Statistics

Monday, March 3, 2025 3:05 PM

median and mode for grouped data.

14.2 Mean of Ungrouped Data

We know that the mean (or average) of observations is the ratio of sum of the values of all the observations divided by the total number of observations. Let $x_1, x_2, \ldots, x_n$ be observations with respective frequencies $f_1, f_2, \ldots, f_n$. This means that observation x_1 occurs f_1 times, x_2 occurs f_2 times, and so on.

Now, the sum of the values of all the observations $= f_1x_1 + f_2x_2 + \ldots + f_nx_n$, and the number of observations $= f_1 + f_2 + \ldots + f_n$.

So, the mean $\bar{x}$ of the data is given by

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots\ldots + f_nx_n}{f_1 + f_2 + \ldots\ldots\ldots + f_n}$$

Recall that we can write this in short, using the Greek letter $\sum$ (read as sigma) which means summation i.e., $\bar{x} = \frac{\sum f_ix_i}{\sum f_i}$

Example-1. The marks obtained in mathematics by 30 students of Class X of a certain school are given in table the below. Find the mean of the marks obtained by the students.

Marks obtained (x_i)	10	20	36	40	50	56	60	70	72	80	88	92	95
Number of student (f_i)	1	1	3	4	3	2	4	4	1	1	2	3	1

Solution : Let us re-organize this data and find the sum of all observations.

Marks obtained (x_i)	Number of students (fi)	f_ix_i
10	1	10
20	1	20
36	3	108
40	4	160
50	3	150
56	2	112
60	4	240
70	4	280
72	1	72
80	1	80
88	2	176
92	3	276
95	1	95
Total	$\sum f_i = 30$	$\sum f_ix_i = 1779$

So, $\bar{x} = \frac{\sum f_ix_i}{\sum f_i} = \frac{1779}{30} = 59.3$

But irl data is just too large for us to do this, so we group it and allocate frequencies to each class interval.

- Frequency which is equal to any upper class boundary would be considered in next class.

Now, for each class-interval, we require a point which would serve as the representative of the whole class. ***It is assumed that the frequency of each class-interval is centred around its mid-point.*** So, the *mid-point* of each class can be chosen to represent the observations falling in that class and is called the class mark. Recall that we find the class mark by finding the average of the upper and lower limit of the class.

$$\text{Class mark} = \frac{\text{Upper class limit} + \text{Lower class limit}}{2}$$

For the class 10 -25, the class mark is $\frac{10+25}{2} = 17.5$.

Class interval	Number of students (f_i)	Class Marks (x_i)	f_ix_i
10-25	2	17.5	35.0
25-40	3	32.5	97.5
40-55	7	47.5	332.5
55-70	6	62.5	375.0

25-40	3	32.5	97.5
40-55	7	47.5	332.5
55-70	6	62.5	375.0
70-85	6	77.5	465.0
85-100	6	92.5	555.0
Total	$\sum f_i$=30		$\sum f_i x_i$=1860.0

This is called direct mean method.

Assumed mean method:
Step 1: consider assumed mean
Step 2: find deviations

Class interval	Number of students (f_i)	Class Marks (x_i)	$d_i = x_i - 47.5$ $d_i = x_i - a$	$f_i d_i$
10-25	2	17.5	-30	-60
25-40	3	32.5	-15	-45
40-55	7	47.5 (a)	0	0
55-70	6	62.5	15	90
70-85	6	77.5	30	180
85-100	6	92.5	45	270
Total	$\sum f_i$=30			$\sum f_i d_i$ = 435

So, from the above table, the mean of the deviations, $\bar{d} = \frac{\sum f_i d_i}{\sum f_i}$

Since, in obtaining d_i we subtracted 'a' from each x_i so, in order to get the mean $\bar{x}$ we need to add 'a' to $\bar{d}$. This can be explained mathematically as:

Mean of deviations, $\bar{d} = \frac{\sum f_i d_i}{\sum f_i}$

So, $\bar{d} = \frac{\sum f_i (x_i - a)}{\sum f_i}$

$= \frac{\sum f_i x_i}{\sum f_i} - \frac{\sum f_i a}{\sum f_i}$

$= \bar{x} - a\frac{\sum f_i}{\sum f_i}$

$\bar{d} = \bar{x} - a$

Therefore $\bar{x} = a + \frac{\sum f_i d_i}{\sum f_i}$

Substituting the values of a, $\sum f_i d_i$ and $\sum f_i$ from the table, we get

$$\bar{x} = 47.5 + \frac{435}{30} = 47.5 + 14.5 = 62$$

Therefore, the mean of the marks obtained by the students is 62.

