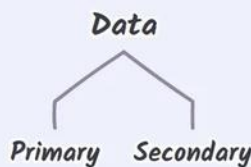


collecting data into tables, the three averages for ungrouped data, and the pictures

Data, and the words for handling it *

- primary data is collected by the investigator directly, and secondary data is taken from a source someone else compiled
- raw data is the list exactly as gathered, in no particular order
- arranging it in order of size gives an array, which makes the smallest and largest visible at once
- the frequency of a value is how many times it occurs
- the range is the largest value minus the smallest, and it is the simplest measure of spread
- a tally mark table is the usual way of counting frequencies without losing count



the two sources data can come from, branching from one heading: primary is gathered by the investigator directly, secondary is taken from a record someone else compiled, and a question naming a survey or a census is pointing at one of the two

Grouping data into classes

- when there are many values they are collected into class intervals such as 10 to 20
- the class mark is the midpoint, found as $\frac{\text{lower} + \text{upper}}{2}$
- the class size is the upper limit minus the lower limit, and it is usually kept the same throughout
- exclusive classes are written 10 to 20 and 20 to 30, where a value of exactly 20 goes into the second class
- inclusive classes are written 10 to 19 and 20 to 29, with no overlap but a gap between them
- an inclusive table must be made continuous before most calculations, by subtracting 0.5 from each lower limit and adding 0.5 to each upper one

The three averages for ungrouped data

- mean of n values $\bar{x} = \frac{\sum x_i}{n}$
- mean when values have frequencies $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$
- median position, n odd the $\frac{n+1}{2}$ th value
- median, n even, the mean of two middles $\frac{1}{2} \left(\frac{n}{2} \text{th} + \left(\frac{n}{2} + 1 \right) \text{th} \right)$
- the range as a measure of spread $R = x_{\max} - x_{\min}$

the median position formula gives a position, not the answer itself. For nine values the median is the fifth value, and answering with 5 instead of that value is a standard slip

The three averages compared

- the mean adds every value and divides by how many there are, so every value affects it
- the median is the middle value once the data is arranged in order, so it ignores how extreme the ends are
- the mode is the value occurring most often, and a set may have more than one mode or none at all
- a single very large value drags the mean but leaves the median almost untouched
- so for incomes, house prices and similar skewed data the median is the fairer summary
- the mode is the only average usable for non numerical data, such as the most common blood group in a class



a set of values marked along a line with one far out to the right: the median stays among the crowd while the mean is pulled towards the outlier, which is why the two can differ sharply on skewed data