Statistics

Two-way and tally tables, bar charts, pie charts, pictograms, scatter diagrams, histograms and cumulative frequency, then the mean, median, mode, range and quartiles, with grouped data and exam-style questions.

Types of tables, charts and diagrams

Two-way table

	Glasses	No glasses	Total
Boys	6	10	16
Girls	8	6	14
Total	14	16	30

The total at the end of each row and column is the total for that individual row or column.

Tally table

Colour	Tally	Frequency
Red	𝍸 ||||	9
Blue	|||	3
Green	𝍸 𝍸 ||	12

Each mark (|) represents one of that thing. One complete group (shown here with a line through it) represents 5.

Frequency is the number of times a particular value or event occurs within a set of data.

Bar charts

In a bar chart, the bars are separated.

The bar chart below is called a **stacked bar chart**. This is used to show how a total is split.

Note Academy

The last type of bar graph is the **dual (side-by-side) bar graph**. It is used when comparing two different groups side by side.

Pie charts

Pie charts show how a whole is divided into parts. Each slice represents a category's share out of the total.

$$\text{Angle of slice} = \frac{\text{category frequency}}{\text{total frequency}} \times 360^\circ$$

TRY IT

Laila asks a group of people what type of exercise they prefer. The results are shown in the pie chart. (i) 90 people prefer running. Calculate the total number of people Laila asks. (ii) Work out the fraction of the group who prefer yoga. Write your answer in its simplest form.

(i) Running is 135° out of 360°, and that is 90 people. Let the total be n:

$$\frac{135}{360} \times n = 90 \quad \Rightarrow \quad n = 90 \times \frac{360}{135} = 240$$

(ii) Angles around a point add up to 360°. The cycling slice has a right angle (90°). So the yoga angle is:

$$360 - (135 + 90 + 60) = 360 - 285 = 75^\circ$$

$$\text{Fraction} = \frac{75}{360} = \frac{5}{24}$$

Pictograms

Pictograms use pictures or icons to represent data. There will always be a **key** that tells what each symbol represents.

Tank	Number of fish
A	3 symbols (6 fish)
B	2 symbols (4 fish)
C	4 symbols (8 fish)

Key: 1 symbol $=$ 2 fish.

Note Academy

Scatter diagrams

A scatter diagram is a graph that shows the relationship between two variables by plotting points. Each point represents one person or thing and has two values.

Draw a **line of best fit**: a straight line that goes through the middle of the points, to check the trend of the points.

In this example, the line of best fit shows a **negative correlation**, which means that an increase in one thing means a decrease in the other.

Histograms

A histogram shows grouped data. The bars touch in this graph because the data is **continuous**. Continuous data is divided into intervals and then used in the histogram to make better sense of the data.

$$\text{Frequency density} = \frac{\text{frequency}}{\text{class width}}$$

NOTE

Memorise the formula above. This formula is used to calculate frequency density, which is on the y-axis of the histogram. Remember that **the frequency is the area of the bar**, not the value on the y-axis.

Speed (s km/h)	$20 < s \leq 30$	$30 < s \leq 40$	$40 < s \leq 45$	$45 < s \leq 50$	$50 < s \leq 70$
Frequency	24	46	44	38	28
Class width	10	10	5	5	20
Frequency density	2.4	4.6	8.8	7.6	1.4

TRY IT

Laila records the number of steps she walks each day for 60 days. (i) Complete the histogram to represent the data. (ii) Her target is 11 000 steps every day. Find an estimate for the percentage of these 60 days that she met her target.

Note Academy

Steps (n thousand)	$5 < n \leq 8$	$8 < n \leq 10$	$10 < n \leq 12$	$12 < n \leq 15$	$15 < n \leq 20$
Frequency	9	16	14	15	6
Class width	3	2	2	3	5
Frequency density	3	8	7	5	1.2

(i) Use the formula for frequency density. For example, $8 < n \leq 10$ has $\frac{16}{2} = 8$. The bar heights are 3, 8, 7, 5 and 1.2.

(ii) We only want values from 11 to 20.

- In $10 < n \leq 12$, only part of the interval (from 11 to 12) counts. Half of the frequency in this group belongs to $11 < n \leq 12$, so we take half of $14 = 7$ days.
- From 12 to 15: all values count, 15 days.
- From 15 to 20: all values count, 6 days.

$\frac{7+15+6}{60} \times 100 = 46.7\%$

Cumulative frequency diagrams

A cumulative frequency diagram shows how many data values are less than or equal to a certain number. It helps us see totals building up.

For example, this graph shows us that most of the people's heights lie between 1.3 and 1.4 m (the curve is steepest there).

This graph is useful in determining:

- **Median:** halfway up the y-axis.
- **Lower quartile (Q_1):** $\frac{1}{4}$ of the total frequency.
- **Upper quartile (Q_3):** $\frac{3}{4}$ of the total frequency.
- **Inter-quartile range** $= Q_3 - Q_1$.

> **TRY IT**
>
> A shop sells two varieties of apple tree. The cumulative frequency diagram shows the heights of 80 Variety A trees. Use the diagram to estimate (a) the median, (b) the 30th percentile, and (c) the value of y if trees taller than y m are graded Class I and $\frac{2}{5}$ of the 80 trees are Class I. (d) Complete the frequency table.

(a) Median. $\frac{80}{2} = 40$th value. Read across from 40 on the vertical axis to the curve, then down: **1.33 m**.

