Introduction to Descriptive Statistics

1 Introduction

Since Statistics involves the collection and interpretation of data, we must first know how to understand, display and summarise large amounts of quantitative information, before undertaking a more sophisticated analysis. Statistical analysis of quantitative data is important throughout the pure and social sciences.

Example. Survival of cancer patients: A cancer patient wants to know the probability that he will survive for at least 5 years. By collecting data on survival rates of people in a similar situation, it is possible to obtain an empirical estimate of survival rates. We cannot know whether or not the patient will survive, or even know exactly what the probability of survival is. However, we can estimate the proportion of patients who survive from data.

Example. Car maintenance: When buying a certain type of new car, it would be useful to know how much it is going to cost to run over the first three years from new. Of course, we cannot predict exactly what this will be — it will vary from car to car. However, collecting data from people who bought similar cars will give some idea of the *distribution* of costs across the *population* of car buyers, which in turn will provide information about the likely cost of running the car.

2 Data presentation

There is lots of information contained in the data, but it is hard to see. The use of *graphs* and *summary statistics* for understanding data is an important *first* step in the undertaking of any statistical analysis. For example, it is useful for understanding the main features of the data, for detecting *outliers*, and data which has been recorded incorrectly. Outliers are extreme observations which do not appear to be consistent with the rest of the data.

2.1 Frequency tables

It is important to investigate the *shape* of the *distribution* of a random variable. This is most easily examined using *frequency tables* and *diagrams*. However, if there are a large number of different observations, consecutive observations may be grouped together to form combined categories.

Example 1. Germinating seeds. We can construct the following frequency table.

No. germinating	85	86	87	88	89	90	91	92	93	94
Frequency	3	1	5	2	3	6	11	4	4	1

Since we only have 10 categories, there is no need to grouped them.

Example 2. Survival times. We have the following table.

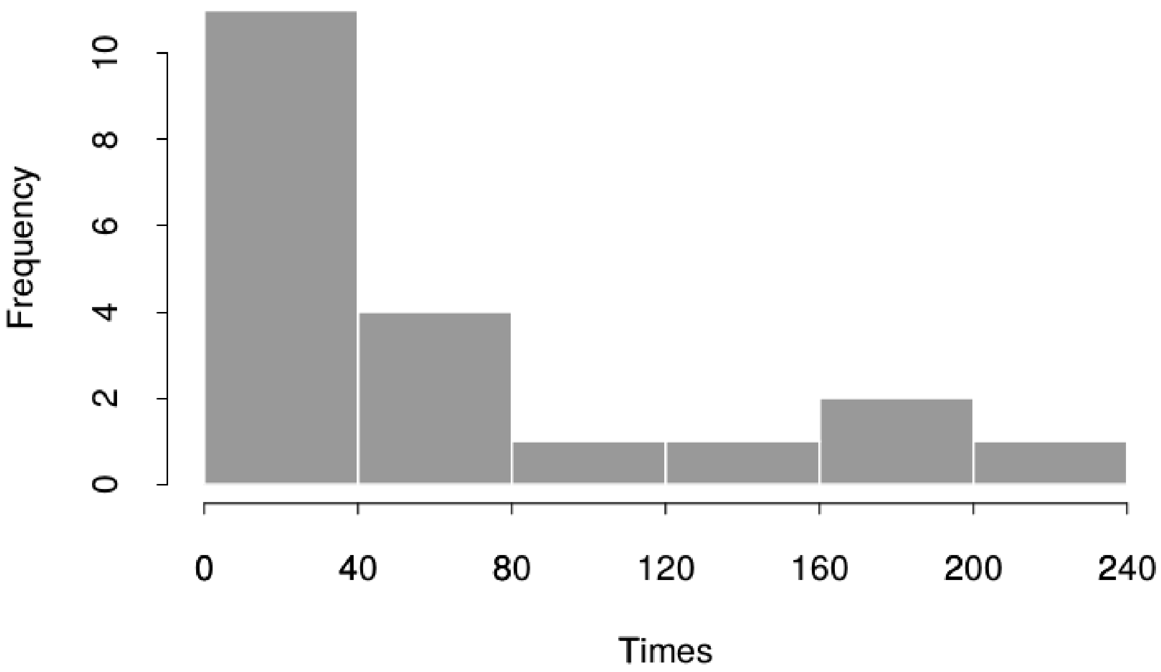
Range	Frequency
0 — 39	11
40 — 79	4
80 — 119	1
120 — 159	1
160 — 199	2
200 — 240	1

