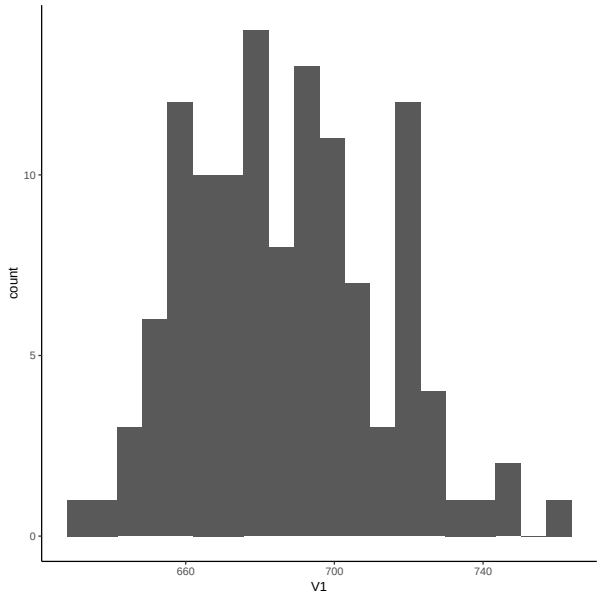
Question 1 – Rise of the machines

A semiconductor manufacturer produces devices used as central processing units in personal computers. The speed of the devices (in megahertz) is important because it determines the price that the manufacturer can charge for the devices. The files (6-42.csv) contains measurements on 120 devices. Construct the following plots for this data and comment on any important features that you notice.

(a) Histogram

Solution: Looking at the histogram it can be seen that it is slightly right skewed with a peak around 670 Megahertz. Also the histogram shows a fairly wide peak.

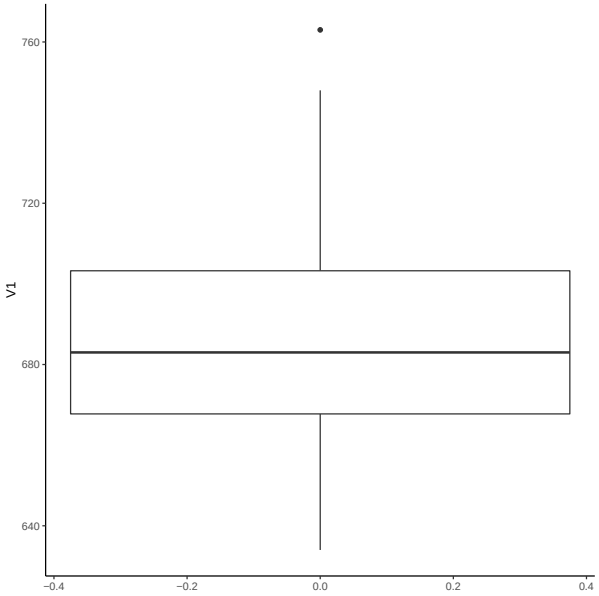
```
> library(tidyverse)
> speeds=read.csv("6-42.csv",header=FALSE)
>
> ggplot(speeds,aes(V1))+geom_histogram(bins=20)+theme_classic()
```



(b) Boxplot

Solution: As with the histogram it can be seen that there is a right skew to the data, and that this skewness is also present in the interquartile range. There is one outlier above 760 Megahertz that should be checked.

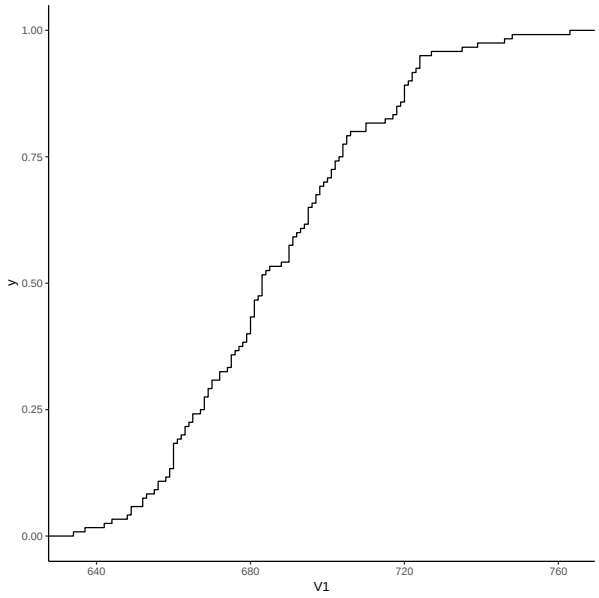
```
> ggplot(speeds,aes(y=V1)) + geom_boxplot()+theme_classic()
```



(c) Empirical cumulative distribution function

Solution:

```
> ggplot(speeds,aes(V1)) + stat_ecdf(geom="step")+theme_classic()
```



Question 2 – Samples

Consider the same data set as in Question 1, compute:

(a) the sample mean, the sample standard deviation and the sample median.

