7 STATISTICS

Mark and grade

The marks that 45 students of a class got in the math exam are given below:

46	99	92	48	53	49	84	45	73
46	56	40	50	35	55	83	59	69
15	73	30	55	26	54	74	60	74
45	54	35	79	40	65	64	23	69
64	89	68	55	59	66	49	37	38

These must be converted to grades and the number of students coming under each grade must be found out.

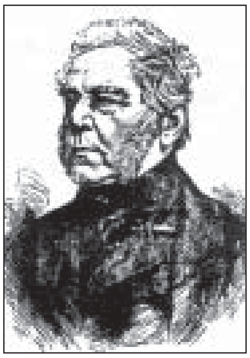
Let's first write down the range of marks each grade specifies:

Grade	Mark
A+	90 - 100
A	80 - 89
B+	70 - 79
B	60 - 69
C+	50 - 59
C	40 - 49
D+	30 - 39
D	20 - 29
E	1 - 19

Next, let's extend the table by putting tallies against each grade. For this, we look at each mark in the first table and put a tally against the grade in which it falls. (Do you remember the section **Tabulation** of the lesson **Statistics** in the Class 8 textbook?)

A bit of history

We have noted that statistics is the study of information presented as numbers. Such studies were started by John Graunt, who lived in England during the seventeenth century.

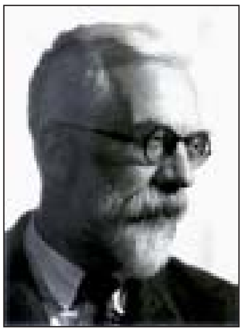


He made his living by selling small articles for sewing, such as buttons and ribbons. In 1663, he published an extensive study of public health in England and the way epidemics like plague spreads among the population. Impressed by this, the king Charles II made him a member of the Royal Society, a prestigious association of learned men in England. Many members of the society did not like the idea of a mere tradesman being included among them, but the king ignored their objections. His last days were spent in abject poverty, without any job.

Statistics

The word statistics is used in two senses: either to denote numerical data or the scientific study of such data. Though this science makes use of mathematical techniques and computations, it arises from, and is applied to, practical situations. Because of this, statistics is considered an independent science, rather than a branch of mathematics.

The fundamental principles of statistics were developed by the biologist, Ronald Fisher who lived in England during the last century.



His studies in statistics led to the synthesis of Darwin's theory of evolution and modern genetics.

Grade	Marks	Students	
A+	90 - 100	||	
A	80 - 89	|||	
B+	70 - 79	卌	
B	60 - 69		
C+	50 - 59		
C	40 - 49		
D+	30 - 39		
D	20 - 29		
E	1 - 19		
Total			

Did you complete the table? Then here are some questions:

- How many got A+?
- How many got E?
- Which grade did most get?
- Which grade did the least number of students get?

Can we form an idea regarding the math learning level of this class?

Try to tabulate the marks of the students in your class for any subject half yearly examination.

Table of Income

From the data collected from fifty persons in a locality, their daily incomes are as given below:

168	185	148	175	165	157	174	160	175	145
198	162	170	145	175	162	170	190	188	158
174	195	165	176	150	171	175	179	154	170
163	192	155	175	163	180	167	188	178	165
186	152	175	185	152	160	176	173	172	168

Do you remember tabulating a similar collection of figures in Class 8?

The above table has numbers from 145 to 198; but not all numbers between these are in it.

So, if we list all numbers from 145 to 198 and write against each, the number of people with that number as daily income, then the resulting table does not help to form any general idea about the incomes.

Instead, let’s split the income into various classes such as 145–155, 155–165, . . . , 195–205, as in the case of marks in the first example. (Here we take the last class as 195–205, even though there are no incomes above 198; this is to make the income difference of all classes equal, namely 10 rupees.)

