Tutorial Letter 501/3/2019

Statistics Education in Intermediate and Senior Mathematics

MAE202N

Semesters 1 and 2

Department of Mathematics Education

This tutorial letter contains important information about your module.

Define tomorrow.

UNISA university of south africa

Open Rubric

CONTENTS

PAGE

STUDY UNIT 1 HOW TO GATHER, ORGANISE AND DISPLAY DATA 4

STUDY UNIT 2 DESCRIPTIVE STATISTICS AS AN APPROPRIATE METHOD FOR ANALYSING DATA.. 60

STUDY UNIT 3 PROBABILITY: THE STUDY OF RANDOMNESS 86

STUDY UNIT 4 FROM SAMPLE TO POPULATION 124

ANNEXURE: CURRICULUM ASSESSMENT POLICY STATEMENTS 133

KEY TO ICONS

The following icons are used throughout the study guide to indicate specific functions:

ACTIVITY

This icon indicates that you are required to complete certain activities which will assist you with your studies.

EXAMPLE

Examples are given for further clarification and are indicated by this icon.

NB/TAKE NOTE

Information of a particular importance is indicated by this icon.

STUDY UNIT 1

HOW TO GATHER, ORGANISE AND DISPLAY DATA

CONTENTS PAGE

INTRODUCTION 5

1.1 HOW TO GATHER DATA 5

1.2 STATISTICAL TERMINOLOGY 14

1.3 HOW TO GATHER DATA TO ANSWER QUESTIONS 17

1.4 HOW TO ORGANISE DATA 26

1.5 HOW TO PRESENT DATA GRAPHICALLY 40

1.5.1 Graphs for qualitative and categorical variables 41

1.5.2 Graphs for quantitative variables 47

INTRODUCTION

Although Sherlock Homes was not a statistician, in his adventure of the Copper Beeches he cried out: "Data! Data! Data! I can't make bricks without clay." We have to answer questions, add information, draw conclusions and make recommendations ("make bricks") – and we need data ("clay") to do this.

What are sources of data and where do we find it? Learners should be given the opportunity to generate their own data by asking questions, deciding on appropriate data to help answer their questions and determining methods of collecting data. The data handling concepts of the Curriculum Assessment Policy Statements states that the learner should "collect, organize and summarize data; represent data; interpret, analyse, and report data" – see the Curriculum Assessment Policy Statements in the annexure.

As a facilitator, it will be your responsibility to make sure that realistic investigations are made and **appropriate** data is gathered. In this study unit we look at gathering data in a statistically correct manner and at organising and displaying data to get the maximum information from it.

1.1 HOW TO GATHER DATA

You can collect data in several ways:

By asking questions – either in person or by means of a written questionnaire. We will take a closer look at the design of a questionnaire later on.

By observing things and recording what happens.

By doing an experiment and counting or measuring the data.

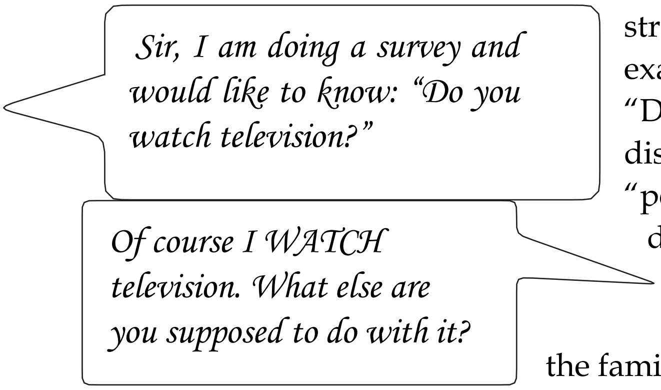
By finding someone else's data in books, newspapers, magazines and (of course the biggest source of data) on the Internet – the website of South Africa's government statistical offices (www.statssa.gov.za) is a rich source of demographic, economic and social data.

Questionnaire design

Begin your questionnaire with a short introduction that tells the respondents the aims of the questionnaire and gives them clear instructions on how to complete it.

Re-read your questionnaire before you start the survey so that you can be sure that the respondent will be able to read and understand all the questions easily and that he/she will be guided throughout. You might want to consider doing a short pilot or trial run amongst your friends to make sure that your questions are clearly understood. A pilot run will often show you the need to alter questions, expand the instructions or introduce questions that you did not think were necessary.

Asking questions is not always as straightforward as it seems. For example: Learners were asked: “Do you have a pet?” A follow-up discussion revealed that the term “pet” means different things to different learners and that some learners answered “yes” if other family members had pets or if the family kept horses and cows.


Sir, I am doing a survey and would like to know: “Do you watch television?”
Of course I WATCH television. What else are you supposed to do with it?

To begin, you have a choice between open-ended questions or closed questions. **Open-ended questions** require respondents to answer the questions in their own words. **Closed questions** provide a variety of possible responses that the respondents can choose from. It is good to pose an open question at the start of your questionnaire to get the different options for your closed question. For example, in the following questionnaire the open question at the bottom revealed that almost everyone also go to church in their free time. We could have included “Go to church” as a closed option and could have left out the open question.

	Daily	Several times a week	Several times a month	Several times a year	Never
Read books					
Get together with friends					
Go to the movies					
Watch TV					
Listen to music					
Spend time on the Internet					
Get together with family					
Do handcrafts such as needle work, wood work					
Play card or board games					
Take part in physical activities such as: sport, gym etc					
Apart from what is mentioned at the top, what do you do in your free time?					

Avoid double-barreled questions: Your questions should be mutually exclusive. In other words, do not include two items in one question. For example:

Do you drive or take the bus to school?
Yes ( ) No ( )

Avoid asking leading questions. Leading questions are actually statements that are disguised as questions. For example:

Would you say that you are not in favour of school on Saturday mornings?

Define terms clearly. For example: the question "How many children are in your family?" can have different answers, depending on what age one is considered a "child".

Your closed questions should provide answers that **cover all the possibilities.** They should be exhaustive. This means that when you come up with a question, you also have to come up with all the possible answers to your question. Most researchers insert the category below to make sure that the answers that are contemplated cover all the possibilities.

*Other*____________________*(Please specify)*

Your questions should be **relevant** – avoid asking questions that deviate from the purpose of your questionnaire.

Avoid using abbreviations or jargon in your questions, for example: How often do you surf the www?

Putting it all together: The layout of your questionnaire

Is this important? Look at the following layouts and decide which one you would prefer to use:

Do you agree, disagree or have no opinion that this company has:

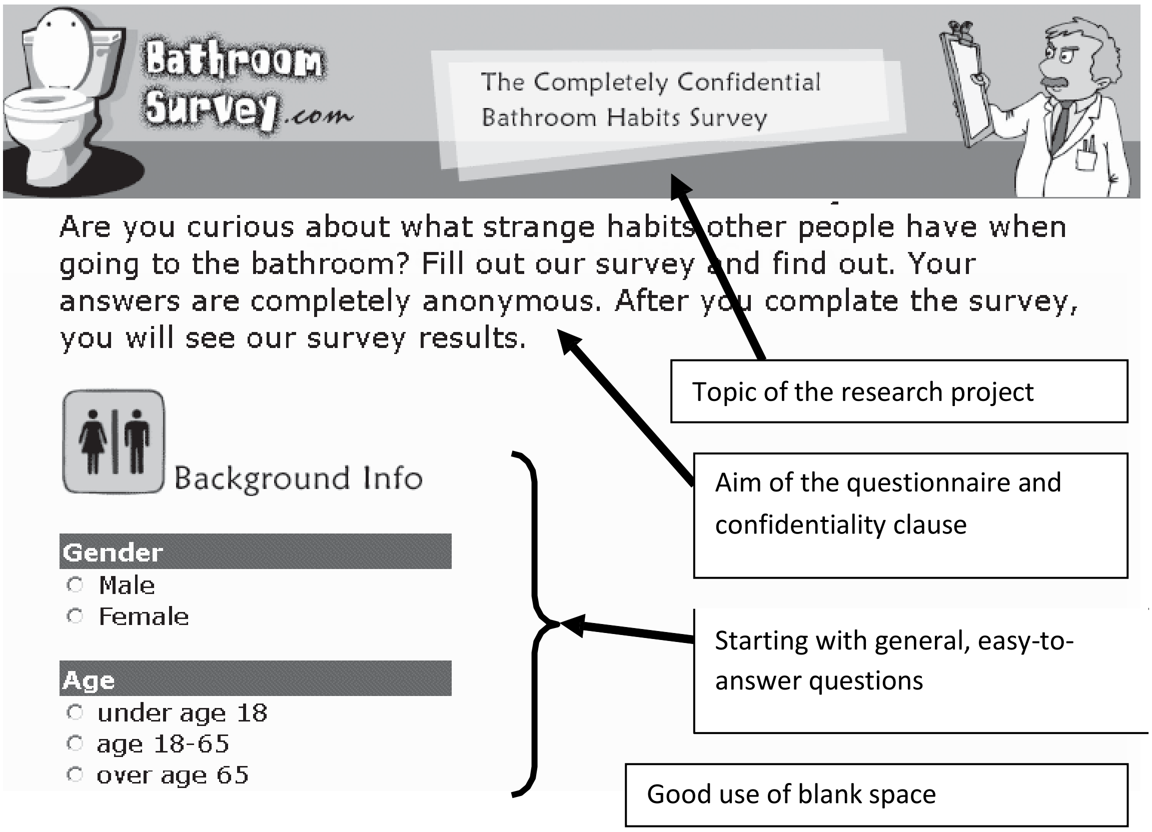
A good vacation policy - agree/not sure/disagree
Good management feedback - agree./not sure/disagree
Good medical insurance - agree./not sure/disagree
High wages - agree./not sure/disagree

An alternative layout is:

Do you agree, disagree or are not sure that this company has:

	Agree	Not sure	Disagree
A good vacation policy	□ 3	□ 2	□ 1
Good management feedback	□ 3	□2	□ 1
Good medical insurance	□ 3	□ 2	□ 1
High wages	□ 3	□ 2	□ 1

Let us take a closer look at the following online questionnaire:



The Completely Confidential Bathroom Habits Survey

Are you curious about what strange habits other people have when going to the bathroom? Fill out our survey and find out. Your answers are completely anonymous. After you complate the survey, you will see our survey results.

Background Info

Gender
○ Male
○ Female

Age
○ under age 18
○ age 18-65
○ over age 65

Topic of the research project

Aim of the questionnaire and confidentiality clause

Starting with general, easy-to-answer questions

Good use of blank space

When you design your questionnaire, keep to the "KISS" principle:

Keep **I**t **S**hort and **S**imple.

Enough has been said about questionnaire design; let us now learn more about the actual survey. Learners should be taught how to approach someone and politely ask him/ her to do the questionnaire. They can practise by surveying one another.

Response bias

In our discussion on sampling techniques we will talk more about bias. We would however like to touch on "response bias" at this stage.

The synonyms for "bias" include "prejudice", "influence", "unfairness", "favouritism" and "partiality". You should try not to introduce bias when you collect data and should eliminate it by wording your questionnaire correctly. However, if you use an interviewer, questionnaire bias can arise from how he/ she asks the questions. The tone of voice or body language of the interviewer can influence the response. A new interviewer should receive basic training.

ACTIVITY 1.1

1.1.1 Criticise the following questions and rephrase/reconstruct them to correct the problems that you have identified:

(1) Don't you think that teenagers who are caught with cigarettes should be fined to prevent them from smoking?

- ♦ Yes
- ♦ No

(2) Do you agree or disagree that teens should not be fined for disobeying the local outdoor smoking ordinance?

(3) Was the presenter knowledgeable about the subject?

- ♦ Yes
- ♦ No

(4) How many times have you been arrested by police?

1.1.2 You want information on:

- ♦ **favourites,** for example TV shows, games, ice-cream and rugby teams
- ♦ **numbers,** for example the number of pets, sisters or brothers
- ♦ **measures,** for example height, hand span or shadow length

Design a questionnaire to get data on the topics that are listed above. You will be evaluated on the

- ♦ layout (20%), and
- ♦ Phrasing of questions (80%)

NOTE: Indicate your target population (for example learners or adults) clearly.

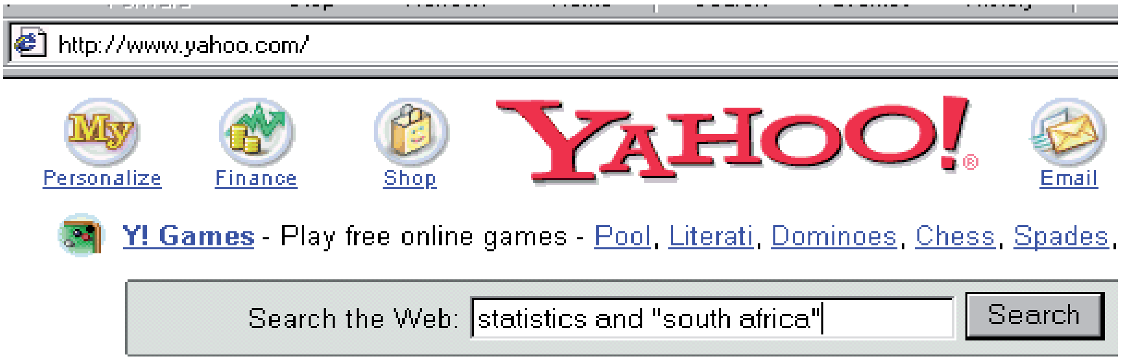
The Internet as a source of real-life data

If you are unfamiliar with doing a search on the Internet, you can follow these easy steps:

Use any search engine, for example

www.yahoo.com, www.ananzi.co.za, www.google.co.za or www.dogpile.com.

Use the key words "statistics" and "South Africa" as illustrated in figure 1.1.


Personalize
Finance
Shop
YAHOO!
Email
Y! Games - Play free online games - Pool, Literati, Dominoes, Chess, Spades,
Search

http://www.yahoo.com/

Search the Web: statistics and "south africa"

Figure 1.1

Visit a few websites and note that the common factor of all the sites is DATA and that in most cases data is presented in tables.

For example, the following table was accessed on the Statistics South Africa website (www.statssa.gov.za):

	Eastern Cape	**Free State**	**Gauteng**	**KwaZulu-Natal**	**Mpuma-langa**	**Northern Cape**	**Northern Province**	**North West**	**Western Cape**	**South Africa**
Energy source for cooking										
Electricity direct from authority	306,964	261,311	1,429,910	756,333	213,890	97,247	188,876	241,967	750,190	4,246,688
Electricity from other source	2,308	1,190	2,795	4,278	1,033	652	2,627	1,509	2,225	18,617
Gas	44,603	24,827	34,285	52,691	14,323	17,753	16,555	33,527	48,093	286,657
Paraffin	390,765	223,265	379,994	296,017	104,321	33,091	120,393	264,253	131,761	1,943,862
Wood	503,438	57,611	18,083	490,122	155,675	34,458	620,960	148,532	44,341	2,073,219
Coal	3,785	43,874	82,696	38,877	106,621	2,573	21,122	20,621	662	320,830
Animal dung	71,371	9,660	255	10,533	2,842	109	5,059	6,206	34	106,068
Unspecified /Other	9,113	3,275	16,150	12,085	5,304	1,100	6,865	4,028	5,709	63,629
Total	1,332,348	625,011	1,964,168	1,660,934	604,010	186,984	982,457	720,643	983,015	9,059,571

Many questions can be compiled on the basis of the information in this table, for example: "In which province or provinces is wood the major energy source for cooking?"

ACTIVITY 1.2

Search the internet for data on crime in South Africa. Copy any table with data from these websites and compile three questions that can be answered by using the data in the table. Remember to reference the website.

OR

Search newspapers, books or magazines for data. Copy the data and compile three questions that can be answered by using the data. Remember to reference the source.

From the examples that we got on the Internet, we can note the following:

- ♦ Data (numbers) is often called "statistics".
- ♦ In most cases, data is presented in tables.
- ♦ Graphical presentations are seldom used.
- ♦ Data has to be processed to reach conclusions. In other words, the collected data has to be analysed in order to
  - ❖ see the data as numbers in context
  - ❖ describe, decide and defend
  - ❖ forecast or predict the future on the grounds of the past pattern in the data
  - ❖ To be able to do all this, we use STATISTICS.

The following figure illustrates what statistics do:

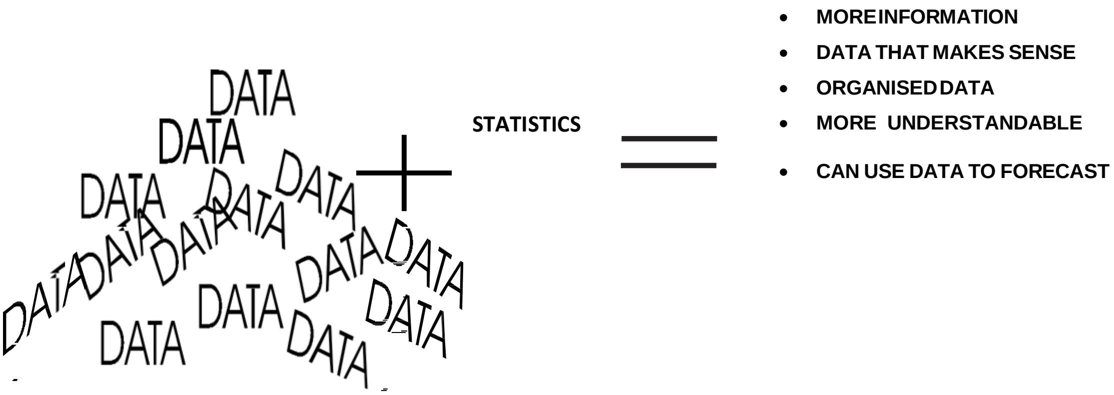


Figure 1.2

Statist~~ics~~

It is interesting to note how the definition of **statistics** has changed over the years. The following are definitions from different authors.

- Statistics can be regarded as the study of: (1) populations, (2) variation, and (3) methods for the reduction of data (Fisher 1925).
- Statistics is a scientific discipline that is concerned with the collection, analysis and interpretation of data which has been obtained from experiment or observation. The subject has coherent structure that is based on the theory of probability and includes many different procedures which contribute to research and development throughout the whole of Science and Technology (Pearson 1936).
- Statistics is the name for that science and art which deals with uncertain inferences and which uses numbers to find out something about nature and experience (Weaver 1952).
- Statistics has become known in the 20th century as the mathematical tool for analysing experimental and observational data (Porter 1986).
- Statistics is concerned with understanding the real world through the information we derive from classification and measurement. Its distinctive characteristic is that it deals with variability and uncertainty which is everywhere. (Bartholomew 1995).
- Statistics is the art of learning from data (Ross 2000).
- Statistics is a way to get information from data (Keller & Warrack 2003).

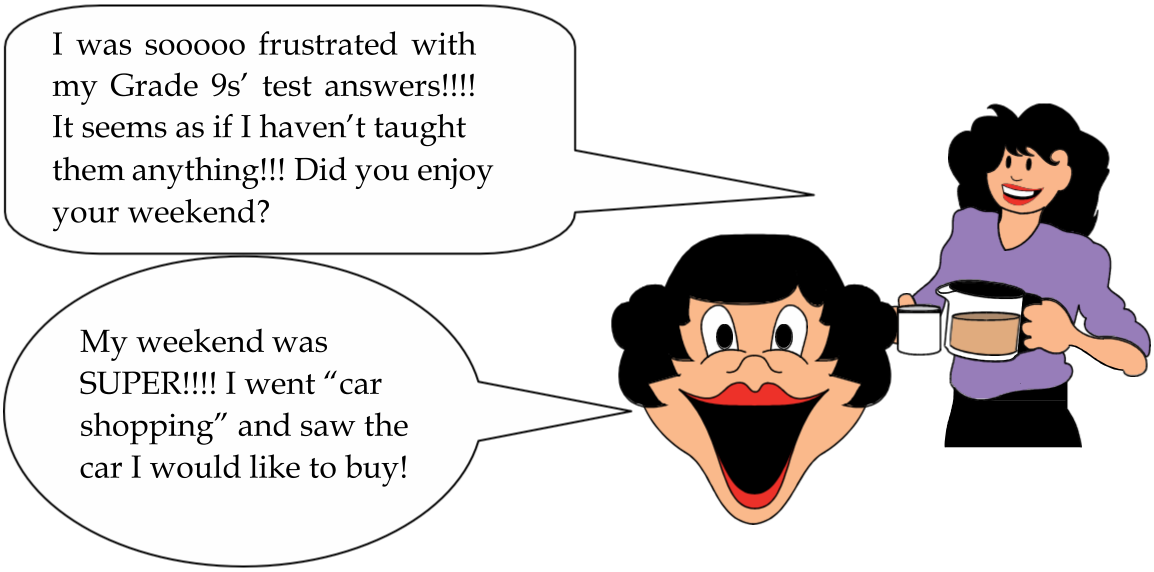
We believe that you are now well aware that data is gathered to answer questions and that statistics are used to get the most reliable information from the data. Before we look at all the wonderful methods for gathering and analysing data, you should be aware that you are entering a world with its own terminology and concepts. The following section is an introduction to some of the concepts and we will add to our vocabulary as we go along.

1.2 STATISTICAL TERMINOLOGY

Have you ever had the "misfortune" of speaking to people who are Internet boffins? When they start to talk about XML, APPLETS, COOKIES, JAVA, SPAM, et cetera, we soon feel as if we are on a strange planet. You will soon use words such as "confidence intervals", "levels of significance", "medians" and "distributions", and you will have to be able to communicate your statistical knowledge to someone who is not from your "planet". Note how simple concepts like "size" and "colour" make communication possible between Ms A and Ms B in case study 1.1

Case study 1.1

It is teatime on a Monday morning and Ms A and Ms B are sharing their weekend experiences.

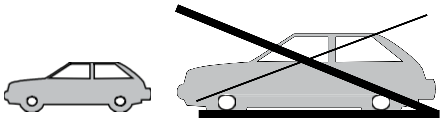

I was sooooo frustrated with my Grade 9s' test answers!!!! It seems as if I haven't taught them anything!!! Did you enjoy your weekend?
My weekend was SUPER!!!! I went "car shopping" and saw the car I would like to buy!

Ms A: Wow! What does it look like?

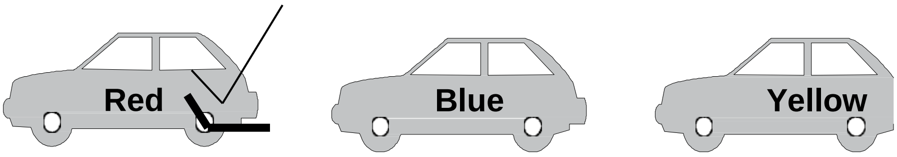
Ms B It's a small red car with the cutest little wings!

In Ms A's mind:

Size:



Colour:



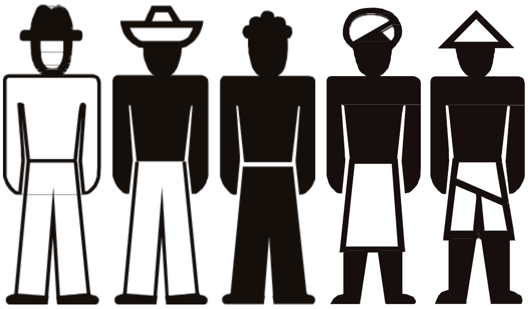
Wings:

Ms A: Wings? I don't understand.

Ms B: Hmmm, it's very difficult to explain. Let me rather draw you a picture.

Ms B roughly draws a picture of what she means by the car having "wings" and Ms A understands immediately.

When you enter the world of statistics, you enter a world with its own terminology and concepts where graphical presentations help to convey information better than mere data. For example, if you use the word "population", the man in the street will immediately see the following in his mind's eye:



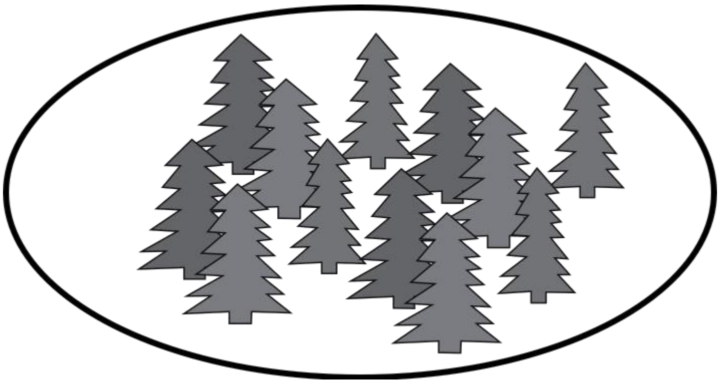
Population

But a statistician knows that "population" refers to an entire set of "things" that we want information on and he/she knows that this can range from people to objects. The logical question that he/she will ask is: "What are we investigating?"

In case study 1.2 the concepts "sample", "random" and "inference" are explained. Your "statistical vocabulary" will expand as you work through this module.

Case study 1.2

A farmer wants to measure the effect of a new fertiliser on tree growth and visits a statistical consultant. The consultant explains that the population refers to ALL the trees on which the fertiliser was used. He roughly draws the following picture to illustrate the population of trees:

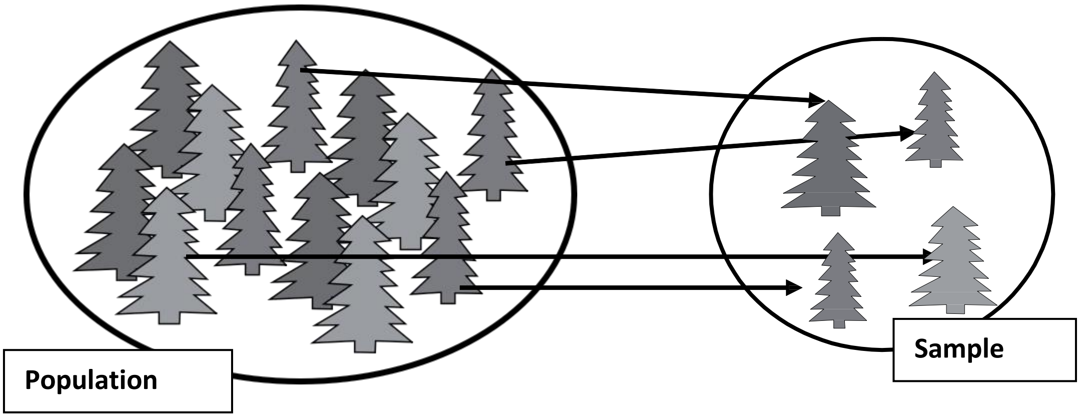


Farmer: Are we going to measure the height of ALL the trees? Wow! Think of the cost it will involve and the time it will take!

Statistician: No, that would be impractical. We will actually examine only a portion or subset of the trees … in statistics this is called a sample.

Sample

Schematically, we can present these concepts as follows:



Farmer (very disappointed): Oh, then we can't say what the effect of the fertiliser was on the whole tree plantation?

Statistician: Yes we can. Based on the sample data, we can draw conclusions or inferences about the whole population on condition that the sample was random.

Farmer: Random?

Random sample

Statistician: Yes, randomness implies that ALL the trees have the same chance to be included in the sample. If this condition is not met, we might in drawing the wrong conclusions from the sample. This is the "beauty" of statistics – statistical inference permits us to draw conclusions about a population on the basis of a sample that is quite small in comparison to the size of the population.

Statistical inference

To recap: **A population** is the set of all the elements of interest in a particular study and a **sample** is a subset of the population. When you collect data on the entire population, you have a **census**.

ACTIVITY 1.3

Write a lesson plan for Grade 7 learners with the following outcome: "At the end of this lesson, you should be able to distinguish between samples and populations."

Guidelines

The task: Describe the task(s) that will enable the learners to discover the differences between samples and populations.

Plan the *before phase*: How will you introduce/present the task?

Plan the *during phase*: List possible hints that you might give to assist the learners.

Plan the *after* phase: How will the learners report their findings? What questions will you ask to assess their understanding?

The lesson plan should be in such a format that we will be able to apply it without any inputs of our own.

We will address many more statistical concepts in this course, including "average", "variance" and "probability".

In the same way that it was easier for Ms B to draw a picture of the car that she planned to buy, we will look at different ways to represent data graphically in order to make more sense of the data. Before we reach that fun part of statistics, we first have to look at the statistically correct way to acquire data. We have emphasised enough that a sample that is taken from a population has to be representative. The question that we now have to answer is: How can we make sure that our sample is representative of the population?

1.3 HOW TO GATHER DATA TO ANSWER QUESTIONS

What makes an investigation process meaningful? Your investigation becomes meaningful when it allows you to answer certain questions that are based on the data. The most important part of any research process or investigation is to clearly formulate the need that arises from a real-life problem. To find the answer to your research problem, the next step is to ask yourself: "What should I measure?" and "How should I measure it?"

Example 1.3

A small school in a rural area caters for learners from the surrounding community and from farms and small towns along the road to the west of the school. It is difficult to teach children to be punctual when they have to walk long distances to school and when they depend on someone else to give them a lift. A bus company has offered to provide the school with cheap transport if enough children will make use of it. As the school principle, how would you approach this problem?

The problem: How many children will use the bus transport to get to school

In trying to solve this problem, the following question arises: What has to be measured? This is easy to answer: How far away do the learners live from the school? Because the school has never collected information on the actual distances which the children have to travel, this leads to the next problem that has to be solved: "How should the distances be measured?" Hmmm, is this a major problem. Should we work with a sample? How are we going to choose a sample?

I think you now realise that planning is of the utmost importance when we do a statistical investigation.

NOTE:

You can ask your learners what they would like to investigate. Bear in mind that this can be the first step of an ongoing project and make sure that the learners have "realistic" research questions, for example do not accept a study of HIV occurrence at the school. Stress the point that they should ask themselves how they will gather the data. From past experience, we know that they can come up with the most creative ideas. To make it more fun, you could allow the learners to work in groups and at the end of the term have a poster session in the school hall. To motivate them even more, let them display their work at the poster session during the conference of the South African Statistical Association that is held every November. For more information and contact numbers, visit the website www.sastat.org.za.

ACTIVITY 1.4

Your last assignment is a project which is an investigation into a "real-life" question that is relevant to your situation. The choice of a "research topic" is the most crucial part of your planning and at this early stage, we want you to:

1. Formulate a research topic. Always remember that you should be able to gather data to "answer your question". Hint: Keep it simple.

2. Answer the following questions: How will you gather the data? From existing sources, for example the Internet? Are you going to use questionnaires?

It is VERY important to remember that in order to do a proper statistical analysis, you need data – data that was collected through sound sampling techniques. Many studies are meaningless because the data was not reliable. Statistical inference permits us to draw conclusions about a population on the basis of studying a sample that is quite small in comparison to the size of the population. It is therefore of the utmost importance that the sample is representative of the whole population.

We cannot stress enough that learners should be taught to be very critical about the technique that they use to collect their data. For example: in a study about the abuse of drugs among Grade 12 learners, the learners might simply claim to have used hard drugs to impress their peers.

The following are some basic points and techniques to consider when collecting data:

♦ When collecting data involves asking sensitive questions, make sure that the learners can complete their questionnaires in private rather than in a classroom where friends are looking over their shoulders.

♦ You can gather data by using data that has been collected by others. The websites of the South African Data Archive (http://www.nrf.ac.za/sada) and Statistics South Africa (http://www.statssa.gov.za) are rich online sources of data.

Bias

♦ Be careful of "convenient sampling" where only elements that are easily collected are included in the sample, for example when one uses the telephone directory to do a survey. Not all people have telephones, some have unlisted phone numbers and some only have cell phone numbers that are not listed. The term "bias" is used if some sections of a population are favoured over other sections in a sampling scheme. Bias refers to the sampling technique and not to the samples that you get from it. A sample technique is biased if it tends to give non-representative samples.

Examples for discussion: Do you think the sampling techniques in the following examples will lead to bias?

♦ In order to answer the question "What percentage of the students in your class plan to go to work immediately after graduation?" you simply ask your friends their opinions over lunch.

