Statistics

Introduction

Statistics is a branch of Mathematics which deals with collection, organisation, analysis and interpretation of data.

Statistics deals mainly in the communication and analysis of facts and figures using statistical methods. Collection, classification, tabulation, representation, reasoning, testing and drawing inferences are parts of the statistical method. Graphs, tables, reasoning, estimation and prediction are the means of stastical methods.

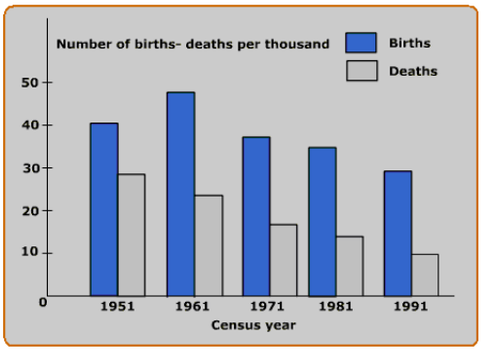
DR.P.K. Bose and Prof.C.R. Rao are eminent Indian statisticians.

Statistics helps in making predictions and estimates. Rainfall patterns of a particular city over a period of time can be analyzed and a fair estimate about next season can be arrived at, with the help of figures (data) collected over a period of time.

The word statistics is used with two meanings.

- Collecting data systematically and presenting numerical data
- Processing the numerical data and drawing conclusions

The following figure shows some information about population. Observe the figure and try to answer the questions below it.



- What information about population does the figure represent?
- State the time interval in which the information has been collected .
- The time interval has been divided into how many parts? How many years does each part contain?
- Does the figure show that the birth rate is constantly declining?
- Does the figure show a constantly declining death-rate?
- State the census year which shows the maximum birth rate. How much is it?
- State the time interval in which the death-rate has fallen suddenly.

We realize that the figure contains a variety of useful information that is easy to understand and analyze. This is what statistics helps us to do.

Today, statistics has become a part of all branches of knowledge. It is used to study problems in Biology, Psychology, Economics, Education, Sociology, Trade, Industry etc.

The statistical method of studying a problem broadly consists of the following steps:

- To collect numerical data about the problem
- To present the collected data systematically
- To analyze the data and
- To interpret the data and draw conclusions from it

The numerical expressions which represent the characteristic of a group (a large number of numerical data) are called measures of central tendency (or averages).

An average which is used to represent a whole series should neither have the lowest value nor the highest value in the group but a value somewhere between the two extremes, possibly in the centre, where most of the items of the group cluster.

There are many types of statistical averages, some of them are the mean, median and mode.

Numerical data and its representation

A class consists of 64 students. The teacher gives a test in English. The scores each student obtained out of 100 are as follows:

Table 1

40	72	14	40	68	46	62	58	37	40
58	38	52	47	16	50	61	37	44	55
38	49	44	52	67	51	33	48	23	51
56	61	46	41	65	43	71	29	50	56
68	25	55	49	44	73	23	63	41	42
66	59	52	28	50	56	60	38	40	73
45	30	47	40						

These marks are the numerical data called **raw data** collected with the purpose of knowing attainment of the class in English.

The way in which the scores are presented hardly gives any idea about attainment of the class in English. It is necessary to present any such numerical data in a systematic manner so as to know some meaning of it.

There are different methods of presenting numerical data systematically.

Arranged data
Arranged numerical data
The same scores have now been arranged in ascending order, in Table 2. Observe the table.

Table 2

14	16	23	23	25	28	29	30	33	37
37	38	38	38	40	40	40	40	40	41
41	42	43	44	44	44	45	46	46	47
47	48	49	49	50	50	50	51	51	52
52	52	55	55	56	56	56	58	58	59
60	61	61	62	63	65	66	67	68	68
71	72	73	73						

Information gathered from the table:

The minimum and maximum marks are 14 and 73 respectively. Some students have scored equal marks. The score 40 has occurred five times which is the maximum repetition.

Drawbacks of this method:

- The method is very tedious.
- Often the numerical data is large and arranging them in the ascending order would take lot of time.
- Even though we do it, we do not get any prominent information from it.

Therefore to get more information, it becomes necessary to put it in a condensed form.

Ungrouped frequency distribution table

The table 3 given below gives the ungrouped frequency distribution of the information in table 1. The table is prepared as follows:

Start from the smallest number in the data and write the numbers one below the other till the largest number. We shall now track the repeated occurrence of numbers by making a **tally mark** like this '|' next to that number. The fifth mark is drawn diagonally crossing the first four marks (𝍸 in this way). This makes counting of tally marks easy. Make the sixth tally mark a small distance from the first five. The total number of tallies corresponding to a number in the data is called the frequency of the number.

