NASCA Mathematics Materials Draft 1

## Topic 4: Statistics

## Unit 1: Measures of central tendency

## Learning Outcomes

By the end of this unit, learners should be able to

* Determine the mean, mode, median and range of ungrouped data

## Activity 1: Working with measures of central tendency

**Purpose**:

* Determine the mean, mode, median and range of ungrouped data

**Resources**: A pen or pencil, eraser, calculator and paper

**Suggested time:** 45 minutes

**Introduction**

Whenever we watch television, listen to the radio, or read newspapers, magazine, or books we find statistics. We can find statistics in articles on business, the state of the economy, politics, science, education, sports and many other subjects. Some understanding of statistics is important if we are to understand the information that is presented.

One way to interpret data is to consider the measures of central tendency, which are values that tell you where the middle or centre of the data lies. The measures of central tendency are mean, median and mode.

The mean (or average) of a numerical dataset is the sum of all the data values divided by the number of data values.

The median of a dataset is the middle value after the data have been arranged in ascending order

The mode (or modal value) of a dataset is the most common number or value that occurs the most often.

**Range** – the highest number subtract the lowest number

**Example**

1. The following marks were obtained by learners in a mathematics test (out of 20)

15; 16; 13; 8; 19; 17; 11; 13; 13; 12; 9; 13; 15; 13; 17; 10; 11.

Determine:

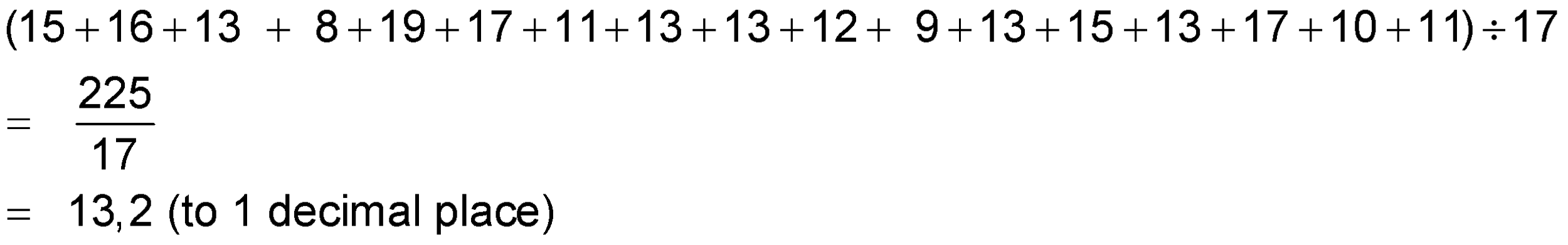
* 1. the mean
  2. the mode
  3. the median

1. Find the median from the following set of numbers:

6; 7; 11; 10; 15; 13; 14; 8

**Solution**

1.1 Add all up and divide by the number



* 1. The mode is 13 (occurs 5 times)

3 Arrange from highest to lowest:

19; 17; 17; 16; 15; 15; 13; 13; 13; 13; 13; 12; 11;11; 10; 9; 8

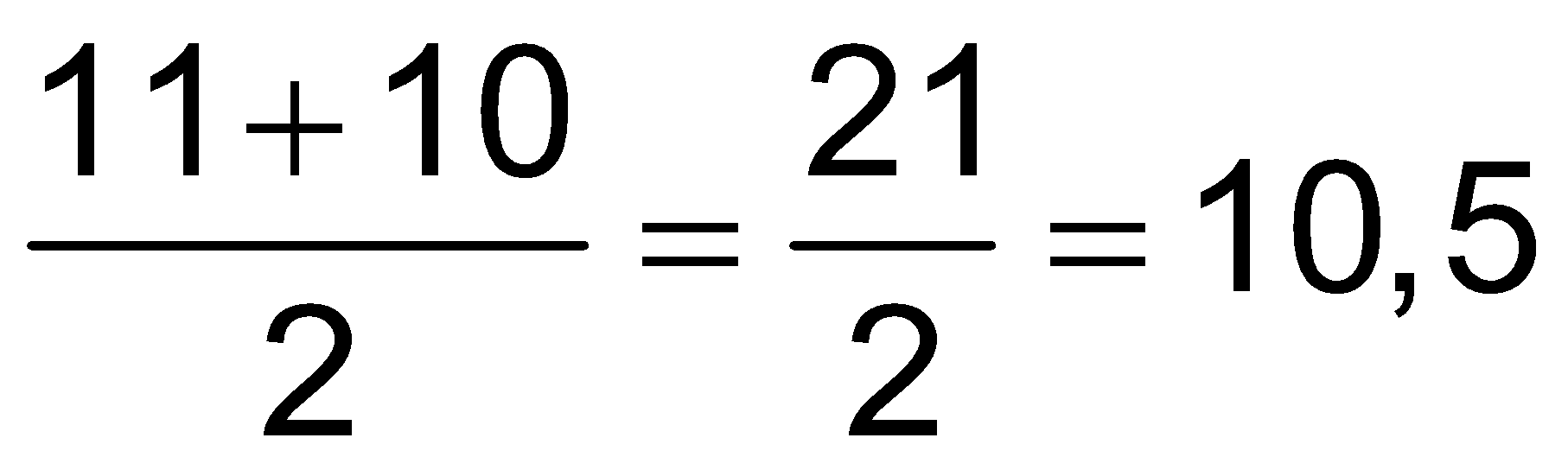
The 9th number from the top will be the median. Thus, the median is 13.

1. There are 8 numbers in this set.

Arrange the numbers from highest to lowest:

15; 14; 13; 11; 10; 8; 7; 6

We can break up this set into two parts of four. The fourth and fifth number are added together and divided by 2:

Median = 

**Task 1**

PQR High School has two cricket teams, a junior and a senior team. The junior team consists of 17 players (including reserves) and the senior team consists of 16 players (including reserves).

The masses (in kg) of the junior team are:

56 60 67 45 51 53 64 49 56 48 42 51 64 52 64 49 50

The masses (in kg) of the senior team are:

88 81 53 62 83 68 70 62 91 78 64 74 73 54 62 62

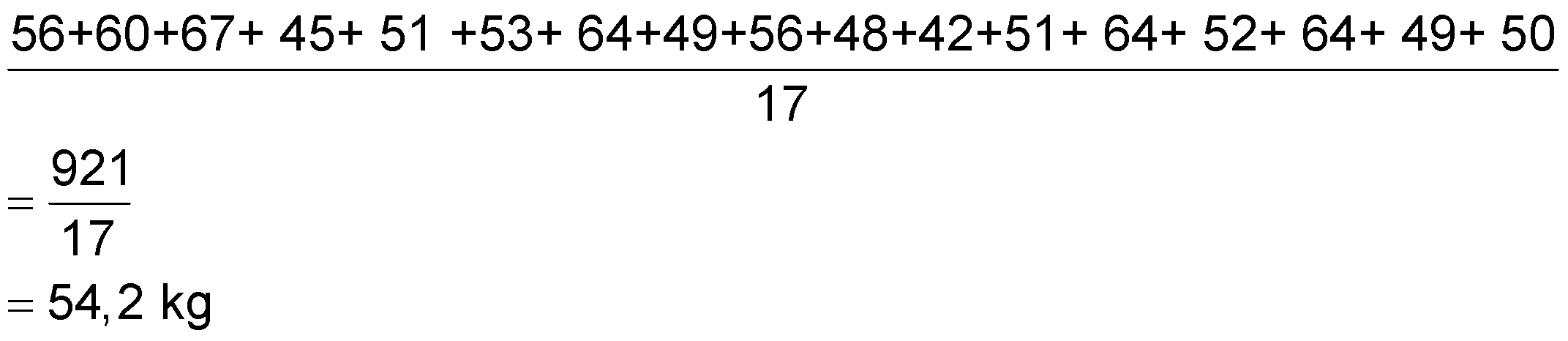
1. Arrange the masses of the senior team in ascending order.
2. Determine the mode of the senior team.
3. Determine the median of the senior team.
4. Calculate the mean of the masses of the junior team correct to one decimal digit.
5. Calculate the median of the masses of the junior team

#### Guided reflection on Activity 1

|  |
| --- |
| * + - 1. To determine the median of a set of data values, what should you first do?   Arrange the data values in ascending order   * + - 1. Rubi says to find the median of dataset with 16 data values, you arrange the data   values in ascending order, and then take the data value that is in the 8th position.  Do you agree with Rubi. If you do not, then explain how you will calculate the  median.  I do not agree with Rubi. I will calculate the median by arranging the data in  ascending order, then find half of the sum of data values that are 8th and 9th  positions   * + - 1. Khomotso says to calculate the mode you add all the data values and divide the sum by the number of data values in the dataset. This is not a correct way to calculate the mode. Explain how you will calculate the mode of a data set.   I will look for the most common number or value that occurs the most often. |

**Answers to Task 1**

1. 53; 54; 62; 62; 62; 62; 64; **68;** **70**; 73; 74; 78; 81; 83; 88; 91
2. Mode of senior team: 62 kg (occurs 4 times)
3. Median of the senior team : = 69 kg
4. Mean of junior team masses:



Arrange masses in ascending order:

42;45;48;49;49; 50;51; 51; **52**; 53; 56; 56; 60; 64;64; 64; 67

As the data set has an uneven number of values, the median will be exactly the middle data value

In this case it is the 9th data value , which is 52kg

1. Median is 52 kg

**Summary Assessment**

The maximum recorded life spans (to the nearest year) of 10 different animals are listed below:

Baboon 36 Polar bear 35

Cat 29 Elephant 60

Gorilla 39 Beaver 20

Dog 20 Horse 46

Lion 25 Cow 30

For this set of life spans find the following:

* 1. mean
  2. median
  3. mode

#### Guided reflection on Summary Assessment

|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Have you been able to correct errors you made? |

**Answers for summary assessment**

* 1. Add all the years up to get 340. Now 340 10 = 34 years
  2. To get the median we arrange the data from highest to lowest: 60; 46; 39; 36; 35; 30; 29;25; 20; 20. Since this is an even number of data items, we add the 5th and 6th items and divide by 2 so we have: Median = (35 + 30) = 35,5 years
  3. We find there are two data items which are 20 years so the mode = 20 years

**Unit 2: Measures of dispersion**

**Learning Outcomes**

By the end of the lesson, the learner should be able to determine:

* the median, lower (first) and upper (third) quartiles in a given data set
* the percentiles (measures of position) of a given data set
* the inter-quartile range and semi-interquartile range

Measures of dispersion are values that tell you how spread out (or grouped) the day values are. They provide a more complete picture than just the mean, median, and mode. In this unit, we will study the following measures of dispersion: range; interquartile range and semi-interquartile range. Before you can understand the concepts of interquartile range and semi-interquartile range, you need to know how to calculate percentiles and quartiles.

**Activity 1: Percentiles**

**Purpose:**

Determine the percentiles of a given data set

**Resources**: A pen or pencil, eraser, calculator and paper

**Suggested time:** 45 minutes

Let us consider the position of Dean and Lulu who are applying for a mathematics tutor job. Dean was ranked 30th in his Mathematics class at university while Lulu was ranked 15th in her mathematics class at another university.

Using rank, as a one of the criteria, who has the higher ranking?

At the moment it would seem that Lulu has a higher ranking. However, you need more information to arrive at an informed conclusion. In order to describe their relative positions better, we can find the **percentile rank or percentile** for Dean and Lulu

Suppose there were 75 students in Lulu's class and 120 students in Dean’s class.

In Lulu's class, there were 75-15 = 60 people ranked below her.

Therefore,

 of Lucy's classmates were ranked below her.

We may say that Lulu was in the 80th percentile in her class

For Dean, 120 –30 = 90 people were ranked below him. Therefore,

 of Dean's classmates were ranked below him

Dean was in the 75th percentile in his class

Thus, we may conclude that Lulu’s ranking is higher than that of Dean’s.

Percentiles divide sets of data into 100 equal parts. Hence, 100% is the basis of measure.

**Examples**

1. In a class of 50, Themba has a rank of 12. What is Themba's percentile ranking in the class?

There are 50 – 12 = 38 learners ranked below Thembu.

His percentile rank is 

**Task 1**

1. In a geography class of 50 learners, an examination was given and Sean scored at the 80th percentile. How many learners scored lower than Sean?
2. In a mathematics class test, Devi is ranked at the 15th percentile. If there were 60 learners in Devi’s class, how many learners scored higher than Devi?

#### Guided reflection on Activity 1

|  |
| --- |
| 1. What does 80th percentile mean to you?   The 80th percentile is the point in a distribution of values (or scores) below which 80% of the values fall.   1. For Q2 of Task 1, explain how you will calculate how many learners scored higher than Devi.   Since Devi is ranked at the 15th percentile, it means that 15% of learners scored less than Devi. This actually means that 85% of learners scored more than Devi.  I will the find 85% of 60 learners. |

**Answers to Task 1**

1. 80% of the learners scored less than Sean.

This is equal to .

40 learners scored lower than Sean.

1. 15% of the learners scored less than Devi. Thus, 85% scored higher than her. This is equal to 85% of 60 = 51

51 learners scored higher than Devi

**Activity 2: Quartiles**

**Purpose:**

Determine the quartiles, interquartile range and semi-interquartile range of a given data set

**Resources**: A pen or pencil, eraser, calculator and paper

**Suggested time:** 45 minutes

We looked at Percentiles in the previous activity. We noted that Percentiles divide sets of data into 100 equal parts. Hence, 100% is the basis of measure. From the percentiles, we refer to three key percentiles.

The 25th percentile; 50th percentile and the 75th percentile are the most commonly used percentiles in educational testing.

**Quartiles**

The data is divided into quarters, which we call quartiles. Each data set has 3 quartiles.

The first quartile (also called the lower quartile) has 25% of data (scores) below it; it is the median of the lower 50% of the data.

The second quartile is also known as the median, which we know is the middle value of data set. It has 50% of the data (scores) below it.

The third quartile (also called the upper quartile) is that value which has 75% of the data (scores) below it; it is the median of the upper 50% of the data.