♦ You use your statistics class to estimate the percentage of students in your university who study at least two hours a night.

- A student wants to determine the average size of farms in South Africa. He drops some rice randomly on a map of South Africa and uses the farms that are hit by the grains of rice as the sample.

Sampling techniques

I think that you are now convinced of the importance of using proper sampling techniques that will minimise bias. There are four probability sampling techniques for collecting a representative sample from a population. Example 1.4 illustrates three basic probability sampling techniques.

Remember: Statistical theory is based on randomness and the theory can be applied ONLY if the sample is random.

Example 1.4

Moriah, the headquarters of the Zionist Christian Church in the Northern Province, attracts more than a million pilgrims every Easter. Near Moriah is a small village with 50 adults (20 males and 30 females). The 50 adults can only afford to hire a taxi that will transport ten of them to Moriah. They decide that each of them, regardless of gender or status, should have an equal chance of going to Moriah. How would you, as a statistician, choose a representative sample of these adults to attend the church at Moriah?

Technique 1: A simple random sample

Simple random sample

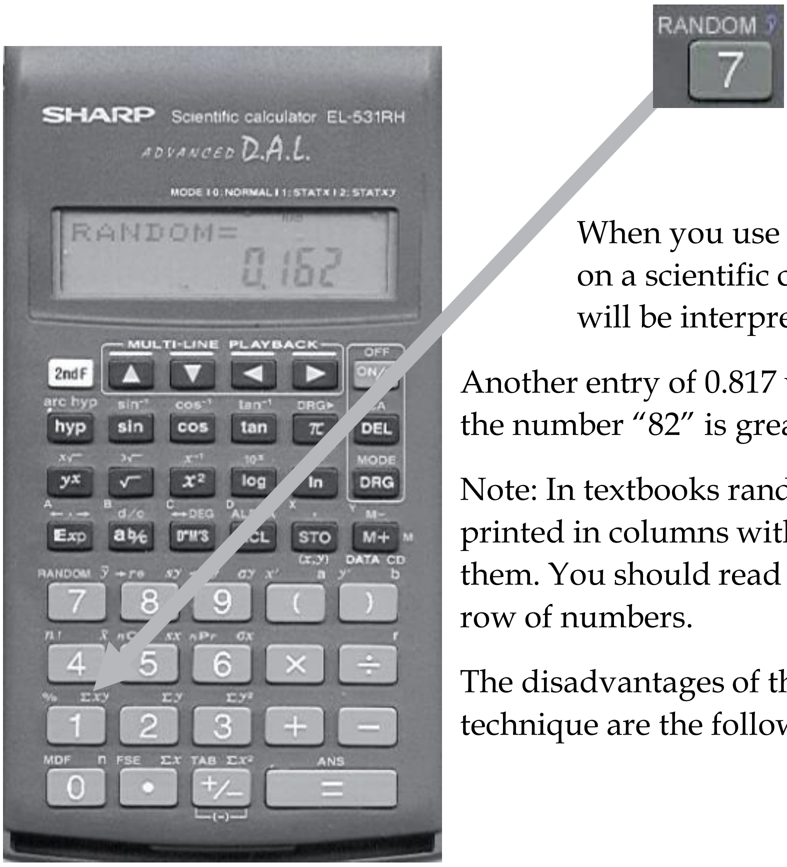
The simple sampling design amounts to placing the names of the inhabitants in a container (the population), shuffling them thoroughly and then selecting ten names (the sample) without peeping. Care should be taken that the names are written on identical pieces of paper. (Have you seen draws done on television where some contestants had sent in their answers in large envelopes and you instinctively felt that it was unfair towards those who had sent in their answers in smaller envelopes?)

A more sophisticated method of doing a random selection is to use random digits that have been generated by a computer in such a way that there is no pattern in the numbers: it is a completely random sequence of the numbers 0, 1, 2, 3, 4, 5, 6, 7, 8 and 9. Most pocket calculators have a “RANDOM” key. Depending on the size of the population, the numbers can be used as single digits or in sets of three, four, et cetera (see the example below).

Let us assume that we will use the following random numbers:

83	20	37	15	72	22	21	78	47	04	14	51
94	38	68	69	47	58	04	45	29	85	48	58

Because our population is 50, we will look at sets of two and ignore all the numbers that are greater than 50. If a number is drawn that has already been selected, it will be discarded. A sample of ten will then consist of the numbers 20, 37, 15, 22, 21, 47, 4, 14, 38 and 45.



When you use the "RANDOM" option on a scientific calculator, the result 0.162 will be interpreted as the number "16".

Another entry of 0.817 will be discarded because the number "82" is greater than 50.

Note: In textbooks random numbers are often printed in columns with spaces in-between them. You should read it as just a continuous row of numbers.

The disadvantages of the simple random technique are the following:

- Numbering all the elements in the population can be time consuming, especially if the population is large.
- A complete list of the population elements that is required to draw the simple random sample is often not available.
- If the population is spread out over a large geographical area, one can include elements from remote areas in the sample and this will raise the costs of the sample taking considerably.
- Since all possible samples from the same population have an equal chance of being selected, sampling from a heterogeneous population can lead to large sampling errors.

Let us take a closer look at the last point because we do have a heterogeneous population (that is, we have 20 males and 30 females). (We would have had a homogeneous population if it consisted of ONLY males or ONLY females.) If the sample contains six males and only four females, the women might feel that the selection was unfair. In this case stratified random sampling would be more appropriate.

Technique 2: A stratified random sample

Stratified random sampling

If the population is heterogeneous (diverse) in terms of the characteristic that is being observed, the population can be divided into segments (called *strata*) where the sampling units in each stratum are relatively homogeneous (of the same kind). Thereafter, random samples are selected separately from each stratum by means of the simple random technique that is described above. For the stratified sample technique, N1 represents the size of the population in the

first stratum and N2 represents the size of the population in the second stratum (and so on), while n_1 represents the size of the sample that is drawn from the first stratum, n_2 represents the size of the sample that is drawn from the second stratum (and so on).

N and n

NOTE: N (uppercase n) is usually used to indicate the population size and n (lowercase n) to indicate the sample size. The subscript 1 and 2 indicate the sample number.

The size of the sample that has to be drawn from each stratum can be calculated by means of the following formula: $n_i = (N_i/N) \times n$, where i represents the i-th stratum.

In our example we can define the two strata as "male" and "female" with $N_{male} = 20$ and $N_{female} = 30$. The size of the samples that will be drawn from each stratum will be:

$n_{male} = (20/50)10 = 4$ and $n_{female} = (30/50)10 = 6$

The four males and six females will then be drawn by using the simple random technique or systematic sampling (described in the following paragraph).

Systematic sampling

Technique 3: Systematic sampling

When a population is homogeneous, a systematic sampling technique might be more convenient. It is easy to administer and saves time and labour. A complete list of the population is not always necessary when this technique is used.

In systematic sampling, sampling begins by randomly selecting the first element and thereafter selecting subsequent observations at a uniform interval. To cover the complete population, we first have to calculate the sampling ratio = N/n.

In our case, the sampling ratio is 50/10 and we will include every fifth value in our sample. Let the inhabitants hold hands and form a big circle. Start randomly anywhere and choose every fifth person.

Example 1.5

The formal problem-solving framework is applied in the following example:

You were notified that any five of your learners can go on a field trip.

Phase 1: *Understand the problem or become aware of the problem*

You can pose the problem to your class and ask them how they would choose the five learners. Make sure that they identify the problem as "how are we going to choose a representative sample of five".

Phase 2: *Design a plan*

- You have to decide on a sampling technique. The choice of a sampling technique is a function of the population (simple and systematic sampling for a homogeneous population and stratified sampling for a heterogeneous population). This leads to the following questions: What is the population? Do we have a homogeneous or heterogeneous population? Does it make a difference whether we are working with a homogeneous or heterogeneous population? Since 25 of the learners in the class of 40 are girls, the learners might feel that more girls than boys should go. This is very important because it will directly influence the choice of the sampling technique. Let us assume that the class has decided that boys and girls should be represented according to the number of boys and girls in the class (in other words, a heterogeneous population).
- Sampling technique: Stratified sampling with two strata (25 girls in the one stratum and 15 boys in the other stratum).
- How will you implement the technique? Let the class write down a strategy. (The discussions can be conducted in groups and the "reporter" from each group can come up with a methodology. From the reports, a strategy can be chosen and written on the board. This is only a suggestion – we know that you can be very creative!)

Phase 3: *Implement the plan*

Let us assume that you have the following class list and that the first five rows (in bold) contain the surnames of the girls in the class:

Baloyi	**Behle**	**Buys**	**Fakude**	**Mabena**
Kgabo	**Lediga**	**Llale**	**Maboane**	**Sepheu**
Mahlangu	**Mahlangu**	**Mahlangu**	**Makubu**	**Mapadimeng**
Masango	**Mmamphoku**	**Mokgata**	**Monyamane**	**Mphahlele**

Mashilo	**Mashimbye**	**Matsho**	**Mchonu**	**Nokeri**
Mohlala	Molondolozi	Moobi	Moraba	Sebelebele
Nkosi	Pitje	Rakgahla	Ramatlo	Talafala
Thaba	Tloubatla	Tsoka	Twala	Zulu

From the girls, you will draw a sample of (25/40) x 5 = 3,125 = 3 and from the boys you will draw a sample of (15/40) x 5 = 1,875 = 2 (or simply 5 – 3, because the total number should be 5).

To draw three girls from the 25 girls and two boys from the 15 boys, you can either use a random sampling technique or a systematic sampling technique (this would have been decided in phase 2). We will assume that you have decided to use a systematic sampling technique. You will then use the "random" mode of your pocket calculator to choose the random starting point (let us say the 10th girl). That will be Sepheu. Depending on the strategy that you have planned, you will choose 25/3=8,3 (say every eighth girl). This will be easy if they are standing in a circle. However, let us say that you are using the class list from left to right. Your second choice will be Mokgata and your third one will be Baloyi.

Baloyi (3)	**Behle**	**Buys**	**Fakude**	**Mabena**
Kgabo	**Lediga**	**Llale**	**Maboane**	**Sepheu (1)**
Mahlangu	**Mahlangu**	**Mahlangu**	**Makubu**	**Mapadimeng**
Masango	**Mmamphoku**	**Mokgata (2)**	**Monyamane**	**Mphahlele**
Mashilo	**Mashimbye**	**Matsho**	**Mchonu**	**Nokeri**
1 Mohlala	2 Molondolozi	3 Moobi	4 Moraba	5 Sebelebele
6 Nkosi	7 Pitje	8 Rakgahla	9 Ramatlo	10 Talafala
11 Thaba	12 Tloubatla	13 Tsoka	14 Twala	15 Zulu

Let us assume further that you have access to random numbers and have decided to use a simple random technique to choose the boys.

- First of all, you will number the boys (small population and not unrealistic). Remember that there is no rule on how they should be numbered. You can start at the bottom of the list or number them as they are seated in the class, in which case you should make sure that every boy notes his number.
- Because the size of the population is two digits, we will look at two digits in the list of random numbers:

  83|20|37|15|72|02|21784752145194386869475804452985485879

You select number 15 and 2: Zulu and Molondolozi.

Phase 4: *Overview*

Are you satisfied that this is indeed a representative sample?

Although there are many more sophisticated sampling techniques, these will suffice for the purpose of introducing you to drawing representative samples.

ACTIVITY 1.5

1.5.1 We identified research questions or topics in activity 1.4. The "what" that had to be measured was identified and now the appropriate sampling technique has to be identified. Now apply the formal problem-solving framework to identify and implement the appropriate sampling technique.

List all the practical problems that you encountered while obtaining the data, for example the girls did not want their hair to be flattened when their height was measured.

1.5.2 Try to find a research article in a newspaper, magazine or on the Internet that states the sampling technique that was used and the sample size. Note the relative frequency of the number of articles that state the sample size or the sampling technique.

We suggest that a discussion session is held before the actual data is gathered. During this session the group leader of each group should present the research question and the sampling technique that his/her group has chosen. This might be time consuming but the class will benefit from the different approaches that their peers used. They should be allowed to criticise the different techniques constructively and a group might even decide to go back to the "drawing board" before implementing their sampling strategy.

We are sure that when you and your learners look at published data, you are already thinking critically of how the data was gathered and asking yourself if the sample was indeed representative of the population. Many research studies in newsletters and magazines do not report the sample size or the sampling technique that was used. Do you agree that this is unacceptable?

From the practical problems that we experience when we measure sampled data, we realise why it is necessary to work with a sample and not a complete population.

Okay, so we have representative data. What now? The next step is to organise and present our data.

1.4 HOW TO ORGANISE DATA

It is important that the learners realise that **data are NOT just numbers,** but numbers with a context. The number 60 is meaningless on its own; however, hearing that a Grade 6 learner weighs 60 kg engages our background knowledge and brings immediate meaning (that he/she is overweight!).

Because context makes numbers meaningful, your examples and exercises should (as far as possible) use real data with real contexts that will encourage the learners to always consider the meaning of their calculations. Encourage them to always state a brief conclusion in the context of the problem. This helps to build their data sense and the communication skills that are valued by employers.

The type of data determines the amount of information that is contained in the data and indicates the data summarisation and statistical analyses that are most appropriate. Before we continue with organising data, we have to expand our "vocabulary" and take a look at the different types of data. This is important because we will refer to a "continuous distribution" later on and then you should have a mental picture of what "continuous data" looks like and what is meant by a "continuous variable".

In real life you sort your dirty clothes before you wash them, putting all the light colours in one pile, the darker colours in another pile, et cetera. So, imagine sorting your data into "piles" as is shown in figure 1.3.

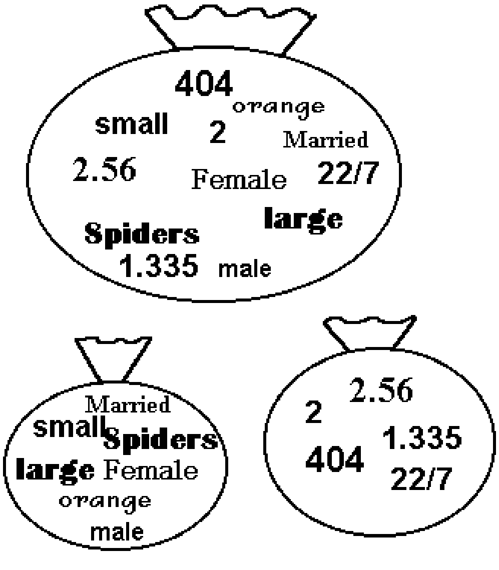


Figure 1.3

This data can be sorted into two broad categories: **quantitative** data (data that is numerical) and **qualitative** data (not numerical but more descriptive data). You should be careful when you sort data because sometimes qualitative data has numerical codes but is still considered qualitative. For example, a "small" jersey might be coded as size 10–12 and the colour of eyes might be coded as 1 = blue and 2 = brown.

Classification of data

The further breakdown of the classification of data is presented in the following diagram:

	Quantitative/numerical/ interval data	**Qualitative data**
	Real numbers/ measurements/ counts	Descriptive – values can be put into categories
Example	Age	Marital status

	Discrete	**Continuous**
Definition	Data that can assume only fixed isolated values in an interval, usually only integer values	No matter how close two values are, it is always theoretically possible to find a value in-between
Example	Number of customers Number of cars at an intersection	Weight/height of people

	Nominal scales	**Ordinal scales**
	Observations are classified into categories	Rank orders are assigned to categories
Example	Gender: Male/Female	Size of T-shirt: Small, medium, large

The following is an example of a lesson with the objective that the learners have to "discover" the concepts "qualitative data" and "quantitative data":

Lesson objective	Students have to "discover" the following concepts: ♦ Qualitative (nominal/ordinal) data/variables ♦ Quantitative data/variables: Continuous (ratio/interval scale) Discrete data	
Examples already sorted into two groups (before stage)	Marital status Job description HIV status Room temperature Size of shoe Cell phone number Opinion on taste of ice-cream (1: good …. 5: bad) Type of residence Smoking status (0:no; 1:yes)	Number of 50c in purse Height of boys in Grade 10 How many students with cell phones? Height of girls in Grade 10 Number of traffic lights in town Distance of Olympic high jumps Weight of newborn babies Number of bedrooms in house How many children in a family?
Reflecting and the explaining (during phase)	♦ How are they different? (distinguish between the properties of variables) ♦ How are the two columns alike? (distinguish between the properties of the variables) ♦ Can you subdivide column A further? (variables with properties they have in common) ♦ Can you subdivide column B further? (variables with properties they have in common) ♦ Can you do calculations (like add and divide) with the properties in Column A? ♦ Can you do calculations (like add and divide) with the properties in Column B?	
Generalising/ Defining (after phase)	**Qualitative data** are labels or names that identify/describe an attribute of each element	

	♦ **Nominal scale:** when the data are labels or names, they are used to identify an attribute of the element ❖ **Non-numeric** labels: GND (=Gender) ❖ **Numeric:** 1 denotes married, 2 denotes divorced, etc. ♦ **Ordinal scale:** if the data has the properties of nominal data and the order or rank of the data is meaningful, eg 1 denotes excellent, 2 denotes good and 3 denotes poor **Quantitative data** are numerical values and are obtained by measuring/counting/calculating something ♦ **Discrete:** something that can be counted ♦ **Continuous** (ratio/interval scale): variables assume values involving fractions or decimals; measured between any two values, another value can be added, eg 2,5 m and 2,6 m, we can add 2,55m
Verifying	Do exercise on p ….

NOTE: Initially, we had one column with all the different variables mixed up and asked the learners to try to differentiate between them according to some common attributes. This took too long and required too much guidance. In another class, we gave the learners a head start with the task by having them already grouped and it worked much better.

Any set of data contains information about some variable. What is a variable? Read through the following definitions and make sure that you understand the term.

Variable **A variable is**

- a category that people construct mentally by creating a class of specifics which has a common set of characteristics
- a set with more than one element or value
- any characteristic of objects that is described by a set of data
- something that can take on different values

If a variable can be described by a set of data, it makes sense that we can refer to a **discrete variable** or a **continuous variable.** For example: Given a random sample of ten houses in a certain area, these are the number of people who are living in each of the houses: 5 3 7 2 5 2 6 8 4

These numbers can be described by the variable "number of inhabitants", which is a discrete variable because the outcomes of the variable are discrete data.

Do activity 1.6 to make sure that you have a thorough understanding of the different types of data.

ACTIVITY 1.6

1.6.1 Sort the data in figure 1.3 into the different data categories.

1.6.2 Here is a small part of the data set that a company keeps to record information about its employees:

Name	*Age*	*Gender*	*Race*	*Salary*	*Job type*
.					
.					
Masango AB	39	Male	Black	52 100	Management
Naidoo CD	27	Female	Asian	20 000	Clerical
.					
.					

List the qualitative and the quantitative variables.

1.6.3 Give two examples of a discrete variable and three examples of continuous variables.

We are now ready to continue with how to organise data.

Tables

Tables are often used to give structured numeric information and to convey information to the reader quickly. It is good practice to abide by the following basic principles (Statistical Services Centre, University of Reading, http://www.rdg.ac.uk/ssc/dfid/booklets/toptgs.html):

- Tables should ideally be self-explanatory. The reader should be able to understand them without having to make detailed reference to the text. In other words, users should be able to identify things in the tables without reading the whole text.
- The title should be informative. It should state what was measured so there is no uncertainty about the definition and the units, where the data was collected (so that the extent of the coverage is clear), when the data was collected (so that the time period that is represented is clear) and whether the data is quoted from another source.

- The rows and columns of tables or axes of graphs should be clearly labelled.
- If relatively large counts in a table with several rows and columns have to be compared, it is often helpful to present them as percentages. The size of the sample on which a percentage table is based should be made explicit. (Can you give a reason for this?)
- It is much easier for a reader to compare numbers that are in a column than numbers that are in a row. Therefore, if the purpose of the table is to demonstrate differences between treatments or groups for a number of variables, the groups should define the rows of the table and the variables should define the columns.
- Units of measurement can be changed to make numbers more manageable. For example, numbers can be multiplied or divided by factors such as 1 000 or 1 000 000 for presentation. Most people find it much quicker and easier to take in a statistic of 72 HIV deaths per 1 000 population per year than a rate of 0,07189.

ACTIVITY 1.7

Study the following table and then answer the questions on the basis of the facts and conclusions that you can draw from the information in it.

Leading causes of death

Cause of death	*Number*	*Percent*
Heart disease	734 090	32
Cancer	536 860	24
Stroke	154 350	7
Chronic obstructive lung disease	101 870	4
Accidents	90 140	4
Pneumonia and influenza	82 090	4
Diabetes mellitus	55 390	2
HIV infection	41 930	2
Suicide	32 410	1
Cirrhosis and chronic liver disease	25 730	1

Cause of death	*Number*	*Percent*
Nephritis	23 630	1
Homicide	23 730	1
All other causes	383 780	17
TOTAL	**2 286 000**	**100**

(1) Criticise the heading of this table.

(2) 56% of the deaths at all ages in 1994 was caused by

___ and____.

(3) Homicide was the cause of____percent of all deaths.

(4) How many people died of HIV infection?

(5) Which causes of death were not directly associated with health problems?

(6) Is there a strong chance that you will die from diabetes?

(7) Which three categories of health-related causes of death should we find easiest to reduce? Why?

(8) Which three categories of health-related causes of death should we find most difficult to reduce? Why?

(9) What percentage of the listed causes are ones other than heart disease and cancer?

One-way table

The table in activity 1.7 is called a **one-way table** because we have one categorical variable, namely “causes of death”.

Quantitative variables such as occupation, size and gender are inherently categorical. Other categorical variables are created by grouping values of quantitative variables (such as the number of children in a family or the height of people in a sample) into classes. To analyse categorical data, we use the counts (also known as the frequency of the observation in that interval) or percentage (also called relative frequencies) of individuals that fall into various categories. **The resulting table is called a frequency table.**

Frequency table

The following data set is an example of a data set that can be presented in a frequency table.

Example 1.6

Lobola (bride wealth) is an age-old African custom that is as alive today as it was 100 years ago. The price that is paid for bride wealth depends on the appearance, education and financial circumstances of the bride. The following data show the amount (in 100s of rands) that was paid for lobola for 44 marriages (this is not real data).

15	25	25	30	33	33	40	40	41	45	46
53	55	55	56	57	58	58	61	62	63	64
65	65	65	66	66	73	74	74	75	76	76
79	81	81	84	85	93	97	97	100	100	110

Compare this data set with the following table and ask yourself which presentation gives the most information.

Frequency table of the price paid for lobola

Amount (in 100s rands)	**Frequency**
15 ≤ x < 30	3
30 ≤ x < 45	6
45 ≤ x < 60	9
60 ≤ x < 75	12
75 ≤ x < 90	8
90 ≤ x < 105	5
105 ≤ x < 120	1
Total	**44**

Constructing a frequency table

Let us take a quick look at how this frequency table was constructed if we bear in mind that it represents all the data.

In order to accommodate ALL the data, we divide the **range** of the data into classes of **equal width.**

The range

♦ The range is simply the largest value minus the smallest one (in our case: 110 – 15 = 95).

- To get classes of "equal width", we have to know how many classes to choose. The following guidelines can be used:
  - For fewer than 50 observations, we can use five to seven classes.
  - 50 to 200 observations = nine to ten classes.
  - The following formula can also be used: the number of classes = $1 + 3.3 \log_{10}n$, rounded to an integer and where n is the number of observations.
- Remember that it is YOUR table and YOU have to decide what is the best way to display your data. If you compile a frequency table of your test marks (as %), you might choose $0 \leq x < 10$, $10 \leq x < 20$ ... $90 \leq x \leq 100$. In this example we chose seven classes and the range divided by 7 is approximately 14. Because multiples of 5 are easier, we simply decided to choose a class length of 15 instead of 14. Do you get it? There is no right or wrong – just guidelines for these choices.
- Why do we have the classes as $15 \leq x < 30$ and not $15 \leq x \leq 30$? Remember that we have to specify the classes so that each observation falls into exactly ONE class. For example, an observation of 29 falls into the first class and an observation of 30 into the second.
- The next step is to count the number of observations in each class. These counts are called frequencies.
- It is very important that your frequency table has an informative title.
- Note that relative frequencies are calculated and presented as percentages to make interpretation of the data easier.

Two-way tables

Study the following table in which data is organised into two categorical variables, namely "Interest in Mathematics" and "Gender". Such two-way tables are often used to summarise large amounts of data by grouping outcomes into categories. It is important for you to be able to grasp the information that is contained in tables with this format. We will say more about these tables when we look at probabilities.

Survey results: Grade 12 learners at Mamelodi High School

		Interest in Mathematics		
		Low	Average	High
Gender	Male	15	50	25
	Female	10	-35	-15

So far, we have looked at tables as the most common way of organising data. The following case study (case study 1.3) is used so that you can revise the construction of a frequency table. It gives a better way of displaying data in an organised fashion.

Case study 1.3

In example 1.4 we looked at the annual Easter pilgrimage to Moriah, the headquarters of the Zionist Christian Church. As part of its community outreach programme, a bus company makes the following offer to the surrounding community: it will provide cheap transport to people who have to walk long distances if enough people need it. A local organising committee measures the distances that 150 families have to travel and presents the bus company manager with the following data set:

6.1	5.7	6.1	6.2	8.3	4.1
0.5	2.3	5.2	1.9	5.4	6.4
5	4	1.5	5.1	0.8	5.1
1.5	5.1	2.3	6.2	6.2	5.7
2.5	1.2	5.5	5.4	5.5	1
0.6	5	0.6	0.7	0.6	2.6
2.4	5.4	3.1	5.3	5.1	1.8
1.7	0.4	5.3	5.6	1.4	2.2
2.6	5.3	0.9	0.8	5.1	5.9
0.4	0.8	5.1	5.2	6.2	2
5.1	0.8	0.5	5.3	1	5.2
6	2.2	2.4	0.7	2.1	1.6
0.7	5	5.2	0.6	5.6	6.1
5.5	0.5	0.7	0.7	0.8	2

2.1	6.2	5.2	6.2	5.3	6.3
4.2	5	0.5	5.2	4.2	6.6
0.4	0.6	2.7	2.3	6.1	2.1
5.7	5.3	5.3	5.8	5.2	6.4
0.5	5.6	0.5	6.1	2.6	6.1
6.3	1.1	1.3	5.2	5.9	3.1
0.9	0.4	6.2	3.1	5.5	6.5
3.1	3.3	0.7	6.3	6.4	0.5
5.4	1.5	5.2	6.3	6.6	4.4
0.7	5	5.2	0.6	5.6	6.1
5.5	0.5	0.7	0.7	0.8	2

A week later, at a meeting between the bus company manager and the local organising committee, the following takes place.

> **Manager:** You don't need my company's help. I added up all the distances that you gave me on the list and divided the total by the 150 families. This tells me that a family will have to walk only an average of 3,6 km to Moriah. That is not so bad. Sorry, we will not be able to help you!
>
> **Chairperson of the organising committee:** Sir, we think that the average hides the long distances that many families will have to travel. We realise that we need to organise the information so that the real story is clear. Will you postpone your final decision until we are able to present our data better?

The manager agrees and the chairperson suggests that the committee ask John, a community member who did a statistics course at university, to help. John suggests that they organise their data in a **stem-and-leaf diagram** (also called a stem plot). In order for them to understand this very effective way of organising data, he summarises the process as follows:

Stem-and-leaf diagram

- Separate each observation into a "stem" that consists of all but the final (rightmost) digit and a "leaf" which is the final digit. Stems may have as many digits as needed, but each leaf can contain only a single digit. For example, the number 223 will have a stem of 22 and a leaf of 3.
- Write the stems in a vertical column with the smallest digit at the top and draw a vertical line to the right of this column.

- Write each leaf in the row to the right of its stem. This will result in an unordered stem plot.
- Arrange the leaves in increasing order, moving out from the stem.

John then helps the committee to organise their data and is surprised at the results that they get in a remarkably short time. Read the data in the columns from top to bottom. For example, 6.1 will have the stem of 6 and a leaf of 1; 0,5 will have a stem of 0 and a leaf of 5; 5 will have a stem of 5 and a leaf of 0; et cetera.

The unordered stem-and-leaf plot (remember that you read the values and add the leaves next to the appropriate stem):

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

The ordered stem-and-leaf plot (you simply sort the values from small to large):

0	4 4 4 4 5 5 5 5 5 5 5 5 6 6 6 6 6 6 77 7 7 7 7 7 7 7 8 8 8 8 8 8 9 9
1	0 0 1 2 3 4 5 5 5 6 78 9
2	0 0 0 1 11 2 2 3 3 3 4 4 5 6 6 6 7
3	1 1 1 1 3
4	0 1 2 2 4
5	0 0 0 0 0 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 44 4 5 5 5 5 5 6 6 6 6 7 77 8 9 9 9
6	0 1 1 1 1 1 1 1 2 22 2 2 2 2 3 3 3 3 4 4 4 5 6 6
7	
8	3

Chairperson of the committee: Oh John, I am so excited! This illustrates exactly what I knew intuitively! Now I can show the bus company that almost half of the families have to travel five or more kilometres to Moriah. Isn't statistics magical!

A few comments on the stem-and-leaf diagram:

- Study the example on page 565 in Van de Walle et al. (2016).
- When you wish to compare two related distributions, a back-to-back stem plot with common stems is useful. The leaves on each side are ordered from the common stem as illustrated in example 1.6.
- Stem plots do not work well for large data sets where each stem has to hold a large number of leaves. Fortunately, there are several modifications to the basic stem plot that are helpful when plotting the distribution of a moderate number of observations. You can increase the number of stems in a plot by splitting each stem into two: one with leaves 0 to 4 and the other with leaves 5 to 9. When the observed values consist of many digits, it is often best to round off the numbers to just a few digits before making a stem plot. You should use your judgment when deciding on whether to split stems or to round off.

Example 1.7

In 16 days a restaurant had the following numbers of orders for chicken and steak, and the manager would like to compare the popularity of the two meat dishes. From the back-to-back stem plot, it is evident that chicken was more popular than steak during the time period.

Chicken	46	55	43	48	54	65	36	40
	51	53	64	32	41	46	53	47

Steak	39	41	25	30	46	36	37	23
	30	33	50	44	41	28	35	37

Back-to-back ordered stem-and-leaf display of orders for 16 days at the restaurant

Leaf: Chicken	Stem	Leaf: Steak
6 2	**2**	3 5 8
8 7 6 6 3 1 0	**3**	0 0 3 5 6 7 7 9
5 4 3 3 1	**4**	1 1 4 6
5 4	**5**	0
	6	

I think that at this stage you cannot wait to use these new methods to display data. After you have done the following activity, sit back and congratulate yourself! You have set yourself apart from the man in the street.

ACTIVITY 1.8

1.8.1 Use the data set in case study 1.3 to construct a frequency table with approximately ten classes.

1.8.2 The following statistics are taken from a newspaper cutting entitled "Enough is enough" that appeared in the Sandton Chronicle. It presents the hijacking statistics in two Johannesburg suburbs, namely Sandton and Bramley.

Hijacking statistics 1996

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sandton	23	19	17	11	18	24	42	23	22	14	19	14
Bramley	70	64	53	42	53	45	41	41	38	37	32	24

Compare the hijacking incidents in the two suburbs by using a back-to-back stem-and-leaf diagram. From this display, what conclusion can you draw?

1.8.3 Use any test results and present them in a stem-and-leaf diagram and in a frequency table. From this display, what conclusion can you draw?

For task 1.8.1, you probably came up with a class length of approximately 0,8 and the first class would then be $0 \leq x < 0{,}8$. It would have been more convenient to round 0,8 off to 1. What we are trying to say is that there is no "right" or "wrong" class length. Let us assume that we use a class width of one. After the classes are constructed, you should go through your data set **once,** indicating in your table in which class the value lies. This is called a “tally” and the way that you do this is to make four marks and a fifth one that is a horizontal line through the first four. For example: 𝍸 (This technique makes it more convenient to count.)