You should define the intervals to the same accuracy of the data. Thus, if the data is defined to the nearest integer, the intervals should be (as above). Alternatively, if the data is defined to one decimal place, so should the intervals. Consequently, if the data has been rounded to the nearest integer, then the intervals are actually $0 - 39.5$, $39.5 - 79.5$, etc.

2.2 Histograms

A histogram is a graphical representation of a frequency distribution. Once the frequency table has been constructed, pictorial representation can be considered. For most continuous data sets, the best diagram to use is a *histogram*. In this the classification intervals are represented to scale on the x-axis of a graph and rectangles are drawn on this base with their *areas* proportional to the frequencies. Hence the y-axis is frequency per unit class interval. Note that the *heights* of the rectangles will be proportional to the frequencies if and only if class intervals of equal width are used. Unlike bar charts, the rectangles are joined together for histogram.

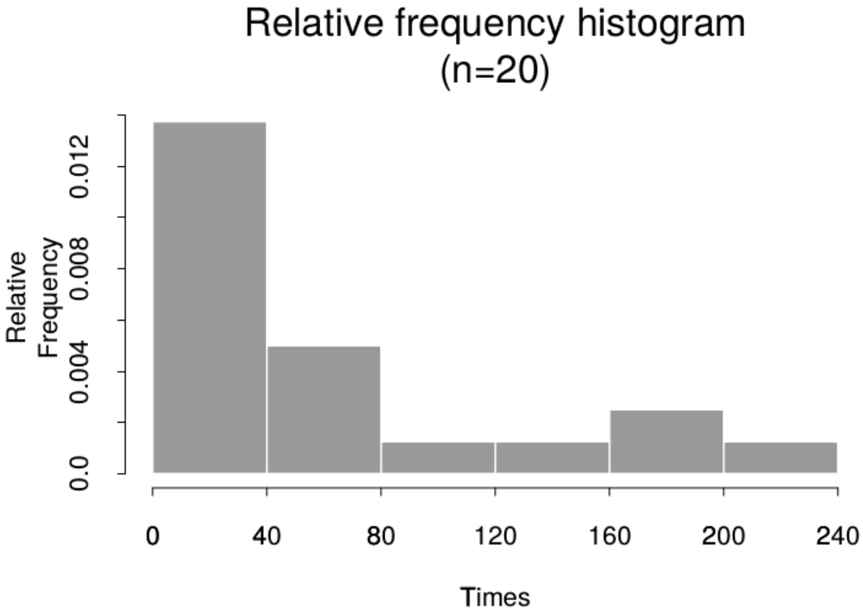
The histogram for Example 2 is as follows.



