The availability of this product is due to the financial support of the National Department of Agriculture and the AgriSETA. Terms and conditions apply.
Dear Learner - This Learner Guide contains all the information to acquire all the knowledge and skills leading to the unit standard:

<table>
<thead>
<tr>
<th>Title</th>
<th>US No: 9015</th>
<th>NQF Level: 4</th>
<th>Credits: 5</th>
</tr>
</thead>
</table>

The full unit standard will be handed to you by your facilitator. Please read the unit standard at your own time. Whilst reading the unit standard, make a note of your questions and aspects that you do not understand, and discuss it with your facilitator.

This unit standard is one of the building blocks in the qualifications listed below. Please mark the qualification you are currently doing:

<table>
<thead>
<tr>
<th>Title</th>
<th>ID Number</th>
<th>NQF Level</th>
<th>Credits</th>
<th>Mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Certificate in Animal Production</td>
<td>48979</td>
<td>4</td>
<td>120</td>
<td></td>
</tr>
<tr>
<td>National Certificate in Plant Production</td>
<td>49009</td>
<td>4</td>
<td>120</td>
<td></td>
</tr>
</tbody>
</table>

Please mark the learning program you are enrolled in:

Are you enrolled in a:       Y   N
Learnership?                  [ ] [ ]
Skills Program?               [ ] [ ]
Short Course?                 [ ] [ ]

Your facilitator should explain the above concepts to you.

This Learner Guide contains all the information, and more, as well as the activities that you will be expected to do during the course of your study. Please keep the activities that you have completed and include it in your Portfolio of Evidence. Your PoE will be required during your final assessment.

What is assessment all about?

You will be assessed during the course of your study. This is called formative assessment. You will also be assessed on completion of this unit standard. This is called summative assessment. Before your assessment, your assessor will discuss the unit standard with you.

Assessment takes place at different intervals of the learning process and includes various activities. Some activities will be done before the commencement of the program whilst others will be done during programme delivery and other after completion of the program.

The assessment experience should be user friendly, transparent and fair. Should you feel that you have been treated unfairly, you have the right to appeal. Please ask your facilitator about the appeals process and make your own notes.

Your activities must be handed in from time to time on request of the facilitator for the following purposes:
The activities that follow are designed to help you gain the skills, knowledge and attitudes that you need in order to become competent in this learning module.

It is important that you complete all the activities, as directed in the learner guide and at the time indicated by the facilitator.

It is important that you ask questions and participate as much as possible in order to play an active roll in reaching competence.

When you have completed all the activities hand this in to the assessor who will mark it and guide you in areas where additional learning might be required.

You should not move on to the next step in the assessment process until this step is completed, marked and you have received feedback from the assessor.

Sources of information to complete these activities should be identified by your facilitator.

Please note that all completed activities, tasks and other items on which you were assessed must be kept in good order as it becomes part of your Portfolio of Evidence for final assessment.

Enjoy this learning experience!
How to use this guide …

Throughout this guide, you will come across certain re-occurring “boxes”. These boxes each represent a certain aspect of the learning process, containing information, which would help you with the identification and understanding of these aspects. The following is a list of these boxes and what they represent:

**What does it mean?** Each learning field is characterized by unique terms and definitions – it is important to know and use these terms and definitions correctly. These terms and definitions are highlighted throughout the guide in this manner.

**Activity**

You will be requested to complete activities, which could be group activities, or individual activities. Please remember to complete the activities, as the facilitator will assess it and these will become part of your portfolio of evidence. Activities, whether group or individual activities, will be described in this box.

**Example**

Examples of certain concepts or principles to help you contextualise them easier, will be shown in this box.

**How am I doing?**

The following box indicates a summary of concepts that we have covered, and offers you an opportunity to ask questions to your facilitator if you are still feeling unsure of the concepts listed.

**My Notes …**

You can use this box to jot down questions you might have, words that you do not understand, instructions given by the facilitator or explanations given by the facilitator or any other remarks that will help you to understand the work better.

.................................................................

.................................................................

.................................................................

.................................................................

.................................................................

.................................................................

.................................................................

.................................................................

.................................................................
What are we going to learn?

What will I be able to do? ............................................................... 6
Learning outcomes ........................................................................ 6
What do I need to know? ............................................................... 6
Session 1 Collecting data ............................................................... 7
Session 2 Organising and analysing data ...................................... 20
Appendixes Appendix 1 & 2 ......................................................... 42
Session 3 Probability ................................................................. 48
Glossary ..................................................................................... 64
Checklist for Practical assessment .............................................. 68
Paperwork to be done ............................................................... 69
Terms and conditions ................................................................. 70
Acknowledgements ................................................................. 71
SAQA Unit Standard
What will I be able to do?

