A 'Mean' Question

MATHEMATICS CO-DEVELOPMENT GROUP

In the first three parts of this series, we unpacked the median and the mode formulas. Comparatively the formula for the mean is easier to understand and not as counter intuitive. However, while computing it for grouped data, we use the midpoints of the class intervals and compute the mean as if all data in each class interval is exactly the midpoint. Consider Table 1.

Class Intervals	Midpoints	Frequencies
0 – 10	5	5
10 – 20	15	10
20 – 30	25	25
30 – 40	35	30
40 – 50	45	20
50 – 60	55	10
Total		100

Table 1

Data values	Frequencies
5	5
15	10
25	25
35	30
45	20
55	10
	100

Table 2

The mean is computed as $(5 \times 5 + 15 \times 10 + 25 \times 25 + 35 \times 30 + 45 \times 20 + 55 \times 10)/100$. This is identical to computing the mean for the ungrouped data given in Table 2. Note that the actual data for the class interval 0-10 can be 1, 1, 2, 2, 4 (adding up to 10) or 5, 6, 8, 8, 9 (adding up to 36). But we are assuming that they sum to $5 \times 5 = 25$. Since we can't get the actual data values in each interval, we must approximate. So, **why do we choose the midpoints?** This article tries to unpack that.

Keywords: Mean, grouped, ungrouped, fair-share, fulcrum, moment, modelling, analysing.

But before answering that question, we considered two models of mean and could link the two for ungrouped data. The two models are (i) the fair-share model and (ii) the fulcrum model. In the **fair-share model**, the data values are all pooled together and then shared equally. Here is an example: 10 couples who are also parents were surveyed for number of children. Figure 1 shows the no. of children for each of these couples (C_1, C_2 … C_{10}). So, there are six couples with one child, viz., C_1, C_2, … C_6; three couples with two children, i.e., C_7, C_8 and C_9 and the last couple C_{10} with four children.

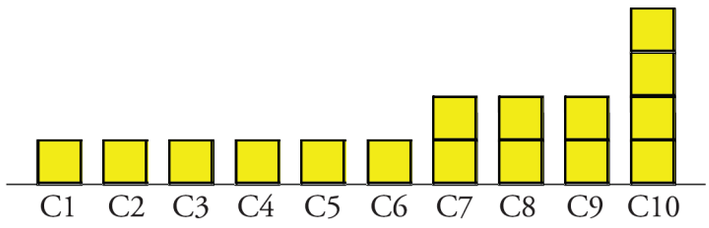

Figure 1

No. of children	No. of couples
1	6
2	3
3	0
4	1

Table .

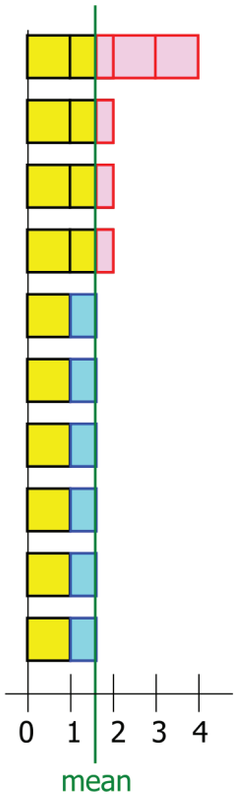

Figure 2

Now, in the fair-share model, all the yellow squares must be shared equally among the 10 couples. Then the resulting common height of the rectangle for each couple is the 'mean'. Figure 2 represents this mean (calculated to be 1.6) with the rows and columns flipped, i.e., each row now represents a couple. [The reason for the flip would become clear soon.] The pink parts have been redistributed to form the blue parts, so they have equal areas. Also, the mean is the common length of each row after redistributing the rectangles. So, the common length is the total yellow area redistributed equally among the rows, i.e., $(6 \times 1 + 3 \times 2 + 0 \times 3 + 1 \times 4) / 10 = 1.6$. Now the total blue area is $6 \times (1.6 - 1)$ and the total pink area is $3 \times (2 - 1.6) + 1 \times (4 - 1.6)$ when computed row by row.

Compare this to the stick representation (Figure 3) illustrating the fulcrum of the distribution. Now, in the **fulcrum model**, the 'mean' is where the fulcrum must be placed to balance the distribution (Figure 3 based on the frequency distribution given in Table 3). So, the total moments on either side of the fulcrum must be equal.