A frequency table of distance travelled to Moriah

Distance (in km)	Tally	Frequency (number of families)
0 ≤ x < 1	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ||||	35
1 ≤ x < 2	𝍸 𝍸 |||	13
2 ≤ x < 3	𝍸 𝍸 𝍸 |||	18
3 ≤ x < 4	𝍸	5
4 ≤ x < 5	𝍸	5
5 ≤ x < 6	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 |||	48
6 ≤ x < 7	𝍸 𝍸 𝍸 𝍸 𝍸	25
7 ≤ x < 8		0
8 ≤ x < 9	|	1
TOTAL		**150**

ACTIVITY 1.9

Present the results of one of your class tests in a frequency table.

When you organised the data in tables and stem-and-leaf diagrams, you already presented it in a better, more understandable format. In the following section, we introduce several graphical methods to display data.

1.5 HOW TO PRESENT DATA GRAPHICALLY

The purpose of this section is primarily to understand data better. In other words, presenting data graphically is not an end in itself. After you make a graph, always ask: “What do I see?” Try and identify important features. Many typing errors can, for example, be identified as an individual value that falls outside the overall pattern.

1.5.1 Graphs for qualitative and categorical variables

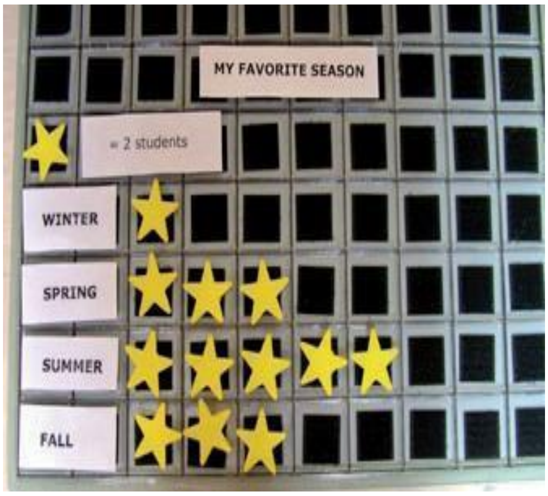
In your textbook, Van de Walle et al. (2016) gives some ideas to present categorical variables graphically that can be used in the classroom.

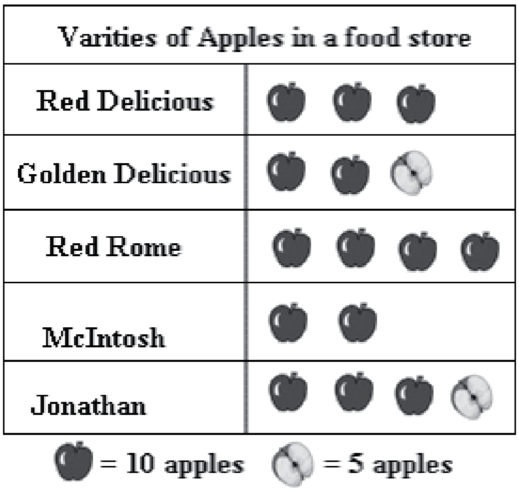
The values of a categorical variable are labels for the categories, such as "male" and "female". The distribution of a categorical variable lists the categories and gives the counts of the individuals who fall into each category. It is always easier to interpret the data if it is presented as percentages and we recommend that you convert your frequencies to percentages.

Picture graphs

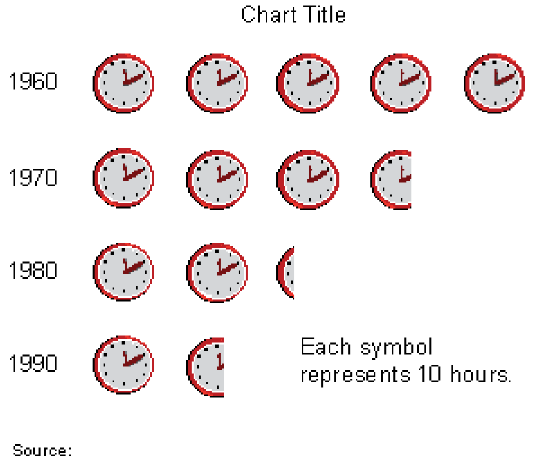
Picture graphs use a drawing of some sort that represents what is being graphed. Learners can make their own drawings and it can be quite a lot of fun, for example the picture graphs of "Clip paper picture" in Van de Walle et al. (2016).

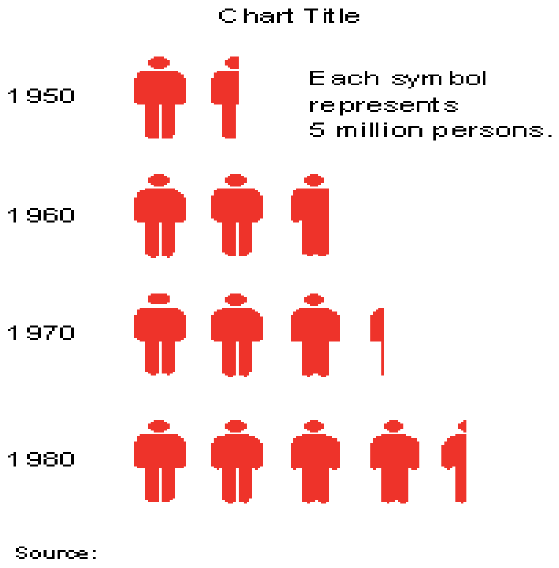
More examples:





The following examples illustrate the shortcomings of picture graphs:





Pie charts

See Van de Walle et al. (2016)'s description of circle graphs.

Bar charts

Bar charts are more flexible than pie charts because pie charts require that you include all the categories that make up a whole. For example, you can use a bar graph to compare the numbers of educators who are married or divorced. You cannot make this comparison with a pie chart because not all educators (for example those who have never been married) fall into one of these categories.

The bar chart

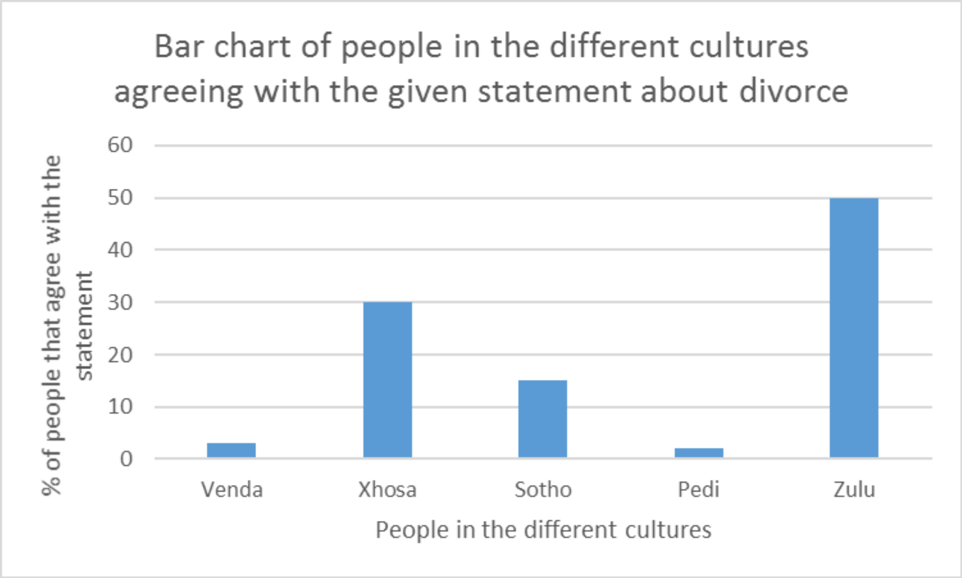
A bar chart is constructed as follows:

- Draw a rectangle to represent each category.
- The height of the rectangle represents the frequency.
- The base is arbitrary, but it is important to use the same base for all the rectangles because your eyes respond to the area of the bars as representing the frequency.
- Because a bar chart compares the size of different items, the bars are usually drawn with a blank space between them.
- Make sure that you give your bar chart or pie chart a descriptive heading.

Example 1.8 illustrates the use of a bar chart to convey information quickly.

Example 1.8

The following bar chart represents the percentage of people in the different cultures who agree with the following statement: *Black couples do not divorce easily, mainly because of one factor – lobola.*



- Which culture agrees the least with this statement? Pedi.
- Is it true that more Venda people than Sotho people agree with the statement? No.

Because the use of bar charts and pie charts is so prevalent, we would like you to do the following activity.

ACTIVITY 1.10

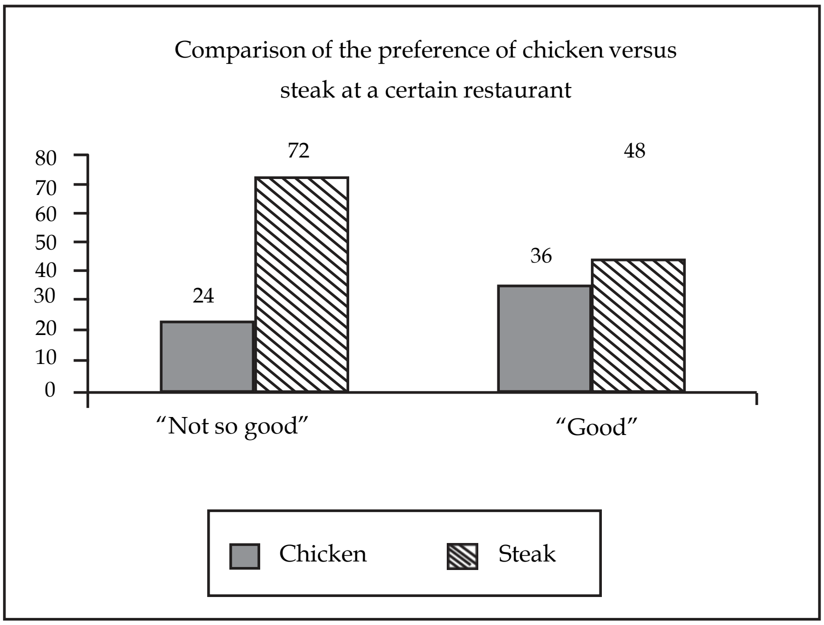
1.10.1 Collect at least five articles from any newspaper or magazine where a bar chart and/or pie chart is used to present data. Critically look at the presentation and discuss its effectiveness with a fellow student.

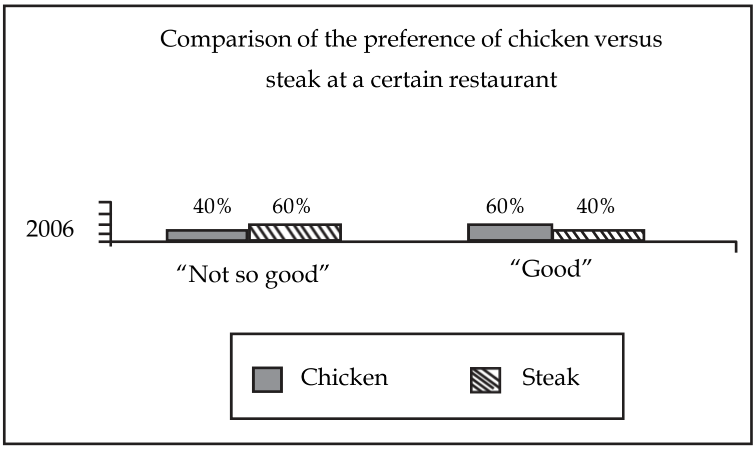
1.10.2 The manager of a restaurant conducts a survey of the clientele’s preference for the chicken or steak dishes (he classified the opinions on the dishes as “not so good” and “good”) and obtains the following data:

	Chicken	**Steak**
Not so good	24	72
Good	36	48
Total	**60**	**120**

Note that the survey included 60 clients who had tasted the chicken dishes and 120 who had tasted the steak dishes.

The following presentations are bar charts that were drawn with the same data. In your opinion, which one is a true picture of the opinions of the clients? Give reasons for your answer and say what important lesson you can learn from this.





1.10.3 Take any test results of your learners and create three classes (or more, if you wish): 1–40: Bad; 41–70: Good; 71–100: Very good. Then construct a bar chart to compare the performance of boys with the performance of girls. Remember to use relative frequencies if the number of boys and girls are not the same. What conclusions can you draw from this graphical display? Would you have reached the same conclusions by just looking at the raw data?

Task 1.10.2 is very important. It is so easy to convey the wrong message when sample sizes are not taken into account, especially when comparing different results. If, for example, we say that 20 learners in a class passed their final examination, what does this mean? Nothing! But if we say that 50% (the relative frequency, that is 20/(total) x 100) of the learners in the class passed, we are giving useful information.

There is no blueprint or "right and wrong" for choosing classes. It is entirely up to you to decide what is best to satisfy your need. You could, for example, have created four classes for activity 1.10 (3): say, "very bad", "bad", "good" and "very good". We would like you to stress this fact to your learners. They should realise that they have to think about and decide what will be the best way to get the most information from their data. Challenging, isn't it?

A word of warning: When you design classes for a questionnaire, try not to have a "middle" category. You know, that category that is neither bad nor good. We think that you can all recall filling in a questionnaire and (perhaps because of time pressure) simply choosing the "safe" middle option without thinking carefully what it was all about.

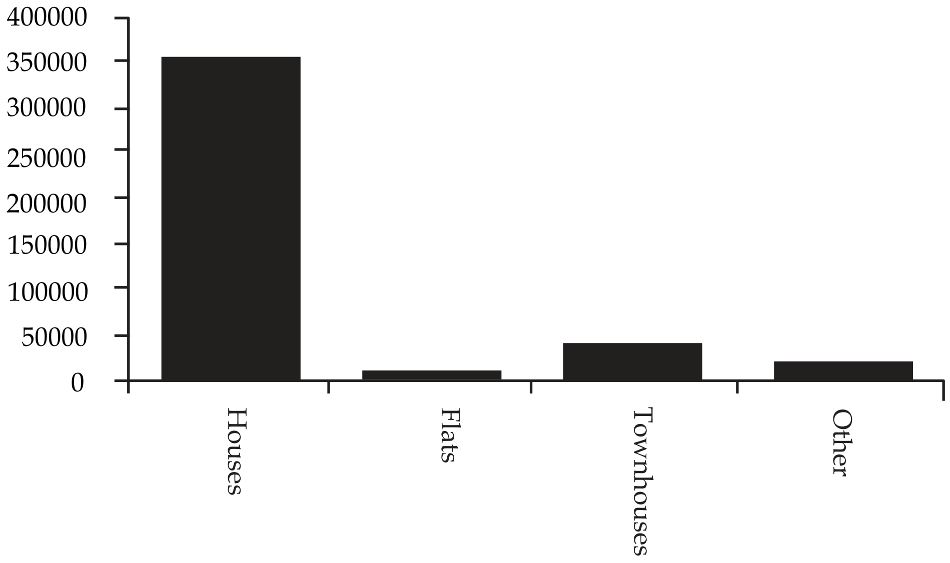
Topic for discussion

The following example is aimed at helping you to improve your learners' ability to interpret, analyse, and extrapolate from graphs. Educators should give tasks that require learners to notice trends in data and to make generalisations or predictions. The following are multiple visual displays that depict "The dwellings we live in" and were made with data that was obtained from a survey which was done in 1998 in the Eastern Cape (Source: StatsSA).

Display 1

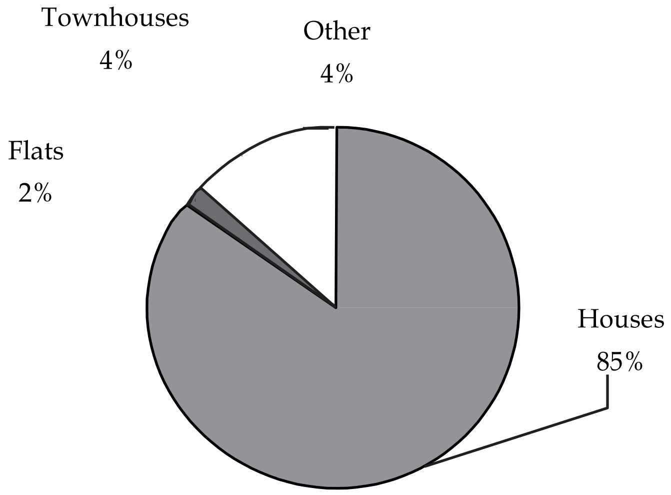
Types of homes	Number of people
Houses	57 792
Flats	9 394
Townhouses	38 153
Other	18 902

Bar chart depicting "The dwellings we live in", made with data obtained from a survey done in 1998 in the Eastern Cape (Source: StatsSA)



Pie chart depicting "The dwellings we live in", made with data obtained from a survey done in 1998 in the Eastern Cape (Source: StatsSA)

<u>Display 3</u>



Possible questions:

1. Describe the advantages and disadvantages of each display. (This question might lead the learners to focus on the data in numeric form instead of attending to the value of the visual displays for indicating relationships between the data.)

2. If you knew that 200 people were planning to move to the Eastern Cape, how many of the 200 people would be expected to move into a townhouse? Which display was the most helpful to answer this question? (This question requires the learners to read "beyond" the data; to see the visual representations not as mere displays of actual data but as instruments that can be used and extrapolated to other problems.)

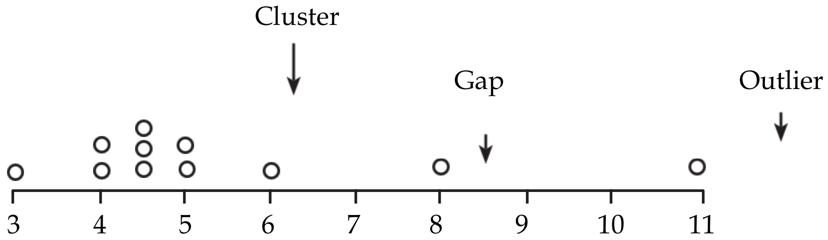
1.5.2 Graphs for quantitative variables

The dot plot

On the dot plot, numerical data are plotted as dots above a number line that reflects the span of the data. Dot plots are good for showing features such as clusters (data points that are grouped close together), outliers (data points that fall far outside the range of the rest of the data) and gaps (an area in the range of the data with no data points). Study figure 21.22 on page 578 in Van de Walle et al. (2016). For example, the dot plot of the data set

3 4 4 4,5 4,5 4,5 5 5 6 8 11 will be:

Dot plot showing special features



The histogram

Drawing a histogram

It is very easy to draw a histogram after you grouped your data in a frequency table. Although the frequency distribution provides information about how the numbers are distributed, the information is more easily understood and imparted in a histogram. **A histogram is created by drawing rectangles whose bases on the horizontal line (X-axis) are the classes and whose heights are the frequencies (or relative frequencies, whichever you prefer)**. Because a histogram shows the distribution of frequencies (or relative frequencies) among the values of a single variable, there is no space between the bars (unless a class is empty, that is the bar has 0 height).

Examine the histogram in example 1.9 and ask yourself whether it adds value to the presentation of the data.

Example 1.9

Using example 1.6, we compiled the following frequency table:

Frequency table of the price paid for lobola

Amount (in 100s rands)	Frequency	Relative frequency	Relative %
15 ≤ x < 30	3	3/44 = 0.0682	7
30 ≤ x < 45	6	6/44 = 0.1364	14
45 ≤ x < 60	9	9/44 = 0.2046	21
60 ≤ x < 75	12	12/44 = 0.2727	27
75 ≤ x < 90	8	8/44 = 0.1818	18
90 ≤ x < 105	5	4/44 = 0.1136	11
105 ≤ x < 120	1	1/44 = 0.02272	2
Total	**44**	**A**	**B**

When we ask you for a frequency table, it is not necessary to calculate columns A and B. You will calculate them when we ask you to draw the ogive.

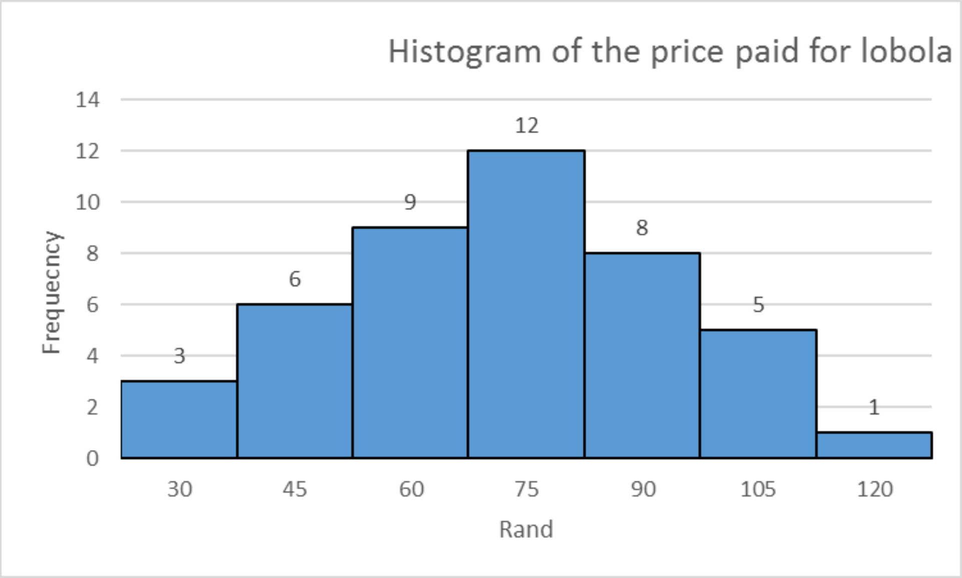
NOTE:

15 ≤ x < 30
30 ≤ x < 45
Etc
105 ≤ x < 120

OR in the following table the ")" means "excluded" and "]" means "included"

[15;30)
[30;45)
Etc
[105;120)

This frequency distribution can be represented by the following histogram. (Note the value that this graphical presentation added to the information.)

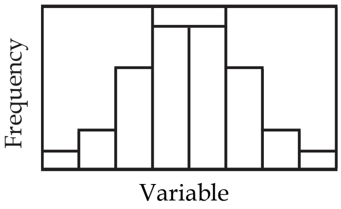


ACTIVITY 1.11

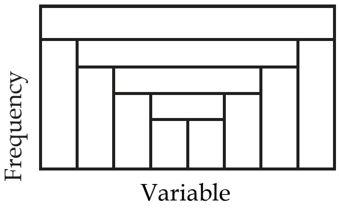
Draw the histogram of the frequency table that you made in activity 1.8

Stress the fact to your learners that we study a histogram to make conclusions about the behaviour of our data. We are interested on whether our data is, for example, approximately symmetrically distributed. In the figure below histograms 1, 2 and 3 are symmetric. A histogram is said to be symmetric when the graph can be divided in half and the left-hand side of the graph is a mirror image of the right-hand side of the graph (which means that the two sides are identical in size and shape). In real life a histogram is seldom exactly symmetrical, but only approximately so. Another important feature of histogram 1 is that it is bell shaped. We will discuss this feature when we discuss the characteristics of the normal distribution. A skew histogram has a tail that extends either to the right or to the left. Histogram 4 is negatively skewed (or skewed to the left) because the "tail" extends to the negative side and histogram 5 is positively skewed (or skewed "to the right") because the "tail" extends to the positive side. Histograms 1, 4 and 5 are all unimodal histograms because they have a single peak. Histograms 2 and 6 are bimodal histograms because they have two peaks. Note that the two peaks do not necessarily have to be the same height.

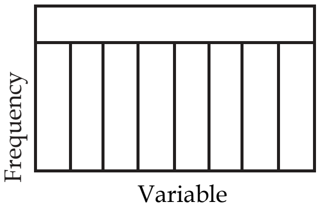
Histogram 1



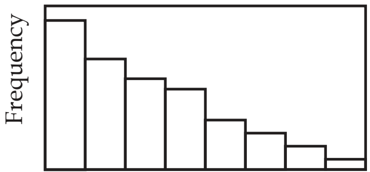
Histogram 2



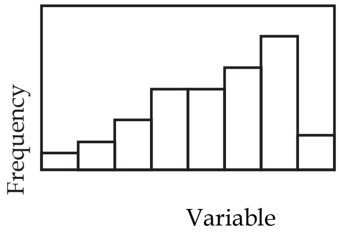
Histogram 3

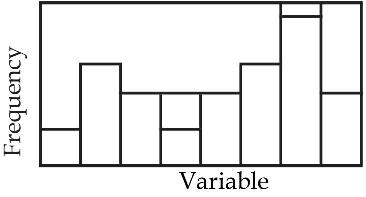


Histogram 5



Histogram 4





Histogram 6

The ogive

Another question that we would, for example, like to have answered is what percentage of people paid less than R7500? To answer this question, let us add another column to our table in which we add (or cumulate) relative frequencies (%). This column will be called the cumulative relative frequencies.

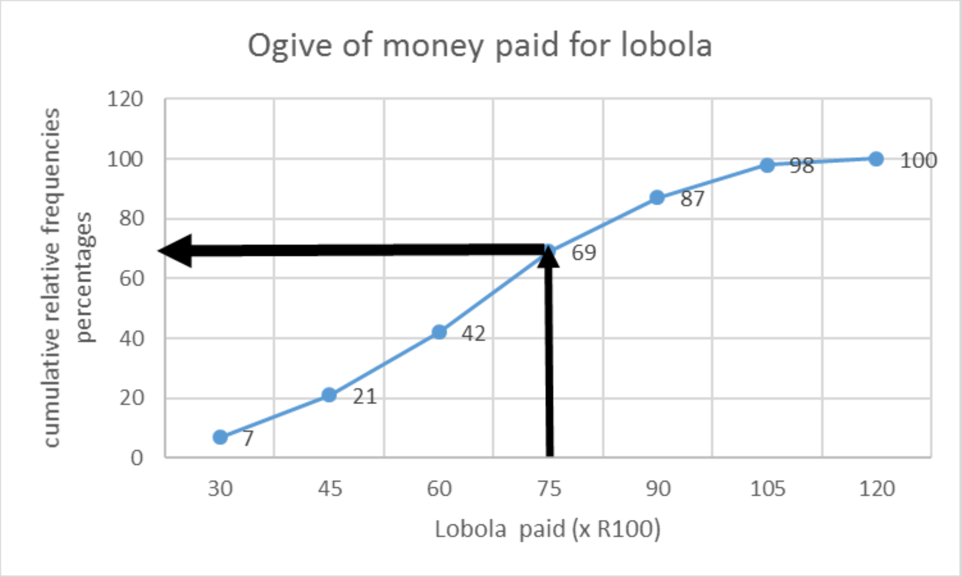
Frequency table of price paid for lobola

Amount (in 100s rands)	**Frequency**	**Relative frequency**	**Relative frequency %**	**Cumulative Relative frequency %**
15 ≤ x < 30	3	0.0682	7	7
30 ≤ x < 45	6	0.1364	14	21
45 ≤ x < 60	9	0.2046	21	42
60 ≤ x < 75	12	0.2727	27	69
75 ≤ x < 90	8	0.1818	18	87
90 ≤ x < 105	5	0.1136	11	98
105 ≤ x < 120	1	0.02272	2	100
Total	**44**	**1**	**100**	

Now it is easy to answer the question, namely less than R7 500 for lobola was paid for 69% of the marriages.

A graph of the cumulative distribution is called an ogive. The data values are shown on the horizontal axis and **the cumulative frequencies, the cumulative relative frequencies or the cumulative percent frequencies** are shown on the vertical axis. We prefer to use the **cumulative relative frequencies** as a **percentage**. The reason for this will become clear in the following study unit.

The ogive (see the following figure) is constructed by plotting a point that corresponds to, for example, the cumulative relative frequencies as a percentage against the upper class limit. Note that at less than 15 (R1 500) the frequency was "0". Note how the above question can also be answered from the graph.



In the back-to-back stem-and-leaf plot that we created in example 1.7 and also in the bar chart in activity 1.10 (2), we compared two variables graphically. There are many situations in which we can depict the relationship between variables graphically, for example marketing managers have to understand the relationship between sales and advertising. Your learners might like to see the relationship between the length of the shadow of a flagpole and the time of the day (summer days can be compared with winter days). The technique that is used to describe the relationship between two variables graphically is the scatter diagram or the line graph.

The scatter plot

We wish that we could look you in the eyes and ask you whether you are starting to experience the magic of statistics! When they practise using statistics, your learners will need a sound mathematical background. When they are creating scatter plots, they are using their knowledge of a Cartesian plane to look at relationships – magic!

Let us recap some mathematical knowledge and learn some new statistical terminology by looking at the following example.

Example 1.10

Are our athletes jumping higher? The gold medal performance of athletes in the men's high jump (measured in cm) for the modern Olympic series was noted (starting in the year 1900) and summarised in the following table:

Year	0	4	8	12	16	20	24	28	32	36
Height	175	183	181	184	186	187	191	187	190	196
Year	44	48	52	56	60	64	68	72	76	80
Height	197	204	208	210	216	215	217	227	227	229

Scale:0:1900

NOTE: There is a missing value in the year 1940 – no Olympic Games were held because of the Second World War.

Show your learners this data and ask them if they can answer the above question by looking at the data for just one second (if you have an overhead projector, "flash" the data by showing it and then quickly removing the transparency).
Remember that we are still looking at means to present data to get the most information in the easiest way. In the follow-up discussions, lead your learners to remember that when they graphed functions, they looked at relationships between x and y. The first important question is: What am I going to choose as my x-variable and what am I going to choose as my y-variable?

Dependent and independent variables

When plotting variables, it is conventional to use the horizontal axis to represent the independent X-variable (also called the explanatory variable) and the vertical axis to represent the dependent Y-variable (also called the response variable). When we look at relationships, one variable depends to some degree on the other variable. For example, learners' test results will depend on the number of hours that they studied. Accordingly, we identify test results as the dependent (Y) variable and the number of hours studied as the independent (X) variable. In an experimental set-up, the independent variable will be the one that you have control over and the dependent variable, the one whose behaviour you are investigating. So, in our example, how are we going to allocate our variables?

This can result in another lively discussion. Usually “time” is presented on the x-axis.

After a suitable scale is chosen that will cover the range of the dependent and independent variables, the co-ordinates are plotted (figure 1.4).

Your learners might be used to the fact that the scales on the x-axis and the y-axis always start at zero. Note how, in this case, it would have resulted in a scatter plot with unnecessary white space (figure 1.5) that would have made it difficult to interpret the graph.

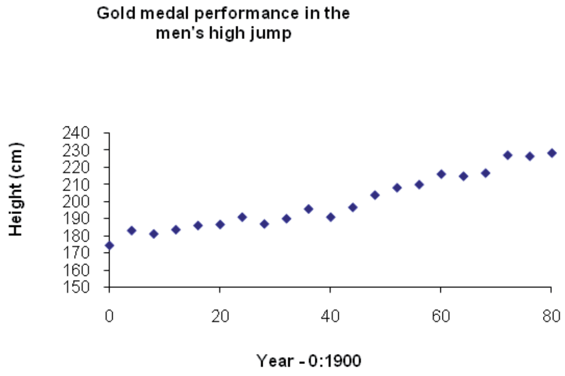


Figure 1.4

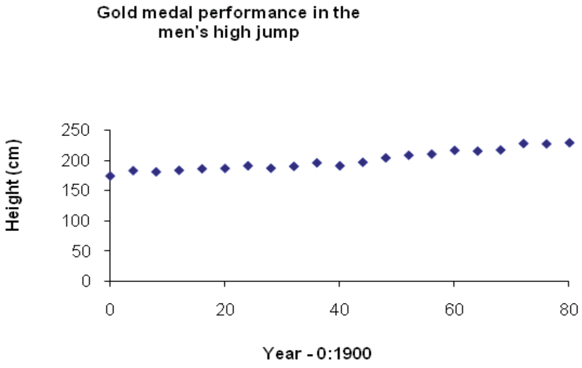
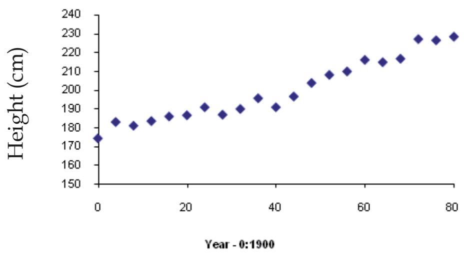
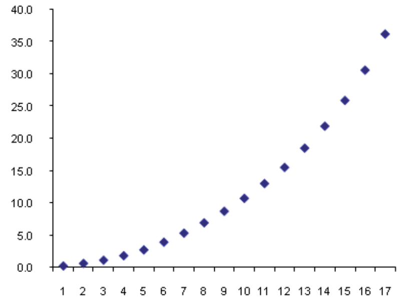


Figure 1.5

After we have drawn the scatter plot, we have to describe in words how the two variables are related. We look for the overall pattern of the scatter plot and describe it in terms of its **shape, direction and strength.** These terms are best described by means of examples.

