Histograms, box plots and cumulative frequency

S2.1 · Statistics · revision notes

Drawing and interpreting histograms, frequency polygons, cumulative frequency diagrams and box plots for single-variable data. Questions usually ask you to use the fact that area represents frequency in a histogram, to estimate values from a diagram, or to compare two distributions in context.

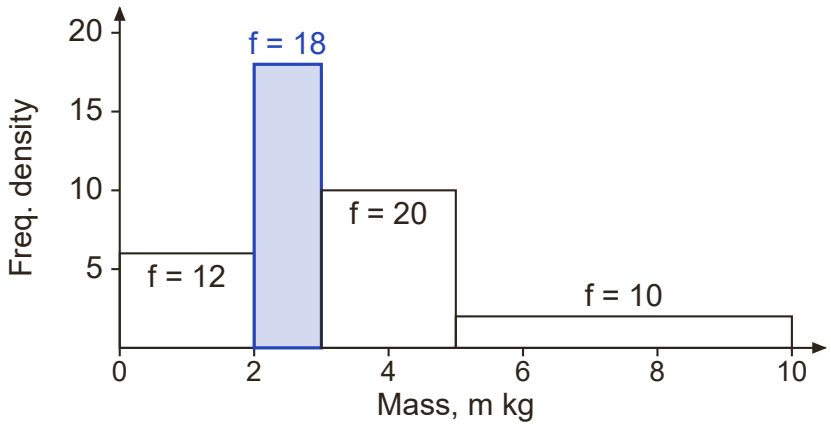
Key facts

- Frequency density = frequency ÷ class width
- Cumulative frequency: plot at the upper class boundaries
- IQR $= Q_3 - Q_1$ (the middle 50%)
- Box plot: minimum, Q_1, median, Q_3, maximum; outliers marked ×
- Histogram: area of a bar is proportional to frequency (frequency = k × area)
- Grouped data: median at $\frac{n}{2}$, Q_1 at $\frac{n}{4}$, Q_3 at $\frac{3n}{4}$
- 10th to 90th interpercentile range $= P_{90} - P_{10}$

Notes

Histograms

- Histograms show grouped continuous data. The bars touch, and the **area** of each bar is proportional to the frequency.
- Usually the vertical axis is **frequency density** = frequency ÷ class width, so that area = frequency.

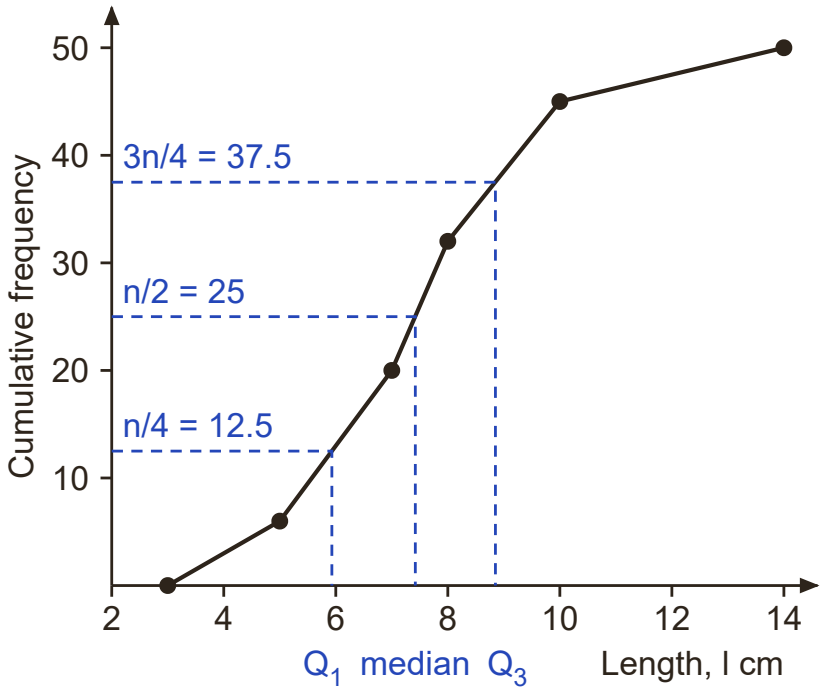


Bar area = frequency. The tallest bar (2 to 3 kg) is not the class with the greatest frequency

- Use class boundaries for widths: lengths recorded to the nearest cm in the class 10–14 have boundaries 9.5 and 14.5, so the width is 5.
- If you are given one bar's frequency, find the scale first: frequency per unit of area (per cm² or per grid square) = that bar's frequency ÷ its area.
- To estimate the frequency in part of a class, assume the values are spread evenly: 3 of the 5 units of width contain $\frac{3}{5}$ of the frequency.
- A **frequency polygon** joins the midpoints of the tops of the bars with straight lines.

