Statistics

IGCSE Mathematics

This handout covers Topic 9, Statistics. Parts marked **(Extended)** are only tested on the Extended papers; everything else is for both levels.

Collecting and showing data

Data is collected from people —a sample drawn from a population.

Image: Yann Kemper, CC0 (commons.wikimedia.org)

To organise **statistical** 统计 **data** 数据, use:

- a **tally table** 计数表—make a mark for each value, then count.
- a **two-way table** 双向表—sorts data by two features at once (for example, boys/girls against walk/bus).

When you read a diagram, only draw conclusions the data really supports.

Averages and range

Three averages describe a typical value of the data, and the range shows how spread out it is. Each is used for a different purpose:

- **mean** 平均数 $= \dfrac{\text{sum of all values}}{\text{how many values}}$.
- **median** 中位数 = the middle value when the data is put in order.
- **mode** 众数 = the value that appears most often.
- **range** 极差 = largest value − smallest value (it shows how spread out the data is).

Worked example. Find the mean, median, mode and range of $4, 7, 7, 2, 5$.

Order the data: $2, 4, 5, 7, 7$.

$$\text{mean} = \frac{4+7+7+2+5}{5} = \frac{25}{5} = 5, \quad \text{median} = 5, \quad \text{mode} = 7, \quad \text{range} = 7-2 = 5.$$

Averages from a frequency table

When data is listed with its **frequency** 频数 (how many times each value occurs), the mean is

$$\text{mean} = \frac{\sum(\text{value} \times \text{frequency})}{\sum \text{frequency}}.$$

Worked example. Values $1, 2, 3$ occur with frequencies $4, 5, 1$. Find the mean.

$$\text{mean} = \frac{1(4)+2(5)+3(1)}{4+5+1} = \frac{17}{10} = 1.7.$$

Grouped data (Extended)

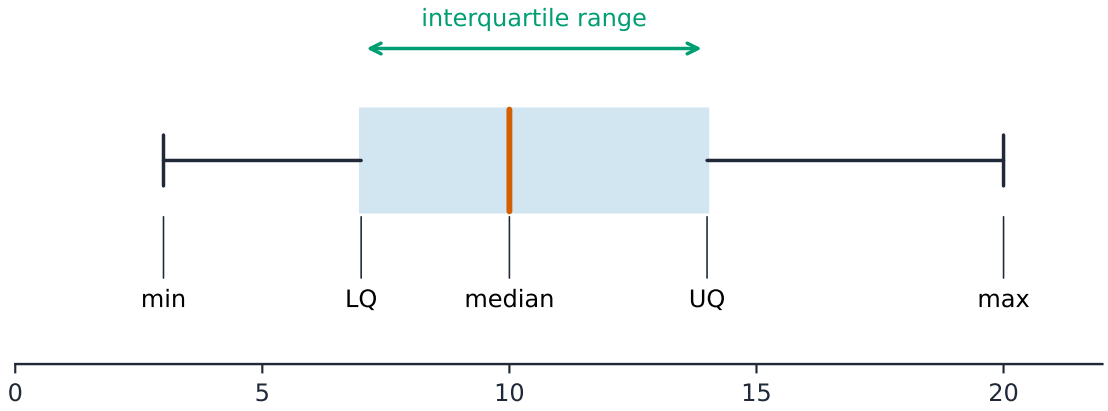
For **grouped data** 分组数据, you cannot find the exact mean, so estimate it using the **midpoint** of each group as the value. The **modal class** 众数组 is simply the group with the highest frequency.

