20. STATISTICS

Statistics is a discipline that provides methods for making informed judgments based on information, facts or **data**. The type of data we collect will depend on the characteristic we are investigating. There are basically two types of data:

1. **Qualitative Data** – this type of data is non-numerical in nature. It comprises categories such as data on the type of transport, favourite subject and religious denomination.
2. **Quantitative Data** – this type of data is numerical in nature. It comprises measures such as distance, time, mass and scores on a test.

When we carry out an investigation, we collect data from a **population**. This is a well-defined collection of objects and may comprise all the students in a school or all the manufactured vehicles in a company, over a period.

Since it is always difficult and costly to collect data from the entire population, we select a **sample** to participate in our study. This is a group of objects taken from a population. If every member of the population has an equal chance of being selected, we have a ***random*** sample.

We are usually interested in a particular characteristic of the objects in a population. This characteristic whose value may change from one object to another, in a population is referred to as the **variable** under investigation. The value of the characteristic may be a category such as male or female (qualitative in nature) or a number such as a score on a test (quantitative in nature).

Discrete and continuous variables

Variables are categorised as **discrete** or **continuous**. **Discrete** variables take fixed values and are not necessarily whole numbers. Furthermore, they can be listed as a finite set or an infinite set. For example, a set of shoe sizes may take the form, $\{1, 1\frac{1}{2}, 2, 2\frac{1}{2}, ...12\}$. Between any two values, there is no value that belongs to the set. This is a property of a discrete data set. The most common form of discrete variables is frequency counts. For example, the number of children in a household {0,1,2,3,..}

Continuous variables

A variable is said to be continuous if it can take any value within a finite or infinite interval. Measures such as mass, time, length and temperature are some examples of continuous variables since they can take any value between any two given values. The heights of students in a school or the ages of the people of a town are examples of continuous variables. For example, in collecting data on heights of students, between any two heights, say, 56 and 57 cm, there are countless values that belong to the data set.

Raw Data

Raw data is defined as data that have just been collected and which remains in a disorganised form. In such a form it is difficult to observe any patterns or trends in the data. Hence, we can not make too many deductions from it, nor can we draw any feasible conclusions or inferences.

An example of raw data is the following record of goals scored by a football team during a season. The data is recorded as it comes starting with the first match and ending with the last match.

1 3 2 0 1 0 2 1 3 4 1 3 4 4 0
2 2 0 4 1 2 2 3 1 2 4 1 4 3 2

Frequency Tables

One of the simplest ways of organising raw data is in the form of **frequency tables.** A frequency table records the ***frequency*** or number of times each value occurs. There are usually two columns in the table. The first column lists the values of the variable while the second column records the frequency of each value.

If the variable has a small range, we may choose to list all the values in the first column. In such a table, the scores are **ungrouped**. The above data on goals scored by a football team has a small range, 0 – 4 and we can use the following frequency table to summarise the data.

No. of goals	Frequency
0	4
1	7
2	8
3	5
4	6
	30

Frequency Table for Ungrouped Data

If the scores are spread over a large range, it is best to group the scores so as to reduce the number of rows in the table. For example, the scores of 30 students in a test out of 50 are shown below:

1 31 21 12 10 48 25 11 30 9
29 33 39 38 34 7 25 15 22 13
36 20 23 36 18 24 45 22 40 6

Notice that the range is 1-50 and as such we cannot use an ungrouped frequency table which will require 50 rows. We must group the scores conveniently and so reduce the number of rows. In so doing, a more meaningful summary is presented. We refer to the table as a grouped frequency table.

Marks (X)	Frequency
1 – 10	5
11 – 20	6
21 – 30	9
31 – 40	8
41 – 50	2
	30

Frequency Table for Grouped Data

The choice of the table depends on the range of the variable and not the number of scores in the distribution. When we group data in this manner we no longer have the raw scores and this makes it impossible to obtain exact values of statistical indices such as measures of central tendency and dispersion from the information in the table. However, we can obtain reasonable estimates of these measures using the procedures outlined in the next section.

Class intervals for continuous data

In interpreting class intervals for continuous data we need to differentiate between the **class limits** and the **class boundaries**. The class limits specify the smallest and largest data values that fall within an interval. To understand the difference between class limits and class boundaries, consider the table below in which the heights of a number of seedlings were recorded.

The height of seedling, in cm, x		The number of seedlings. (frequency, f)
Class Limits	Class Boundaries	
6 – 10	$5.5 \le x < 10.5$	50
11 – 15	$10.5 \le x < 15.5$	75
16 – 20	$15.5 \le x < 20.5$	120
20 – 25	$20.5 \le x < 25.5$	240
26 – 30	$25.5 \le x < 30.5$	15
		$\sum f = 500$

The class limits 6-10, 11-15, 16-20, and so on do not accommodate a measure of 10.6 or 15.5. It treats the data as if it were discrete. Hence, we must create class boundaries, which take into consideration all possible data values without excluding any possible measures. So, for continuous data, we need to include the class boundaries to realistically capture the actual data. For discrete data, this is not necessary as all the data is accounted for.