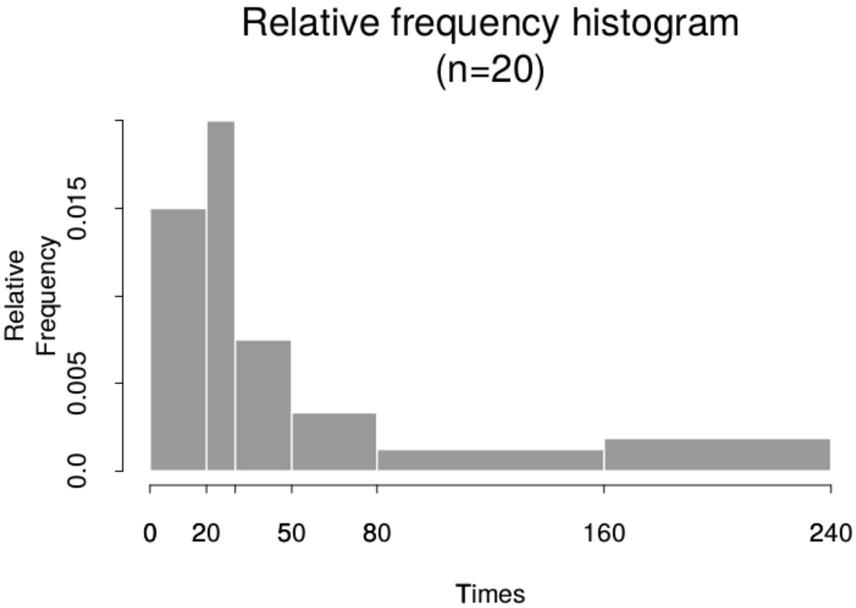
Note that here we have labeled the y-axis with the raw frequencies. This only makes sense when all of the intervals are the same width. Otherwise, we should label using relative frequencies, as follows.

The y-axis values are chosen so that the area of each rectangle is the proportion of observations falling in that bin. Consider the first bin (0–39). The proportion of observations falling into this bin is 11/20 (from the frequency table). The area of our rectangle should, therefore, be 11/20. Since the rectangle has a base of 40, the height of the rectangle must be $11/(20 \times 40) = 0.014$. In general therefore, we calculate the bin height as follows:

$$Height = \frac{Frequency}{n \times BinWidth}$$



This method can be used when the interval widths are not the same, as shown below.