We can use our first quartile and third quartile to compute the **interquartile range.**

**Interquartile range**

= Third quartile – First Quartile

=  - (this is 50% of the data)

We can use the interquartile range to compute the semi-interquartile range

**Semi-interquartile range**

= one-half of the interquartile range

= 

**Example**

The dataset below gives the heights (in mm) of seedlings 4 weeks after germinating:

47; 52; 56; 62; 71; 74; 78; 86; 89; 92; 93; 95

Determine the following measures of dispersion for the data set:

1. Median
2. First quartile
3. Third quartile
4. Interquartile range
5. Semi-interquartile range

.

**Answers**

1. The data is arranged from lowest to highest and is an even number. To find the median we add the sixth and seventh number and divide by 2:

mm

1. The first quartile is the median of the lower half of the data which is = 59mm
2. The third quartile is the median of the upper half of the data which is = 90,5mm
3. The interquartile range is  - = 90,5 – 59 = 31,5 mm
4. The semi-interquartile range is mm

**Task 2**

1. In a class of 150 students, Peter is ranked 30. What is his percentile rank?
2. Given 16 scores: 65; 68; 67; 66; 56; 47; 69; 74; 48; 76; 80; 89; 57; 58; 60; 90

* 1. What is the rank (from the top) of 76?
  2. What is the percentile rank of 56?
  3. Determine the first quartile?
  4. Determine the third quartile?
  5. What is the percentile rank of 89?
  6. Determine the interquartile range

#### Guided reflection on Task 2

|  |
| --- |
| 1. Is it true that the second quartile is the middle value of the data set?   Yes (it is also called the median)   1. How will you calculate the third quartile of a given data set?   Consider the upper 50% of the data. Then find the median for this data set.   1. Explain how you will calculate the inter-quartile range of a dataset?   I will calculate the first quartile and third quartile. Then I will subtract the first quartile from the third quartile. |

**Answers to Task 2**

1. Peter’s percentile **rank is**



1. Arrange scores from lowest to highest

47; 48; 56; 57; 58; 60; 65; 67; 66; 68; 69; 74; 76; 80; 89; 90

* 1. 76 is ranked at 4
  2. 56 has 2 scores below it. Now its percentile rank will be



* 1. First quartile = = 57,5
  2. Third quartile = = 75 .
  3. 89 has 14 scores below it. Thus, the percentile rank of 89 is calculated as follows:



2.6 Interquartile range = 75 – 57,5 = 17,5

**Summary assessment**

1. Ntombi scored at the 80th percentile in a certain life skills test in her class of 30 students. What is her rank in her class?
2. In a class of 60 students, Jane has a percentile rank of 75. What is Jane's rank in the class?
3. The following scores were made by a batsman during a cricket season.

27; 45; 69; 77; 45; 13; 65; 34; 56; 87; 12; 50.

* 1. Determine his average score
  2. Determine his median score
  3. Determine the lower quartile () and upper quartile ()

3.4 Describe the batsman’s performance in terms of his average and quartiles

#### Guided reflection on Summary Assessment

|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Write down the errors you made. 4. Have you been able to correct errors you made? |

**Answers for summary assessment**

1. 80% of 30 = 24. Thus, Ntombi will be ranked 5th in her class
2. 75% of 60 = 45. Jane will be ranked at 15th in her class
   1. Average = 48,33
   2. Median score = 
   3. = = 30,5 = = 67
   4. He has a good average of 48,33. Only three of his scores are below the first quartile.. All the other scores are above it. Thus, 75% of the time he is scoring at above the first quartile. He is doing reasonably well as a batsman.

**Unit 3: Five Number summaries and box-and-whisker plots**

**Learning Outcomes**

By the end of this unit, learners should be able to:

* Complete a 5-number summary for a data set.
* Determine the fences
* Use the five number summary to draw a box-and-whisker plot.
* Interpret the meaning of the representation of the box and whisker diagram

**Activity 1:** **A 5- number summary for a data set and its corresponding box-and-whisker-plot**

**Purpose**:

* Determine the 5- number summary for a data set
* Draw a box-and whisker plot for a data set

**Resources**: A pen or pencil, eraser, calculator, paper and graph or dotty paper

**Suggested time:** 30 minutes

We use the some of the measures of data from the previous units to compile a 5-number summary for a data set.

The 5 number summary for univariate data set are the following::

**Maximum**: the highest value in the set

**Minimum**: the lowest value in the set

**Median**: the middle value of the data set

**First quartile** (lower quartile): the middle value of the lower half of the data

**Third quartile** (upper quartile): the middle value of the upper half of the data.

**Example**

Determine the 5-number summary of the data set below. Use your 5-number summary to draw the box and whisker plot.

62; 56; 71; 78; 89; 92; 86 ; 74

**Solution**

Arrange the values from lowest to highest: 56; 62; 71; 74; 78; 86; 89; 92

Maximum value = 92

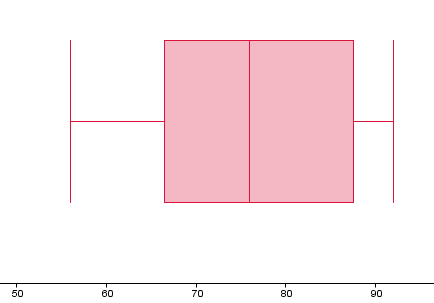
Minimum value = 56

First quartile = = 66,5

Third quartile = = 87,5

Median = = 76

**Box-and-whisker plot**



**Activity 2: Fences and outliers of a data set**

**Purpose**:

* Determine the outliers of a data set
* Determine the fences of a data set

**Resources**: A pen or pencil, eraser, calculator, paper and graph or dotty paper

**Suggested time:** 30 minutes

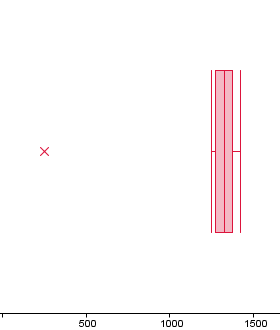
**Outliers**

Outliers are items in a **data set** that lie well above or below the majority of the scores in the **set**.

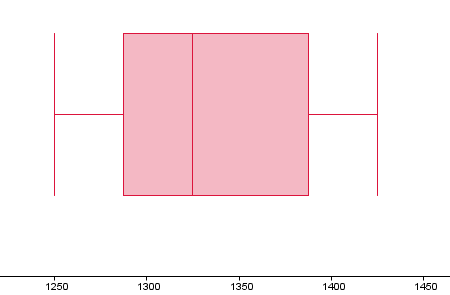
For example, consider the following data set

250; 1250; 1275; 1300; 1325; 1325; 1350; 1375; 1400; 1425

Here were see that that 9 of the 10 values lie from 1250 to 1450. The other value of 250 is very far away from the second lowest score. Thus, 250 is an outlier. We can draw the box plot for this data set:



We note that 250 (the outlier) is not part of the box plot. If we leave the 250 out then our box-plot will look as follows:



**Fences of a data set**

We can also determine the fences of the above data set. We first find the first and third quartiles. The formula for the upper fence is  while the formula for the lower fence is  where IQR is the interquartile range.

Let us consider the following data set again:

250; 1250; 1275; 1300; 1325; 1325; 1350; 1375; 1400; 1425.

We have 10 values. The first quartile is the median of the lowest 5 values of the set. Thus, our first quartile is 1275.

The third quartile is the median of the top 5 values of the set. Thus, our third quartile is 1375.

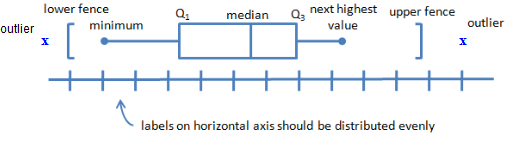
The interquartile range (IQR) = 1375 – 1275 = 100

Our lower fences is 1275 – 1,5×100 = 1275 – 150 = 1125

Our higher fence is 1375 + 1,5 × 100 = 1375 + 150 = 1525

This means that any value below our lower fence and above our upper fence is regarded as an outlier.

The diagram below shows the various key values which occur in a box-and-whisker plot



**Task 1**

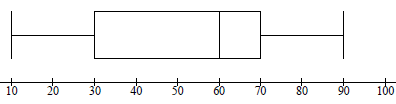
Two Mathematics classes, A and B are in competition to see which class performed best in the June examination. The marks of the learners in class A are given below and the box-and-whisker plot for class B illustrates the results of class B. Both classes have 25 learners. Marks are given as percentages:

**Marks of class A**

9; 14; 14; 19; 21; 23; 33; 35; 37; 37; 42; 45; 55;

56; 57; 59; 68; 75; 75; 75; 77; 78; 80; 81; 92

The box and whisker diagram for the learners in class B is:



1. Write down the five-number summary for class A
2. Are there any outliers in the data for class A? Explain
3. Draw the box and whisker diagram (box plot) for class A. Show all relevant values
4. Determine which class did better in the June Examination and give reasons for your conclusion

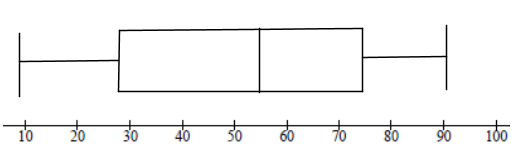
#### Guided reflection on Task 1

|  |
| --- |
| 1. List the five values which give a complete description of the dispersion or spread of a dataset.   The 5-number summary values: minimum value; lower quartile ; median; upper quartile ; and maximum value.   1. Write down steps to follow when drawing a box-and –whisker plot   Step 1: Arrange the data in ascending order  Step 2: Complete the 5-number summary  Step 3: Draw a number line which includes the minimum and maximum values.  Step 4: Draw a short vertical just above the number line which shows the median  Step 5: Draw a box using the lower and upper quartiles as endpoints  Step 6: Draw the whiskers by extending lines from the quartiles to minimum and  maximum values. |

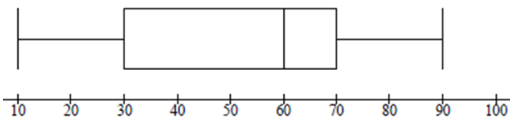
**Answers to task 1**

1. Maximum = 92; minimum = 9; first quartile = 28; median = 55; third quartile = 75
2. Interquartile range: 75 – 28 = 47 Lower fence = 28 -1,5×47 = -42,5; Higher fence = 75 + 1,5 ×47 = 145,5 All values fall within the lower and upper fence. There are no outliers

Class A



Class B



1. Class B did marginally better than class A. Its median 60 while for class A it is 55. Further, the first quartile for class B is higher than that of class A

**Summary Assessment**

In an experiment a group of 23 girls were presented with a page containing 30 coloured triangles. They were asked to name the colours of the rectangles correctly in the quickest possible time. The time, in seconds, taken by each of the girls is given in the table below:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 12 | 13 | 13 | 14 | 14 | 16 | 17 | 18 | 18 | 18 | 19 | 20 |
| 21 | 21 | 22 | 22 | 23 | 24 | 25 | 27 | 29 | 30 | 36 |

1. Determine:
   1. the mean
   2. the interquartile range
2. Draw a box-and-whisker diagram to represent the above data
3. The five number summary of the times taken by a group of 23 boys in naming the colours of the rectangle correctly is (15; 21; 23,5; 26; 38).
   1. Which of the two groups, boys or girls, had the lower median time to correctly name the colours of the rectangles
   2. The first three learners who named the colours of all 30 rectangles correctly in the shortest time will each receive a prize, How many boys will be among these prize winners? Motivate your answer

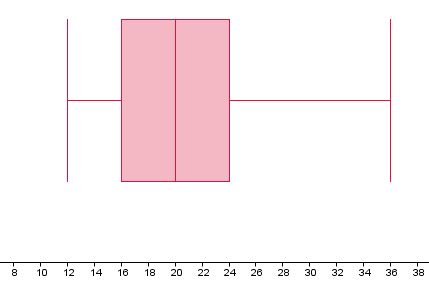
#### Guided reflection on Summary Assessment

|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Write down the errors you made. 4. Have you been able to correct errors you made? |

**Answers for summary assessment**

* 1. Add all up to get: 472. Now 472 23 = 20,52 (the mean)
  2. First quartile is the median of the lower half of the data items and this is 16; the third quartile is the median of the upper half of the data which is 24. Thus, Interquartile range = 24 – 16 = 8

2.



3.1 Median of girls = 20 seconds; Median of boys = 23,5 seconds. The girls had the lower time.

3.2 No boys will get a prize as the lowest time taken by the boys is 15 seconds. The top 3 of the girls is lower than 15 seconds (12 seconds; 13 seconds; 13 seconds)

**Unit 4: Frequency distribution table and histograms and frequency polygons**

**Learning Outcomes**

By the end of this unit, learners should be able to

* Compile a frequency distribution table by grouping data into classes;
* Construct histograms using tabulated grouped data
* State whether a given set of data value is skewed to the left or right

**Activity 1: Frequency tables and histograms**

**Purpose**:

* Grouping raw data into classes and compiling the frequency distribution table.
* Draw an histogram from tabulated group data

**Resources**: A pen or pencil, eraser, calculator, paper (and graph or dotty paper)

**Suggested time:** 45 minutes

**Grouping data**

Thus, far we have been working with univariate data. We will now group univariate data into classes.