Cumulative frequency

- Plot cumulative frequency against the **upper class boundary**, starting at the lowest boundary with cumulative frequency 0, and join the points with a smooth curve or straight lines.
- For n values, read across from $\frac{n}{2}$ for the median, $\frac{n}{4}$ for Q_1, $\frac{3n}{4}$ for Q_3 and $\frac{k}{100}n$ for the kth percentile.

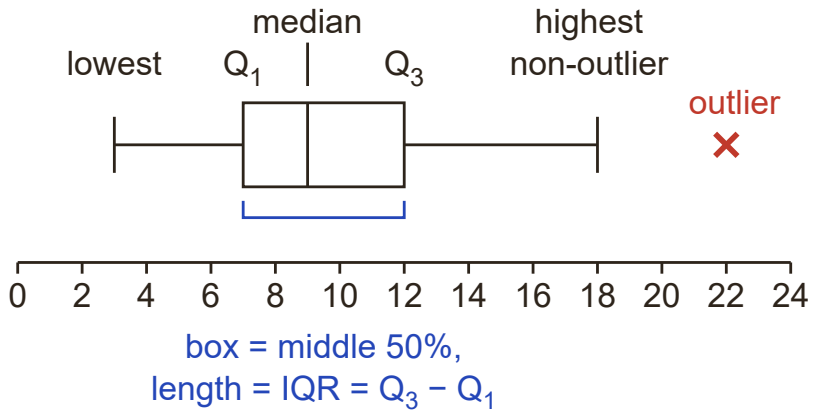


Plot at upper class boundaries, then read across from n/4, n/2 and 3n/4 and down to the axis

- To estimate how many values are above a given value, read up from it and subtract that cumulative frequency from n

Box plots

- A box plot shows the minimum, Q_1, median, Q_3 and maximum. Outliers are plotted as crosses, and the whisker goes to the most extreme value that is not an outlier.



Box plot: the whisker stops at the most extreme value that is not an outlier

- The box holds the middle 50% of the data; its length is the interquartile range, $Q_3 - Q_1$

Comparing distributions

- Compare a measure of location (median or mean) **and** a measure of spread (IQR, range or standard deviation), quoting values and interpreting them in context.
- Use the median and IQR when the data are skewed or contain outliers, as they are not affected by extreme values.

How to answer each type of question

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

Calculate frequency densities and use areas

4 to 6 marks **Grade D**

1. Find each class width from the class boundaries.
2. Frequency density = frequency ÷ class width.
3. For part of a class, use (fraction of the class width) × frequency.

Example. The table shows the masses, m kg, of 60 parcels.

Mass, m kg	Frequency
$0 \le m < 2$	12
$2 \le m < 3$	18
$3 \le m < 5$	20
$5 \le m < 10$	10

(a) Calculate the frequency density of each class, and state which class has the tallest bar on a histogram.
(b) A parcel is chosen at random. Estimate the probability that its mass is between 4 kg and 6 kg.

Model answer.

(a) Widths 2, 1, 2, 5, so frequency densities 6, 18, 10, 2 (M1 A1). The tallest bar is $2 \le m < 3$ (B1), even though $3 \le m < 5$ has the greatest frequency.

(b) 4 to 5 kg: $\frac{1}{2} \times 20 = 10$ (M1); 5 to 6 kg: $\frac{1}{5} \times 10 = 2$ (M1); probability $= \dfrac{12}{60} = 0.2$ (A1)

Compare two distributions from box plots

2 to 3 marks **Grade C**

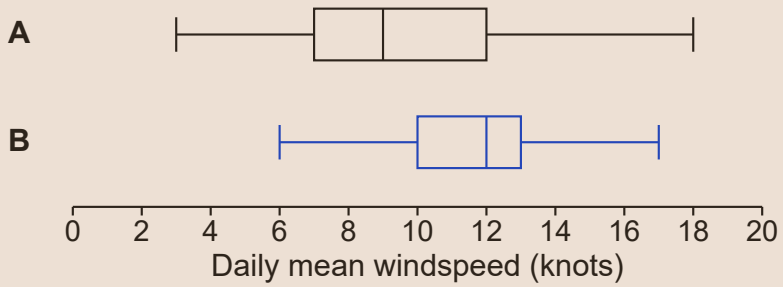
1. Compare the medians (location), with values, in context.
2. Compare the IQRs or ranges (spread), with values, in context.
3. Mention outliers or skew only if they are relevant.

Example. A student's samples of daily mean windspeed (knots) from two stations give these box plot values.