Measures of spread (Extended)

When data is in order, the **quartiles** 四分位数 cut it into four equal parts. The lower quartile (LQ) is one quarter of the way up; the upper quartile (UQ) is three quarters of the way up. The **interquartile range** 四分位距 measures the spread of the middle half:

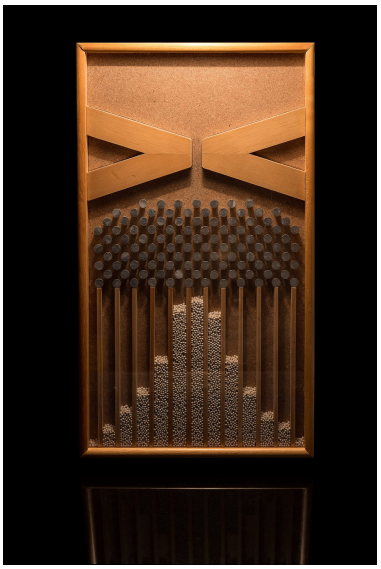
$$\text{interquartile range} = \text{UQ} - \text{LQ}.$$

The interquartile range is useful because, unlike the range, it ignores extreme values.



A box-and-whisker plot shows the five-number summary; the box length is the interquartile range.

Charts and diagrams

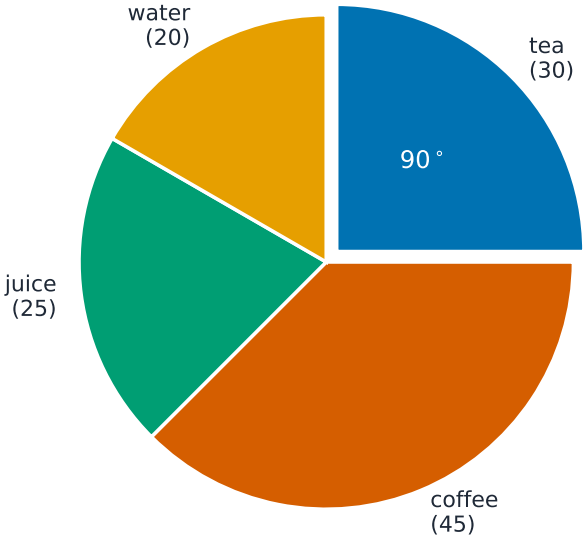


A Galton board shows how data piles up into a distribution.

Chart	What it shows
bar chart 条形图	a bar for each category; height is the frequency
pie chart 饼图	a circle split into slices, each a fraction of 360°
pictogram 象形图	uses a symbol to stand for a number of items
stem-and-leaf diagram 茎叶图	keeps the digits of ordered data, with a key
frequency distribution 频数分布	a table of values and their frequencies

Worked example (pie chart). Out of 120 people, 30 chose tea. Find the angle of the tea slice.

$$\frac{30}{120} \times 360^\circ = 90^\circ.$$

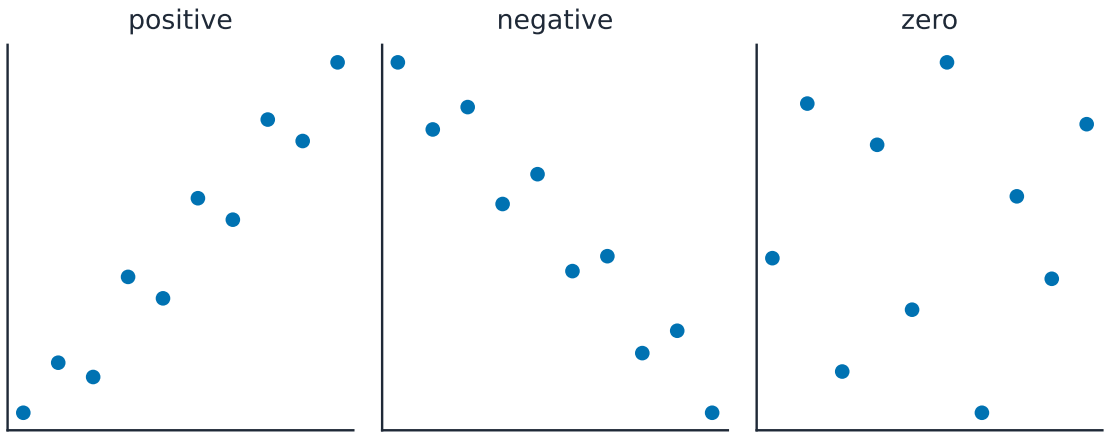


Each slice is its fraction of 360°*; tea is* $\frac{30}{120} \times 360° = 90°$.