The method discussed above is called the **Assumed Mean Method**.

Step deviation method:

Let $\bar{u} = \frac{\sum f_i u_i}{\sum f_i}$.

Observe that in the table given below the values in Column 4 are all multiples of 15. If we divide all the values of Column 4 by 15, we would get smaller numbers which we then multiply with f_i. (Here, 15 is the class size of each class interval.)

So, let $u_i = \frac{x_i - a}{h}$, where a is the assumed mean and h is the class size.

Class interval	Number of students (f_i)	Class Marks (x_i)	$d_i = x_i - a$	$u_i = \frac{x_i - a}{h}$	$f_i u_i$
10-25	2	17.5	-30	-2	-4
25-40	3	32.5	-15	-1	-3
40-55	7	47.5	0	0	0
55-70	6	62.5	15	1	6
70-85	6	77.5	30	2	12
85-100	6	92.5	45	3	18

55-70	6	62.5	15	1	6
70-85	6	77.5	30	2	12
85-100	6	92.5	45	3	18
Total	$\sum f_i = 30$				$\sum f_i u_i = 29$

Here again, let us find the relation between $\bar{u}$ and $\bar{x}$.

We have $u_i = \frac{x_i - a}{h}$

So $\bar{u} = \frac{\sum f_i u_i}{\sum f_i}$

$$\bar{u} = \frac{\sum f_i \frac{(x_i - a)}{h}}{\sum f_i}$$

$$= \frac{1}{h}\left[\frac{\sum f_i x_i}{\sum f_i} - \frac{\sum f_i a}{\sum f_i}\right]$$

$$= \frac{1}{h}(\bar{x} - a)$$

$$h\bar{u} = \bar{x} - a$$

$$\bar{x} = a + h\bar{u}$$

Therefore, $\bar{x} = a + h\left[\frac{\sum f_i u_i}{\sum f_i}\right]$.

or $\bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) \times h$

Substituting the values of a, $\sum f_i u_i$ h and $\sum f_i$ from the table, we get

$$\bar{x} = 47.5 + \frac{29}{30} \times 15$$

$$= 47.5 + 14.5 = 62$$

So, the mean marks obtained by a student are 62.

The method discussed above is called the **Step-deviation** method.

We note that:

- The step-deviation method will be convenient to apply if all the d_i's have a common factor.
- The mean obtained by all the three methods is same.
- The assumed mean method and step-deviation method are just simplified forms of the direct method.
- The formula $\bar{x} = a + h\bar{u}$ still holds if a and h are not as given above, but are any non-zero numbers such that $u_i = \frac{x_i - a}{h}$.

14.3 Mode

A mode is that value among the observations which occurs most frequently.

Before learning, how to calculate the mode of grouped data, let us first recall how we found the mode for ungrouped data through the following example.

Example-4. The wickets taken by a bowler in 10 cricket matches are as follows: 2, 6, 4, 5, 0, 2, 1, 3, 2, 3. Find the mode of the data.

Solution : Let us arrange the observations in order i.e., 0, 1, 2, 2, 2, 3, 3, 4, 5, 6

Clearly, 2 is the number of wickets taken by the bowler in the maximum number of matches (i.e., 3 times). So, the mode of this data is 2.

In a grouped frequency distribution, it is not possible to determine the mode by looking at the frequencies. Here, we can only locate a class with the maximum frequency, called the modal class. The mode is a value inside the modal class, and is given by the formula.

Mode $= l + \left(\frac{f_1 - f_0}{\quad}\right) \times h$

at the frequencies. Here, we can only locate a class with the maximum frequency, called the modal class. The mode is a value inside the modal class, and is given by the formula.

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

where, l = lower boundary of the modal class,

h = size of the modal class interval,

f_1 = frequency of the modal class,

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class.

Let us consider the following examples to illustrate the use of this formula.

Example-5. A survey conducted on 20 households in a locality by a group of students resulted in the following frequency table for the number of family members in a household.

Family size	1-3	3-5	5-7	7-9	9-11
Number of families	7	8	2	2	1

Find the mode of this data.

Solution : Here the maximum class frequency is 8, and the class corresponding to this frequency is 3-5. So, the modal class is 3-5.