For the class limit, 6 – 10, we refer to 6 as the lower-class limit (LCL) and 10 as the upper-class limit (UCL). In creating the class boundaries, we simply compute half of the smallest unit of measure - in this example, $\frac{1}{2}\ of\ 1\ cm = 0.5\ cm$. We then obtain our class boundaries as follows. $LCB = LCL - 0.5$

$$UCB = UCL + 0.5$$

where 5.5 is the lower-class boundary (LCB) and 10.5 is the upper-class boundary (UCB).

The **class width** is computed as follows:
Class width = UCB – LCB
= 10.5 – 5.5 = 5

The **mid-point** of a class interval is a useful measure when computing statistical indices from frequency tables. It is the median of the class interval and is

computed using either the class boundaries or the class limits.

The mid-point of the class interval

$$\frac{LCL + UCL}{2} = \frac{LCB + UCB}{2}$$

For example, for the interval, 11-15,

$$\text{Mid Point} = \frac{10.5+5.5}{2} = 8$$

OR

$$\text{Mid Point} = \frac{11+5}{2} = 8$$

Measures of central tendency

In statistics, we are often required to summarise a set of scores by obtaining a single score to represent them. Measures of central tendency are commonly used in statistics when analysing numerical data. A review of these measures is presented below.

Mean

The mean of a set of scores is the sum of a set of scores, divided by the number of scores in the set. When data is presented in raw form, we can calculate the exact value of the mean by adding up the scores and dividing the total by the number of scores.

The mean, $\overline{X}$ of a set of n scores, is $\overline{X} = \frac{\sum x}{n}$

where $\sum x$ represents the sum of all data values.

Example 1

Calculate the mean of the set of scores:
6, 7, 8, 10, 12, 13, 13, 15, 16, 20

Solution

The mean, $\overline{X}$, is calculated as

$$\overline{X} = \frac{6+7+8+10+12+13+13+15+16+20}{10} = 12$$

Mean from frequency tables - Ungrouped data

When data is presented in a frequency table, we can determine the mean without having to extract the raw data. Consider the frequency table mentioned earlier in this chapter, in which the number of goals scored by a team in 30 matches was recorded.

No. of goals (x)	Frequency	fx
0	4	0
1	7	7
2	8	16
3	5	15
4	6	24
	$\sum f = 30$	$\sum fx = 62$

The mean number of goals per match is calculated as follows:

$$\text{Mean} = \frac{\text{Total number of goals}}{\text{Total number of matches}} = \frac{\sum fx}{\sum f} = \frac{62}{30} = 2.1$$

Mean from grouped data

While we can reconstruct the raw data from ungrouped frequency tables, this cannot be done when the data is grouped. Consider the data in which the masses of peas were recorded.

Mass in grams	Mid-Point (x)	No. of peas (f)	$f \times x$
3 – 7	5	3	15
8 – 12	10	8	80
13 – 17	15	12	180
18 – 22	20	10	200
23 – 27	25	7	175
		40	650

If we do not have the raw scores in each interval, we need to estimate their values in each interval. This is done by assuming, on average, that the set of values in an interval will approximate to the midpoint of the interval. Consider the first interval where the masses range from 3 to 7 grams. The mid-point is 5 grams and it is reasonable to assume that some of the seedlings will be under 5 grams and others will be over 5 grams. We estimate that all the values are at the mid-point, which is 5 grams. We continue this process for the other intervals.

We now treat the midpoint as the variable score, x, and proceed to use the same formula as was done for ungrouped data. The computations are shown in the table.

Mean mass, $\overline{X} = \frac{\sum fx}{\sum f} = \frac{650}{40} = 16.3$ grams

It should be noted that this method would give rise to only an estimate of the mean.

The mean, $\overline{X}$ from a frequency table, is $\overline{X} = \frac{\sum fx}{\sum f}$ where x represents the variable and f the frequency of x.

Mode

The mode is 'the most frequent' occurrence in a distribution. For categorical data, it is the category that occurs the most. For numerical data, it is the value that occurs the most.

It is possible to have more than one mode, and it is possible to have no mode. The mode of a distribution is determined by inspection. No computation is required. When data is organised in a frequency table, the mode is the score or group with the highest frequency.

For the scores 0, 1, 1, 2, 2, 2, 3, 4, 4, 5, the mode is clearly 2.

For the data below,

Score (x)	0	1	2	3	4	5
Frequency	20	15	18	14	25	17

The mode is 4, as it has the highest frequency.

