Statistics

I

Measures of Averages

- Mean :
- mode
- Median

Measures of variation

- range
- variance
- standard deviation
- interquartile range.

Averages

The mean is the sum of the values divided by the total number of values. The symbol $\bar{X}$ represents the mean.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n} = \frac{\sum X}{n}$$

e.g. The mean of 1, 2, 2, 6, 5, 4, and 8 is

$$\frac{1+2+2+6+5+4+8}{7} = \frac{28}{7} = 4$$

Averages

The mode is the value that occurs most often in a data set.

Please note that a data set can have more than one mode or no mode at all.

e.g.

The mode of 1, 2, 2, 6, 5, 4 and 8 is 2 since it occurs twice and the other numbers each occur once.

The modes of 1, 2, 2, 6, 5, 4, 4 and 8 are 2 and 4.

The data set 1, 2, 6, 5, 4 and 8 has no mode.

Averages

The median is the midpoint of the data set, once it has been ordered.

The median can either be a specific value in the data set or fall between two values.

e.g. The median of 1, 2, 2, 6, 5, 4 and 8

is 4 since 1 2 2 (4) 5 6 8 (data set in order)

The median of 1, 2, 2, 6, 5, 5, 4 and 8

is 4.5 since 1 2 2 4 5 5 6 8 (data in order).

↑

median between 4 and 5.

Variation

The range is equal to the highest value of the data set minus the lowest value of the data set.

e.g. The range of 1, 2, 2, 6, 5, 4 and 8 is 7 since $8 - 1 = 7$.

Variation

We calculate the variance in the following way:-

Given a data set,

1) we calculate the mean of the data set.

2) For each number of the data set, we subtract the mean, and then we square the resulting number (to get rid of the negative sign and make everything positive).

3) We then take each of these numbers, which have been squared in the previous step, and calculate their mean.

4) The answer from the calculation of the mean in Step 3 is the value of the variance.

Variation

We recall that the symbol $\bar{X}$ represents the mean of the data set.

We often use the symbol σ^2 to represent the variance.

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$$

where $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum x}{n}$

Variation

Can regard

$$\sigma^2 = \frac{\sum Y}{n},$$

where $Y_1 = (X_1 - \bar{X})^2$, $Y_2 = (X_2 - \bar{X})^2$, ...

... $Y_n = (X_n - \bar{X})^2$.

So can regard variance σ^2 as mean of the data set $Y_1, Y_2, \ldots, Y_n$, where Y data set is constructed from the (X data set) $X_1, X_2, \ldots, X_n$.

Variation

example

Variance of 1, 2, 2, 6, 5, 4 and 8 can be calculated using the following table:-

(Recall that mean $\bar{X} = 4$

X	1	2	2	6	5	4	8
$X - \bar{X}$	-3	-2	-2	2	1	0	4
$(X - \bar{X})^2$	9	4	4	4	1	0	4

mean of last row of table :

$$\frac{9+4+4+4+1+0+4}{7} = \frac{26}{7}$$

$$\sigma^2 = 3\tfrac{5}{7}$$ variance.

The <u>standard deviation</u> is the square root of the variance of a data set.

We often represent the standard deviation by the symbol σ.

e.g. We have seen that $3\frac{5}{7}$ is the variance of 1, 2, 2, 6, 5, 4 and 8.

So the standard deviation $\sigma = 1.927$ (3 d.p.).

Variation

- The lower quartile is the quarter way point of the data set, once it has been ordered from lowest to highest

- The upper quartile is the three-quarter way point of the data set, once it has been ordered from lowest to highest.

- The interquartile range is the upper quartile of the data set minus the lower quartile of the data set.

