Boxplots (8–9)

Contents

introduction

Boxplots (or Box & Whisker diagrams) are a useful way of comparing data e.g. boys' heights against girls' heights.

1 How to draw a boxplot

To draw a boxplot, we need 5 key pieces of data:

- The lowest data value
- The lower quartile Q_1
- The median Q_2
- The upper quartile Q_3
- The highest data value

The lowest and highest data values are easy to find, but how do we find the quartiles Q_1, Q_2 and Q_3?
The quartiles are found one quarter, half and three quarters of the way through a set of data. If the data set is small, we can simply count along and find these three positions:

$$\begin{array}{cc|cc|cc|cc} 3 & 4 & 6 & 8 & 9 & 9 & 10 & 12 \end{array}$$

$$Q_1 = 5 \quad Q_2 = 8.5 \quad Q_3 = 9.5$$

Otherwise, we can use the method below to locate the position of the term we require and then count through to find this term:

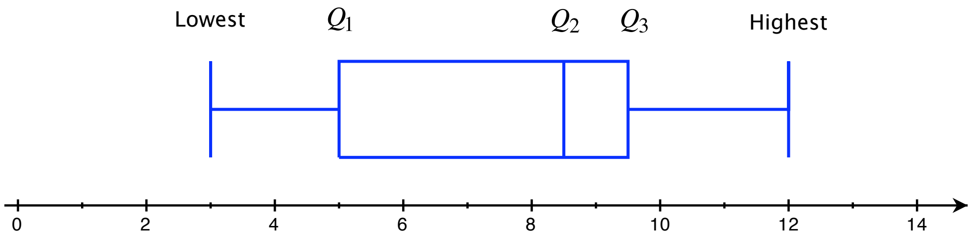
Assuming n is the number of data values (in the above example 12)

Q_1 Find $\frac{n}{4}$. If this is a whole number, round to the next half. If it is a decimal, round to the next whole. E.g. $\frac{1}{2}4 = 3\ldots$ so round to 3.5 (we picked the term between position 3 and position 4 above i.e. between 4 & 6)

Q_2 Do the same but using $\frac{n}{2}$.

Q_3 Do the same but using $\frac{3n}{4}$.

The boxplot is then drawn as follows

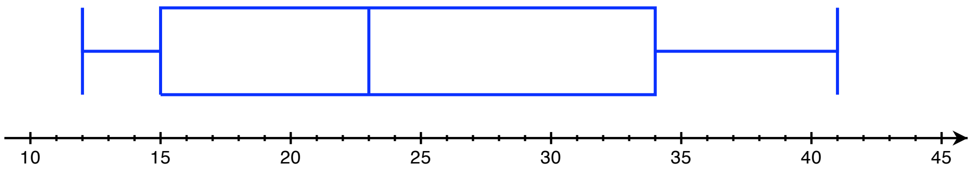


Example. Draw a boxplot to show the following data:

$$12 \quad 14 \quad 15 \quad 15 \quad 16 \quad 21 \quad 23 \quad 27 \quad 31 \quad 34 \quad 35 \quad 36 \quad 41$$

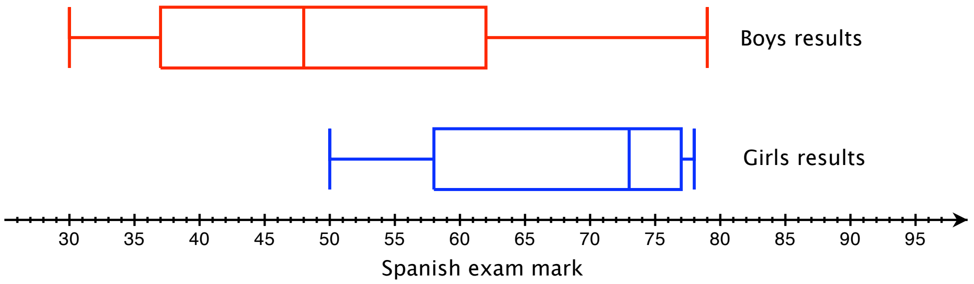
- Lowest$=12$.
- $Q_1 : \frac{13}{4} = 3.25$ so we need the 4th position, which is 15.
- $Q_2 : \frac{13}{2} = 6.5$ so we need the 7th position, which is 23.
- $Q_3 : \frac{3\times 13}{4} = 9.75$ so we need the 10th position, which is 34.
- Highest $= 41$.

This gives us the following boxplot:



2 Comparing boxplots

We need to be able to discuss what boxplots show. Look at these boxplots that show the test results of boys and girls in a Spanish exam:



We shouldn't make sweeping comments that are untrue. E.g.

Girls did better than boys in Spanish

This comment is not true since not all girls did better than all boys e.g. the lowest girl scored 50 but the highest boy scored 79, so here is an example of a boy that did better. We should try to focus our comments, talking about average and spread in each case. The key measures of average and spread used are:

Average: Mean, median or mode.

Spread: Range, Interquartile range (IQR $= Q_3 - Q_1$).

If we have boxplots, it is sensible to talk about the median as an average, since this is shown on the boxplot. For the spread, either is acceptable. However, if you have one or a few really high or really low values, the interquartile range "chops" these off, concentrating on the middle 50% of the data. E.g. here we have one girl who got 50%, much lower than most and one boy who got 79%, much higher than most, so we may wish to look at the middle 50%. A convenient way to focus your comment is to use the following headings:

Average	Spread
Mathematical statement. On average, girls scored more highly than boys in Spanish. **Evidence.** This is shown by a median of 73% for girls but only 48% for boys. **Real life Meaning.** On average, girls were more knowledgeable about the material in this Spanish exam and may have revised harder than the boys.	**Mathematical statement.** The middle 50% of marks are less spread for girls than for boys. **Evidence.** This is shown by an IQR of 18 for girls and 23 for boys. **Real life Meaning.** The marks are more consistent for girls, meaning that girls were scoring more similarly to one another. There was a bigger variety of boys' marks: perhaps some had revised well but others had done nothing at all.