

Unit 5: Statistics

Mathematics for Primary School Teachers

Work through this course to learn the mathematics you need to know for primary school teaching!

Unit 5: Statistics	3
Introduction	3
Random numbers, population and sample	5
Graphical forms of data representation	8
Measures of central tendency	16
Misleading statistics	19
Statistics in the classroom	21
Unit summary	22
Assessment	23

© Saide, with the Wits School of Education, University of the Witwatersrand



Permission is granted under a Creative Commons Attribution license to replicate, copy, distribute, transmit, or adapt this work freely provided that attribution is provided as illustrated in the citation below. To view a copy of this licence visit <http://creativecommons.org/licenses/by/3.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California, 94305, USA.

Citation

Sapire, I. (2010). *Mathematics for Primary School Teachers*. **Saide** and the Wits School of Education, University of the Witwatersrand, Johannesburg.
ISBN: 978-0-9869837-5-7

Saide
Fax: +27 11 4032813
E-mail: info@saide.org.za
Website: www.saide.org.za
P O Box 31822
Braamfontein
2017 Johannesburg
South Africa



Unit 5: Statistics

Introduction

Statistics is the name given to that mathematical field which tries to give numeric representations and assists us to interpret situations. If, for example, someone quite simply says to you, 55% of the children at their school are boys, while only 5% of the teaching staff are males, they have given you some statistical information. In statistics, we are usually not so much concerned with the exact original figures as we are with comparisons, which often involve percentages. Statistical results are often shown in a graphical form rather than purely in words and numbers. These graphs are also meant to make interpretation and analysis of the information represented easier for us.



Reflection

Which do you think would be easier to interpret – graphical or written information?

In this very short introductory course on statistics, we will outline the most important terminology which is often used in statistics, we will look at ways of representing information and then we will look very simply at some statistical interpretations of given information.

Upon completion of this unit you will be able to:



Outcomes

- *Define* and cite examples of key statistical concepts to be used in primary schools.
- *Identify* graphical forms of data representation.
- *Differentiate* between different measures of central tendency.
- *Explain* and cite examples of how statistics can be used in misleading ways.

Statistics lends itself to group work and to projects – it is useful to remember this when you plan your work. In the earlier grades, statistics will simply be the collection, organisation and description of data. The interpretation, analysis and use of data will be developed in the later grades.

**Reflection**

Why does statistics lend itself to group work and projects?

First, let us have a look at some important statistical terminology. You need to be sure that you know the meaning of the terms described below. Examples are given for each term, but an example for you to work through (relating to each of the terms) is given after all of the terms have been explained.

Data

This is information collected relating to a given topic. For example, at a certain pet shop there are 205 goldfish, 6 puppies, 15 kittens, 37 budgies, 17 hamsters and 4 cockatiels.

Raw data

Raw data is data which has been collected but not yet sorted out in any way, such as into categories. For example, you might want to find out information about birthdays of the learners in your class. You could use a class list to do so – as you ask each person what day their birthday will fall on this year, you record the day next to their name. All you then have is a list of names with the corresponding days on which their birthdays fall, for example, X. Zulu – Thursday, etc. You cannot easily tell from just looking at the list if more birthdays fall on a Wednesday (or whatever other day) since you have not yet counted up the number of birthdays which fall on each day of the week. Raw data needs to be sorted.

Tally

This is a very common method used to sort through data. You could use tallying to sort the data in the raw data example given above. To do so, you would write a list of the days (Monday to Sunday) on a page, and then go through the class list, making a mark next to the correct day for each name on the list. The tallies could then be counted up.

Frequency

The totals that you get when you add your tallies give you the frequencies for the particular data collected. You could check that the total number of children in the class is the same as the total you get if you add up all of the frequencies, to be sure that your tallying has been correct.

Grouped data

Sometimes it is not practical to itemise each category for which we have collected data, because this will lead to too many categories. If we collect and sort birthdays according to days of the week, we would have seven categories, which is a reasonable number. If we wanted to record actual dates of birth (e.g. 12 September, 27 July, 25 December, etc), there is very little chance that we would have many children in one class with the exact same dates of birth, and so there would be almost as many categories as there are children

in the class. In this instance, we would choose to sort the information according to the **month** of birth. This would **group** the data in a more sensible manner and still lead to meaningful comparisons. We would then have 12 categories, which is still quite a large number of categories, but it is a meaningful and manageable number to deal with and to use.