When you have achieved this unit standard, you will be able to:

- This Unit Standard is designed to provide credits towards the mathematical literacy requirement of the NQF at Level 4. The essential purposes of the mathematical literacy requirement are that, as the learner progresses with confidence through the levels, the learner will grow in:
  - A confident, insightful use of mathematics in the management of the needs of everyday living to become a self-managing person
  - An understanding of mathematical applications that provides insight into the learner’s present and future occupational experiences and so develop into a contributing worker
  - The ability to voice a critical sensitivity to the role of mathematics in a democratic society and so become a participating citizen.
- People credited with this unit standard are able to:
  - Critique and use techniques for collecting, organising and representing data.
  - Use theoretical and experimental probability to develop models, make predictions and study problems.
  - Critically interrogate and use probability and statistical models in problem solving and decision making in real-world situations

What do I need to know?

It is expected of the learner attempting this unit standard to demonstrate competence against the unit standard:

- The credit value is based on the assumption that people starting to learn towards this unit standard are competent in Mathematical Literacy and Communications at NQF level 3.
Session 1

Collecting Data

After completing this session, you should be able to:

SO 1: Critique and use techniques for collecting, organising and representing data.

In this session we explore the following concepts:

♦ What is data?
♦ Business Decisions and Data
♦ Statistics and the Scientific Method
♦ Elements of Data Collection
♦ Steps in Data Collection
♦ Methods for Collecting Data

1.1 Data, what's that

Data – the word is everywhere! Data processing, databases, numerical data, data-entry operation, data banks, and data just-about-everything else. It's the "in" thing these days – sort of instant prestige (take the word "data" add another noun or two, tack it onto your title, and you've got it). Well despite their prominent place on most office pedestals, data are really sort of everyday fellows.

Data are actually just bits of factual information. Numerical data are bits of factual information stated as numbers – indicating quantities, such as counts or measurements. Data that are somehow related to each other, enough so they can be compared, can be statistical data.

For example your salary figure, by itself, with nothing to compare it with, probably would be of little use to a statistician (some of us feel our salary figure is of little use to anyone).
1.2 Business Decisions and Data

Today, changes occur rapidly. To take advantage of these changes, decisions with long-term implications must often be made quickly. Information is vital. Farmers, for example, must be conscious of market needs and must determine how to react to changing economic, political, and social situations to maintain a competitive advantage and remain profitable.

The question is whether our decisions, based on the information we receive, are making things better, having no effect, or making things worse. We will argue that decisions based on numerical information or data, if the data are collected and processed correctly, typically make things better. Consider the following situations where decisions seem to naturally depend on data.

- A survey of customers indicates that a bank should improve the quality of its customer service. What should the bank do? The first step might be to concentrate on the teller-customer interaction. To improve service, the bank might use data on the lengths of time for teller transactions, the number of customers waiting in line, the teller error rates, the costs of handling exceptional requests, the amount of teller turnover, and so forth. These data could then be used to examine the entire process of providing customers with exceptional service.

- Capital for company projects can be raised by selling bonds or issuing additional shares of stock. The choice of one or the other (or some combination of the two) depends, to some extent, on the likely behaviour of interest rates, inflation, tax policy, and other economic variables. These variables are described by sets of numbers or data. Interest rates may be given by a set of short-term (6-month certificate-of-deposit) and long-term (10-year Treasury note, 30-year Treasury bond) rates. Tax policy may be characterized by a set of tax brackets.

- A drug company has developed a drug for combating the HIV virus that, according to the company, is more effective than current drugs and has fewer side effects. Before the company can produce and market the drug, however, it must receive government approval. Approval from the Medical Council requires that the company successfully complete a set of clinical trials. That is, the proposed drug must be tested, under carefully controlled conditions, on groups of human subjects. Data on the effectiveness and side effects must be collected, analyzed, and reported as part of the company's case. As part of its review, the Medical Council must decide whether the company's claims are sound. Were the data collected properly? Were the data interpreted correctly?

Sound decisions involve the collection of pertinent data and the application of appropriate techniques for extracting the information contained in the data.
1.3 **Statistics and the Scientific Method**

The original idea of "statistics" was the collection of information about and for the "state". The word statistics derives directly, not from any classical Greek or Latin roots, but from the Italian word for state.

The birth of statistics occurred in mid-17th century. A commoner, named John Graunt, who was a native of London, began reviewing a weekly church publication issued by the local parish clerk that listed the number of births, christenings, and deaths in each parish. These so called Bills of Mortality also listed the causes of death. Graunt who was a shopkeeper organized this data in the form we call descriptive statistics, which was published as Natural and Political Observations Made upon the Bills of Mortality. Shortly thereafter he was elected as a member of Royal Society. Thus, statistics has to borrow some concepts from sociology, such as the concept of Population. It has been argued that since statistics usually involves the study of human behaviour, it cannot claim the precision of the physical sciences.

