Deciphering the Median Formula

Part 1: From the Histogram

MATHEMATICS CO-DEVELOPMENT GROUP, Azim Premji Foundation

At the secondary level, data handling becomes statistics. There is a new graph – the ogive and two quite complicated formulas (i) $M = l + \frac{\frac{N}{2} - m}{f} \times c$ and (ii) $m = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$ for computing the median and the mode respectively from grouped data, i.e., data that is organized in class intervals. Textbooks usually do not explain how those formulas work even though the logic can be completely worked out within secondary mathematics content. In this series, we are going to decipher these formulas. We shall also investigate a question related to the mean. So, there are four parts as follows:

1. Median formula from the histogram
2. Median formula from the ogive, and why the ogives intersect at the median
3. Mode formula from the histogram
4. A 'Mean' question

For ungrouped quantitative data, median splits the entire data set in two parts – each with the same number of data points.

If the daily salaries of 7 employees in a company (in decreasing order) are ₹5000, ₹500, ₹450, ₹430, ₹400, ₹150 and ₹100, then the median is the 4^{th} salary ₹430 which splits the salaries in two groups {₹5000, ₹500, ₹450} and {₹400, ₹150, ₹100} – each with 3 salaries (or data points).

Similarly, if the salaries are ₹5000, ₹500, ₹450, ₹430, ₹410, ₹400, ₹150 and ₹100, then the median is the mean of the 4^{th} and the 5^{th} salaries i.e. ₹420 which splits the salaries in two groups {₹5000, ₹500, ₹450, ₹430} and {₹410, ₹400, ₹150, ₹100} – each with 4 data points.

Keywords: Data, Median, Histogram, Developing the formula

So, if there are an odd number $2n + 1$ data points, then we order them and pick the $(n + 1)^{th}$ data point as the median. If there are an even number $2n$ of data points, then we take the mean of the nth and the $(n + 1)^{th}$ data points as the median. For both cases i.e. odd and even number of data points, the first n data points have values lower than the median and the remaining n points have values higher than it. Note that, for the even case, any value in between the n^{th} and the $(n + 1)^{th}$ data points would have worked as median in terms of splitting the data set in two subsets with equal cardinality. The above examples illustrate these with $n = 3$ for odd and with $n = 4$ for even number of data points.

Marks	No. of students
20	6
25	20
28	24
30	28
33	15
38	4
42	2
43	1

Table 1

Marks	Cumulative frequency
≤ 20	6
≤ 25	$6 + 20 = 26$
≤ 28	$26 + 24 = 50$
≤ 30	$50 + 28 = 78$
≤ 33	$78 + 15 = 93$
≤ 38	$93 + 4 = 97$
≤ 42	$97 + 2 = 99$
≤ 43	$99 + 1 = 100$

Table 2

Let us consider the frequency table (Table 1) for marks (out of 50) obtained by 100 students. The marks of these 100 students can be put in decreasing order: 43, 42, 42, 38, 38, 38, 38, 33,… 25, 20, 20, 20, 20, 20, 20. So, it is possible to find the 50^{th} and the 51^{st} marks. We can find it more easily with a cumulative frequency table (Table 2). The 50^{th} mark = 28 and the 51^{st} mark = 30. So, the median is 1/2(28 + 30) = 29.

Marks		No. of students	Marks	Cumulative frequency
150-154	149.5-154.5	5	≤ 154.5	5
155-159	154.5-159.5	2	≤ 159.5	$5 + 2 = 7$
160-164	159.5-164.5	6	≤ 164.5	$7 + 6 = 13$
165-169	164.5-169.5	8	≤ 169.5	$13 + 8 = 21$
170-174	169.5-174.5	9	≤ 174.5	$21 + 9 = 30$
175-179	174.5-179.5	11	≤ 179.5	$30 + 11 = 41$
180-184	179.5-184.5	6	≤ 184.5	$41 + 6 = 47$
185-189	184.5-189.5	3	≤ 189.5	$47 + 3 = 50$

Table 3

Now consider another frequency table (Table 3) for marks (out of 100) obtained by 50 students. In this case, we have no way of finding the exact marks, the x_1, x_2, … x_{50}. So, we can't order them to find the median using the above method. So, we turn to the graphs. As mentioned above, in this article we are going to use the histogram. So, we make the histogram (Figure 1) for the grouped data in Table 3 (and make the class intervals continuous i.e. 149.5-154.5, 154.5-159.5, etc., to draw the graph).