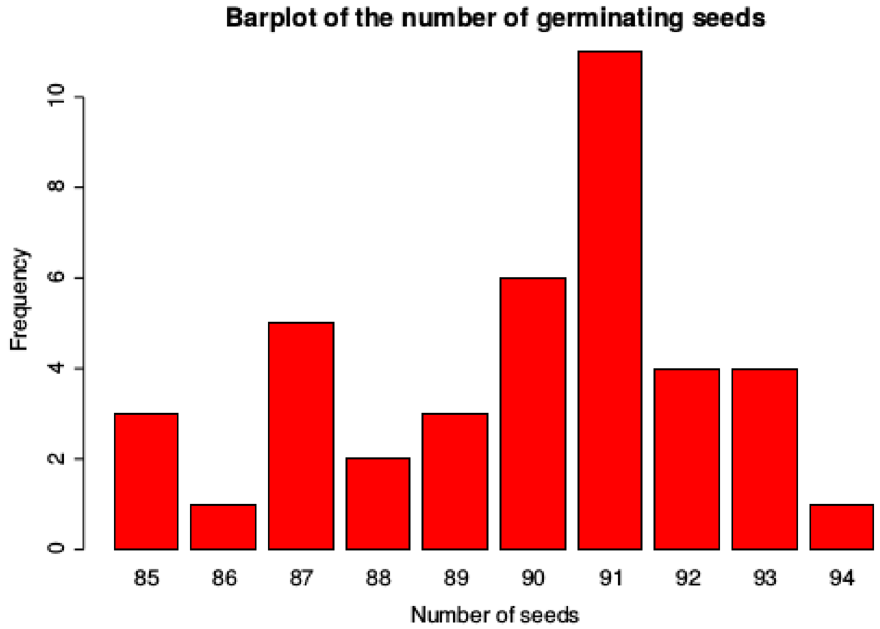


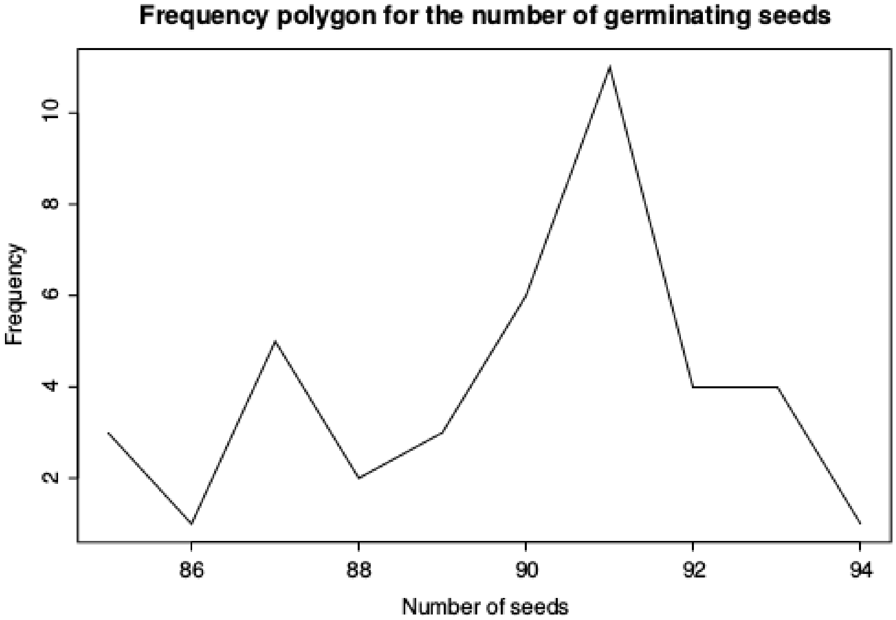
Note that when the y-axis is labeled with relative frequencies, the area under the histogram is always one. Bin widths should be chosen so that you get a good idea of the distribution of the data, without being swamped by random variation.

2.3 Bar charts and frequency polygons

When the data are discrete and the frequencies refer to individual values, we display them graphically using a *bar chart* with heights of bars representing frequencies, or a *frequency polygon* in which only the tops of the bars are marked, and then these points are joined by straight lines. Bar charts are drawn with a gap between neighboring bars so that they are easily distinguished from histograms. Frequency polygons are particularly useful for comparing two or more sets of data.

Example: Consider again the number of germinating seeds from Example 1. Using the frequency table constructed earlier, we can construct a Bar Chart and Frequency Polygon as follows.





2.4 Pie chart

A third method which is sometimes used for *qualitative* data is called a *pie chart*. Here, a circle is divided into sectors whose *areas*, and hence *angles* are proportional to the *frequencies* in the different categories. Pie charts should generally not be used for *quantitative* data – a bar chart or frequency polygon is almost always to be preferred. Whatever the form of the graph, it should be clearly labeled on each axis and a fully descriptive title should be given, together with the number of observations on which the graph is based.

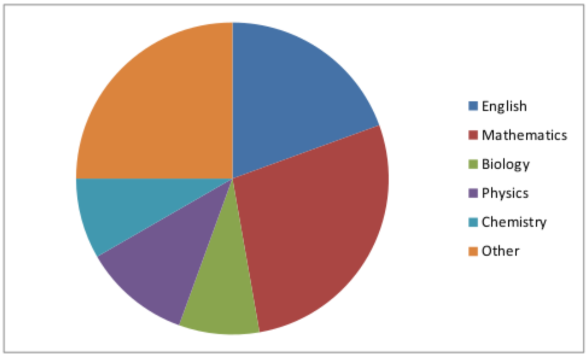
Example 3. In ABC international school, the lesson periods for each week are given below.

English 7, Mathematics 10, Biology 3, Physics 4, Chemistry 3, others 9. Draw a pie chart to illustrate this information.

Solution. Total no of periods in a week = 7+10+3+4+3+9=36

Subject	No of period	Angle of sector
English	7	$\frac{7}{36} \times 360^\circ = 70^\circ$
Mathematics	10	$\frac{10}{36} \times 360^\circ = 100^\circ$
Biology	3	$\frac{3}{36} \times 360^\circ = 30^\circ$
Physics	4	$\frac{4}{36} \times 360^\circ = 40^\circ$
Chemistry	3	$\frac{3}{36} \times 360^\circ = 30^\circ$
others	9	$\frac{9}{36} \times 360^\circ = 90^\circ$
Total	36	360°