If we knew exactly what was going to happen, when it was going to happen, and to whom it was going to happen, we could prepare for it. Knowing everything makes decision making easy but life less interesting. If Eskom knew how cold it was going to be next winter, it would be simple to plan for the amount of electricity to have available. On the other hand, operating with complete uncertainty is frustrating and often costly. Fortunately, there is a middle ground. Often we can collect or generate numerical information that, although not eliminating uncertainty entirely, will allow us to learn enough about the underlying situation to perform effectively.

Statistics is the body of methodology concerned with the art and science of gathering, analyzing, and using data to identify and solve problems, and to make decisions.

Statistical methods should be regarded as valuable tools. They do not replace critical thinking and common sense. However, if used correctly, statistical methods enable us to generate and assemble numerical information in a way that will help us pick out the signals in the fog, make better decisions, and create more rapid improvements in processes and products.

Statistics is a science assisting you in making decisions under uncertainties based on some numerical and measurable scales. It is about obtaining facts from figures, but moreover it is the ability to transform raw numbers (data) into knowledge that can be acted upon. Today's business decisions are driven by data. Statistical skills enable you to intelligently collect, summarise, analyse and interpret data relevant to your decision-making process. Statistical concepts and statistical thinking enable you to:

- Represent data in meaningful ways that are understandable by others;
- Identify functional relationships between variable quantities;
- Justify decisions on the basis of data: the decision making process must be based on the data and not on personal opinion or belief.
- Solve problems in a diversity of contexts.

Some would say, particularly from a business perspective, that statistics is the study of variation. The Director of Statistical Methods for Nashua Corporation...
maintains that the central problem of management, in all its forms, including planning, research and development, procurement, manufacturing, sales, personnel, accounting, and law, is the failure to completely understand variation. What are its causes? What does it mean?
The director is referring to the differences in numbers of the same type and the failure to understand the information contained in these differences. Variation can often be summarized in simple graphs. Pictures are often easier to interpret than tables or lists of numbers. Spreadsheet programs can be used to construct familiar displays like bar charts and pie charts. For the former, amounts are indicated by the heights of bars. For the latter, amounts are indicated by the areas of pie slices.

1.4 Elements of Data Collection

Collecting data from a complex environment can be time-consuming and expensive. Consequently, it is necessary to have a plan and a systematic collection method that gains maximum information at minimum cost. Every effort should be made to comply with sound data collection principles so that the information received is as accurate and relevant as possible.

There are two types of data:
Primary data and
Secondary data.
Secondary data are collected by someone else and are available in published sources. Quarterly profits published in the Finance Week are secondary data. For our purposes, primary data will refer to data collected directly by the investigator or by the organization employing the investigator. Primary data are collected by a variety of methods: simple observation, personal interview self-Enumeration, check sheets, electronic data capture, experiments simulated on a computer, and controlled laboratory or field experiments.

1.5 Steps in Data Collection

Certain steps help to ensure data quality. Collecting data properly, particularly data collected as part of a survey, is often an intricate process. Data collection requires careful reflection on the complexities involved in a population structure, the practical feasibility of sampling methods, the coordination and supervision of field work, and finally the processing, analyses, and reporting of the data. We briefly introduce these issues by examining the principal steps involved in the collection of primary data.

- **Determine the purpose of the study.**
In order to have a clear goal of the investigation we need to decide on the following:
What do we hope to learn from the data?
How are we going to collect the data?
By defining the purpose of the study as specifically as possible before we even start ensures that we are unlikely to overlook vital information.
Determine the data to be collected.

There are two main ways of collecting data: questionnaires and recording of experimental results

- Questionnaires:

In sampling human populations, the main method for gathering data is the questionnaire. A well-designed questionnaire is crucial to the success of a survey.

A well designed questionnaire is usually anonymous, but allows you to collect background information about the person that is answering the questions. We usually ask questions regarding age, gender, income bracket, number of children or race. These are called demographic questions. Demographic data helps you paint a more accurate picture of the group of persons you are trying to understand. For example, if you want to find out what type of chocolate is the most popular, you will probably get very different answers from males or females, teenagers or elderly people.

Two types of questions can be asked in a questionnaire: open ended questions or closed questions. If you ask: “How old are you?” the person answering has to write an answer. This is an open ended question. The advantage of these questions is that you get a very accurate reflection. The disadvantages are that this type of questionnaire takes longer to complete and that you will find it very difficult to group the responses when you analyse the data later.

You could ask the same question as follows:

Mark your response with a cross.

<table>
<thead>
<tr>
<th>Your age</th>
<th>10-20 years</th>
<th>20-30 years</th>
<th>30-40 years</th>
<th>40-50 years</th>
<th>Older than 50 years</th>
</tr>
</thead>
</table>

The person answering the questions does not have to write anything, but merely place a cross. This type of question is answered more easily and quickly. Analysis of data is also much faster.

