

Outliers

Definitions

There are two definitions of outliers depending whether we're working with box plots or normal distributions.

Box Plots

The definition is to take $1\frac{1}{2}$ times the IQR and then add that value to the UQ and subtract that value from the LQ. Then examine and decide whether to include or exclude.

- 1) Is it an obvious transcription error or miscalculation? Remove and redraw the boxplot to show highest and lowest values.
- 2) Is it "genuine" but to include would distort figures? Remove but note.
- 3) Is it "genuine" and therefore valid to include? Retain.

Normal Distribution Curve

Examine any values that fall outside 3 standard deviations, which will enclose 99.73% of the distributed values. So we're looking for anything outside of say 3 in a thousand. Now that does occur - approximately three times in a 1000 normally distributed results - so don't just discount automatically. Go through the same process as for box plot outliers.

Don't forget if you remove it for reason 1) you need to RECALCULATE the standard deviation. You don't recalculate for reasons 2) or 3).

Are They Equivalent?

Use normal cumulative distribution function tables to look up z values.

Identify $\Phi(z) - \Phi(-z) = 0.500$ and read off $z = 0.675$

Calculate 0.675×2 (two tailed) $\times 1.5$ (box plot criteria) and add to UQ.

So we get the result ± 2.7 standard deviations which is of the same order and encompass 99.31% of the distributed values.

For exactly 3 s.d. we need to amend 1.5 to 1.7 (*not a lot of people know this*)

∞ rg

Outliers

Definitions

There are two definitions of outliers depending whether we're working with box plots or normal distributions.

Box Plots

The definition is to take $1\frac{1}{2}$ times the IQR and then add that value to the UQ and subtract that value from the LQ. Then examine and decide whether to include or exclude.

- 1) Is it an obvious transcription error or miscalculation? Remove and redraw the boxplot to show highest and lowest values.
- 2) Is it "genuine" but to include would distort figures? Remove but note.
- 3) Is it "genuine" and therefore valid to include? Retain.

Normal Distribution Curve

Examine any values that fall outside 3 standard deviations, which will enclose 99.73% of the distributed values. So we're looking for anything outside of say 3 in a thousand. Now that does occur - approximately three times in a 1000 normally distributed results - so don't just discount automatically. Go through the same process as for box plot outliers.

Don't forget if you remove it for reason 1) you need to RECALCULATE the standard deviation. You don't recalculate for reasons 2) or 3).

Are They Equivalent?

Use normal cumulative distribution function tables to look up z values.

Identify $\Phi(z) - \Phi(-z) = 0.500$ and read off $z = 0.675$

Calculate 0.675×2 (two tailed) $\times 1.5$ (box plot criteria) and add to UQ.

So we get the result ± 2.7 standard deviations which is of the same order and encompass 99.31% of the distributed values.

For exactly 3 s.d. we need to amend 1.5 to 1.7 (*not a lot of people know this*)

∞ rg