Median

The median of a set of scores is the middle value when the scores are arranged in either ascending or descending numerical order. If there are two middle values, then the median is the mean of the middle two values.

To determine the median of the following set of scores:

4, 5, 7, 3, 9, 0, 6, 8, 7

We first arrange the scores in either ascending or descending order.

0, 3, 4, 5, 6, 7, 7, 8, 9.

The median is 6, the middle score.

If two scores are in the middle and we must compute the mean of the two middle scores

8, 12, 10, 15, 16, 16, 17, 18, 19, 19, 19, 23

Median = $\frac{16+17}{2} = 16.5$

To calculate the median from grouped data, a cumulative frequency curve must be drawn. This is done in a later section of this chapter.

Measures of spread or dispersion

The variation between values in a distribution is called spread or dispersion. If a data sample has all its values close together, there is little variability and we say that the scores are homogenous. If the scores are widely dispersed, there will be considerable variability and we say that the scores are non-homogenous.

Range

The range of a data set is a measure of the spread (or dispersion) of the observations.

The range is the difference between the largest and the smallest observed value in a distribution.

For the data set: 75, 73, 88, 56, 73, 52, 11, 56

The range is 88 - 11 = 77.

For data in a frequency table, the smallest and largest scores are taken at the mid-points of the two end intervals.

Mass in grams	3 – 7	8 – 12	13 – 17	18 – 22	23 – 27
No. of peas	3	8	12	10	7

Mid-point of the lowest interval, 3-7 is 5.

Mid-point of the highest interval, 23-27 is 25.

Hence, the range is: 25 – 5 = 20 grams.

The range ignores the remaining data in a distribution and considers only the two end scores. It is greatly influenced by the presence of just one unusually large or small score in the sample. It is, therefore, a crude measure of spread and as such, we need other measures to compute dispersion.

Quartiles

Quartiles divide the set of scores into four equal sets. The three quartiles are called the lower quartile, denoted by Q_1, the middle quartile, or the median Q_2 and the upper quartile denoted by Q_3. To determine the quartiles for a given set of scores, we first divide the set into two equal parts then further divide each half into two equal parts.

Consider the eleven raw scores shown below:

1 2 2 3 3 4 6 6 7 8 9
↑ ↑ ↑
Q_1 Q_2 Q_3

The score 4 divides the set into two equal parts with five scores on each side, so 4 is the median, Q_2.
The score 2 divides the lower set into two equal parts with two scores on each side, so 2 is the lower quartile, Q_1.
The score 7 divides the upper set into two equal parts with two scores on each side, so 7 is the upper quartile, Q_3.

Notice in this example, the quartiles all belong to the data set. This is not always the case as we can see in the following example.

Consider the set of 8 scores:

2 2 3 3 4 6 6 7
↑ ↑ ↑
Q_1 Q_2 Q_3
2.5 3.5 6

In this set, there is no single score in the middle, so the median is the average of the two middle scores, 3 and 4. The median Q_2, is 3.5.
The lower set has 4 scores, so there are two middle scores 2 and 3. Hence, the lower quartile, Q_1 is 2.5.
The upper set has 4 scores, so there are two middle scores 6 and 6. Hence, the upper quartile, Q_3 is 6.

Inter-quartile Range

The inter-quartile range is a measure of the spread or dispersion, within a data set. The Inter Quartile Range (IQR) is the width of an interval that contains the middle 50% of the sample. It is calculated by taking the difference between the **upper** and the **lower** quartiles.

Inter-Quartile Range $= Q_3 - Q_1$

The **Semi Inter Quartile Range** is one half of the IQR.

Semi Inter-Quartile Range $= \frac{1}{2}(Q_3 - Q_1)$.

The IQR is not affected by extreme scores and is a better measure of dispersion than the range.

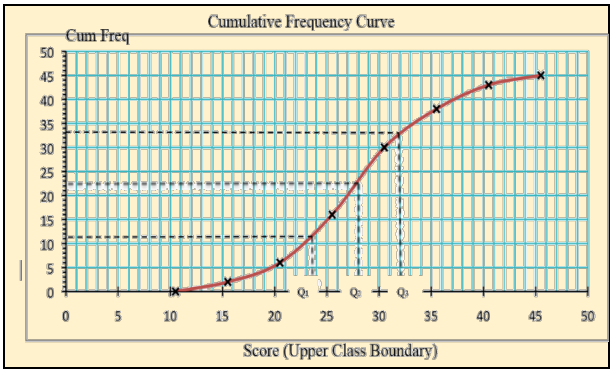
Quartiles from cumulative frequency curve

To determine the quartiles from a grouped frequency table, we use a cumulative frequency curve. Consider the data on scores of students in a test out of 45 shown in the table below.