Range

This is the difference between the highest score and the lowest score in data which has been collected. For example, if we are analysing the test results of a class, an interesting piece of information would be to find out the difference between the highest result and the lowest result obtained by learners in the class. This would be called the range. It gives us some idea of the spread over which the results occurred.

Data collection

The data has to be selected using an appropriate strategy. This will depend on factors such as the data required, the time available for the research, the funding available for the research, etc. We could use questionnaires, observation, experiments, telephonic research, group or individual interviews which we record on tape, etc.

Random numbers, population and sample

Random numbers

Statisticians have lists of random numbers which they use in various ways to ensure that the information they collect is random and not biased in any way. You will not use random numbers, but you need to try to be aware of the necessity for a random selection of data, which is not influenced (and therefore biased) by any personal beliefs, or by simple laziness. If, for example, you want to find out information about who likes what food at the tuck shop at school, be sure not just to ask your friends (who might all have similar tastes) or people from only one class (if you want to speak about the whole school). You need to think of a way of getting the information from as broad a selection of learners as possible. Another way of selecting items randomly would be to draw them out of a hat – good mixing of the items in the hat is then very important, or otherwise the items which are placed last in the hat will have a better chance of selection than those which were placed first into the hat. This will lead to a bias in your data.

Population

The population is the whole group of people that you want to speak about. In the example above about what the favourite foods from the tuck shop are, if you want to speak about favourite foods of the learners in your class then your class would be the population, but if you want to speak about favourite foods of the learners in the whole school then all of the learners in the school would be the population.

Normally the population is quite large, but it depends on your statistical research. The larger your population (for example, you might want give information relating to "male South Africans") the more impossible it becomes to ask each member of that population

the questions you want to ask. You then need to develop some form of random selection of individuals from your population, and you will give a generalised result based on information obtained from your sample.



Reflection

How big is the population of "all male South Africans"?

Sample

We collect information from a certain population, which can sometimes be large. If we want information from a whole school, we will not have time to ask everyone in the school the questions that we want to ask. Therefore, as randomly as possible, we ask a selection of learners whose answers we will use to make conclusions about the whole school. We would hope in our sample to have asked learners from each grade in the school, with a good spread across the school.



Activity

Activity 5.1

Now work through the following example, referring to the information in the previous pages. To complete the example, you need a small packet of Jelly Tots. (If you were to do a similar exercise with a class, it could be costly to buy several bags of Jelly Tots, so for them you could possibly make up little envelopes with a selection of different coloured squares of paper all mixed up together, for example.)

1. What data do you want to collect?
2. If you collected data initially as raw data, how would you record this data?
3. Make a tally of your data – use the table below. Add lines as necessary.

Colour	Tally	Frequency

4. On the table above, as indicated in the third column, record the frequencies of the colours (answer in same table).
5. Is it necessary to group your data? Explain your answer.
6. Can you find the range for this data? Explain your answer.
7. What form of data collection did you use?
8. Did you use random selection to find your data?
9. What is the population of the data that you have collected?
10. Did you need to choose information from a sample when you collected your data?

Once you have collected and recorded data in a table like you have in questions 3 and 4 of the activity above, you have a **frequency table** of your data. If you look at the frequency table, you can start to answer descriptive and interpretive questions about your data such as the questions in the following activity.

**Activity****Activity 5.2**

1. Which colour was the most common in your packet of Jelly Tots?
2. Which colour was the least common in your packet of Jelly Tots?
3. Do you think that other packets of Jelly Tots would have the same numbers of different colours as you have found?
4. How could you find out the general distribution of colours in Jelly Tots?

You can design other such "research" situations for yourself, and work through them in relation to each of the terms discussed.

**Discussion**

Discuss some other statistical research problems that you think of with a group of colleagues. Record your ideas. You will be able to use them for your own studying as well as for exercises for your learners to work through in class or at home, in groups or individually. Try to think of problems which have slightly different natures, such as:

- A problem that requires random selection to obtain the sample.
- A problem for which you are able to calculate the range of the data.
- A problem for which you are required to group the data.