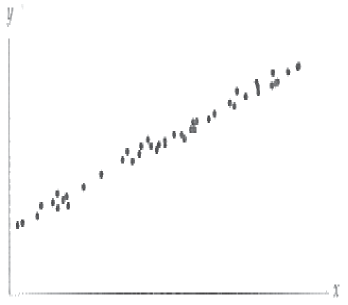
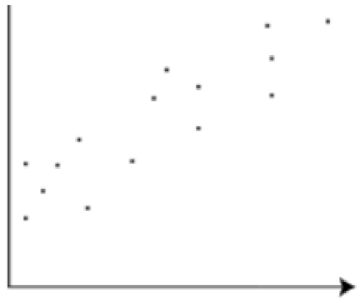
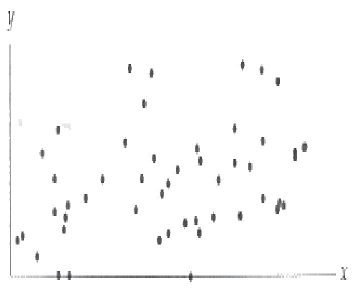
Shape: We distinguish between a linear relationship, a non-linear relationship and no relationship.

Linear relationship	Non-linear relationship
	

"No relationship" will be a random scatter of points as in figure 21.16 in your textbook.

Direction: A linear relationship with a positive slope is referred to as a positive relationship. If the slope is negative, we refer to it as a negative relationship.

Strength: We only refer to a linear relationship. The closer the scatter is around an imaginary straight line, the stronger the relationship.

Strong linear relationship	Medium-strength linear relationship	Weak linear relationship
		

We can now describe the relationship between time and the gold medal performance in the high jump as a strong, positive, linear relationship. In other words, over the years there has been an increase in the height that gold-medal athletes are jumping.

At this stage, you might ask: "If there is an obvious linear relationship, can't we draw a line that will mathematically describe this relationship?" Study pages 569-570 in Van de Walle et al. (2016).

ACTIVITY 1.12

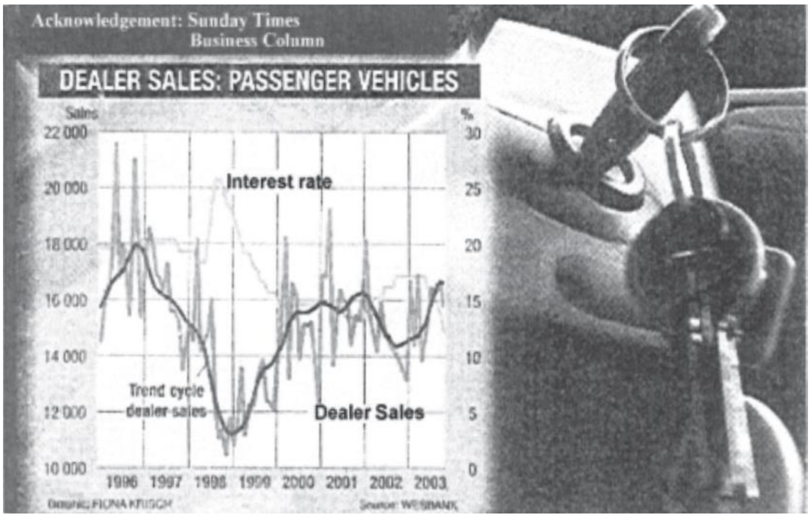
Use the following winning times in seconds for the men's 1 500-metre race at the Olympic Games and a scatter plot to answer the following question: "Are our athletes running faster?"

Year	1900	1904	1908	1912	1920	1924	1928	1932
Time	246.0	245.4	243.4	236.8	241.9	233.6	233.2	231.2
Year	1936	1948	1952	1956	1960	1964	1968	1972
Time	227.8	229.8	225.2	221.2	215.6	218.1	214.9	216.3
Year	1976	1980	1984	1988	1992	1996		
Time	219.2	218.4	212.5	216.0	220.1	215.8		

Time plots

Whenever data is collected over time, it is a good idea to plot the observations in time order. Many interesting data sets are time series (measurements of a variable that are taken at regular intervals over time). Government economic and social data are often published as time series. If you turn to the business section of a newspaper, you will often see something like the following:

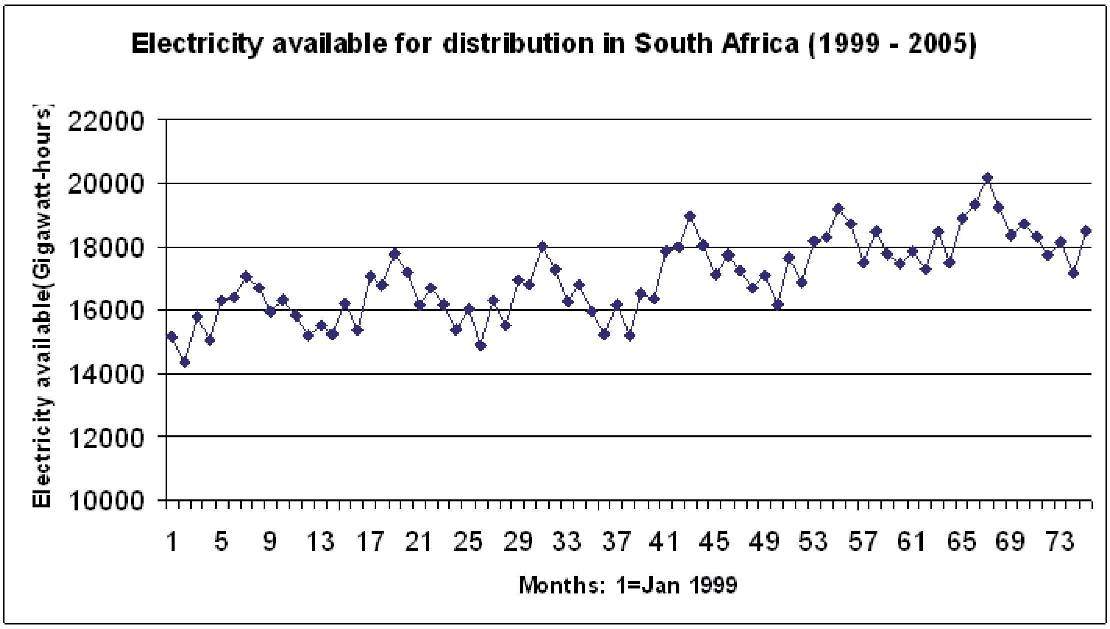
A line graph of dealer sales of passenger vehicles and interest rates between 1996 and 2003 (*Business Time* Column: *Sunday Times*, 18 April 2004)



Time plots can reveal the main features of a time series. An overall rise or fall in the time plot is called a **trend.** If there is a pattern that repeats itself at regular intervals, we call it a **seasonal variation.** Can you recognise a trend and seasonal variation in the following figure?

Trend

Seasonal variation



You should now be able to organise and present a data set in a meaningful way so that the reader can easily understand the meanings that are conveyed by the data set.

ACTIVITY 1.13

Study the data set that you gathered in activity 1.4. Apply the knowledge that you gained in this study unit to organise and present your data in a meaningful way. Bear your research question in mind and ask yourself if your method of presenting the data contributes to the answer of the study.

Study unit 2 will give you more "tools" to describe your data set.

Answers to some activities

Activity 1.8

1.8.1

Frequency table for the distances walked to Moriah

Distance (in km)	Frequency	Relative frequency	Relative %
0 ≤ x < 1	35	0.23	23
1 ≤ x < 2	13	0.086	9
2 ≤ x < 3	18	0.12	12
3 ≤ x < 4	5	0.033	3
4 ≤ x < 5	5	0.033	3
5 ≤ x < 6	48	0.32	32
6 ≤ x < 7	25	0.166	17
7 ≤ x < 8	0	0	0
8 ≤ x < 9	1	0.006	1
Total	150	**A**	**B**

When we ask you for a frequency table, it is not necessary to calculate columns A and B. You will calculate them when we ask you to draw the ogive.

1.8.2 Back-to-back ordered stem-and-leaf display for hijacking statistics in 1996

Leaf: Sandton	Stem	Leaf: Bramley
9 9 8 7 4 4 1	1	
4 3 3 2	2	4
	3	278
2	4	1125
	5	33
	6	4
	7	0

We can conclude that there were far more hijackings in Bramley than in Sandton in 1996.

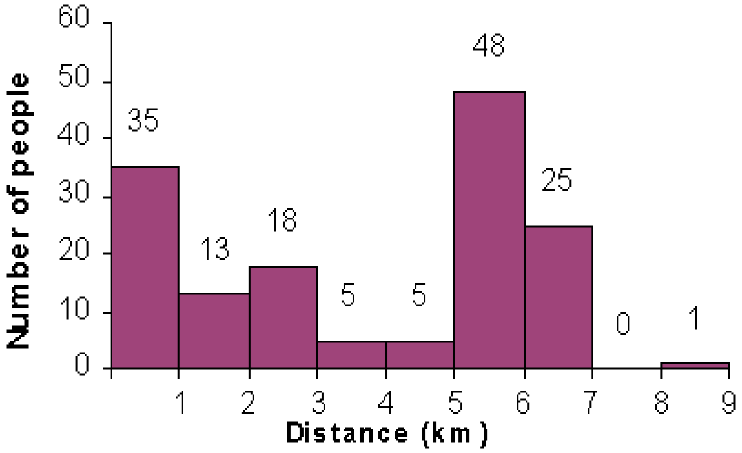
Activity 1.10

1.10.2 The second bar chart that expresses the frequencies as percentages is the true reflection of the clients' opinions. This is due to the different sample sizes. Since double the amount of people tasted the steak dishes than the chicken dishes, the comparison figures in the first chart are incorrect. The difference in the "Not so good" frequency (in chart 1) is 48, whereas the difference in percentage (in chart 2) is 20. Thus in the first chart, it appears that many more people disliked the steak dishes than there should be. The information in the first bar chart is therefore meaningless.

This teaches us that we should never ignore the sample size in our investigations and resulting calculations, especially when we are **comparing** conclusions from separate samples. When you compare samples of different sample sizes, use the relative frequency (as a percentage).

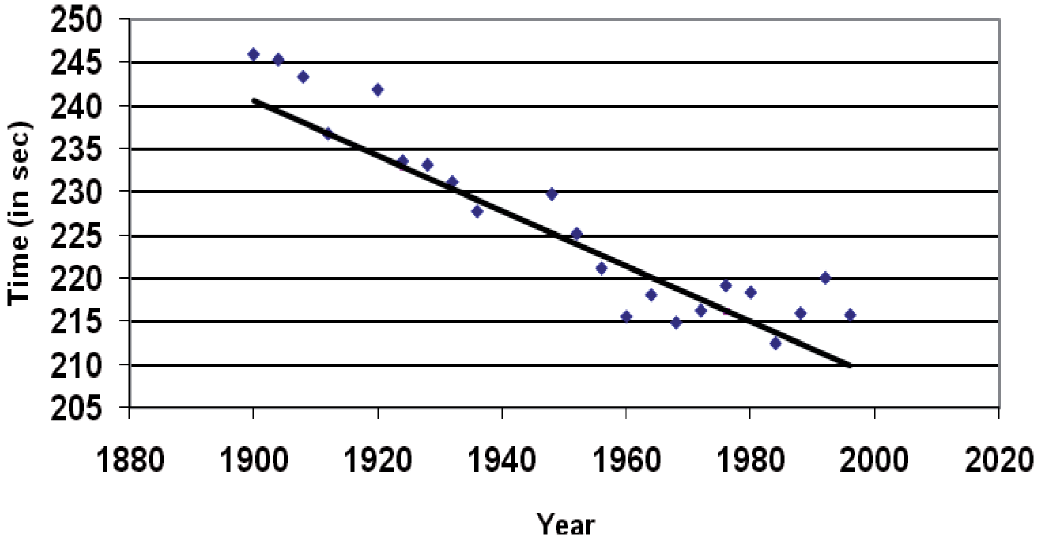
Activity 1.11

Histogram of distances travelled to Moriah



Activity 1.12

Scatter plot of winning times in seconds for men's 1500 m race at the Olympic Games



Yes, athletes are running faster. We can see this from the negative slope of the scatter plot: because the slope is negative, it is decreasing and therefore the time in seconds is getting shorter.

STUDY UNIT 2

DESCRIPTIVE STATISTICS AS AN APPROPRIATE METHOD FOR ANALYSING DATA

CONTENTS PAGE

INTRODUCTION 61

2.1 MEASURES OF CENTRALITY OR CENTRAL LOCATION 61

2.1.1 The mean 62
2.1.2 The median 65
2.1.3 The mode 67

2.2 MEASURE OF SPREAD 69

2.2.1 The range 70
2.2.2 The inter-quartile range and box-plots 72
2.2.3 The standard deviation 75

2.3 TECHNOLOGICAL WAYS TO CALCULATE THE MEAN AND STANDARD DEVIATION 81

INTRODUCTION

A week ago, while waiting at the international arrivals section of the Oliver Tambo International Airport, I recognised an acquaintance and asked him whom he was expecting. He answered that he was waiting for the arrival of Professor Smith, a key speaker from England who was going to deliver a paper at their symposium. I was puzzled because he did not have a notice or anything to identify him and asked him how he was going to recognise the person. He laughingly said that it was not a problem because he had a **description** of the man and, furthermore, the man would be accompanied by his Chinese wife and their daughter of three. To pass the time, I studied the people who arrived as they came through the barrier and, yes, I did recognise the professor and his family immediately! Even though I did not have the description of the professor, I knew what a Chinese lady and a three-year-old look like. If it is so easy to describe a person, is it as easy to describe data?

In study unit 1 we gave you several graphical techniques for describing data. In this study unit we introduce you to **numerical descriptive techniques** that will help you to be more precise when you describe various characteristics of a sample or population. We will look at:

- ♦ measures of centrality
- ♦ measures of spread and box plots
- ♦ technological ways to calculate the mean and standard deviation

Parameters

Before we look at these measures, we have to introduce a few new terms. You are aware that data can be gathered from a complete population (if it is small) or from samples that are drawn from a population. Measures that are taken by using all the data values in the population are called **parameters** and are denoted by **Greek letters;** measures that are obtained by using the data values of samples are called statistics.

2.1 MEASURES OF CENTRALITY OR CENTRAL LOCATION

When you looked at your stem-and-leaf plot or histogram, you would have noticed that most of the time there is one "peak" (the interval with the highest frequency), around which all the other values are gathered. This is called the centred value and we describe it by means of the mean, mode and median.

2.1.1 The mean

The mean

The concept "mean" (or "average") is used so often that we would start this lesson with a group discussion on what the learners understand by the concept "mean" (or "average"). We are sure that if you ask a learner to choose a learner of "average" height from the class, he/she would not choose the tallest or the shortest learner. In our mind's eye, average is a "typical" value. Usually, measurements are clustered around a "typical" value with a few exceptions ranging far away from this midpoint.

ACTIVITY 2.1

2.1.1 Read through the "levelling concept" of the mean as described in Van de Walle et al. (2016). Use this approach in a lesson plan to help your learners to "discover" the algorithm for computing the mean.

2.1.2 Give an example of measurements that are clustered around a centred value.

2.1.3 Read through the "balance point" concept of the mean as presented in Van de Walle et al. (2016). Ask yourself whether this approach will make an easy concept seem complicated to your learners. Discuss this with your colleagues. There are many applets on the Internet that illustrate this concept with animated data points. If you have access to the Internet, visit the NCTM standards website at the following address:
http://standards.nctm.org/document/examples/chap6/6.6/

The following examples illustrate the use of the mean as a measure of centrality.

Example 2.1

A newspaper article that appeared in *Beeld* (7 June 2003) stated that the weight of bread does not measure up to the standard that is prescribed by law. Is this true if a standard loaf of brown bread has to weigh 800 g?

An investigation was made where ten loaves of bread were taken from four randomly chosen shops and weighed at Weighing Instrument Services in Booysens, Johannesburg. The following table summarises the results of the study (every row represents a different shop):

735	725	715	760	745	760	750	735	730	720
810	755	770	725	740	780	735	745	755	745
745	760	680	715	740	725	690	680	715	680
755	710	750	745	780	715	735	750	755	745

The average weight is 738 g (check the answer) and this substantiates what was stated in the newspaper. If we look at the stem-and-leaf plot of the data on the following page, we see that the mean is between the 18th observation (that is 735) and the 19th observation (that is 740).

Stem-and-leaf plot of the weight of brown bread

68	0	0	0					
69	0							
70								
71	0	5	5	5	5			
72	0	5	5	5				
73	0	5	5	5	5			
74	0	0	5	5	5	5	5	5
75	0	0	0	5	5	5	5	
76	0	0	0					
77	0							
78	0	0						
79								
80								
81	0							

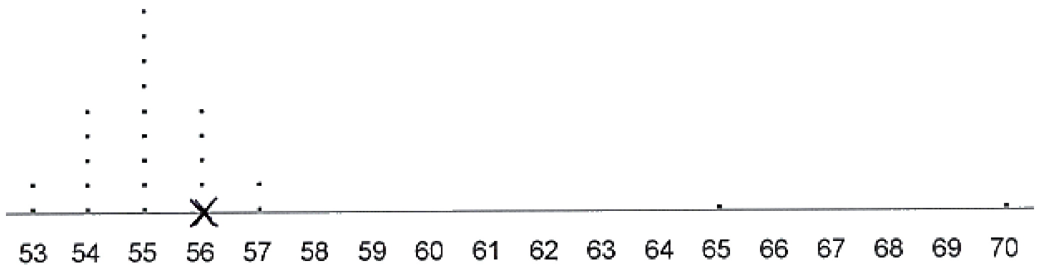
Study the following example carefully and ask yourself whether you agree with the argument that follows thereafter.

Example 2.2

The average weight of a class of 25 boys is 56 kg. Because we know that the mean is a “centred” value, we will picture a class where the majority of the boys weigh 56 kg or very close to 56 kg. Let us take a closer look at the real data: 53, 53, 54, 54, 54, 54, 54, 55, 55, 55, 55, 55, 55, 55, 55, 55, 56, 56, 56, 56, 56, 57, 57, 65 and 70

Let us “stack” the same weights on top of each other and present the data on a number line as follows (this is called a dot plot – it is seldom used but is convenient if one has a small data set). The calculated average is indicated as a cross.

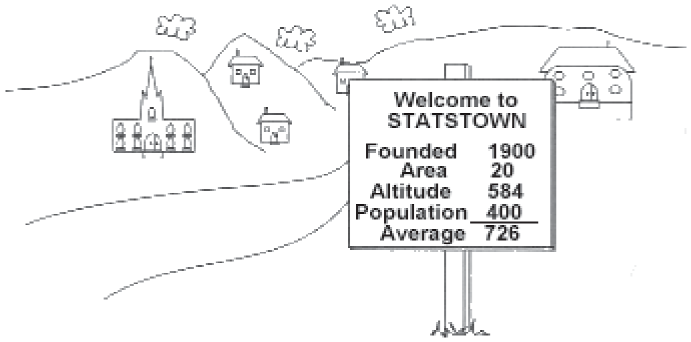
Dot plot of the weight of boys



Do you think that this plot truly reflects the centre value? Do you think that 55 kg would have been a better reflection of the distribution?

Example 2.2 illustrates an important weakness of the mean as a measure of centrality: it is sensitive to the influence of a few extreme observations. When there are a relatively small number of extreme observations, we prefer to use another measure to describe the centred value. This measure is called the median.

Before we take a closer look at the median, a word of warning: Beware of inappropriate averaging! Some people seem to have the philosophy "if there are numbers, average them". We know that you would not do something as crazy as what is presented in the cartoon, but some people come very close to doing it.



ACTIVITY 2.2

Give two examples that illustrate the following statement: An average of a **qualitative variable** is totally meaningless.

The median

2.1.2 The median

The sign in the figure below will probably not appear on a traffic island between our roads, but we have already seen a "Do not cross over median" sign on the N3 highway. Here the median is something that separates two roads, in other words it is IN THE MIDDLE. In statistics the median is the middle value in an ordered set of data: half of all the observations are smaller than the median value and the other half of the observations are larger.

To find the median of a distribution of data:

1. Arrange all the observations in order of size, from smallest to largest. A stem-and-leaf arrangement of data is very convenient.

2 If the number of observations is odd, the median is the centre observation in the ordered list.

3. If the number of observations is even, the median is the mean of the two centre observations in the ordered list.

ACTIVITY 2.3

The results of two classes for a Mathematics test are presented in the following back-to-back stem-and-leaf plot (unit: tens):

Class A		Class B
2	1	2 3
9	2	
8	3	
9	4	
8 7	5	
8 7	6	
9 8	7	0 0 0 1 2 2 3 4 6
7 6 5 3 2 2 1 0	8	0 1 2 4 4 8
9 1	9	0 1 3 9

Calculate both the mean and the median of this data set. Which one do you prefer? Motivate your answer.

Example 2.3

Because we have to arrange our observations in order of size from smallest to largest, it is convenient to have our data sorted in an ordered stem-and-leaf plot before we calculate the median. For the bread survey in example 2.1, we have the following stem-and-leaf plot:

Stem-and-leaf plot of the weight of brown bread

68	0	0	0					
69	0							
70								
71	0	5	5	5	5			
72	0	5	5	5				
73	0	5	5	5	5			
74	0	**0**	**5**	5	5	5	5	5
75	0	0	0	5	5	5	5	
76	0	0	0					
77	0							
78	0	0						
79								
80								
81	0							

Location of the median: The median will be the (n+1)/2 –th observation, in other words the (40+1)/2 = 20.5 value. Please make sure that you understand that this is not the median but simply the position of the median in the ordered data set.

Because we have an even number of observations (n = 40), the median will be the mean of the 20th and the 21st values (the two numbers appear in bold in the stem-and-leaf plot).

Median = (740 + 745)/2 = 742.5 and this value does not differ much from the mean.

ACTIVITY 2.4

2.4.1 You might recall that in case study 1.3 in study unit 1, the manager of the bus company calculated the mean distance that a family would have to travel to Moriah and was of the opinion that the mean distance of 3,5 km is still a walkable distance. Calculate the median distance and use this value to convince the bus manager that a bus is indeed necessary to transport the families to Moriah. Remember that the manager is not a statistician and that you will have to explain to him what the median value represents. Stress the fact that 50% of the observations lie above the median value.

2.4.2 Can you find examples in the popular press where the mean of a data set is cited and other examples where the median is cited? Why do you think the authors of the articles chose to cite those particular measures of centrality?

The mode

2.1.3 The mode

The mode is seldom used to describe the central value of a distribution. Take note of the definition of the mode as it is defined by Van de Walle et al. (2016) on page 571. Most distributions have one mode and these are called unimodal distributions.

You should by now realise that you must have a thorough understanding of the advantages and disadvantages of each measure to be able to decide on an appropriate one to solve a particular problem.

To summarise: How do you determine which measure of central location to use – the mean, the median or the mode? If the data is qualitative, the only appropriate measure of central location is the mode; if the data is ranked, the most appropriate measure of central location is the median. For quantitative data, however, it is possible to compute all three measures. The measure that you will use will depend on your objective. The mean is most popular because it is easy to compute and to interpret. (In particular, the mean is generally the best measure of central location for purposes of statistical inference – as you will see in later study units.) However, it has the disadvantage of being unduly influenced by a few very small or very large measurements. To avoid this influence, you might choose to use the median. This could well be the case if the data consisted, for example, of salaries or house prices. The mode, which represents the value that occurs most frequently (or the midpoint of the class with the largest frequency) should be used when the objective is to indicate the value (such as a shirt size or house price) that is most popular with consumers.

One last comment before we look at measures of spread: At this stage, you might ask why the average gave the manager of the bus company the wrong impression of the distance that people have to travel to Moriah. Statistical measures and the methods that are based on them are generally meaningful if we look at the **complete picture.** It is irresponsible to draw conclusions on the basis of single statistics. Your analysis starts with a visual investigation of the data and the other measures that add value to the analysis. The visual investigation of the stem-and-leaf plot showed that the distribution had two "humps" and here other measures than the mean are more appropriate to describe the data set.

The balance approach to the mean (which you studied in Van de Walle et al. (2016)) clearly illustrates that many different distributions can have the same mean. For example, learners should recognise that the statement "the mean score on a test was 50% can cover several situations, including the following: all scores are 50%; half the scores are 40% and half the scores are 60%; half the scores are 0% and half the scores are 100%; one score is 100% and 50 scores are 49%. In other words, a measure of centrality is not enough to describe a data set accurately.

Quartiles

At this stage, we want to introduce the first and third quartiles. A quartile is not a "measure of centrality", but is calculated in the same manner as the median and we will use it in the next study unit.

As the term indicates, "quartiles" divide an ordered set of observations into quarters. 25% of our ordered observations will be less than the first quartile, 50% will be less than the second quartile and 75% will be less than the third quartile. Yes, you have come to the correct conclusion: the median is nothing other than the second quartile.

To calculate the quartiles:

- Arrange all the observations in order of size from smallest to largest and locate the median M in the ordered list of observations.
- The first quartile Q_1 is the median of the observations below the median M.
- The third quartile Q_3 is the median of the observations above the median M.

Example 2.4

With reference to the bread survey in example 2.1, we have the following stem-and-leaf plot:

Stem-and-leaf plot of the weight of brown bread

68	0	0	0					
69	0							
70								
71	0	5	5	5	5			
72	**0**	**5**	5	5				
73	0	5	5	5	5			
74	0	**0**	**5**	5	5	5	5	5
75	0	0	0	**5**	**5**	5	5	
76	0	0	0					
77	0							
78	0	0						
79								
80								
81	0							

$Q_1 = (720 + 725)/2 = 722{,}5$

$Q_3 = (755 + 755)/2 = 755$

2.2 MEASURES OF SPREAD

In the previous paragraph we concluded that a measure of centrality is not enough to describe a data set accurately. The simplest useful numerical description of a distribution consists of both a measure of centrality and a **measure of spread.**

To demonstrate the necessity of using a measure of spread, let us consider the following example.

Example 2.5

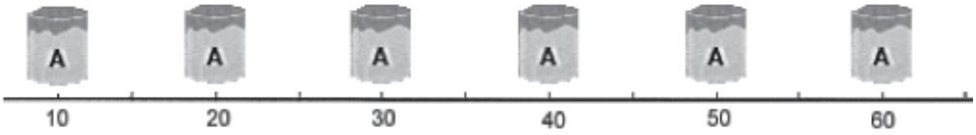
Two experimental brands of outdoor paint were tested to see how long they would last before fading. The results (in months) follow:

Brand a	10	60	50	30	40	20
Brand B	35	45	30	35	40	25

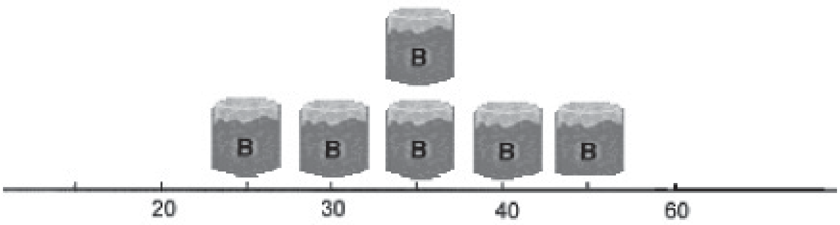
The mean for brand A is 210/6 = 35 months and for brand B it is also 35 months

Since the means are equal, one might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion can be drawn.

Variation of Paint Brand A



Variation of Paint Brand B



These figures show that brand B performs more consistently (that is, it is less variable). Note that the words **"scatter", "variability"** and **"dispersion"** have the same meaning as "spread".

2.2.1 The range

Range

ACTIVITY 2.5

It is very important that the learners are aware of the fact that the mean by itself is insufficient to describe data numerically. The simplest measure of spread is the **difference between the largest and the smallest observation,** which is called the "**range".** Incorporate the following example in your lesson plan. The table summarises the height (in cm) of five players from soccer team A and five players from soccer team B:

Team A	183	185	193	193	198
Team B	170	183	193	193	213

- Calculate the mean height of each team.
- Calculate the median height of each team.
- What is the mode of each team?
- Present this data on a number line (encourage your learners to be creative as we did with the paint – this can be a fun group activity).
- From the values of the average, median and mode, and also the graphical presentation of the data, what can we conclude?
- Can you suggest a measure that will, in your opinion, describe the difference in the data?
- Do you think that this measure has any disadvantages?

NOTE:

The interpretation of measures of centrality and spread are very important. At the International Conference on Teaching Statistics (ICOTS), six lecturers from the University of Tasmania emphasised the following (Watson & Kelly 2002):

> Variation is at the heart of all statistical investigation. *If there were no variation in data sets, there would be no need for statistics.* Although statistical variation is taken for granted by statisticians, school students often have little concept of appropriate variation and many tertiary students also fail to understand the variation behind the formulae they learn to measure it. Traditionally, standard deviation is the common formula used to measure spread; *however, due to its complex nature it is often avoided by educators* and definitely not included in the primary and middle school mathematics curriculum. The concept of variation, however, is an important element of basic understanding of statistics and "real world" functioning and does not necessarily rely on the understanding of complex formulae to be taught effectively.

The learners should discover that even though the advantage of the range is its simplicity, this simplicity is also its disadvantage. Because the range is calculated from just two observations, it tells us nothing about the other observations. Use the following two data sets that are completely different and yet have the same ranges as an example.

Set 1: 4, 4, 4, 4, 5, 6 and 50 Set 2: 4, 8, 10, 24, 28, 31 and 50

ACTIVITY 2.6

Design a lesson plan that will demonstrate to the learners the restrictions of the range as a measure of spread.

2.2.2 The interquartile range and box plots

The nature of data is usually such that we have a concentration of data around a "centred value" and then a "flattening" out to some extreme values. Take, for example, the scores of 30 learners. You have an average score with scores that are clustered around it, and then the very bright scores and the very low scores. So does it not make sense to describe the centre, say 50%, of our observations? This is precisely what the interquartile range (IQR) does. The IQR is the distance between the first and the third quartiles. The IQR is not affected by changes in the tail of the distribution that will definitely influence the range.

At this stage, we would like to introduce the five-number summary of data and the box plot.

In a five-number summary, the following five numbers are used to summarise the data:

1. smallest value
2. first quartile
3. median
4. third quartile
5. largest value

A box plot (also called a box-and-whisker plot) is a graphical summary of data that is based on the five-number summary. The following steps are used to construct the box plot:

1. A box is drawn with the ends of the box located at the first and third quartiles. This box contains the middle 50% of the data.

2. A vertical line is drawn in the box at the location of the median.

3. By using the IQR, the following limits are calculated: lower limit is Q_1 – 1,5(IQR) and upper limit is Q_3 + 1.5(IQR). Data outside these limits are considered *outliers*.

4. Lines are drawn from the ends of the box to the smallest and largest values inside the limits that were calculated in step 3. This is called *whiskers*.

