Deciphering the Mode Formula

MATHEMATICS CO-DEVELOPMENT GROUP

In Parts 1 and 2 of this series on measures of central tendencies (published in Jul 2021 and Nov 2021 respectively) we discussed the median formula $M = l + \frac{\frac{N}{2} - m}{f} \times c$ for grouped data from the corresponding histogram and the ogives. In Part 3, we will explore the mode formula $m = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$ …(1) from the corresponding histogram.

Mode can be obtained for both quantitative and qualitative data. For ungrouped data, it is simply the data value with the highest frequency. So, if data is ordered (according to some way for categorical data, for the rest there is an obvious order), then the longest run implying maximum frequency indicates the mode. For example, when we consider the choice of fruits for a class of 20 students and order it alphabetically, we get:

apple	apple	apple	apple	apple			
banana	banana	banana	banana	banana	banana	banana	banana
guava	guava	guava	guava				
orange	orange	orange					

Therefore, banana is the mode because it has the highest frequency. Note that neither mean, nor median work for categorical data.

Keywords: data, mode, modal class, frequency

Mode can be obtained from a frequency table also. For example, consider the academic qualifications of mothers of a class of 40 students (Table 1):

Academic qualification	Number of mothers
Primary school	7
Middle school	8
Secondary	10
Higher secondary	6
Graduate	5
Postgraduate	4

Table 1

Clearly, 'secondary' is the mode with the highest frequency of 10.

It is possible for a data set to have more than one mode.

Along the same lines, for grouped data, the class interval with the highest frequency is called the **modal class**. But how to figure out the exact location of the mode within this interval? That's what we are going to explore in this article.

Let us consider the marks obtained by two groups of students (Table 2):

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of students in Group 1	20	21	27	16	11	10	7	3	1
No. of students in Group 2	20	21	27	24	11	10	7	3	1

Table 2

The modal class for both groups is the same, i.e., 20-30 with frequency 27 each. However, the frequency distributions are different. So, the shapes of the histograms (Figure 1 and Figure 2) are different and therefore the modes should be different too. So, the midpoint of the modal class should not be the mode since it is independent of the frequencies. In fact, for Group 2, the class after the modal class has a much higher frequency (24) than the class before (21). So, the mode for Group 2 should be higher than that of Group 1, where the frequency of the class before (21) is higher than the class after (16).

It is easy to identify the modal class from the histogram. It corresponds to the tallest rectangle. Now, there are six points on that rectangle. Let us consider the histogram for Group 1 (Figure 1) for now. So, the three points on the left side of the rectangle are G (20, 27), F (20, 21) and T (20, 0) while those on the right side are H (30, 27), K (30, 16) and U (30, 0).

Since T and U have 0 as y-coordinates, they are independent of the frequencies and hence provide less information than the other four, i.e., G, F, H and K.

The y-coordinates of the top two points, i.e., G and H, are the highest frequency 27 and those of the middle two points, i.e., F and K, are the frequencies of the neighbouring classes, i.e., 21 and 16 respectively.

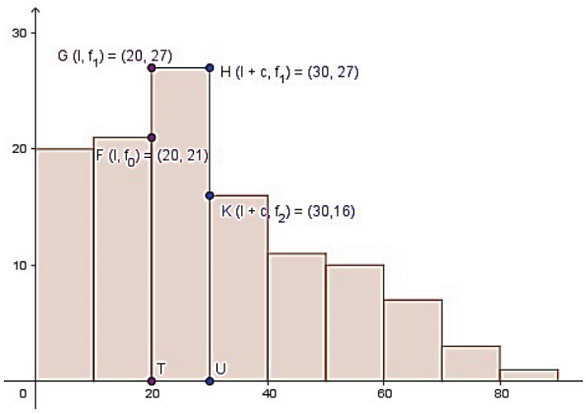

Figure 1. Marks from Group 1

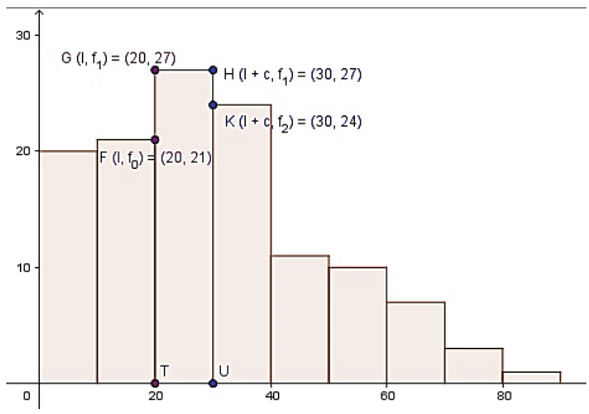

Figure 2. Marks from Group 2

It is understood that the mode lies somewhere within the modal class, i.e., mode is some x-value between 20 and 30 and is not necessarily the midpoint, i.e., (20 + 30)/2 = 25. So, the only way to get a x-value within the modal class or a point in the modal class rectangle is to draw two lines, each connecting points on either vertical side of the rectangle. Let's call these lines 'diagonals'. Since the choice of points must be uniform on both sides, there are three choices:

1. GU and HT: the diagonals intersect at the centre of the modal class rectangle making 25 the mode regardless of the frequencies, which is not desirable as mentioned already
2. FU and KT: modal frequency is not considered and therefore not desirable
3. GK and HF: only option left

So, mode is defined as the x-coordinate of the intersection point E of these two 'diagonals' GK and HF.

Let us look at the modal class in the histogram (Figure 3). Let the top border of the modal class be GH with G (20, 27) and H (30, 27) while F (20, 21) is the top-right corner of the rectangle for the class just before modal class and K (30, 16) is the top-left corner of the rectangle for the class just after the modal class.

Mode is the x-coordinate of the point of intersection E (m, f) of the line segments GK and HF. Let J be the point on GH directly above E, i.e., EJ $\perp$ GH and J $(m, f_1) = (m, 27)$.

$$\begin{array}{lll} \text{Now} & \triangle \text{GJE} \approx \triangle \text{GHK} & \text{and} \quad \triangle \text{HJE} \approx \triangle \text{HGF} \\ \Rightarrow & \text{JG} : \text{HG} = \text{JE} : \text{HK} & \text{and} \quad \text{HJ} : \text{HG} = \text{JE} : \text{GF} \\ \therefore & \text{JG} \cdot \text{HK} = \text{JE} \cdot \text{HG} = \text{HJ} \cdot \text{GF} & \text{i.e.,} \quad \text{JG} \cdot \text{HK} = \text{HJ} \cdot \text{GF} \ldots . (2) \end{array}$$

Now, JG $= m - 20$, HK $= 27 - 16$, HJ $= 30 - m$ and GF $= 27 - 21$

So,

$$\begin{array}{ll} (m - 20)(27 - 16) = (30 - m)(27 - 21) & \\ \Rightarrow 11(m - 20) = 6 \times 10 - 6(m - 20) & \text{since} \quad \text{HJ} = \text{HG} - \text{JG} = 10 - (m - 20) \\ \Rightarrow (11 + 6)(m - 20) = 6 \times 10 & \Rightarrow m = 20 + \frac{60}{17} = 23.53 \end{array}$$

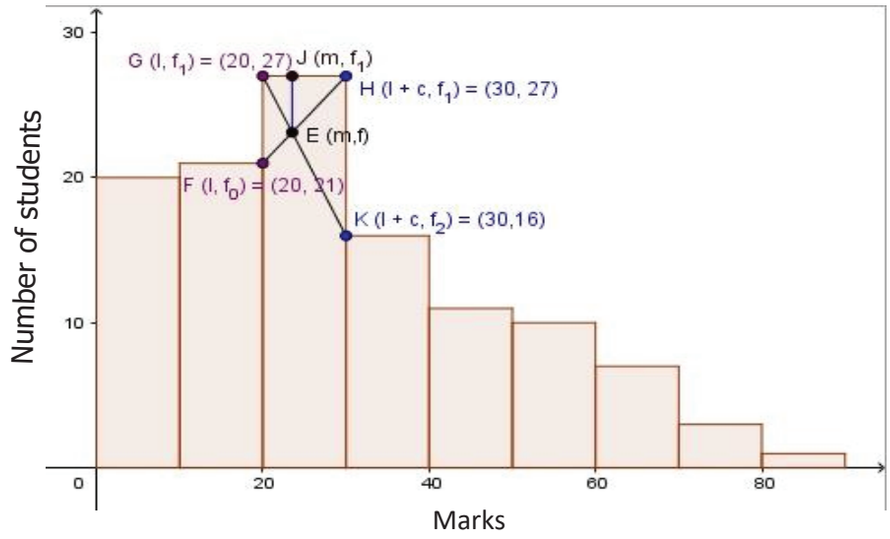

Figure 3. Group 1

Observe that for Group 2, the same process yields the mode to be:

$$m = 20 + \frac{27-21}{(27-21)+(27-24)} \times 10 = 20 + \frac{60}{9} = 26.67$$

which is higher than the mode for Group 1

Symbol	Meaning	In the examples
c	(Uniform) class-width	10
l	Lower limit of modal class	20
f_1	Frequency of modal class	27
f_0	Frequency of the class before the modal class	21
f_2	Frequency of the class after the modal class	16 (G1), 24 (G2)

Table 3

Note that this looks similar to the formula mentioned at the beginning. Now, let us generalize by algebraizing as follows:

So, G $= (l, f_1)$, H $= (l + c, f_1)$, F $= (l, f_0)$ and K $= (l + c, f_2)$. As before mode m is the x-coordinate of the point of intersection E (m, f) of the line segments GK and HF. Also, J (m, f_1) is the point on GH directly above E, i.e., EJ $\perp$ GH.