Table 3

Tally Marks	Frequency	Marks	Tally Marks	Frequency	Marks
14	|	1	29	|	1
15			30	|	1
16	|	1	31		
17			32		
18			33	|	1
19			34		
20			35		
21			36		
22			37	||	2
23	||	2	38	|||	3
24	-	-	39	-	-
25	|	1	40	𝍸	5
26	-	-	41	||	2
27	-	-	42	|	1
28	|	1	43	|	1

Tally Marks	Frequency	Marks	Tally Marks	Frequency	Marks
44	|||	3	59	|	1
45	|	1	60	|	1
46	||	2	61	||	2
47	||	2	62	|	1
48	|	1	63	|	1
49	||	2	64		
50	|||	3	65	|	1
51	||	2	66	|	1
52	|||	3	67	|	1
53	-	-	68	||	2
54	-	-	69		
55	||	2	70		
56	|||	3	71	|	1
57	-	-	72	||	2
58	||	2	73	||	2

Such a frequency distribution table makes the numerical data more informative. One glance at the table shows us that the number of students scoring marks between 44 and 58 is very large.

Drawback

The table has a large span.

This drawback is removed in a grouped frequency distribution table.

Grouped frequency distribution table

In this method, the numerical data is classified into convenient groups or classes using tally marks.
In the table given below, the data is **classified into groups** 11 to 20, 21 to 30,….,71 to 80.

Table 4

Class	Tally marks	Frequency
11 to 20	||	2
21 to 30	𝍸 |	6
31 to 40	𝍸 𝍸 |	11
41 to 50	𝍸 𝍸 𝍸 |||	18
51 to 60	𝍸 𝍸 ||||	14
61 to 70	𝍸 ||||	9
71 to 80	||||	4
	Total	**64**

Note the following important points of this table:

- The method of preparing this table is easy.
- The numerical data have got a concise form.
- The individual nature of the data has disappeared. The table represents the nature of group.
- With the help of this table, it is possible to analyse the data to some extent.

Some terms used in statistics

Raw numerical data

Table 1 shows marks of each student in a class. In terms of statistics, they are the scores of individuals of a group. This is the primary information collected. Such an information is called Raw numerical data.

Range of the data

In table 2, the numbers are written in order. Let us call such data as the data presented in order. The difference between the largest and the smallest number in the data is called the Range of the data. The range of data in table 2 is 73-14=59.

Class **limit**

In table 4, the numerical data are presented dividing into groups. Each of them is called a class. The end values of a class are called the limits of the class or the class-limits. The smaller of the two values is called the lower class-limit and the larger is called the upper class-limit.

Class interval

The range of a class is called its class interval.

Frequency of the class

The number of tally marks corresponding to a class is called the frequency of the class.

Cumulative frequency table

The cumulative frequency less than the upper limit of a certain class, is the sum of the frequency of that class and the frequencies of all classes preceeding it.

Table 6 is a cumulative frequency table prepared from table 4.

Table 6

Class	Frequency (No. of students)	Cumulative frequency (Less than the upper class limit)
11 to 20	2	2
21 to 30	6	2+6 = 8
31 to 40	11	8+11 = 19
41 to 50	18	19+18 = 37
51 to 60	14	37+14 = 51
61 to 70	9	51+9 = 60
71 to 80	4	60+4 = 64

In this table, the column of cumulative frequency shows the number of scores less than the upper class limit of the corresponding class. Hence such a table is called **'a cumulative frequency less than'** table.

Similarly, the cumulative frequency more than the lower limit of a class is the sum of the frequency of that class and the frequencies of all the class succeeding to it. The table 7 given below shows the cumulative frequency of this type.

Table 7

Class	Frequency (No. of students)	Cumulative frequency (Less than the upper class limit)
11 to 20	2	62+2 = 64
21 to 30	6	56+6 = 62
31 to 40	11	45+11 = 56
41 to 50	18	27+18 = 45
51 to 60	14	13+14 = 27
61 to 70	9	4+9 = 13
71 to 80	4	4

In this table, the column of cumulative frequency shows the number of scores more than the lower class limit of the corresponding class. Hence, such a table is called a cumulative-frequency-more-than table.

To frame such a table, record classes and their corresponding frequencies in a table. Write the cumulative frequencies from bottom to top of the table. The last class is 71 to 80 and its corresponding frequency is 4. Therefore, the cumulative frequency of that class is 4. The class preceding is 61 to 70 and its frequency is 9.

Hence the cumulative frequency of that class is 4 + 9 = 13.

Representation of statistical data

Numerical data can be represented by two methods:

- Diagramatic representation
- Graphical representation

Diagramatic representation

There are different forms of diagrammatic representation of numerical data.