5. Finally, the location of each outlier is shown with a * or dot.

Example 2.6

For many coffee shops, the time that customers linger over coffee negatively affects profits. The ideal is that customers will leave after 35 minutes. To learn more about this variable, a sample of 50 customers was observed and the amount of time (minutes) that they spent in the restaurant (after they were served) was recorded as follows:

25	28	29	23	32	41	42	32	28	33	26	25	29	26	28	25	27	34	28	27	23	32	32	37	
37	25	30	28	34	27	28	40	30	37	31	29	36	26	25	29	36	26	36	27	33	29	29	28	33

To make more sense from this data, we decided to summarise it in a box plot. We do realise that we will have to explain it carefully to the coffee shop owner.

Before quartiles can be calculated, the data has to be sorted in increasing order. If it is possible, it is convenient to arrange the data in a stem-and-leaf plot.

Stem-and-leaf plot of the time spent (minutes) over coffee in a coffee shop

Stems	**Leaves**
2	3355555666677778888888999||999
3	001222233344666777
4	0123

The median: The position of the median is the (n+1)/2 = 51/2 = 25.5 position (indicated with ||). Since halfway between 29 and 29 is 29, the median is 29.

The first quartile: (Remember that the median was defined in such a manner that 50% of the observations are below the median and 50% of the observations are above the median.) The lower (first) quartile = the median of the lower 50% of the observations = the 13 value = 27.

The upper (third) quartile = the median of the upper 50% of the observations = 33

Do we have an outlier? To test whether we have outliers (remember, an observation that falls outside the overall pattern), we calculate the **boundaries** of the "overall pattern" as follows.

The lower boundary: The first quartile – 1.5 x IQR = 27- 1.5(33 – 27)

= 27 – 9 = 18

The upper boundary: The third quartile + 1.5 x IQR = 33 – 1.5(33 – 27)

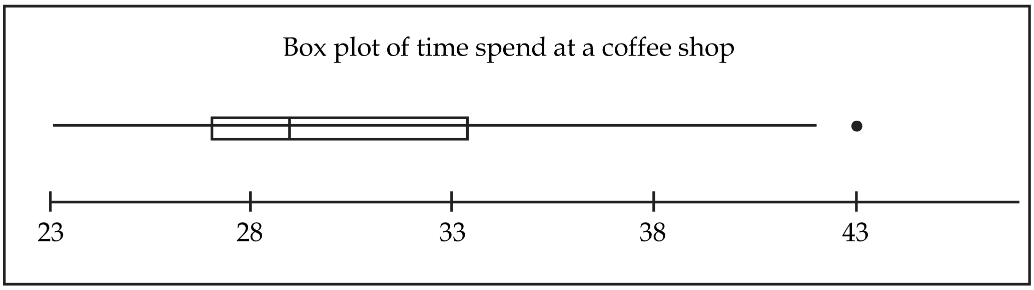
= 33 + 9 = 42

Because the last value of 43 lies outside the upper boundary, we can define it as an outlier.

Schematically:

_____[____________________________________]_________
18 42 43

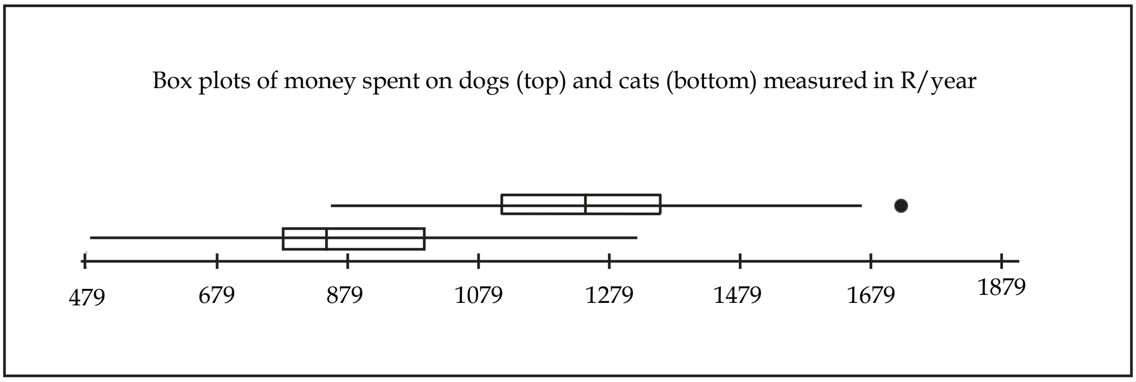
The resulting box plot will be:



When you explain the box plot to the manager of the coffee shop, stress the point that 75% of his customers spend at most 33 minutes drinking their coffee, 50% (the IQR) spend between 27 and 33 minutes, and half of his customers spend at most 29 minutes.

ACTIVITY 2.7

A random sample of dog and cat owners was taken to compute the amount of money that they spent on their pets per year and the following box plots were drawn:



Box plots of money spent on dogs (top) and cats (bottom) measured in R/year

List **all** the information that you can get from the box plot and the conclusions that you can draw. What “research question” can be answered from this information?

These measures of spread are intuitive and easy to calculate, but their main disadvantage is that they do not take into account all the observations. The standard deviation is a measure of spread that does take into account all the observations, but is a little trickier to interpret. It is also always part of the summary statistics that any software package, for example Excel, provides and we have to be able to interpret it.

2.2.3 The standard deviation

ACTIVITY 2.8

Compare hand spans: How far are you from the mean?

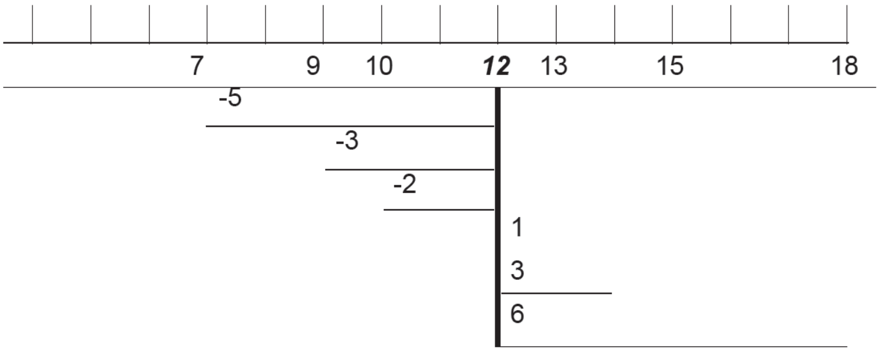
1. Spread your hand out on a ruler and measure your hand span as the distance from the tip of your thumb to the tip of your little finger when you spread your fingers. Measure it to the nearest half centimetre.
2. Find the mean hand span for your group.
3. Make a dot plot of the results for your group. Mark the mean on a number line.
4. Give two reasons why the measurements are not all the same.
5. How far from the mean are the hand spans of your group?

6. Calculate the difference from the mean. Invent a method that will give a "typical" distance from the mean.
7. Compare your results with the other groups' results.

This activity is an attempt to lead the learners to "discover" variation and possible ways to measure it. Let them work in groups. Give them a hint by telling them that they can measure the deviations from the mean and perhaps get the "average deviation". From the feedback sessions, they should "discover" that the average deviation from the mean is always zero.

Example 2.7

Consider the following data set and present it on a number line: 7, 11, 9, 15, 13, 10 and 18.



The average of the data set is 12. The average and the deviations from the average are indicated in the above diagram. Depending on your learners' mathematics background, they should find it relatively easy to prove the fact that the total deviations from the mean will always be zero:

$$\sum_{i=}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Squaring the deviations makes them all positive and leads to the definition of the **variance** as the average of the squares (with the exception that instead of dividing the total by n, we divide it by n-1):

Variance

$$s^2 = \frac{\Sigma\,(x - \bar{x})^2}{n - 1}$$ and for the population $$\sigma^2 = \frac{\Sigma\,(x - \mu)^2}{N}$$

Standard deviation

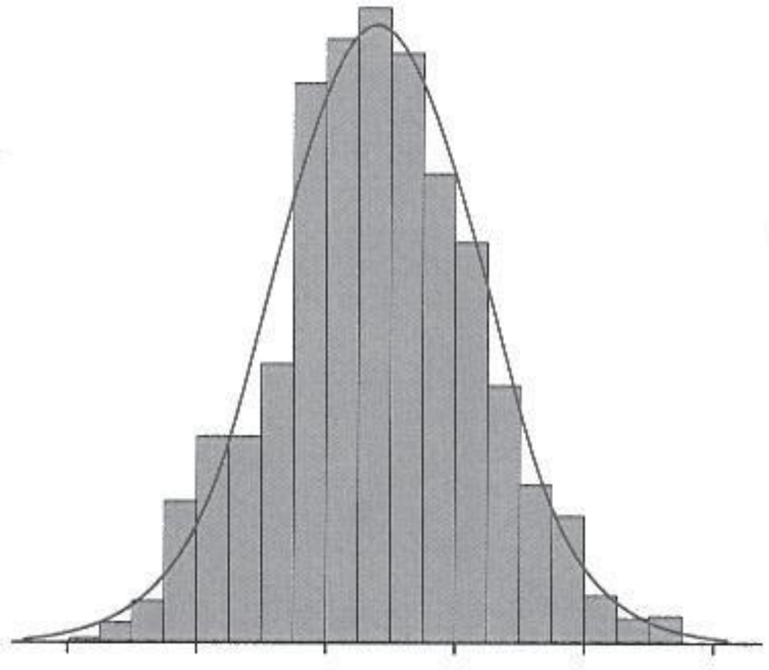
The reason why we divide by (n-1) if it is a sample and by N if it is a population is too technical to explain at this stage. Make the learners aware of this fact because some pocket calculators have sn and s(n-1) instead of sx and σx.

Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. The variance of people's height is, for example, measured in squared centimetres. The remedy for this is to take the square root and the square root of the variance is called the **standard deviation.** The standard deviation measures the spread about the mean in the original scale.

Properties and interpretation of the standard deviation

- The standard deviation is zero only if there is no spread and all the observations are identical.
- As the observations become more spread out about the mean, s gets larger. Remember to look at the magnitude of the observations in the data set when interpreting the size of s. If a test was for example measured in percentages, a standard deviation of 5 will be considered a small number. On the other hand, if the test was out of 20, a standard deviation of 5 will be considered a large number.
- The standard deviation is used to determine the consistency of a variable. For example, when one-litre milk containers are filled, the variation in the content should be small.
- Depending on the shape of the histogram, useful information can be extracted from the mean and standard deviation. If the histogram is approximately symmetrical or "bell shaped" (as illustrated in the following figure), we can use the 68–95% rule:

  > Approximately 68% of the data lie within **one standard deviation** of the mean and approximately 95% of the data lies within **two standard deviations** of the mean.



Other statistical procedures make use of these characteristics but we will refer to them when we look at inference.

- If the shape of the histogram does not suggest a symmetrical distribution, we can apply a theorem that was developed by a Russian mathematician named Chebyshev who states that the proportion of observations in any sample that lie within k standard deviations of the mean is at least (1-1/k^2) for k>1. When for example k=2, at least three quarters or 75% of all observations lie within two standard deviations of the mean.

  For an example where Chebychev's theorem is used, suppose that the test scores of 100 students in a statistics course had a mean of 70 and a standard deviation of 5. How many students had test scores between 58 and 82? For the test scores between 58 and 82, we see that (58-70)/5 = -2.4 indicates that 58 is 2,4 standard deviations below the mean and that (82-70)/5 = +2,4 indicates that 82 is 2,4 standard deviations above the mean. If we apply Chebyshev's theorem with k = 2,4 and we have (1-1/k^2) = (1-1/(2,4)^2) = 0,826, at least 82,6% of the students must have test scores between 58 and 82.

  In the same way that extreme values (outliers) influence the mean, it also influence the standard deviation and in this case it is better to use the IQR.

- An important application of the standard deviation is yet another measure of variability, the **coefficient of variation.**

  Coefficient of variation

  Sometimes we have to compare the variability that is present in two sets of data. This can usually be done by comparing the two variances or standard deviations directly, provided that the data satisfies two conditions: (1) the same unit of measurement is used in both data sets and (2) the means of the two data sets are approximately equal. If either of these conditions is not met, we need a relative measure of spread to compare the variability of the two data sets.

The following example emphasises the need for a measure that gives the relative variation of the mean.

Example 2.8

Suppose that we want to determine the accuracy of two different instruments that measure distance. (Measuring an object a number of times and checking the variability of the measurements can determine the accuracy of the instrument. If the instrument is accurate, the measurements should all be more or less equal [that is, there should be little variability in the measurements]. If however the measurements vary to a great extent, the instrument does not seem to make reliable measurements.)

Suppose that the length of a needle is measured five times with an instrument. The average of the five measurements is 3,2 cm and the standard deviation is 1 cm. Another instrument is used to measure the length of a room five times. The average of these five measurements is 350 cm and the standard deviation is 1 cm.

Can we conclude that since the two data sets have the same standard deviation, the two instruments are equally accurate? Certainly not! The variation in the measurements of the needle is 1 cm and the needle itself is only about 3 cm long! On the other hand, the measurements of the room vary by just 1 cm, while the room itself is about 350 cm long. The measurement of the room seems much more accurate than the measurement of the needle.

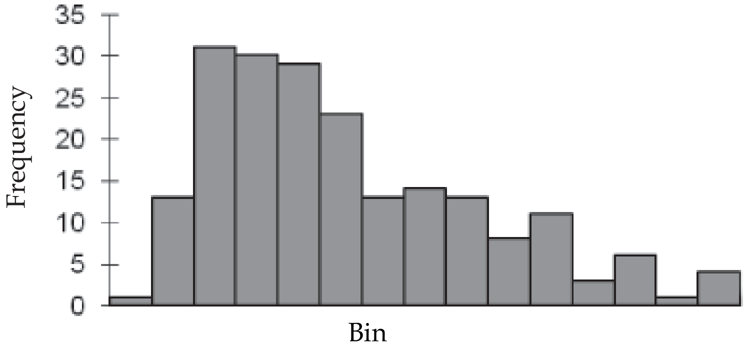
It is clear that that we need a measure that gives the relative variation with respect to the mean. Such a measure is the coefficient of variation (cv): cv = standard deviation/mean.

ACTIVITY 2.9

2.9.1 Many traffic experts argue that the most important factor in accidents is not the average speed which cars travel but the amount of variation in the speed. Suppose that the speeds of a sample of 200 cars were taken over a stretch of highway where there has been numerous accidents.

You have Excel and get the following statistics (units are km/hour) very easily and quickly:

Histogram of speed of 200 cars



Smallest = 104	Q1 = 112	Median = 116.8	Q3 = 128
Largest = 150.4	Mean = 119.65	Standard deviation = 10.21	

Apply your knowledge of spread to interpret the results.

2.9.2 The question that we would like you to answer is: Do customers at a supermarket favour the checkout counter that is closest to the entrance? Suppose that a consultant for the supermarket counts the number of arrivals per hour during a sample of 150 hours at two checkout points: checkout point A is close to the entrance and checkout point B is not close to the entrance. He calculates the mean and standard deviation for both sets of data but does not have time to write a report for the manager of the supermarket. He gives you with the statistics and asks you, as a favour, to report back to the manager. Briefly describe to the manager what these statistics tell you.

	Counter A (close to the entrance)	**Counter B (not close to entrance)**
Mean	98	92
Standard deviation	15	19

2.3 TECHNOLOGICAL WAYS TO CALCULATE THE MEAN AND STANDARD DEVIATION

Until now, our emphasis has been on the interpretation of statistics. Now we have to show you how to calculate the mean and standard deviation.

If you have access to a computer and Excel, study the procedures that are given in Keller to calculate the descriptive statistics and to draw the graphical presentations. If you have a computer lab at your school, organise workshops and show your learners how to do it.

If you do not have access to a computer, the following procedure describes how to do the calculations on a SHARP scientific calculator. Although calculators might differ a little, basically the procedure is the same. Let us use the following small data set: 10 15 12 27 and 18.

1. Put your calculator in the STAT mode. The procedure to do this is usually indicated somewhere on the calculator.
2. You will see the word "DATA" printed underneath the M+ key. Key in your data one by one as follows: 10 and press the M+ key, 15 and M+, et cetera. If you realise that you keyed in the wrong number **before** you press the M+ key, you can press the On/C key and re-enter the data. If you realise that you keyed in the wrong number only **after** you pressed the M+ key, you will have to exit the STAT mode and start from the beginning.
3. To find the mean, look for the key above which a small x-bar is printed. To access the value of the mean, you usually have to press the "RCL" and then the key with the small x-bar printed at the top. Answer 16.4
4. To find the standard deviation, look for the key on top of which the small sx is printed. To access the value of the standard deviation, you usually have to press the "RCL" and then the key with the small sx printed at the top. Answer 6.655824517
5. Easy, isn't it! Remember to exit the STAT mode and to re-enter it again before you start with a new set of data.

ACTIVITY 2.10

Practise the use of the "STAT-mode" with the following data sets:

2 13 16 17 22 25

Mean = 15.833 Standard deviation = 8.03

150 203 256 267 220 234 247

Mean = 225.2857   Standard deviation = 39.63

ACTIVITY 2.11

Refer back to the research problem that you started with in activity 1.4 and worked on in activity 1.13. Apply the knowledge that you have gained in this study unit to get the maximum information from your data set. We are sure that the end result is something to be proud of.

At this stage, you should know the importance of a representative sample and techniques to obtain a representative sample. You learned graphic and numeric methods that have given you the tools to summarise and describe data. All of this is part of descriptive statistics. However, at the end of the day, you have to draw conclusions from the information that gained from the SAMPLE regarding the POPULATION. To be able to do this, you need inferential statistics ("infer" means to deduce, conclude, surmise or assume). A study of probability, the subject of the next study unit, will give you a foundation for developing the theory behind inferential statistics.

Answers to some activities

Activity 2.4

2.4.1 Dear Bus Manager

> Although we calculated earlier that the average distance for families to travel to Moriah is 3.6 km, we have investigated our data in more depth and now have a clearer picture of the real situation. The median is the midpoint of the data set and is not influenced by outlier values as the mean/average is. It is calculated by finding the half-way point, in other words by making sure that 50% of the data is above it and the other 50% below it.
>
> The median's position is (n + 1)÷2 = (150 + 1)÷2 = 75.5
>
> Thus the median falls between 4.2 and 4.4 and can be calculated by (4.2 + 4.4) ÷2 = 4.3. The median distance is 4.3 km, which is significantly different from the mean distance of 3.6 km. This basically means that half the people have to walk a distance of **more than** 4.3 km. We hope that you will reconsider your offer in the light of this new information.

(Student)

Activity 2.5

2.5.1 Team A: mean = (183+185+193+193+198)÷5 = 190.4 cm.
Team B: mean = (170+183+193+193+213)÷5 = 190.4 cm.

2.5.2 Team A: 183 185 **193** 193 198 - median height is 193 cm.
Team B: 170 183 **193** 193 213 - median height is 193 cm.

2.5.3 Team A: mode is 193
Team B: mode is 193

2.5.4 Variation of team A

170 180 190 200 210 220

Variation of team B

170 180 190 200 210 220

2.5.5 According to the average, median and mode, the height distribution of both teams could be exactly the same. However, the number line shows us that team B's height distribution is far more varied/dispersed than team A's.

2.5.6 The range shows the difference in distribution clearly and is simple to work out. It is calculated by subtracting the smallest value from the largest value.

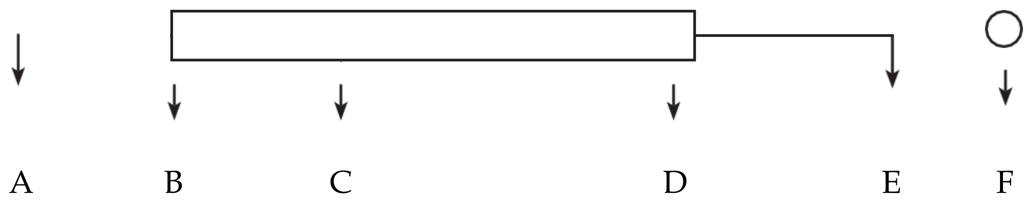
The range of team A is: Range = 198 – 183 = 15.

The range of team B is: Range = 213 – 170 = 43.

2.5.5 Yes. Since it is calculated by using only two observations, it tells us nothing about the other observations that lie between the largest and smallest values.

Activity 2.7

Let us quickly revise box plots.



A: smallest value of the data set

B: first quartile

C: median value

D: third quartile

E: biggest value of the data set

F: outlier

- Because we have the smallest and the biggest values, we can calculate the range.
- Because we have the first and the third quartiles, we can calculate the inter-quartile range.

Student:

I can read the following information from the box plot:

- the range of money spent on dogs and cats respectively
- the median amount spent on dogs (± R1 229) and cats (± R839)
- the lowest (R479) and highest (R1 269) amounts spent on cats
- the lowest (R849) and highest (R1 769) amounts spent on dogs
- 50% of the cats cost between R779 and R989 to keep (inter-quartile range)
- 50% of the dogs cost between R1 129 and R1 349 to keep (inter-quartile range)

I drew the following conclusions from the box plot:

- ♦ More than three quarters of the cats cost less than one quarter of the dogs.
- ♦ One quarter of the dogs costs more to keep than all the cats.
- ♦ It is more expensive to keep dogs as pets than it is to keep cats as pets.
- ♦ The dogs' values are more evenly spread out than the cats' values.
- ♦ The cats' values are more spread out in the upper half than in the lower half.

The research question could be: Is it cheaper to have a dog or a cat for a pet?

Activity 2.9

It is easy to compare the mean of counter A with the mean of counter B, but the question that is more difficult to answer is: How does the standard deviation of 15 of a data set with a mean of 98 compare with a standard deviation of 19 of a data set with a mean of 92? This is why we calculate the **relative variation**, namely the coefficient of variation (see example 2.7).

2.9.2 Counter A: cv = 15÷98 = 0.153

Counter B: cv = 19÷92 = 0.207

The measurements at counter A is less variable than those at counter B, the data is clustered closer to the average of 98 than the average of 92, and we can conclude that customers favour a checkout point that is closer to the entrance over one that is further away.

If we use Chebyshev's theorem, which states that at least 75% of all observations lie within two standard deviations of the mean (in other words for counter A, between 98 – 2(15) = 68 and 98 + 2(15) = 128 and for counter B between 92 – 2(19) = 54 and 92 + 2(19) = 130), counter A (close to the entrance) seems to be the most popular.

STUDY UNIT 3

PROBABILITY: THE STUDYOF RANDOMNESS

CONTENTS PAGE

INTRODUCTION 87

3.1 RANDOMNESS 88

3.2 HOW TO DESCRIBE EVENTS 91

3.2.1 The single event 91
3.2.2 Two events 92
3.2.3 Tree diagrams 94

3.3 PROBABILITY: THREE APPROACHES TO ASSIGNING PROBABILITY 99

3.3.1 The relative frequency approach 99
3.3.2 The classical approach 103
3.3.3 The subjective approach 116

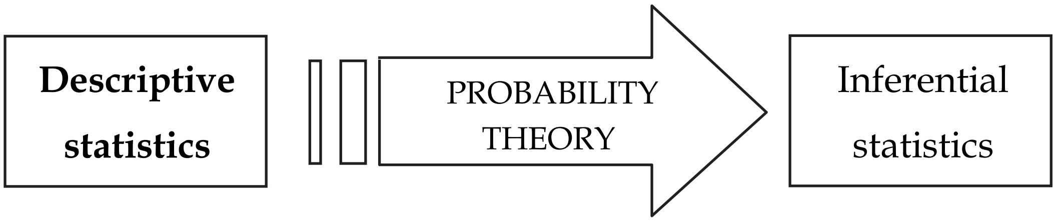
INTRODUCTION

The husband of one of our colleagues did a course in statistics during his studies and would always says that a statistician is a very intelligent person because in order to understand probabilities, one must have "more" brains than the average person. This, of course, is not true. However, it is not difficult to make this part of statistics "seem" complicated. Probability problems are "word problems" and we know that our learners do have problems with these. Because of this, it is extremely important that they always understand the "story" behind a problem (but we will come back to this aspect at a later stage). We will show you techniques to "break down" a problem in such a way that at the end of the day, you will ask yourself "what was so difficult?"

To put the study of probability in context, you should ask yourself the following questions (with regard to your knowledge of statistics):

Where am I? Where do I want to go? What do I need to get there?

At this stage, you know the importance of a representative sample and techniques for finding a representative sample. You have learned graphic and numerical methods that gave you tools for summarising and describing data – all part of descriptive statistics. However, **at the end of the day, you want to draw conclusions from the information that you gained from the SAMPLE of the POPULATION.** To be able to do this, you need inferential statistics ("infer" means to deduce, conclude, surmise or assume). Probability, the subject of this study unit, gives you a foundation for developing the theory behind inferential statistics. Schematically, this can be presented as follows:



When you work through this study unit, bear in mind the concepts and skills that are stipulated in the Curriculum Assessment Policy Statements for Grades 6 to 9) (see the annexure).

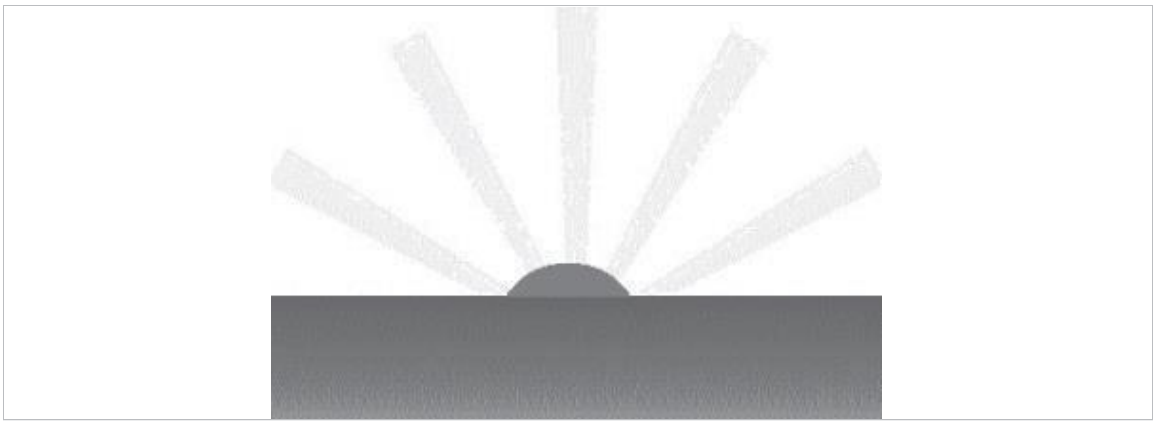
3.1 RANDOMNESS

We cannot stress enough that your learners must have a clear understanding of what probability is: A study of randomness! Because probability attempts to **quantify uncertainty** as a tool for decision making, it is one of the most difficult concepts to explain. On the one hand, it is not difficult to understand the formal rules of probability and the ability to apply them can be acquired through enough exercises; however, on the other hand, chance is also part of everyday life – in games of dice, lotteries or predicting the weather. Does this "informal" probability that is firmly established in common culture and the fact that we have **multicultural** classrooms have an effect on the understanding of "formal" probability? Amir and Williams* investigated selected cultural influences on the probability thinking of 11-year-old to 12-year-old children in England. They found that language, beliefs and experience do influence children's "informal" knowledge of probability. The educator has to remember that the learners bring this knowledge to the classroom and that it can have an influence when they interpret probability.

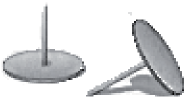
*Reference: Amir, GS & Williams, JS. 1999. Cultural Influences on children's probability thinking. *Journal of Mathematical Behaviour* 18(1):85–107.

The following activities can be build into a lesson so that your learners can "discover" randomness:

Divide your learners into three groups. Group 1 should come up with examples of events where the outcome remains the same independent of how many times the event occurs. For example: If we release a ball while standing on a chair, it will fall to the ground. Which "certain" event is depicted in the following picture?

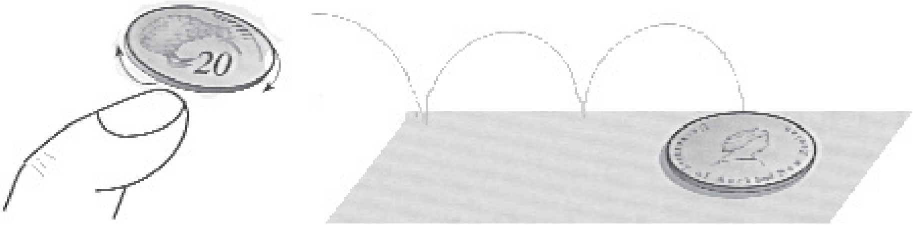


Group 2 should give examples of "impossible" events, for example one cannot be in the class and at home at the same time.

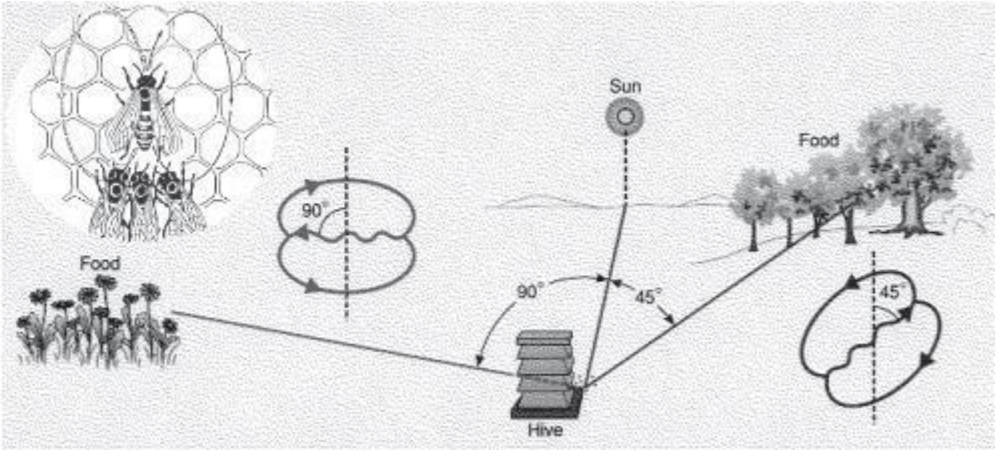


Group 3 should give examples of events where the individual outcomes are uncertain or unpredictable. Consider, for example, shaking a thumb tack in a cup and tossing the tack out onto a table. It can land in one of two ways (see picture) and we cannot say which way it will land.

The following picture also describes an event where the outcome is unpredictable.



The events that should be described by group 3 are called random events. So, is "random" in statistics a synonym for "haphazard"? No! If you look at a beehive, it will at first seem to be a very unordered community. Yet, did you know that on their return to the hive with pollen and nectar, the worker bees perform an elaborate dance on the vertical surface of a comb to communicate the location of the food to their colleagues. If the source is relatively distant from the hive (as it generally is), the dance takes the form of a figure eight. The angle of the straight run or "waggle" from a vertical position is equal to the angle from the hive between the sun and the nectar or pollen source. If the flowers are located 45 degrees to the right of the sun, the dance will be oriented 45 degrees to right of the vertical position (see the following figure). The distance of the straight waggle run is proportional to the distance from the hive to the source. And no, you will not be examined on entomology – what we want to illustrate is that what seems to be a haphazard event is actually a very orderly event with definite patterns (this information might also be useful in another area of Mathematics).



Many events (such as those that are described for group 3 above) **have individual outcomes that are uncertain, but there is nevertheless a kind of order that emerges in the long run after a large number of repetitions of the event. These events are called "random".** In the case of tossing a coin, for example, some diligent people have in fact made thousands of tosses. The classical example is that of the South African mathematician John Kerrick. While imprisoned by the Germans during World War II, he tossed a coin 10 000 times and noted the number of times that heads fell upwards: 5067 (that is almost 50%). Fortunately, in this age of technology, what took ages to accomplish in the past can be simulated within seconds.

