## Frequency Distribution

It may be noted that numerical data may arise in two ways:

* In some cases, the data are numerical to start with, e.g. when we record the height for each of a group of men or the number of rooms in each house of a town.
* In other cases, numbers arise only secondarily. When we record the sex of each newborn baby during a month or the language of each book in a library, the data are not numbers initially. We get numbers if, subsequently, we note the number of male and female babies, or the number of books written in English, Hindi, Bengali etc.

**Attribute:** The qualitative character (e.g. sex of babies, colour of eye, language of books etc.) which cannot be expressed in numerical terms is referred to as an attribute.

**Variable:** The quantitative character (e.g. Height, weight, age etc.) which can be expressed in numbers is referred to **as a variable.**

**Variate:** A particular value of a variable is called variate (observation).

**Variables** may be of two principal types:

**Discrete variable:** The character may take only some isolated values. For example, the number of letters in a word, number of petals in a flower, number of members in a family and so forth.The variable of this type is called discontinuous or discrete variable.

**Continuous variables:** The character may conceivably take any value within its range of variation. For example, the height, weight of a man, the diameter of a bobbin, the temperature, rainfall or humidity in a region, etc. The variable of this type is called continuous variable.

## Often the raw data will be so numerous that their significance will not be readily comprehended. For examples,

* To know the awareness about stock market among Indian, 150 people were interviewed and each was asked among other question whether he/she knew about Bombay Stock Exchange. The responses of 150 people are recorded as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | U | U | A | A | U | U | A | A | A | A | U | A | U | U |
| U | A | A | U | A | A | U | A | U | U | A | A | A | A | A |
| U | U | A | A | U | U | A | A | U | U | U | U | A | A | A |
| A | U | A | A | A | U | A | A | A | A | A | U | U | A | A |
| A | A | A | A | A | A | A | U | A | U | U | A | U | A | U |
| A | A | U | A | A | A | U | U | A | A | A | A | A | U | U |
| U | U | U | U | U | A | A | A | U | A | A | A | A | A | A |
| A | U | A | U | A | A | U | A | U | A | A | U | U | A | A |
| A | A | A | A | A | U | A | A | A | U | A | A | A | U | A |
| A | U | U | A | A | A | A | U | A | U | U | A | U | A | U |

A = Aware and U = Unaware

## Rajesh bought some pea-pods in the marker. On returning home he found there are 198 pea-pods in his bag. He took each pod and counted the number of peas it contained. The figures thus obtained are given below:

4 3 5 3 5 2 4 5 2 4 4 4 5 3 5 3 6 3 2 2 3 4 3 2 3 3 4 3 4 6 4 3 3 3 1 3 2 4 3 3 3 3 3 2 4 5 3 4 3 2 4 3 3 2 2 6 1 3 5 2 4 4 3 3 5 4 2 3 3 3 7 6 4 4 3 3 2 3 4 4 3 3 2 3 6 3 4 2 4 4 3 3 2 2 3 5 3 4 2 3 2 3 4 5 3 4 5 2 5 3 3 4 3 5 5 6 4 5 4 3 5 4 3 3 3 5 5 4 4 4 3 3 6 4 4 4 1 4 4 3 2 2 4 3 2 3 5 3 4 3 2 6 3 5 4 4 3 2 2 5 3 3 4 3 2 2 3 3 3 3 4 3 5 4 3 4 5 2 3 3 5 3 3 4 3 2 2 3 4 4 1 5 5 3 2 2 4

* The ages (in years) of 200 unemployed persons who registered their names in an Employment Exchange on a city are shown below:

## 33 26 25 20 21 21 24 36 27 31 22 24 25 22 22 31 24 17 31 24 26 17 20 40 49 20 26 19 19 18 27 46 22 20 24 18 18 18 24 22 33 29 21 20 34 23 29 28 24 25 17 27 24 37 17 22 15 25 23 30 30 22 29 20 26 29 29 26 19 23 22 20 32 19 29 19 18 31 21 19 24 18 27 27 29 23 20 31 39 19 43 26 24 24 24 33 24 26 24 25 25 40 26 19 31 17 17 52 23 30 33 20 20 21 17 25 19 29 18 21 28 20 19 20 25 23 32 25 24 20 22 21 22 28 23 24 22 24 21 21 22 33 37 20 33 27 23 57 25 19 17 24 15 27 21 25 22 31 19 25 17 21 32 29 23 43 24 16 17 50 21 22 34 19 21 23 16 23 30 45 30 25 21 20 16 22 25 26 19 40 25 21 22 31 31 21 21 18 24 24