So,

$$\text{GJ} \cdot \text{HK} = \text{HJ} \cdot \text{GF} \Rightarrow (m-l) \cdot (f_1 - f_2) = (l + c - m) \cdot (f_1 - f_0)$$
$$\Rightarrow (m-l) \cdot (f_1 - f_2) = c \cdot (f_1 - f_0) - (m-l) \cdot (f_1 - f_0)$$
$$\Rightarrow (m-l) \cdot ((f_1 - f_2) + (f_1 - f_0)) = c \cdot (f_1 - f_0)$$
$$\Rightarrow (m-l) \cdot (2f_1 - f_0 - f_2) = c \cdot (f_1 - f_0) \qquad \Rightarrow m - l = \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$
$$\Rightarrow m = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

It may make sense to let different groups of students work with different histograms, collate their work in a table like Table 3 and then crystalize the algebraic form of the formula from them.

But what happens when the modal class is at either end of the histogram? That is either the previous class with frequency f_0 is missing or the class after with frequency f_2 is not there. Let us consider the following case where modal class is the lowest class:

Age	Percentage of population
65-70	29.2
70-75	22.6
75-80	18.7
80-85	14.8
85-90	9.2
90-95	4.0
95+	1.5
Total	100.0

Table 4. Age distribution of US population of age 65yrs and over in 2008

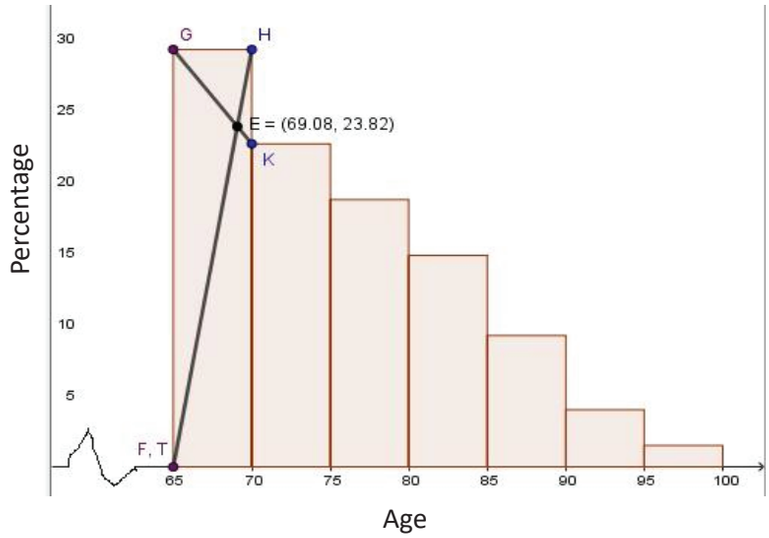

Figure 4. Age distribution

Note that in this case F and T have coincided and $f_0 = 0$ because there is no class before the modal class. So, the mode is

$$m = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c = 65 + \frac{29.2 - 0}{2\,(29.2) - 0 - 22.6} \times 5 = 69.08.$$

So, if the modal class is the lowest class, then the formula holds with $f_0 = 0$. Similarly, if the modal class is the highest class, then the same applies to f_2 i.e., $f_2 = 0$.

Another related question: What if there are two (or more) consecutive modal classes (with the same frequency)? In that case, the modal classes should be combined to form a wider modal class with the class interval twice (or more times) that of other classes. Then this combined (and fatter) modal class would be

the rectangle GHUT in the histogram with the point F and K on either vertical side, E as the intersection of GK and HF, and J as the point on GH just above E. The same process holds.

As we observed for the median formula, the derivation of the mode formula is also not difficult. It uses histogram and basic coordinate geometry along with a little bit of similar triangles – all very much part of secondary school syllabus. So, instead of prescribing a pretty complicated formula (without any clue to why or how it formed), it is better to teach the students the underlying reasoning involving the line segments GK and HF. The same principle can enable them to tackle cases like modal class at an end or multiple consecutive modal classes.

In the next and final article in this series we would discuss why the midpoints of the class intervals are used to compute the mean of a grouped data set. It would be particularly relevant since we argued that the midpoint (of the modal class) is not an ideal choice for the mode.

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 yashvendra@azimpremjifoundation.org