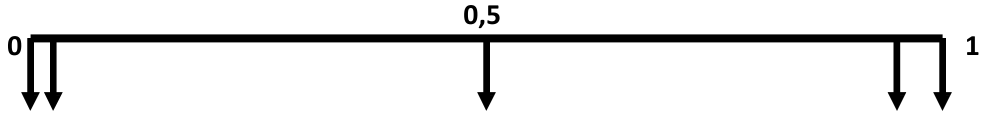
Random

Although the outcome of a random event cannot be predicted in advance, there is still a regular pattern in the results – a pattern that emerges clearly only after many repetitions. This remarkable fact is the basis for the idea of probability. It is important to understand that we can never observe a probability exactly. A probability gives us an idea of what will happen "in the long run".

In the literature we quite often read about a random "experiment". A random experiment is simply an action (like tossing of a coin) or a process that results in a random outcome. **The set of all possible outcomes of a random experiment is called the sample space (S)** of the experiment. For example, the sample space of the experiment "Toss a coin" is S = {heads, tails}.

ACTIVITY 3.1

You want to explain to your class the concept "probability" as the likelihood that an event will occur which ranges between impossible and certain. You draw the following "probability line" on the board:



When preparing the lesson, you define ten events that you expect the learners will place at the arrow positions on the probability line. Number the arrows from left to right as 1.1, 1.2, 1.3, 1.4 and 1.5, and define two events per arrow.

We said that we cannot observe a probability exactly and yet we have to define a mathematical expression or model for randomness. The basis for all probability models has two parts:

1. a list of all possible outcomes, and
2. a probability for each outcome

How do I know that I listed all the possible outcomes? The task of finding a solution can be a difficult one if it is done by guessing alone. In the following paragraph we look at techniques to get "all the possible outcomes" in a systematic way.

3.2 HOW TO DESCRIBE EVENTS

3.2.1 A single event

Example 3.1

(a) Suppose that a salesman can travel from Cape Town to Johannesburg by plane, train or bus. List all the possible ways that he can travel.

What is the "event" that we want to describe? Answer: the means of travel. The sample space of this event is: S = {plane, train, bus}.

(b) Suppose that one can be married, never married, divorced or widowed. List the possible "status" of a person.

What is the "event"? Answer: the marital status of a person. The sample space of this event is: S = {married, never married, divorced, widowed}.

(c) A mother has two R2 pieces, five R1 pieces and four 50c pieces in her purse. Her daughter may take one coin from the purse. List all the possible coins that can be chosen.

What is the "event"? Answer: choosing a coin. The sample space of this event is: S = {R2 piece, R1 piece, 50c piece}.

Note that we should always ask ourselves: What is the event? We cannot stress enough that one has to understand what the problem is all about. The statistical consultant is confronted with problems from all spheres of life and it is not a show of "stupidity" if you ask questions until you understand the "story". Confront your learners from time to time with something that you know they are not supposed to have experience of and encourage them to ask questions until they understand it. We will, for example, use the Internet to research the background of a problem before we meet with clients.

List at least five random experiments in which the outcome is a single event. Clearly define the event and list the sample space of the event.

3.2.2 Two events

Two events can be described easily by means of a two-way table:

	List all the possible outcomes of the one event
List all the possible outcomes of the other event	**The sample space**

Example 3.2

(a) A die is tossed twice (or two dice are tossed). List all the possible outcomes.

		Outcome of one die (or first outcome)					
		1	2	3	4	5	6
outcome of the other die (or second outcome)	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

(b) Toss a die and a coin. List all the possible outcomes.

		outcome of the die					
		1	**2**	**3**	**4**	**5**	**6**
outcome of the coin	**Heads (H)**	(1,H)	(2,H)	(3,H)	(4,H)	(5,H)	(6,H)
	Tail (Ts)	(1,T)	(2,T)	(3,T)	(4,T)	(5,T)	(6,T)

NOTE:

1. Outcome (1,H) is the same as (H,1).
2. (1,H) is read as outcome “1” with the die **AND** “heads” with the coin.
3. We may use abbreviations for naming an outcome, for example ““H” for the outcome “heads”. Make sure that you show or define the abbreviations clearly. The reader should not have to guess what “H” stands for.

(c) A person is drawn randomly and his/her marital status (never married, married, widowed or divorced) is recorded. List all the possible outcomes.

		Marital status			
		Never married (NM)	**Married (M)**	**widowed (W)**	**Divorced (D)**
Gender	**Male (m)**	(m, NM)*	(m, M)	(m, W)	(m, D)
	female (f)	(f, NM)	(f, M)	(f, W)	(f, D)

* Remember to always read the cells as the one outcome **and** the other outcome, for example the randomly drawn person is "male" **and** "never married".

ACTIVITY 3.2

3.2.1 A learner can be male or female and right-handed or left-handed. In a table, list the sample space that will describe these attributes.

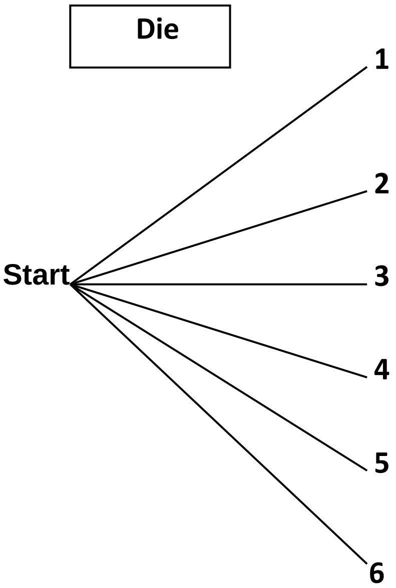
3.2.2 Is there a difference between age and preference for cool drink? A survey was conducted on the school grounds. The learners were aged from 10 to 14 and they had to choose between fruit juice, cold milk or gassy cool drink. Describe the sample space in a table.

3.2.3 Tree diagrams

An effective method for describing events, especially if there are more than two, is to represent them with lines. The resulting figure resembles a tree, hence the name "tree diagram". This graphical method for obtaining the outcomes of a random experiment is again best explained by means of examples.

Example 3.3

3.3.1 In example 3.2 (b) a die and a coin are tossed (or we could have said a coin and a die are tossed – the order is not important). Start at a point and show all the possible outcomes of the first event by means of lines.



Start at the end of each branch and repeat the procedure by showing all the possible outcomes of the second event, drawing line segments that "fan" out from the tip of the branch. **To reach all the possible outcomes, you simply "climb" EVERY branch from "start" until you reach the end (see figure 3.3.1).**

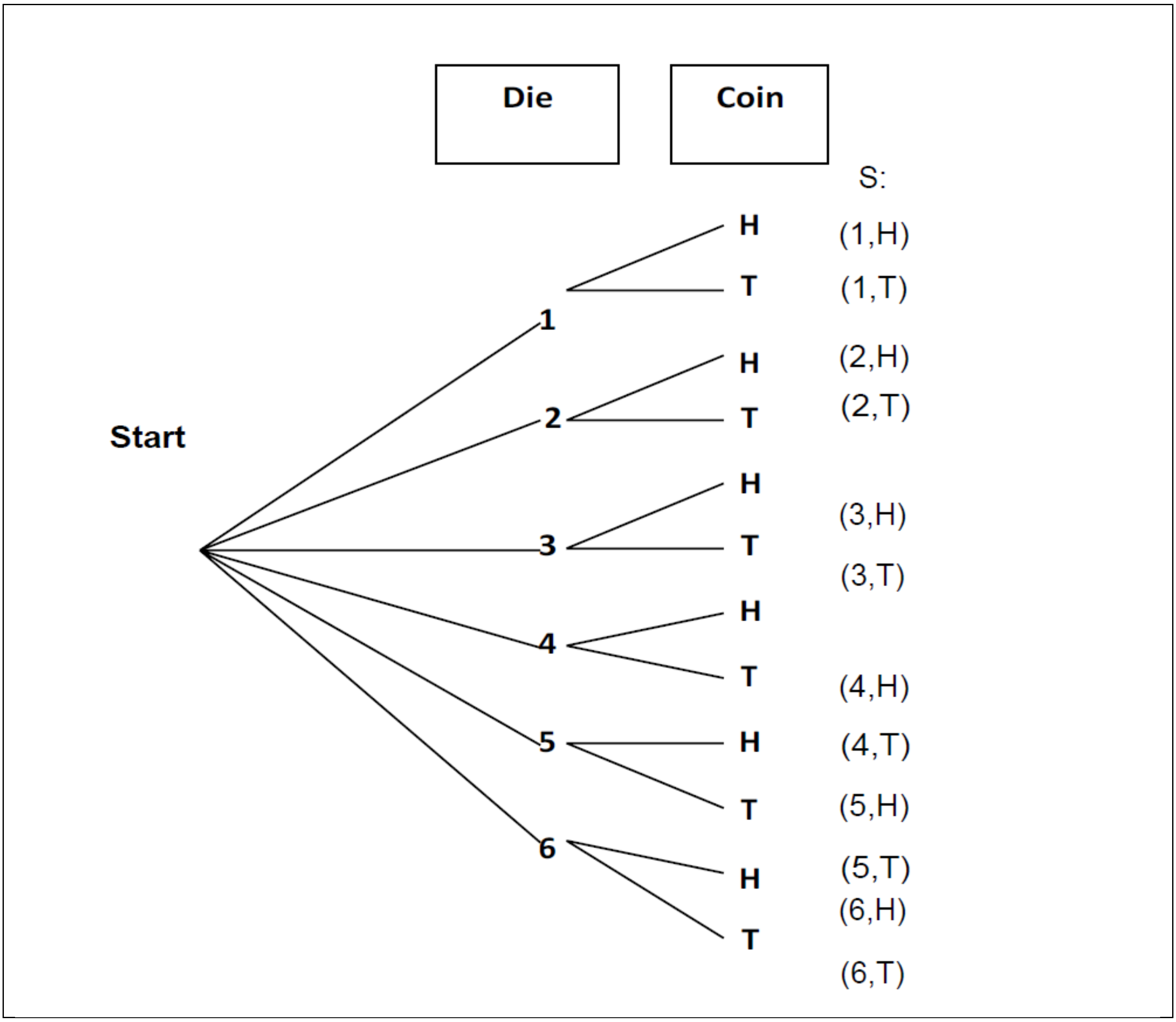


Figure 3.3.1

This method might seem far more cumbersome than a table, but you will see the reason for its use when we describe more than two events.

Suppose that a learner can be either male (M) or female (F); right-handed (RH) or left-handed; (LH) and has blue (Bl), brown (Br) or green (Gr) eyes. List the sample space that describes the learner in terms of these characteristics.

We think that you will agree that the "beauty" of tree diagrams is that one really does not have to think very hard. Let us start with the fun (figure 3.3.2).

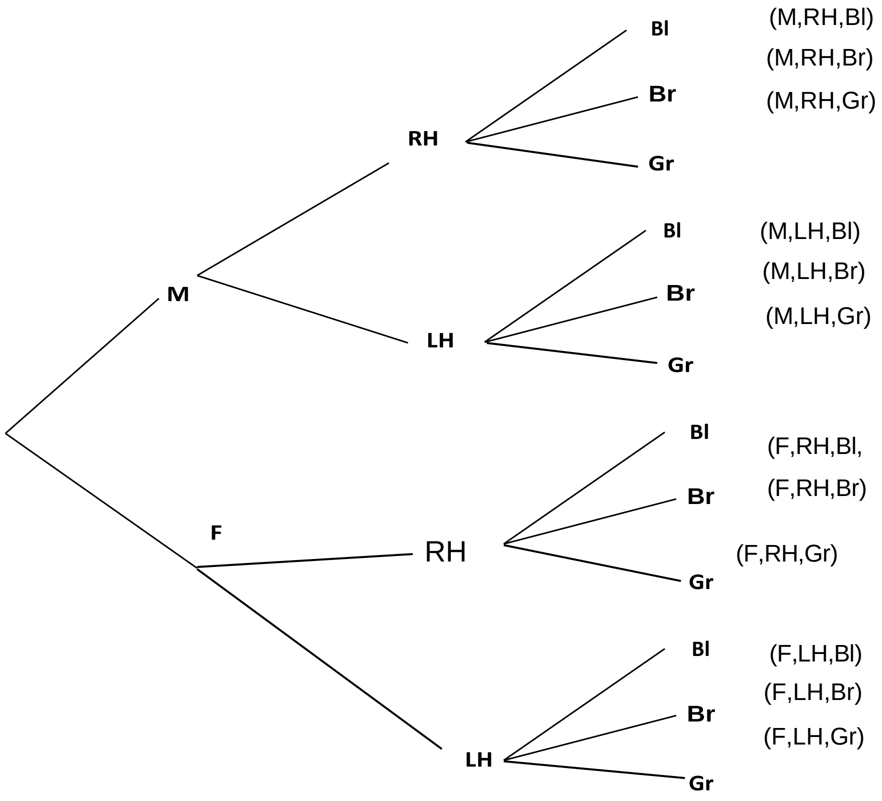


Figure 3.3.2

The following example illustrates a tree with branches that are not all the same length: Sue (S) and Tom (T) play in a tennis tournament. The first person to win two games out of three wins the tournament. Find all the possible outcomes.

Note that a branch ends the moment a person has won two games (see figure 3.3.3).

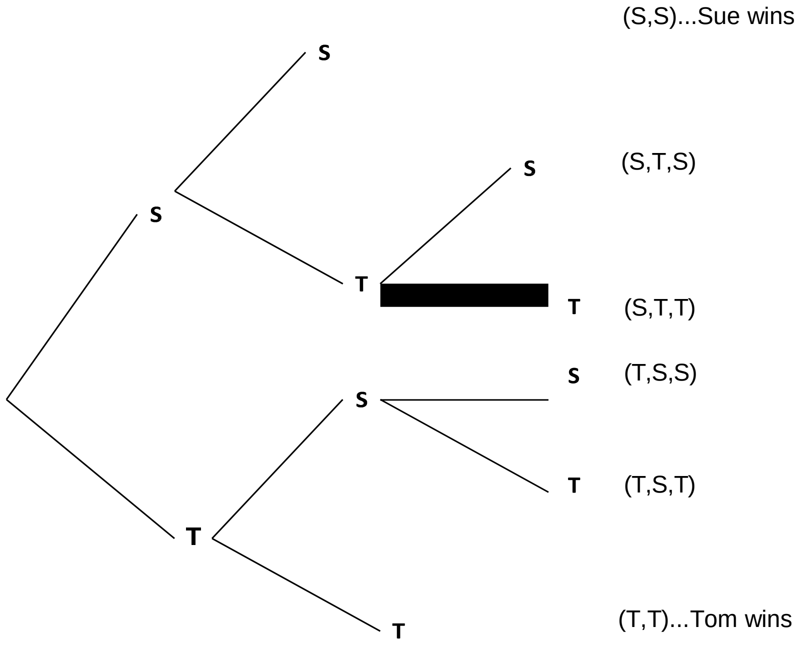


Figure 3.3.3

The use of a tree diagram can become quite cumbersome when we are looking at more than three events. The following example, taken from *Elementary statistics* by AG Blumen, illustrates this point.

A breakfast menu consists of the following items (one from each category):

Juice: Orange, grapefruit, cranberry

Toast: White, wholewheat

Eggs: Scrambled, fried, poached, hardboiled

Beverage: Coffee, tea

List all the possible breakfast choices (see figure 3.3.4).

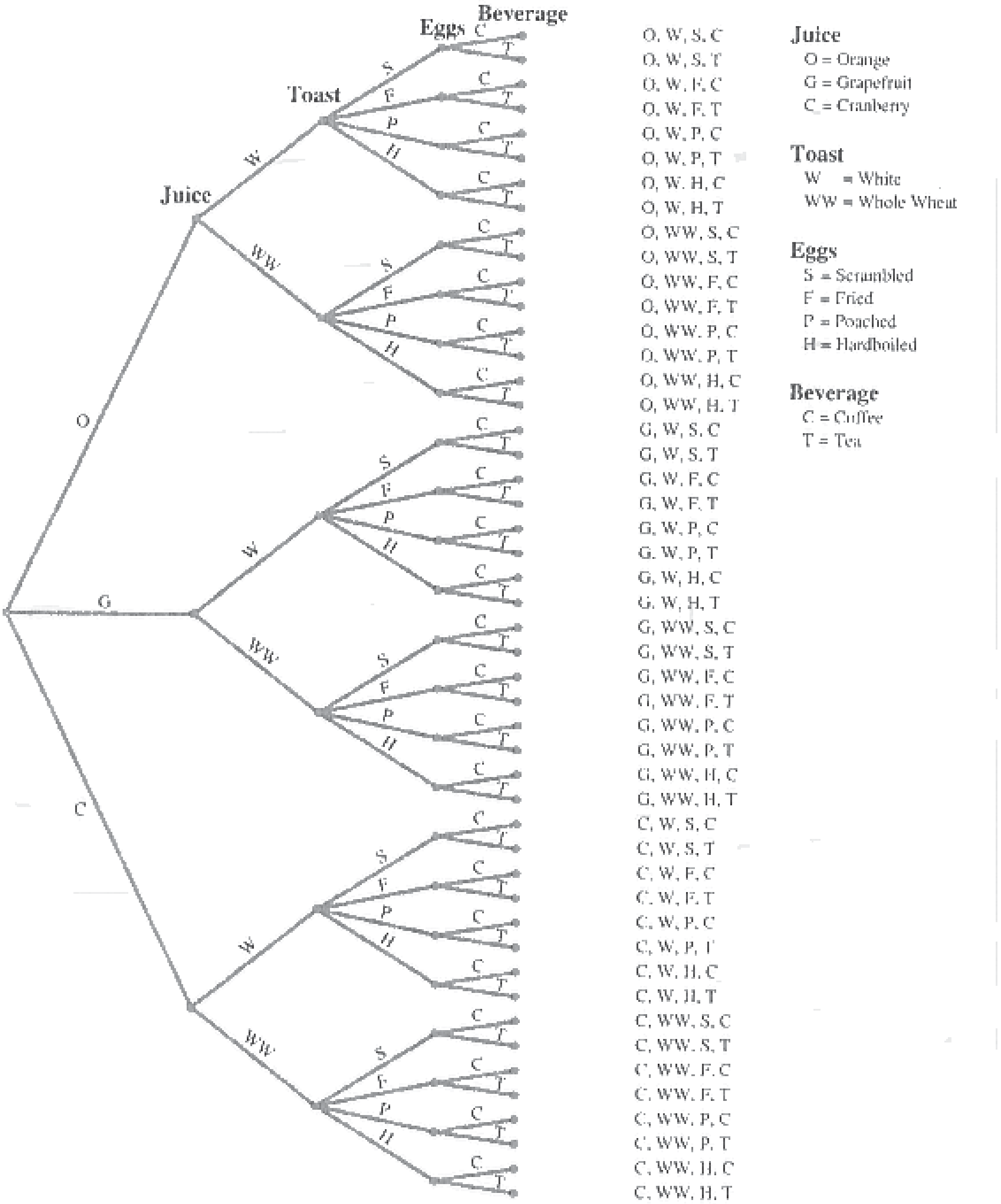


Figure 3.3.4

This figure might not be very clear, but it illustrates the point that it can become "messy".

ACTIVITY 3.3

3.3.1 Suppose that a salesman can travel from Cape Town to Johannesburg by plane or train; from Johannesburg to Durban by plane or car; and from Durban to Cape Town by ship, plane or car. List the sample space of all the possible ways that the salesman can travel from Cape Town to Johannesburg to Durban and back to Cape Town again.

3.3.2 Use a tree diagram to find all the possible outcomes for the sexes of the children in a family that has three children.

The picture below is the view from one of our studies and it suddenly struck us why the above diagrams are called "tree-diagrams". Do you see the similarity?

So far, we have looked at methods to describe all the possible outcomes of a random experiment. Each possible outcome (or groupings of outcomes) can be assigned a numerical value (probability) from which we can deduce the "chance" that that particular outcome (or groupings of outcomes) will occur.

3.3 PROBABILITY: THREE APPROACHES TO ASSIGNING PROBABILITY

3.3.1 The relative frequency approach

Most people are familiar with tossing a coin to decide, for example, which player should be the first to serve in a tennis match. They know intuitively that both players have an equal chance of serving first. In the first part of this study unit, where we looked at randomness, we saw that the ratio of the number of times that "heads" appeared to the total number of times that the coin was tossed is not exactly 0,5; however, because the tossing of the coin was repeated a large number of times, the relative frequency of the outcome "heads" came closer and closer to 0,5.

The relative frequency approach defines probability as the long-run relative frequency with which an outcome (or set of outcomes) occurs. Mathematically speaking, we will say that "in the limit" the relative frequency approaches the probability. **This is an empirical or experimental approach to probability.** We can never know the exact probability in this way, but we can get a pretty good estimate of it.

Event

Before we continue, let as define an "event". **An event is an outcome or a set of outcomes of a random experiment** (that is, an event is a subset of the sample space). For example: when a die is tossed, the outcome "even number" is the set {2, 4, 6}. Usually we allocate a symbol to events, for example E = {2, 4, 6}.

In terms of an event A, we can define the relative frequency as:

$$\frac{\text{number of times event A occured}}{\text{number of times experiment was done}}$$

And this is an estimate of the probability that this event will occur in the long run.

What properties of probabilities can we deduce from this approach to probability?

- Since any proportion is a number between 0 and 1, any probability is also a number between 0 and 1: an event with probability of 1 occurs in every trial (for example the event of letting a ball drop to the ground or the sun rising), an event with a probability of 0 never occurs (like the sun not rising in the morning) and an event with a probability of 0,5 occurs in half the trials in the long run.

- All the possible outcomes together must have a probability of 1. From this, we can say that the probability that an event **does not** occur is 1 minus the probability that the event **does** occur. If, for example, when we flip a coin 1 000 times, we have 480 times the outcome "heads" with a relative frequency of 480/1000, then we have 1 – 480/1000 = 520/1000 as the relative frequency of "not heads". The probability of an event occurring and the probability of it not occurring always adds up to 1. This is nothing new. When the weatherman predicts a 40% chance of rain, you will most properly leave your umbrella at home because there is a far greater chance (that is a 60% chance) that it will not rain.

NOTE: Probabilities can be expressed as fractions, decimals or percentages. If one asks "What is the probability of getting a head when a coin is tossed?" typical responses can be any of the following three: "one half", "point five" or "fifty percent". These answers are all equivalent.

A major source of quoted probabilities for events is data on the relative frequencies of the same events in the past. If, for example, in recent years roughly 600 people from a population of about 3 000 000 were killed on the roads, most people will be fairly comfortable with a statement that takes the following form: "The probability that a randomly selected person will die on the roads next year is about 600/3000000." One should bear in mind that statements like these are only valid if the underlying process is stable over time, for example the speed limit stays the same.

Because this is an experimental approach to probability, it is an ideal topic for a lesson where the learners have to simulate a situation and "discover" that "in-the-long-run" the relative frequency of an event occurring approaches a constant value.

The following lesson demonstrates how a real-life situation can be simulated.

LESSON OBJECTIVE	Learners will discover how to use simulations for estimating probabilities
Problem	Is it "fair" to flip a coin to see who will serve first in a tennis match?
Experimenting	In small groups of two (one learner will flip the coin and the other will note the outcome), flip a coin 50 times and note every outcome. Add the number of outcomes "head". Calculate the relative frequency of the outcome "head". See whether your pocket calculator cannot simulate the flipping of a coin. For example on the Sharp EL-531: To simulate a coin flip, 0 (head) or 1(tail) can be randomly generated by pressing "2ndF" and "Random" and "2" and "ENT". To generate the next random coin number, press "ENT".
Reflecting/ Discussion	Compare the different groups' outcomes. How close is it to 0,5? What happens when you combine the results of two (or more) groups?

Conclusion/ hypothesising

- ♦ In the long run, the relative frequencies approach 0,5 and we refer to this as the “probability of the outcome ‘head’ when a coin is flipped”.
- ♦ In other words, we can never observe a probability exactly.
- ♦ Mathematical probability is an idealisation that is based on imagining what will happen in an indefinitely long series of trials.
- ♦ THIS IS THE BEAUTY OF STATISTICS! Although a phenomenon can be random and individual outcomes are uncertain, there is nevertheless a regular distribution of outcomes in a large number of repetitions and this pattern can be described/modelled by means of mathematics.

To test whether the learners understood what they were investigating, ask the following question: You will all agree that we can say that “the probability to get the outcome ‘head’ when we flip a coin is 0,5”. Does this mean that I will definitely get five heads if I flip a coin ten times (because 0,5 x 10 = 5)? (For your sake, we sincerely hope that they do not answer “yes”!)

ACTIVITY 3.4

3.4.1 The probability of getting the outcome “heads” when a coin is flipped is 0,5. The educator asks the class: “Do you think that we will definitely get five heads if we flip a coin ten times (because 0,5 x 10 = 5)?” What would you expect the learners to answer?

3.4.2 There is a 30% chance of rain tomorrow. What exactly does this mean?

3.4.3 Ten patients have just been given medication because they are HIV positive. They have been told that there is a 40% chance that they will be cured if they follow up the treatment by having a series of prescribed injections. Will four of the ten patients (10 x 0.4 = 4) be cured if they get the full treatment?

3.4.4 The weather forecast in various cities predicts the following for tomorrow:

- ♦ 0,3 probability of rain in Pretoria
- ♦ 0,7 probability of rain in Cape Town
- ♦ 0,5 probability of rain in Durban

In which of these cities will it most likely rain tomorrow?

3.4.5 What is the probability that a family with three children has two female children and one male child? Discuss how you will use three coins in a box lid and flipping to model the problem (or use the random option on a pocket calculator). Execute your plan and write a short report on your findings.

NOTE: It is better to talk about the **expected** frequency (an estimate or an educated guess) when we calculate the product of the probability that an event will occur and the number of repetitions of the event.

At this stage, you should be well aware that to assign correct probabilities to individual outcomes often requires long observation of the random phenomenon. Is there an alternative method? If we are willing to assume that individual outcomes are equally likely, we can use the following "classical" approach to probability. You may well ask when events will not be "equally likely"? Let us think about card games. The probability for card games depends critically on the cards being "well shuffled" so that their order is random. From experience, we know that children and beginners cannot shuffle cards well and some of the cards do not get mixed up. Experts are so clever at shuffling that it is hard for a layperson to know whether they are shuffling for randomness or for their own advantage. Some dice might be "loaded" in order for a "six" to appear more often than the other outcomes. These are instances of cheating and we cannot use our classical approach to predict outcomes for these.

3.3.2 The classical approach

Classically probability uses sample spaces to determine the numerical probability that an event will happen. One does not actually have to perform an experiment to determine this probability. Classical probability is called by this name because it was the first type of probability that was studied formally by mathematicians in the seventeenth and eighteenth centuries.

Probabilities can be calculated with the following formula:

The probability of any event A is P(A)

$$= \frac{\text{number of ways A can occur}}{\text{total number of outcomes in the sample space}}$$

We immediately realise that the "tricky" part of calculating probabilities is to be able to get the "total number of outcomes in the sample space". This is why we studied the section on describing events prior to this section.

Let us start by looking at a **single event.**

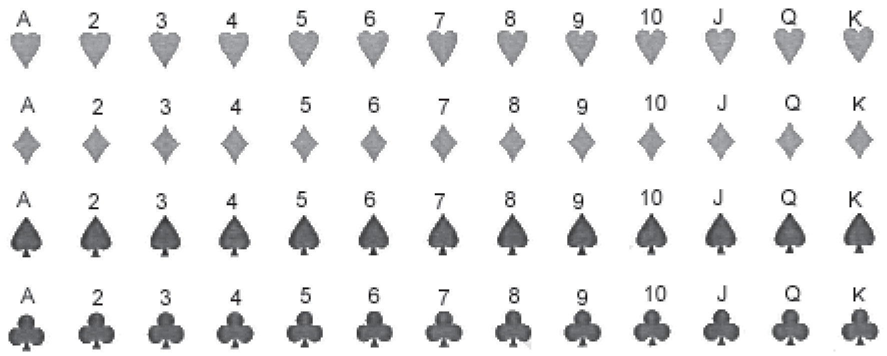
Example 3.4

3.4.1 The random experiment

Roll a die: S = {1, 2, 3, 4, 5, 6}

Event	Possible outcomes of the event	P(the event)
1	{1}	1/6
Even number	{2, 4, 6}	3/6
Number smaller than 5	{1, 2, 3, 4}	4/6

3.4.2 The random experiment is to draw a card from an ordinary deck of cards. If you are not sure how familiar your learners are with cards, it might be a good idea to either display all the cards on your desk or to draw the following figure on a transparency. Note that jokers are not part of the "possible outcomes". The first two rows (hearts and diamonds) are red and the other two rows (spades and clubs) are black.



A ♥	2 ♥	3 ♥	4 ♥	5 ♥	6 ♥	7 ♥	8 ♥	9 ♥	10 ♥	J ♥	Q ♥	K ♥
A ♦	2 ♦	3 ♦	4 ♦	5 ♦	6 ♦	7 ♦	8 ♦	9 ♦	10 ♦	J ♦	Q ♦	K ♦
A ♠	2 ♠	3 ♠	4 ♠	5 ♠	6 ♠	7 ♠	8 ♠	9 ♠	10 ♠	J ♠	Q ♠	K ♠
A ♣	2 ♣	3 ♣	4 ♣	5 ♣	6 ♣	7 ♣	8 ♣	9 ♣	10 ♣	J ♣	Q ♣	K ♣

If you use this example, ask as many learners as possible for probabilities so that they can gain confidence in calculating probabilities. For example:

P(King) = 4/52 P(6) = 4/52 P(Red King) = 2/52

P(3 of hearts) = 1/52 et cetera

3.4.3 The following example illustrates the important concepts of "less than", "more than", "at least" and "at most".

Hospital records indicate that maternity patients stay in the hospital for the number of days that are shown in the distribution:

Number of days	3	4	5	6	7	Total
Number of patients	15	32	56	19	5	127

If a maternity patient is randomly chosen, what is the probability that she stayed for the following days?

(a) exactly 5 days (b) at most 4 days (c) at least 5 days

(d) less than 6 days (e) more than 4 days

(a) P(5) = 56/127

(b) P(at most 4 days means 3 or 4 days) = (15 +32)/127 = 47/127

(c) P(at least 5 days means 5, 6 or 7 days) = (56 + 19 +5)/127 = 80/127

(d) P(less than 6 days means 3, 4 or 5 days) = (15 +32 + 56)/127 = 103/127

(e) P(more than 4 days mean 5, 6 or 7 days) = (56 + 19 +5)/127 = 80/127

This illustrates an important rule of probability theory. Let us take a closer look at (c):

the P(at least 5 days) = P(5 OR 6 OR 7 days)

= (56 + 19 + 5)/127

= 56/127 + 19/127 + 5/127

= P(5) + P(6) + P(7)

This is called the addition rule of mutually exclusive events.

Mutually exclusive

Two (or more) events are **mutually exclusive** (also called disjoint) if they have no outcomes in common and **can never occur at the same time.**

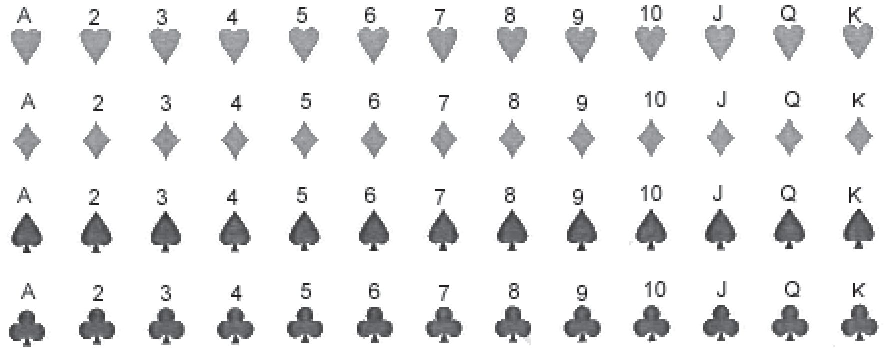
The addition rule of mutually exclusive events

The addition rule of mutually exclusive events:

Assume events A and B are mutually exclusive, then

P(A **or** B) = P(A) + P(B)

What if the two events are not mutually exclusive? Let us take a look at our pack of cards:



P(red card) = 26/52 P(jack) = 4/52

P(red OR jack) = 28/52 # 26/52 + 4/52 # P(red) + P(jack)

Note that the event "red card" and "jack" are **NOT** mutually exclusive because we have two cards that can be red AND a jack. If you look carefully, you will see that when you add the events "red" and "jack", the jack of hearts and the jack of diamonds are counted **twice.**

The addition rule

The addition rule for two events that are not mutually exclusive:

If A and B are not mutually exclusive,

then P(A or B) = P(A) + P(B) – p(A and B)

ACTIVITY 3.5

3.5.1 If an event is certain to occur, what is the probability?

3.5.2 If an event cannot happen, what value is assigned to its probability?

3.5.3 If the probability that it will rain tomorrow is 0,8, what is the probability that it will not rain tomorrow?

3.5.4 If 25 tickets are sold for a lottery and one person buys three tickets, what is the probability of that person winning a prize?

