

The five-number summary and the box plot

Sunday, 10 February 2019 4:07 pm

★ By the end of the lesson I would hope that you have an understanding and be able to apply to questions the following concepts:

- What does it mean to have a five number summary
- What does the five number summary tell us
- What is a box plot
 - How do we create a box plot
 - How do we read a box plot
- How do we identify outliers on box plots
- How do I use my CAS to draw a box plot

RECAP

This whole section of the Further course requires us to learn how to display and analyse data. There are a number of ways to describe data in terms of how it looks and where important numbers might lie.

Mostly, we do this to help us describe data and to compare it.

So far, we have been dealing with:

- The median
- The range
- The interquartile range

Ignoring the range (as it can be susceptible to problems when there are outliers) we can use the **median** and the data which gives us the **interquartile range** to help us describe the key points of our data.

The daily five

It's now a "thing" that we should, to be healthy, be eating 5 pieces of fruit and vegetables a day!



Whist, not as good for you, Mathematics has 5 things we need to do to describe some data. It's called the **five-number summary**.

The five-number summary

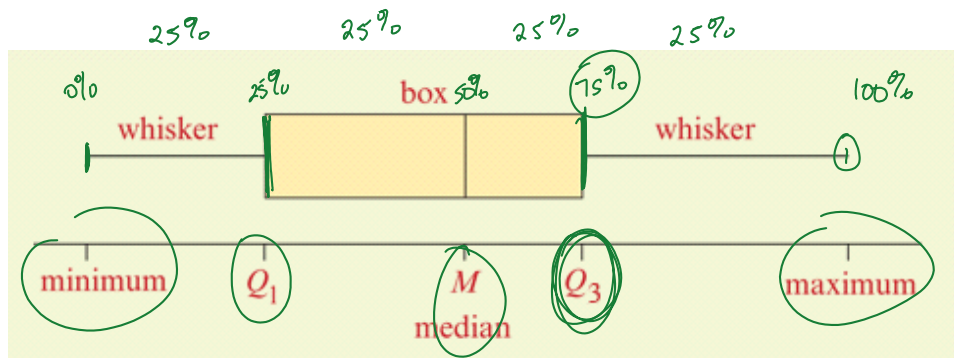
The following are the MOST important things in this chapter (so far). They help us describe some very important things about a data set.

- ✓ Minimum value
- ✓ Lower quartile (Q_1)
- ✓

- ✓ Minimum value
- ✓ Lower quartile (Q_1)
- ✓ Median (M)
- ✓ Upper quartile (Q_3)
- ✓ Maximum value

Why do I need to know this?

You're about to meet the most awesome of pictures!
 Welcome to the **boxplot!!!!**
 It's also called the box and whisker plot.



THE MOST IMPORTANT THING TO KNOW IS: Each section of the boxplot shows 25% of the data.
 Hence, each section stands for a quarter (or quartile).

So many people don't know or understand this!

Drawing a box plot

To be able to draw one of these we (obviously) need some data!

Having the data, we then calculate (or find) the following:

- Minimum value
- Lower quartile (Q_1)
- Median (M)
- Upper quartile (Q_3)
- Maximum value

We then draw a **sensible scale**.
 Finally, we draw the box plot.

Example:

Extracted from the Cambridge Further Mathematics Textbook

The stem plot shows the distribution of life expectancies (in years) in 23 countries.

Key: 5|2 = 52 years
 5 | 2 ← minimum
 5 | 5 6



✓ Min = 52
 ✓ LQ = 66

1

- ✓ Min = 52
- ✓ LQ = 66
- ✓ Med = 73
- ✓ UQ = 75
- ✓ Max = 77

A hand-drawn box plot on a horizontal axis labeled 'life expectancy (years)'. The axis has tick marks at 50, 55, 60, 65, 70, 75, and 80. The box plot shows a minimum at approximately 52, a first quartile (Q1) at approximately 63, a median at approximately 73, a third quartile (Q3) at approximately 77, and a maximum at approximately 79.

Using the CAS to draw a boxplot

Example:
Extracted from the Cambridge Further Mathematics Textbook

28	21	21	3	22	31	35	26	27	33
43	31	30	34	48	36	35	23	24	

How did the calculator know what an outlier is?

How did the calculator know what an outlier is?

We are pretty lucky, being humans, as we can generally tell which items might be outliers.