Please see appendix 1 for more information about the different ways you can phrase questions in a questionnaire.
Please complete Activity 1:
Below is a questionnaire provided by a restaurant in Nelspruit to see if the service is good enough. Study the questionnaire and answer the questions:

How is our service?
Please take a few minutes to complete our survey. Mark your response with a cross.

<table>
<thead>
<tr>
<th>1. Your age</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below 20 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20-30 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30-40 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40-50 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Older than 50 years</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. Your gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Nelspruit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 10 km of Nelspruit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 20 km from Nelspruit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 30 km from Nelspruit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than 30 km from Nelspruit</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Where do you live?</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 10 km of Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 20 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 30 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>More than 30 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. How often do you visit a restaurant?</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>More than once a week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Once a week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Once a month</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A few times a year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hardly ever</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5. How friendly are our waiters?</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 30 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 20 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 10 km of Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6. How quickly did you get served?</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 30 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 20 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 10 km of Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. Appearance of your meal</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 30 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 20 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 10 km of Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8. Quality of the food</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 to 30 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 to 20 km from Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within 10 km of Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>In Nelspruit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 9. What did you like best about your visit to our restaurant? | | |
|---------------------------------------------------------------||

1. What kind of information do questions 1-4 provide?
2. Why do you think the restaurant included question 3?
3. Why is question 4 important?
4. What other question could be included in this category?
5. Which questions are
   a. closed
   b. open ended?
6. In question 7 why was “excellent” listed as the first choice and not the last?

**Experimental results**

If we perform an experiment in the field or in a laboratory, then we need to plan how we are going to record our data. We draw up a sample table during the planning phase. The table is then used during the collection of data. We need to decide what our independent, dependent and controlled variables are. For example, Mr Rose, a flower farmer wants to see if a new brand of fertiliser will give him more flowers. He will plant two plots with exactly the same plants, but give them different fertiliser. The variables are as follows:

Independent variable: fertiliser type (i.e. either type A or type B)
Dependent variable: Number of flowers produced per plant.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture
NQF Level 4
Unit Standard No: 9015

Controlled variable: How much water is given and how often, soil type must be the same in both plots, types of plants must be the same, the two plots must receive the same amount of light etc.

By deciding on the variables, farmer Rose knows how to conduct the experiment and what results to collect.

Please complete Activity 2
Farmer Brown currently uses SuperDip to ensure that his cattle is free of parasites. A rep has given him a free sample of ParaDip to try. Farmer Brown decides to do a trial to see which dip works better. He decides to divide his cattle into three groups. Group 1 will receive no dip at all, Group 2 will be dipped with SuperDip and Group 3 will be treated with ParaDip. After a week he will stick a piece of clear tape onto the fur on the neck. The tape will be removed quickly so that the parasites adhere to the tape. The number of parasites on the tape will then be counted under the microscope.

a. Name the independent variable in this investigation.
b. Name the dependent variable in this investigation.
c. List three controlled variables.
d. Why would Farmer Brown include a group of undipped cattle in his experiment?
e. Farmer Brown has more than 4000 cattle. He does not have the time or the money to count the parasites on 4000 strips of tape. Suggest a way in which he can get reliable results at a minimum of time and money.

Determine the plan for collecting the data.
The method used for collecting data depends on the money and equipment available, and also on the purpose of the investigation. You must chose a sample design and a sample size.

The choice of the sampling design is based on such factors as the structure of the population, the type of information sought, and the administrative facilities and personnel available to carry out the plan. In conjunction with choosing the appropriate data collection method, we determine the required sample size by specifying the degree of precision desired in the sample summary measure. Data collection costs money, thus you must chose a sample procedure that you can afford.

Farmer Rose would select two small plots close to each other to conduct his fertiliser trial. He would need to repeat the experiment a few times in order to collect sufficient data for analysis. Perhaps he could select six plots of 1 m² each, so that he could conduct the trial three times at the same time. Farmer Brown (Work book Activity 2), on the other hand, could select 30 cattle that are similar in size for each treatment. In this way he could get a reasonably good idea which dip works best without spending huge amounts of money and time.

Train personnel.
Training is often needed for the people responsible for actually recording the observations and organizing them in files. Training may take several forms and may be ongoing if the data collection takes place over an extended period of time.
Everyone should understand the importance of the data so that data collection is taken seriously. Why are the data being collected? What difference does it make to the organization? What happens if the data contain errors?

Please complete Activity 3
a. What kind of training does Farmer Rose need to give his staff?
b. What kind of training does Farmer Brown need to give his staff?
c. What kind of information does each farmer need to give to his workers?

