Outliers and cleaning data

S2.4 · Statistics · revision notes

Identifying outliers with a rule given in the question, deciding whether to remove them, cleaning data (errors, missing values, and ‘n/a’ and ‘tr’ in the large data set), and choosing or criticising a way to present data. Usually short parts linked to box plots or measures of spread.

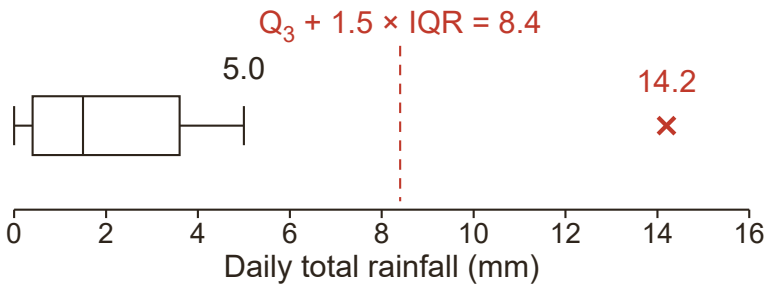
Key facts

- Quartile rule: outlier if below $Q_1 - k(Q_3 - Q_1)$ or above $Q_3 + k(Q_3 - Q_1)$
- Remove only errors; keep genuine extreme values
- Hardly affected by outliers: median, IQR. Strongly affected: mean, standard deviation, range, PMCC
- Mean rule: outlier if outside $\bar{x} \pm k\sigma$
- Removing x_0: $n - 1$, $\Sigma x - x_0$, $\Sigma x^2 - x_0^2$
- n/a = not available (leave out); tr = trace of rain, less than 0.05 mm

Notes

Identifying outliers

- An **outlier** is a value that is very different from the rest of the data.
- The rule to use is always given in the question. Common rules: an outlier is less than $Q_1 - k(Q_3 - Q_1)$ or greater than $Q_3 + k(Q_3 - Q_1)$, often with $k = 1.5$; or it is more than k standard deviations from the mean, e.g. outside $\bar{x} \pm 2\sigma$
- Only the smallest and largest values need checking against the limits.
- On a box plot, mark each outlier with a cross; the whisker ends at the most extreme value that is not an outlier.



Only 14.2 is beyond the limit, so it is a cross; the whisker ends at 5.0

Keep or remove?

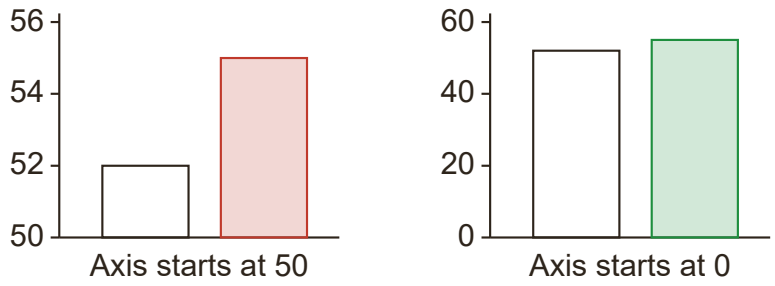
- Remove a value only if it is an **error**: impossible in context (e.g. negative rainfall), a recording or typing mistake, or not from the population being studied.
- A genuine extreme value (e.g. a very wet day during a storm) is real data and should be kept; removing it would misrepresent the population.
- Outliers strongly affect the mean, standard deviation, range and PMCC; the median and IQR hardly change.

Cleaning data

- **Cleaning** means dealing with errors, missing values and outliers before analysing the data.
- Leave out missing values (n/a in the large data set) and reduce n. A rainfall value of 'tr' (a trace, less than 0.05 mm) is often treated as 0 in calculations; say what you have done.
- To remove a value x_0 using summary statistics: new $n = n - 1$, new $\Sigma x = \Sigma x - x_0$, new $\Sigma x^2 = \Sigma x^2 - x_0^2$

Choosing and criticising diagrams

- Histogram: the shape of grouped continuous data. Box plots: comparing distributions and showing outliers. Cumulative frequency: medians, quartiles, percentiles and how many values lie above or below a value. Scatter diagram: the relationship between two variables.
- Criticise a diagram by spotting what misleads, e.g. unequal class widths plotted as frequency, or a vertical axis that does not start at zero and exaggerates differences.



Same data (52 and 55): a cut-off axis exaggerates the difference

How to answer each type of question

M = method mark, A = answer mark, B = a mark on its own.

Show that a value is an outlier (quartile rule)

3 to 4 marks | Grade D

1. Find Q_1 and Q_3 using the rule for an ordered list.
2. Calculate the limit using the rule given.
3. Compare the value with the limit and state the conclusion.
4. If asked whether to remove it, use the context: is the value possible?

Example. The daily total rainfall, in mm, for 15 days in a student's sample, in order, is
0, 0, 0.2, 0.4, 0.6, 1.0, 1.2, 1.5, 2.0, 2.4, 3.0, 3.6, 4.1, 5.0, 14.2
An outlier is a value greater than $Q_3 + 1.5(Q_3 - Q_1)$.
(a) Show that 14.2 is an outlier.
(b) State, with a reason, whether 14.2 should be removed from the data.