- Bar diagram
- Pie diagram

Graphical representation

Some of the graphical methods of representing numerical data are

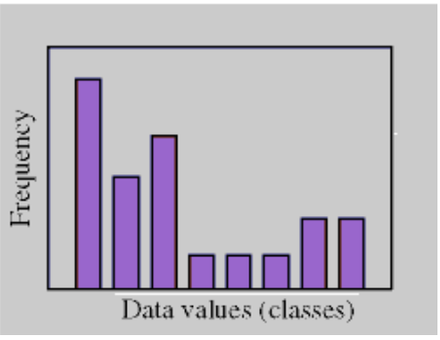
- Histogram
- Frequency polygon
- Ogive curve

Graphical representation of statistical data

Histogram

A histogram is a two-dimensional graphical representation of a continuous frequency distribution.

A histogram is a special type of bar diagram.



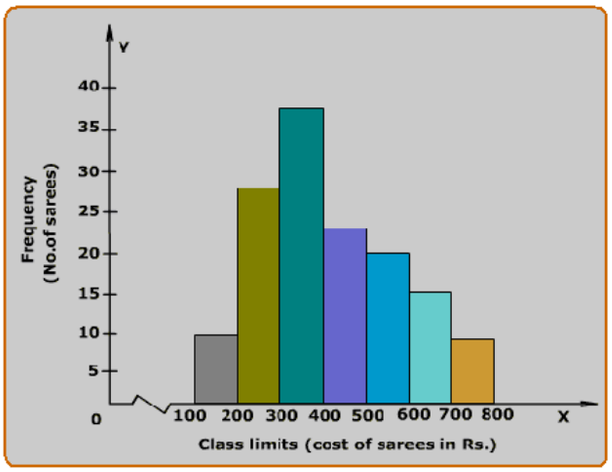
A histogram is a way of summarizing data pictorially. Histograms show the distribution of the data. They are constructed from a frequency table, which is a summary of the data. The general format for a histogram is a vertical scale that demonstrates frequencies and a horizontal scale that represents the individual intervals, sometimes called classes. Bars are used to represent each individual interval with the height of the bar corresponding to the frequency.

A histogram is drawn according to the steps given below.

- Prepare a grouped frequency distribution table of the given data.
- Show class-limits on X-axis with a suitable scale.
- Show frequencies on Y-axis with a suitable scale.
- Draw rectangles having base equal to the class limits and heights proportional to the frequencies. These rectangles should be joined to each other.

A frequency distribution table is shown below.

Class (Cost of saree in Rs.)	Frequency (No. of sarees sold in a week)
100 - 200	12
200 - 300	28
300 - 400	37
400 - 500	23
500 - 600	20
600 - 700	14
700 - 800	09

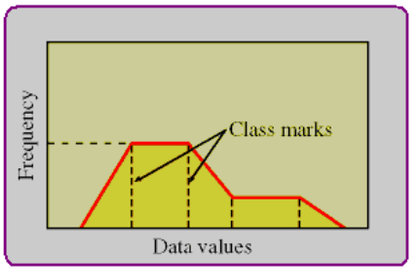


The following points can be inferred from the above histogram.

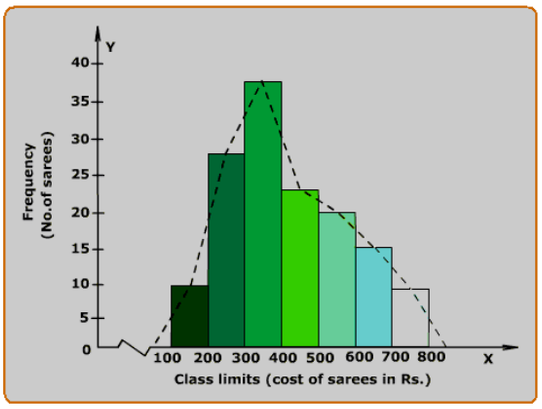
- It is not necessary that the scale on the X-axis and the Y-axis be the same. Different scales may be taken on the axes considering the nature of the data, size of the paper etc. A histogram should look neat and attractive.
- The position of origin on the Y-axis is according to the scale, which is not so on the X-axis. This is indicated by drawing '⌐' mark on the X-axis near the origin. If necessary, the mark can be made on the Y-axis or on both axes.
- In a histogram, it is necessary that the adjacent rectangles be attached to each other. Therefore, if the given classes are not continuous, it is necessary to make them continuous e.g.; if the classes are 2 to 5, 6 to 9, 10 to 13,…. It should be as 1.5 to 5.5, 5.5 to 9.5, 9.5 to 13.5…

Frequency Polygon

A frequency polygon is a way of picturing data by plotting the class mark on the horizontal axis and the frequency of the class on the vertical axis and connecting the points. The polygon is completed by extending the class marks one-class width on either end with a frequency of zero for both.