(b) 30th percentile. $\frac{30}{100} \times 80 = 24$th value. Read across from 24 to the curve, then down: **1.28 m**.

Note Academy

(c) Value of y. $\frac{2}{5} \times 80 = 32$ trees are in Class I, so the other $80 - 32 = 48$ trees are not. Read across from 48 to the curve, then down: $y =$ **1.36 m**.

(d) Frequency table. The frequency is the difference between consecutive cumulative frequencies, read from the curve at $h = 1.1, 1.2, 1.3, 1.4, 1.5, 1.6$:

Height (h m)	$1.1 < h \leq 1.2$	$1.2 < h \leq 1.3$	$1.3 < h \leq 1.4$	$1.4 < h \leq 1.5$	$1.5 < h \leq 1.6$
Cumulative frequency	6	30	58	76	80
Frequency	6	24	28	18	4

(For example, $58 - 30 = 28$ and $80 - 76 = 4$.)

TRY IT

The frequency table shows the heights of 50 Variety B trees. Using the midpoints of the intervals, the estimated mean height is 1.81 m. Calculate the value of p and the value of q.

Height (h m)	$1.5 < h \leq 1.7$	$1.7 < h \leq 1.8$	$1.8 < h \leq 1.9$	$1.9 < h \leq 2.3$
Frequency	p	15	17	q
Midpoint	1.6	1.75	1.85	2.1

Because there are 2 unknown variables, we need to make **simultaneous equations**.

The frequencies add up to 50: $p + 15 + 17 + q = 50$, so

$$p + q = 18 \qquad (1)$$

The mean is the sum of $\text{frequency} \times \text{midpoint}$ divided by the total:

$$1.81 = \frac{1.6p + 1.75(15) + 1.85(17) + 2.1q}{50}$$

$$1.81 \times 50 = 1.6p + 26.25 + 31.45 + 2.1q \quad \Rightarrow \quad 1.6p + 2.1q = 32.8 \qquad (2)$$

Solving: from (1), $p = 18 - q$. Substituting into (2): $1.6(18 - q) + 2.1q = 32.8$, so $28.8 + 0.5q = 32.8$, so $q = 8$ and $p = 10$.

Frequency polygons

A frequency polygon is a line graph made by joining the midpoints of the top of each bar from a bar chart or histogram. It is used to show trends or patterns in data and is great for comparing the shapes of data.

Note Academy

Averages and spread

Mean

The mean is one of the ways to get the average of a dataset. You add all the numbers and divide by how many there are.

$$\text{Mean} = \frac{\text{sum of all values}}{\text{number of values}}$$

TRY IT

The heights of 5 kids are 140, 145, 150, 150 and 155 cm. Find the average height.

$$\text{Mean} = \frac{140+145+150+150+155}{5} = \frac{740}{5} = 148 \text{ cm}$$

NOTE

The mean is affected by very large or very small values.

Median

The median is the middle number when the data is in order. Use the formula below to get the **position** of the median term.

$$\text{Median position} = \frac{1}{2}(n+1)$$**, where n is the number of values.**

If the answer has $.5$ in it, take the average of the two terms either side. For example, if the position is 4.5, take the average of the 4th and 5th terms.

TRY IT

The heights of 5 kids are 140, 145, 150, 150 and 155 cm. Find the median.

The position is $\frac{1}{2}(5+1) = 3$, so the median is the 3rd value: **150 cm**. (Cut one term from each side repeatedly.)

NOTE

Note Academy

If there are two middle values, take the average of them. The median is great when there are **outliers** (values that are much smaller or larger than the others).

Mode

The value that comes up the most.

> **TRY IT**
>
> The heights of 5 kids are 140, 145, 150, 150 and 155 cm. Find the mode.

150 appears twice, so the mode is **150 cm**.

> **NOTE**
>
> There can be more than one mode, or no mode at all!

Range

It is the difference between the biggest and smallest value. It shows how spread out the data is.

> **TRY IT**
>
> The heights of 5 kids are 140, 145, 150, 150 and 155 cm. Find the range.

$$\text{Range} = 155 - 140 = 15 \text{ cm}$$

Mean of grouped data

When values are in groups, we don't know the exact numbers. So we use the **midpoint** of each group.

$\text{Mean} = \frac{\sum fx}{\sum f}$**, where** f **is the frequency,** x **is the midpoint of the class,** $fx = f \times x$ **and** $\sum$ **means "sum of".**

> **TRY IT**
>
> Calculate the mean mark.

Note Academy

Marks	Frequency (f)	Midpoint (x)	fx
0–10	4	5	20
10–20	6	15	90
20–30	10	25	250
30–40	5	35	175

$$\sum fx = 535, \quad \sum f = 25 \quad \Rightarrow \quad \text{Mean} = \frac{535}{25} = 21.4$$

Modal class

The class (range) with the highest frequency. From the table above, the highest frequency is 10, so the modal class is **20–30**.

Cumulative frequency table

We add up frequencies step by step to get a cumulative frequency table.

Marks	Frequency	Cumulative frequency
0–10	4	4
10–20	6	10
20–30	10	20
30–40	5	25

Quartiles and IQR

- Q_1 (lower quartile): the 25% point, term $= \frac{1}{4}(n+1)$.
- Q_2 (median): the 50% point.
- Q_3 (upper quartile): the 75% point, term $= \frac{3}{4}(n+1)$.
- **Inter-quartile range** $= Q_3 - Q_1$. It shows the spread of the middle 50% of the data.

Note Academy