DATA HANDLING

Data is the collection of information for a specific purpose.

When data is collected by counting, it is called **discrete data** .
An example of discrete data is, e.g. shoe sizes : A person wears a number 5 shoe, or $5\frac{1}{2}$ or 6 or.....There are no shoe sizes in between these numbers.

Data that is collected by measurement is called **continuous data** . If your height is 1,5 metres, it does not imply that it is exactly 1,5 metres, it may be 1,531 metres, but rounded off to one decimal digit it is 1,5 metres.

1. Ordening data

Data must always be ordered before any deductions can be made. Data can be ordered by arranging the data from low to high.

15 Learners achieved the following percentages in a test :

55, 68, 54, 77, 88, 71, 68, 51, 92, 48, 80, 75, 61, 70, 69.

The data arranged from low to high will be:

48, 51, 54, 55, 61, 68, 68, 69, 70, 71, 75, 77, 80, 88, 92.

Stem and leaf diagrams

Data can also be ordered by a stem and leaf diagram. The tens digits form the stem and the units digits form the leaves, e.g. for the number 26, 2 will be the stem and 6 the leaf, and for the number 135, 13 will be the stem and 5 the leaf. The stem is written on the left-hand side of a vertical line and the leaves on the right-hand side.
The leaves are then ordered numerically.

Example 1

The following are the scores of a rugby team for 24 of their matches :

15	22	12	31	41	17	28	16	25	36	42	47
12	34	44	14	19	21	7	18	26	24	35	13

Order the data using a stem and leaf diagram.

Solution

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

Now order the leaves numerically.

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

Now it is easy to order the data from low to high.

7 12 12 13 14 14 16 17 18 19 21 22

24 25 26 28 31 34 35 36 41 42 44 47

2. Measures of central tendency

2.1 Ungrouped data

An example of ungrouped data is the test marks of 20 learners. The data is not grouped into intervals; instead the learners' individual test marks are used.

2.1.1 **The mean** ((the average) is calculated by dividing the total of the values by the number of the values. The mean is indicated by $\bar{x}$ (pronounced x bar) and the following formula is used :

$$\boxed{\bar{x} = \frac{\sum x}{n} = \frac{\textit{The sum of all the observations } (\sum x)}{\textit{The number of observations } (x)}}$$

2.1.2 **The mode** is the number that occurs most often. When there are two numbers that occur an equal number of times, we say the data is bimodal. If there are more than two modes, then the data has no mode.

2.1.3 **The median**, the middle value, divides the data into two parts, 50% of the data lies below the median and 50% of the data lies above the median.

Example 2

The following table shows the mass in kilogram of 12 Grade 10 learners.

51	48	55	41	61	58	62	53	66	59	46	51

2.1 Find the mean.
2.2 What is the mode ?
2.3 Calculate the median.

Solution

Data must always be ordered first. Thus, first order the data from low to high.

41	46	48	51	51	53	55	58	59	61	62	66

2.1 $\textit{The mean } (\bar{x}) = \dfrac{\textit{The sum of all the observations}}{\textit{The number of observations}}$

$$= \frac{41+46+48+51+51+53+55+58+59+61+62+66}{12}$$

$$= \underline{54,25}$$

2.2 *The mode* = $\underline{51}$ [*51 occurs twice*]

The median can be determined in two ways

2.3 *Method* 1 (*Counting the data*)

The median is the middle value of the ordered data. There are 12 *learners, therefore, the median lies between* 53(*the 6th value*) *and* 55 (*the 7th value*)

41 46 48 51 51 53 ↓ 55 58 59 61 62 66

and is thus the average of the 6th and the 7th value $\therefore \dfrac{53+55}{2} = 54$

$\therefore$ *The median* = 54

Method 2 (*Using the formula*)

Position of the median $= \frac{1}{2}(n+1) \quad = \frac{1}{2}(12+1) = 6\frac{1}{2}$

Remember : $6\frac{1}{2}$ *is only the position of the median and not the value of the median.*
The median thus lies between the sixth and seventh number.

The sixth number = 53 *and the seventh number* = 55

$\therefore$ *The median* $= \dfrac{53+55}{2} = \underline{54}$

2.2 Grouped data

When we work with a large amount of data, the data is grouped in class intervals.