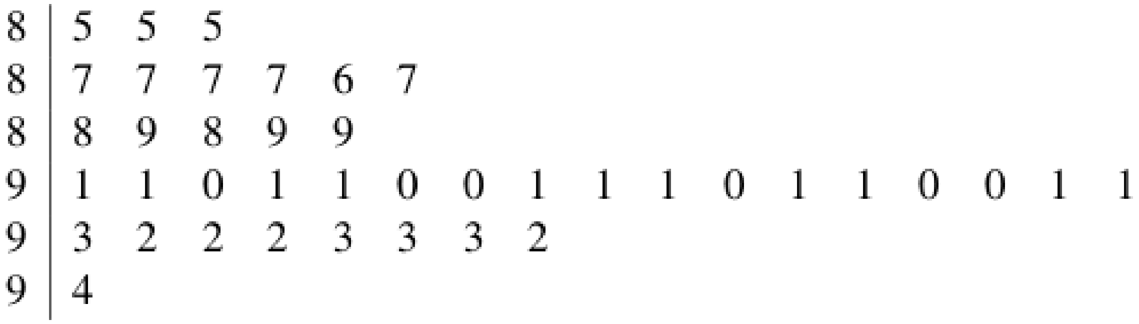
The pie chart showing the lesson period in ABC international school



2.5 Stem-and-leaf plots

A good way to present both continuous and discrete data for sample sizes of less than 200 or so is to use a *stem-and-leaf plot.* This plot is similar to a bar chart or histogram, but contains more information. As with a histogram, we normally want 5–12 intervals of equal size which span the observations. However, for a stem-and-leaf plot, the widths of these intervals must be 0.2, 0.5 or 1.0 times a power of 10, and we are not free to choose the end-points of the bins. They are best explained in the context of an example.

Example. Recall again the seed germination Example 1. Since the data has a range of 9, an interval width of 2 $(= 0.2 \times 10\hat{}\ 1\ )$ seems reasonable. To form the plot, draw a vertical line towards the left of the plotting area. On the left of this mark the interval boundaries in increasing order, noting only those digits that are common to all of the observations within the interval. This is called the *stem* of the plot. Next go through the observations one by one, noting down the next significant digit on the right-hand side of the corresponding stem.



```
8 | 5 5 5
8 | 7 7 7 7 6 7
8 | 8 9 8 9 9
9 | 1 1 0 1 1 0 0 1 1 1 0 1 1 0 0 1 1
9 | 3 2 2 2 3 3 3 2
9 | 4
```

For example, the first stem contains any values of 84 and 85, the second stem contains any values of 86 and 87, and so on. The digits to the right of the vertical line are known as the *leaves* of the plot, and each digit is known as a *leaf.* The main advantages of using a stem-and-leaf plot are that it shows the general shape of the data (like a bar chart or histogram), and that the all of the data can be recovered (to the nearest leaf unit).

2.6 Summary

Using the plots described in this section, we can gain an empirical understanding of the important features of the distribution of the data.

- Is the distribution *symmetric* or *asymmetric* about its central value?
- Are there any *unusual* or *outlying* observations, which are much *larger* or *smaller* than the main body of observations?
- Is the data *multi-modal*? That is, are there *gaps* or *multiple peaks* in the distribution of the data? Two peaks may imply that there are two different groups represented by the data.
- By putting plots side by side with the *same scale*, we may compare the distributions of different groups.

3 Summary measures

3.1 Measures of location

These are measures of the center of a distribution. They are single values that give a description of the data. They are also referred to as measure of central tendency. Some of them are arithmetic mean, geometric mean, harmonic mean, mode, and median. In addition to the graphical techniques encountered so far, it is often useful to obtain quantitative summaries of certain aspects of the data. Most simple summary measurements can be divided into two types; firstly quantities which are "typical" of the data, and secondly, quantities which summarize the variability of the data. The former are known as *measures of location* and the latter as *measures of spread*. Suppose we have a sample of size n of *quantitative* data. We will denote the measurements by $x_1, x_2, ...x_n$.

