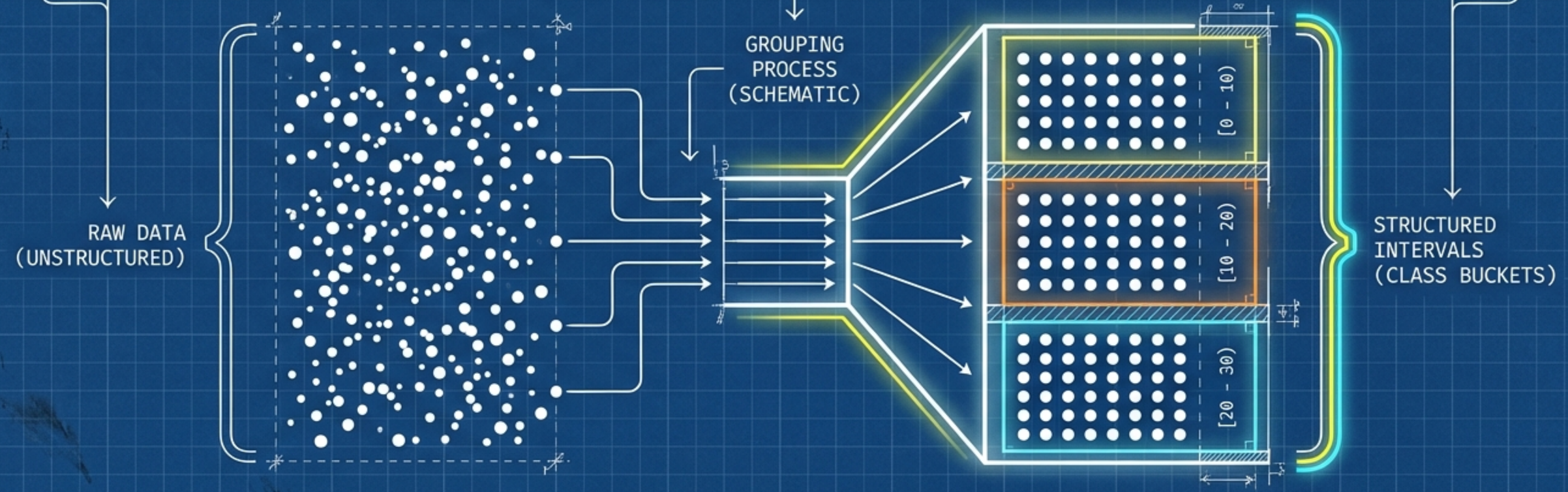
$$\bar{x} = \frac{\sum(f_i * x_i)}{\sum f_i}$$

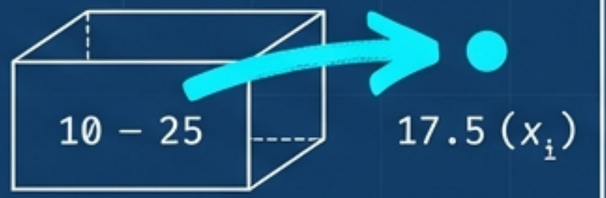
Architecting Data: The Grouped Statistics Toolkit

Finding the center in a world of scattered numbers.



Three Tools, Three Different Centers

THE PROXY



Grouped data hides individual points. We use the midpoint (x_i) as a proxy for the entire class.

MEAN ($\bar{x}$)

What it finds: The mathematical center of gravity.

Best Used When: You need to account for every single observation.

Vulnerability: Pulled heavily by extreme outliers.

MODE

What it finds: The most popular occurrence.

Best Used When: You want to know what happens most often.

Vulnerability: Ignores the rest of the data spread.

MEDIAN

What it finds: The exact middle value when sorted.

Best Used When: Data is skewed or has extreme outliers.

Vulnerability: Ignores the actual values of the extremes.

The Evolution of the Mean

1. Direct Method

$$\frac{\sum(f_i * x_i)}{\sum f_i}$$

Best for small numbers. Calculations become crushing as numbers grow.