**Analyze and report the data.**
Once the data collection plan is established and the data are collected, the full force of graphic and numerical techniques can be used to interpret the results. Ingenuity in creating plots and careful data analysis can suggest interesting relationships and conclusions that may be considered in additional studies. Data may be reported as lists of numbers, as summary measures, in graphical form, in tables, or as an equation, or they may simply be described verbally. In general, data should be reported so the information is readily apparent to those who will use it to make decisions. If the appropriate presentation method is unclear, ask the user. Alternatively, imagine yourself as the decision maker. Which methods work for you?

**Solution and Discussion.**
- State the purpose of the study.

The objective is to get information from subordinates about the performance of their managers. This project is important because changes in organizational structure, salary adjustments, and advancement opportunities will depend on the outcomes. Data will be collected with a questionnaire given to all workers. Responses will be confidential.

- Determine the data to be collected.

Performance variables might include the following: ability to communicate goals and objectives; accessibility; leadership ability; ability to delegate; commitment to quality; management of resources; interpersonal skills; availability of feedback; and overall effectiveness. Questions related to these variables will be included on the questionnaire. The response to each performance variable question will be indicated on a 5-point scale.

![Scale](1 2 3 4 5)

Almost never  Almost always

A sample question follows.
Describe the plan for collecting the data.

The data will be collected from all workers. A two-person team has been designated to construct and distribute the questionnaire. The questionnaire will be limited to a single page. The team will help workers to fill in the questionnaire, and tabulate and summarize the data.

Train personnel.

Each of the team members has completed a short training course on questionnaire construction and ways of dealing with people. Farmer Green has instructed each of the team members on the importance of this project. A draft version of the questionnaire will be tested with a small group of people to be sure the questions are clearly stated and provide the information required.

Analyze and report the data.

After the questionnaires are produced and distributed, the team will allow a week for the questionnaires to be completed. As the questionnaires are returned, they will be checked for the manager's name and to see if all the questions have been answered. Data will be entered into a file in a spreadsheet program as the completed questionnaires are received. One member of the personnel team is responsible for data entry. Another member of the team will check a printout of the file for errors. Reports for each manager, including a sample questionnaire and the data summaries, will be prepared and forwarded. Each report will also indicate where the raw data are stored and how they can be accessed. The members of the team will be identified as the authors of the reports.

1.6 Methods for Collecting Data

A variety of procedures exist for collecting data. Three commonly used ones are:

- Observation
- Personal interview
- Self-enumeration

- **Observation**
  Observation entails direct examination and recording of an ongoing activity.
1. In a study of family decision making, a researcher observed and recorded interactions between husband and wife as they decided on the brand of colour television to buy.
2. In an engineering study, data about the internal temperature of an oven were obtained by reading all instruments inserted in the oven.
3. In a marketing study, an analyst monitored customer flow in a department store by means of closed-circuit television.

The observation procedure has certain advantages:

- The directness of the procedure avoids problems such as incomplete or distorted recall.
- Data can be gathered more or less continuously over an extended time period.

Limitations of the method include the following:

- The observer (or the instrument) must be free of bias and must accurately record the events of interest. Human observers usually require thorough training so that they will record precisely what they observe and so that different observers will record the same events in the same manner.
- Individuals who are under observation and aware of this fact may alter their behaviour and, as a result, observations of their behaviour may be biased.

**Personal Interview.**

In a personal interview, an interviewer asks questions that are printed on a questionnaire and records the respondent's answers in designated spaces on the questionnaire form.

1. A household member was interviewed at home about purchases of toothpastes and mouthwashes and about family characteristics, such as income and family size.
2. A household member was interviewed over the telephone about television viewing, including viewing at the moment of call, the station viewed, and number of persons viewing.
3. A company's financial executive was interviewed at the office by a representative of a trade association about the firm's plans for capital expenditures in the coming year.

Both the advantages and limitations of securing data through personal interviews arise from the direct contact between the respondent and the interviewer. Advantages include the following:

- Persons will tend to respond when they are approached directly; hence, the personal-interview procedure usually yields a high proportion of usable returns from those persons who are contacted.
- The direct contact generally enables the interviewer to clear up misinterpretations of questions by the respondent, to observe the respondent's reactions to particular questions, and to collect relevant supplementary information.

There are several limitations of the personal-interview method:
The interviewer may not follow directions for selecting respondents. For instance, if a member of the family other than the one designated is interviewed, a bias may be introduced into the results.

The interviewer may influence the respondent by the manner in which the questions are asked or by other actions. A slight inadvertent gesture of surprise at an answer, for example, can exert subtle, undetected pressures on the respondent.

The interviewer may make errors in recording the respondent's answers;

**Self-Enumeration.**

With self-enumeration, the respondent is provided with a questionnaire to complete, which often also contains necessary instructions.