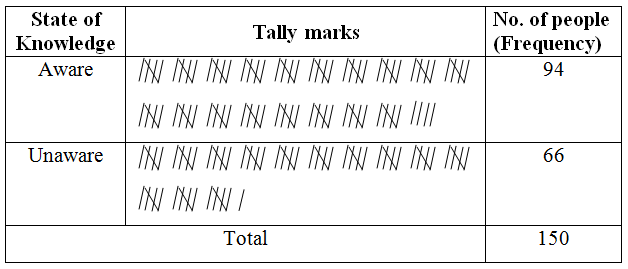
## In a test run of 50 cars of the same model, the mileage covered per litre of fuel consumed by each car was as follows:

18.6 33.4 25.3 27.8 30.6 31.9 33.0 26.3 24.9 29.4 20.0 26.2 28.1 33.1 37.5 22.5 39.1 32.9 33.8 52.6 32.5 34.6 32.7 09.5 38.5 29.6 25.3 49.5 30.1 27.9 26.9 23.8 36.0 38.0 27.5 32.3 34.2 23.1 34.7 29.0 34.1 38.6 25.9 40.6 53.8 29.3 36.8 27.1 34.9 31.6

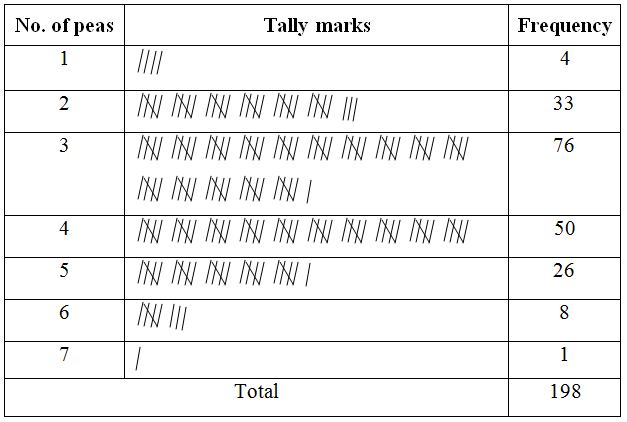
It may be noted that in all the above cases, a large number of observations are available in the form of a simple series and the mind cannot properly grasp their significance. Therefore, some method of **condensation** of data is necessary. One way of data condensation or data summarization is to write down the values of the variable in order of magnitude and show the corresponding frequencies side by side in a tabular form. Such a table is called Frequency table or Frequency distribution of the variable.

***Frequency distribution or Frequency table*** *is a statistical table which shows the values of the variable arranged in order of magnitude, either individually or in groups, and also the corresponding frequencies side by side.*

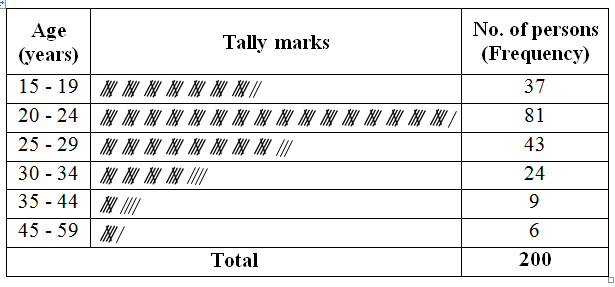
## Table 1: Frequency distribution of survey results



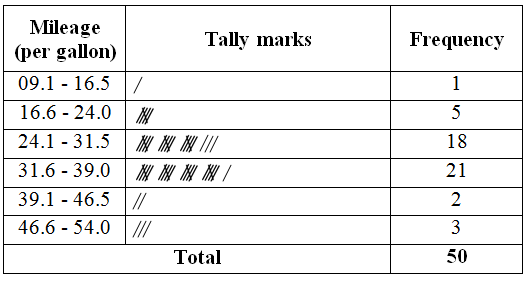
**Table 2:** Frequency distribution of number of peas per pod