Solution:

```
> paste0("The sample mean is ",mean(speeds$V1), "  MHz.")

[1] "The sample mean is 686.775  MHz."
```

(b) the sample standard deviation

Solution:

```
> paste0("The sample standard deviation is ",sd(speeds$V1), " MHz.")
```

```
[1] "The sample standard deviation is 25.6680367235926 MHz."
```

(c) the sample median

Solution:

```
> paste0("The sample median is ",median(speeds$V1), " MHz.")
```

```
[1] "The sample median is 683 MHz."
```

As can be seen from the above summary statistics there is a slight right skewness to the data with the median being lower than the mean.

Question 3 – The thickest rod

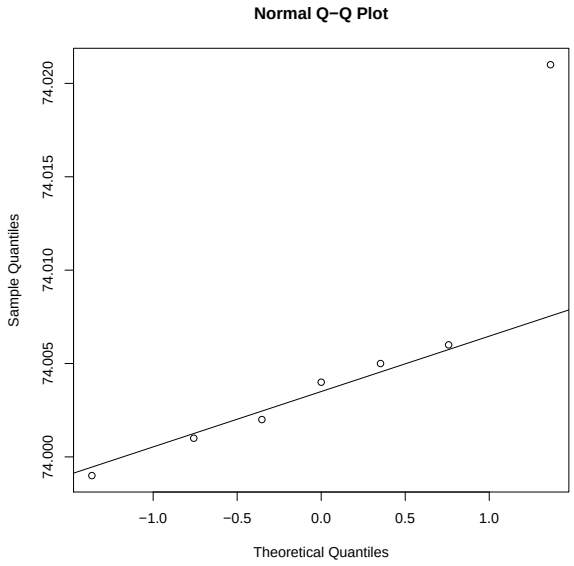
Eight measurements were made on the inside diameter of forged piston rings in an automobile engine. The data (in millimetres) is:

74.004, 73.999, 74.021, 74.001, 74.006, 74.002, 74.005.

Use R to construct a `qnorm` plot of the piston ring diameter data. Does it seem reasonable to assume that piston ring diameter is normally distributed? How about if you remove a single observation that is potentially an outlier?

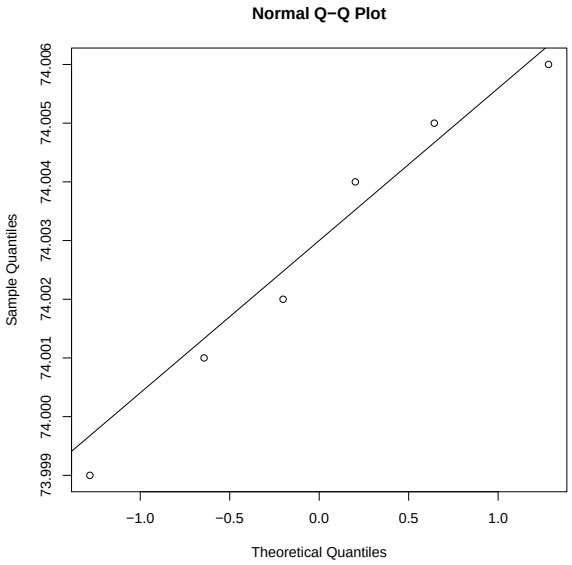
Solution:

```
> rods = tibble(thick=c(74.004, 73.999, 74.021, 74.001, 74.006, 74.002, 74.005))
> qqnorm(rods$thick)
> qqline(rods$thick)
```



As the data does not follow the line of the plot, the data can not be said to be distributed normally. There is a potential outlier at the highest point, so this should be removed and then the plot created again.

```
> rods2 = tibble(thick=c(74.004, 73.999, 74.001, 74.006, 74.002, 74.005))
> qqnorm(rods2$thick)
> qqline(rods2$thick)
```



With such a small data set it is hard to determine any properties, however a wave pattern is appearing suggesting that the data is not normally distributed.

Question 4 – A non-flat earth

In 1789, Henry Cavendish estimated the density of the Earth by using a torsion balance. His 29 measurements are in the file (*6-122.csv*), expressed as a multiple of the density of water.

