Statistics
Part 2

Aviral Janveja

2022

In this part, we shall extend our study of the three measures of central tendency, namely mean, median and mode from ungrouped data to grouped data.

1 Mean of Grouped Data

Mean, as we know is the sum of all observations divided by the total number of observations. Let us look at an example of marks obtained out of 100 by 30 students of class 10 :

Marks Obtained	**Number of Students**	**Product**
10	1	10
20	1	20
36	3	108
40	4	160
50	3	150
56	2	112
60	4	240
70	4	210
72	1	72
80	1	80
88	2	176
92	3	276
95	1	95
Total	30	1779

Table 1: Ungrouped Frequency Distribution Table

Therefore, the mean mark obtained is :

$$\bar{x} = \frac{1779}{30} = 59.3$$

However, in most real life situations, data is usually so large that to make a meaningful study, it needs to be condensed into grouped data. Let us therefore convert the above ungrouped frequency distribution table into grouped one by forming class-intervals of width say 15.
Remember while allocating frequencies to each class interval, students having

marks equivalent to any of the upper class-limits, would be counted in the next higher class. For example, a student who has scored 40 marks will be counted in the class-interval 40-55 and not 25-40.
With this convention in mind, let us form a grouped frequency distribution table :

Class Interval	**Number of Students**
10-25	2
25-40	3
40-55	7
55-70	6
70-85	6
85-100	6

Table 2: Grouped Frequency Distribution Table

Having converted the the above ungrouped data into grouped data, we need to now devise some method to find its mean.
Firstly, to calculate the sum of marks obtained, we need a representative mark for each class. It is assumed that the frequency of each class is centered around its mid-point. This mid-point, called the **class mark** can be chosen to represent the observations falling in that class.
The class mark is calculated as follows :

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

In the above table for example, for the class 10-25, the class mark is $\frac{10+25}{2} = 17.5$
We can now proceed to calculate the mean in the same way as before :

Class Interval	**Number of Students(f_i)**	**Class mark (x_i)**	$f_i x_i$
10-25	2	17.5	35
25-40	3	32.5	97.5
40-55	7	47.5	332.5
55-70	6	62.5	375
70-85	6	77.5	465
85-100	6	92.5	555
Total	$\Sigma f_i = 30$		$\Sigma f_i x_i = 1860$

Table 3: Final Grouped Frequency Distribution Table

Multiplying the number of students with the class mark gives us the sum of observations as shown in the above table. The mean is thus given by :

$$\bar{x} = \frac{\Sigma f_i x_i}{\Sigma f_i} = \frac{1860}{30} = 62$$

This method for finding mean of grouped data is known as the **Direct Method**.
We **observe** that, even though we are using the same data as before, the mean obtained is different this time. The difference in the two values is due to the class mark assumption, with 59.3 being the exact mean, while 62 being the approximate mean.

2 Mode of Grouped Data

Recall from Part 1, that mode is the observation that occurs most often. Here, we shall discuss ways of obtaining mode for grouped data.
In a grouped frequency distribution, it is not possible to determine the mode directly by looking at the frequencies. We can only locate the class with the maximum frequency, called the **modal class**. The mode is then 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 limit of modal class.
h = size of the class-interval.
f_0 = frequency of the class preceding the modal class.
f_1 = frequency of the modal class.
f_2 = frequency of the class succeeding the modal class.

2.1 Example

A survey conducted regarding family size, on 20 households, by a group of students resulted in the following data :

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

Table 4: Survey Data

Find the mode of this data.
Solution : It is clear from the data that the modal class is 3-5 with frequency 8.
Therefore as per the formula for mode above, $l = 3$, $h = 2$, $f_1 = 8$, $f_0 = 7$, $f_2 = 2$.
Substituting these values in the formula, we get mode of the above data :

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

3 Median for Grouped Data

In part 1, we have learned that the median is a measure of central tendency, which gives the middle most observation of the data, hence dividing the data into two equal parts.

We have looked into how to obtain median for ungrouped data in part 1. Here, we will see how to obtain median for grouped data through the following example.

3.1 Example

Consider the following grouped frequency distribution of marks obtained out of 100, by 53 students in a certain examination :

Marks	**Number of Students(f)**
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
Total	53

Table 5: Grouped Frequency Distribution

For ungrouped data, we began by first arranging the marks in ascending order. In the above table, the class intervals are already arranged in ascending order.
Further, To aid our calculation process, we make use of **cumulative frequencies** alongside normal frequency.
In the table above, we see that 5 students fall in the interval 0-10 and 3 students fall in the interval 10-20, therefore a total of 8 students fall in the interval 0-20. Hence, we say that the cumulative frequency up to class 10-20 is 8. Calculating the cumulative frequencies for all class intervals in the same manner, we get the following table :

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

Table 6: Cumulative Frequency Table

Now, as it is grouped data, we are not able to directly locate that middle most observation, as it is some value within the middle class-interval.
To find this middle class-interval, we compute $n/2$, then locate the class-interval whose cumulative frequency is greater than and nearest to $n/2$. Having computed cumulative frequencies beforehand, makes it easy to readily locate that middle class-interval which is also called the **median class**.

In the above distribution, $n = 53$ thus $n/2 = 26.5$. Therefore, the class-interval 60-70 with cumulative frequency = 29, is the median class.
After finding the median class, we use the following formula to find the median value within the median class :

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

where,
l = Lower limit of median class.
n = Total number of observations.
cf = Cumulative frequency of class preceding the median class.
f = Frequency of median class.
h = Class size.

Substituting the values in the above formula : $n/2 = 26.5$, $l = 60$, $cf = 22$, $f = 7$, $h = 10$, we get :

$$\text{Median} = 60 + \left(\frac{26.5 - 22}{7} \right) \times 10 = 66.4$$

This means about half the students have scored less than 66.4 and other half have scored more than 66.4

4 Which measure would be best suited for a given situation ?

Having studied about all the three measures of central tendency, let us discuss which measure would be best suited for a particular requirement.

1. Mean
The mean takes the average of all observations, thereby enabling us to compare two or more distributions. For example, by comparing the average (mean) results of students of different schools for a particular examination, we can conclude which school has a better performance.
However, extreme values in the data affect the mean. For example, if one student has scored, say 15 out of 100, and the other five have scored 80, 100, 93, 96, 90, then the mean will certainly not reflect the true nature of the data. So, in such cases, the mean is not a good representative of the data.

2. Median
In such situations, where each individual observation need not be considered, and we instead wish to find out a typical observation, the median is more appropriate. For example, finding the typical productivity rate of workers, average wage in a country, etc. These are example where extreme values may be present. So, rather than the mean, we take median as a measure of central tendency.

3. Mode
In situations which require establishing the most frequent value or the most popular item, the mode is the best choice. For example, to find the most popular TV Programme or the vehicle model in greatest demand, etc.

Remark : There is an empirical relationship between the three measures of central tendency : **Mode = 3 Median - 2 Mode**.

5 References

1. Class 10 - Chapter 14 : Statistics.
   NCERT Mathematics Textbook, Version 2020-21.
   As per Indian National Curriculum Framework 2005.