Marks	Upper class boundary	Frequency	*Cumulative frequency
6 – 10	10.5	0	0
11 – 15	15.5	2	2
16 – 20	20.5	4	6
21 – 25	25.5	10	16
25 – 30	30.5	14	30
31 – 35	35.5	8	38
36 – 40	40.5	5	43
41 – 45	45.5	2	45
		45	

To determine the quartiles, we use the following steps:

1. Determine the Upper Class Boundaries.
2. Calculate the cumulative frequencies - the Cumulative frequency of a particular score, x is the number of scores less than or equal to x.
3. Plot the cumulative frequency against the upper class boundary of each interval and join the points with a smooth curve. The cumulative frequency curve is shown below.



4. On the cumulative frequency axis, locate the half-mark position, in this case, it will be at ½ (45) = 22.5. Draw a horizontal line to meet the curve and then a vertical line to meet the *x*-axis. Read-off Q_2 on this axis.
5. Locate the lower quartile at ¼ (45)= 11.25. Draw a horizontal line to meet the curve and read off Q_1 on this axis.
6. Locate the upper quartile at ¾ (45)= 33.75. Draw a horizontal line to meet the curve and read off Q_3 on this axis.

Interpretation of the median and the quartiles

A median score of 28 indicates that one half of the number of students in the class scored below 28 on the test.

The lower quartile of 23.5 means that one-quarter of the students scored below 23.5 on the test.

The upper quartile of 32 means that, three-quarters of the students scored below 32 on the test.

We can also use a cumulative frequency curve to find out**:**

1. The number of students scoring below or above any target score.
2. A specific score that divides the group into two parts. For example, if the top ten students qualify for the next round in the competition, what is the pass mark to qualify?

For example,

a) How many students scored below 20 marks on the test?
b) What is the pass mark if 10 students qualify for the next round?

To answer part (a), we draw a vertical line starting at 20 marks to meet the curve and a horizontal line to meet the *y*-axis. The frequency of 5 indicates that 5 students scored below 20 marks on the test.

To answer part (b), we must think of the number of students who will not qualify for the next round. This is calculated as 45 – 10 = 35 students.

We must then draw a horizontal line at 35 from the *y*-axis until it meets the curve, then draw a vertical line from this point to meet the *x*-axis. This line cuts the axis at 33 marks. Hence, students scoring over 33 marks will qualify for the next round.

Variance and Standard Deviation

The standard deviation is the best measure of spread and is used extensively in statistics. The variance, s^2 is the square of the standard deviation, *s*. If data is in raw form, the variance and standard deviation are obtained by using the formulae:

$$S = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} \quad \text{and} \quad S^2 = \frac{\sum(x_i - \bar{x})^2}{n}, \text{ where}$$

n is the number of values

x_i – represents the value of the variable, *i* of the sample,

$\bar{x}$ is the mean $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Standard deviation from raw data

Consider the set of raw scores obtained from marks on a quiz out of 10.

x_i = 4, 0, 6, 8, 5,7. $n = 6$

1. Calculate the mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{4+0+6+8+5+7}{6} = 5$$

2. Calculate the deviations from the mean, $x - \bar{x}$. See column 2 in the table below.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
0	(0 – 5)	25
4	(4 – 5)	1
5	(5 – 5)	0
6	(6 – 5)	1
7	(7 – 5)	4
8	(8 – 5)	9
		$\sum(x_i - \bar{x})^2 = 40$

3. Square the deviations and find their sum. See column 3.
4. Compute the variance

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{40}{6} = 6.7 \quad \text{(to 1 d.p)}$$

5. Compute the standard deviation

$$S = \sqrt{\frac{40}{6}} = 2.6 \text{ (to 1 d.p).}$$

A standard deviation of 2.6 indicates that on the average the scores differed by 2.6 units from the mean. These scores can be considered as fairly homogenous since the standard deviation is relatively small.

The above computation illustrates that the standard deviation actually measures the average deviation from the mean. It has the advantage over other measures of spread in that it considers each data value.

Standard deviation from frequency distributions (ungrouped data)

In frequency distributions, calculation of the standard deviation involves the same basic steps except that the frequency of each score must be considered when computing the deviations. The formula for variance and standard deviation becomes:

$$S^2 = \frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n} \text{ and } S = \sqrt{\frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n}} \text{ where}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} fx_i}{\sum_{i=1}^{n} f_i}$$

Consider the data below, which represents scores of 60 students on a quiz out of 10.