Daily Income	Number of People
145 -155	||
155 - 165	||
165 - 175	|||
175 - 185	||
185 - 195	|
195 - 205	

Here, in which class do we include the number 185? In 175 – 185 or 185–195?

Usually, in tables of this kind, the last number of each class (except the last class) is not included in that class.

So, 185 is included in the class 185–195. Now can’t you complete the table? We can condense it as shown below:

Daily Income	Number of People
145 -155	7
155 - 165	9
165 - 175	14
175 - 185	11
185 - 195	7
195 - 205	2

On tables

To draw conclusions from a collection of data, we have to first put them in order. One method of such an arrangement is to classify them and form a table. A type of table used in statistics is the frequency table.

When we tabulate data like this, some information is lost. For example, when the entire data collected on incomes is presented as income groups and the number of people belonging to each group, we cannot find the actual income of each person from it.

But from such a table, we can get a general idea of how the various incomes are distributed among the people. Such a general view cannot be readily gained from the entire unorganised collection of data.

Classification methods

We classify and tabulate data for a concise presentation, from which it is easy to draw general conclusions. We have noted that some information is lost when we do this. This loss can be reduced by forming a large number of classes with small widths. But then the table will not be concise. On the other hand, if we form few classes of large widths, the presentation would be compact, but the loss of information would be so great that no valid inferences could be drawn.

For example, in our example on incomes, suppose we divide the incomes into classes of width 1 rupee. All the collected information would be in the table; but there is no condensation of data. At the other extreme, if we consider the entire range of incomes as a single class, from the lowest income to the highest, then we have maximum condensation; but no general conclusions can be drawn from it.

From this table, we can get much information on the general nature of incomes of these fifty people. For example,

- Most people have income between 175 rupees and 185 rupees.
- 50% of these people have income between 165 rupees and 185 rupees.
- Only 4% of these people have income above 200 rupees

Can't you draw some more conclusions like these?

As mentioned in Class 8, the general name for the number of entries in each class in a table of this sort is called the *frequency* of that class. (The word "frequency" usually refers to repetition.)

In the table just made, the frequency of the class 145–155 is 7 and the frequency of the class 155–165 is 9.

Again, in this table, the difference of incomes in each class is 10 rupees. It is called the *class width*, in general. We can also classify our original data on incomes as a table with each class width equal to 5 rupees instead of 10 rupees. (Try it!)

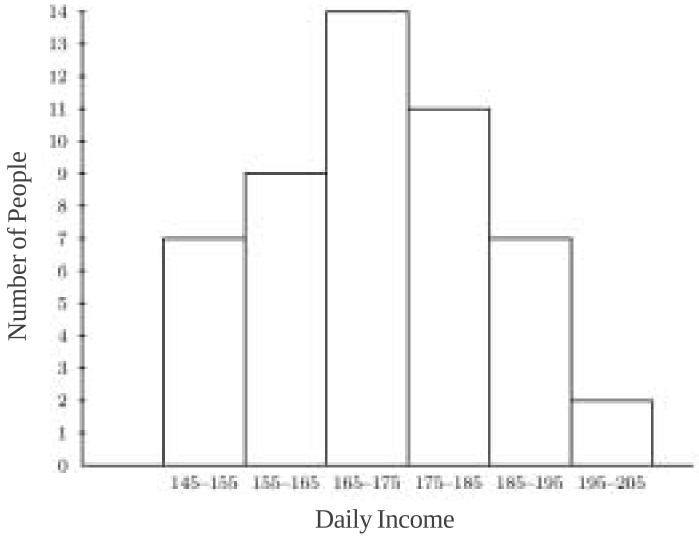
Now look at this problem: the heights of 40 people, in centimetres, are given below. Make a frequency table of this data.

175	143	177	163	170	168	156	168	145	160
159	154	148	166	168	150	164	160	176	163
140	170	165	168	169	174	158	148	155	165
162	167	171	165	171	167	159	171	167	180

Data in pictures

We have seen how we can use pictures, just as tables, to form general ideas about numerical data.