**Example**

Suppose 20 students obtained the following marks out of 30 for a mathematics test:

25 26 18 24 27

9 11 15 22 4

8 12 19 20 13

15 16 18 13 7

We can organise the data in ascending order:

4 7 8 9 11

12 13 13 15 15

16 18 18 19 20

22 24 25 26 27

This information can be **grouped** in a frequency table of 6 classes:

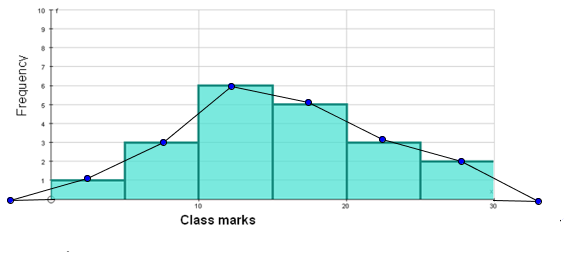
|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Test mark** | **Tally** | **Frequency** |
| 1 | (0; 5] | / | 1 |
| 2 | (5;10] | /// | 3 |
| 3 | (10;15] | ~~////~~  / | 6 |
| 4 | (15;20] | ~~////~~ | 5 |
| 5 | (20;25] | /// | 3 |
| 6 | (25;30] | // | 2 |
|  |  |  | Sum = 20 |

Note: If  is a test mark then (0; 5] means; (25; 30] means , etc

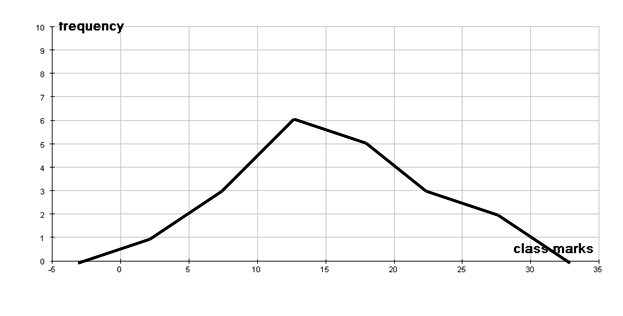
We can now construct a histogram for the above grouped data:



We can draw the frequency polygon by joining the mid-points of top of the bars Note that extra classes are included; one on the left and one on the right.



We can leave our the histogram bars to get the following frequency polygon



**Task 1**

The following marks were scored by learners in a Science test (out of 50)**.**

08, 13, 16, 17, 18, 19, 19, 22, 22, 23, 23, 23, 25, 25, 27, 28, 28, 29, 31, 31, 32, 33, 33, 34, 35, 36, 37, 38, 39, 39, 39, 40, 40, 41, 41, 42, 43

* 1. Complete the frequency table below

|  |  |
| --- | --- |
| **Class mark** | **Frequency** |
| 0 – 10 |  |
| 11 – 20 |  |
| 21 – 30 |  |
| 31 – 40 |  |
| 41 – 50 |  |

* 1. Now draw the histogram for the above data.
  2. Complete the frequency polygon

#### Guided reflection on Activity 1

|  |
| --- |
| 1. How do you draw a frequency polygon through using a histogram?   Draw a line from the mid-class grouping before the first histogram bar to pass through the mid-class value of each bar in the histogram and then end at the mid-class value of the grouping after the last bar. |

**Answers to task 1**

1.1

|  |  |
| --- | --- |
| **Class mark** | **Frequency** |
| 0 – 10 | 1 |
| 11 – 20 | 6 |
| 21 – 30 | 11 |
| 31 – 40 | 15 |
| 41 – 50 | 4 |

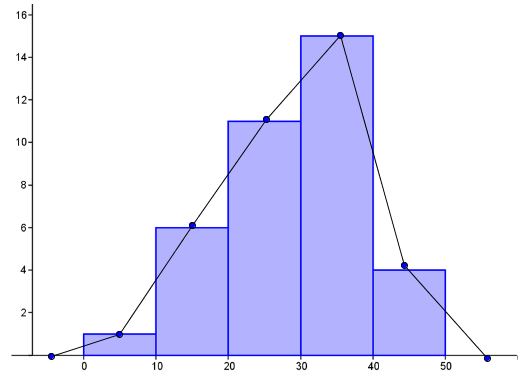
1.2



Class intervals

Frequency

1.3



Frequency

Class intervals

**Activity 2: Determining how given data is skewed**

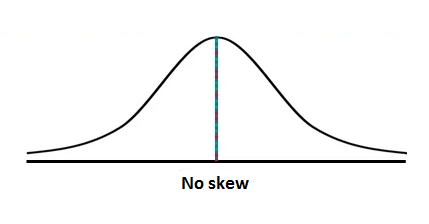
**Purpose**:

* Determine how given data values are skewed

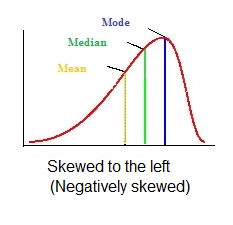
**Resources**: A pen or pencil, eraser, calculator, paper (and graph or dotty paper)

**Suggested time:** 45 minutes

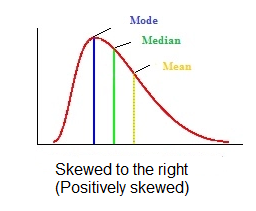
The diagram below shows data which is symmetric or normally distributed. The median, mean and mode are equal. It is very rare to get data which is symmetric in nature.



The diagram below shows data which is skewed to the left or negatively skewed.



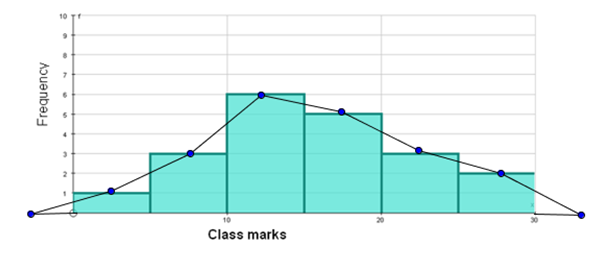
For negatively skewed data, mean < median < mode



For positively skewed data, mode < median < mean

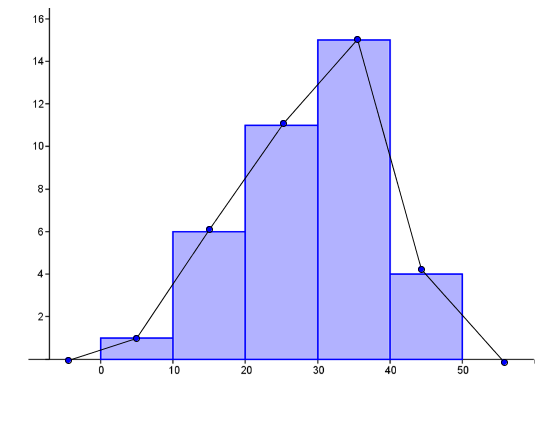
**Examples**

1. Consider the histogram and frequency polygon drawn earlier in this unit in Activity 1.



Our mode is 6 and both the median and mean are to the right of the mode. We say that this data is skewed to the right.

1. Examine the histogram and frequency polygon drawn for Task 1.



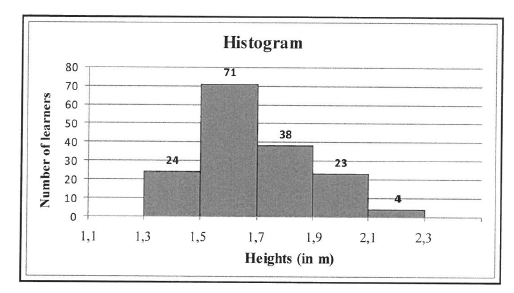
Frequency

Class intervals

Our mode is 15 and both the mean and median are to the left of the mode. We say that this data is skewed to the left.

**Task 2**

The heights of 160 learners in a school are measured. The height of the shortest learner is 1,39 m and the tallest learner is 2,21 m. The heights are represented in the histogram below:



1. Write down the modal class for the above histogram
2. Calculate the range of the heights
3. Describe the skewness of the data

#### Guided reflection on Activity 2

|  |
| --- |
| 1. When the data values are more to spread to the right of the median, what can you conclude about the skewness of the data?   We say the data is skewed to the right   1. Is it true that the mean is less than the median when the data is skewed to the right?   No. When the mean is greater than the median, the data is skewed to the right. |

**Answers to Task 2**

1. The mode is 71 and it lies in the class (1,5;1,7]
2. Range = 2,21 m – 1,39m = 0,82 m
3. The data is skewed to the right

**Summary assessment**

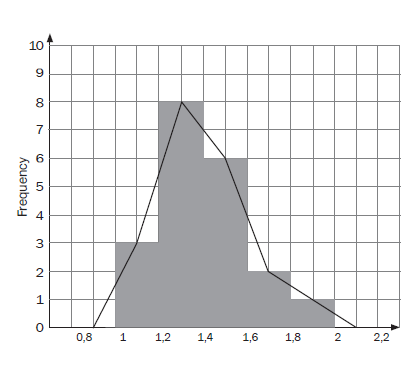
1. The marks of 35 grade-eight learners obtained the following marks out of 50 for a mathematics test is given below

46, 40, 12, 10, 47, 23, 26, 8, 29, 34, 37, 17, 40, 50, 18, 23, 33, 23, 24, 15, 35, 23, 19, 22, 28, 35, 27, 42, 29, 26, 46, 33, 27, 19, 28

* 1. Complete the table below for the above marks

|  |  |
| --- | --- |
| **Interval of scores** | **Frequency** |
| 0 – 10 |  |
| 11 – 20 |  |
| 21 – 30 |  |
| 31 – 40 |  |
| 41 – 50 |  |

1.2 Now construct the histogram. What can you conclude about the marks?

1. Study the following histogram and frequency polygon

Heights of Weeds

Describe the skewness of the data.

#### Guided reflection on Summary Assessment

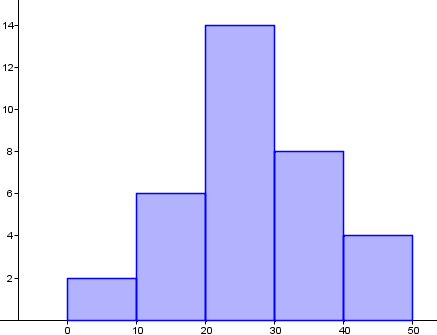
|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Write down the errors you made. 4. Have you been able to correct errors you made? |

**Answers for summary assessment**

1.1

|  |  |
| --- | --- |
| **Interval of scores** | **Frequency** |
| 0 – 10 | 2 |
| 11 – 20 | 6 |
| 21 – 30 | 14 |
| 31 – 40 | 8 |
| 41 – 50 | 5 |

1.2



Class intervals

Frequency

The learners did satisfactorily in the test with 27 learners getting more than 20 out of 50; the modal class interval is [21;30] and both the mean and median lie in this interval.

1. The mode is 8 and both the median and mean are to the right of the mode. We say that this data is skewed to the right.

**Unit 5: Ogive Curves**

**Learning Outcomes**

By the end of this unit, learners should be able to:

* Compile the cumulative frequency for a given set of grouped data
* Draw the Ogive curve from the cumulative frequency values
* Use the Ogive curve to estimate quartile values
* Calculate the mean, median and modal values of grouped data using the Ogive curve

**Activity 1:** The cumulative frequency graph and the Ogive curve

**Purpose**:

* Compiling the cumulative frequency table
* Drawing the Ogive curve

**Resources**: A pen or pencil, eraser, calculator, paper and graph paper

**Suggested time:** 60 minutes

A **Cumulative Frequency Graph**is a graph plotted from a cumulative frequency table. A cumulative frequency graph is also called an **ogive** or **cumulative frequency curve.**

It is best to understand how to compile the cumulative frequency table by considering the following example:

The marks of learners (as a percentage) obtained in a grade 12 examination are shown in the table below:

|  |  |  |
| --- | --- | --- |
| **%** | **Frequency**  **(Number of learners)** | **Cumulative frequency** |
| 0 < x ≤ 10 | 0 | 0 |
| 10 < x ≤ 20 | 2 | 2 |
| 20 < x ≤ 30 | 6 | 8 |
| 30 < x ≤ 40 | 7 | 15 |
| 40 < x ≤ 50 | 14 | 29 |
| 50 < x ≤ 60 | 20 | 49 |
| 60 < x ≤ 70 | 35 | 84 |
| 70 < x ≤ 80 | 29 | 113 |
| 80 < x ≤ 90 | 6 | 119 |
| 90 < x ≤ 100 | 1 | 120 |

We know how to work out the frequency as we have done previously. To work out the cumulative frequency, we add the frequencies of the intervals in a chronological order. For example, the cumulative frequency of the interval 0 < x ≤ 10 is the same as the frequency for this interval, which is 0.

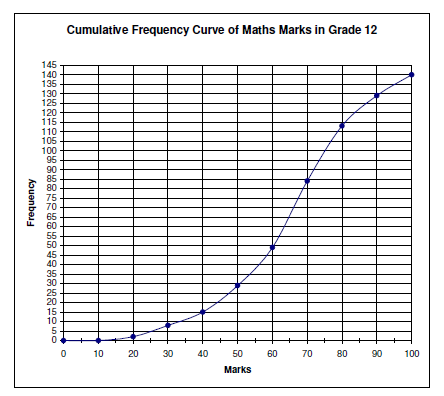
For the next interval 10 < x ≤ 20, the frequency is 2. So the cumulative frequency is 0 + 2 = 2.

For the interval 20 < x ≤ 30, the frequency is 6. To get the cumulative frequency we add this 6 to the previous frequency 2 so we get 8.

We proceed in this manner until we arrive at a cumulative frequency of 120 for the interval 90 < x ≤ 100, even though the frequency for this interval is 1.

Please note that the 120 is the total number of learners (cumulative)

We use the cumulative frequency to draw the Ogive curve.



**Task 1**

In an English class, 30 learners completed a test out of 20. Here is a list of heir results:

14; 10; 11; 19; 15; 11; 13; 11; 9; 11; 12; 17; 10; 14; 13;

17; 7; 14; 17; 13; 13; 9; 12; 16; 16; 9; 11; 11; 13; 20

1.1 Complete the table below:

|  |  |  |
| --- | --- | --- |
| **Class interval** | **Frequency** | **Cumulative frequency** |
| 1 < x ≤ 5 |  |  |
| 5 < x ≤ 10 |  |  |
| 10 < x ≤ 15 |  |  |
| 15 < x ≤ 20 |  |  |

**NB:** 1 < x ≤ 5 is the same as (1;5]

1**.**2. Draw the Ogive curve from your cumulative frequency.