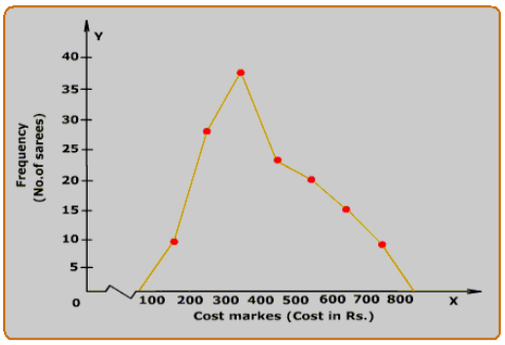
A frequency polygon can be drawn according to the following steps.



- Draw a histogram from the given data.
- Join the consecutive mid-points of the upper horizontal sides of the rectangles in the histogram.
- It is assumed that the class preceding the first class and the class succeeding the last class in the classification exists and the frequency of each of them is zero. Class marks of these classes are joined with the mid-points of the upper horizontal sides of the extreme rectangles of the histogram.

The figure above shows the frequency polygon drawn with the help of the histogram. Observe it carefully. A frequency polygon can be drawn without drawing a histogram.

Applying the method of point plotting, a frequency polygon can be drawn as follows:



Recap

Graphical representation of statistical data includes construction of

Histogram

Class-intervals on the x-axis and cumulative frequencies on the y-axis, the corresponding rectangles are drawn.

Frequency polygon

A histogram is drawn and the midpoints of the rectangles are joined by straight lines.

Arithmetic Mean

(a) Arithmetic mean for ungrouped data

The arithmetic mean of a set of raw data is obtained by adding all the values of the variable given and dividing this sum by the total number of values.

Let 'n' be the total number of values and x_1, x_2, x_3 … x_n be the values recorded in the data. Then the arithmetic mean is defined as follows,

$$\text{Arithmetic Mean} = \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots . x_n}{n}$$

$$\text{or } \bar{x} = \frac{\Sigma x_n}{n}$$

The symbol Σ denotes: ' Sum of '

(b) Arithmetic mean for ungrouped data by direct method

In the previous example, all the 50 marks are not distinct, for example, 3 students have 7 marks each, 4 students have 19 marks each etc.

Therefore the sum can be written as

$$\Sigma x_n = 3 \times 7 + 4 \times 19 + \ldots . + 4 \times 91$$

Marks	Frequency f_i	$f_i x_i$
7	3	21
19	4	76
31	5	155
40	7	280
49	9	441
62	7	434
73	6	438
83	5	415
91	4	364
	n=50	$\sum f_i x_i = 2624$

$$\text{Arithmetic Mean} = \frac{\sum fx}{n} = \frac{2624}{50}$$

= 52.48

(c) Arithmetic mean for grouped data by direct method

Example:

Find the arithmetic mean for the following frequency distribution.

Marks	F	Mid-pt x_i	fx
5-15	3	10	30
15-25	4	20	80
25-35	5	30	150
35-45	7	40	280
45-55	9	50	450
55-65	7	60	420
65-75	6	70	420
75-85	5	80	400
85-95	4	90	360
	n=50		2590

$$\text{Mean} = \frac{\sum fx}{N}$$

$$= \frac{2590}{50}$$

Mean = 51.80

Assumed Mean or short-cut method for calculating the mean

(a) Short cut method for ungrouped data

In this method, an assumed mean (A) is taken from the scores, usually about the middle. If there are two middle scores, the one with the higher frequency is taken as the assumed mean and then the arithmetic mean is obtained by using the formula.

$$\text{Mean} = A + \frac{\Sigma fd}{\Sigma f}$$

where A is the assumed mean, d is the deviation of x from the assumed mean A.

(b) Short-cut method for grouped data

In this method, an assumed mean (A) is taken from the mid-values near about the middle of the table and then the Arithmetic Mean is obtained by using the following formula,

$$\text{Mean} = A + \frac{\Sigma fd}{\Sigma f}$$

where, 'A' is the assumed mean, 'd' is the deviation of 'x' from assumed mean 'A'.

(c) Step-deviation method

According to this method,

$$\text{Mean} = A + \frac{\Sigma fu}{\Sigma f} \times i,$$

$$\text{Where } A = \text{Assumed Mean, } u = \frac{x - A}{i} \text{ and}$$

i= class size [i.e., upper limit - lower limit]

Median and Mode

Median

If the given statistical data be arranged in ascending or descending order of their values, then the value of the middle term is called the median.

Let 'n' be the number of scores in ascending or descending order.

$$\text{Then, Median} = \left(\frac{n}{2} + 1\right)^{th} \text{ term, 'n' is odd}$$

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} \text{ term} + \left(\frac{n}{2} + 1\right)^{th} \text{ term}}{2}, \text{ when 'n' is even.}$$

Mode

The number which appears maximum times in the given statistical data is called mode.

In other words, mode is the number whose frequency is maximum.