Scatter diagrams and correlation

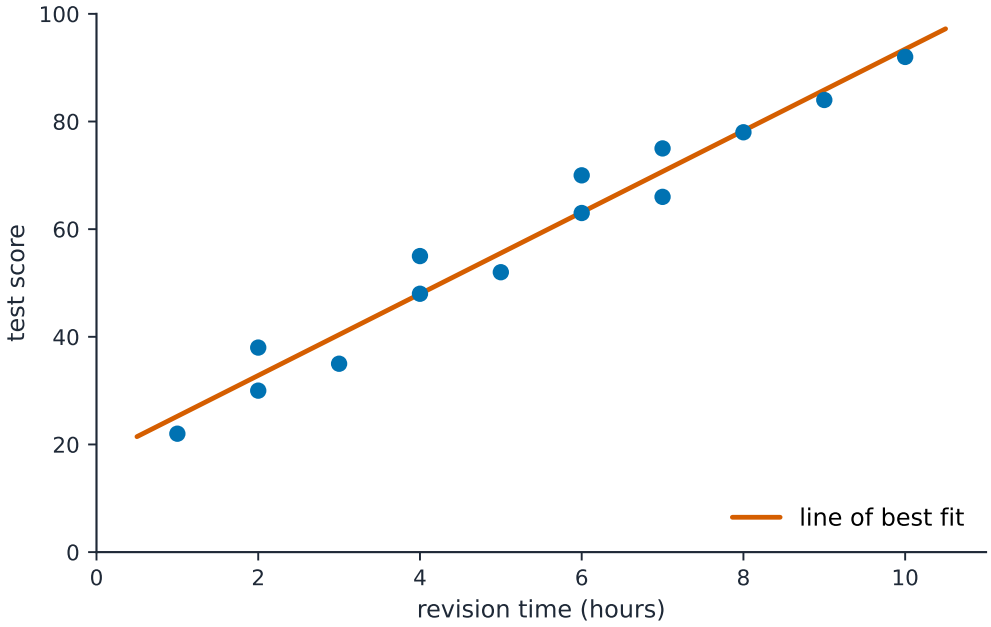
A **scatter diagram** 散点图 plots pairs of values as points to show whether two things are linked. The link is the **correlation** 相关性:

- **positive correlation** 正相关—as one goes up, the other goes up.
- **negative correlation** 负相关—as one goes up, the other goes down.
- **zero correlation** 零相关—no clear link.



Positive correlation rises together; negative falls as the other rises; zero shows no clear link.

If there is correlation, draw a **line of best fit** 最佳拟合线: one straight ruled line through the middle of the points, with about the same number of points on each side. Use it to predict values.