2.2.1 **The mean** : To find the mean of grouped data, the mean (midpoint) of each class interval is calculated. This **midpoint of the class interval** (x) is multiplied by the **frequency**(f) of the class interval. The sum of this values (midpoint × frequency) is then divided by the number of observations.

The mean of grouped data is calculated using the following formula

$$\boxed{\bar{x} = \frac{\sum fx}{n} = \frac{\textit{The sum of (midpoint of class interval} \times \textit{frequency)}}{\textit{total number of observations}}}$$

2.2.2 **The mode** or modal class is the class interval with the highest frequency.

2.2.3 **The median** : The **position** of the median is $\frac{1}{2}n$ where n = the number of observations.

Example 3

The following table shows the mass of 50 Grade 10 girls.

Class interval	Midpoint of interval(x)	Frequency(f)	(x) × (f)
$35 \le x < 40$	37,5	1	37,5
$40 \le x < 45$	42,5	13	552,5
$45 \le x < 50$	47,5	18	855
$50 \le x < 55$	52,5	11	577,5
$55 \le x < 60$	57,5	5	287,5
$60 \le x < 65$	62,5	2	125
		Sum of (f)(x)	2435

3.1 Find the mean.
3.2 What is the modal class?
3.3 Identify in which interval the median lies.

Solution

$$\textit{Mean} = \frac{\textit{The sum of } (x \times f)}{\textit{total number of observations}}$$

$$= \frac{37,5 + 552,5 + 855 + 577,5 + 278,5 + 125}{50}$$

$$= \frac{2435}{50} \qquad = \underline{48,7}$$

The Modal class = *the interval* $45 \le x < 50.$

The Median : *Position of the median* $= \frac{1}{2}n = \frac{1}{2}(50) = 25$

The 25th value occurs in the class interval [45;50)

$\therefore$ *Median lies in the interval* [45; 50)

3. Measures of dispersion

The median divides the data into two parts. The data can, however, also be divided into **quartiles** , where 25% of the data lies below the first quartile and 75% of the data lies below the third quartile.

3.1 Calculating measures of dispersion of ungrouped data

3.1.1 **The first (lower) quartile,** Q_1: Position of **first quartile,** Q_1, is $\frac{1}{4}(n+1)$.

3.1.2 **The third (upper) quartile,** Q_3: Position of **third quartile,** Q_3, is $\frac{3}{4}(n+1)$.

3.1.3 **The range** = largest value – smallest value.

3.1.4 **The interquartile range** $= Q_3 - Q_1$.

Exept for quartiles, data can also be divided into **percentiles**. Percentiles divide the data into **hundredths**. For example, 10% of the data will lie under the 10th percentile. Calculate percentiles the same way as quartiles. Therefore, the position of the 10th percentile will be : $\frac{1}{10}(n+1)$.

Example 4

Look again at the data in example 2

41	46	48	51	51	53	55	58	59	61	62	66

Determine

4.1 the first quartile
4.2 the third quartile
4.3 the range
4.4 the interquartile range.

Solution

Like the median, the quartiles can also be determined in two ways

4.1 *Method* 1 (*Counting the data*)

41 46 48 ↓ 51 51 53 ↓ 55 58 59 ↓ 61 62 66

The median divides the data into two equal parts.

The first quartile divides the first half of the data into two equal parts.
To the left of the median there are 6 *numbers.*

Therefore, the first quartile lies between the 3*rd and the* 4*th number.*

$$\therefore\ Q_1 = \frac{48+51}{2} = \underline{49,5}$$

4.1 *Method* 2 (*Using the formula*)

Position of $Q_1 = \frac{1}{4}(n+1) = \frac{1}{4}(13) = 3,25$

This is the position of Q_1. *Now determine* Q_1.