One way to picture a frequency table is to draw rectangles with heights proportional to the frequency of each class. For example, see how this is done for our data on the incomes of fifty people:



In this picture, each rectangle represents the daily income of a class, the heights of the rectangles being proportional to the frequencies. Here we take 0.5 centimetres for each occurrence in a class so that the first rectangle has height 3.5 centimetres, the second has height 4.5 centimetres and so on.

Such a picture is called a *histogram*.

Didn't you form a frequency table of heights? Now draw a histogram for it also.

Look at this table. It classifies the days of the months of June, July and August according to the rainfall received in a locality.

Rain (mm)	**Days**
10 - 20	8
20 - 30	10
30 - 40	14
40 - 50	20
50 - 60	15
60 - 70	8
70 - 80	7
80 - 90	6
90 - 100	4

Draw a histogram for this data.

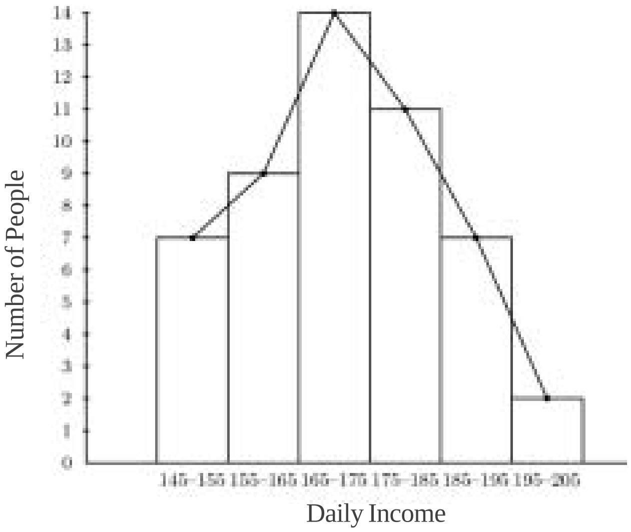
Pictures of data

Pictures convey information more easily than tables. We have seen how data can be represented as pictures in Class 7 itself.

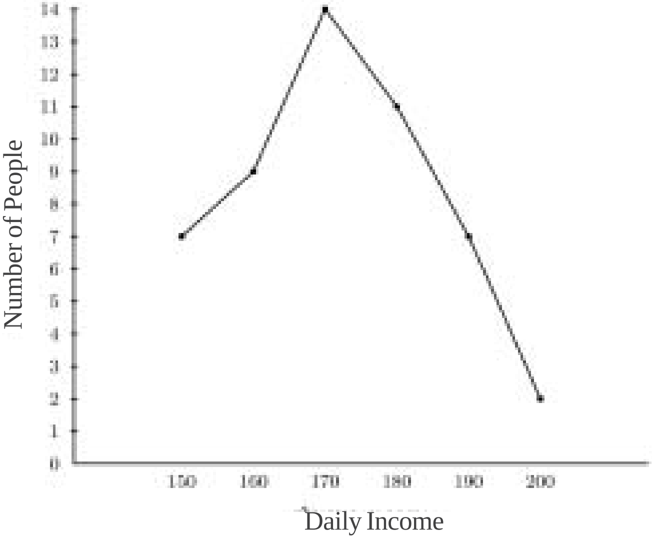
The information contained in a frequency table can be more easily seen in its various pictorial representations. Also, the rise and fall of frequencies across the classes can be readily perceived from the ups and downs in the picture.

Another picture

There is another way to picture a frequency table. For example, let's again take the histogram of the incomes. Mark the midpoints of the top of each rectangle and connect these points using straight lines.



We can do this without drawing the histogram first.



Here we represent each class by its mid-value.

