



three averages for grouped data, and the pictures that show a distribution at a glance

Grouped data, and the words it brings

- when there are many observations they are collected into **class intervals**, such as 10 to 20 and 20 to 30
- the **frequency** f of a class is how many observations fall in it
- the **class mark** x is the midpoint of a class, found as $\frac{\text{lower} + \text{upper}}{2}$, and it stands in for every value in that class
- the **class size** h is the width of a class, upper limit minus lower limit, and it is usually the same throughout
- cumulative frequency** is a running total of the frequencies, either from the top down or from the bottom up
- grouping loses the individual values, so every average below is an **estimate**, not an exact figure

The three averages *

mean, direct method $\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$

mean, assumed mean method $\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i}$

mean, step deviation method $\bar{x} = A + \frac{\sum f_i u_i}{\sum f_i} \times h$

median of grouped data $M = L + \frac{\frac{n}{2} - cf}{f} \times h$

mode of grouped data $Z = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

the empirical relation between them $Z = 3M - 2\bar{x}$

the three mean methods give exactly the same answer. They differ only in how heavy the arithmetic is, so the step deviation method is chosen for large class marks

Which average, and when *

- the **mean** uses every observation, so it is the most complete, but a single extreme value pulls it badly
- the **median** is the middle value once the data is ordered, so extreme values do not move it. Prefer it for incomes and similar skewed data
- the **mode** is the most frequent value, and it is the only average that makes sense for **non numerical** data such as shoe sizes or favourite colours
- the **modal class** is the class with the highest frequency, and the mode lies inside it
- the **median class** is the class in which the cumulative frequency first reaches or passes $n/2$



the three averages marked along one scale: for a distribution stretched out to the right the mode sits lowest, the median next and the mean furthest along, because only the mean is dragged by the extreme values at the far end

the least arithmetic of the three

The step deviation method in order

write the class marks x_i for every class, each the midpoint of its interval

choose an **assumed mean** A , usually the class mark near the middle of the table, and the class size h

compute $u_i = \frac{x_i - A}{h}$ for each row, which gives small

whole numbers such as -2, -1, 0, 1, 2

multiply each u_i by its frequency and total the column to get $\sum f_i u_i$

total the frequency column to get $n = \sum f_i$

substitute into $\bar{x} = A + \frac{\sum f_i u_i}{n} \times h$, remembering that h multiplies the whole fraction

