Outliers and cleaning data

Starter

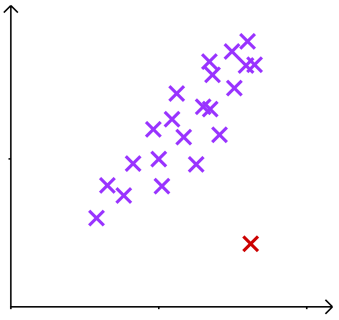
1. ***(Review of last lesson)*** On applying for a job at a small firm you are told that the average salary is about £3200. You take the job and imagine your horror when you are told your salary will be £2000. Believing you have been lied to you discover the salaries of all employees at the firm as:

£2000 £2000 £2000 £2000 £2100 £2100
£2400 £2600 £3400 £5900 £8500

Did the firm lie to you?

Notes

Outliers are very small or very large values in a data set that ***differ greatly from the majority*** of the other values.



Often outlying values are ***ignored*** since they can ***distort*** data — maybe there was an ***error in data collection***.

However, ***outliers can also inform***: for example, an athlete generally performs well but every second week of a month their times are outliers — a good coach would look into the reason for the dip in form.

While outliers can often be see visually from a scatter graph, there are two definitions of an outlier that should be followed:

1. More than $1.5 \times$ interquartile range from the nearer quartile.
2. More than $2 \times$ standard deviation away from the mean.

More than $1.5 \times$ ***interquartile range from the nearer quartile***

1. Find the quartiles, Q_1 and Q_3 and the interquartile range for the data.
2. A value is an outlier if it is $< Q_1 - 1.5 \times$ IQR.
3. A value is an outlier if it is $> Q_3 + 1.5 \times$ IQR.

E.g. 1 Using the definition "more than $1.5 \times$ interquartile range from the nearer quartile", determine if there are any outliers in this data set.

11, 6, 16, 5, 8, 5, 7, 8, 9, 13
9, 8, 15, 6, 5, 8, 5, 8, 8, 10

More than $2 \times$ ***standard deviation away from the mean***

1. Find the mean and standard deviation for the data.
2. A value is an outlier if it is $< \mu - 2\sigma$.
3. A value is an outlier if it is $> \mu + 2\sigma$.

E.g. 2 Jake recorded the times, to the nearest minute, that it took him to cook dinner. Using the definition "more than $2 \times$ standard deviation away from the mean", determine if there are any outliers in this data set.

48 56 71 63 68 49
45 61 65 79 54 32

Video: Outliers
Video: Outliers: using quartiles and interquartile range
Video: Outliers: using the mean and standard deviation

Solutions to Starter and E.g.s

Exercise
p350 16E Qu 1i, 2i, 3-10, (11-12)

Summary
Outliers are very small or very large values in a data set that ***differ greatly from the majority*** of the other values.
There are two definitions of an outlier that should be followed:

1. More than $1.5 \times$ interquartile range from the nearer quartile.
2. More than $2 \times$ standard deviation away from the mean.