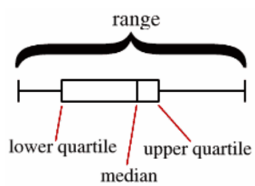
MDM4U: Mathematics of Data Management

Measures of Spread (Part 1)

Quartiles and Percentiles

MDM4U: Data Management



Measures of Spread

The *range* of a data set is the difference between the lowest and highest data.

This is of limited use, however, since only two values are being used to describe variation within the set.

A better option is to partition the data set into smaller intervals.

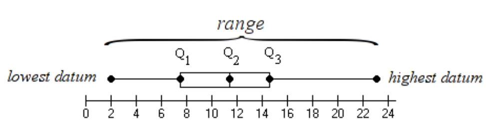
The two main methods of doing this is to use *quartiles* or *percentiles*.

Quartiles

Quartiles uses three points to divide a data set into four groups, each with an equal number of values.

These points are the first quartile ($Q1$), the second quartile ($Q2$) and the third quartile ($Q3$).

Since the second quartile divides the data set in half, the second quartile is the median.



Quartiles

Example

A salesperson records the shoe sizes of the last 10 sales.

6	7	9	9	9	10	10	12	12	18

Determine the median and the first and third quartiles.

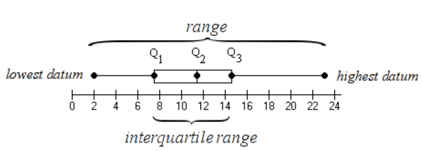
The median is the mean of the fifth and sixth values, or 9.5.

$Q1$ is the median of the lower half of the data, or the third value, 9.

$Q3$ is the median of the upper half of the data, or the eighth value, 12.

Interquartile and Semi-Interquartile Ranges

The *interquartile range* is the range of the central half of the data. Therefore, the interquartile range is $Q3 - Q1$.



A larger interquartile range reflects a larger spread of data within the central half of the data.

The *semi-interquartile range* is half of the interquartile range.

Both measures indicate how closely the data are centred around the median.

Interquartile and Semi-Interquartile Ranges

Example

Determine the range of the data, the interquartile range, and the semi-interquartile range for the earlier shoe example.

The lowest datum is 6, while the highest is 18. The range of the data is $18 - 6 = 12$.

Since $Q1 = 9$ and $Q3 = 12$, the interquartile range is $12 - 9 = 3$.

The semi-interquartile range is half of the interquartile range, or 1.5.

Box-and-Whisker Plots

Quartiles can be illustrated using *box-and-whisker plots*.

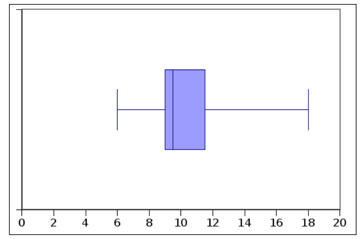
A box shows $Q1$, the median, and $Q3$. So the box shows the interquartile range.

Two whiskers extend to the lowest and the hightest data. This shows the range of the data set.

Box-and-Whisker Plots

Example

Illustrate the data from the shoe example using a box-and-whisker plot.



Box-and-Whisker Plots

A *modified box-and-whisker plot* is used when there are outliers in the data.

Outliers are not included in the whiskers, but are indicated as separate points.

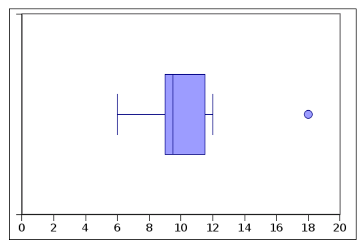
By convention, an outlier is any data point whose distance from the box is at least 1.5 times the width of the box.

Modified box-and-whisker plots typically illustrate the spread of the data more effectively.

Box-and-Whisker Plots

Example

Illustrate the data from the shoe example using a modified box-and-whisker plot.



Percentiles

Percentiles divide a data set into one hundred equally-sized intervals.

Therefore, the nth percentile, P_n, contains n% of the data.

It follows that $(100 - n)$% of the data are greater than or equal to P_n.

Percentiles

Example

The marks of 40 students who wrote a standardized test are below.

35	38	38	44	47	53	54	56	57	59
62	62	63	65	65	68	68	69	70	71
72	72	72	74	75	79	81	83	85	85
88	89	91	93	94	94	95	97	97	98

- Determine the 80th percentile of the data.
- What mark would a student have to score to be at the 60th percentile?
- What percentile corresponds to a test score of 77?

Percentiles

The 80th percentile is the boundary between the lower 80% of the scores and the top 20%.

80% of 40 is 32, so the 80th percentile is the mean of the 32nd and 33rd data, or 90.

To be at the 60th percentile, a student would have to score better than 60% of his/her classmates.

60% of 40 is 24, so the 60th percentile is the mean of the 24th and 25th data, or 74.5.

A test score of 77 lies between the 25th and 26th data. Since $\frac{25}{40} = 62.5\%$, the test score corresponds to the 63rd percentile.

Questions?