$\therefore\ Q_1 = 3rd\ number + 0,25(\ 4th\ number - 3rd\ number.)$

$\therefore\ Q_1 = 48 + 0,25(51-48)$ $\quad$ $(4th\ number = 51\ :\ 3rd\ number = 48\ )$

$= \underline{48,75}$

4.2 *Method* 1 (*Counting the data*)

The third quartile divides the second half of the data into two equal parts. To the right of the median there are 6 *numbers.*

Therefore, the third quartile lies between the 9*th and the* 10*th number.*

$$\therefore\ Q_3 = \frac{59+61}{2} = \underline{60}$$

4.2 *Method* 2 (*Using the formula*)

Position of $Q_3 = \frac{3}{4}(n+1) = \frac{3}{4}(13) = 9,75$

$\therefore\ Q_3 = 9th\ number + 0,75(10th\ number - 9th\ number.)$

$= 59 + 0,75(61-59)$ $\qquad$ $= \underline{60,5}$

4.3 *Range* = *Maximum value* − *minimum value*

$= 66 - 41$ $\qquad$ $= \underline{25}$

4.4 *Interquartile range* $= Q_3 - Q_1$

$$= 60 - 49,5 \qquad = \underline{10,5}$$

and if we use the values of the second method :

$$= 60,5 - 48,75 \qquad = \underline{11,75}$$

Both answers will be accepted.

3.2 Calculating measures of dispersion of grouped data

3.2.1 **The first quartile, Q_1** : The position of the first quartile, Q_1, is $\frac{1}{4}n$.

3.2.2 **The third quartile, Q_3** : The position of the third quartile, Q_3, is $\frac{3}{4}n$.

Example 5

Concider the data in example 3.

5.1 Identify the class interval in which the first quartile lies.
5.2 Identify the interval in which the third quartile lies.

Solution

5.1 *Position of* $Q_1 = \frac{1}{4}n \qquad = \frac{1}{4}(50) \qquad = 12,5$

The 12,5*th value is found in the interval* [40;45).

$\therefore$ *Q_1 lies in the interval* [40;50)

5.2 *Position of* $Q_3 = \frac{3}{4}n \qquad = \frac{3}{4}(50) \qquad = 37,5$

The 37,5*th value is found in the interval* [50 – 55).

$\therefore$ *Q_3 lies in the interval* [50;55)

3. Listing a five number summary

The five number summary gives the minimum value, the lower quartile (Q_1), die median, the upper quartile (Q_3) and the maximun value.

Example 6

The following are the test marks of 10 learners :

68	49	54	65	42	87	44	47	65	85

Give a five number summery of the data.

Solution

First order the data, i.e., arrange from low to high.

42	44	47	49	54	65	65	68	85	87

Method 1 (*Counting the data*)

1. *The minimum value* = 42

2. *The first quartile*

Divide your data into four equal parts.

		↓			↓		↓		
42	44	47	49	54	65	65	68	85	87

To the left of the median there are 5 *numbers.*

Q_1 is the third number [*The number in the middle of the* 5 *numbers*]

$\therefore\ Q_1 = 47$

3. *The median*

There are 10 *numbers, therefore the median lies between the* 5*th and the* 6*th number.*

$\therefore$ *The median* $= \dfrac{54+65}{2} = 59{,}5$

4. *The third quartile*

To the right of the median there are 5 *numbers.*

$\therefore$ Q_3 *is the* 8*th number* [*The number in the middle of the* 5 *numbers*]

$\therefore Q_3 = 68$

5. *The maximum value* = 87

$\therefore$ *The five number summery is* : 42; 47; 59,5; 68; 87.

Method 2 (*Using the formulae*)

1. *The minimum value* = 42

2. *The first quartile*

Position of the first quartile is $\frac{1}{4}(n+1)\ \therefore\ \frac{1}{4}(10+1) = 2\frac{3}{4}$

$\therefore Q_1$ = 2*nd number* + $\dfrac{3}{4}$(3*rd number* – 2*nd number*)

$\therefore Q_1 = 44 + \frac{3}{4}(47-44)$ [2*nd number* = 44 ; 3*rd number* = 47]

$= 46{,}25$

3. *The median*

Position of the median is $\frac{1}{2}(n+1)\ \therefore\ \frac{1}{2}(10+1) = 5\frac{1}{2}$

So the median lies between the 5*th number the* 6*th number.*

$\therefore$ *The median* $= \dfrac{54+65}{2} = 59{,}5$

4. *The third quartile*

Position of third quartile $= \frac{3}{4}(n+1)\ \therefore\ \frac{3}{4}(10+1) = 8\frac{1}{4}$

So Q_3 *is the* 8*th number* + $\dfrac{1}{4}$(9*th number* – 8*th number*).