Graphical forms of data representation

A frequency table is not the most elegant way of presenting your data. In this unit, we study various graphical forms of data representation which are often used by statisticians. Some forms of representation are used more commonly than others (as you would have noticed if you read newspapers or magazines). Some are useful very generally while others are better used for particular data in particular situations.

We will study each form of data representation separately.



Reflection

What types of data representation have you noticed in newspapers and magazines?

Are you able to read this graphically represented information easily? If not, why not?

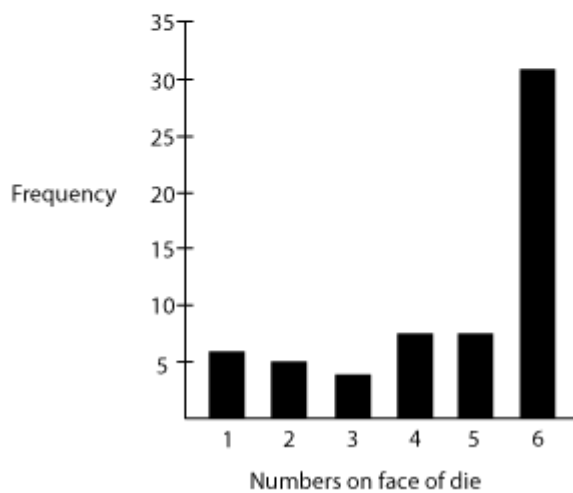
Why do you think that it is generally accepted that information presented in a graph is easier to read and interpret quickly than information that is presented in paragraph style writing?

We will now have a look at five different forms of graphs used by statisticians. These are bar graphs, line graphs, pie charts, pictograms and stem and leaf displays. A worked example will be given for each one followed by an example for you to try on your own.

Bar graphs

A bar graph is a graph made of vertical columns. When the bars are right next to each other, the graph is called a histogram. Here is an example of a bar graph to represent the data from the following table.

Occurance of the numbers thrown with a die	
Number	Frequency
1	6
2	5
3	4
4	7
5	7
6	31



Try the following on your own, using the given frequency table. Remember to label the axes of the graph correctly.

Favourite cakes in GRADE 4B	
Type	Frequency
chocolate	15
vanilla	11
carrot	6
coconut	7
apple pie	6
coffee	3

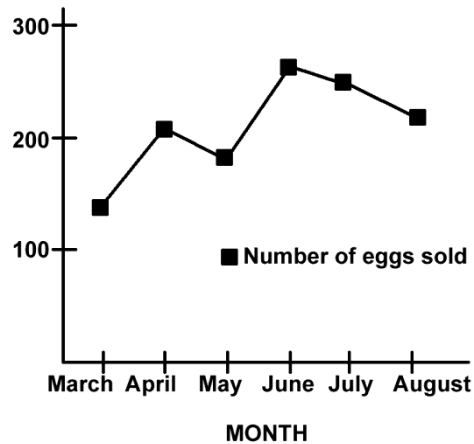
Use graph or grid paper to draw your bar graph on, as this makes it much easier to keep your scale consistent without having to go to any effort measuring.

Line graphs

A line graph is a graph made up of straight line segments joined together between points which are marked according to frequencies for the various categories under consideration. The line starts from the point marking the frequency of the first category and ends at the point marking the frequency of the last category.

MONTH	NUMBER OF EGGS SOLD
March	135
April	203
May	178
June	267
July	250
August	211

EGGS SOLD AT CORNER SPAR MARCH TO AUGUST



Now try the following: using the given frequency table, plot and draw a line graph representing the following information. Once again, use graph or grid paper.

ACTIVITY	AVERAGE RESULTS (%)
Test 1	56
Project	68
Classwork	65
Test 2	61
Oral	63
Practical	72

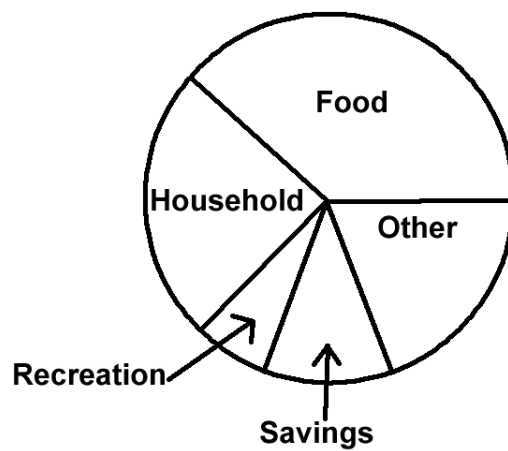
Pie charts

A pie chart is a circular representation, a bit like a pie which has been cut up into various sizes of slices. The different segments (the slices) represent the various relative frequencies of the data. The segments are usually labelled in percentages, and the total of all the percentages should be 100%.