Now,

modal class = 3-5, boundary limit (l) of modal class = 3, class size (h) = 2

frequency of the modal class (f_1) = 8,

frequency of class preceding the modal class (f_0) = 7,

frequency of class succeeding the modal class (f_2) = 2.

Now, let us substitute these values in the formula-

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$$= 3 + \left(\frac{8-7}{2\times 8 - 7 - 2}\right) \times 2 = 3 + \frac{2}{7} = 3.286$$

Therefore, the mode of the data above is 3.286.

Example-6. The marks distribution of 30 students in a mathematics examination are given in the adjacent table. Find the mode of this data. Also compare and interpret the mode and the mean.

Class interval	Number of students (f_i)	Class Marks (x_i)	$f_i x_i$
10-25	2	17.5	35.0
25-40	3	32.5	97.5
40-55	7	47.5	332.5
55-70	6	62.5	375.0
70-85	6	77.5	465.0
85-100	6	92.5	555.0
Total	$\sum f_i = 30$		$\sum f_i x_i = 1860.0$

Solution : Since the maximum number of students (i.e., 7) have got marks in the interval, 40-65 the modal class is 40 - 55.

The lower boundary (l) of the modal class = 40,

the class size (h) = 15,

the frequency of modal class (f_1) = 7,

the frequency of the class preceding the modal class (f_0) = 3 and

the frequency of the class succeeding the modal class (f_2) = 6.

Now, using the formula:

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$$= 40 + \left(\frac{7-3}{2\times 7 - 6 - 3}\right) \times 15 = 40 + 12 = 52$$

Interpretation : The mode marks is 52. Now, from Example 1, we know that the mean marks is 62. So, the maximum number of students obtained 52 marks, while on an average a student obtained 62 marks.

14.4 Median of Grouped Data

Median is a measure of central tendency which gives the value of the middle-most observation in the data. Recall that for finding the median of ungrouped data, we first arrange the data values or the observations in ascending order.

Then, if n is odd, the median is the $\left(\frac{n+1}{2}\right)^{th}$ observation and

if n is even, then the median will be the average of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations.

Suppose, we have to find the median of the following data, which is about the marks, out of 50 obtained by 100 students in a test :

Marks obtained	20	29	28	33	42	38	43	25
Number of students	6	28	24	15	2	4	1	20

First, we arrange the marks in ascending order and prepare a frequency table as follows :

Marks obtained	Number of students (frequency)
20	6
25	20
28	24
29	28
33	15
38	4
42	2
43	1
Total	100

Here $n = 100$, which is even. The median will be the average of the $\left(\frac{n}{2}\right)$ and the $\left(\frac{n}{2}+1\right)$ observations, i.e., the 50^{th} and 51^{st} observations. To find the position of these middle values, we construct cumulative frequency.

Marks obtained	Number of students	Cumulative frequency
20	6	6
upto 25	6 + 20 = 26	26
upto 28	26 + 24 = 50	50
upto 29	50 + 28 = 78	78
upto 33	78 + 15 = 93	93
upto 38	93 + 4 = 97	97
upto 42	97 + 2 = 99	99
upto 43	99 + 1 = 100	100

Now we add another column depicting this information to the frequency table above and name it as *cumulative frequency column*.

From the table above, we see that :

50^{th} observation is 28 (Why?)

51^{st} observation is 29

$$\text{Median} = \frac{28+29}{2} = 28.5 \text{ marks}$$

Remark : The above table is known as *Cumulative Frequency Table*. The median marks 28.5 conveys the information that about 50% students obtained marks less than 28.5 and another 50% students obtained marks more than 28.5.

Consider a grouped frequency distribution of marks obtained, out of 100, by 53 students, in a certain examination, as shown in adjacent table.

Marks	Number of students
0-10	5
10-20	3
20-30	4
30-40	3
40-50	3
50-60	4
60-70	7
70-80	9
80-90	7
90-100	8

Now in a grouped data, we may not be able to find the middle observation by looking at the cumulative frequencies as the middle observation will be some value in a class interval. It is, therefore, necessary to find the value inside a class that divides the whole distribution into two halves. But which class should this be?

To find this class, we find the cumulative frequencies of all the classes and $\frac{n}{2}$. We now locate the class whose cumulative frequency exceeds $\frac{n}{2}$ for the first time. This is called the median class.

