Statistics
Part 1

Aviral Janveja

2022

Everyday we come across a lot of information in the form of **facts, figures, tables & graphs**. These facts or figures, which may be numerical or otherwise, when collected with a definite purpose are called **data**.
Every part of our lives utilizes data in one form or another. So it becomes essential for us to know, how to extract meaningful **information** from such data.
This extraction of meaningful information, is studied in the branch of mathematics called **statistics**. Statistics deals with collection, organization, presentation, analysis and interpretation of data.

1 Collection of Data

1.1 Primary Data

The data, when collected by the investigators themselves for a defined objective, is called primary data.

1.2 Secondary Data

However, when gathered from another source which already had the data collected beforehand, then it is called secondary data.
Such data, which has been collected by someone else, possibly in a different context, needs to be used with great care, ensuring that the source is reliable and relevant towards our objective.

2 Presentation of Data

As soon as the work related to collection of data is over, the investigator has to find ways to present it in a form which is meaningful, easily understood and gives its relevant features at a glance.

2.1 Frequency Distribution Table

One of the ways to make data more easily understandable, is to write it in a tabular form.

For Example,
Consider the marks obtained out of 100 by 30 students of class IX of a school :

$$10, 20, 36, 92, 95, 40, 50, 56, 60, 70$$
$$92, 88, 80, 70, 72, 70, 36, 40, 36, 40$$
$$92, 40, 50, 50, 56, 60, 70, 60, 60, 88$$

Data in the form written above is called **raw data**. To make it more easily understandable, we write it in a table as show below :

Marks	**Number of Students (Frequency)**
10	1
20	1
36	3
40	4
50	3
56	2
60	4
70	4
72	1
80	1
88	2
92	3
95	1
Total	30

The above table is called an ungrouped frequency distribution table or simply a frequency distribution table. Where, the number of students who have obtained a certain number for marks is called the **frequency** of those marks.

2.2 Grouped Frequency Distribution Table

In the next example,
We see a tree plantation drive where 100 saplings each, were planted in 100 schools. After one month, the number of plants that survived were recorded as follows :

$$95, 67, 28, 32, 65, 65, 69, 33, 98, 96$$
$$76, 42, 32, 38, 42, 40, 40, 69, 95, 92$$
$$75, 83, 76, 83, 85, 62, 37, 65, 63, 42$$
$$89, 65, 73, 81, 49, 52, 64, 76, 83, 92$$
$$93, 68, 52, 79, 81, 83, 59, 82, 75, 82$$
$$86, 90, 44, 62, 31, 36, 38, 42, 39, 83$$
$$87, 56, 58, 23, 35, 76, 83, 85, 30, 68$$
$$69, 83, 86, 43, 45, 39, 83, 75, 66, 83$$
$$92, 75, 89, 66, 91, 27, 88, 89, 93, 42$$
$$53, 69, 90, 55, 66, 49, 52, 83, 34, 36$$

To make sense of such a large amount of raw data, we condense the number of surviving plants into groups like 20-29, 30-39 ... 90-99 since our data ranges from

23 to 98. These groups are called **classes** and their size is called **class-size** which is 10 in this case.
Thereby, the data above can be condensed into tabular form as follows :

Number of Surviving Plants	Number of Schools (Frequency)
20-29	3
30-39	14
40-49	12
50-59	8
60-69	18
70-79	10
80-89	23
90-99	12
Total	30

The above is called a grouped frequency distribution table. Presenting data in this form simplifies and condenses data and enables us to observe certain important features at a glance.
For instance, in the above table we can easily observe that 50 or more plants survived in 8+18+10+23+12 = 71 schools.
Note : You can chose the class-size as per the range and type of data. There is no hard and fast rule about that, just that the classes should not overlap.

3 Graphical Representation of Data

It is well said that "a picture is worth a thousand words". Bar Graphs, Histograms and Frequency Polygons are some basic types of graphs which are used to represent data. Various other graphical and visual representations are possible as per the given context and situation.
They are especially useful when comparing two different sets of data. For example, comparing the performance of two different classes in a school.

4 Measures of Central Tendency

After looking at collection and presentation of data in previous sections. Let us now look at how to efficiently analyse data.
The question is - Do we always need to study the entire data to make sense of it, or can we find out some important features by looking at certain representatives of the data ?
This is possible, by using the **measures of central tendency**. A measure of central tendency is a single value that attempts to represent a set of data by identifying the central position within that data-set. The most common measures of central tendency are the **mean**, the **median** and the **mode**.
In order to understand this better, let us take an **example** of two students Shreyas and Siddharth. They obtained the following marks out of 10 across 5 tests :

Shreyas	Siddharth
7	4
8	7
8	10
9	10
10	10

Who had a better performance of the two ? We will try and answer that by using the three measures of central tendency.

4.1 Mean

The mean of a number of observations is the sum of all observations divided by the number of observations. It is denoted by $\bar{x}$, read as "x bar".

$$\bar{x} = \frac{\text{Sum of Observations}}{\text{Number of Observations}}$$

Hence, Shreyas' mean is :

$$\frac{7+8+8+9+10}{5} = 8.4$$

Similarly, Siddharth's mean :

$$\frac{4+7+10+10+10}{5} = 8.2$$

The mean seems to indicate that Shreyas performed better as he has a higher mean score. However, let us look at what the other two measures have to say.

4.2 Median

The median is the middle most observation of the data, which divides the data into exactly two parts. We calculate the median as follows :

- First the data is arranged in ascending (or descending) order.
- If the number of observations is odd, then $(\frac{n+1}{2})^{th}$ observation is the median.
- If the number of observations is even, then median is the mean of $(\frac{n}{2})^{th}$ and $(\frac{n}{2}+1)^{th}$ observations.

Therefore, the median mark for Shreyas is 8, whereas the median for Siddharth is 10. The median suggests that Siddharth's performance should be rated better.

4.3 Mode

The mode is that observation which occurs most frequently.
In Shreyas' case the mode would therefore be 8. Whereas, Siddharth's mode is 10.
Looking at all the 3 measures for Shreyas and Siddharth together :

Measures of central tendency	Shreyas	Siddharth
Mean	8.4	8.2
Median	8	10
Mode	8	10

We see that these measures of central tendency are not sufficient for concluding which student is better. We require some more information to conclude this, which will be studied in further chapters on statistics.

5 Solved Exercises

Question : Consider a small factory of 5 employees. One Supervisor and four labourers. The labourers draw a salary of 5000 rupees per month while the supervisor gets 15000 per month. Calculate the mean, median and mode of the salaries.
Answer : Mean is :

$$\frac{5000 + 5000 + 5000 + 5000 + 15000}{5} = 7000$$

The median salary is 5000, the mode is 5000 as well as it occurs the maximum number of times.
It is interesting to observe here that the mean salary doesn't give an accurate estimate of any of their salaries, while the median and mode represent the data more effectively. **Note** that extreme values in the data affect the mean. This is one of the weaknesses of the mean. Median and mode give a better estimate of the average in such situations.

6 References

1. Class 9 - Chapter 14 : Statistics.
   NCERT Mathematics Textbook, Version 2020-21.
   As per Indian National Curriculum Framework 2005.