Score (x_i)	4	5	6	7	8
Frequency (f)	10	8	13	10	9

To facilitate ease of computation, the above data is presented in a vertical form in the first two columns.

x_i	f	$f_i x_i$	$x - \bar{x}$	$(x - \bar{x})^2$	$f(x - \bar{x})^2$
4	10	40	(4 −6)	4	40
5	8	40	(5 −6)	1	8
6	13	78	(6 −6)	0	0
7	10	70	(7 −6)	1	10
8	9	72	(8 −6)	4	36
	$\sum_{i=1}^{n} f_i$ =50	$\sum_{i=1}^{n} f_i x_i$ =300			$\sum_{i=1}^{n} f(x_i - \bar{x})^2$ =94

We use the following steps to fill out the rest of the table column by column.

1. Compute the mean, using the formula
$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{300}{50} = 6$$
2. Calculate the deviations from the mean, $x - \bar{x}$.
3. Square the deviations, $(x - \bar{x})^2$.
4. Multiply the squared deviations by the frequencies.
5. Compute the variance and standard deviation using
$$S^2 = \frac{\sum_{i=1}^{n} f(x_i - \bar{x})^2}{n} = \frac{94}{50} = 1.88$$
$$S = \sqrt{\frac{94}{50}} = 1.37$$

Alternate formula for standard deviation

The variance and standard deviation can also be computed from the following formulae:

$$S^2 = \frac{\sum_{1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - (\bar{x})^2 \quad \text{and } S = \sqrt{\frac{\sum_{1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - (\bar{x})^2}$$

Using this formula, we compute the standard deviation for the data in the above example.

x_i	f_i	$f_i x_i$	x_i^2	$f_i x_i^2$
4	10	40	16	160
5	8	40	25	200
6	13	78	36	468
7	10	70	49	490
8	9	72	64	576
	$\sum_{i=1}^{n} f_i$ =50	$\sum_{i=1}^{n} f_i x_i$ =300		$\sum_{1}^{n} f_i x_i^2$ =1894

1. Compute the mean as before, $\bar{x} = 6$
2. Compute $f x_i$, see column 3 above.
3. Square the x values, see column 4 above.
4. Compute $f_i x_i^2$, see column 5 above.
5. Compute the variance and standard deviation using

$$S^2 = \frac{\sum_{1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - (\bar{x})^2 = \frac{1894}{50} - (6)^2 = 37.88 - 36 = 1.88$$

$$S = \sqrt{1.88} = 1.37$$

Standard deviation from frequency distributions (grouped data)

When we have data in grouped frequency distributions, we can compute the standard deviation using the same procedure as ungrouped data with one small adjustment. In previous sections when computing the mean from grouped data, we used the midpoint of the interval for the score. We do the same when computing standard deviation.

The following table gives the marks obtained in a test by 70 students. We will estimate the variance and standard deviation using the alternate formula given above.

Marks	Mid-point (x_i)	f	$f_i x_i$	x_i^2	fx_i^2
$0 \le x < 10$	5	6	30	25	150
$10 \le x < 20$	15	16	240	225	3600
$20 \le x < 30$	25	24	600	625	15000
$30 \le x < 40$	35	17	595	1225	20825
$40 \le x < 50$	45	7	315	2025	14175
		$\sum_{i=1}^{n} f_i$ =70	1780		$\sum_{1}^{n} f_i x_i^2$ =53750

1. Compute the mean,

$$\bar{x} = \frac{\sum_{i=1}^{n} fx_i}{\sum_{i=1}^{n} f_i} = \frac{1780}{70} = 25.4$$

2. Square the x values then multiply these by the frequencies (Columns 5 and 6)

3. Compute the variance and standard deviation by substituting values from the table into the following formulae.

$$S^2 = \frac{\sum_{1}^{n} f_i x_i^2}{\sum_{i=1}^{n} f_i} - (\bar{x})^2$$
$$= \frac{53750}{70} - \left(\frac{1780}{70}\right)^2$$
$$= 767.86 - 646.6$$
$$= 121.26$$

$$S = \sqrt{121.26} = 11.01$$

Interpretation of standard deviation

The standard deviation can be interpreted as a measure of homogeneity of a group. It can tell us how close or how different the group members are with respect to a given variable. It is basically a measure of the spread of a set of data.

When the standard deviation is small, the group is said to be homogenous with respect to the variable measured; that is, all the scores are close to the mean. The larger the standard deviation, then further away are the scores from the mean. Thus, the standard deviation is a useful tool when comparing groups on the variability of scores.

We can also use standard deviation to compare two or more individuals' performance on a variable. Suppose we know that the mean scores of two batsmen over five innings were both 50. We want to select the better one for a new game. To make a fair judgment, we use both the mean and the standard deviation.

Assume that we had their raw scores and we calculated the standard deviation for each batsman as shown.

	Batsman 1	Batsman 2
Mean	50	50
Standard Deviation	2.3	10.5

The smaller standard deviation for batsman 1 reveals that the scores over the 5 innings were close to the mean, indicating that batsman 1 was consistent in performance. Batsman 2 would have had scores that varied considerably. Having a mixture of high and

low scores suggests that he is not consistent in his performance. Hence, there will be more at risk involved in choosing Batsman 2.