Model answer.

(a) $\frac{15}{4} = 3.75$, so Q_1 is the 4th value, 0.4; $\frac{45}{4} = 11.25$, so Q_3 is the 12th value, 3.6 (B1)

Limit $= 3.6 + 1.5 \times (3.6 - 0.4) = 8.4$ (M1); $14.2 > 8.4$, so 14.2 is an outlier (A1)

(b) No: 14.2 mm is a realistic amount of rain for a very wet day, so it is probably a genuine value (B1)

Choose a diagram and deal with missing data

2 to 3 marks | Grade B

1. Name a diagram suited to the purpose (comparing, showing shape, showing a relationship).
2. Give a reason linked to what the diagram shows.
3. For missing or invalid values, say what you would do and the effect on n

Example. A student wants to compare the daily mean temperatures at Leeming in 1987 and in 2015, and to show any outliers.
(a) Suggest a suitable diagram, giving a reason.
(b) Some days have ‘n/a’ recorded for the daily mean temperature. Explain how the student should deal with these.

Model answer.

(a) Two box plots on the same scale (B1): they show the medians, spreads and any outliers side by side, so the two years are easy to compare (B1)

(b) Leave these days out, as the data are missing, and reduce the sample size accordingly (B1)

Mean and standard deviation rule, then clean and recalculate 6 to 7 marks **Grade A**

1. Find $\bar{x}$, σ and the limits $\bar{x} \pm k\sigma$
2. Compare the value with the limit and conclude.
3. Remove the value from n, Σx and Σx^2
4. Recalculate the mean and standard deviation.

Example. The times, x minutes, taken by 12 students to solve a puzzle give $\Sigma x = 504$ and $\Sigma x^2 = 22\,032$. One of the times is 62 minutes. An outlier is a value more than 2 standard deviations from the mean.
(a) Show that 62 is an outlier.
(b) The time of 62 minutes was recorded wrongly and is removed. Find the mean and the standard deviation of the remaining times.

Model answer.

(a) $\bar{x} = \dfrac{504}{12} = 42$ and $\sigma = \sqrt{\dfrac{22\,032}{12} - 42^2} = \sqrt{72} = 8.49$ (M1)

Upper limit $= 42 + 2 \times 8.49 = 58.97$ (M1); $62 > 58.97$, so 62 is an outlier (A1)

(b) $n = 11$, $\Sigma x = 504 - 62 = 442$, $\Sigma x^2 = 22\,032 - 62^2 = 18\,188$ (M1)

Mean $= \dfrac{442}{11} = 40.2$ minutes (A1)

$\sigma = \sqrt{\dfrac{18\,188}{11} - \left(\dfrac{442}{11}\right)^2}$ (M1) $= 6.24$ minutes (A1) (6.54 using $n - 1$)

Don’t lose marks

- Using a different outlier rule from the one given in the question.
- Removing an outlier just because it is extreme, when it is a genuine value.
- Forgetting to reduce n after removing a value or leaving out missing data.
- Drawing the whisker out to an outlier instead of marking the outlier with a cross.
- Saying that removing an outlier changes the median and IQR a lot.

More tips

Memory tricks

- Error or extreme? Ask whether the value is possible here. Impossible means an error, so remove it; possible means keep it.
- Removing a value: subtract it from Σx, subtract its square from Σx^2, and reduce n by 1.
- Only test the most extreme values: if the largest value is not an outlier, no other value above the median is.

Exam technique

- In ‘show that’ outlier questions, calculate the limit, compare it with the value and write a conclusion.
- When deciding whether to remove a value, refer to the context: could this value really happen?
- Know how the large data set records missing values (n/a) and very small rainfall (tr), and say how you would handle them.

What each grade needs

What you need to be able to do, from the first marks up to the top grade.

Grade E — **State what an outlier is**
An outlier is a value that is very different from (much larger or smaller than) the rest of the data.

Grade D — **Find outlier limits using quartiles and the IQR**
Use the rule given, e.g. below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$

Grade C — **Find outlier limits using mean and standard deviation**
Use the rule given, e.g. outside $\bar{x} \pm 2\sigma$

Grade C — **Draw a box plot showing outliers**
Mark outliers with crosses and end the whisker at the most extreme value that is not an outlier.

Grade B — **Choose a suitable diagram for the data**
Match the diagram to the data and the purpose, e.g. box plots to compare distributions.

Grade B — **Decide whether an outlier should be removed**
Remove it only if it is an error; keep a genuine extreme value.

Grade A — **Clean data and recalculate summary statistics**
Take errors and missing values out of n, Σx and Σx^2, then recalculate.

Grade A* — **Evaluate how cleaning affects the conclusions**
Explain which statistics change a lot (mean, standard deviation, range, r) and which barely change (median, IQR).

Free notes, Full packs, practice papers and mock exams at chhetriacademy.com. Written for this site; not taken from past papers. Stuck? 1-to-1 tutoring with a free 30-minute trial lesson.