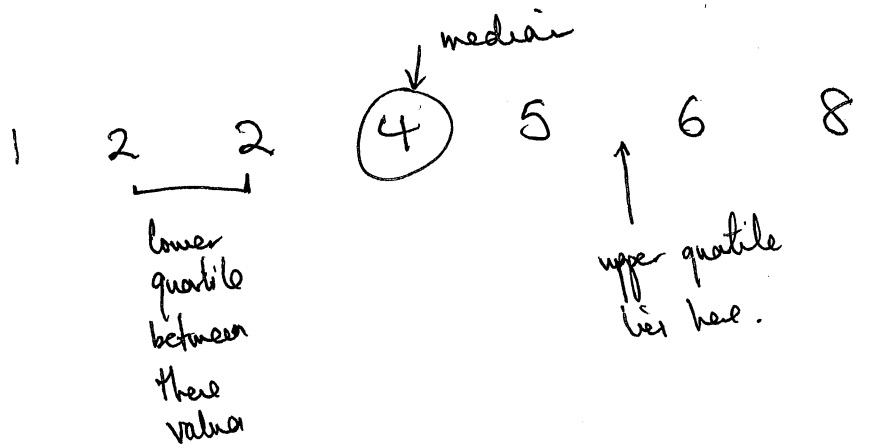
Note that the lower and upper quartiles can each be a specific value in the data set or fall between two values.

Variation

example

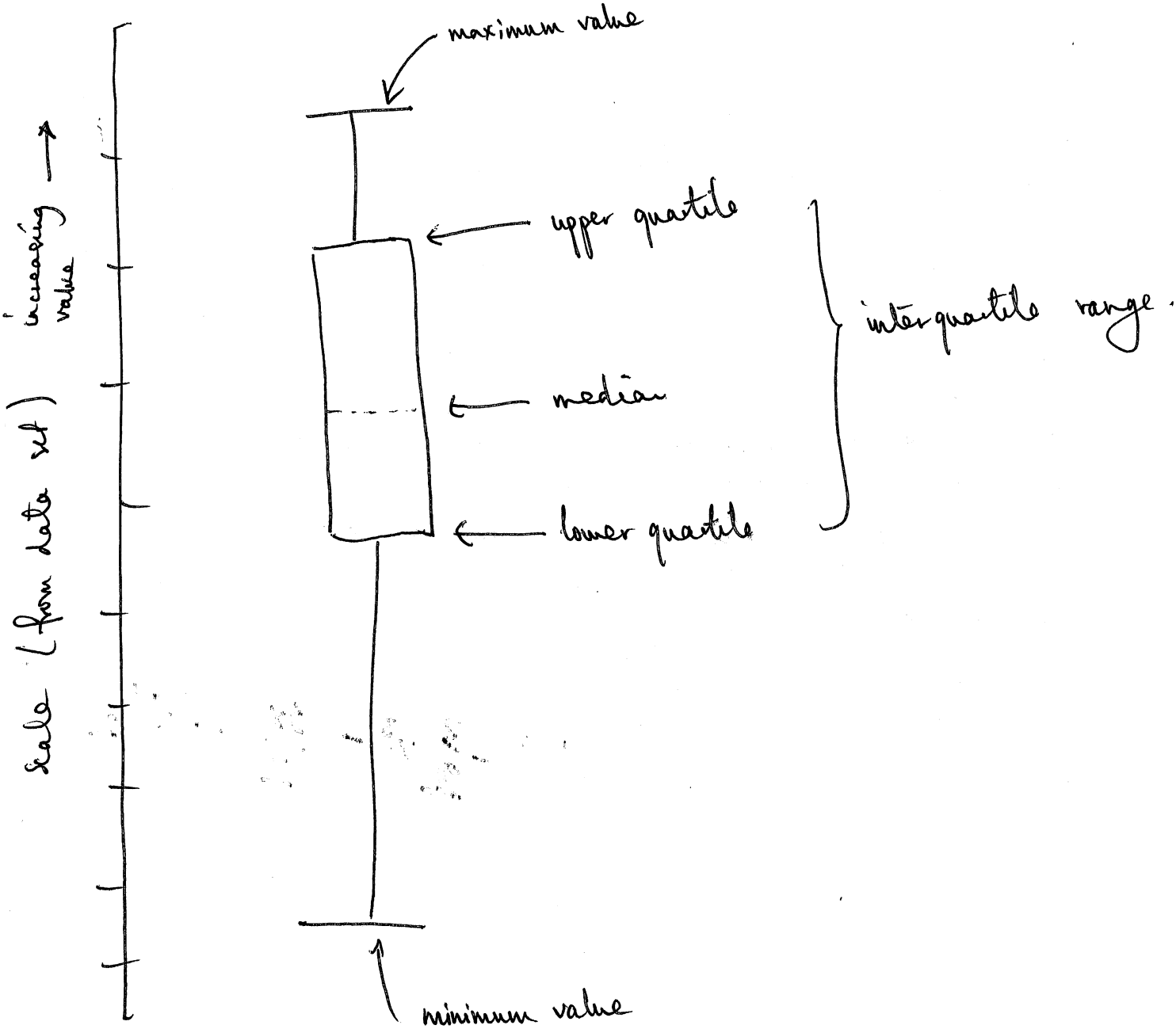
The lower ~~quartile~~ quartile and the upper quartile of the data set 1, 2, 2, 6, 5, 4 and 8 are