Percentiles

Another useful statistic that is widely used to compare the performance of a group on any variable is percentiles. Percentiles are useful for giving the relative standing of an individual in a population. They express the rank position of an individual.

Percentiles are like quartiles, except that they divide the data set into 100 equal parts instead of four equal parts. A **percentile rank** is the percentage of scores that fall at or below a given score. For example, if Sammy took the Secondary Entrance Assessment and scored at the 96th percentile, this means that he scored as high as or higher than 96% of the candidates who wrote the same test.

The following should be noted about percentile ranks:

1. The lowest score is the 1st percentile since there is no zero rank.
2. The highest score is the 99th percentile since there is no rank of 100.
3. Percentile ranks are written to the nearest whole number.
4. The median is the same as the 50th percentile.
5. The lower quartile, Q_1 is the same as the 25th percentile.
6. The upper quartile, Q_2 is the same as the 75th percentile.

Calculating percentiles from raw scores

To calculate percentile rank, we must first sort the data in ascending order. The percentile of a score is the percentage of scores that fall ***at or below*** the given score.

If we have a small number of scores and there are not many repeated scores, we can calculate the percentile rank (PR) from the formula:

$PR = \frac{S_{AB}}{n} \times 100$ where S_{AB} is the number of scores at or below the desired score and n is the number of scores.

Example 2

Find the percentile rank of the score 25 given the set of scores
16, 25, 20, 23, 14, 12, 17, 30, 28, 20, 19, 13

Solution

1. First, we must rank the scores from lowest to highest:
   12, 13, 14, 16, 17, 19, 20, 20, 23, 25, 28, 30
2. Calculate S_{AB} the number of scores at or below the desired score.

By counting, there are 10 scores at or below 25, $S_{AB} = 10$

3. Calculate the percentile rank, PR, using the formula.

$$PR = \frac{S_{AB}}{n} \times 100 = \frac{10}{12} \times 100 = 83.3$$

So the score of 25 is at the 83rd percentile. This means that 83% of the scores are lower than or equal to 25.

Estimating the value of a score (S_{AB}) given the percentile rank

It is also possible to estimate the position and actual value of a score in a data set given its percentile rank.

By re-arranging this formula $PR = \frac{S_{AB}}{n} \times 100$ to calculate S_{AB}, the number of scores below the desired score, we obtain $S_{AB} = \frac{PR}{100} \times n$

Example 3

The mathematics test scores for a group of 20 students are shown below.

50	65	70	72	72
78	80	82	84	84
85	86	88	88	90
94	96	98	98	99

i. Determine the percentile rank for a score of 84 on this test.
ii. Determine the score whose percentile rank is 30.

Solution

i The number of scores at or below 84 is 10.
Using the formula for percentile rank,
we substitute, $S_{AB} = 10, n = 20$

$$PR = \left(\frac{S_{AB}}{n}\right) \times 100$$

$$= \frac{10}{20} \times 100 = 50^{th} \text{ percentile}$$

The score of 84 is at the 50th percentile.

ii To find the position of the score whose percentile is 30. First, use S_{AB} to find the number of scores below the desired score.

$$S_{AB} = \left(\frac{PR}{100} \times n\right) \Rightarrow S_{AB} = \frac{30}{100} \times 20 = 6$$

This means that there are 6 scores at or below the desired score. We examine the raw scores and note that the 6th score is 78 and the 7th score is 80. Hence, 80 is the desired score.

Example 4

Neil placed 18th on a competition taken by a group of 90 students. Calculate his percentile rank. Javed placed 24th in the same competition taken by a group of 150 students, calculate his percentile rank.
With respect to their group, who ranked higher, Neil or Javed?

Solution

The number of students whose score is at the same, or below Neil's score is 90-18 = 72.
His percentile rank is therefore

$$\frac{72}{90} \times 100 = 80^{th} \text{ percentile.}$$

The number of students whose score is at, or below Javed's score is 150-24 = 126
His percentile rank is therefore

$$\frac{126}{150} \times 100 = 84^{th} \text{ percentile.}$$

Therefore, Javed did better than Neil since he scored better than 84% of the group while Neil scored better than 80% of the group.

Data Displays

In addition to frequency tables, data sets can be represented using a variety of visual techniques. Some of these are Pie Charts, Bar Charts, Histograms and Frequency Polygons. In this section we will focus on two other techniques of displaying statistical data, these are stem and leaf diagrams and box and whisker plots.