3.5.5 A box contains ten red, 15 yellow and six blue Smarties. If one of the Smarties is selected at random, what is the probability that it is (a) red, (b) blue and (c) red or blue?

Example 3.5

Suppose a learner can be either male (M) or female (F); right-handed (RH) or left-handed (LH); and has blue (Bl), brown (Br) or green (Gr) eyes. What is the probability that a randomly chosen learner is a right-handed boy with brown eyes?

P(male and right-handed and brown eyes) = P(M, RH, Br) = 1/12 (from the tree diagram in figure 3.3.2 and the definition of probability)

Now let us have a look at P(male) x P(right-handed) x P(brown eyes)

= 1/2 x 1/2 x 1/3 and that is also 1/12

So, P(M AND RH AND Br) = P(M) x P(RH) x P(Br)

This is called the **multiplication rule** and the events "male", "right-handed" and "brown eyes" are said to be **independent events.**

Independent events

Two events (A and B) are **independent** if knowing that one occurs does not change the probability that the other occurs. If A and B are independent, P(A and B) = P(A) x P(B).

When the outcome or occurrence of the first event affects the outcome or occurrence of the second event in such a way that the probability is changed, the events are said to be dependent. Examples of dependent events are: selecting a ball from an urn, not replacing it and then selecting a second ball; being a lifeguard and getting a suntan; parking in a no-parking zone and getting a parking ticket; having high grades and getting a scholarship.

Multiplication rule

This is the **multiplication rule** for independent events.

What if the events are **not** independent? Before answering this question, we have to know a little about conditional probabilities.

In a questionnaire the question that was asked was: "Do you approve of abortion when the child is not wanted?"

A sample of 200 people took part in the survey and the following table summarises the results (this is not actual data):

Summary of males and females' responses to abortion

	Yes	No	Don't know	Total
Males	20	40	20	**80**
Females	60	30	30	**120**
Total	**80**	**70**	**50**	**200**

If we choose a questionnaire at random, what is the probability that a male answered it? This is simply the proportion of males, that is P(male) = 80/200 = 0,4.

If we choose a questionnaire at random, what is the probability that a male answered it **and** said "yes" to abortion? Because we know that 20 males said "yes", we have:

P(male AND "yes") = 20/200 = 0,1

Let us assume that we sorted the questionnaire into two piles: the one pile is all the questionnaires that were answered by males and the other pile is all the questionnaires that were answered by females. If we randomly draw a questionnaire from the "male" pile, what is the probability that the male said "yes" to the question? In this pile we have only 80 questionnaires and we know that of the 80 men, 20 said "yes" to the question – the answer is 20/80 = 0,25. This is the probability of getting a "yes" answer **given** the information that a male answered the questionnaire. We write this as follows:

P("yes" | male) = 20/80

This is called a **conditional probability**, that is it gives the probability of one event (the answer to the question was "yes", under the condition that we know another event [a male answered the questionnaire]). You can read the bar | as "given the information that".

If we have a look at P(male) = 80/200 = 0,4 , P(male AND "yes") = 20/200 = 0,1 and P("yes" | male) = 20/80, there is a relationship among these three:

P(male AND "yes") = 20/200 =(80/200) x (20/80) = P(male) x P("yes" | male)

Multiplication rule

This relationship is the fundamental **multiplication rule of probability:**

The probability that both of two events (A and B) will happen together can be found by:

P(A and B) = P(A)P(B | A)

where P(B | A) is the conditional probability that B occurs given the information that A occurs (In other words, this rule says that for both of two events to occur, first one has to occur and then – given that the first event has occurred – the second has to occur.)

Conditional probability

From this expression follows the definition of **conditional probability**:

When P(A)>0, the conditional probability of B given A is:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

If you are starting to think "now it is getting complicated!" we want you to relax. Let us go back to our table and define all the probabilities that we can calculate from the data in it:

Summary of males and females' responses to abortion

	Yes	**No**	**Don't know**	**Total**
Males	20	40	20	**80**
Females	60	30	30	**120**
Total	**80**	**70**	**50**	**200**

(a) If we choose a questionnaire at random, what is the probability that it is a "yes" answer. We are given no information about the gender and find the probabilities from the "total" row at the bottom of the table:

Yes	**No**	**Don't know**	**Total**
80	**70**	**50**	**200**

P("yes") = 80/200

(b) If we are told that a woman answered the questionnaire, we look only at the "female" row:

(c)

	Yes	No	Don't know	Total
Females	60	30	30	**120**

(d) Thus the conditional probability that an answer was "yes", given the information that a woman answered, is simply 60/120.

And we should not forget that all the cells in the table read "the one event and the other event".

The probability, for example, of having a questionnaire that was answered by a female and her answer is "Don't know" is 30/200.

	Yes	No	Don't know	Total
Males	20	40	20	**80**
Females	60	30	30	**120**
Total	**80**	**70**	**50**	**200**

Previously, we used tree diagrams to get all the possible outcomes of a random experiment. In these cases the different "stages" were independent. For example, gender was independent of colour of eyes or being right-handed or left-handed (example 3.5). Consider the following random experiment.

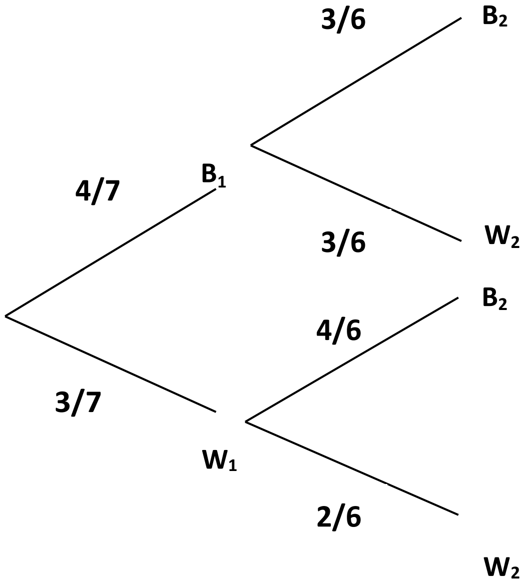
Suppose that we draw two balls at random, one at a time **without replacement,** from an urn that contains four black balls and three white balls. When we make the second draw, the chances of drawing a black ball (for example) depend on the colour of the ball that was removed in the first draw. Because the first ball was set aside, the composition of the urn changed. The colour of the second ball will be conditional on the colour of the first ball that was drawn. The outcomes can again be represented with a tree diagram, but let us add the probability that the event will occur beside each line segment.

Let B_1 = Black ball with the first draw $P(B_1) = 4/7$

Let W_1 = White ball with the first draw $P(W_1) = 3/7$

Let B_2 = Black ball with the second draw

W2 = White ball with the second draw



What is the probability that both balls will be black? We "climb" the branch that will result in two black balls (the top one) and as we go along, we multiply the probabilities: 4/7 x 3/6.

In terms of the conditional probabilities:

$P(B_1 \text{ and } B_2) = P(B_1) \text{x} P(B_2 \mid B_2) =$

4/7x3/6

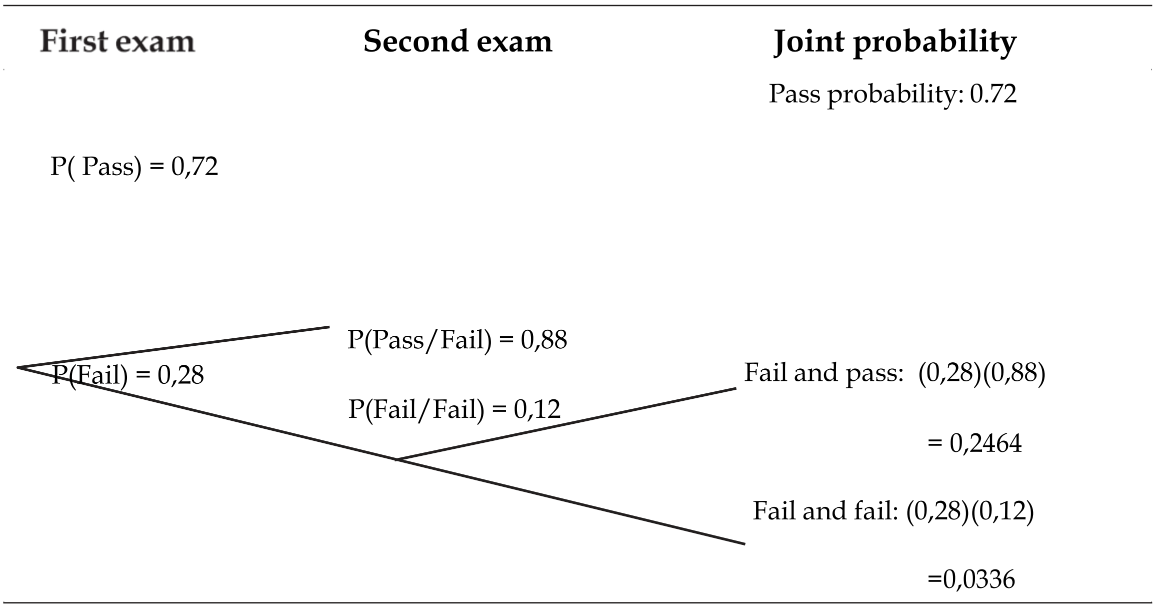
What is the probability of getting a black ball in the second draw? Now we have to "climb" all the branches that will result in a black ball in the second draw. This is the first and the third branches. Because the two events are mutually exclusive, we can add the two products:

$$P(B_2) = 4/7 \text{ x } 3/6 + 3/7 \text{ x } 4/6 = 4/7$$

In a previous example we looked at events that are not equally likely to occur and saw that it is still convenient to use tree diagrams and contingency tables to list probabilities.

Another example: Students who graduate from law school still have to pass a bar exam before they can become lawyers. Suppose that in a particular jurisdiction the pass rate for first-time test takers is 72%. Candidates who fail the first exam may take it again several months later. Of those who failed their first exam, 88% pass their second attempt. Find the probability that a randomly selected law school graduate will become a lawyer. Assume that candidates cannot take the exam more than twice.

This might seem like a complicated problem; however, when we approach it systematically, it becomes quite easy.



It is important to notice that the second set of "branches" represents **conditional probabilities.** It makes sense, doesn't it? Students will be allowed to write the second exam only if they failed the first one. To calculate the probability of the combined event, you just climb the tree and multiply the probabilities of the branches as you go along.

Finally, we have to apply the addition rule for mutually exclusive events to find the probabilities of passing the first or second exam.

P(that a randomly selected law school graduate becomes a lawyer)

= 0,72 + 0,2464 = 0,9664

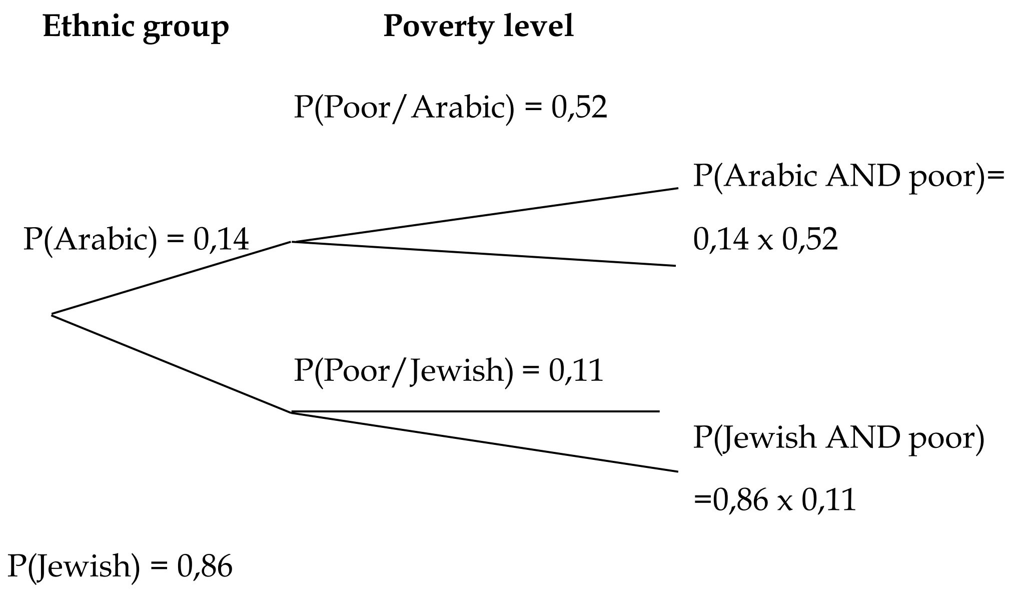
= P(pass first exam) + P(fail first exam AND pass second exam)

Thus 96,64% of the applicants becomes lawyers by passing the first or second exam.

To demonstrate the relationship between a tree diagram and a contingency table, let us consider the following example:

In 1992 a news report stated that 11% of Israel's Jewish population and 52% of its Arabic citizens lived below the poverty line. Arabic citizens were reported to make up 14% of the population of Israel. We shall assume that these two groups account for the whole population of Israel.

First, construct a probability tree diagram. Two factors are at work here: (1) ethnic group and (2) poverty level (where we define "living below the poverty line" as "poor").



The same information can be represented in a two-way table where the two factors, ethnicity and poverty, become the two dimensions of the table:

		Ethnicity		
		Arabic (A)	**Jewish (J)**	Total
Poverty	**Poor (P)**	P(P and A) **=P(P/A)P(A)** =0,52x0,14 =0,0728	P(P and J) **=P(P/J)P(J)** =0,11x0,86 =0,0946	= 0,728+0,0946 = 0,8226
	Not poor (N)	=0,14 – 0,0728	= 0,86 – 0,0946	=1 – 0,8226
	Total	**0,14 (given)**	0,86 (given)	**1**

We think that you will agree that we can answer more probability questions from the table than from the tree diagram; however, in most cases, it is much easier to construct a tree diagram.

ACTIVITY 3.6

3.6.1 A random sample of 200 people from rural and urban communities was asked whether or not they would like the death penalty to be reinstated. This was the outcome:

	Favours death penalty	**No comment**	**Does not favour death penalty**	**Total**
Rural	50	15	35	100
Urban	40	10	50	100
Total	90	25	85	200

For a randomly selected person, calculate the probability that

(a) the person favours the death penalty

(b) the person favours the death penalty **and** is from a rural community

(c) the person favours the death penalty **given** that he/she is from a rural community

If we are looking only at people from urban communities, what is the probability that the randomly selected person is not in favour of the death penalty?

3.6.2 From a group of 40 learners, 20 are above average in Mathematics and 20 are above average in Music, while 15 are above average in both Mathematics and Music.

The following table summarises the data:

		RATING IN MUSIC		
		Below average	Above average	Total
RATING IN MATHEMATICS	Below average	15	5	**20**
	Above average	5	15	**20**
Total		**20**	**20**	**40**

A learner is selected at random from these learners.

(a) What is the probability that the learner has an above average rating in Music?

(b) What is the probability that the learner will have an above average rating in Mathematics **given** that he/she has an above average rating in Music?

(c) What is the probability that the learner will have an above average rating in both Music **and** Mathematics?

(d) Are the events (the selected learner is above average in Mathematics and the selected learner is above average in Music) independent? Give reasons for your answer.

3.6.3 Suppose that you have a sample of couples who have children and in 2% both the father and the mother are left-handed, in 20% one parent is left-handed, and in the rest neither parent is left-handed. The chances of a child being left-handed are 1 in 2 if both parents are left-handed, 1 in 6 if one parent is left-handed, and 1 in 16 if neither parent is left-handed. What is the probability that neither parent of a left-handed child is left-handed?

(a) Complete the tree diagram by adding the probabilities to the different branches.

(b) What is the probability of a randomly chosen child being left-handed? Hint: The different events are indicated with an arrow. Remember that these events are mutually exclusive and we can apply the addition rule for mutually exclusive events.

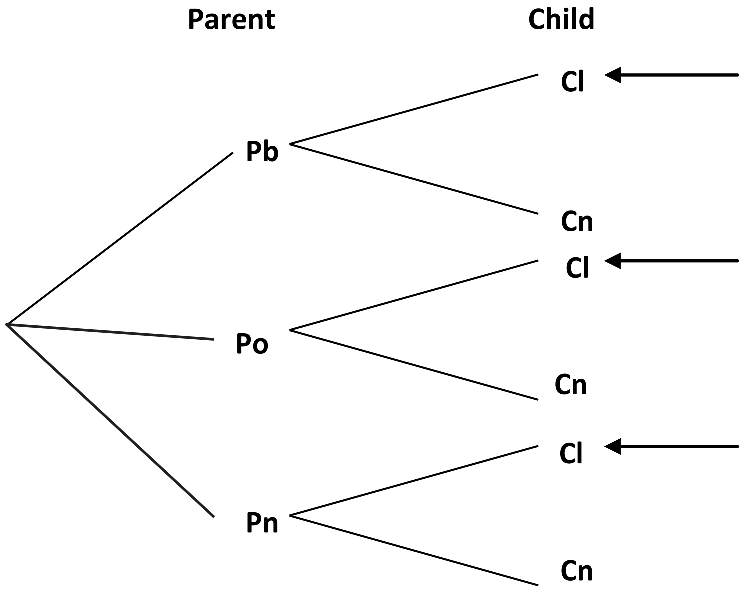
Key to symbols:

Pb: Parents both left-handed

Po: One parent left-handed

Cl: Child left-handed

Cn: Child not left-handed



Now we have looked at the calculation of probabilities from data, namely the relative frequencies of the occurrence of an event. We determined probabilities from models and we realised that these probabilities will only be approximately true for the real experiment. However, in the media probabilities are often quoted that are backed up neither by data nor probability models. These are subjective probabilities that often masquerade as frequency probabilities.

3.3.3 The subjective approach

Quite often no data will be available and the opinions of experts will be used to subjectively predict the probability that an event will occur. For example, consider a situation where a new product is being introduced. In this case, no historical data for the product is available. Expert opinion, which can be supplied by members of the sales force and the market research team, can be used to estimate a probability that the product sales will be high in the first month after the launch of the new product.

We think you will agree with us that people who quote probabilities in the media should report how they obtained it. The following is an example of a "relative frequency probability" that was actually a subjective probability.

In 1977 a Pan Am jumbo jet and a KLM jumbo jet collided on an airport runway in the Canary Islands. One jet was taxiing after landing, while the other was taking off. BBC news reported: "At least 560 people died when two jumbo jets collided on a runway in what is thought to be the world's worst disaster involving aircraft on the ground."

Charred remains of airliner in Tenerife crash. Most passengers burned to death: a lucky few survived.

Soon after, Terry Speed (a well-known Australian statistician) noticed the following report in *The West Australian:*

> NEW YORK, Mon: Mr. Webster Todd, Chairman of the American National Transportation Safety Board, said today that the chances of two jumbo jets colliding on the ground were about 6 million to one... – AAP

Professor Speed, who had strong research interests in probability, was intrigued by this statement and wondered how the board had calculated their figure. Speed wrote to the chairperson. The reply stated that the figure (6 000 000:1) had no statistical validity nor was it intended to be a rigorous probability statement. The statement was made to emphasise the intuitive feeling that such an occurrence indeed has a very remote, but not impossible, chance of happening. At best, the quoted probability was a subjective assessment; at worst, it was a negligible small number that was plucked out of thin air to reassure the public.

ACTIVITY 3.7

Search for at least three newspaper or magazine articles in which probabilities is quoted and give your opinion on what approach was used to reach the probability.

By now, you should have a firm background knowledge of probabilities. The next step, before we look at statistical inference, is to take a look at probability distributions.

Answers to some activities

Activity 3.2

3.2.1

		Dominance	
		Right-handed (RH)	**Left-handed(LH)**
Gender	**Male (M)**	(M, RH)	(M, LH)
	female (F)	(F, RH)	(F, LH)

3.2.2

		AGE				
		10	**11**	**12**	**13**	**14**
DRINK PREFERENCE	**FRUIT JUICE (FJ)**	(FJ, 10)	(FJ, 11)	(FJ, 12)	(FJ, 13)	(FJ, 14)
	COLD MILK (CM)	(CM, 10)	(CM, 11)	(CM, 12)	(CM, 13)	(CM, 14)
	GASSY Cool DRINK (GCD)	(GCD, 10)	(GCD, 11)	(GCD, 12)	(GCD, 13)	(GCD, 14)

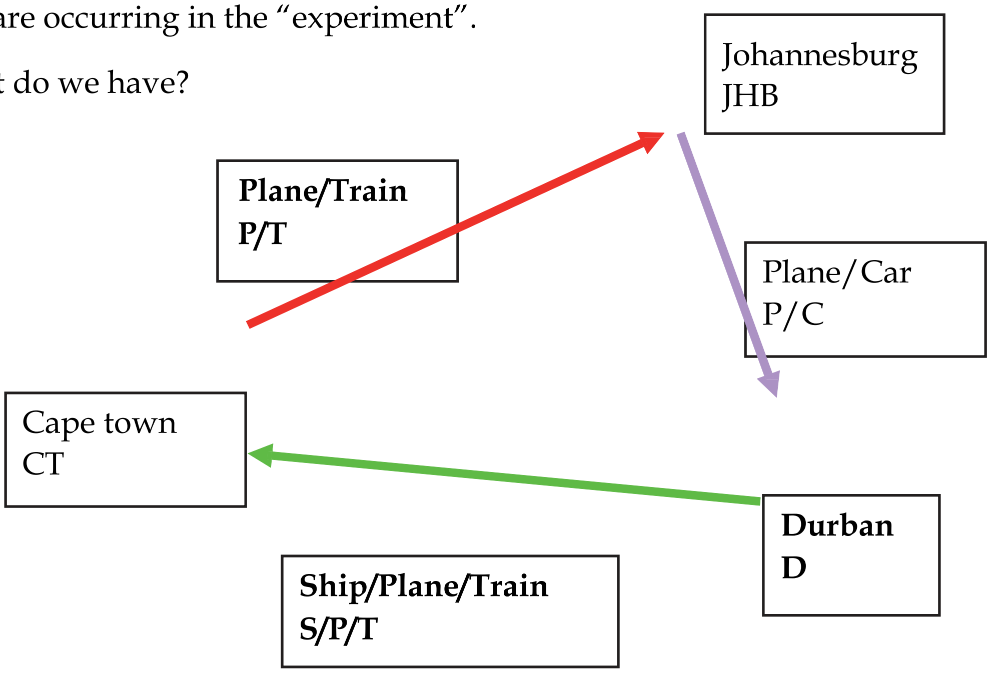
Activity 3.3

3.3.1

It is obvious that you still have problems with summarising all the outcomes of a random experiment by using a tree diagram.

First of all, make sure that you **CLEARLY** understand the different events that are occurring in the "experiment".

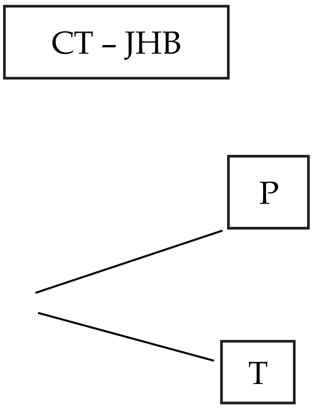
What do we have?



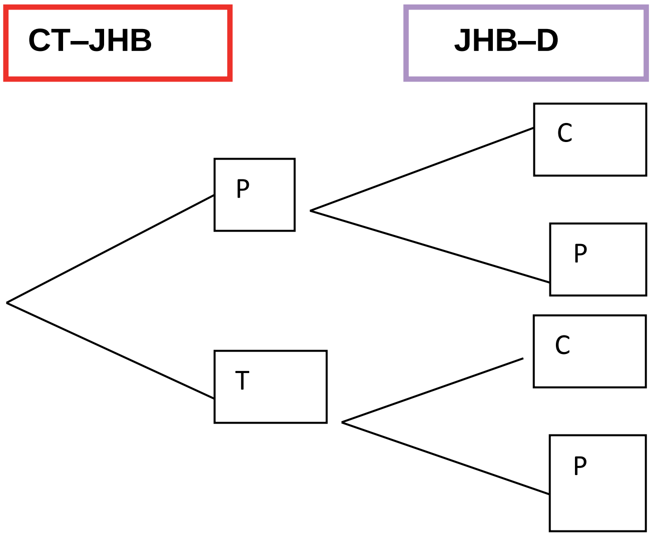
So there are THREE events:

CT–JHB JHB–D D–CT

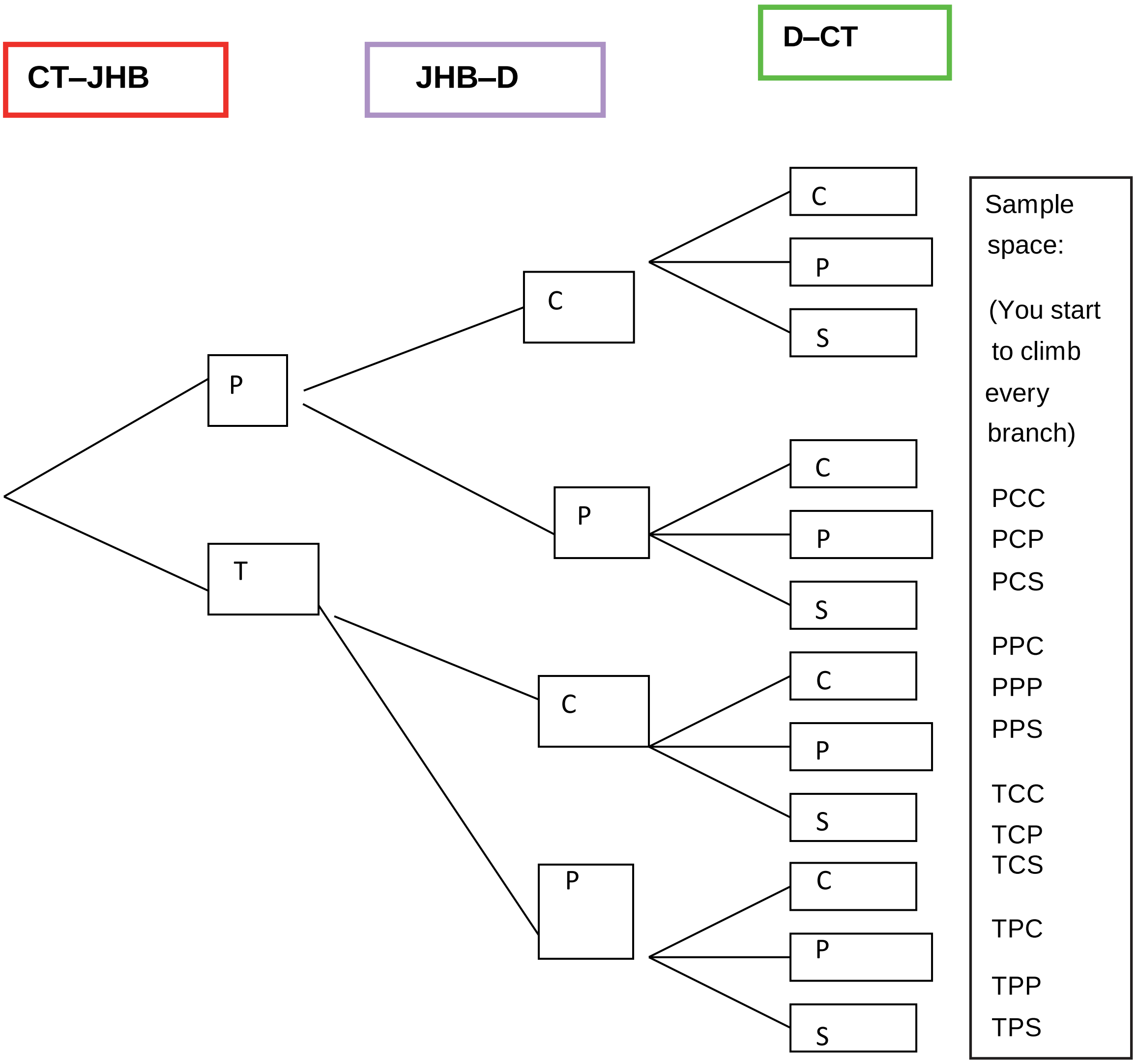
You should start by describing all the possible outcomes of the first event:



Then you should begin at the end of each branch and describe the outcomes of the second event:



And then, you should describe the third event:



Sample space:

(You start to climb every branch)

PCC
PCP
PCS

PPC
PPP
PPS

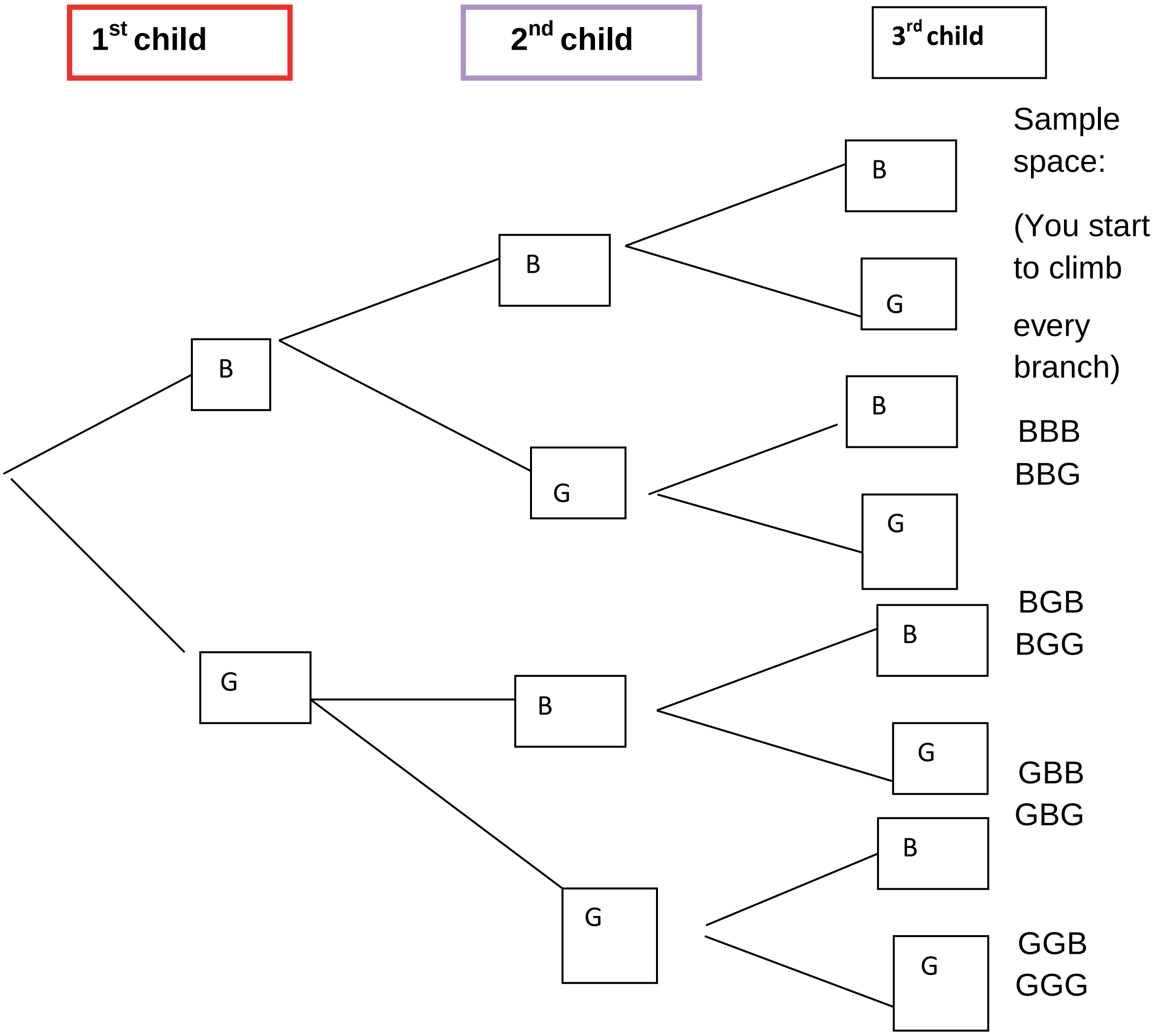
TCC
TCP
TCS

TPC

TPP
TPS

3.3.2

Let B: Boy and G: Girl



Activity 3.4

3.4.2

Weather forecasters know that with a certain weather pattern, they can expect **in the long run** that in 30 out of 100 such patterns it will rain.