The pie chart represents amounts spent by a family on various items in the month of September in a certain year. Take note of how the angle sizes of the segments are calculated.

ITEM	AMOUNT SPENT (Rands)	AMOUNT SPENT (%)	ANGLE SIZE OF SEGMENT
Food	2000	40	$\frac{40}{100} \times 360 = 144$
Household	1250	25	$\frac{25}{100} \times 360 = 90$
Recreation	250	5	$\frac{5}{100} \times 360 = 18$
Savings	500	10	$\frac{10}{100} \times 360 = 36$
Other	1000	20	$\frac{20}{100} \times 360 = 72$

AMOUNTS SPENT FROM FAMILY BUDGET ON VARIOUS ITEMS




































Pictograms

A pictogram looks and functions a bit like a bar graph does. The difference is that the bars are made up of little icons (pictures) which represent certain numbers of things as is indicated in the key, which must accompany the pictogram. The picture usually relates in some way to the data being represented. Below is a frequency table for the number of cars passing through various tollgates on 24 September in a certain year.

TOLL GATE	NUMER OF CARS PASSED
Moorriver	40 500
Grassmere	52 000
Kranskop	43 000
Middelburg	32 000
Kroonvaal	40 500

Below is a pictogram representing the number of cars passing through the tollgates:

Tollgate							
Moorriver							
Grassmere							
Kranskop							
Middleburg							
Kroonvaal							

KEY	
	Represents 10 000 cars
	Represents 1000 cars
	Represents 500 cars

Now make up your own pictogram to represent the following information relating to numbers of aeroplanes taking off and landing at various airports in South African cities:

AIRPORT CITY	NUMBER OF PLANES PER DAY
JHB International	240
Cape Town	160
East London	45
Durban	130
Port Elizabeth	115

Stem and leaf displays

Stem and leaf displays are not all that common, but they are useful for teachers and so they have been included in this module. They are also very simple to set up, as you will see. They function in a similar way to bar charts, excepting that the bars are made up of the actual data (recorded in a particular way) and no detail is lost since every item of data is used to draw up the bars.

In the table below are Grade 4F marks for Test 3, Term 3: (given as raw data from the class list)

88	67	94	72	77	64	77	83	75	85
87	63	74	80	95	81	84	81	81	70
84	63	68	71	52	56	74	69	65	48
56	41	82	65	70	69	81	53	40	64
63	69	78	96	45	75	58	59	52	54

First we draw the **unsorted** stem and leaf display.

The "stem" consists of the categories into which we group the data. In this case it will be the symbol ranges, such as 30-39, 40-49, 50-59, 60-69, etc. The "leaf" part is taken from the actual data, recording the units within the correct percentage range. Look carefully at the diagram.

Below is the unsorted stem and leaf display of Grade 4F marks for Test 3, Term 3:

4	8 1 0 5 (which came from scores 48, 41, 40, 45)
5	2 6 6 3 8 9 2 4
6	7 4 3 3 8 9 5 5 9 4 3 9
7	2 7 7 5 4 0 1 4 0 8 5
8	8 3 5 7 0 1 4 1 1 4 2 1
9	4 5 6

Now we sort the "leaf" detail, for easier analysis, to obtain a sorted display.

Below is the **sorted** stem and leaf display of marks obtained by Grade 4F, Test 3, Term 3:

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



Activity 5.3

Activity

Provide a sorted stem and leaf display for the following data:

Marks obtained by learners in Grade 7G, Test 6, Term 4.

51 64 56 53 72 45 42 46 57 41

51 63 50 45 64 53 48 47 34 48

39 52 36 55 58 46 50 39 48 44

57 47 54 41 36 54 46 44 57 52

49 77 70 49 41 60 54 69 53 65



Reflection

Earlier in this unit you were required to think of some other statistical research problems. For further exercises, calculate the necessary frequencies (etc.) and select a form of data representation from the above five forms to represent your data. Try to select a different form of data representation for each problem.