Stem and Leaf Diagrams

A quick way to obtain an informative and visual representation of numerical data is to construct a stem and leaf plot. The display summarises the data in such a form that one could quickly obtain the numerical raw scores in an ordered fashion. Essentially, it splits the digits of the number into two parts, a leaf and a stem, then, separates each leaf from its stem by using two columns. For example, if the data set consists of two-digit numbers a score of 45 can have a stem of 4 (tens digit) and a leaf of 5. The choice of digits for the stem and leaf depend on the value of the numbers in the data set.

Constructing a stem and leaf plot

Consider the following scores of students in a test out of 100.

50, 96, 72, 83, 92, 98, 85, 74, 64, 50, 58,
58, 64, 67, 68, 52, 61, 76, 78, 79, 78, 92

To construct a stem and leaf diagram, we use the following steps:

1. Select the leading digit(s) for the stems and the trailing digits for the leaves. In this case, we choose the tens-digit as the stems: 5, 6, 7, 8 or 9. The ones-digit will form the leaves.
2. Order the numbers in rows in ascending or descending order. The numbers in each row should have the same stem.
   92, 92, 96, 98
   83, 85
   72, 74, 76, 78, 78, 79
   61, 64, 64, 67, 68
   50, 50, 52, 58, 58
3. List the stem values in a vertical column in order.
4. Record the leaf for every observation next to the corresponding stem value. Leave equal intervals between the numbers.

Stem	Leaf
9	2 2 6 8
8	3 5
7	2 4 6 8 8 9
6	1 4 4 7 8
5	0 0 2 8 8

5. State the Key

Key 6|4 = 64
Stem unit = 10.0
Leaf unit = 1.0

From the diagram, we can easily deduce the following:

- The majority of students got marks in the 70's.
- The lowest mark is 50 and the highest mark is 98.
- The number of students who took the test is the number of leaves.

Displaying two sets of data

To compare two sets of data, we may use a back to back stem and leaf plot. The stems are listed in the middle of the display. One set of leaves is placed on the left and the other set is placed on the right.

The scores obtained by two batsmen in ten innings are as follows:

Batsman 1: 30, 33, 35, 37, 39, 41, 42, 48, 51, 53

Batsman 2: 32, 32, 43, 45, 45, 54, 56, 58, 58, 59

Batsman 1		Batsman 2
Leaf	Stem	Leaf
9 7 5 3 0	3	2 2
8 2 1	4	3 5 5
3 1	5	4 6 8 8 9

Key: Leaf 3|2 = 32
Stem unit = 10
Leaf unit = 1.0

We can deduce the following:

- Both batsmen played 10 games each.
- Batsman 2 had most of his scores in the 50's while Batsman 1 had most of his scores in the 30's.

Example 5

The stem and leaf diagram displays the daily temperatures in degrees Fahrenheit for the month of June in a certain year.

Stem	Leaf
5	0 1 7 7 9
6	1 2 2 4 5 5 7 8 9
7	0 1 1 3 6 7 7 9 9
8	0 0 0 2 2 3 7

Key: 5|0 = 50
Stem unit = 10, Leaf unit = 1.0

(i) Determine the median and quartiles of the distribution.

(ii) Comment on the spread of the distribution.

Solution

(i) There are 30 leaves, indicating that the number of observations is 30. The median will lie between the 15th and 16th values.

The 15th observation is 70

The 16th observation is 71

$$\text{The median} = \frac{70+71}{2} = 70.5$$

The lower quartile is the 8th score in the distribution (the middle of the lower set of 15 observations), which is 62.

The upper quartile is the 22nd score in the distribution (the middle of the upper set of 15 observations), which is 79.

(ii)The distribution is almost symmetrical, with the majority of scores in the middle range.

Advantages of stem and leaf diagrams

1. In stem and leaf diagrams, the value of each individual data point can be easily recovered from the display. Hence, the raw scores are not lost.
2. The data is arranged compactly and the stem is **not** repeated for multiple data values.
3. The shape of the distribution is readily observed.
4. The information provided is similar to data obtained from a histogram.
5. The display can easily identify the mode and any outliers (unusually high or unusually low values in the data).
6. It is easy to construct.

Disadvantages of stem and leaf diagrams

1. If there is a large number of observations it becomes tedious to construct because every single item of data is to be written on the diagram.
2. It cannot be used for categorical data.

3. Ideally, it requires the leaves to be ordered on each stem, and this requires much re-organising of the data.
4. When the range of values is very large, there will be many stems with no leaves. In this case, it is best not to use this type of representation.
5. When the values vary in magnitude from tens to hundreds to thousands, and so on, this representation is not ideal.

Box and whisker plots

Box-and-whisker plots provide a pictorial summary of the five most prominent features of a data set. These features are:

the smallest value
the lower quartiles
the median,
the upper quartile
the largest value.

This display combines statistical indices with graphical displays and is, therefore, a powerful tool for analysing data.