#### Guided reflection on Activity 1

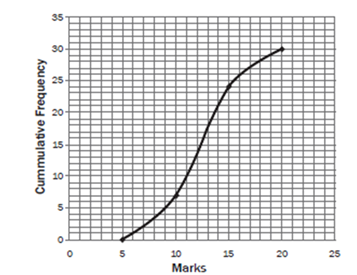
|  |
| --- |
| 1. What is cumulative frequency?   A running total of the frequencies.   1. Describe the important points to bear in mind when drawing an ogive curve.   The graph is a smooth curve, and not straight lines joined together.  Unlike frequency polygons, which work with mid-class values, this curve uses the endpoints of each group.  While the curve is connected to the -axis at the end point of the previous group (the lower limit of the first group, the curve does not return to a point on the -axis like the frequency polygon does. |

**Answers to task 1**

1.1

|  |  |  |
| --- | --- | --- |
| **Class interval** | **Frequency** | **Cumulative frequency** |
| 1 < x ≤ 5 | 0 | 0 |
| 5 < x ≤ 10 | 7 | 7 |
| 10 < x ≤ 15 | 17 | 24 |
| 15 < x ≤ 20 | 6 | 30 |

1.2



**Activity 2:** Using the cumulative frequency graph ( Ogive) curve to determine the median and quartiles of grouped data

**Purpose**:

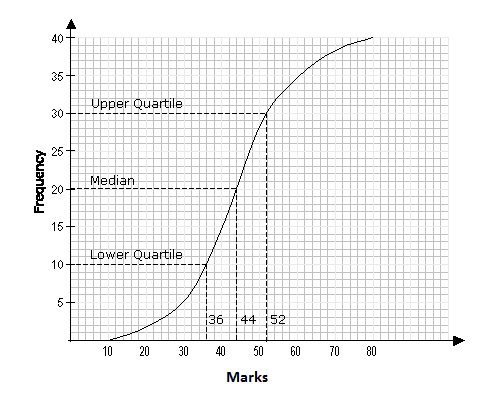
* Compiling the cumulative frequency table
* Drawing the Ogive curve
* Working out the median, quartiles and interquartile range

**Resources**: A pen or pencil, eraser, calculator, paper and graph paper

**Suggested time:** 64 minutes

**Example**

The following cumulative frequency graph shows the distribution of marks scored by a class of 40 students in a test.



Use the graph to estimate

1. the median mark
2. the upper quartile
3. the lower quartile
4. the interquartile range

**Answer:**

1. The Median corresponds to the 50th percentile, that is, 50% of the total frequency. Now 50% of the total frequency = 20. From the graph, 20 on the vertical axis, corresponds to 44 on the horizontal axis. The median mark is 44.
2. The upper quartile corresponds to the 75th percentile i.e. 75% of the total frequency. Now 75% of the total frequency = 30. From the graph, 30 on the vertical axis, corresponds to 52 on the horizontal axis. The upper quartile is 52.
3. The lower quartile corresponds to the 25th percentile i.e. 25% of the total frequency. Now 25% of the total frequency = 10. From the graph, 10 on the vertical axis corresponds to 36 on the horizontal axis. The lower quartile is 36.
4. The interquartile range

= upper quartile – lower quartile

= 52 – 36 = 16

**Task 2**

Consider the frequency table below:

|  |  |
| --- | --- |
| **Length (in mm)** | **Frequency** |
| 11 – 15 | 2 |
| 16 – 20 | 4 |
| 21 – 25 | 8 |
| 26 – 30 | 14 |
| 31 – 35 | 6 |
| 36 – 40 | 4 |
| 41 – 45 | 2 |

1. Determine from the above table the following:
   1. the modal class
   2. the mean
2. Draw a cumulative frequency graph (ogive curve) for the frequency table above.
3. Use the ogive curve (from 2 above) to determine the following measures:
   1. the median
   2. upper quartile
   3. lower quartile
   4. semi-interquartile range
4. Use your graph to determine the approximate number of data items which lie between the upper quartile and the maximum?
5. Which data item lies at the 60th percentile?

#### Guided reflection on Activity 2

|  |
| --- |
| 1. How will you determine the modal class from a frequency table?   Look for the group with the highest frequency.   1. How will you determine which data item lies in the 60th percentile using the ogive curve?   Find 60% of 40. This will give you 24. Then draw a line from 24 on the vertical axis parallel to the horizontal axis. Where the line touches the curve drop a perpendicular onto the -axis and read off the -value |

**Answers to Task 2**

* 1. The modal class is the class interval 26 – 30. This interval consists of the most number of data items (14)
  2. To find the mean from this table we add two columns to the table as shown below. The class mark is found by adding the end points of each interval and dividing by 2. The class mark is multiplied by the frequency to give the approximate sum of the data items. This is written in the last column. The total of 1100 is then divided by the total number of data items (40).

|  |  |  |  |
| --- | --- | --- | --- |
| **Length (in mm)** | **Class mark** | **Frequency** | **Class mark Frequency** |
| 11 – 15 | 13 | 2 | 26 |
| 16 – 20 | 18 | 4 | 72 |
| 21 – 25 | 23 | 8 | 184 |
| 26 – 30 | 28 | 14 | 392 |
| 31 – 35 | 33 | 6 | 198 |
| 36 – 40 | 38 | 4 | 152 |
| 41 – 45 | 43 | 2 | 86 |
| **Total** | | **40** | **1100** |

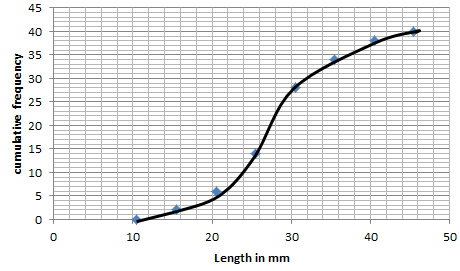
Thus, the mean is =27,75 mm

2**.** In the original frequency table we need to add a class with 0 frequency before the first class and then find the upper boundary for each class interval. The upper class boundary is midway between the upper value of one interval and the lower value in the next interval. For example, 10,5 is calculated by adding 10 to 11 and dividing by 2.

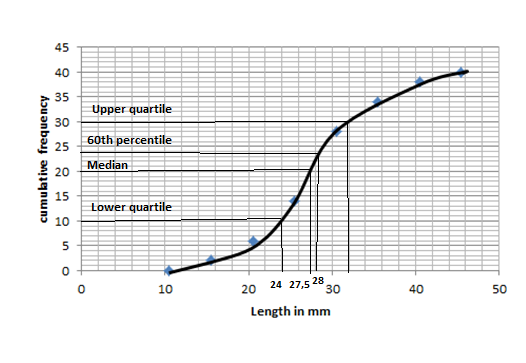
**NB**: Sometimes, we just use the upper limit on the horizontal axis and plot these values with the cumulative frequency on the vertical axis

|  |  |  |  |
| --- | --- | --- | --- |
| **Length (in mm)** | **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 | 8 | 14 |
| 26 – 30 | 30,5 | 14 | 28 |
| 31 – 35 | 35,5 | 6 | 34 |
| 36 – 40 | 40,5 | 4 | 38 |
| 41 – 45 | 45,5 | 2 | 40 |

The cumulative frequency graph (ogive curve) is shown below.



To answer questions 3, 4 and 5, we draw horizontal lines from the vertical axis to the the ogive as shown below**.**



3**.**

* 1. The median is at 20 (50% of 40) on the vertical axis. From here we draw a line to the curve. It corresponds to a length value of 27,5 mm
  2. The upper quartile is at 30 (75% of 40) on the vertical axis. From here we draw a line to the curve. It corresponds to a value of 32 mm.
  3. The lower quartile is at 10 (25% of 40) on the vertical axis. From here we draw a line to the curve. It corresponds to a value of 24 mm
  4. Semi-interquartile range =

=

= 4 mm

1. 10 items lie between the upper quartile and the maximum?
2. For the 60th percentile, we draw a line from 24 (60% of 40) on the vertical axis to the curve. It corresponds to a value of 28.

**Summary Assessment**

1. The following is a list of maximum temperatures (in degree Celsius) a South African town for the first three weeks of July 2018.

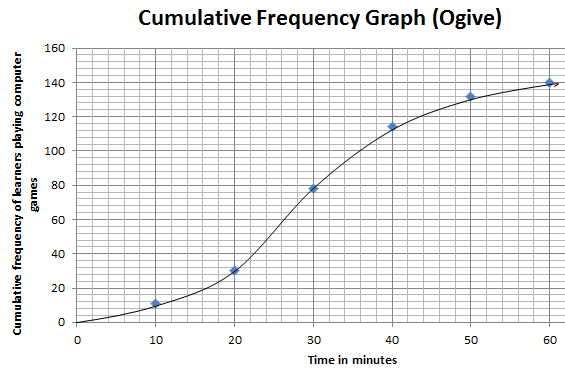
13; 14; 15; 14; 18; 16; 19; 11; 13; 16; 17; 15; 13; 19; 21; 23; 21; 15; 14; 17; 21

* 1. Complete the following frequency table:

|  |  |  |
| --- | --- | --- |
| **Temperature** | **Frequency** | **Cumulative frequency** |
| (10;12] |  |  |
| (12;14] |  |  |
| (14;16] |  |  |
| (16;18] |  |  |
| (18;20] |  |  |
| (20;22] |  |  |
| (22;24] |  |  |

* 1. Draw the Ogive curve from your cumulative frequency
  2. How would you describe the distribution of temperatures**?**

1. The cumulative frequency graph (ogive) drawn below shows the number learners spending time playing computer games.



* 1. Write down the total number of learners who played computer games during this hour
  2. Write down the modal class of the data
  3. How many learners played computer games in the first 30 minutes?
  4. What was the median for the data collected?
  5. How many learners came to play computer games in the last 15 minutes?
  6. Determine the 75th percentile for the data
  7. Calculate the interquartile range of the data

#### Guided reflection on Summary Assessment

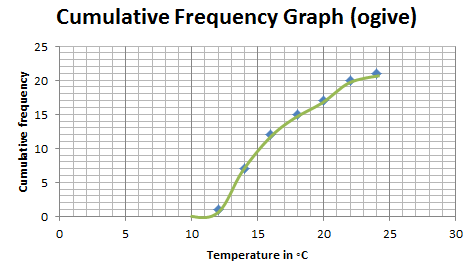
|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Write down the errors you made. 4. Have you been able to correct errors you made? |

**Answers for summary assessment**

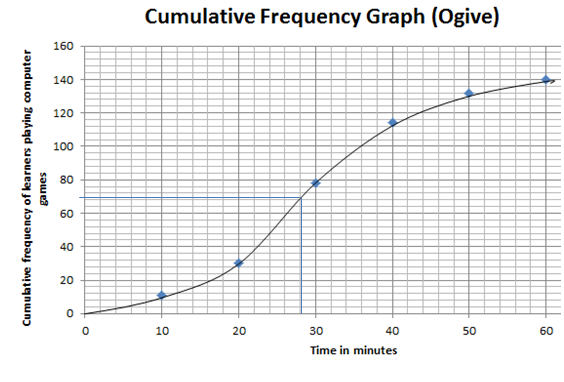
**1.1**

|  |  |  |
| --- | --- | --- |
| **Temperature** | **Frequency** | **Cumulative frequency** |
| (10;12] | 1 | 1 |
| (12;14] | 6 | 7 |
| (14;16] | 5 | 12 |
| (16;18] | 3 | 15 |
| (18;20] | 2 | 17 |
| (20;22] | 3 | 20 |
| (22;24] | 1 | 21 |

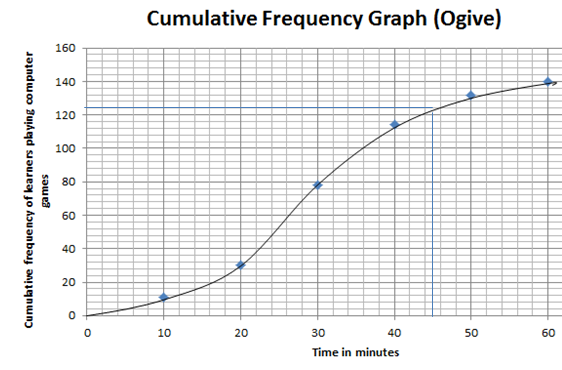
1.2



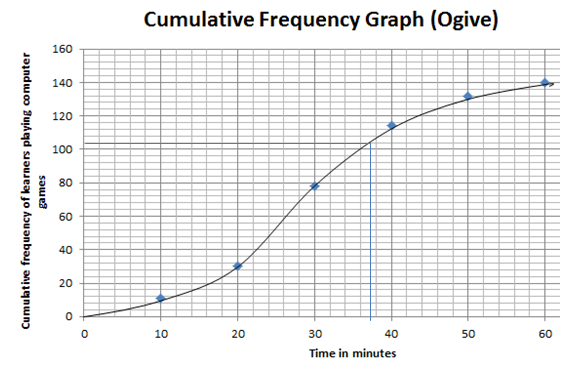
* 1. There are 15 days where the temperature is below . This is typical for the month of July as it is winter and higher temperatures are not normally expected. South Africa.
  2. The graph, the end of 60 minutes, goes to 140. Thus, 140 learners played computer games.
  3. The biggest number came from the 20th to the 30th minute. This is 79 – 30 = 49
  4. 79 learners played computer games in the first 30 minutes
  5. We draw a line from 70 on the vertical axis to the curve and then a line to the horizontal axis. The median is 28.



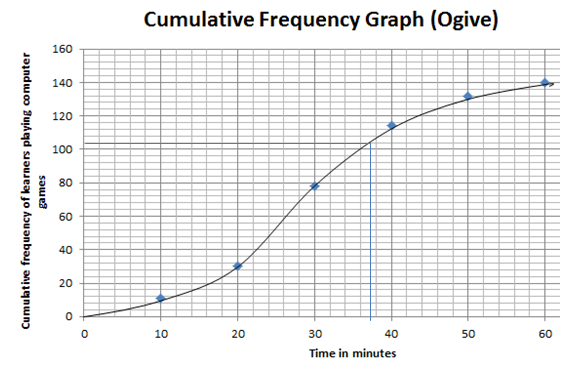
* 1. We draw a line from 45 on the horizontal axis to the curve below. We see it corresponds to 122 on the vertical axis. Thus, 140 – 122 = 18 learners came to play in the last 15 minutes,



2.6. The 75th percentile: we draw a line from 105 on the vertical axes to the curve; It corresponds to 37,6 minutes .



2.7 We know the 75th percentile is the upper quartile which is 37,6. We then draw a line from 35 on the vertical axis to the curve. It corresponds to 20 minutes on the horizontal axis.