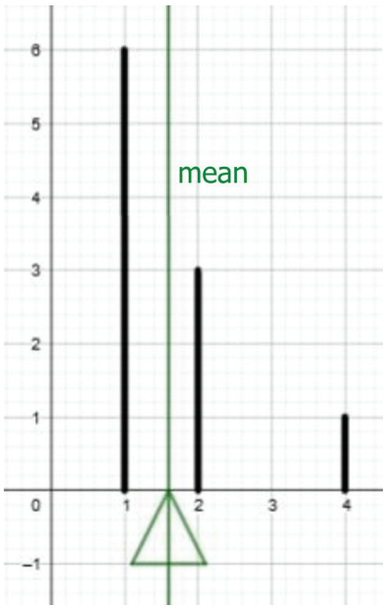

Figure 3

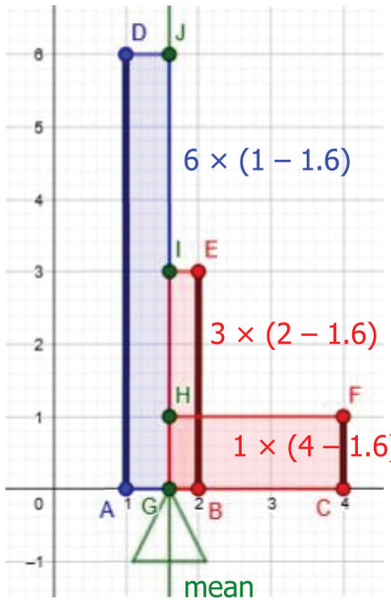

Figure 4

So, the total moment on the left of the fulcrum is indicated by the area of the blue rectangle, i.e., $6 \times (1.6 - 1)$ in Figure 4. The height of each rectangle is the frequency while the base of each one is the difference 'mean – data value'. Similarly, the total moment on the right is given by the areas of the two pink rectangles, i.e., $3 \times (2 - 1.6)$ and $1 \times (4 - 1.6)$. Again, the heights are the frequencies, but the bases are 'data value – mean'. However, to combine the areas in one formula, we need to write them as 'data value - mean' (or 'mean - data value') for each rectangle. Therefore, in Figure 4, since the data value $= 1 <$ $1.6 =$ the mean, 'data value - mean' is negative. Thus, the area of the blue rectangle can be considered to be negative. Note that this area is of equal magnitude to the total pink area. In other words, $6 \times (1.6 - 1) = 3 \times (2 - 1.6) + 1 \times (4 - 1.6)$. Also observe the following:

- The blue rectangle AGJD in Figure 4 has the same area (3.6) as that of all the blue rectangles in Figure 2. In fact, if the blue rectangles are lined up by removing the gaps among them, then they would form the same blue rectangle AGJD.
- The pink rectangle GCFH in Figure 4 has the same area (2.4) as all the pink rectangles in top row of Figure 2. They in fact have the same dimensions.
- The pink rectangle GBEI in Figure 4 has the same area (1.2) as all the remaining pink rectangles in rows 7-9 in Figure 2. Like the blue one, if the pink rectangles are joined by removing the gaps among them, then they would form a rectangle with the same dimension as GBEI.

So, combining Figures 2 and 4, we observe that the fair-share and the fulcrum model have a deep connection. [This is why we flipped the graph in Figure 2.] We strongly encourage the reader to try to recreate Figure 2 and Figure 4 with any ungrouped data in order to understand this connection.

Algebraically speaking, if $x_1, x_2 \ldots x_k$ are the data values with frequencies $f_1, f_2 \ldots f_k$ respectively then the fulcrum is located at 'm' if the total moment from 'm' is 0, i.e., $\sum_{i=1}^{k} f_i (x_i - m) = 0$. Note that this generates $m = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$ which is the same formula derived from the fair-share model. For the above

example, this is $6(1-m)+3(2-m)+1(4-m)=0$, which reduces to $6\times 1+3\times 2+1\times 4=(6+3+1)\,m$, i.e., $m=\dfrac{6\times 1+3\times 2+1\times 4}{6+3+1}=\dfrac{16}{10}=1.6$, i.e., what we got earlier.

However, for a grouped frequency distribution, it is impossible to know the individual data values as mentioned earlier. So, fair-share model can't be applied directly. However, the fulcrum model can be adopted. Instead of a stick diagram as in Figures 3 and 4, we would use the histogram and the fulcrum would be the point where it can be balanced.