---

1. A student who graduated from high school recently received a self-enumeration questionnaire through the mail, a page of which is shown in Figure 1.1, to provide information about educational activities since graduation.

2. A new magazine subscriber received a questionnaire through the mail to provide information about age, type of job held, income, and amounts of money spent last year on specified recreational activities.

3. A person completed a certificate of registration for a motor vehicle, supplying information on make, model, and year of car.

4. A purchaser of a toaster filled out the warranty card, giving information on family characteristics and on the primary method by which attention was directed to this appliance (e.g., word of mouth, television commercial).

Both the advantages and limitations of the self-enumeration procedure arise from the elimination of interviewers. The type of interviewer error discussed earlier is thus avoided. On the other hand, the absence of interviewers creates two serious problems:

- When a questionnaire is sent to a household or an organization, there is no control over which person answers the questions.

- The absence of interviewers can lead to low response rates. A low response rate can be a source of serious bias in survey results, because the persons who do answer the questionnaires are often not representative of the entire group contacted. The user of data collected by a self-enumeration procedure should therefore know the rate of non-response as one factor affecting the magnitude of the potential bias.

Where there is a material rate of non-response some follow-up of non-respondents will be valuable. In most well-conducted mail surveys, some or all of the non-respondents are contacted as a routine procedure. Non-respondents may be contacted by means of "reminder" letters, telephone calls, or special personal interviews.
### Please complete Activity 4

Complete the table below. Make sure that you summarise the information.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Personal Interview</th>
<th>Self-Enumeration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>What is it?</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Example</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(only one needed)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Since you left high school, have you taken any courses in a technical or trade school, hospital school, beauty school, business school, other vocational school or attend a learnership?
   - [ ] Yes – Continue with question 2
   - [ ] No – Continue with question 6

2. When did you first begin classes at a vocational, technical, trade or business school or the learnership?
   - Month: 116
   - Year: 20
<table>
<thead>
<tr>
<th>Question</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. What was your field of study when you began taking these classes?</td>
<td></td>
</tr>
<tr>
<td>(for example, beautician, auto mechanics, IT Technical Support, Customer Management, etc)</td>
<td>□ or No specific field of study</td>
</tr>
<tr>
<td>4. What is the full name and address of the learning institution you attend?</td>
<td>Name _________________________________ Address: _________________________________</td>
</tr>
<tr>
<td>5a. At the institution you entered in question 4, were you enrolled in a program leading to a formal qualification?</td>
<td>□ Yes – Continue with question 5b □ No – Continue with question 6</td>
</tr>
<tr>
<td>5b. How many years of continuous full-time study does it usually take to complete the requirements for the programme you were enrolled?</td>
<td>□ Less than 6 months □ 6 months to 1 year □ 2 years □ 3 years</td>
</tr>
<tr>
<td>5c. Did you complete this programme?</td>
<td>□ Yes – When did you complete that program? Go to question 6 □ No – When did you last attend classes? Continue with question 5d</td>
</tr>
<tr>
<td>5d. What are the reasons that you did not complete the programme? (Mark X to all that apply)</td>
<td>□ Financial reasons □ Academic problems □ Took a job □ Other – specify ________________</td>
</tr>
<tr>
<td>6. Since you left high school, have you been enrolled in a college or university?</td>
<td>□ Yes – Continue with question 7 □ No – Go directly to question 15a</td>
</tr>
</tbody>
</table>

**Figure 1.1 Example of part of a self-enumeration questionnaire**

<table>
<thead>
<tr>
<th>Concept (SO 1)</th>
<th>I understand this concept</th>
<th>Questions that I still would like to ask</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situations or issues that can be dealt with through statistical methods are identified correctly.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriate methods for collecting, recording and organising (data are used so as to maximise efficiency and ensure the resolution of a problem or issue.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data sources and databases are selected in a manner that ensures the representativeness of the sample and the validity of resolutions.

Activities that could result in contamination of data are identified and explanations are provided of the effects of contaminated data.

Data is gathered using methods appropriate to the data type and purpose for gathering the data.

Data collection methods are used correctly.

Calculations and the use of statistics are correct.

Graphical representations and numerical summaries are consistent with the data, are clear and appropriate to the situation and target audience.

Resolutions for the situation or issue are supported by the data and are validated in terms of the context.

My Notes ...

organisation and analysing data

After completing this session, you should be able to:

SO 2: Use theoretical and experimental probability to
In this session we explore the following concepts:
- The Use of Measurement
- Data Organisation and Presentation
- Range
- Measures of Central Tendency
- How spread out is the data?
- Scatterplots
- The Correlation Coefficient – a Measure of Linear Relation

2.1 Introduction

Among the first activities in statistical analysis is to count or measure: Counting/measurement theory is concerned with the connection between data and reality. A set of data is a representation (i.e., a model) of the reality based on numerical and measurable scales.