The interquartile range is 37,6 minutes – 20 minutes = 17,6 minutes

**Unit 6: Variance and Standard Deviation of ungrouped data**

**Learning Outcomes**

By the end of this unit, learners should be able to

* Calculate the variance and standard deviation manually for small sets of data
* Calculate the variance and standard deviation through using calculators for larger sets of data;
* Interpret the meaning of the calculated variance and standard deviation of given numerical data.

**Activity 1:** Standard deviation and variance (for small sets of data)

**Purpose**:

* Show how the standard deviation 9and variance) is calculated for small sets of data

**Resources**: A pen or pencil, eraser, scientific calculator and paper

**Suggested time:** 45 minutes

*NB*: *The scientific calculator used in the calculations is the Casio fx-82ZAPlus (or similar). Please check the manual of your own calculator. There may be similar or slightly different calculator steps to the ones indicated here*

**Standard deviation** is a statistical measure of spread or variability. The standard deviation is the root mean square (RMS) deviation of the values from their arithmetic mean.

The formula for standard deviation is given by:



where Σ = Sum of  
              d = Individual score – mean   
              n = Sample size (Number of scores)

**Variance**  is the square of the standard deviation. It gives a measure of the degree of spread among a set of values; a measure of the tendency of individual values to vary from the mean value.

      **Variance = s2**

**Example**

The ages of learners in a school’s art club are:

14; 15; 16; 14; 16; 17; 15; 13.

1. Calculate the standard deviation.
2. How many learners’ ages are within one standard deviation of the mean?

**Solution**

Add the ages and divide by 8

1. Mean = 15

|  |  |  |
| --- | --- | --- |
| **Age** | **Deviation from the mean** | **(Deviation from the mean)2** |
| 14 | 14 – 15 = -1 | 1 |
| 15 | 15 – 15 = 0 | 0 |
| 16 | 16 – 15 = 1 | 1 |
| 14 | 14 – 15 = -1 | 1 |
| 16 | 16 – 15 = 1 | 1 |
| 17 | 17 – 15 = 2 | 4 |
| 15 | 15 – 15 = 0 | 0 |
| 13 | 13 – 15 = -2 | 4 |
| **Total** |  | **12** |

The **standard deviation** is given by: 

The **variance** = (1,22)2 =1,5

1. To get within one standard deviation of the mean we do the following calculations: 15 – 1,5 = 13,5 and 15 + 1.5 = 16,5. We thus looks at ages between 13,5 and 16,5. We see that 6 of the 8 learners’ ages are within one standard deviation of the mean.

**Task 1**

The following scores were obtained by learners in a grade 11 Accounting class:

71; 88; 47; 59; 93; 90; 74; 81; 52; 75.

1. For this set of scores find the
   1. mean
   2. standard deviation
2. Determine the percentage of learners with marks that are within one standard deviation of the mean

#### Guided reflection on Activity 1

|  |
| --- |
| * 1. What is the relationship between standard deviation and variance?   Standard deviation is the square root of variance   * 1. How would you determine the marks that are within one standard deviation of the mean?   Add the standard deviation value to the mean value to get the upper limit.  Subtract the standard deviation value from the mean to get the lower limit  So any mark that falls within the lower limit an upper limit lies within one standard deviation of the mean mark. |

**Answers for task 1**

1. Mean = = = 73

|  |  |  |
| --- | --- | --- |
| **Mark** | **Deviation from the mean** | **(Deviation from the mean)2** |
| 71 | 71 – 73 = -2 | 4 |
| 88 | 88 – 73 = 15 | 225 |
| 47 | 47 – 73 = -26 | 676 |
| 59 | 59 – 73 = -14 | 196 |
| 93 | 93 – 73 = 20 | 400 |
| 90 | 90 – 73 = 17 | 289 |
| 74 | 74 – 73 = 1 | 1 |
| 81 | 81 – 73 = 8 | 64 |
| 52 | 52 – 73 = -21 | 441 |
| 75 | 75 – 73 = 2 | 4 |
| **Total** |  | **2300** |

Standard deviation: 

1. We choose the marks between (73 – 15,16) and (73 + 15,16), that is, between 57,84 and 88,16. We see that 6 of the learners are in this interval; this equates to a percentage of ×100% = 60%

**Activity 2:** Standard deviation and variance (for larger sets of data)

**Purpose**:

* Show how the standard deviation and variance) is calculated for larger sets of data

**Resources**: A pen or pencil, eraser, calculator and paper

**Suggested time:** 45 minutes

In the previous activity, we worked with small data sets and we were able to determine the standard deviation and variance with simple calculations. However, there is no need to do these calculations manually as all scientific calculators are equipped to calculate the standard deviation and variance. Please study the manual of your scientific calculator carefully and practice using smalls sets of data.

Let us look at example 1 from the first activity. Here the ages of learners were given as follows:

We can use our scientific calculator to compute the standard deviation::

14; 15; 16; 14; 16; 17; 15; 13

Set your calculator to STATS mode; enter the above data. After entering each number, press =. To retrieve the data press AC shift 1, 4, 3

We find that the standard deviation is 1,22

The variance is (1,22)2 = 1,5

**Example**

The following set of data, which was used in unit 4 (task 1), represents the marks of learners in a Science test.

08, 13, 16, 17, 18, 19, 19, 22, 22, 23, 23, 23, 25, 25, 27, 28, 28, 29, 31,

31, 32, 33, 33, 34, 35, 36, 37, 38, 39, 39, 39, 40, 40, 41, 41, 42, 43

**Determine**

1. Determine the standard deviation, using your calculator

2. Hence, calculate the variance

3. How many learners’ marks were within one standard deviation of the mean? Write this as a percentage of the total data

**Answers**

1. Set the calculator to STATS mode; enter the above data (37 items)

After entering each number, press =.

To retrieve the data press AC shift 1, 4, 3

We find our standard deviation to be 9,09

1. Variance = (9,09)2 = 82,63
2. We add all the numbers and divide by 37 or we retrieve our mean from our calculator by pressing press AC shift 1, 4, 2. Our mean is 29,43

We subtract 9,09 from 29,43 to get 20,34; we add 9,09 to 29,43 to get 38,52. So we look at marks between 20,34 and 38,52; there are 22 learners’ marks which were within one standard deviation of the mean (59,46% of the data) .

**Task 2**

The marks of 20 learners’ mathematics marks (in a test out of 30) are shown below:

4 7 8 9 11 12 13 13 15 15

16 18 18 19 20 22 24 25 26 27

* 1. Determine the standard deviation, using your calculator
  2. How many learners’ marks were within two standard deviations of the mean? Write this as a percentage of the total data?

#### Guided reflection on Activity 2

|  |
| --- |
| Briefly write down the steps that you will use on your calculator to determine the standard deviation. |

**Answers for Task 2**

* 1. We set our calculator to STATS mode; enter our data. Then press AC shift 1, 4, 3. Our standard deviation is 6,46
  2. To get our mean, in STATS mode, press AC shift 1,4,2. Our mean is 16,1

Now we subtract 6,46 twice (2 standard deviations) from 16,1 to get 3,18

We add 6,46 twice (2 standard deviations) to 16,1 to get 29,02 . All the data values lie between 3,18 and 29,02 . All 20 learners (100%) are within two standard deviations of the mean.

**Summary assessment**

* 1. Two mathematics examinations were given. In the first one, the mean was 64 and the standard deviation was 6. In the second one, the mean was 56 and the standard deviation was 4. The scores of 4 learners who wrote both examinations are listed below:

|  |  |  |
| --- | --- | --- |
| **Name** | **Examination 1** | **Examination 2** |
| Mary | 73 | 68 |
| Andile | 79 | 71 |
| Sipokazi | 67 | 58 |
| Johan | 58 | 64 |

* 1. Who improved in the second examination?
  2. Who improved the most in the second examination?
  3. Who did not improve in the second examination?
  4. Considering both examinations, which student did the poorest?
  5. Who performed the same in both examinations?
  6. Taking both examinations into account, which student has the best overall mark?
  7. The following marks were obtained by learners in a Geography test.

9, 10, 12, 15, 17, 18, 19, 20, 22, 23, 23, 23, 23, 24, 26,

26, 27, 27, 29, 28, 29, 29, 33, 33, 34, 35, 35, 37, 39, 40

**Determine**

2.1 Determine the standard deviation, using your calculator

2.2 Hence, calculate the variance

2.3 How many learners’ marks were within one standard deviation of the mean?

Write this as a percentage of the total data

#### Guided reflection on Summary Assessment

|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Write down the errors you made. 4. Have you been able to correct errors you made? |

**Answers for summary assessment**

* 1. Johan improved in the second examination as he went from a ranking of 4 to a ranking of 3
  2. Johan improved the most in that his marks went up by 6
  3. Sipokazi did not improve in the second examination; her ranking was the same
  4. In the first examination Johan was one standard deviation from the mean with the lowest mark; in the second examination his mark was above the mean. However, when adding both marks, his total of 122 is below that of Sipokazi (125). So we may conclude that Johan did the poorest
  5. Both Mary and Andile performed the same in both examinations (in terms of their ranking)
  6. In both examinations, Andile was more than one standard deviation above the mean in both examinations
  7. We set our calculator to STATS mode; enter our data. Then press AC shift 1, 4, 3. Our standard deviation is 8,13
  8. Variance = (8,13)2 = 66,1
  9. The data is already in the memory of the calculator. Press AC shift 1,4, 2 to get the mean. The mean is 25,5. Since the standard deviation is 8,13, we look at marks between 25,5 – 8,13 and 25,5 + 8,13, that is, between 17,37 and 33,63. We count these marks; there are 19 of these marks. The percentage is = 63,3%

**Unit 7: Represent, Analyse and Interpret Data Using Various Techniques**

**Learning Outcomes**

By the end of this unit, learners should be able to

* Identify situations or issues that can be dealt with through statistical analysis:
* Discuss the use of appropriate and efficient methods to record, organise and interpret given data.
* Justify and apply statistics to answer questions about problems.

**Activity 1: Identification of situations or issues which may be analysed statistically**

**Purpose**:

* Identification of situations where data is collected
* Decide how to analyse data in a statistical manner

**Resources**: A pen or pencil, eraser, calculator, paper, newspaper, access to internet

**Suggested time:** 45 minutes

**Introduction**

There are various issues or questions which we may face in our daily lives. Some of these issues may be personal or family related. Other issues may be faced by you at school or in the community. It is important for us to pose questions about these issues and investigate them. In mathematics we tend to focus more on the numerical data that emerge from our investigations. Numerical and other data should be used to support claims made in our investigations

**Collection of mathematical data**

Mathematical or numerical data can be collected in several ways. These include:

* Examining and interpreting figures/graphs/tables etc. that appears in newspapers, magazines and journals.
* Information from radio and television news.
* Conducting surveys of groups of people by questionnaires or interviews. These people may be your neighbours or members of your community, members of your class/school, your family, church, etc.
* Consulting with experts such as teachers, university lectures, and others.
* Doing your own or group experiments and making notes of the results in table or graphic form.
* Using internet search engines.
* Using relevant books from the library.
* Information from government departments e.g.) Health, education, safety and security, communication, etc.
* Information from government agencies such as HSRC.
* Getting information from non-governmental agencies; International agencies of the United Nations, etc.

The data that has been collected must be handled in such a way that it can convey to meaningful messages to masses of people.

**Discussion of different issues**

Study the following issues with friends or family members. Suggest ways in which the questions posed for each issue could be addressed.

For each issue discussed, your response should include:

* The relevant numerical data that should be collected
* How this data should be collected
* To whom should your findings be reported

1. You have observed that some children in your city/town are standing at busy street intersections or at different shops in your community and are begging for food or money. It is clear to you that these children should be at school. However, due to their home circumstances, they are forced to beg. Thus, one of their basic human rights, that is, their right to education is being violated. How would you go about investigation this situation?

1. One of many social problems facing communities in South Africa is alcohol abuse. Many children know of relatives or other people that have abused alcohol. It would appear that alcohol abuse often leads to child abuse. How would you go about investigating the incidence of child abuse in your community, city or town?
2. News reports show that the South African economy is booming and that investment in our country has been increasing steadily over the past few years. However, it would appear that there are large numbers of people in your community or town that are unemployed. It would appear that this unemployment is contributing to an increase in burglaries of homes in your neighbourhood. How would you investigate this issue?
3. A river/pond or pool of water in your community is polluted and that small children are playing in these polluted waters. As a result, some children have developed certain skin diseases and other waterborne illnesses. How would you investigate this issue? What would you recommend to the relevant authorities?
4. One of the major problems in our country at the moment is service delivery to the people, especially the poor people. It has been discovered that money earmarked for poverty relief in a certain community has been “stolen” by a corrupt government official. The people in community are upset and demand action from government. What should the government do?