3.1.1 Mean

This is the most important and widely used measure of location. The *mean* of a set of data is

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

This is the location measure often used when talking about the *average* of a set of observations. If we have *discrete* quantitative data, tabulated in a frequency table, then if the possible outcomes are $x_1, ..., x_k$, and these occur with frequencies $f_1, ..., f_k$, so that $\sum f_i = n$, then the mean is

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + ... + x_k f_k}{n}$$

For *continuous* data, the sample mean should be calculated from the original data if this is known. However, if it is tabulated in a frequency table, and the original data is *not* known, then the mean can be *estimated* by assuming that all observations in a given interval occurred at the mid-point of that interval. So, if the mid-points of the intervals are $m_1, ..., m_k$, and the corresponding frequencies are $f_1, ..., f_k$, then mean can be approximated using

$$\overline{x} \simeq \frac{f_1 m_1 + f_2 m_2 + ... + f_k m_k}{n}$$

Calculation of mean from grouped data

If the items of a frequency distribution are classified in intervals, we make the assumption that every item in an interval has the mid-values of the interval and we use this midpoint for x.

Example. The table below shows the distribution of the waiting items for some customers in a certain petrol station in some city.

Waiting time (in mins)	1.5 – 1.9	2.0 – 2.4	2.5 – 2.9	3.0 – 3.4	3.5 – 3.9	4.0 – 4.4
No. of customers	3	10	18	10	7	2

Find the average waiting time of the customers.

Solution.

Waiting (in min)	No of customers	Class mark mid-value(x)	fx
1.5 – 1.9	3	1.7	5.1
2.0 – 2.4	10	2.2	22
2.5 – 2.9	18	2.7	48.6
3.0 – 3.4	10	3.2	32
3.5 – 3.9	7	3.7	25.9
4.0 – 4.4	2	4.2	8.4
Total	$\sum f = 50$		$\sum fx = 142$

$$\begin{aligned} \overline{x} &= \frac{\sum fx}{\sum f} \\ &= \frac{142}{50} = 2.84 \end{aligned}$$

Use of Assume mean

Sometimes, large values of the variable are involve in calculation of mean, in order to make our computation easier, we may assume one of the values as the mean. This if A= assumed mean, and d= deviation of from A, i.e. $d = x - A$. Then we have

$\overline{x} = \frac{\sum fx}{n} = \frac{\sum f(A+d)}{n} = \frac{A \sum f}{n} + \frac{\sum fd}{n} = A + \frac{\sum fd}{n}$

If a constant factor C is used then, $\overline{x} = A + \left(\frac{\sum fU}{\sum f}\right) C$ where $U = \frac{x-A}{C}$.

Advantage of mean

The mean is an average that considers all the observations in the data set. It is single and easy to compute and it is the most widely used average.

Disadvantage of mean

Its value is greatly affected by the extremely too large or too small observation.

3.1.2 The harmonic mean (H.M)

The H.M of a set of numbers $x_1, x_2, ..., x_n$ is the reciprocal of the arithmetic mean of the reciprocals of the numbers. It is used when dealing with the rates of the type x per d (such as kilometers per hour, Naira per liter). The formula is expressed thus:

H.M= $\frac{1}{\frac{1}{n}\sum \frac{1}{x_i}} = \frac{n}{\sum \frac{1}{x_i}}$

Example. Find the harmonic mean of $2, 4, 8, 11, 4$.

Solution. H.M= $\frac{n}{\sum \frac{1}{x_i}} = \frac{5}{\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{11}+\frac{1}{4}} = 4.112$

Note

1. Calculation takes into account every value.
2. Extreme values have least effect.
3. The formula breaks down when “0” is one of the observations.

3.1.3 The geometric mean (G.M)

The G.M is an analytical method of finding the average rate of growth or decline in the values of an item over a particular period of time. The geometric mean of a set of number $x_1, x_2, ..., x_n$ is the root of the product of the n^{th} number. Thus

G.M= $\sqrt[n]{x_1 \times x_2 \times ... \times x_n}$

Example. The rate of inflation in fire successive year in a country was $5\%, 8\%, 12\%, 25\%$ and 34%. What was the average rate of inflation per year?