How data is presented and analysed depends on the type of data collected. Certain statistical methods are valid for certain data types only. Any characteristic being measured or observed (e.g. assessment results, choice of canteen meal) is referred to as a variable. Since data collected for statistical purposes can take on a variety or spread of values that are not pre-determined, these variables are termed random variables. E.g. When measuring assessment results, the values of the data can range from 0% to 100%. The random variable under study is Assessment Result. The individual data values are the values between zero and one hundred. They are also referred to as observations.

2.2 The use of measurement

Data may be classified in different ways according to the nature of the random variable under study. Random variables may be classified as:

- **Qualitative**: which yield non-numeric answers. Examples are eye colour, gender, marital status (i.e. married, divorced, single, widowed, living together) This type of information cannot be manipulated mathematically.

- **Quantitative**: which yield numeric responses that can be mathematically manipulated. Data is measured by Quantitative random variables (e.g. Income, Prices, Assessment results).

Data of random variables can also be classified as:

- **Discrete Data** (or discrete random variables): are data that can assume specific values only, (usually whole numbers). Discrete random variables are characterised by data that is countable. E.g. The number of employees in an organisation.

- **Continuous Data** (or continuous random variables): are data that can assume any numerical value. Continuous random variables are characterised by data that is measurable. E.g. income, age, time to complete a task, the height of a person.
2.3 Data organisation and presentation

There are many ways of presenting data. No one style is appropriate for all types of data. You should select the style so that the observer will be able to understand the features of the data that you want to illuminate. Here are a few types of data presentation styles:

- **Tables**
  Lists of numerical quantities. If the data set is "small" just listing the data can be done. As the data set becomes larger this is not a good technique. Remember that people can only look at a very small set of number without "turning-off".

<table>
<thead>
<tr>
<th>YEAR</th>
<th>CERTIFICATES</th>
<th>DIPLOMAS</th>
<th>DEGREES</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>50</td>
<td>20</td>
<td>0</td>
<td>70</td>
</tr>
<tr>
<td>2000</td>
<td>70</td>
<td>35</td>
<td>10</td>
<td>115</td>
</tr>
<tr>
<td>2001</td>
<td>100</td>
<td>35</td>
<td>15</td>
<td>150</td>
</tr>
<tr>
<td>2002</td>
<td>130</td>
<td>50</td>
<td>20</td>
<td>200</td>
</tr>
<tr>
<td>2003</td>
<td>200</td>
<td>100</td>
<td>30</td>
<td>330</td>
</tr>
<tr>
<td>TOTAL</td>
<td>550</td>
<td>240</td>
<td>75</td>
<td>865</td>
</tr>
</tbody>
</table>

- **Pictograms**
  The simplest way of visually representing data is to use pictograms. Pictures are used to compare quantities. This can have the look as a bar graph as well as add some interest to the viewer. Pictograms were dealt with in great detail in earlier modules and you are invited to refer back to work from previous years.

- **Pie Charts**
  Used to illustrate proportions of the total amount. The size of each slice shows how much of the total the slice represents. This type of graph is seen quite often in economics. The disadvantage is that the viewer has no idea how big the actual sample size was.

My Notes ...

---------------------------------

Educational Levels in Industry

- Certificates
- Diplomas
- Degrees

63% 28% 9%
Bar Graphs

(Bar Charts). The length of a bar represents the quantity that is under comparison. There are many types of bar charts, but they all should try to have a uniform width, spacing, and a zero value. Remember that bar charts have gaps between them. (Learning tip: bar graph has a gap in its name!). Bar graphs are used when there is no continuity between data groups. For example: if you are comparing the sales if a number of different types of bakkies you will use a bar graph. There is no continuity between a Nissan bakkie and a Toyota bakkie.

My Notes ...

.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................

Histogram

A histogram has no spaces between the bars. It is used when there is continuity between the categories selected. A histogram is also used in showing frequency distributions.
Graphs used to illustrate some characteristic over time. These and regression plots study two random variables simultaneously. There are many examples of this in economics. E.g. Daily share prices.

♦ Graphs must be clearly labelled so as to avoid any ambiguity as to what information is being conveyed.

♦ The reader must also be informed of the source of the presentation so that the data may be authenticated if necessary.

The ultimate choice of presentation depends on the data type:

- For qualitative data (nominal and ordinal) use tables, pie and bar charts.
- For quantitative data (interval and ratio) use histograms or line graphs. The latter is useful for looking at trends over time.