**Possible responses**

|  |  |  |  |
| --- | --- | --- | --- |
| **Issue** | **Relevant numerical data to be collected** | **How data should be collected** | **To whom should findings be reported** |
| 1 | Number of children begging | Observation; going to various busy intersections and shops and counting the number of children that are begging | Child welfare; Department of Social Development |
| 2 | Number of children affected by child abuse | Visit places of safety; hospitals, etc; interview children that are known to have alcoholic parents/relatives; interview social workers/nurses to find out the link between alcohol abuse and child abuse | Child Welfare, Department of Social Development; Police |
| 3 | Incidence of burglaries in your neighbourhood | Interviewing the local police; checking police statistics for your neighbourhood, etc; see if there is a link between unemployment and burglaries in your area. | Local councillors, labour officers; members of parliament, etc |
| 4 | Children affected by the polluted water | Interviewing doctors and nurses or sending out questionnaires to medical centres; | Environmental affairs |
| 5 | Corrupt government officials | Using the internet; checking newspaper reports; interviewing people affected by poor service delivery | Local councillors; members of parliament; human rights commission |

**Task 1**

1. For each of the following questions, you are given issues that need to be investigated. For each issue, you should indicate what numerical data is required in order to investigate the issue, how you would collect the data, and to whom should the data be reported. Complete your response in a table form as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Question** | **Data to be collected** | **How the data should be collected** | **To whom should your findings be reported** |
| 1.1 Vandalism in schools |  |  |  |
| 1.2 The high price of petrol |  |  |  |
| 1.3 The state of roads in your neighbourhood |  |  |  |
| 1.4 The incidence of homeless people in your city/town |  |  |  |
| 1.5 Drug abuse among teenagers in your community |  |  |  |

2. List any other issues that should be investigated in your

2.1 school

2.2 community

2.3 city/town

2.4 province

**Possible answers to Task 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **Question/issue** | **Data to be collected** | **How the data should be collected** | **To whom should your findings be reported** |
| 1.1 Vandalism in schools | Incidence of vandalism at schools | Interviews; newspaper reports, observations, etc | School principal; SGB; District office; police |
| 1.2 The high price of petrol | Price of petrol for a say, 6 months | Internet, newspaper, personal experiences | Councillor; member of parliament; ratepayers association |
| 1.3 The state of roads in your neighbourhood | Roads that are in need of repair | Through observations; speaking to other motorists from your area, etc | Councillor; municipal; roads department |
| 1.4The incidence of homeless people in your city/town | Number of homeless people in city/town | Through observations; social agencies; internet | Local; provincial and national housing departments (human settlements) |
| 1.5 Drug abuse among teenagers in your community | Identification of teenagers who indulge in drugs | Through observation; social agencies; medical reports | Drug treatment centres |

2.1 Learners mode of travel to school

2.2 Incidence of poverty

2.3 Road safety

2.4 Unemployment

**Activity 2: Specific data collection methods and sources: Survey from the internet**

**Purpose**:

* Accessing a survey from a website
* Analysing the statistical data in the survey

**Resources**: A pen or pencil, eraser, calculator, paper, newspaper, access to internet

**Suggested time:** 45 minutes

We have seen that data can be collected in a variety of ways.

The following report Victims of Crime Survey (VOCS) was extracted from the internet using the website [www.statssa.gov.za](http://www.statssa.gov.za).

*“The general level of crime as estimated by VOCS has been declining during the past five years but increased in 2016/17 and 2017/18. Household crimes increased by 5% to a total of 1,5 million incidences of crime while individual crime also increased by 5% to a total of 1,6 incidences, affecting 1,4 million individuals aged 16 and above. Northern Cape had the highest increase in both household and individual crimes. Housebreaking or burglary was the most dominant (54%) crime category among crimes measured by the Victims of Crime Survey (VOCS). An estimated total of 830 thousand incidences of housebreaking occurred in 2017/18, affecting 4,25% of all South African households. Nearly 32% of items stolen during housebreaking were clothes, followed by cell-phones (24%) and food (22%).*

*An estimated 156 thousand home robberies occurred last year, affecting 0,8% of all South African households. This was an increase of 3% compared to the previous year. Theft of livestock, poultry and other animals which occurred in 2017/18, is estimated at 159 thousand incidences, affecting 0,77% of households in South Africa. The number of incidences increased by 1% compared to the previous year.*

*Murder increased during the past three years both in terms of the total number of incidences (VOCS) and the number of cases reported to the police (SAPS). It is estimated that 16 809 incidences of murder occurred in 2017/18, which is an increase of 4% compared to the previous year. Usually VOCS figures are higher than SAPS figures because not all crime incidents are reported to the police. This is not the case for murder, where SAPS figures are double the VOCS figures. Not all murders are known to households, for example murders of immigrants or street people, which is one of the reasons that SAPS figures are higher than those of VOCS.*

*The percentage of households who think that the levels of violent crimes increased during the past three years is greater for 2017/18 (42,1%) when compared to the figure for 2016/17 (39,4%). The percentage of households who feel safe walking in their neighbourhoods during the day declined from 84,8% in 2016/17 to 79,1% in 2017/18. The level of satisfaction with the police and courts continued to decline, more rapidly in the case of courts.*

*The percentage of households who were satisfied with the police services in their area decreased from 57,3% in 2016/17 to 54,2% in 2017/18. The percentage of households who were satisfied with the way courts generally deal with perpetrators decreased from 44,9% in 2016/17 to 41,1% in 2017/18. Satisfaction with the police declined in every province except the Western Cape and Free State, while satisfaction with the courts declined in every province except the North West”.*

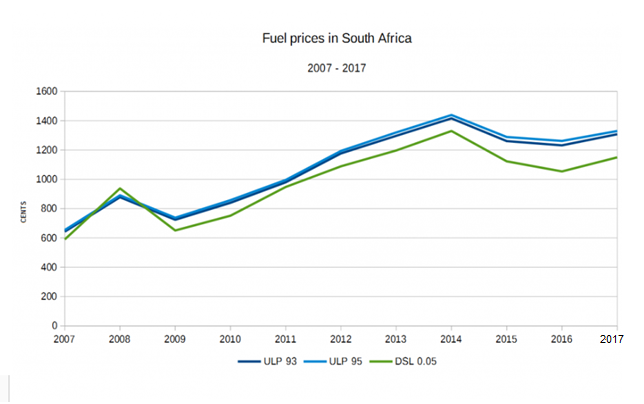
Study the above survey. It makes reference to crimes which have been committed in South Africa using actual numbers and percentages. List 5 of these and explain what the numbers or percentages convey to the reader.

**Possible answers**

|  |  |
| --- | --- |
| **Details from survey** | **Meaning/ interpretation** |
| Declining crime rate in past 5 years, although there was an increase from 2016/2017 to 2017/2017 | Overall decrease but increase in the last year; residents should be aware about crime in their area as well as in the country |
| 1,5 million incidences of household crime; 1,6 million incidences of individual crime | High number of household and individual crimes; residents should take precautions in order not to fall victim to crime |
| Housebreaking or burglary (54%) was the dominant category | More than half the crime was linked to households and other places – housebreaking and burglary; need for precaution |
| 16809 incidences of murder in 2017/2018 | High rate of murder in the country; people need protection |
| Different in figures given by VOCS as opposed to SAPS | Calculations are done differently; but there is still crime, irrespective which agency is providing the data |
| Decrease in satisfaction with police services | This decreased from 57,3% in 2016/17 to 54,2% in 2017/18; Police services needs to regain the confidence of the people |
| Satisfaction with the courts to deal with perpetrators of crime | This decreased from 44,9% in 2016/17 to 41,1% in 2017/18; or courts need to be more strict on perpetrators of crime |

**Task 2**

Study the following graph which was accessed from the website businesstech.co.za**.** it shows the price of different brands of fuel in South Africa in the period 2007 – 2017**.**



1. Besides 2007, in which year was the average price of diesel the lowest?
2. During which period did the fuel price (all brands) increase significantly? Explain how you arrived at your answer.
3. In 2018, the price of ULP 93 was an average of R16,50 per litre, compared to an average price of the same brand of R13,00 per litre in 2017. Calculate the percentage increase in the price of ULP 93

**Answers to Task 2**

1. We see that the price went up in 2008 and then dipped in 2009
2. After dipping in 2009; the prices went up until 2014, when there was a dip. All three graphs have a positive gradient
3. Percentage increase =

**Activity 3: Unemployment rates**

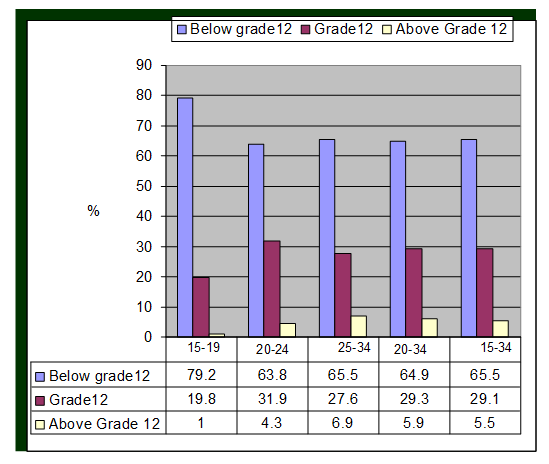
**Purpose**:

* Analysis of data presented in bar graphs and tables

**Resources**: A pen or pencil, eraser, calculator, paper, newspaper, access to internet

**Suggested time:** 45 minutes

Examine the unemployment statistics in South Africa presented in both graph and table form and answer the questions that follow. (from (Unemployed Youth in South Africa – The Disturbed Generation: Dr, Reinette Du Toit)



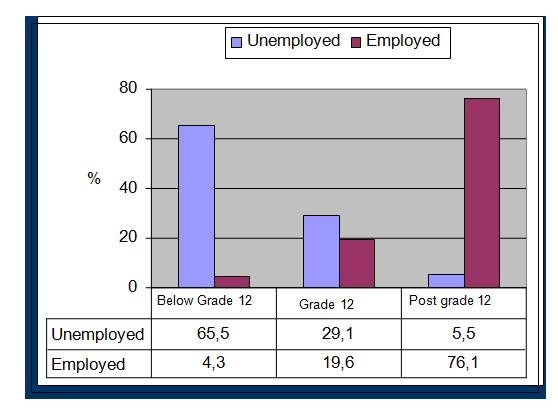
1. Name the type of graph that is used
2. What is presented on the vertical axis?
3. What is presented on the horizontal axis?
4. What age groups are represented in the data?
5. On which of the two factors, age or education, does employment depend more? Give a reason for your answer.
6. Why is the first column for the first interval longer than the first column for the other intervals?
7. Name an important life decision that you ought to take based on the above.

**Answers**

1. Bar graphs show three bars depicting three categories of qualifications those below grade 12; those with grade 12 and those above grade 12
2. Unemployment percentage
3. Ages of the youth
4. 15 – 19; 20 – 24; 25 – 34; 20 – 34; 15 – 34
5. Employment depends on education; we note that percentage unemployment of youth with grade 12 and above qualification is the lowest followed by those with grade 12
6. It represents those without grade 12; their unemployment rate is significantly higher than those with grade 12 and above
7. It is best to complete your schooling career and then study further so that your prospects of employment is increased appreciably

**Task 3**

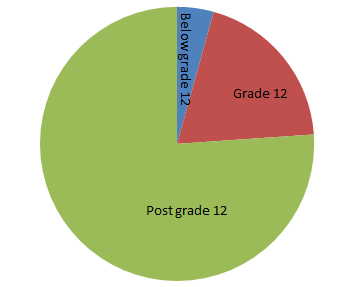
Study the graph below which presents another perspective on unemployment rates in South Africa (among youth) and then answer the questions that follow: From (Unemployed Youth in South Africa – The Disturbed Generation: Dr, Reinette Du Toit)



1. In what way is the form of the second graph different from the first?
2. What is represented on the vertical axis?
3. Give a reason why the percentages of unemployed and employed do not add up to 100% in any of the three cases.
4. Suppose a cousin of yours approaches you with the news that he intends leaving school at the end of Grade 10. How would you use this graph to convince him otherwise?
5. Use the above data to draw a pie chart comparing the percentages of employed persons in the groups represented in the diagram

**Answers to Task 3**

1. It looks at unemployed/employed for the three sets of qualifications, without the age categories
2. Percentage of employed/unemployed youth
3. In the groups at grade 12 and below, some of the youth will still be at school, possibly in grades 9 to 12. Those with a post grade 12 qualification may also be studying further for post-graduate or other qualifications; they are full time students.
4. Advise him/her to complete grade 12 and study further; the chances of employment increases appreciably
5. We will look at the following percentages of employment: 4,3%; 19,6%; 76,1% and draw the corresponding pie chart for this data:



**Activity 4: Misleading data and “bias” graphs**

**Purpose**:

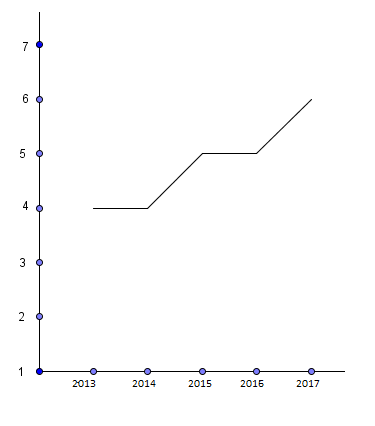
* Comparing graphs for misleading data and bias

**Resources**: A pen or pencil, eraser, calculator, paper, newspaper, access to internet

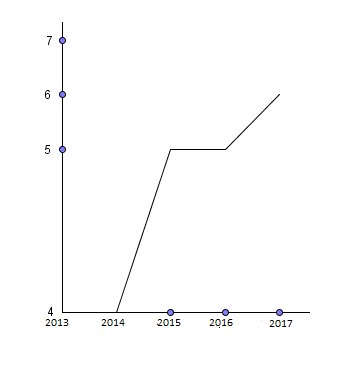
**Suggested time:** 45 minutes

Data is often used in a misleading way by unscrupulous people. The following two b line graphs both represent the profit of a small company in units of R100 000. Compare the two graphs in terms of how the data is presented and in terms of the overall impression created by each graph.

**Graph 1**



**Graph 2**



1. Briefly discuss how the two graphs differ from each other.
2. Which one of the two graphs do you think the company would use in its advertising campaign? Why?
3. Which one of the graphs presents a more realistic picture of the true state of affairs? Why?