But ... sometimes we might get it wrong just using our eyes.

There must be a better way to define what an outlier is!

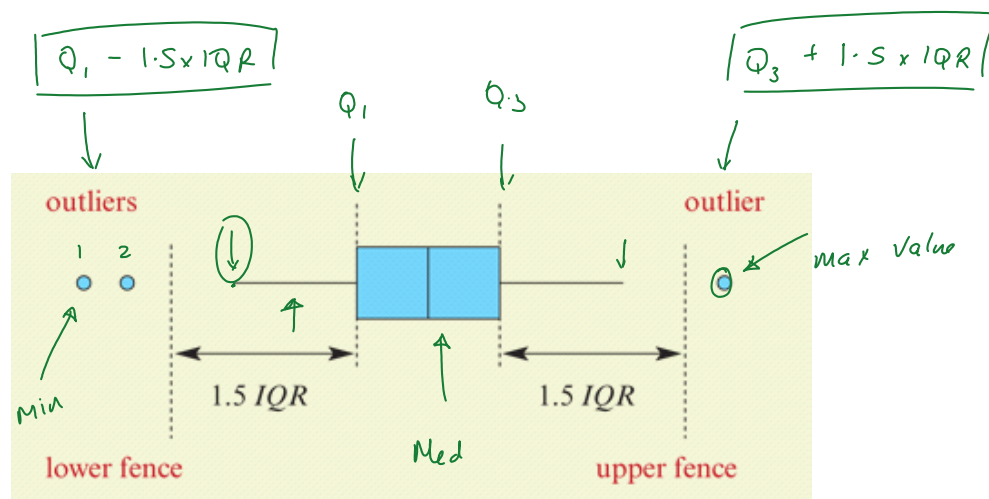
And there is!

Barry has defined an outlier to be the following:

- Greater than $Q_3 + 1.5 \times IQR$
- Less than $Q_1 - 1.5 \times IQR$

He has even given them names - to continue to confuse us!

- Upper fence
- Lower fence



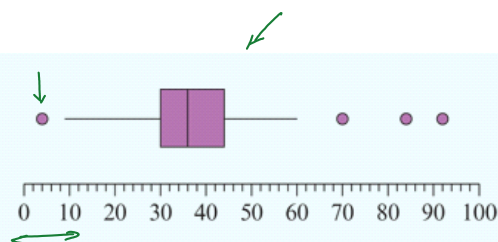
Knowing what they are ... how do I read them?

Always good to learn through examples!

Example taken from the Cambridge ^{Further} ~~Essentials~~ Textbook Series

For the box plot shown, write down the values of:

- the median 36
- the quartiles Q_1 and Q_3 $Q_1 = 30$ $Q_3 = 44$
- the interquartile range (IQR) $IQR = 44 - 30 = 14$
- the minimum and maximum values $Min = 4$ $Max = 92$
- the values of any possible outliers $(4), 70, 84, 92$
- the smallest value in the upper end of the dataset that will be classified as an outlier
- the largest value in the lower end of the dataset that will be classified as an outlier.



Before answering any questions, we need to know how to read the boxplot.

It's always a good idea to just understand the diagram first, before answering any questions!

$$\begin{aligned} \text{Upper fence} &= Q_3 + 1.5 \times IQR \\ &= 44 + 1.5 \times 14 \end{aligned}$$

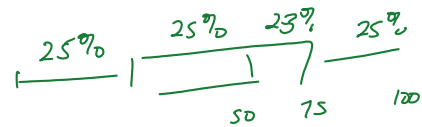
$$\begin{aligned}
 &= 65 \\
 \text{Lower Fence} &= Q_1 - 1.5 \times IQR \\
 &= 30 - 1.5 \times 14 \\
 &= 9 \\
 &=
 \end{aligned}$$

Trying to trick us with Percentages (or percentiles)

Barry really has done his finest work with Further Mathematics!

If we understand that the box plots are nothing more than sections which show where 25% of the data lies, there is nothing he can throw at us which will confuse.

Example taken from the Cambridge ^{Further Maths} Essentials Textbook Series



For the box plot shown, estimate the percentage of values:

- a less than 54 25%
- b less than 55 50%
- c less than 59 75%
- d greater than 59 25%
- e between 54 and 59 50%
- f between 54 and 86. 75%