Measures of central tendency

In the previous two units, we have looked at basic statistical terminology and graphs. We have asked a few analytical questions, but not many. We need to begin interpreting information and asking questions that will lead to answers which give some insight into the information. One of the most simple yet frequently used ways of describing and analysing data is to use measures of central tendency. You will have heard of an average. This is one of the measures of central tendency. There are also two other ways in which statisticians discuss central tendency. What they are looking for is the score, or a number, which best describes all of the scores or data items. The average (or mean) is not always the ideal measure of central tendency, as you will see.

Mean or average

A single number, obtained through a mathematical calculation which is very often used to describe a set of data. To calculate the mean, we add all of the data items together and then divide that sum by the number of terms which we added.

You should now try to find the mean of 56, 34, 25, 38, 49, 80 and 73.

To do this you should add 56, 34, 25, 38, 49, 80 and 73 and then divide the sum of these numbers by 7.

The mean is not always useful as it does not give information about extreme scores – that is, the highest and lowest scores. Its value is influenced by the extremes and so often does not even reflect one actual recorded central score. We need to find a way of minimising the effect of extreme scores. The next two measures of central tendency minimise these extremes in different ways, but before we look at them, here are some exercises for you to try where you have to calculate the mean (or average). Note that statisticians use the word mean more often than they use the word average.

**Activity****Activity 5.4**

1. Find the mean of the following numbers: 32, 24, 14, 18, 11, 10, 31
2. If the rainfall in the Karoo was 12mm in January, 16mm in February, 80mm in March and in the remaining months of the year, no rainfall was recorded, calculate the mean rainfall for the Karoo for that year.
3. Is this mean useful or misleading? Explain your answer.

Median

The median of a set of data is the value which divides the set into two equal numbered parts when the data has been ranked according to size. To rank data, we put it into ascending numeric order. If the number of scores is odd, then the central most score will be the median. If the number of scores is even, then the median will be the average of the two central most scores.

Study the following two examples.



Example

1. Calculate the median of the following data:

2, 4, 9, 10, 13, 15, 19

The data has been ranked (it is in ascending numeric order), there are seven data items (scores) and so 10, which is the central most score (the one in the middle), is the median.

2. Calculate the median of the following data:

12, 16, 80, 0, 0, 0, 0, 0, 0, 0, 0

Rank the data: 0, 0, 0, 0, 0, 0, 0, 0, 12, 16, 80.

There is an even number of data items. The two central most data items are both zero. The average of these two numbers is zero and so the median is zero. You can see through this example how the median eliminates the value of extreme scores. This is a better way of describing the Karoo rainfall, which was 0mm each month for most of the year.



Activity

Activity 5.5

1. Calculate the median of 32, 24, 14, 18, 11, 10 and 31.
2. Calculate the median of 50, 65, 35, 60, 40, 90, 50, 48, 63, 27, 68 and 53.

Mode

The mode is the score that appears the most often. It is the most common score. The mode is also a measure of central tendency which eliminates the effect of extremes. Not all sets of data have a mode, as there is not always one score which recurs. As educators, it is worth considering whether we should calculate modes more regularly than we do means of our class marks.



Reflection

Why do you think educators might find the mode a useful measure of central tendency?

**Example**

What is the mode of 1, 2, 2, 3, 3, 3, 3, 4 and 15?

The 3 appears more often than any other score and so it is the mode.

What is the mode of 2, 4, 9, 9, 10, 10, 13, 13, 13, 15, 19 and 19?

The 13 appears more often than any other score and so it is the mode.

Data may have two modes, in which case it is called **bi-modal**. We also talk about tri-modal data but we do not consider data with more than three modes worthy of discussion (too many repetitive scores do not have any particular significance).

**Activity****Activity 5.6**

1. What is the mode of 1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4, 5 and 6?
2. What is the mode of 50, 65, 35, 60, 40, 90, 50, 48, 60, 27, 60 and 53?

Misleading statistics

Statistics can often be used to confuse the reader. You should not believe everything that you read, *even* if you are told that what you are reading has statistical backup. There could be faults or built-in flaws in the data. Some of the things that can go wrong or result in misleading statistics are the following:

The sample may not be representative of the population.