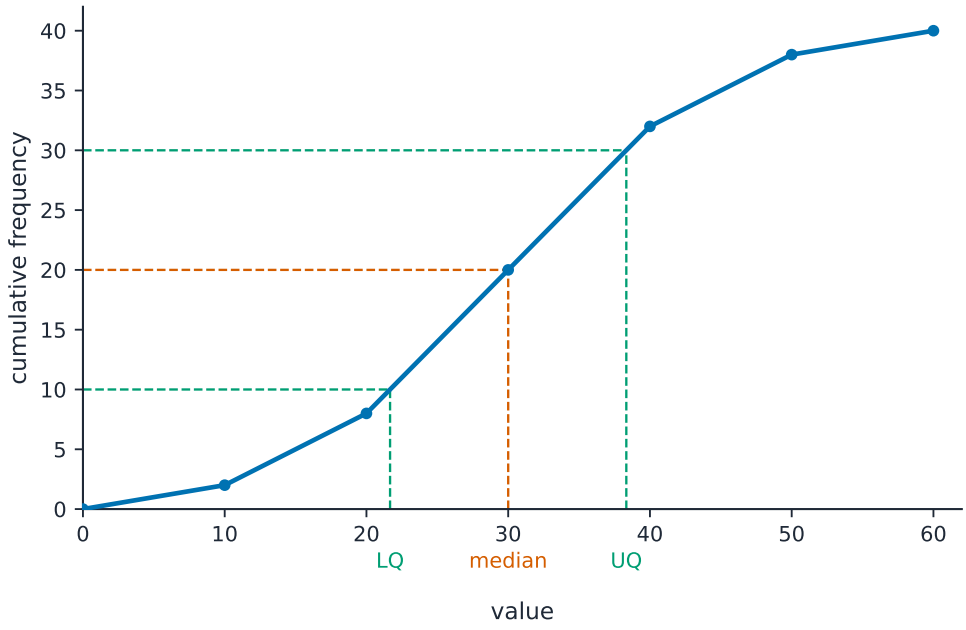


A line of best fit is one straight ruled line through the middle of the points; use it to predict.

Cumulative frequency diagrams (Extended)

The **cumulative frequency** 累积频数 is a running total of the frequencies. Plot it against the upper end of each class and join the points with a smooth curve.

From the curve you can read the median (at half the total), the quartiles (at one quarter and three quarters), and any **percentile** 百分位数 (for example, the 90th percentile is at 90% of the total).

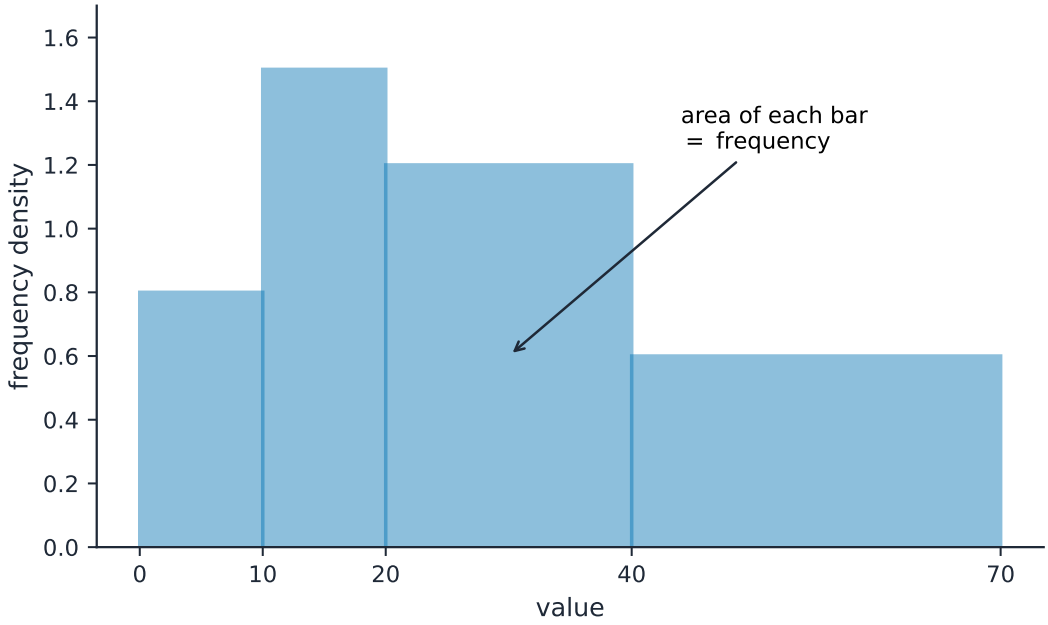


Read the median at half the total, and the quartiles at one quarter and three quarters: across to the curve, then down.

Histograms (Extended)

A **histogram** 直方图 looks like a bar chart, but the bars can have different widths and the area of each bar (not its height) shows the frequency. The vertical axis is the **frequency density** 频数密度:

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



With unequal class widths the bar AREA (not its height) is the frequency, so the axis is frequency density.

Worked example. A class has **class width** 组距 10 and frequency 25. Find the frequency density.

$$\text{frequency density} = \frac{25}{10} = 2.5.$$