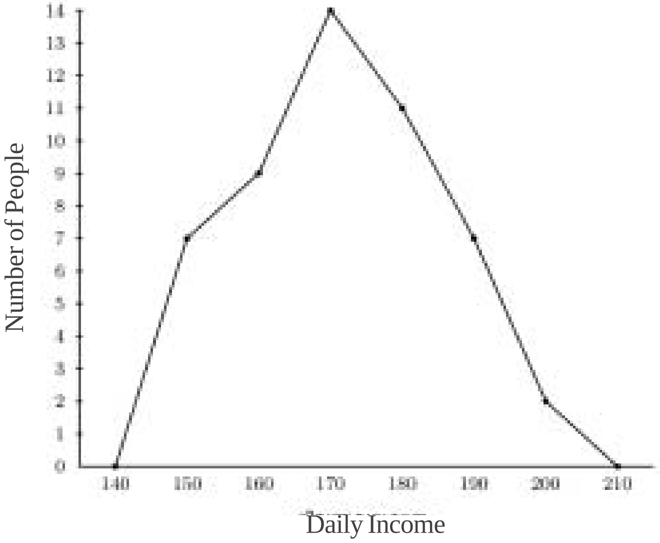
Usually in such pictures, the two ends are joined to the base line as shown below:

Doing it with computers

When the data collected is very large, we can use computers to tabulate and classify them. We can use any spreadsheet program to do this. In GNU/ Linux, some such programs are Open Office Calc and Gnumeric. The Help feature of these programs will show how these can be done. The web-page

www.wikieducator.org/ MathGloss/H/ Histogram

gives information on how to draw histograms with Open Office Calc



Thus, a new class 135–145, before the first one of the table and another new class 205–215 after the last class of the table are introduced with frequency 0 in each.

A picture like this is called a *frequency polygon.*

Now draw frequency polygon of the various data considered earlier, such as heights of 40 people and the rainfall during June to August.

Average

We have studied averages in Class 6. What is the average marks in math exam, in our first example?

How do we compute it?

We sum up the marks of all the 45 students and divide it by 45. This gives 56.8. Since marks are given as natural numbers, we take it as 57. (Why not 56?)

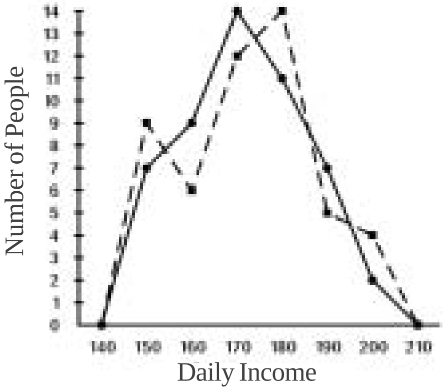
Now suppose instead of giving the marks of all the students as in our example, it is only said that the average marks of the class is 57. Can we form any idea about the learning level of that class? We can form some guesses such as

- most students would have got marks between 50 and 60 (that is C+ grade)
- Not many would have got A+ or E

Comparison

One advantage of a frequency polygon, over a histogram is that we can present two different data in the same picture, so that comparisons become easier.

For example, the picture below shows the frequency polygons of the incomes of two groups of 50 people from different localities.



From this we can make a comparison of the income distribution in the two groups, apart from some conclusion about the income distribution in each group separately.

Can you add any other observation?

Suppose we are told that the average marks in math in another class is 75. What all things can we reasonably guess about its level?

Can you compare it with the first class?

Remember our data on the incomes on 50 people. Compute the average daily income. Discuss what all conclusions we can draw from this number alone; also, what more conclusions we can draw from the table.

Not so correct average

The monthly incomes of ten families in a neighbourhood are as follows:

12000	11500	10500	10800	11000
10900	11200	10750	10250	10800

What is the average monthly income?

A new family moved into this neighbourhood. Their monthly income is 120000 rupees. What is the average now?

$$\frac{109700+120000}{11}\approx 20882.$$

Instead of giving all the details above, if we are told only that the average monthly income of a family in this neighbourhood is 20882 rupees, we would get the impression that most of these families have a monthly income of about 20000 rupees; and this figure is almost double the monthly income of most of them.