2. Assumed Mean Method

$$a + \left[\frac{\sum(f_i * d_i)}{\sum f_i}\right]$$

Subtract an assumed center (a) to shrink values.

3. Step-Deviation Method

$$a + h * \left[\frac{\sum(f_i * u_i)}{\sum f_i}\right]$$

Divide by class size (h) to reduce math to single digits. Maximum formula complexity, minimum calculation friction.

Blueprinting the Mean: Step-Deviation in Action

Wickets Taken (45 Bowlers)

Class	f_i	x_i	u_i	$f_i * u_i$
30 - 50	2	90	-3	-6
50 - 100	4	100	-2	30
100 - 150	3	150	-0	20
150 - 250	12	200	0	0
200 - 300	3	210	1	-10
Totals	45	$\sum f_i = 45$		$\sum f_i * u_i = -106$

Class Size
h = 20

Assumed Mean
a = 200

The Formula Anatomy Engine

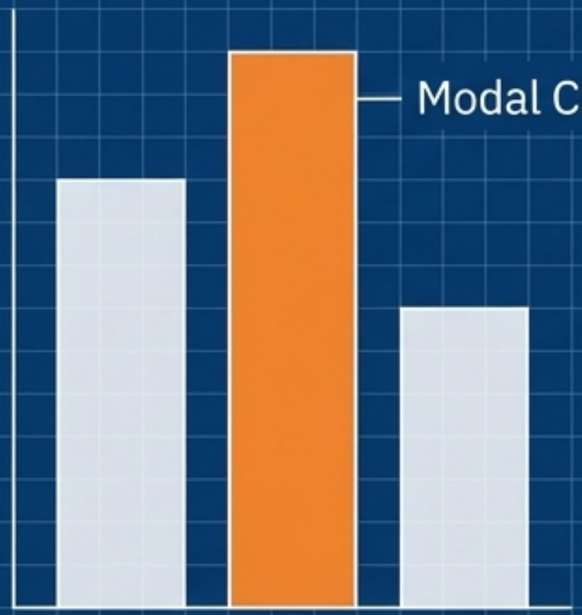
$$\bar{x} = a + h * \frac{\sum f_i u_i}{\sum f_i}$$

Output:

$$200 + 20 * \left(\frac{-106}{45}\right)$$

$$\bar{x} = 152.89 \text{ wickets}$$

The Anatomy of the Mode



$$\text{Mode} = l + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] * h$$

The width of the bar.

The starting line (lower limit of tallest bar).

The Tension Ratio

This calculates how much taller the modal class is compared to its immediate neighbors, pulling the mode away from the center toward the taller neighbor.

Blueprinting the Mode: Finding the Peak

Student Marks Distribution

Marks	Number of Students (f_i)
10 – 25	2
25 – 40	3
40 – 55	7
55 – 70	6
70 – 85	6
85 – 100	6

Modal Class

$l = 40$
$h = 15$
$f_0 = 3$ (Previous)
$f_1 = 7$ (Peak)
$f_2 = 6$ (Next)

The Formula Anatomy Engine

$$\text{Mode} = l + \left[\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right] * h$$