(a) Calculate the sample mean, sample standard deviation, and median of the Cavendish density data

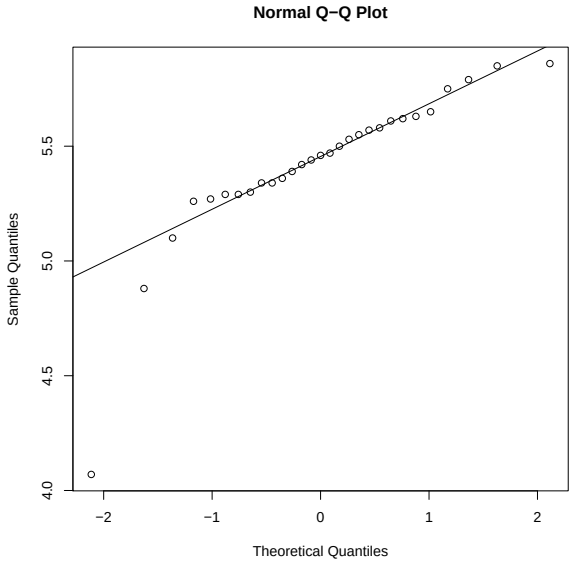
Solution: First read in the data from the file provided using `read.csv` then run the R functions for mean, standard deviation and median.

```
> cavendish = read.csv("6-122.csv",header=FALSE)
> paste0("The sample mean is ", mean(cavendish$V1))

[1] "The sample mean is 5.41965517241379"

> paste0("The sample standard deviation is ",sd(cavendish$V1))

[1] "The sample standard deviation is 0.338879274316941"

> paste0("The sample median is ",median(cavendish$V1))

[1] "The sample median is 5.46"
```

(b) Construct a `qnorm` plot of the data. Comment on the plot. Does there seem to be a "low" outlier in the data?

Solution: Using the function defined in Question 4 the following plot is obtained

```
> qqnorm(cavendish$V1)
> qqline(cavendish$V1)
```



There does appear to be a low outlier on the plot at approximately -4. Also the plot would be linear in the middle if this were removed with some deviation towards the end of the line.

(c) Would the sample median be a better estimate of the density of the earth than the sample mean? Why?

Solution: With the presence of an outlier in the data the median would be a better estimate of centre as it is robust against the presence of outliers.

Question 5 – Choice of Sample Size

The Charpy V-notch (CVN) technique measures impact energy. Assume that the impact energy is normally distributed with $\sigma = 1$ Joules. How many specimens must be tested to ensure that the error between the sample mean and the true mean is at most 0.5 with a confidence of 95%?
Solution: To work out the sample size require first recall the formula for the confidence interval.

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

The part right of the $\pm$ is called the margin of error, and since we know that we want this to be 0.5 we can work out the number of samples required by rearranging this part of the formula. This is shown below:

$$\begin{aligned}
\text{error}(\bar{x} - \mu) &= z^* \frac{\sigma}{\sqrt{n}} \\
0.5 &= 1.96 \frac{1}{\sqrt{n}} \\
\sqrt{n} &= 1.96 \frac{1}{0.5} \\
n &= (2 \times 1.96)^2 \\
&= 15.3664
\end{aligned}$$

This gives us the minimum number of samples needed to get this margin of error. As it is not an integer we need to round up to the nearest integer. So the samples size needed is 16 specimens.

Question 6 – Seeing the CLT with simulation

Consider the following random variables

$$V \sim \text{Exp}(2) \text{ (exponential distribution)}$$
$$W \sim \text{G}(0.4) \text{ (geometric distribution)}$$

(a) What is the mean and variance of each?

Solution:

Exponential

The mean of an Exponentially distributed variable is given by

$$\mu_V = E(V) = \frac{1}{\lambda} = \frac{1}{2} = 0.5$$

The variance is given by

$$\sigma_V^2 = Var(V) = \frac{1}{\lambda^2} = \frac{1}{2^2} = 0.25$$

Geometric

The mean of a Geometric distributed variable is given by

$$\mu_W = E(W) = \frac{1}{p} = \frac{1}{0.4} = 2\frac{1}{2}$$

The variance is given by

$$\sigma_W^2 = \frac{1-p}{p^2} = \frac{1-0.4}{0.4^2} = 3\frac{3}{4}$$

Consider now,

$$S_n = \sum_{i=1}^{n} X_i,$$

where X_i is either V_i or W_i (distributed as V or W) and different X_i are assumed independent.

(b) What is the mean of S_i? Answer this separately for V and W.

(c) What is the variance of S_i? Answer this separately for V and W.