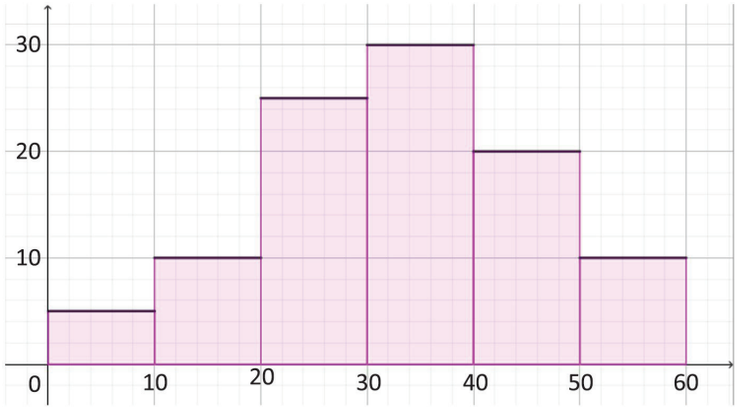

Figure 5

Let us consider the histogram corresponding to the grouped frequency distribution given in Table 1 (Figure 5). Now we can think of the corresponding step function $f(x)$ as follows (Figure 6):

$$
\begin{aligned}
f(x) &= 0 && \text{for } x<0\\
&= 5=f_1 && \text{for } 0\le x<10\\
&= 10=f_2 && \text{for } 10\le x<20\\
&= 25=f_3 && \text{for } 20\le x<30\\
&= 30=f_4 && \text{for } 30\le x<40\\
&= 20=f_5 && \text{for } 40\le x<50\\
&= 10=f_6 && \text{for } 50\le x<60\\
&= 0 && \text{for } x\ge 60
\end{aligned}
$$

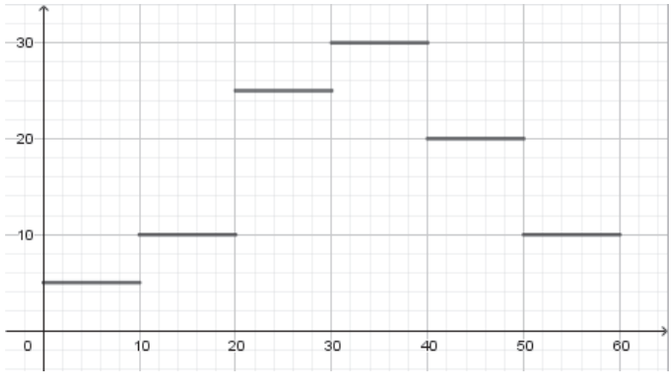

Figure 6

Then the sum for total moment becomes this integral

$$\int_0^{60} f(x)\,(x-m)\,dx=\sum_{i=1}^{6}\int_{10(i-1)}^{10i} f_i\,(x-m)\,dx=5\times\int_0^{10}(x-m)\,dx+10\times\int_{10}^{20}(x-m)\,dx+25$$
$$\times\int_{20}^{30}(x-m)\,dx+30\times\int_{30}^{40}(x-m)\,dx+20\times\int_{40}^{50}(x-m)\,dx+10\times\int_{50}^{60}(x-m)\,dx.$$

So, mean is 'm', the value that makes this integral zero.

Now, any general integral of this form, i.e., $\int_a^b (x-m)\,dx = \left[\left(\frac{b^2}{2} - bm\right) - \left(\frac{a^2}{2} - am\right)\right] = \frac{b^2-a^2}{2} - (bm - am) = (b-a)\left(\frac{a+b}{2} - m\right)$ using $\frac{b^2-a^2}{2} = (b-a)\frac{a+b}{2}$. Note that $b-a = 10 =$ common class width for each of these integrals.

So, this integral becomes

$$\begin{aligned}
&= 10\left\{5\left[\frac{10+0}{2} - m\right] + 10\left[\frac{20+10}{2} - m\right] + 25\left[\frac{30+20}{2} - m\right] + 30\left[\frac{40+30}{2} - m\right]\right. \\
&\quad \left. +20\left[\frac{50+40}{2} - m\right] + 10\left[\frac{60+50}{2} - m\right]\right\} \\
&= 10 \times \{5\,[5-m] + 10\,[15-m] + 25\,[25-m] + 30\,[35-m] + 20\,[45-m] + 10\,[55-m]\}\,.
\end{aligned}$$

So, the integral is 0 if and only if

$$5\,[5-m] + 10\,[15-m] + 25\,[25-m] + 30\,[35-m] + 20\,[45-m] + 10\,[55-m] = 0\ldots \qquad (1)$$

which is possible if and only if

$$m = \frac{5\times 5 + 10\times 15 + 25\times 25 + 30\times 35 + 20\times 45 + 10\times 55}{5+10+25+30+20+10} = \frac{3300}{100} = 33\ldots \qquad (2)$$

Note that (1) is exactly like $\sum_{i=1}^{k} f_i\,(x_i - m) = 0$ while (2) resembles $m = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$; both corresponding to the ungrouped distribution in Table 2 represented in Figure 7 by the stick representation. Note that the sticks are at the midpoint of each class interval and share the same height (i.e., frequency).