**Answers**

1. Graph 1 has the vertical axis clearly demarcated, with equal spacing; graph 2 starts at 400000 and goes to 700000. However, the length from 400000 to 500000 is bigger or more stretched out compared to the units from 500000 to 700000 which are equally spaced.
2. It may use graph 2 as it seems to indicate that its profits has been significantly increasing (except in 2015 and 2016)
3. Graph 1 is more realistic; its axes are well calibrated. The information always portrays a more steady increase in the profits

**Task 4**

1. The statistics for the number of citizens, of a particular country, who had access to clean water and sanitation were as follows:

1998: 65%

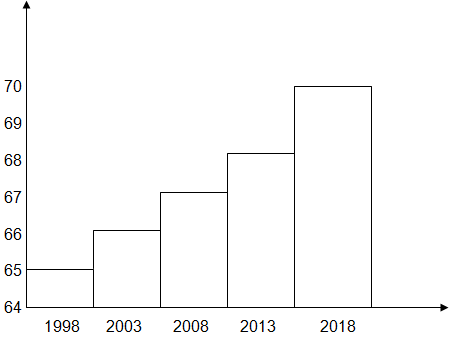
2003: 66%

2008: 67%

2013: 68%

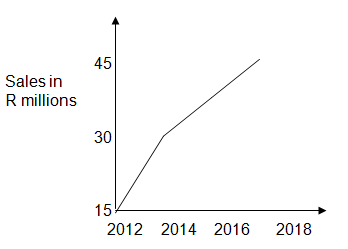
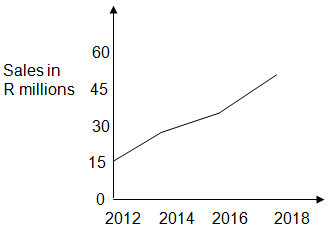
2018: 70%

This information is shown in the bar graph below



The ruling party used the above graph in their election campaign after they had been in power for 25 years. Was this honest? Motivate your answer

2. Below are two line graphs.



2.1 Do these two graphs refer to the same situation? Do they tell the same story?

2.2 Which graph should be used in a newspaper advertisement to boost the sales of the company?

2.3 How has the better image been achieved?

2.4 Is there anything about the more flattering graph which is false?

1. Simunye High School entered two teams for the regional Mathematics Olympiad. The marks (out of 50) were as follows:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Team A | 24 | 27 | 31 | 26 | 30 | 47 | 24 | 19 | 29 | 23 |
| Team B | 22 | 30 | 24 | 27 | 32 | 31 | 23 | 24 | 28 | 29 |

The following measures of central tendency were also given for both teams

|  |  |  |
| --- | --- | --- |
|  | Team A | Team B |
| Range | 28 | 10 |
| Mode | 24 | 24 |
| Median | 26,5 | 27,5 |
| Mean | 28 | 27 |

Study the marks and the measures of central tendency and state which of the two teams performed better in the Mathematics Olympiad.

**Answers to Task 4**

1. The graph gives the impression that there has been a steady upward increase of those with access to clean water and sanitation; it portrays a very favourable image of what has been accomplished. However, the reality is that is has only gone up by 5 % in 20 years. This a not a major accomplishment as the population would have increased in the meantime; as a result more people would need to get access to clean water and sanitation.

2.1 Yes – the graphs refer to the same situation. Both graphs start at 15 million on the vertical axis; then go upwards. However, the graphs do not tell the same story. The first graph shows a “gradual” increase in sales while the second graph shows a “steeper” increase in sales.

2.2 The second graph should be used as it gives a more “favourable” image of the company’s sales; this is likely to influence customers that its products are very popular and it may attract more customers

2.3 The better image has been achieved by starting at 15 million and having a bigger scale on the vertical axis

2.4 The vertical axis starts at 15 million rather than zero; this gives the impression of a substantial increase in sales

1. We may draw the following conclusions from the table of statistics:

* The range of team A is very high because one team member scored very well (47 marks) and one scored poorly (19 marks)
* Both modes are the same
* Team B’s median mark is marginally better
* Team A’s mean mark is slightly better

One could say that team A did slightly better but its performance is distorted by one high mark of 47. This is called an “extreme” value.

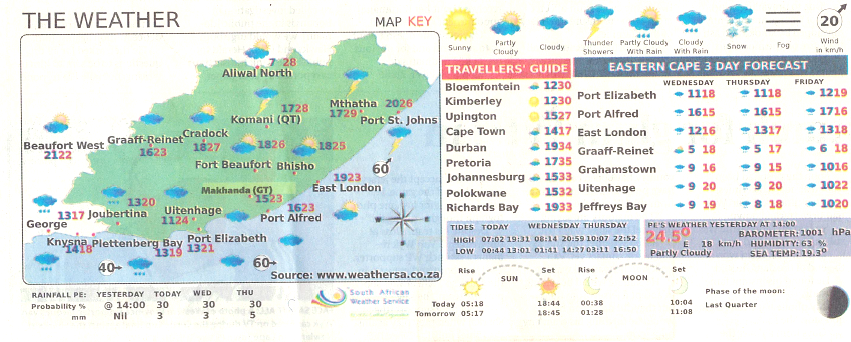
Team B’s performance was more consistent, with a much smaller range and a better median. Although its highest mark is lower than that of team A, all other performances are either on par or better than team A. This evidence is shown in the table below which shows the marks in descending order:

|  |  |
| --- | --- |
| **Team A** | **Team B** |
| 47 | 32 |
| 31 | 31 |
| 30 | 30 |
| 29 | 29 |
| 27 | 28 |
| 26 | 27 |
| 24 | 24 |
| 24 | 24 |
| 23 | 23 |
| 19 | 22 |

On closer scrutiny, we may conclude that Team B did better.

**Summary assessment**

1. Study the weather report which appeared in a local Eastern Cape newspaper on 30 October 2018



* 1. Make a list of some of the key information that is given in this report.
  2. The minimum and maximum temperatures for 10 cities/towns of South Africa are given:

|  |  |  |
| --- | --- | --- |
| **Town/city** | **Minimum temperature in degree Celsius** | **Maximum temperature in degree Celsius** |
| Bloemfontein |  |  |
| Kimberley |  |  |
| Upington |  |  |
| Cape Town |  |  |
| Durban |  |  |
| Pretoria |  |  |
| Johannesburg |  |  |
| Polokwane |  |  |
| Port Elizabeth |  |  |
| Richard’s Bay |  |  |

1.2(a) Determine the mean maximum and minimum temperatures for the 10 cities

(b) Write down the mode of the maximum temperatures

(c) Determine the median of the minimum temperatures

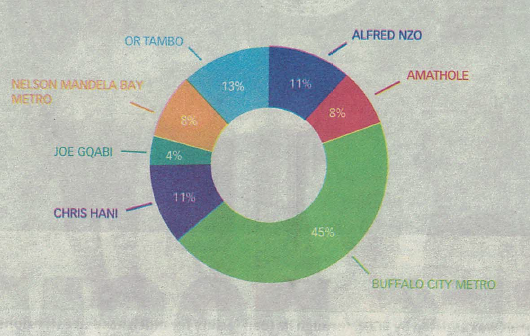
(d) Hence, determine the interquartile range of the minimum temperatures

1.3 Sunrise in Port Elizabeth was at 05:18 and sunset was at 18:44. Determine in terms of hours and minutes, the total amount of sunshine for this particular day

1.4 Which city/town in the Eastern Cape had the biggest range in temperature? What was this range?

1.5 The probability for rainfall for the day was given as 30%. What does this mean?

1. Study the diagram below which shows the allocation of funding and support (as a percentage) by the Eastern Cape Development Corporation to youth owned SMMEs (Small and Medium Enterprises) to a number of municipalities in the Eastern Cape. The total amount allocated was R98,3 million. This information was published in the Khula Nathi newspaper supplement for October 2018.

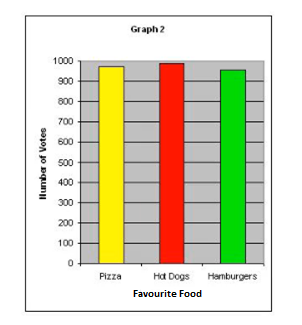
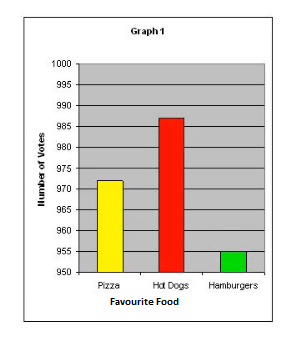


2.1. Determine the range in the percentages of funding and support. Then calculate the actual range.

2.2 Which municipality received the smallest percentage allocation? What was this amount?

2.3 Buffalo City metro (which incorporates the key cities/towns of East London and King Williamstown) received 45% of the allocation. Calculate this amount and state why you think such a large amount was allocated to this municipality?

1. A fast food establishment sells a number of foods. Three of these foods are shown in the two graphs below.



* 1. Do the graphs represent the same information? Explain
  2. Graph 1 accentuates a specific favourite food which is not easily seen in graph 2. Explain how this is done
  3. In order to promote sales of all three foods, which graph will be likely to be used in an advertising campaign? Give a reason for your answer

#### Guided reflection on Summary Assessment

|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Write down the errors you made. 4. Have you been able to correct errors you made? |

**Answers for summary assessment**

* 1. It shows minimum and maximum temperatures, rainfall, sunrise and sunset; moon phase; wind direction; etc
  2. Mean minimum:

Add all up:

+ + + + + + + + + =

Divide by 10 to get a mean minimum temperature of

Mean maximum

Add all up:

+ + + + + + + + + =

Divide by 10 to get a mean minimum temperature of

1.3 Median of the minimum temperatures:

Arrange in ascending order:

; ; ; ; ; ; ; ; ;

Median = =

First quartile =

Third quartile =

Inter-quartile range = - =

* 1. It is Aliwal North. It had a maximum of and a minimum of ; giving a range of
  2. The probability of 30% rainfall means that in the past when there were similar weather conditions to the one predicted for this day, it rained 30% of the time

2.1 Range = 45% - 4% = 41% 41% of R98,3 million = R40,3 million

2.2 Joe Gqabi which received 4% of R98,3 million = R3,93 million

2.3 45% of R98,3 million = R44.235 million. It is possible that there were more youth owned SMMEs which needed support; and the number of youth owned SMMEs may be increasing

* 1. Yes. The information on the horizontal and vertical axes are the same in both cases
  2. Graph 1 accentuates or emphasises the sales of hot dogs. This is done by making the scale on the vertical axis go up in 5s. In graph 2, the scale goes up in hundreds; it shows the sales of the food items to be very close to each other.
  3. Graph 2 will likely be used as it shows the popularity of all three food items; sales of all three food items will boost the overall sales at the establishment

**Unit 8: Bivariate data: Scatter plots, regression lines and correlation**

**Learning Outcomes**

By the end of this unit, learners should be able to

* Represent bivariate numerical data as a scatter plot and suggest intuitively and by simple investigation whether a linear, quadratic or exponential function would best fit the data
* Use a calculator to determine the linear regression line which best fit a given set of bivariate numerical data
* Use the regression line to predict the outcome of a given problem
* Use a calculator to calculate the correlation coefficient of a set of bivariate numerical data and make relevant deductions.

**Activity 1:** Representing bivariate data as a scatter-plot

**Purpose**:

* Draw a free-hand scatter-plot for given bivariate data
* Use excel to represent bivariate data and determine whether the a linear, quadratic or exponential graph would fit the data

**Resources**: A pen or pencil, eraser, calculator and paper

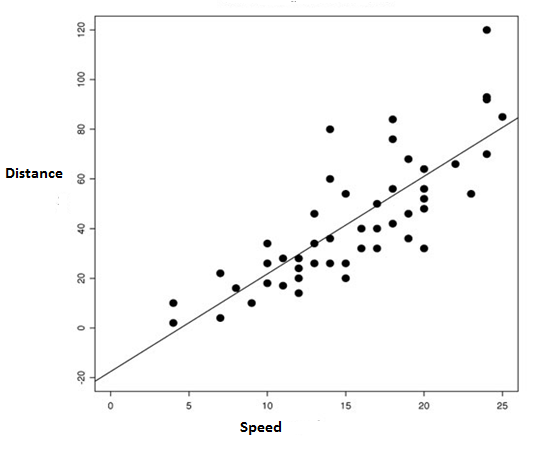
**Suggested time:** 45 minutes

In [statistics](https://en.wikipedia.org/wiki/Statistics), **bivariate data** is data with two variables where each value of one of the variables is paired with a value of the other variable. Typically it would be of interest to investigate the possible relationship between the two variables. This relationship can be studied via a tabular or graphical display, or via sample statistics which might be used for inference. The method used to investigate the relationship would depend on the level of measurement of the variable.

For two quantitative variables (ordinal, interval, or ratio in level of measurement) a scatterplot can be used and a correlation coefficient or regression model can be used to quantify the relationship

**Scatterplots**

Study the scatterplot plot below:



Line of best fit

Outlier

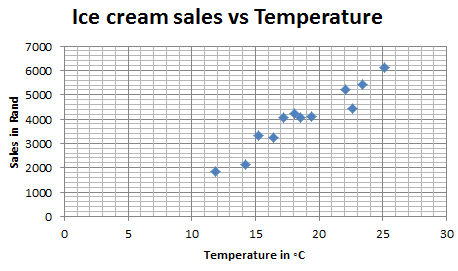
This scatter plot displays the relationship between Distance and Speed (two variables). A **line of best fit** is drawn. There appears to be a positive correlation between distance and speed. One **outlier** is identified.

**Example**

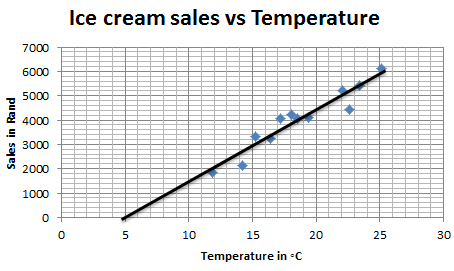
The table below shows the relationship between the temperature in **°C** and ice-cream sales in rand

|  |  |
| --- | --- |
| ***Ice Cream Sales versus Temperature*** | |
| **Temperature °C** | **Ice Cream Sales** |
| 14.2° | R2150,00 |
| 16.4° | R3250,00 |
| 11.9° | R1850,00 |
| 15.2° | R3320,00 |
| 18.5° | R4060,00 |
| 22.1° | R5220,00 |
| 19.4° | R4120,00 |
| 25.1° | R6140,00 |
| 23.4° | R5440,00 |
| 18.1° | R4210,00 |
| 22.6° | R4450,00 |
| 17.2° | R4080,00 |

We can plot this information on a graph and draw the line of best fit manually.