2 and 5.5 respectively. This is because

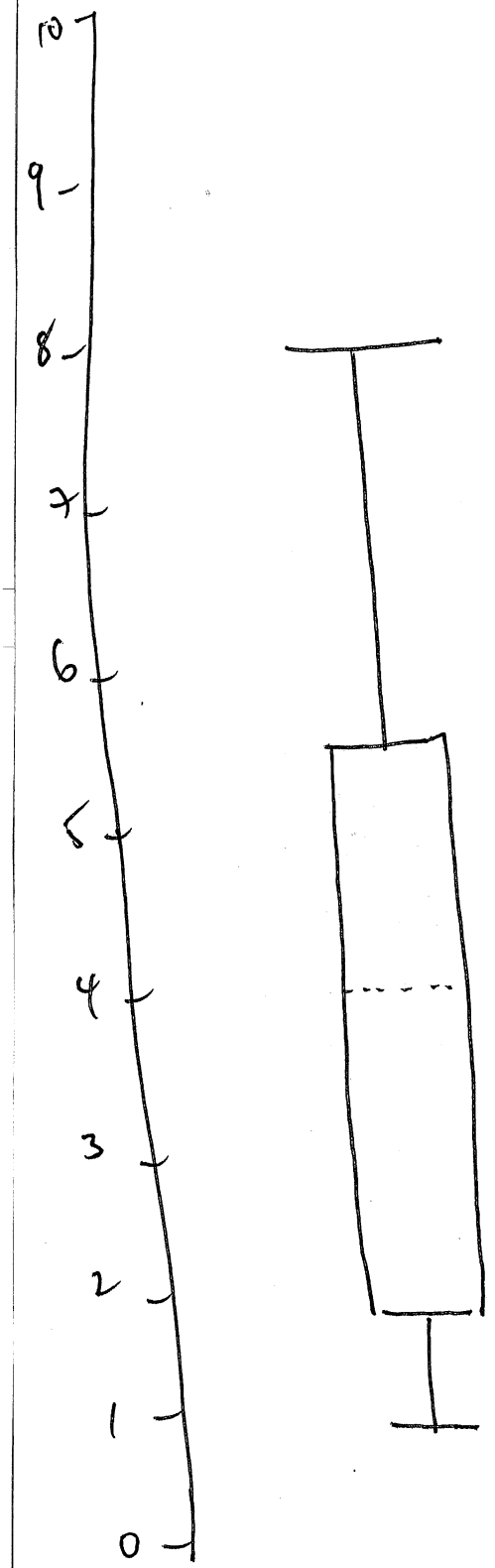


Hence the interquartile range is $5.5 - 2 = 3.5$.

Box-plots (Box & whisker diagrams).



Box-plot of 1, 2, 2, 6, 5, 4 and 8.



The averages, ie mean, mode and median, are used to highlight a value in the data set that best represents the whole of the data set.

The measures of the variation are used to describe how much the data is bunched, or clumped, around the relevant measure of the average.