**Table 3:** Grouped frequency distribution of age



**Table 4:** Grouped frequency distribution of mileage per litre



## It can be noticed that the frequency distribution *gives a quick idea about the central value (average)* in the series, and *shows how the observations vary (‘dispersion’) around the average*. The histogram and frequency polygon drawn from a frequency distribution also through light on the shape (‘skewness’ and ‘kurtosis’) of the frequency curve. Again, all recorded data may be thought of as a sample coming from a certain ‘population’. The frequency distribution may be used to *approximate the ‘probability distribution’ in the population*, which is fundamental in solving statistical problems.

Thus the important uses of a frequency distribution are:

1. Condensation and summarization of a large mass of data
2. Visualizing the important characteristics of data, like average, dispersion, skewness and kurtosis
3. Approximating the nature of probability distribution in the population.

## Some useful terms associated with grouped frequency distribution

**Class interval:** When large numbers of observations varying in a wide range are available these are usually classified in several groups according to the size of values. Each of these groups defined by an interval is called class interval or simply class. It is necessary that these classes are mutually exclusive and exhaustive.

When one end of a class is not specified, the class is called an open-end class. The necessity of open-end classes arises when there are relatively few observations which are far apart from the rest. In such a case it is not considered worthwhile to show several classes with zero frequencies before reaching a class with a very small frequency**.**

**Class frequency:** The number of observations falling within a class is called its class frequency.

**Class limits:** The two numbers used to specify the limits of a class interval for the purpose of tallying the original observations into various classes are called class limits. The smallest of the pair is known as lower class limit and the largest as the upper class limit with reference to the particular class.

**Class boundaries:** When measurements are taken on continuous variable, all data are recorded nearest to a certain unit. Thus if age are recorded to the nearest whole number of years and any age between 14.5 years and 15.5 years is recorded as 15 years. Similarly 19 years denotes an age between 18.5 and 19.5 years. Hence the class interval 15-19 actually includes all ages between 14.5 and 19.5 years. *These extreme values which would ever be included in a class interval are called class boundaries. Class boundaries are, in fact, the real limits of a class interval.* The lower extreme point is called lower class boundary and the upper extreme point is called upper class boundary with reference to any particular class.

Class boundaries may be calculated from class limits by applying the following rule:

where is the common difference between the upper class limit of any class interval and lower class limit of the next class interval.

Note that the upper boundary of any class coincides with the lower boundary of the next class. But the upper limit of any class is different from the lower limit of the next class.

Class limits are used only for the construction of a grouped frequency distribution. But in all statistical calculations and diagrams involving end points of classes (e.g. median, mode, quartiles, histogram, ogive etc.) class boundaries are used.

**Class mark or Mid-value:** The value exactly at the middle of a class interval is called class mark or Mid-value

Class mark is used as a representative value of the class interval for the calculation of mean, standard deviation, mean deviation etc.

**Width or Size of class interval:** Width of a class is the difference between the lower and upper class boundaries (not class limits).

It is generally preferable to have classes of equal width. Unequal width of class is resorted to when some of the observations are few and far away from the rest. The use of equal width in such cases may result in some empty classes.

**Frequency density:** Frequency density of a class is its frequency per unit width. It shows the concentration of frequency in a class.

Frequency density is used in drawing histogram when the classes are of unequal width.

**Relative frequency:** A relative frequency of a class is the ratio (fraction or proportion) of the class frequency to the total frequency. It is useful for comparison of different classes.

**Guidelines for construction of frequency distribution**

The steps in the construction of a frequency distribution from ungrouped data are as follows:

1. Find the largest and smallest observations in the given data and then calculate the range (R), which is difference between the largest and smallest observations.
2. Divide the range into suitable number of class intervals, by means of class limits.

(The number of class intervals should not ordinarily be less than 5 and more than 15. The class intervals should preferably be of the same width. In special circumstances, class intervals of unequal width may also be used.)