3.4.3

Again, the 40% means that in the long run, it can be expected that 40 out of 100 of the people who follow the treatment will be cured.

One can therefore definitely NOT say precisely that four out of ten people will be cured. It gives an indication of how small/big your chance is.

Activity 3.5

The values of a probability can vary from 0 to 1: a probability of 0 means that it is impossible that the event will occur (for an example that an apple will fall "up") and a probability of 1 indicates that the event is certain to happen (for example the sun will rise tomorrow).

3.5.1 P(certain event) = 1

3.5.2 P(event that CANNOT happen) = 0

3.5.3 P(will not rain) = 1 – P(rain) =1 - 0,8 = 0.2

3.5.4 P(winning) = $\dfrac{\text{number of tickets bought}}{\text{number of tickets sold}} = \dfrac{3}{25}$

3.5.5 (a) P (Red) = $\dfrac{\text{number of red smarties}}{\text{total number of smarties}} = \dfrac{10}{31}$

(b) P (Blue) = $\dfrac{\text{number of blue smarties}}{\text{total number of smarties}} = \dfrac{6}{31}$

(c) P (red or blue) = P(red) + P(blue) … because these events are mutually exclusive

=10/31 + 6/31 = 16/31

Activity 3.6

3.6.1 (a) P(favours) = 90/200 = 0.45

(b) P(favours and rural) = 50/200

(c) P(favours/rural) = 50/100

(d) P(not favour/urban) = 50/100

3.6.2 (a) P(music above) = 20/40 = 0.5

(a) P(maths above/music above) = 15/20 = 0.75

(c) P(maths above **AND** music above) = 15/40

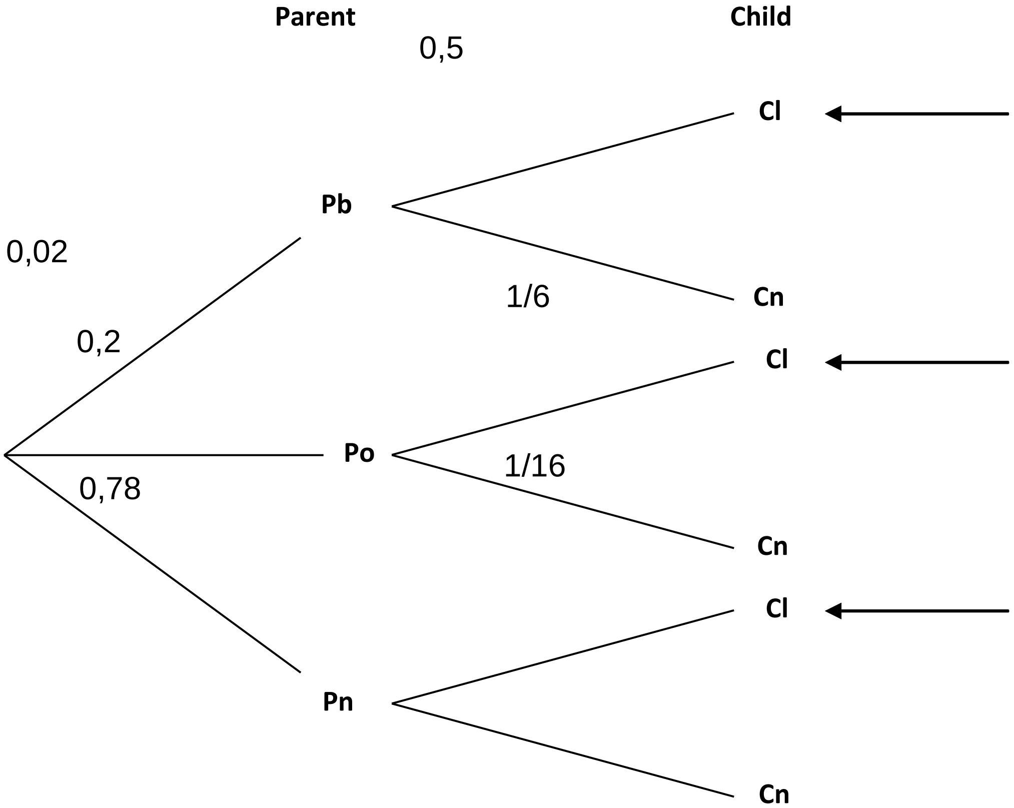
(d) To **TEST** for independence we have to verify that P(A and B) = P(A).P(B)

for all the outcomes of event A and event B

P(maths above AND music above) = 15/40 whereas

P(maths above).P(music above) = 20/40 x 20/40 ≠ 15/40 in other words, the two events are NOT independent

3.6.3



P(Cl) = "climb" the branches that are indicated with an arrow and multiply the probabilities as you go along

= 0,02x0,5 + 0,2x(1/6) + 0,78x(1/16) =1/100 +1/30 + 39/800 = 221/2400 = 0.092

STUDY UNIT 4

FROM SAMPLE TO POPULATION

CONTENTS PAGE

INTRODUCTION 125

ANNEXURE: CURRICULUM ASSESSMENT POLICY STATEMENTS 133

INTRODUCTION

This can be either a very easy or a very difficult study unit, depending on whether you understand the “broader picture”. We would like you to sit back while we recap what we have learned and how you can build on this knowledge. What building blocks do you have?

A RANDOM phenomenon is an action or process that leads to one of several outcomes, for example “measure the time to assemble a computer” or “the flip of a coin”.

You also know that “random” in statistics is not a synonym of “haphazard” but a description of a kind of order that emerges only in the long run.

The **probability** of an event is the proportion of times that the event occurs in many repeated trials of a random phenomenon.

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon. We usually denote random variables by capital letters near the end of the alphabet, such as X or Y. Its lowercase counterpart represents the value of the random variable. For example, if X denotes the time it takes to assemble a computer or if Y denotes the number of outcomes “head” when a coin is flipped three times, then y=2 is an outcome of this variable.

A **discrete random variable** is one that can take on a countable number of values. For example, if we define X as the number of heads observed in an experiment that flips a coin three times, the values of X can be 0, 1,2 and 3.

A c**ontinuous random variable** is one whose values are uncountable. For example: If X denotes the time that it takes to assemble a computer, it can take on any values between say 10 to 12 hours.

A **distribution** is the pattern of variation of a variable. The distribution of a quantitative variable records its numerical values and how often each value occurs, for example the **frequency distributions** that we encountered in the previous study units.

A **probability distribution** is a table, formula or graph that describes ALL THE POSSIBLE values of a random variable and the probability that is associated with these values.

A **discrete probability distribution** lists ALL THE POSSIBLE values and the probabilities of a **discrete** variable, for example:

Value of X (x)	x_1	x_2	x_3	…	x_k
Probability p(x) where p(x) = P(X=x)	p_1	p_2	p_3	…	p_k

The probabilities p_i have to satisfy two requirements:

♦ every probability p_i is a number between 0 and 1

❷ $p_1 + p_2 + p_3 + \ldots + p_k = 1$

For example: Determine whether each of the following distributions is a probability distribution.

x	1	2	3	4
p(x)	1/4	1/4	1/4	1/4

Yes

x	5	10	15	20
p(x)	1/4	1/8	1/16	9/16

Yes

x	0	2	4	6
p(x)	0.3	1.5	0.2	-1

No, the probability cannot be negative.

x	A	B	C
p(x)	0.5	0.3	0.4

No, the total of the probabilities is more than 1.

In the same manner that we can describe a sample by means of the mean and variance, we can describe a probability distribution by means of the mean and variance. The mean is also an average of the possible outcomes, but a weighted average. The weights are the probabilities. Just as probabilities are an idealised description of long-run proportions, the mean of a probability distribution describes the long-run average outcome. We cannot call this mean $\bar{x}$ and therefore need a different symbol. The common symbol for the mean of a probability distribution is the Greek letter mu: μ. To remind ourselves that we are talking about the mean of X, we often write it as μ_x. This parameter is also called the **expected value of X** and is represented by E(X). (You were introduced to these concepts in study unit 7 of Study Guide 1.)

Definition:

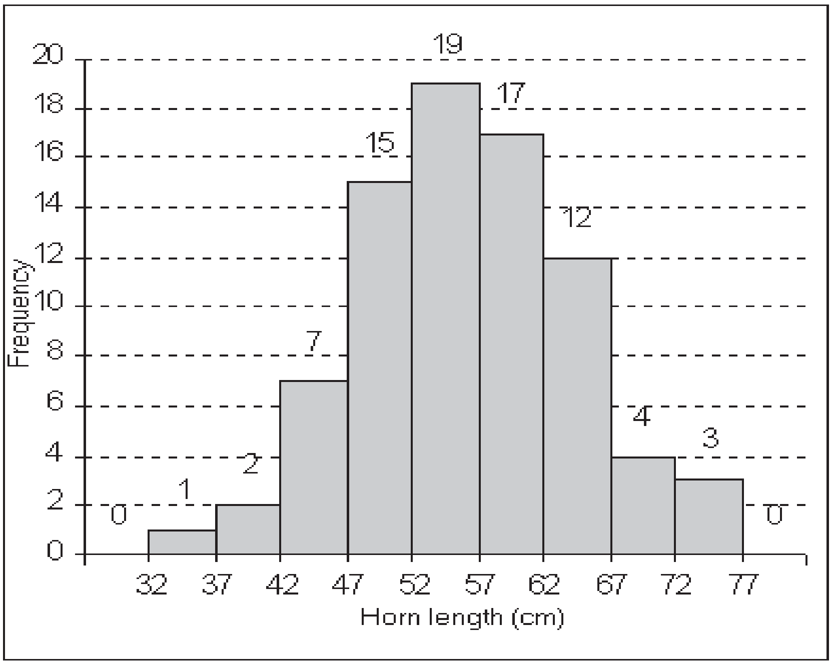
$E(X) = \mu = \sum xp(x)$ and the variance: $^2 = \sum(x - \mu)^2 p(x)$

Our next objective is to understand a continuous distribution.

The following three figures were constructed from data about the horn length of Impala.

Histograms of horn length:

Figure A



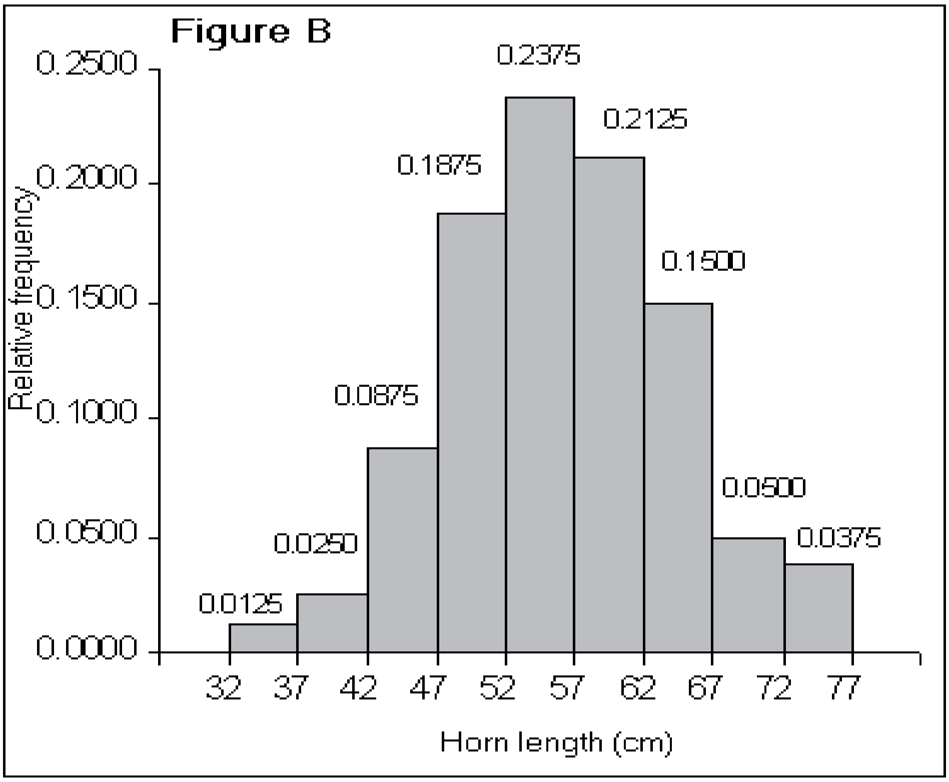


Figure B
Relative frequency
0.2500
0.2000
0.1500
0.1000
0.0500
0.0000
0.0125
0.0250
0.0875
0.1875
0.2375
0.2125
0.1500
0.0500
0.0375
32
37
42
47
52
57
62
67
72
77
Horn length (cm)

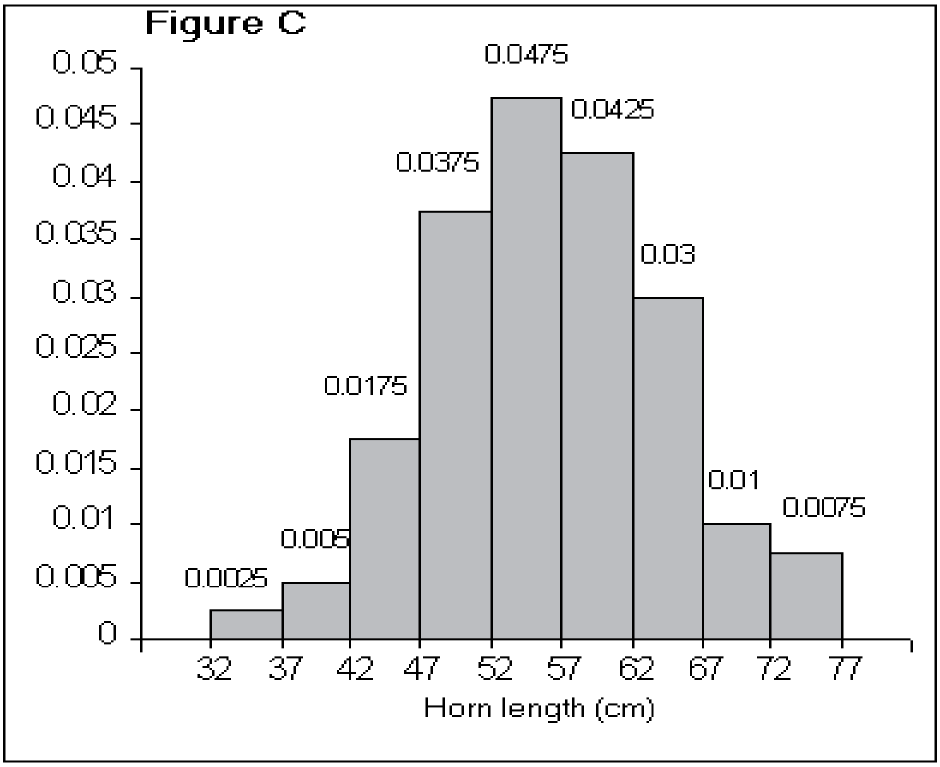


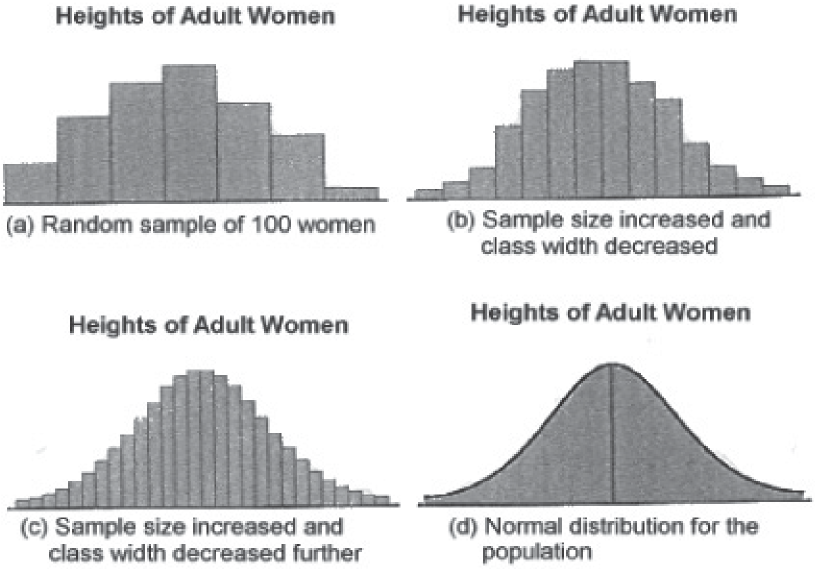
Figure C
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0.0025
0.005
0.0175
0.0375
0.0475
0.0425
0.03
0.01
0.0075
32
37
42
47
52
57
62
67
72
77
Horn length (cm)

Now you must stay with us! First, compare figures A to C. You will agree that the shape of these histograms looks identical. Figure B depicts the relative frequencies and we can say that the "probability" that the length of Impala horns are between 52 cm and 57 cm is 0,2375. If you look at figure C, how do we get the same probability? If you look at the **AREA** under the histogram between 52 and 57: 5 x 0,0475 = 0,2375, the answer is the same as the relative frequency/ proportion/probability.

This correspondence between **AREA** and proportions **(probability)** is the foundation on which the theory of probability for continuous variables is built.

Additionally, we note that the T**OTAL AREA** of the histogram of figure C is **1** (verify this for yourself).

If it was possible to measure the horn length (a continuous variable) of ALL the Impala in, say, the Kruger National Park and draw a "figure C"- type of histogram with VERY small class intervals, it would tend to be well approximated by a smooth curve. The following picture is not our Impala example, but we think that it will help you to get an idea of what is meant by "approximation by a smooth curve":



(a) Random sample of 100 women

(b) Sample size increased and class width decreased

(c) Sample size increased and class width decreased further

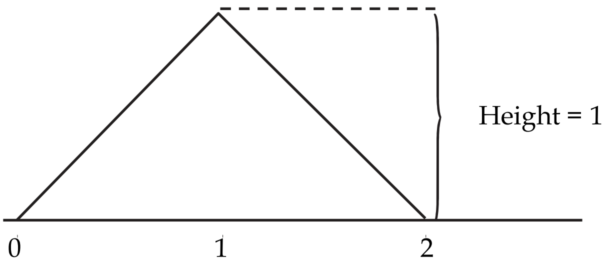
(d) Normal distribution for the population

The way in which the curve describes probabilities is that the probability that a random observation falls between limits **a** and **b** is given by the **AREA UNDER THE CURVE** between those limits. (You are not surprised, are you?)

We call the smooth curve that depicts the behaviour of a continuous random variable X a **density curve** or the graph of a **probability density function.** Again, it is not surprising that the **TOTAL** area under the density curve will be 1, is it?

ACTIVITY 4.1

Assume that the random variable Y is the sum of two random numbers. The random numbers are generated between 0 and 1. Then Y is a continuous random variable that can take on any value between 0 and 2. The density curve of Y is the following triangle:



1. Verify by geometry that the area under the curve is 1. Or verify that this is indeed a density curve.
2. What is the probability that Y is less than 1? Sketch the density curve, shade the area that represents the probability and then find the area. Do this also for question 3.
3. What is the probability that Y is less than 0,5?

ACTIVITY 4.2

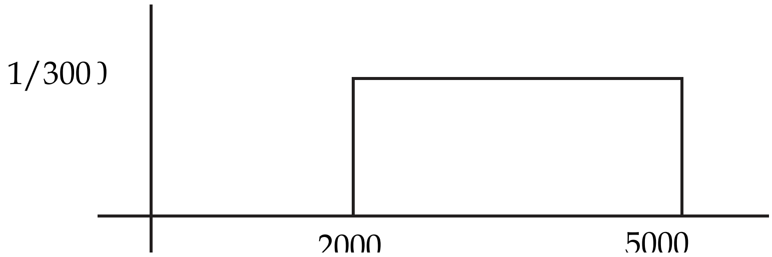
A random variable has the following density function:

f(x) = 1 – 0,5x   0<x<2

1. Graph the density function.
2. Verify that f(x) is indeed a density function.
3. Find the probability that X is more than 1.
4. Find the probability that X is at most 0,5.

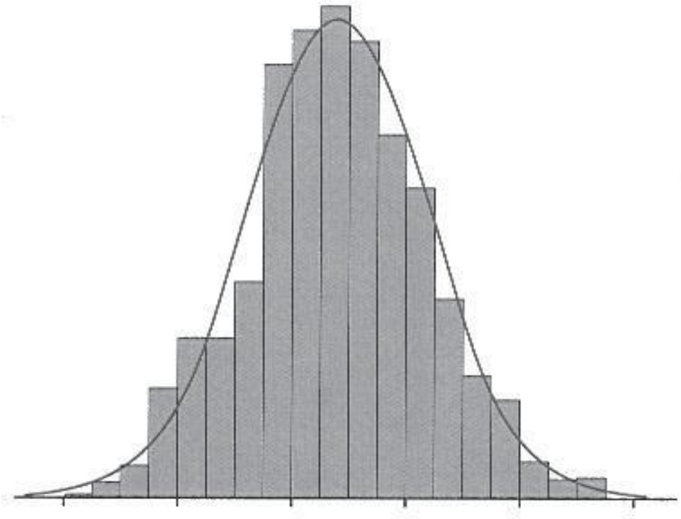
ACTIVITY 4.3

The amount of petrol (measured in litres) that is sold at a service station has the following distribution (also called a uniform distribution):



- What is the minimum amount of petrol that is sold daily?
- What is the maximum amount of petrol that is sold daily?
- Find the probability that daily sales will fall between 2500 and 3000 litres.

We are now ready to introduce the **normal distribution** as a probability density function. Remember that one of the reasons for drawing a histogram of our data was to have an idea of the distribution of our data. Many phenomena in real life will produce a histogram that is approximately symmetric. If you superimposed an approximating smooth curve on the type of histogram in figure C above, the result will be a "bell-shaped" curve. (This curve was introduced when we looked at the properties and interpretation of the standard deviation in study unit 5). In Statistics this class of density curves (do you agree that we can call it density curves?) that are symmetrical and bell shaped are called normal curves and they describe the normal (or Gaussian) distributions.



To calculate the probability that a normal random variable falls into any interval, we have to compute the area in the interval under the curve. Mathematically this implies a rather complicated integration process and we therefore prefer to use existing normal tables or a built-in function in Excel.

Note that in the same way as the probability is an estimate of what will happen "in the long run", no variable fits the normal distribution perfectly since the normal distribution is a picture of what will happen "in the long run". The normal curve is an idealised mathematical description for the distribution

with a specific mathematical equation. You are familiar with the fact that in mathematics curves can be defined by equations (for example the equation of a circle is $x^2 + y^2 = r^2$, where r is the radius). The circle can be used to represent physical objects such as a wheel or a gear. Although it is impossible to manufacture a wheel that is perfect, the equation and the properties of the circle can be used to study the many aspects of the wheel (for example area, velocity and acceleration). In a similar manner, the theoretical curve (called the normal distribution curve) can be used to study many variables that are not perfectly normally distributed but that could "in the long run" be normally distributed.

This introduction to probability distributions is the cornerstone of statistical inference, which is outside the scope of this module

Annexure: Curriculum Assessment Policy Statements

DATA HANDLING INTERMEDIATE PHASE GRADE 4-6

TOPICS	GRADE 4	GRADE 5	GRADE 6
5.1 **Collecting and Organising data**	**Collecting and organising data** • Collect data using tally marks and tables for recording	**Collecting and organising data** • Collect data using tally marks and tables for recording • Order data from smallest group to largest group	**Collecting and organising data** • Collect data -- using tally marks and tables for recording -- using simple questionnaires (yes/no type response) • Order data from smallest group to largest group
5.2 **Representing data**	**Representing data** Draw a variety of graphs to display and interpret data including: • pictographs (one-to-one correspondence between data and representation) • bar graphs	**Representing data** Draw a variety of graphs to display and interpret data including: • pictographs (many-to-one correspondence) • bar graphs	**Representing data** Draw a variety of graphs to display and interpret data including: • pictographs (many-to-one correspondence) • bar graphs and double bar graphs
5.3 **Analysing, Interpreting and Reporting data**	**Interpreting data** Critically read and interpret data represented in • words • pictographs • bar graphs • pie charts **Analysing data** Analyse data by answering questions related to data categories **Reporting data** Summarise data verbally and in short written paragraphs	**Interpreting data** Critically read and interpret data represented in • words • pictographs • bar graphs • pie charts **Analysing data** Analyse data by answering questions related to: • data categories • data sources and contexts **Reporting data** Summarise data verbally and in short written paragraphs that include • drawing conclusions about the data • making predictions based on the data **Ungrouped data** Examine ungrouped numerical data to determine the most frequently occurring score in the data set (mode)	**Interpreting data** Critically read and interpret data represented in • words • pictographs • bar graphs • double bar graphs • pie charts **Analysing data** Analyse data by answering questions related to: • data categories, including data intervals • data sources and contexts • central tendencies – (mode and median) **Reporting data** Summarise data verbally and in short written paragraphs that includes. • drawing conclusions about the data • making predictions based on the data **Ungrouped data** Examine ungrouped numerical data to determine • the most frequently occurring score in the data set (mode) • the middlemost score in the data set (median)

5.4 **Probability**	**Probability experiments** • Perform simple repeated events and list possible outcomes for experiments such as: -- tossing a coin -- rolling a die	**Probability experiments** • Perform simple repeated events and list possible outcomes for experiments such as: -- tossing a coin -- rolling a die -- spinning a spinner • Count and compare the frequency of actual outcomes for a series of trials up to 20 trials	**Probability experiments** • Perform simple repeated events and list possible outcomes for experiments such as: -- tossing a coin -- rolling a die -- spinning a spinner • Count and compare the frequency of actual outcomes for a series of trials up to 50 trials

Curriculum and Assessment Policy Statement (CAPS): Mathematics – Intermediate Phase- DBE, 2011

DATA HANDLING SENIOR PHASE GRADE 7-9

TOPICS	GRADE 7	GRADE 8	GRADE 9
5.1 **Collect, organize and summarize data**	**Collect data** • Pose questions relating to social, economic, and environmental issues in own environment • Select appropriate sources for the collection of data (including peers, family, newspapers, books, magazines) • Distinguish between samples and populations and suggest appropriate samples for investigation • Design and use simple questionnaires to answer questions: -- with yes/no type responses -- with multiple choice responses **Organize and summarize data** • Organize (including grouping where appropriate) and record data using -- tally marks -- tables -- stem-and-leaf displays • Group data into intervals • Summarize and distinguishing between ungrouped numerical data by determining: -- mean -- median -- mode • Identify the largest and smallest scores in a data set and determine the difference between them in order to determine the spread of the data (range)	**Collect data** • Pose questions relating to social, economic, and environmental issues • Select appropriate sources for the collection of data (including peers, family, newspapers, books, magazines) • Distinguish between samples and populations, and suggest appropriate samples for investigation • Design and use simple questionnaires to answer questions with multiple choice responses **Organize and summarize data** • Organize (including grouping where appropriate) and record data using -- tally marks -- tables -- stem-and-leaf displays • Group data into intervals • Summarize data using measures of central tendency, including: -- mean -- median -- mode • Summarize data using measures of dispersion, including: -- range -- extremes	**Collect data** • Pose questions relating to social, economic, and environmental issues • Select and justify appropriate sources for the collection of data • Distinguish between samples and populations, and suggest appropriate samples for investigation • Select and justify appropriate methods for collecting data **Organize and summarize data** • Organize numerical data in different ways in order to summarize by determining: -- measures of central tendency -- measures of dispersion, including extremes and outliers • Organize data according to more than one criteria

5.2 **Represent data**	**Represent data** • Draw a variety of graphs by hand/technology to display and interpret data (grouped and ungrouped) including: -- bar graphs and double bar graphs -- histograms with given intervals -- pie charts	**Represent data** • Draw a variety of graphs by hand/technology to display and interpret data including**:** -- bar graphs and double bar graphs -- histograms with given and own intervals -- pie charts -- broken-line graphs	**Represent data** • Draw a variety of graphs by hand/technology to display and interpret data including: -- bar graphs and double bar graphs -- histograms with given and own intervals -- pie charts -- broken-line graphs -- scatter plots
5.3 **Interpret, analyse, and report data**	**Interpret data** • Critically read and interpret data represented in: -- words -- bar graphs -- double bar graphs -- pie charts -- histograms **Analyse data** • Critically analyse data by answering questions related to: -- data categories, including data intervals -- data sources and contexts -- central tendencies (mean, mode, median) -- scales used on graphs **Report data** • Summarize data in short paragraphs that include -- drawing conclusions about the data -- making predictions based on the data -- identifying sources of error and bias in the data -- choosing appropriate summary statistics for the data (mean, median, mode)	**Interpret data** • Critically read and interpret data represented in: -- words -- bar graphs -- double bar graphs -- pie charts -- histograms -- broken-line graphs **Analyse data** • Critically analyse data by answering questions related to: -- data categories, including data intervals -- data sources and contexts -- central tendencies (mean, mode, median) -- scales used on graphs -- samples and populations -- dispersion of data -- error and bias in the data **Report data** • Summarize data in short paragraphs that include -- drawing conclusions about the data -- making predictions based on the data -- identifying sources of error and bias in the data -- choosing appropriate summary statistics for the data (mean, median, mode, range) -- the role of extremes in the data	**Interpret data** • Critically read and interpret data represented in a variety of ways • Critically compare two sets of data related to the same issue **Analyse data** • Critically analyse data by answering questions related to: -- data collection methods -- summary of data -- sources of error and bias in the data **Report data** • Summarize data in short paragraphs that include -- drawing conclusions about the data -- making predictions based on the data -- making comparisons between two sets of data -- identifying sources of error and bias in the data -- choosing appropriate summary statistics for the data (mean, median, mode, range) -- the role of extremes and outliers in the data

5.4 **Probability**	**Probability** • Perform simple experiments where the possible outcomes are equally likely and: -- list the possible outcomes based on the conditions of the activity -- determine the probability of each possible outcome using the definition of probability	**Probability** • Consider a simple situation (with equally likely outcomes) that can be described using probability and**:** -- list all the possible outcomes -- determine the probability of each possible outcome using the definition of probability -- predict with reasons the relative frequency of the possible outcomes for a series of trials based on probability -- compare relative frequency with probability and explains possible differences	**Probability** • Consider situations with equally probable outcomes, and: -- determine probabilities for compound events using two-way tables and tree diagrams -- determine the probabilities for outcomes of events and predict their relative frequency in simple experiments -- compare relative frequency with probability and explains possible differences

Curriculum and Assessment Policy Statement (CAPS) – Mathematics: Senior Phase- DBE, 2011

NOTES