Solution. G.M= $\sqrt[5]{1.05 \times 1.08 \times 1.12 \times 1.25 \times 1.34} = 1.16$
Therefore average rate of inflation is 16%.

Note

1. Calculate takes into account every value.

2. It cannot be computed when "0" is on of the observation.

Relation between Arithmetic mean, Geometric and Harmonic

In general, the geometric mean for a set of data is always less than or equal to the corresponding arithmetic mean but greater than or equal to the harmonic mean.

$$\text{That is, H.M}\leqslant\text{G.M}\leqslant\text{A.M}$$

The equality signs hold only if all the observations are identical.

3.1.4 Median

The *sample median* is the middle observation when the data are *ranked* in increasing order. We will denote the ranked observations $x_{(1)}, x_{(2)}, ..., x_{(n)}$. If there are an even number of observations, there is no middle number, and so the median is defined to be the sample mean of the middle two observations.

$$\text{SampleMedian=}\begin{cases} x_{\left(\frac{n+1}{2}\right)}, & n\,odd \\ \frac{1}{2}x_{\left(\frac{n}{2}\right)} + \frac{1}{2}x_{\left(\frac{n}{2}+1\right)}, & n\,even \end{cases}$$

The sample median is sometimes used in preference to the sample mean, particularly when the data is asymmetric, or contains outliers. However, its mathematical properties are less easy to determine than those of the sample mean, making the sample mean preferable for formal statistical analysis. The ranking of data and calculation of the median is usually done with a stem-and-leaf plot when working by hand. Of course, for large amounts of data, the median is calculated with the aid of a computer.

Calculation of Median from a grouped data

The formula for calculating the median from grouped data is defined as

$$Median = L + \left(\frac{n/2 - B}{f_m} \right) C.$$

Here $L =$ Lower class boundary of the median class,
$n =$ Total number of values,
$B =$ Cumulative frequency of the groups before the median group,
$f_m =$ Frequency of the median group,
$C =$ Class width of median class.

Example. Compute the median that table below shows the height of 70 men randomly selected at some city.

Height	118-126	127-135	136-144	145-153	154-162	163-171	172-180
Frequency	8	10	14	18	9	7	4

Solution. Median $= 144.5 + \left(\frac{35-32}{18}\right) 9 = 136$

Advantage of the median

1. Its value is not affected by extreme values; thus it is a resistant measure of central tendency.

2. It is a good measure of location in a skewed distribution

Disadvantage of the median

1. It does not take into consideration all the value of the variable.

3.1.5 Mode

The mode is the value which occurs with the greatest frequency. Consequently, it only really makes sense to calculate or use it with discrete data, or for continuous data with small grouping intervals and large sample sizes. For an example the mode of scores 2, 5, 2, 6, 7 is 2.

Calculation of mode from grouped data

From a grouped frequency distribution, the mode can be obtained from the formula.

$$Mode = L + \left(\frac{\triangle_1}{\triangle_1 + \triangle_2}\right) c$$

Where L =lower class boundary of the model class,
$\triangle_1$ = Difference between the frequency of the modal class and the class before it.
$\triangle_2$ = Difference between the frequency of the modal class and the class after it.
c = the class size model class

Example. For the table below, find the mode.

Class	11 – 20	21 – 30	31 – 40	41 - 50	51 – 60	61 – 70
frequency	6	20	12	10	9	9

Solution. $L = 20.5$, $\triangle_1 = 20 - 6 = 14$, $\triangle_2 = 20 - 12 = 8$.
Mode$= 20.5 + \left(\frac{14}{14+8}\right) 10 = 26.9$

Advantage of the mode

1. It is easy to calculate.

Disadvantage of the mode

1. It is not a unique measure of location.

2. It presents a misleading picture of the distribution.

3. It does not take into account all the available data.