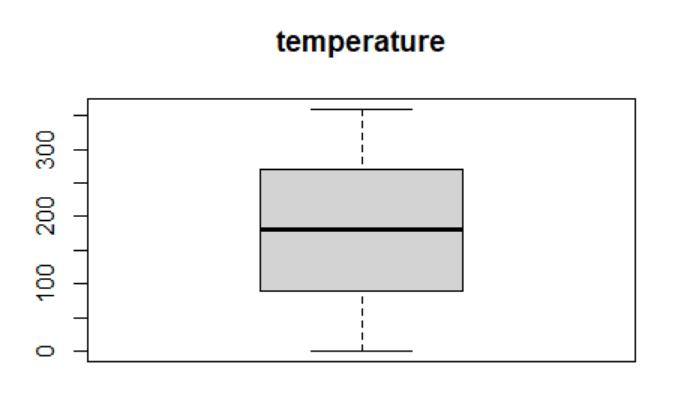
Histogram is one of the graph which we use most often. In histogram we use bar of different heights to represent different data. A histogram is good when we only need to represent one set of data.

e.g. 

To creat a histogram, we need to use hist(x, …), x represent the data we want to create a histogram and … part allow us to change the way this histogram has been represent into the way we need. For example in the diagram above I use hist(x0, freq = FALSE), x0 is the data I want to represent and freq = false means I want to represent this graph using densities instead of frequencies.

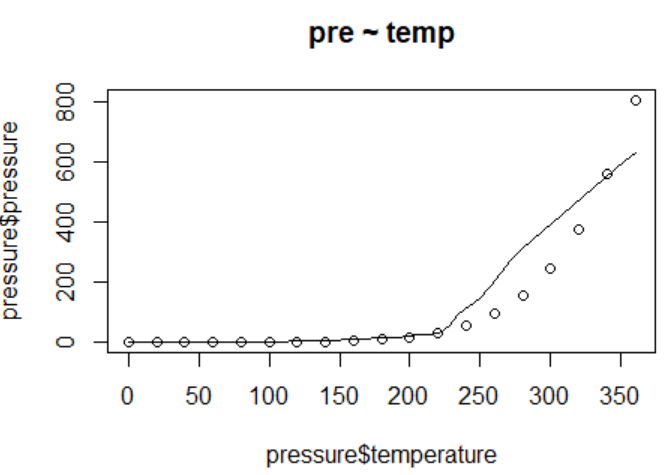
We can create a normal density curve using lines(xfit0, yfit0).

A boxplot is a plot look like a box. A boxplot group the data by their quartiles and there are also lines to represent the data outside the quartiles.

e.g. 

The way to create a boxplot is similar to the way we create a histogram. In this case I use boxplot(pressure$temperature, main="temperature"), which boxplot represent boxplot, pressure$temperture is the data I use and main=“temperature” set the title of this boxplot.

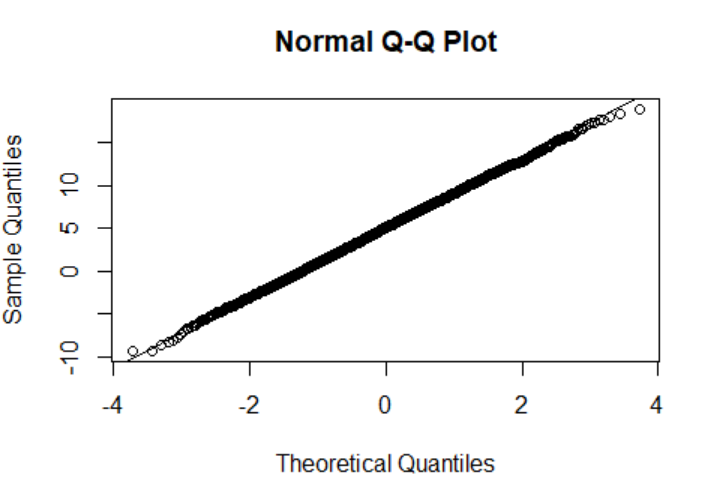
A scatterplot is a plot using dots to show data and each dot represent a single variable. This is useful when we have two set of data.

e.g. 

I create this plot using plot(pressure$temperature, pressure$pressure), which pressure$temperaturem is the x value and pressure$pressure is the y value.

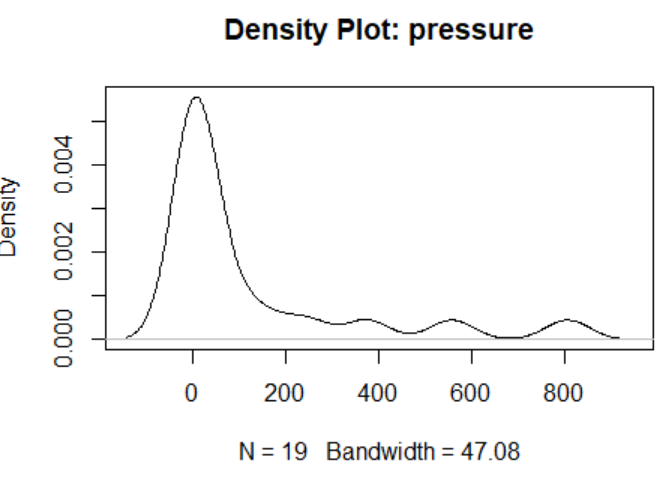
I also use scatter.smooth(x=pressure$temperature, y=pressure$pressure, main="pre ~ temp") to form a smooth line (scatter.smooth), labels for x and y (x=pressure$temperature, y=pressure$pressure )and the title(main="pre ~ temp").

A QQ plot is normally represent by dots with a line. If those dots roughly form a straight and match with the line in the graph we can say that the data we use have the same distribution.

e.g. 

In this case I use qqnorm(x0) to form the QQ plot and qqline(x0) to form the line in the plot which x0 represent the data set.

A density plot is a plot represent the density of the data set. It is similar to a histogram but instead using bars density use line to represent the whole graph.

e.g. 

To form a density plot we need to use plot(density(pressure$pressure), main="Density Plot: pressure"), plot(density) means this is a density graph and pressure$pressure is our data set, the main= “” is our title of this graph.

Q6.

p-hacking(also been called data dredging) mean when people use data analysis to make the data look like significant by keep trying until it gives a significant result and only report the significant result.

Q7.

Prosecutor fallacy is a mistake in statistics people often don’t be aware with. This happen when someone think as P(A|B) is true, P(B|A) is also true. This is called prosecutor fallacy because this often been use to prove someone is guilt for example If person A have a knife in shop B and someone got stab by this knife, if there are 10% of people have this knife than the prosecutor state there is only 10% of chance that this person is innocent. But if there are 1000 people in this town and 100 people of them have this knife the probability person A is guilty is only about 1%. So that’s why this is a fallacy. Bayes law is the way to solve this fallacy, instead of P(A|B) = P(B|A) which has been proved as wrong the formula for Bayes law is P(A|B) = (P(B|A)P(A))/P(B) to allow us to calculate the correct probability event P(A|B) might happened.

Q8.

Normal distribution is the most important distribution because its suitable for many different conditions. For example A person’s height and weight, IQ and EQ can all be represent by normal distribution. For example if the average of people’s IQ is 300 the graph will look like most people are around 300 and only a tiny amount of people is really high or really low. It’s is depend on the mean and it’s spread is depend on the standard deviation so this means that when we know the mean and standard deviation, we can say lots about a variable with this type of distribution. Normal distribution also appear in statistical theory quite a lot which means we need this normal distribution to allow us to use those statistical theory. Normal distribution also have an infinite of range this make is easier to compare with different normal distribution curves.