Marks	Number of students (f)	Cumulative frequency (cf)
0-10	5	5
10-20	3	8
20-30	4	12
30-40	3	15
40-50	3	18
50-60	4	22
60-70	7	29
70-80	9	38
80-90	7	45
90-100	8	53

In the distribution above, $n = 53$. So $\frac{n}{2} = 26.5$. Now 60-70 is the class whose cumulative frequency 29 is greater than (and nearest to) $\frac{n}{2}$, i.e., 26.5.

Therefore, 60-70 is the median class.

After finding the median class, we use the following formula for calculating the median.

$$\text{Median} = l + \left[\frac{\frac{n}{2} - cf}{f}\right] \times h$$

where l = lower boundary of median clas,

n = number of observations,

cf = cumulative frequency of class preceding the median class,

f = frequency of median class,

h = class size (size of the median class).

Substituting the values $\frac{n}{2} = 26.5$, $l = 60$, $cf = 22$, $f = 7$, $h = 10$

in the formula above, we get

$$\text{Median} = 60 + \left[\frac{26.5 - 22}{7}\right] \times 10$$

$$= 60 + \frac{45}{7}$$

$$= 66.4$$

So, about half the students have scored marks less than 66.4, and the other half have scored marks more than 66.4.

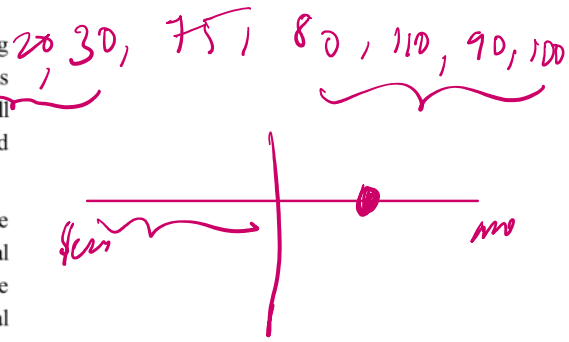
14.5 Which value of Central Tendency

Which measure would be best suited for a particular requirement.

The mean is the most frequently used measure of central tendency because it takes into account all the observations, and lies between the extremes, i.e., the largest and the smallest observations of the entire data. It also enables us to compare two or more distributions. For example, by comparing the average (mean) results of students of different schools of a particular examination, we can conclude which school has a better performance.

However, extreme values in the data affect the mean. For example, the mean of classes having frequencies more or less the same is a good representative of the data. But, if one class has frequency, say 2, and the five others have frequency 20, 25, 20, 21, 18, then the mean will certainly not reflect the way the data behaves. So, in such cases, the mean is not a good representative of the data.

In problems where individual observations are not important, especially extreme values, and we wish to find out a 'typical' observation, the median is more appropriate, e.g., finding the typical productivity rate of workers, average wage in a country, etc. These are situations where extreme values may exist. So, rather than the mean, we take the median as a better measure of central tendency.



Example-2. The table below gives the percentage distribution of female teachers in the primary schools of rural areas of various states and union territories (U.T.) of India. Find the mean percentage of female teachers using all the three methods.

Percentage of female teachers	15 - 25	25 - 35	35 – 45	45 - 55	55 - 65	65 - 75	75 – 85
Number of States/U.T.	6	11	7	4	4	2	1

Source : *Seventh All India School Education Survey conducted by NCERT*

$x_i = \frac{\text{lower} + \text{upper}}{2}$

Class	f_i	x_i	$d_i = x_i - a$	$u_i = \frac{x_i - a}{h}$	$f_i x_i$	$f_i d_i$	$f_i u_i$
15 – 25	6	20	−30	−3	120	−180	−18
25 – 35	11	30	−20	−2	330	−220	−22
35 – 45	7	40	−10	−1	280	−70	−7
45 – 55	4	50 = a	0	0	200	0	0
55 – 65	4	60	10	1	240	40	4
65 – 75	2	70	20	2	140	40	4
75 – 85	1	80	30	3	80	30	3
	35				1390	−360	36

$$\text{direct} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{1390}{35} = 39.71$$

$$\text{Assumed mean} = \bar{x} = a + \frac{\Sigma f_i d_i}{\Sigma f_i} = 50 + \frac{-360}{35} = 39.71$$

Step deviation $\Rightarrow \quad \bar{x} = a + \left(\frac{\sum f_i u_i}{\sum f_i}\right) h = 50 + \left(\frac{-36}{35}\right)(10)$

$= 39.71$