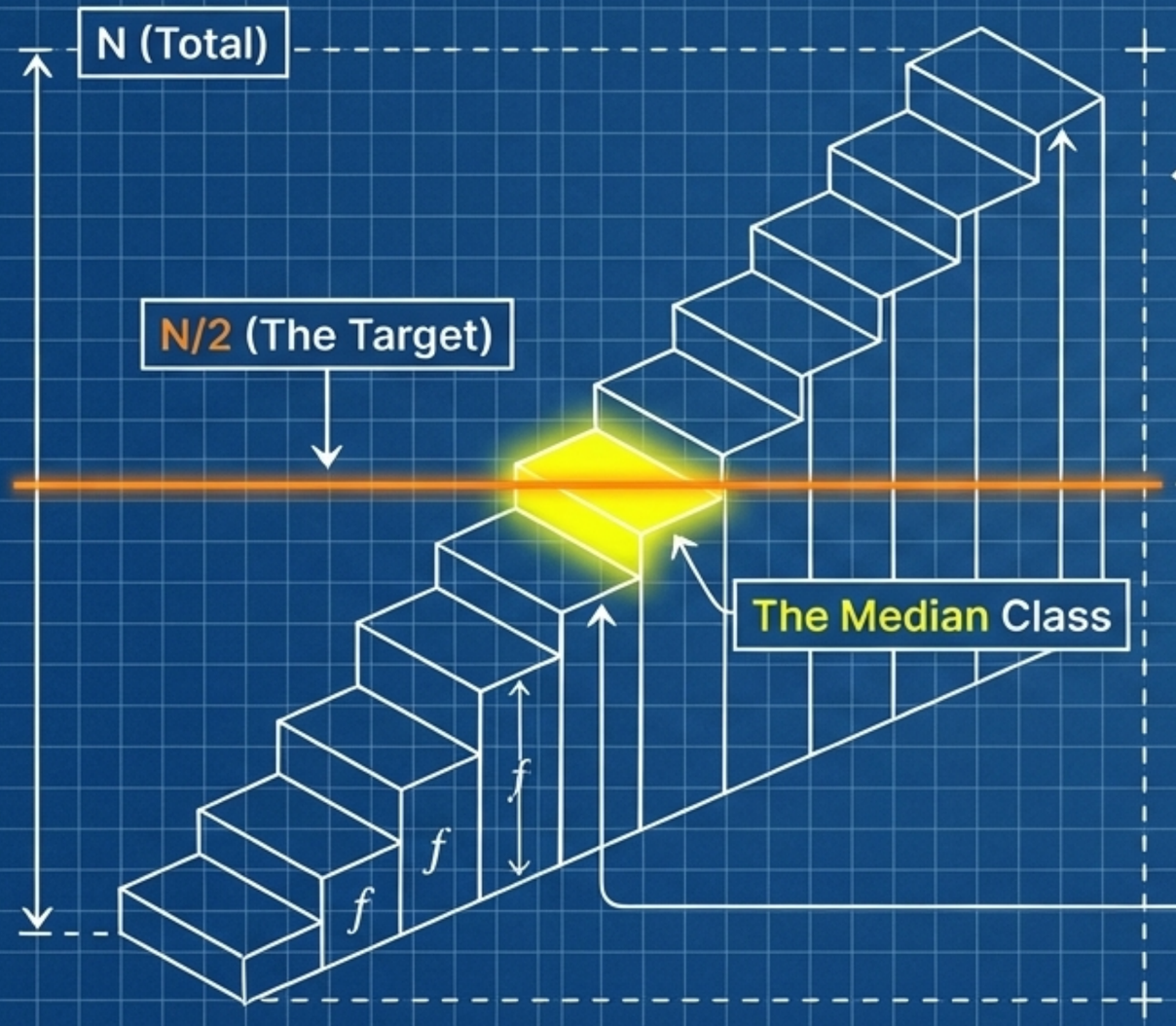
Output:

$40 + [(7 - 3) / (14 - 3 - 6)] * 15$

$40 + [\ 4\ /\ 5] * 15$

Mode = 52 marks

The Median: Climbing the Cumulative Stairs



$$\text{Median} = \ell + \left[\frac{N/2 - cf}{f}\right] * h$$

Blueprinting the Median: Locating N/2

Target Lock: N = 53 → N/2 = 26.5

Marks of 53 Students

Marks	f	Cumulative Freq (cf)
40 – 50	3	18
50 – 60	4	22
50 – 70	7	29
60 – 70	7	29
70 – 80	9	38

First value > 26.5

l = 60
cf = 22
(from row above)
f = 7
h = 10

The Formula Anatomy Engine

$$\text{Median} = l + \left[\frac{N/2 - cf}{f}\right] * h$$

Output:

60 + [(26.5 – 22) / 7] * 10

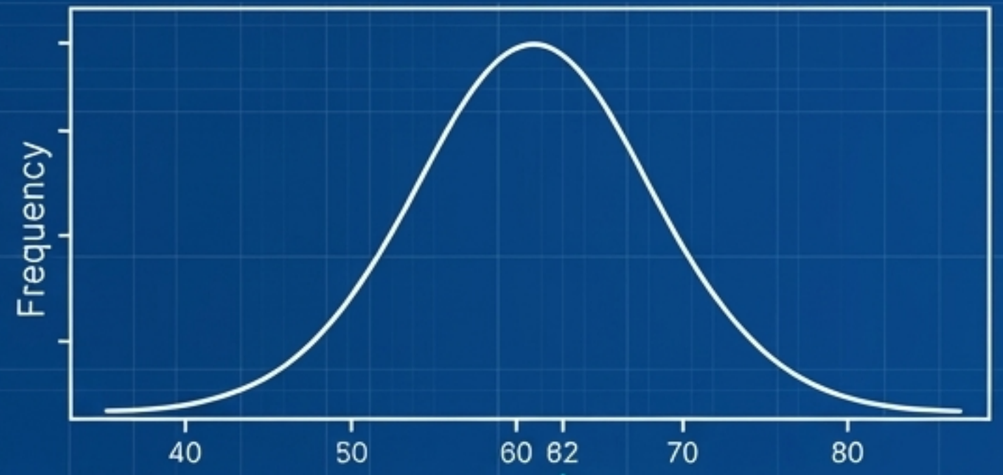
60 + 45 / 7

Median = 66.4 marks

One Dataset. Two Different Stories.

The Mean Story

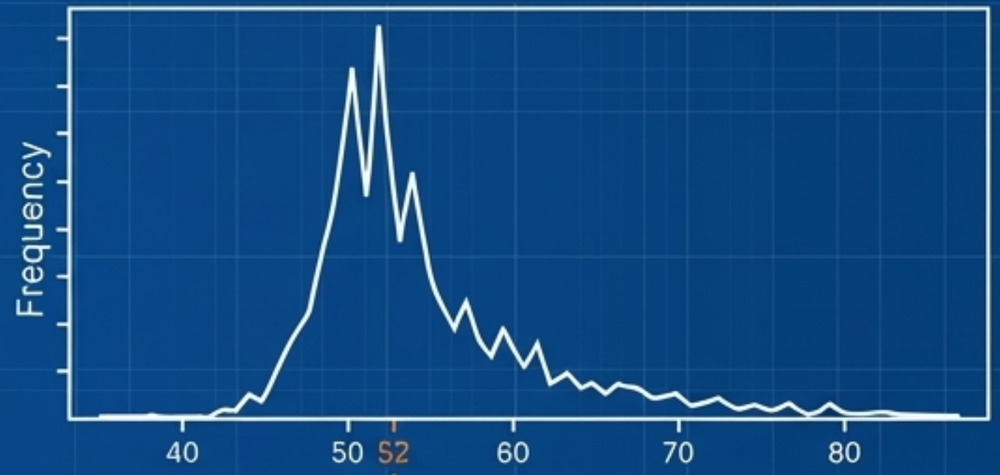
Result: 62 Marks



The Administration's View
On average, the class performed quite well, pulled up by high achievers.

The Mode Story

Result: 52 Marks



The Student's View
Despite the high average, the single largest cluster of students actually scored a much lower 52.

Choosing the right formula isn't just about math; it's about deciding which truth you need to reveal.

The Master Blueprint

MEAN

$$a + h * \frac{\left[\sum(f_i * u_i)\right]}{\sum f_i}$$

TRIGGER: Use when you need the center of gravity and have no extreme outliers.

MODE

$$l + \left[\frac{(f_1 - f_0)}{2f_1 - f_0 - f_2)}\right] * h$$

TRIGGER: Use when you need to find the most frequent occurrence or popular consensus.

MEDIAN

$$l + \left[\frac{(N/2 - cf)}{f}\right] * h$$

TRIGGER: Use when your data is heavily skewed or you need the exact 50th percentile.

Raw data is noise. Grouped statistics are the signal.