So, generally, when we consider the histogram of a grouped frequency distribution, the sum for total moment becomes the area under the histogram which is the integral $\int_{x_0}^{x_k} f(x)\,(x-m)\,dx = \sum_{i=1}^{k} \int_{x_{i-1}}^{x_i} f_i\,(x-m)\,dx = \sum_{i=1}^{k} f_i \times \int_{x_{i-1}}^{x_i} (x-m)\,dx$ where $x_0 - x_1$, $x_1 - x_2$ … $x_{k-1} - x_k$ are the class

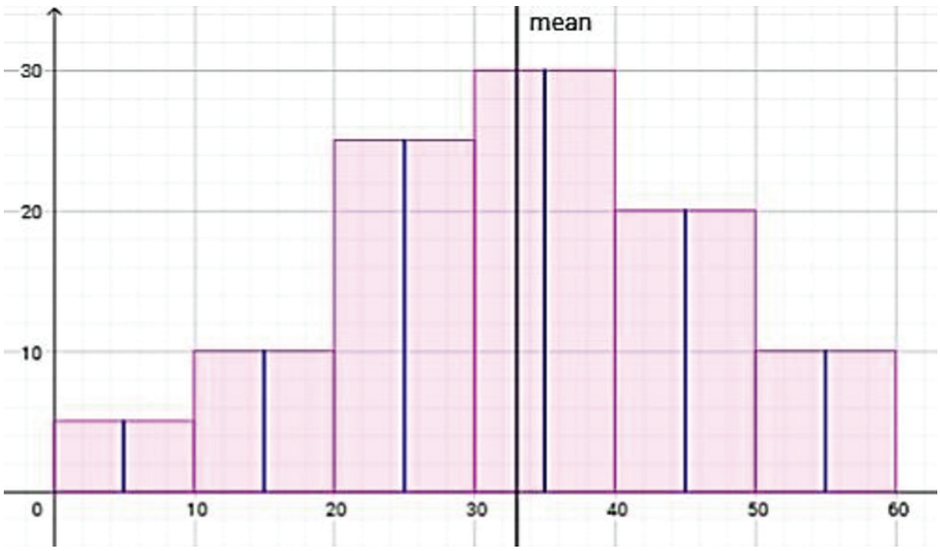

Figure 7

intervals with frequencies $f_1, f_2 \ldots f_k$ respectively where the mean is that value of 'm', which makes this integral zero. Therefore, the integral becomes

$$= \sum_{i=1}^{k} f_i \times \int_{x_{i-1}}^{x_i} (x - m)\, dx = \sum_{i=1}^{k} f_i \left[\left(\frac{x_i^2}{2} - m x_i \right) - \left(\frac{{x_{i-1}}^2}{2} - m x_{i-1} \right) \right]$$

$$= \sum_{i=1}^{k} f_i (x_i - x_{i-1}) \left[\frac{x_i + x_{i-1}}{2} - m \right]$$

$$= h \times \sum_{i=1}^{k} f_i \left[\frac{x_i + x_{i-1}}{2} - m \right]$$

where h is the class interval (which is usually the same for all classes).

Now, $\frac{x_i + x_{i-1}}{2} = y_i$ is nothing but the midpoint of the class interval $x_{i-1} - x_i$ for i = 1, 2 … k. So, the integral becomes $h \times \sum_{i=1}^{k} f_i (y_i - m)$ which is similar to ungrouped frequency distribution with the midpoints $y_1, y_2 \ldots y_k$ as the data values. This converts the histogram into a stick representation.

Note that if we find the fulcrum for each class, then that is also the midpoint of each class by the symmetry of rectangles. This can also be arrived at with integration.

It is interesting to observe that while the formulas for median and mode for grouped data looked very complicated, they required nothing beyond Class 10 syllabus to decipher. However, a much simpler looking process for calculating mean for group data requires a much more sophisticated tool like integration to understand it.

Math Co-dev Group or more elaborately **Mathematics Co-development Group** is an internal initiative of Azim Premji Foundation where math resource persons across states put their heads together to prepare simple materials for teachers to develop their understanding on different content areas and how to transact the same in their classrooms. It is a collaborative learning space where resources are collected from multiple sources, critiqued and explored in detail. Math Co-dev Group can be reached through ashish.gupta@azimpremjifoundation.org