Solution: In all cases this is a sum of random variables and so

$$\begin{aligned}\mu_{S_i} = E(S_i) &= E(\sum_{i=1}^{n} X_i) \\ &= \sum_{i=1}^{n} E(X_i)\end{aligned}$$

$$\begin{aligned}\sigma_{S_i}^2 = Var(S_i) &= Var(\sum_{i=1}^{n} X_i) \\ &= \sum_{i=1}^{n} Var(X_i)\end{aligned}$$

As the variables will be independent identically distributed (i.i.d.) this leads to

$$\mu_{S_i} = E(S_i) = nE(X)$$
$$\sigma_{S_i}^2 = Var(S_i) = nVar(X)$$

Exponential

Using the above working this means if S_i is made of Exponential random variables V then

$$\mu_{S_i} = E(S_i) = nE(V) = 0.5n$$
$$\sigma^2_{S_i} = Var(S_i) = nVar(V) = 0.25n$$

Geometric

Using the above working this means if S_i is made of Geometric ranodm variables W then

$$\mu_{S_i} = E(S_i) = nE(W) = 2\frac{1}{2}n$$
$$\sigma^2_{S_i} = Var(S_i) = nVar(S_i) = 3\frac{3}{4}n$$

Question 7 – The CLT with simulation

Let X_i be independent and exponentially distributed with parameter 10, that is $X_i \sim \text{Exp}(10)$, for $i = 1, \ldots, n$. Define

$$\tilde{Z}_n = \frac{S_n - E(S_n)}{\sqrt{var(S_n)}}.$$

(a) What distribution has $\tilde{Z}$ for large n? (*Hint: Use the Central Limit Theorem.*)

Solution: Calculating the expected value for $\tilde{Z}$ we get

$$\begin{aligned} E(\tilde{Z}) &= E\left(\frac{S_n - E(S_n)}{\sqrt{var(S_n)}}\right) \\ &= \frac{E(S_n) - E(S_n)}{\sqrt{var(S_n)}} \\ &= 0 \end{aligned}$$

as $E(S_n)$ and $var(S_n)$ would be constants for this random variable. Similarly calculating the variance we get

$$\begin{aligned} var(\tilde{Z}) &= var\left(\frac{S_n - E(S_n)}{\sqrt{var(S_n)}}\right) \\ &= \frac{var(S_n)}{\left(\sqrt{var(S_n)}\right)^2} \\ &= 1 \end{aligned}$$

So the variable $\tilde{Z}_n$ will have a mean of zero and variance of 1. This holds for all n. As n increases the CLT states that this random variable gets closer to the standard Normal Distribution.

(b) Generate Monte Carlo estimates of $P\left(\left|\tilde{Z}_n\right| > 2.0\right)$ using no less than 10^6 generations of $\tilde{Z}_n$ for every n.Compare your results to $P(|Z| > 2.0)$ taken from a normal distribution table, where Z is a standard normal variable. Do this for $n = 2, 10, 20$. Tabulate your results neatly and explain your results.

Solution: The following R code will produce the required Monte Carlo simulations with the distributions being in rows and the number of points increasing left to right.

```
> lambda = 10
> meanexp = 1/lambda
> varexp = 1/lambda^2
> j=1
> z=vector()
> for (n in c(2,10,20)) {
+         z[j]=mean(abs(colSums(replicate(10^6,rexp(n,rate=lambda)))-
+         n*meanexp)/sqrt(n*varexp) > 2)
+         j=j+1
+ }
> print(z)

[1] 0.046917 0.041452 0.043076
```

Question 8 – Sample Mean and Sample Variance

Suppose that a sample of size $n = 20$ is selected at random from a normal population with mean 100 and standard deviation 8.

(a) Calculate $P(98 \leq \bar{X} \leq 102)$.

Solution: To calculate this probability we need to find the difference of the cdf between the upper and lower bounds. To do this we need to standardise to the standard normal distribution and then look up the values in a table.

$$\begin{aligned}
P(98 \leq \bar{X} \leq 102) &= P(\bar{X} \leq 102) - P(\bar{X} \leq 98) \\
&= P\left(\frac{\bar{X} - 100}{\frac{8}{\sqrt{20}}} \leq \frac{102 - 100}{\frac{8}{\sqrt{20}}}\right) - P\left(\frac{\bar{X} - 100}{\frac{8}{\sqrt{20}}} \leq \frac{98 - 100}{\frac{8}{\sqrt{20}}}\right) \\
&= P(Z \leq 1.118034) - P(Z \leq -1.118034) \\
&= 0.8682 - 0.1318 \\
&= 0.7364
\end{aligned}$$

(b) Find x such that $P\left(\left|\bar{X} - 100\right| > x\right) = 0.01$

Solution: Here again we need to standardise the distribution to the standard normal distribution and then look up this value in the tables, remembering that as we are looking for a two-sided probability that we need to multiply the value in the table by two.

$$\begin{aligned}
P\left(\left|\bar{X} - 100\right| > x\right) &= 0.01 \\
P\left(\frac{\left|\bar{X} - 100\right| - 0}{\frac{8}{\sqrt{20}}} > \frac{x - 0}{\frac{8}{\sqrt{20}}}\right) &= P(|Z| > 2.5758) \\
\frac{x}{\frac{8}{\sqrt{20}}} &= 2.5758 \\
x &= \frac{8}{\sqrt{20}} \times 2.5758 \\
x &= 4.6077
\end{aligned}$$