We may also use excel to draw the graph and then extend to the x-axis as shown below



Whether we draw the line of best fit manually or using excel, we note that there line of best fit indicates that when the temperature increases (it gets warmer), the sales of ice-cream increases. At the same time, when the temperature drops (it gets colder), the sales drop. When using excel to draw the line of best we are able to use the “add trendline option” to get the “approximate” equation of the line. The equation is

**Activity 2:** The regression line and correlation coefficient

**Purpose**:

* Use a scientific calculator to determine the equation of the line of best fit for a scatterplot
* Use a scientific calculator to determine the correlation coefficient for a line of best fit

**Resources**: A pen or pencil, eraser, scientific calculator and graph paper

**Suggested time:** 45 minutes

*NB*: *The scientific calculator used in the calculations is the Casio fx-82ZAPlus (or similar). Please check the manual of your own calculator. There may be similar or slightly different calculator steps to the ones indicated here*

Let us consider the same data from activity 1 (on temperature and ice-cream sales).

|  |  |
| --- | --- |
| ***Ice Cream Sales versus Temperature*** | |
| **Temperature °C** | **Ice Cream Sales** |
| 14.2° | R2150,00 |
| 16.4° | R3250,00 |
| 11.9° | R1850,00 |
| 15.2° | R3320,00 |
| 18.5° | R4060,00 |
| 22.1° | R5220,00 |
| 19.4° | R4120,00 |
| 25.1° | R6140,00 |
| 23.4° | R5440,00 |
| 18.1° | R4210,00 |
| 22.6° | R4450,00 |
| 17.2° | R4080,00 |

**Regression line**

A linear **regression line** has an equation of the form where is the explanatory variable and is the dependent variable. The regression line is the line of best fit for a scatterplot.

To find the equation of the regression line for the above table of values we go through the following steps:

1. Press [MODE] and then select [2: STAT]
2. Select [2: A+BX] for linear regression
3. Now enter the bivariate data, by entering the [X / Y] value and [=] for all x and y data points
4. Press [AC] to clear the screen and store the data values
5. Press [SHIFT] [1] to get to the stats menu
6. Select [5: REG] for linear regression
7. To get the value of (y-intercept), select [1: A] and [=].

Answer:

1. Press [AC] to clear the screen.
2. Press [SHIFT] [1] to get to the stats menu
3. Select [5: REG] for linear regression
4. To get the value of (gradient), select [2: B] and [=].

Answer:

NB: The equation (obtained by using a scientific calculator) would appear to be a more accurate equation than the equation (obtained from using excel)

**Correlation coefficient**

The **correlation coefficient** (*r*) is a statistical measure that calculates the strength of the relationship between two variables.

The correlation coefficient can assume the following values: … where :

* -1 indicates a negative and strong correlation;
* 0 indicates no correlation; and
* 1 indicates a positive and strong correlation

To find the **correlation coefficient** for the table of values (on ice-cream sales) we go through the following steps:

1. Press [MODE] and then select [2: STAT]
2. Select [2: A+BX] for linear regression
3. Now enter the bivariate data, by entering the [X / Y] value and [=] for all x and y data points
4. Press [AC] to clear the screen and store the data values
5. Press [SHIFT] [1] to get to the stats menu
6. Select [5: REG] for linear regression
7. To get the value of (correlation coefficient), select [3: r] and [=].

Answer:

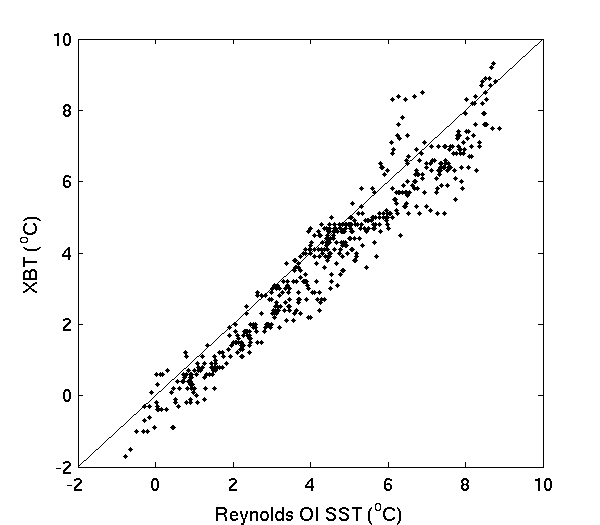
1. Press [AC] to clear the screen.

We have which is very close to 1 so we have a strongly positive correlation.

If r was close to -1 then then the correlation would have been strongly negative.

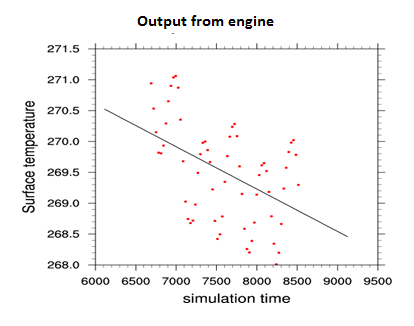
If r is close to 0, then there would be a very negligible or no correlation

Please note the following scatter plots and correlation coefficients for three different scenarios:

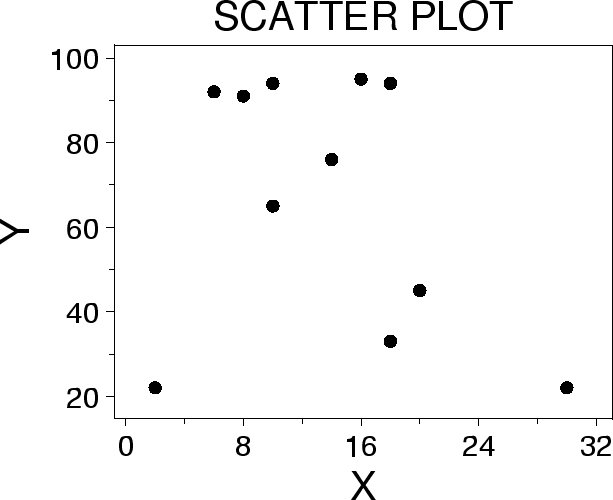


Positive and strong correlation

2**.**



Negative and weak correlation

****

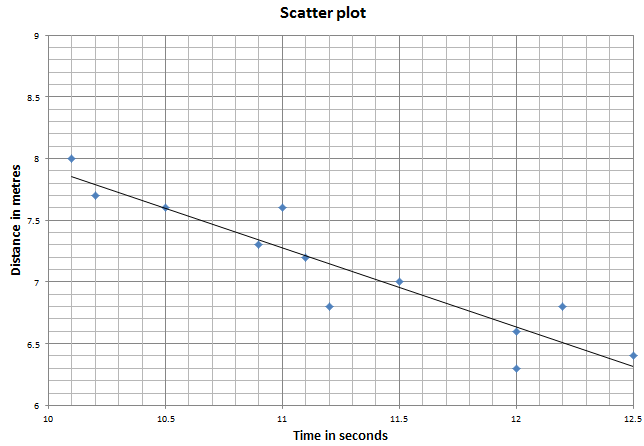
No correlation

**Task 1**

The table below shows the time (in seconds, rounded to one decimal place) taken by 12 athletes to run the 100 metre sprint and the distance (in metres, rounded to one decimal place) of their best long jump.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Time for 100 m sprint (in seconds) | 10,1 | 10,2 | 10,5 | 10,9 | 11,0 | 11,1 | 11,2 | 11,5 | 12,0 | 12,0 | 12,2 | 12,5 |
| Distance of best long jump (in metres) | 8,0 | 7,7 | 7,6 | 7,3 | 7,6 | 7,2 | 6,8 | 7,0 | 6,6 | 6,3 | 6,8 | 6,4 |

The scatter plot representing the above data is shown below:



The equation of the least squares regression line is given by

* 1. Determine the values of *a* and *b*
  2. An athlete runs the 100 m sprint in 11,7 seconds. Use the regression line to predict the best long jump distance of this athlete
  3. Determine the correlation coefficient r of the above data. What can you deduce from your value of r?
  4. Another athlete completes 100 m in 12,3 seconds and has a long jump distance of 7,6 m. If this data is included, will the gradient of the least square regression line increase or decrease? Motivate your answer without any further calculations.

**Answers to Task 1**

1.1 To find the equation of the regression line for the above table of values we go through the following steps:

* Press [MODE] and then select [2: STAT]
* Select [2: A+BX] for linear regression
* Now enter the bivariate data, by entering the [X / Y] value and [=] for all x and y data points
* Press [AC] to clear the screen and store the data values
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (y-intercept), select [1: A] and [=].
  + - Answer:
* Press [AC] to clear the screen.
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (gradient), select [2: B] and [=].

Answer:

1.2 *y* = 14,34 – 0,64(11,7) = 6,85

1.3 To find the **correlation coefficient** we go through the following steps:

* Press [MODE] and then select [2: STAT]
* Select [2: A+BX] for linear regression
* Now enter the bivariate data, by entering the [X / Y] value and [=] for all x and y data points
* Press [AC] to clear the screen and store the data values
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (correlation coefficient), select [3: r] and [=].

Answer:

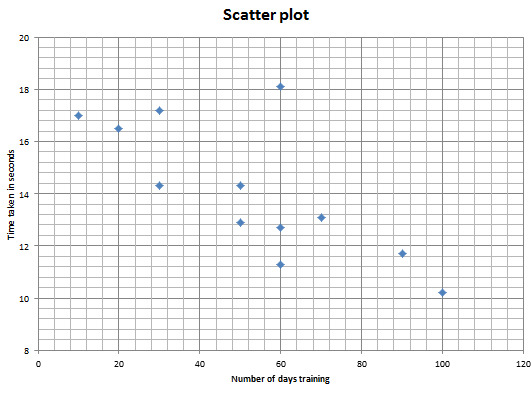
Wemay deduce that there is a strong negative correlation between the time for the sprints and the long jump distance in metres (as the time increases, the distance decreases)

1.4 The point (12,3; 7,6) lies some distance above the current data. This will make the regression line “less steep”; the gradient increases.

**Summary assessment**

1. Twelve athletes are trained to run the 100 m sprint at the local athletic club trials. Some of them took their training more seriously than others. The following able and scatter plot shows the number of days that an athlete trained and the time taken to run the event. The time taken, in seconds, is rounded off to one decimal place:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Number of days of training | 50 | 70 | 10 | 60 | 60 | 20 | 50 | 90 | 100 | 60 | 30 | 30 |
| Time taken in seconds | 12,9 | 13,1 | 17,0 | 11,3 | 18,1 | 16,5 | 14,3 | 11,7 | 10,2 | 12,7 | 17,2 | 14,3 |

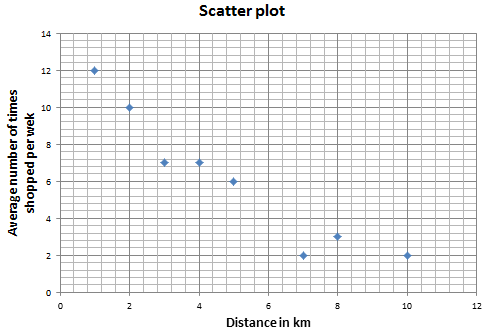


* 1. Discuss the trend of the data in the scatter plot
  2. Identify any outlier(s) in the data
  3. Calculate the equation of the least squares regression line
  4. Predict the time taken to run the 100 m sprint for an athlete who trains for 45 days
  5. Calculate the correlation coefficient
  6. Comment on the strength of the relationship between the variables

1. A survey was conducted at a local supermarket relating the distance that shoppers from the store to the average number of times they shopped at the store in a week. These results are shown in the table below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Distance from the store in km | 1 | 2 | 3 | 4 | 5 | 7 | 8 | 10 |
| Average number of times shopped per week | 12 | 10 | 7 | 7 | 6 | 2 | 3 | 2 |

The scatter plot for the above data is shown below:



* 1. Use the scatter plot to comment on the strength of the relationship between the distance a shopper lived from the store and the average number of times she/he shopped at the store in a week
  2. Calculate the correlation coefficient of the data
  3. Determine the equation of the least squares regression line of the data
  4. Use your answer in 2.3 to estimate the average number of times that a shopper living 6 km from the supermarket will visit the store in a week
  5. Copy the scatter plot and then sketch the least squares regression line on the scatter plot

#### Guided reflection on Summary Assessment

|  |
| --- |
| 1. Use the answers given for the summary assessment to check if your answers were correct. 2. Did you realize why you got some your answers incorrect? 3. Write down the errors you made. 4. Have you been able to correct errors you made? |

**Answers for summary assessment**

* 1. As the number of days that an athlete trained increased, the time taken to run

the 100m event decreased.

**Or**

The fewer number of days an athlete trained, the longer the time he took to complete the 100m sprint.

**Or**

The greater number of days an athlete trained, the shorter the time he ran the 100m sprint.

* 1. (60 ; 18,1) is an outlier as it lies outside the region where the rest of the data is concentrated

To find the equation of the regression line for the above table of values we go through the following steps:

* Press [MODE] and then select [2: STAT]
* Select [2: A+BX] for linear regression
* Now enter the bivariate data, by entering the [X / Y] value and [=] for all x and y data points
* Press [AC] to clear the screen and store the data values
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (y-intercept), select [1: A] and [=].
  + - Answer:
* Press [AC] to clear the screen.
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (gradient), select [2: B] and [=].

Answer:

2.1 As the distance from the supermarket increases, the average number of times shopped per week decreases **or**

As the distance from the supermarket decreases, the average number of times shopped per week increases

2.2 To calculate the **correlation coefficient** we go through the following steps:

* Press [MODE] and then select [2: STAT]
* Select [2: A+BX] for linear regression
* Now enter the bivariate data, by entering the [X / Y] value and [=] for all x and y data points
* Press [AC] to clear the screen and store the data values
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (correlation coefficient), select [3: r] and [=].

Answer:

Wemay deduce that there is a strong negative correlation between the time for the sprints and the long jump distance in metres (as the time increases, the distance decreases)

2.3

To find the equation of the regression line for the above table of values we go through the following steps:

* Press [MODE] and then select [2: STAT]
* Select [2: A+BX] for linear regression
* Now enter the bivariate data, by entering the [X / Y] value and [=] for all x and y data points
* Press [AC] to clear the screen and store the data values
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (y-intercept), select [1: A] and [=].
  + - Answer:
* Press [AC] to clear the screen.
* Press [SHIFT] [1] to get to the stats menu
* Select [5: REG] for linear regression
* To get the value of (gradient), select [2: B] and [=].

Answer:

2.4 We substitute x = 6 into the above equation:

= 4,63

The shopper will visit the store approximately 5 times.

2.5