* Determination of number of classes
* **Sturges rule for number of classes:** N, where is the number of classes and is the total number of observations. **2K**
* **Rule for number of classes**: ,where is the number of classes and is the total number of observations.
* Determination of width of class intervals

(Rounded to a suitable value)

1. For continuous variable, convert the class limits into class boundaries.
2. Determine the number of observations falling in each class interval by tally marks.
3. A table is now prepared showing the class intervals in the first column and the corresponding class frequencies in the second column. This is required frequency distribution.

**Exercise 1:**  Table 1, given below, contains data on Gain of 120 tested amplifiers. Prepare a frequency distribution of these data.

**Table 1: Data on Gain of 120 tested amplifiers**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8.1 | 10.4 | 8.8 | 9.7 | 7.8 | 9.9 | 11.7 | 8.0 | 9.3 | 9.0 |
| 8.2 | 8.9 | 10.1 | 9.4 | 9.2 | 7.9 | 9.5 | 10.9 | 7.8 | 8.3 |
| 9.1 | 8.4 | 9.6 | 11.1 | 7.9 | 8.5 | 8.7 | 7.8 | 10.5 | 8.5 |
| 11.5 | 8.0 | 7.9 | 8.3 | 8.7 | 10.0 | 9.4 | 9.0 | 9.2 | 10.7 |
| 9.3 | 9.7 | 8.7 | 8.2 | 8.9 | 8.6 | 9.5 | 9.4 | 8.8 | 8.3 |
| 8.4 | 9.1 | 10.1 | 7.8 | 8.1 | 8.8 | 8.0 | 9.2 | 8.4 | 7.8 |
| 7.9 | 8.5 | 9.2 | 8.7 | 10.2 | 7.9 | 9.8 | 8.3 | 9.0 | 9.6 |
| 9.9 | 10.6 | 8.6 | 9.4 | 8.8 | 8.2 | 10.5 | 9.7 | 9.1 | 8.0 |
| 8.7 | 9.8 | 8.5 | 8.9 | 9.1 | 8.4 | 8.1 | 9.5 | 8.7 | 9.3 |
| 8.1 | 8.2 | 9.0 | 10.2 | 9.5 | 8.3 | 8.9 | 9.1 | 10.3 | 8.4 |
| 8.5 | 8.2 | 9.0 | 10.2 | 9.5 | 8.3 | 8.9 | 9.1 | 10.3 | 8.4 |
| 8.6 | 9.2 | 8.5 | 9.6 | 9.0 | 10.7 | 8.6 | 10.0 | 8.8 | 8.6 |

**Exercise 2:** Heights (in cm.) of 177 Indian adult males are given below. Construct a frequency distribution of heights.

169.0 166.7 159.9 157.8 169.9 158.4 171.7 160.4 167.5 161.0 168.8 167.8 164.0 167.4 167.8 165.2 163.5 170.4 159.0 158.1 157.6 159.2 167.7 170.2 169.0 157.5 161.5 165.8 159.6 160.3 164.5 166.7 165.0 159.7 158.9 168.9 163.9 162.0 165.2 167.4 172.8 168.2 166.8 163.5 159.3 163.1 168.9 166.3 164.5 162.1 182.0 158.5 183.5 163.5 170.1 167.7 157.4 164.9 168.3 164.0 154.2 161.8 156.5 161.7 145.6 162.2 162.0 165.5 167.2 159.8 169.0 165.3 158.0 169.5 169.2 161.5 166.7 162.6 171.3 166.7 168.0 160.8 168.0 156.0 169.0 172.5 171.1 168.2 172.6 169.3 163.0 163.1 157.6 160.9 152.5 167.9 164.2 167.5 164.2 171.7 167.7 168.0 168.1 164.4 165.2 163.5 176.4 160.9 164.2 161.0 158.4 171.4 159.5 162.7 160.5 171.7 163.7 162.3 171.9 169.7 171.6 167.0 169.5 163.7 162.5 166.8 160.2 163.9 171.0 156.4 170.0 161.4 160.5 173.2 174.2 161.0 161.4 165.4 160.0 156.9 164.9 167.3 159.5 165.2 160.8 170.1 166.4 168.1 168.7 165.4 157.5 160.5 170.3 167.0 171.3 162.2 169.4 170.0 166.6 160.5 160.6 168.0 155.9 162.0 161.4 172.2 156.0 163.2 172.0 152.6 167.1 161.3 170.1 158.7 167.0 178.4 165.5

**Exercise 3:** The ages (in years) of 200 unemployed persons who registered their names in an Employment Exchange on a city are shown below. Construct a frequency distribution of ages.