**Reflection**

When would the sample not be representative of the population?

The average does not always show extremes, and cannot be used to "summarise" information.



Reflection

Give an example of an average which is not a good descriptor of a set of data.

Graphical representation can be distorted to say what the presenter wants it to say. Scales may be chosen to exaggerate information, to make increases look large, or to make decreases look minimal, depending on the needs of the presenter.



Reflection

Think of an example where scale could be used to mislead an audience. Who would often present data in this correct, but misleading manner?

Generalisations may extend beyond acceptable limits. As you are aware, statistical information is based on a sample – the generalisations made can apply only to the population whom this sample represents. For example, if you interview learners at your school only, you cannot make generalisations about the average South African learner based on these interviews.



Reflection

Think of another instance where a generalisation may have extended beyond acceptable limits.

It is not always clear how averages are calculated. The method should be made clear. Different orders of adding, particularly when many groups of data are involved, can alter the average. We need to be aware of this and ensure that the best manner of calculating the mean is chosen.

When you calculate term marks for a learner, could you find different averages by manipulating the marks differently? How would this happen?

Axes on graphs may not be clearly labelled. We have always said "label your axes clearly and correctly". Labels can be chosen to be purposefully misleading.

**Reflection**

Think of an example where labels on axes could mislead the reader of a graph.

Incorrect information might have been used. People can make mistakes – we need to be sure that no human errors have crept into the information which we are reading.

Statistics in the classroom

It is worthwhile to study and teach statistics for the following reasons:

1. Learners will be able to formulate and solve problems that involve collecting, organising, describing and interpreting data.
2. Learners will be able to produce their own statistics. In this way they will be producers instead of consumers of knowledge.
3. Learners can develop an appreciation for statistical methods as a powerful means for communicating ideas and decision making.
4. Learners will become aware of abuses of statistics and be less readily misled by data which is represented and used by salesmen, politicians, insurance agents, etc.

You need to present the learners with appropriate problems that require data collection. Look again at the problems that you formulated at the end of unit 1. Could you use any of them in your classroom?

It would also be useful to give your learners guidelines for data collection. You might point out to them some of the following things depending on the problem which has been set. Data should be found using the correct sample. If the population is large and a random sample is needed then a means of random selection should be used. One cannot estimate, guess, or make up data – actual data must be found. If special instruments are needed in the recording of the data, ensure that your learners know how to use these instruments and check that the instruments are in good working order.

Your learners might not know all of the graphical forms of representation. You should decide whether to instruct them as to which graph they should use, or you could give them a choice if you know that they already know a few graphs.

**Reflection**

What are the possible types of data representation that you have learned about?

Several mathematical concepts and skills are developed through data collection problems. The basic operations will be used and calculations with percentages will often be needed. Rounding off will have to be done.



Reflection

What other skills and concepts do you think could be developed through data collection problems?



Activity

Activity 5.7

1. Here are questions that you could use to get learners to reflect on and interpret the data. Ask these questions of yourself too!
2. Investigate the extreme data items.
3. What do you find?
4. What was the most or least popular score?
5. What prediction could you make based on your findings?
6. What is the average of your scores?
7. What other questions do you think you could set for your learners?

Unit summary



In this unit you learned how to:

1. *Define and cite examples of* key statistical concepts to be used in primary schools.
2. *Identify* graphical forms of data representation.
3. *Differentiate between* different measures of central tendency.
4. *Explain and cite examples of* how statistics can be used in misleading ways.

Assessment



Statistics

This is an exercise for you to try out. You could also use it in your class, if it seems appropriate.

Find an article from a newspaper or magazine which includes a graphical form of data representation. Read the whole article and study the graph carefully before answering the following questions.

1. What was the investigation about?
2. How is the data represented?
3. Is the presentation used to compare sets of data? If it is, then what is the comparison about?
4. Is the data represented clearly? Give reasons for your answer. Explain any abuses of statistics which you feel are present, if there are any.
5. Discuss the usefulness of the data.
6. What technology (such as computers) do you think was used to collect the data?
7. How will you use the knowledge you have obtained from this discussion to prepare a lesson on statistics for primary school learners?
8. Will it be a useful exercise for the learners? Why?
9. Where and at what level do you think data collection and interpretation should fit into the curriculum? Explain your answer.