Note that median halves the dataset. We will use something similar with histogram. The area of each rectangle = the frequency of the corresponding class × the class width. Since the width is uniform across classes, the heights of the rectangles are proportionate to the respective class frequencies. So, the total area of the histogram is proportional to the total frequency.

Therefore, median should be that value of x that cuts the histogram in two parts (blue and pink) with equal areas. In other words, M is the median if the line $x = M$ cuts the histogram in two parts each with half the area of the total histogram. This line cuts through the rectangle corresponding to the class where the cumulative frequency crosses half of total frequency.

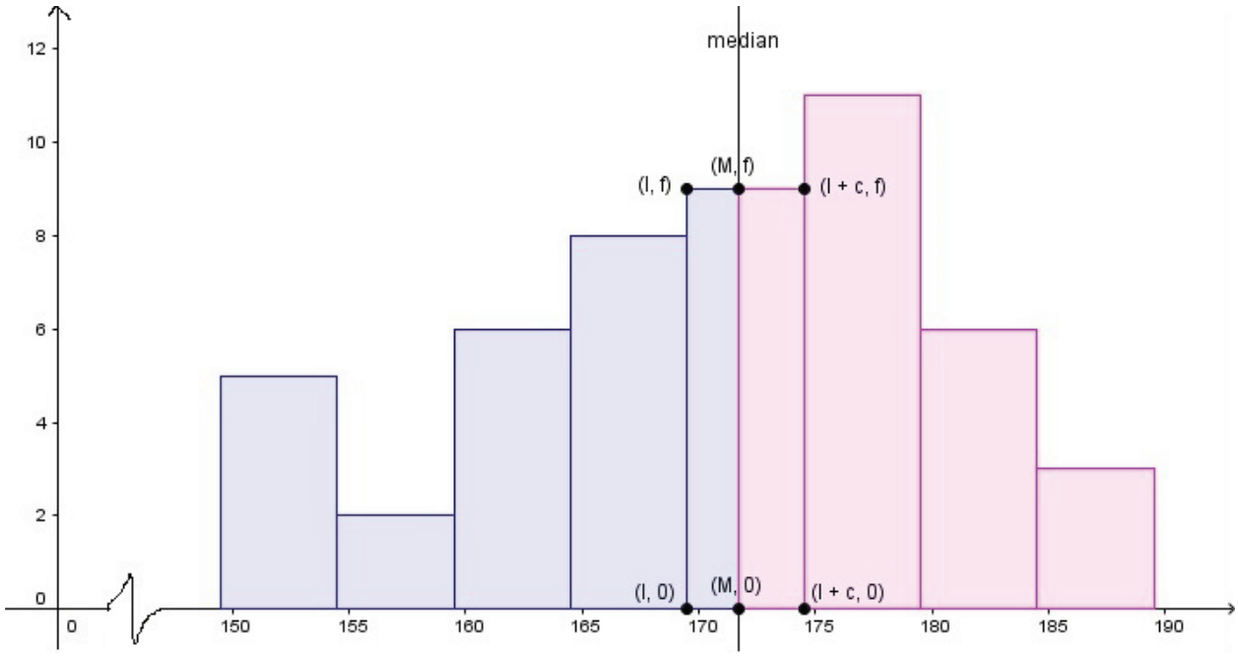


Figure 1

In this case, half of the total frequency is $50/2 = 25$. The cumulative frequency up to the class 169.5-174.5 is $21 < 25$ and after this class it becomes $30 > 25$. So, the median i.e. the 25th data value should be somewhere in the class 169.5-174.5. We call this class, 169.5-174.5, the **median class**. So, the cumulative frequency crosses 25 for 174.5 which is the upper limit of the median class.

The total area of the histogram is the sum of areas of all rectangles.

The area of each rectangle is the class-width × the corresponding frequency.

This histogram has uniform class-width 5.

So, the total area is $5 \times (5 + 2 + 6 + 8 + 9 + 11 + 6 + 3) = 5 \times 50$, i.e., class-width × total frequency.

So, the area of each part (blue or pink) is $1/2 \times 5 \times 50$.

Now the blue part consists of

(i) the full rectangles corresponding to the classes before the median class with area $5 \times (5 + 2 + 6 + 8) = 5 \times 21$

(ii) a part of the median class with area $(M - 169.5) \times 9$

Note that:

- 169.5 is the lower limit of the median class
- 21 is the cumulative frequency (less than) for 169.5
- 9 is the frequency of the median class.