Station	Minimum	Q_1	Median	Q_3	Maximum
A	3	7	9	12	18
B	6	10	12	13	17

Compare the windspeeds at the two stations.

Model answer.



The median at B (12 knots) is higher than at A (9 knots), so on average it was windier at B (B1)

The IQR at A is $12 - 7 = 5$ knots and at B is $13 - 10 = 3$ knots, so the windspeeds at A were more variable (B1)

Use a histogram when you are given one bar's frequency

3 to 4 marks | **Grade B**

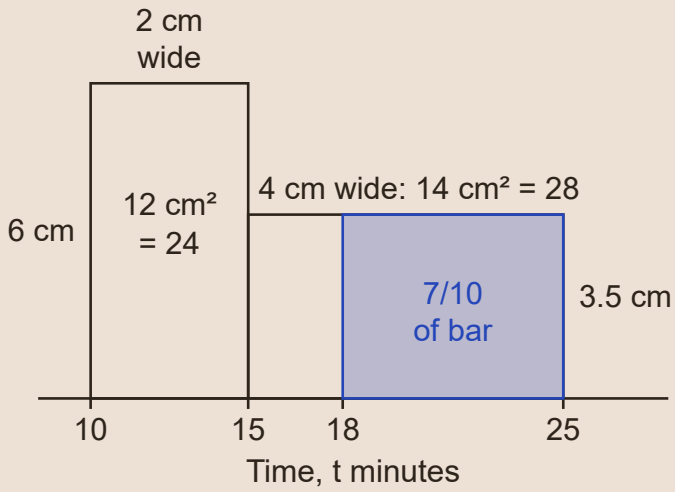
1. Find the area of the bar whose frequency you know.
2. Work out the frequency represented by one unit of area.
3. Multiply the area of each other bar (or part of a bar) by this.

Example. On a histogram of the times, t minutes, that people waited at a clinic, the bar for $10 \leq t < 15$ is 2 cm wide and 6 cm tall and represents 24 people. The bar for $15 \leq t < 25$ is 3.5 cm tall.

(a) Find the number of people who waited between 15 and 25 minutes.

(b) Estimate the number of people who waited between 18 and 25 minutes.

Model answer.



(a) 5 minutes is 2 cm wide, so the second bar is 4 cm wide. First bar: $2 \times 6 = 12$ cm² represents 24 people, so 1 cm² represents 2 people (M1). Second bar: $4 \times 3.5 = 14$ cm², so 28 people (A1)

(b) $\frac{7}{10} \times 28$ (M1) $= 19.6$, so about 20 people (A1)

Don't lose marks

- Plotting frequency instead of frequency density when the class widths are unequal.
- Plotting cumulative frequencies at class midpoints instead of upper class boundaries.
- Using the stated class limits instead of the class boundaries to find widths.
- Comparing only averages, or quoting numbers without saying what they mean in context.

More tips

Memory tricks

- The tallest bar is not always the class with the greatest frequency: area, not height, shows frequency.
- Scale questions: find the frequency per unit of area from one bar, then use it for every bar.
- The last cumulative frequency always equals the total n: a quick check of your table.
- For comparisons think ‘one for the middle, one for the spread’.

Exam technique

- ‘Estimate’ tells you to assume values are spread evenly within a class; round sensibly if the answer is a number of people or items.
- Check the vertical axis: if it is labelled frequency density, area = frequency; if not, find the scale from a bar you know.
- When you draw a box plot, use the scale given, mark outliers with crosses and label which plot is which.

What each grade needs

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

Grade E — **Calculate frequency densities for a histogram**
Frequency density = frequency ÷ class width, using the class boundaries.

Grade E — **Read the median and quartiles from a box plot**
The box shows Q_1, the median and Q_3; the whiskers show the extent of the data and crosses show outliers.

Grade D — **Estimate quartiles from a cumulative frequency graph**
Read across from $\frac{n}{4}$, $\frac{n}{2}$ and $\frac{3n}{4}$ on the cumulative frequency axis.

Grade D — **Draw a histogram with unequal class widths**
Plot frequency density on the vertical axis, with the bars touching at the class boundaries.

Grade C — **Estimate frequencies from histogram bar areas**
Use area to find frequency, assuming values are spread evenly within a class for part of a bar.

Grade B — **Find a histogram's scale from a known bar**
When area is not equal to frequency, use one bar to find the frequency per unit of area.

Grade A — **Compare distributions using location and spread in context**
Compare a median (or mean) and an IQR (or standard deviation), with values and in the context of the data.

Grade A* — **Link a histogram's shape to a probability model**
Recognise that a roughly symmetrical, bell-shaped histogram suggests a normal model, and treat areas as probabilities.

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.