Constructing a box and whisker plot

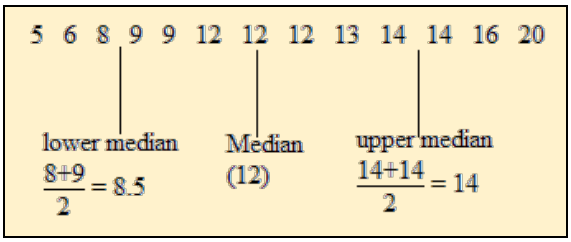
We use the data below to illustrate how a box and whisker plot is constructed.

At a plant nursery, the height of 13 seedlings in a box, are measured in centimetres, and recorded as:
12, 13, 5, 8, 9, 20, 16, 14, 14, 6, 9, 12, 12

Steps for constructing a box and whisker plot.

1. Order the data in ascending order of magnitude.
   5, 6, 8, 9, 9, 12, 12, 12, 13, 14, 14, 16, 20

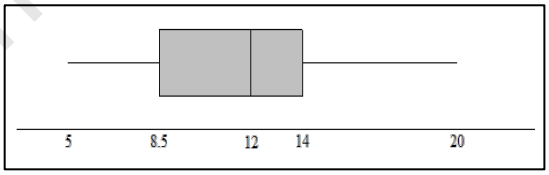
2. Determine the median, lower median (first quartile or lower fourth) and upper median (third quartile or upper fourth).



3. Draw a number line to accommodate all the values in the data set, starting with the smallest and ending with the largest.

4. Draw a rectangle above the number line so that its horizontal length is the distance between the lower and upper median. A vertical line is drawn inside of the rectangle at the median. The rectangle encloses all scores in the interquartile range.

5. Draw the whiskers (horizontal lines) from the vertical endpoints to the lowest and largest values of the data.



Interpretation of box and whisker plots

1. Data range - From the box and whisker plot we can deduce that the seedlings were between 5 cm and 20 cm in height. Hence, the range of the data = 20 – 5 = 15 cm.
2. The median is 12 cm.
3. The plot clearly illustrates the separation of the data values into four equal parts. In other words, it shows that
   - $\frac{1}{4}$ of the seedlings are between 5 cm and 8.5 cm,
   - $\frac{1}{4}$ of the seedlings are between 8.5 cm and 12 cm,
   - $\frac{1}{4}$ of the seedlings are between 12 cm and 14 cm,
   - $\frac{1}{4}$ of the seedlings are between 14 cm and 20 cm.

4. The shaded region shows where $\frac{1}{2}$ of the number of seedlings lie. Half of the seedlings lie between 8.5 cm and 14 cm.

Interquartile range from box and whisker plots

The inter-quartile range (IQR) is a measure of spread and the box and whisker plot visually displays this measure as the length of the box plot. In the above example we can deduce the following:

$IQR = Q_3 - Q_1 = 14 - 8.5 = 5.5$

The semi-inter-quartile range is one half of the interquartile range

$= \frac{1}{2}(Q_3 - Q_1) = \frac{1}{2}(5.5) = 2.75$

Advantages of box and whisker plots

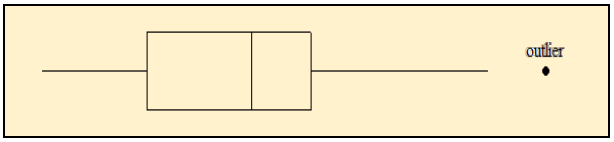
1. Box and whisker plots are ideal for graphically displaying the median, the IQR and hence the variability of a distribution.
2. They are especially useful for indicating whether a distribution is symmetrical or skewed.
3. Box and whisker plots have an advantage over stem and leaf plots in that they can accommodate distributions that have large numbers of observations.
4. They can readily show if there are potential unusual observations (outliers) in the data set.
5. Box-and-whisker plots are useful for comparing multiple sets of data because these diagrams use the same number line for showing five data features- lowest and highest values, lower and upper quartile and median.

Disadvantages of box and whisker plots

1. It does not show every value in a distribution as a stem and leaf plot does.
2. The raw data is not obtainable from a box and whisker plot.

Outliers and Extreme values

When collecting data, we sometimes obtain a result that is unusual in the sense that its value is 'far' from the others. Such values are called outliers. On a box and whisker plot, outliers should be ignored and not plotted on the whisker position of the diagram. However, we may choose to plot them individually and label them as outliers. This is illustrated in the diagram below.



Example 6

Draw the box and whisker plot. Indicate if there are any outliers on your plot.

21, 23, 24, 25, 29, 33, 49

Solution

1. We can easily spot the median and quartiles,

   $Q_2 = 25$ $\quad Q_1 = 23$

   $Q_3 = 33$
2. The $IQR = 33 - 23 = 10$
3. The box and whisker plot is shown below. The arrow shows the approximate position of the outlier, 49.