Can you think of any more situations like this, where wrong impressions are formed by considering the average alone?

Meaning of average

We have seen in Class 6, how some general ideas about quantities can be formed using averages. For example, when we say that a cow gives an average of five litres of milk per day, it does not mean that it gives exactly five litres each day. This may vary from day to day, but we can assume that on most days, it gives not much more or less than five litres.

Similarly, if the average marks in math exam in a class is 57, generally it means most students got marks near 57.

Another number

The average we usually compute is an attempt to represent by a single number, the information given by a collection of numbers. There are other methods of doing this. For example, let’s again look at the problem we have just done. We can write the monthly incomes of the 11 families in order, starting from the least, as follows:

10250 10500 10750 10800 10800 10900

11000 11200 11500 12000 120000

Which is the number in the middle?

The number got like this is called the *median* of these numbers.

In this problem itself, what is the median of the monthly incomes of the 10 families considered first?

If these numbers are written in order as before, there would be two numbers in the middle, right?

10250 10500 10750 10800

10800 10900

11000 11200 11500 12000

The median in this case is taken as the average of these two numbers. That is, the median of these monthly incomes is

$$\frac{10800+10900}{2} = \text{Rs.}10850$$

This is not much different from the average, 10970 rupees, computed earlier, is it?

In statistics, the average we usually compute is called the *arithmetic mean* (or simply the *mean*).

Now compute the mean and median of the various data considered earlier, such as the marks of 45 students, daily incomes of the 50 people, heights of 40 people and so on.

Math in context

The method we use to compute the average may not be suitable to the context. For example, if in a locality were most people have low incomes, there are one or two people with very high incomes, then the average computed would be much higher than the actual income of most of the people there.

In such situations, instead of the usual average, we have to use other methods more suited to the context, to compute numbers which will show the general nature of the data more faithfully. In short, theory formed should be suitable for the application planned.

Third number

Look again at the marks in the math exam given at the beginning of this lesson. Didn't you write them in order to compute the median? This shows that some marks are repeated. Thus we see that there are two students who got 35 and the number of students who got 40 is also two. But there are three who got 55. No other mark is repeated three or more times.

The mark 55, which is repeated most is called the *mode* of this data.

Like the mean and median, the mode may not always be a single number. Thus, in the above example, had one more student got 40 marks, then 40 would also have been another mode.

On the other hand, if no number is repeated, then the data does not have a mode.

Let's look at another example. The distances (in metres) that an athlete jumped during a training session for long jump are as follows:

6.10, 6.20, 6.20, 6.18, 6.20, 6.25, 6.21, 6.15, 6.10.

What is the mean of these? About 6.18 metres, right?

What does this number show? The distances he jumped are somewhere around 6.18 metres.

What about the median? 6.20 metres, isn't it?

What is the meaning of this? Half of his jumps were 6.20 metres or more. (And of course half of his jumps were 6.20 metres or less).

Now what about the mode? It is also 6.20 metres. And its meaning? The distance he jumped most number of times is 6.20 metres.

- The sizes of shirts sold in a shop are as follows:

  36, 38, 30, 40, 34, 42, 30, 46, 40, 38, 34, 40, 38, 40, 42, 30, 40, 38, 44, 40, 32, 40, 32, 44, 38

Compute the mean, median and mode. Among these, which is the most important one for the seller? Discuss.

Ways of statistics

We have seen how pictures and tables can be used to present data concisely. When the data is very large, even these techniques may not be effective. So, statistics aims to represent the general features of the data by means of a few numbers. Also, the using such numbers, more studies on data can be done

In the words of Ronald Fisher

> *The inherent inability of the human mind to grasp in its entirety, a whole mass of numerical data, compels us to seek relatively few constants, which will adequately describe the data.*