$\therefore Q_3 = 68 + \frac{1}{4}(85-68)$ [8*th number* = 68 *and* 9*th number* = 85]

$= 72{,}25.$

5. *The maximum value* = 87

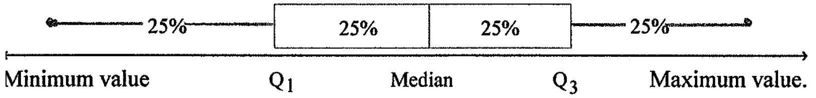
$\therefore$ *The five number summery is* : 42 ; 46,25 ; 59,5 ; 72,25 ; 87.

There is a small difference between the values obtained. Use the method you have been taught in class. Both answers will be accepted as correct.

4. Drawing a box-and-whisker diagram

The box-and-whisker diagram is a graphical representation of the five number summary. A box-and-whisker diagram has 4 parts and each part represents a quarter of the data.

Look at the following representation of the box-and-whisker diagram.



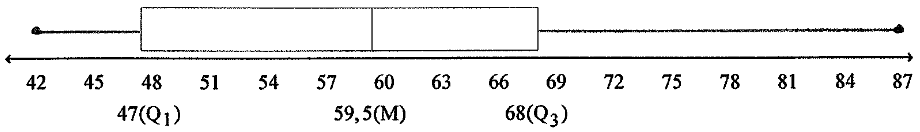
When you have calculated the first quartile, the median and the third quartile, it is easy to draw the box-and-whisker diagram.

Example 7

Draw a box-and-whisker diagram of the data in example 6.

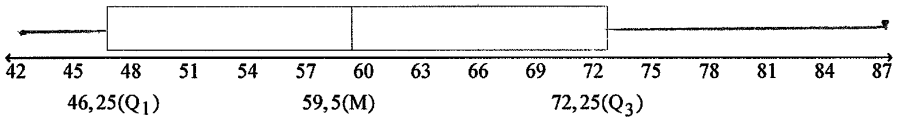
Solution

If you used method 1



Test marks of ten learners.

If you used method 2

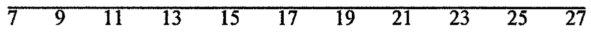


Test marks of ten learners.

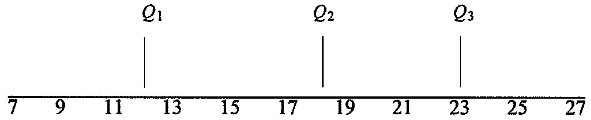
Drawing a box-and-whisker diagram.

Five number summery : 9; 12; 18; 23; 27

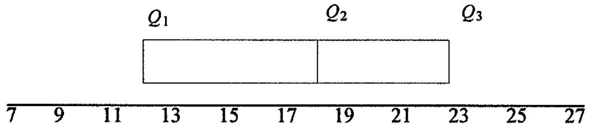
1. Use a convenient scale and draw a horizontal axis starting at the minimum value and ending at the maximum value.



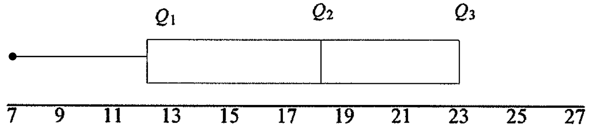
2. Draw short vertical lines above Q_1, Q_2 and Q_3, slightly above the horizontal axis.



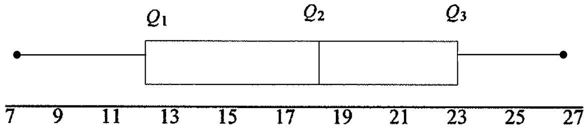
3. Join the vertical lines above Q_1 and Q_3 to form a rectangle.



4. From the left hand side of the box, draw a line to the minimum value.



5. From the right hand side of the box, draw a line to the maximum value.


Q_1
Q_2
Q_3
7 9 11 13 15 17 19 21 23 25 27