## 33 26 25 20 21 21 24 36 27 31 22 24 25 22 22 31 24 17 31 24 26 17 20 40 49 20 26 19 19 18 27 46 22 20 24 18 18 18 24 22 33 29 21 20 34 23 29 28 24 25 17 27 24 37 17 22 15 25 23 30 30 22 29 20 26 29 29 26 19 23 22 20 32 19 29 19 18 31 21 19 24 18 27 27 29 23 20 31 39 19 43 26 24 24 24 33 24 26 24 25 25 40 26 19 31 17 17 52 23 30 33 20 20 21 17 25 19 29 18 21 28 20 19 20 25 23 32 25 24 20 22 21 22 28 23 24 22 24 21 21 22 33 37 20 33 27 23 57 25 19 17 24 15 27 21 25 22 31 19 25 17 21 32 29 23 43 24 16 17 50 21 22 34 19 21 23 16 23 30 45 30 25 21 20 16 22 25 26 19 40 25 21 22 31 31 21 21 18 24 24

## Exercise 4: The data below give the marks secured by 70 candidates in a certain examination. Construct a frequency distribution of marks.

21 31 35 52 64 74 89 53 42 07 22 35 43 67 76 35 46 26 32 40 72 43 38 41 63 71 28 32 45 54 15 18 52 73 86 50 39 55 47 12 44 58 67 85 39 40 50 65 72 69 57 63 05 56 79 37 24 54 82 49 51 54 68 29 34 44 58 62 59 65

## Exercise 5: In a test run of 50 cars of the same model, the mileage covered per litre of fuel consumed by each car was as follows:

18.6 33.4 25.3 27.8 30.6 31.9 33.0 26.3 24.9 29.4 20.0 26.2 28.1 33.1 37.5 22.5 39.1 32.9 33.8 52.6 32.5 34.6 32.7 09.5 38.5 29.6 25.3 49.5 30.1 27.9 26.9 23.8 36.0 38.0 27.5 32.3 34.2 23.1 34.7 29.0 34.1 38.6 25.9 40.6 53.8 29.3 36.8 27.1 34.9 31.6

**Cumulative frequency distribution**

In statistical investigations, sometimes we are interested in the number of observations less than (or greater than) a given value. In such cases, our chief concern is the accumulated frequency up to (or above) some value of the variable. This accumulated frequency is known as cumulative frequency.

The number of observations up to a given value is called less-than cumulative frequency and the number of observations greater than a value is called the more-than cumulative frequency.

When a grouped frequency distribution relates to a variable of the continuous type, the cumulative frequencies calculated therefrom must be shown against the class boundary points. Cumulative frequency expressed as a percentage of total frequency is known as cumulative percentage.

**Uses:** Cumulative frequency distribution is used

1. To find the number of observations less than (or more than) any given value;
2. To find the number of observations falling between any two specified values of the variable;
3. To find median, quartiles, deciles and percentiles graphically, i.e. in general, to find the value of the variable below (or above) which a specified number or percentage of the total frequency lies.

For example the cumulative frequency distributions of age data shown in Table 4 can be presented as follows:

**Table 5:** Cumulative frequency distribution of age

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age in years | Frequency | Relative frequency | Cumulative frequency | |
| Less than | More than |
| 14.5 – 19.5 | 37 | 0.185 | 37 | 200 |
| 19.5 – 24.5 | 81 | 0.405 | 118 | 163 |
| 24.5 – 29.5 | 43 | 0.215 | 161 | 82 |
| 29.5 – 34.5 | 24 | 0.120 | 185 | 39 |
| 34.5 – 44.5 | 9 | 0.045 | 194 | 15 |
| 44.5 – 59.5 | 6 | 0.030 | 200 | 6 |
| Total | 200 | 1.000 |  |  |

It should be noted that the cumulative frequencies of the less-than type correspond to the upper class-boundaries. For instance, the third one, i.e. 161 is the number of persons with age 29.5 or less. Similarly, the cumulative frequencies of the more-than type correspond to the lower class-boundaries.