So, the area of the blue part is

$$5 \times 21 + (M - 169.5) \times 9 = \frac{1}{2} \times 5 \times 50 \Rightarrow (M - 169.5) \times 9 = 5 \times (50/2 - 21)$$

$$\Rightarrow M - 169.5 = 5/9 \times (50/2 - 21) \Rightarrow M = 169.5 + (50/2 - 21) \times 5/9$$

Now, doesn't this look similar to the median formula mentioned at the beginning?

Let us generalize by algebraizing as follows:

Symbol	Meaning	In the example
N	Total frequency	50
c	(uniform) class-width	5
l	Lower limit of median class	169.5
f	Frequency of median class	9
m	(Less than) cumulative frequency for l	21

Table 4

- The total area of the histogram is Nc
- The area of the full blue rectangles is mc and that of the blue part of the median class is $(M - l)f$

Therefore, the total area of the blue part is

$$mc + (M - l)f = \frac{1}{2}Nc \qquad \Rightarrow (M - l)f = \frac{1}{2}Nc - mc = \left(\frac{N}{2} - m\right)c$$

$$\Rightarrow M - l = \frac{\left(\frac{N}{2} - m\right)c}{f} = \frac{\frac{N}{2} - m}{f} \times c \quad \Rightarrow M = l + \frac{\frac{N}{2} - m}{f} \times c$$

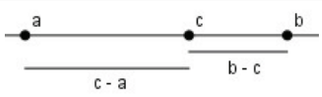
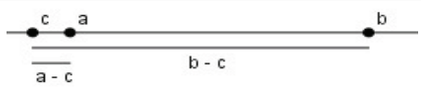
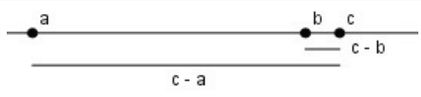
It may make sense to let different groups of students work with different histograms, collate their work in a table like Table 4 and then crystalize the algebraic form of the formula from them.

It is worth mentioning that median also minimizes mean deviation i.e. M = median minimizes $\frac{1}{n}\sum_{k=1}^{n} |x_k - M|$ where $x_1, x_2, x_3 \ldots x_n$ are the data points.

We will prove this in three steps [c represents a constant in the remaining part, not class-width]:

1. Mean deviation from c i.e. $\dfrac{|x_1 - c| + |x_2 - c| + \cdots + |x_n - c|}{n}$ is minimized when the sum of the deviations i.e. $S = |x_1 - c| + |x_2 - c| + \ldots + |x_n - c|$ is minimized since n is independent of c

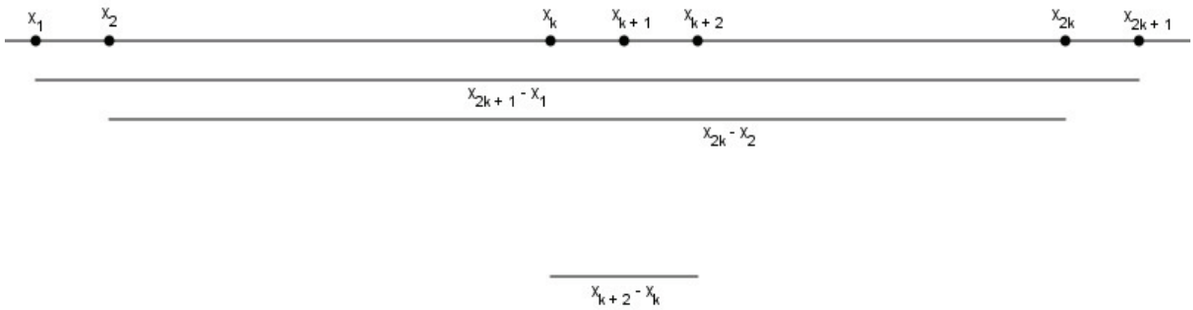
2. Show that if $a < b$ then $|a-c|+|b-c|$ is minimized when a $< c < b$
   There are three possibilities:

$a < c < b$	$c < a < b$	$a < b < c$
$\|a-c\| + \|b-c\|$ $= c - a + b - c$ $= b - a$	$\|a-c\| + \|b-c\|$ $= a - c + b - c = a + b - 2c$ $= b - a + 2a - 2c$ $= b - a + 2(a - c)$ $> b - a$ since $a - c > 0$	$\|a-c\| + \|b-c\|$ $= c - a + c - b = 2c - (a + b)$ $= 2c - 2b + b - a$ $= b - a + 2(c - b)$ $> b - a$ since $b < c$
		

   Therefore $|a-c|+|b-c|$ is minimized when $a < c < b$

3. We now prove that the median minimizes the mean deviation

Let us first look at the situation for odd $n = 2k+1$ with $x_1 < x_2 < \ldots < x_{k+1} < \ldots < x_n$ without loss of generality. Since $(n+1)/2 = k+1$, so x_{k+1} is the median.



We need to show that the sum of the deviations i.e. $S = |x_1 - c|+|x_2 - c|+\ldots+|x_n - c|$ is minimized when $c = x_{k+1}$

Now this sum can be regrouped in pairs with

- the 1^{st} pair corresponding to the extreme values x_1 and x_{2k+1}
- the 2^{nd} pair corresponding to the next two i.e. x_2 and x_{2k} etc. till
- the k^{th} pair corresponding to x_k and x_{k+2} and
- the single term corresponding to the median x_{k+1} i.e.

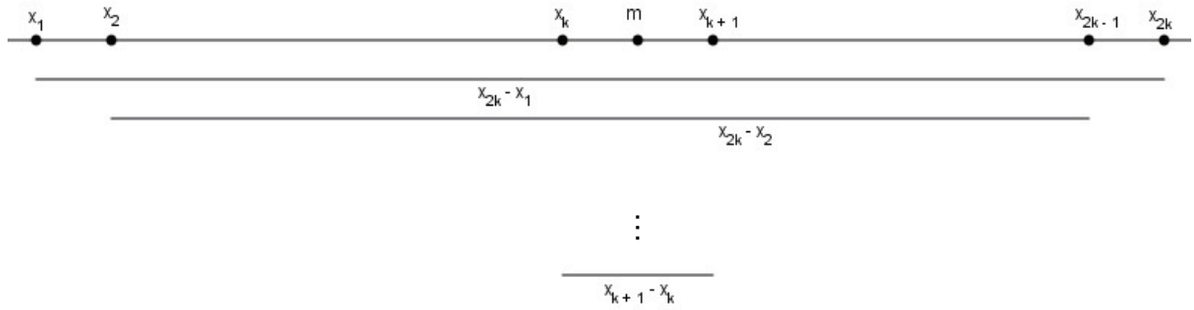
$$S = (|x_1 - c|+|x_{2k+1} - c|) + (|x_2 - c|+|x_{2k} - c|) + \ldots + (|x_k - c|+|x_{k+2} - c|) + |x_{k+1} - c|$$
$$= S' + |x_{k+1} - c|$$

Now S' will be minimized if $x_k < c < x_{k+2}$ from 2. above

And $|x_{k+1} - c|$ will be minimized if $c = x_{k+1}$

Since $x_k < x_{k+1} < x_{k+2}$, $c = x_{k+1}$ will minimize both S' and $|x_{k+1} - c|$ and therefore S, and that minimal value will be $(x_{2k+1} - x_1) + (x_{2k} - x_2) + \ldots + (x_{k+2} - x_k) = (x_{2k+1} + \ldots + x_{k+2}) - (x_1 + \ldots + x_k)$ i.e. the difference between the sums of the upper and lower halves of the data-values

For even $n = 2k$, $S = (|x_1 - c|+|x_{2k} - c|) + (|x_2 - c|+|x_{2k-1} - c|) + \ldots + (|x_k - c|+|x_{k+1} - c|)$ which will be minimized if $x_k < c < x_{k+1}$



In this case, median $m = 1/2(x_k + x_{k+1})$ and obviously $x_k < m < x_{k+1}$ and therefore median minimizes S and that minimal value will be $(x_{2k} - x_1) + (x_{2k-1} - x_2) + \ldots + (x_{k+1} - x_k) = (x_{2k} + \ldots + x_{k+1}) - (x_1 + \ldots + x_k)$ or the difference between the sums of the upper and lower halves of the data-values

Note that for even n, median is not the ONLY number that minimizes the mean deviation, but for odd n, it is.

In the next part, we will look into the relations between the ogives and the median…

MATHEMATICS CO-DEVELOPMENT GROUP

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.