Graphs can be powerful tools to convey information but can also be confusing to understand. Try to keep them simple and use the type of graph which best illustrates information about the data and its characteristics.
Please complete Activity 5

1. At a college 180 learners were interviewed regarding their alcohol drinking habits.
   a) Complete the table below.

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Number of learners</th>
<th>Degrees</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learners who do not drink</td>
<td>110</td>
<td>40</td>
</tr>
<tr>
<td>Male learners who drink</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Female learners who drink</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>180</td>
<td></td>
</tr>
</tbody>
</table>

   b) Construct a pie chart to reflect the data

2. Below is a table showing predicted % of males and females infected with HIV between 2008 and 2011.

<table>
<thead>
<tr>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult males (20-65)</td>
<td>24%</td>
<td>23%</td>
<td>22%</td>
</tr>
<tr>
<td>Adult females (20-65)</td>
<td>25%</td>
<td>24%</td>
<td>23%</td>
</tr>
</tbody>
</table>

   a) Draw a double bar chart to show the prevalence of HIV in males and females between 2008 and 2011.

3. The table below shows the matric pass rates from 1995 - 2004

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass rate (%)</td>
<td>53</td>
<td>54</td>
<td>47</td>
<td>49</td>
<td>49</td>
<td>56</td>
<td>62</td>
<td>69</td>
<td>73</td>
</tr>
</tbody>
</table>

   a) Calculate the mean pass rate for the time period 1995 – 2004
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

b) Construct a line graph to show how the matric pass rate has changed.

c) Write a paragraph describing the changes that have occurred in the matric pass rate. Use your line graph as a guide.

d) Give an explanation for the increase in pass rates from 1998 to 2003.

2.4 Range

The range is defined as the difference between the maximum and minimum data values.

Farmer Green has measured the height (in meters) of his calves and has the following data set, arranged from smallest to biggest:
0,7; 0,8; 1,3; 1,5; 1,5

The range of heights is 1,5m – 0,7m = 0,8m
2.5 Measures of central tendency

Most measurements are distributed randomly in the range that they were collected in, but most of them concentrate around some sort of “central tendency” or “average”. Three measures of central tendency are the mean, median and mode.

<table>
<thead>
<tr>
<th>Arithmetic Mean</th>
<th>Median</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work out the total and divide by the number of items</td>
<td>Write numbers from smallest to biggest</td>
<td>Write numbers from smallest to biggest</td>
</tr>
<tr>
<td>Add up all the wages</td>
<td>R200; R200; R240; R250; R250; R250; R260; R260; R275; R520</td>
<td>R200; R200; R240; R250; R250; R250; R260; R260; R275; R520</td>
</tr>
<tr>
<td>Total = R2705</td>
<td>The two middle numbers are R250 and R250.</td>
<td>The two middle numbers are R250 and R250.</td>
</tr>
<tr>
<td>Mean = total/number of wages</td>
<td>Median = (R250+R250)/2 = R250</td>
<td>Median = (R250+R250)/2 = R250</td>
</tr>
<tr>
<td>Mean = 2705/10 = R270.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The mean is R270.50</td>
<td>The median is R250.00</td>
<td>The mode is R250.00</td>
</tr>
</tbody>
</table>

Thabo employs 10 workers on his farm. Below are the weekly wages of the 10 workers: R250; R275; R200; R520; R250; R260; R250; R260; R200; R240

### The Main Characteristics of the Mode, the Median, and the Mean

<table>
<thead>
<tr>
<th>Fact No.</th>
<th>The Mode</th>
<th>The Median</th>
<th>The Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>It is the value that appears the most often.</td>
<td>It is the middle-most value.</td>
<td>Add all the values and divide the total by the number of items.</td>
</tr>
<tr>
<td>2</td>
<td>A distribution may have 2 or more modes. On the other hand, there can also be no mode.</td>
<td>Each array has one and only one median.</td>
<td>An array has one and only one mean.</td>
</tr>
<tr>
<td>3</td>
<td>It cannot be manipulated algebraically.</td>
<td>It cannot be manipulated algebraically.</td>
<td>Means may be manipulated algebraically.</td>
</tr>
<tr>
<td>4</td>
<td>Individual values need to known to calculate the mode</td>
<td>Individual values need to known to calculate the mode</td>
<td>You can calculate it even if you do not know the individual values. You need to</td>
</tr>
</tbody>
</table>
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

<table>
<thead>
<tr>
<th>5</th>
<th>Values must be arranged from smallest to biggest.</th>
<th>Values must be arranged from smallest to biggest.</th>
<th>Values need not be ordered or grouped for this calculation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Tells you what score occurs the most often.</td>
<td>Provides a better measure of location than the mean when there are some extremely large or small observations. Median income is used as the measure for the SA household income.</td>
<td>Very easy to calculate and used the most often when there is a list of numbers.</td>
</tr>
</tbody>
</table>

Please complete Activity 6

1. Farmers were interviewed to find out how many times they had to call out the local vet over the last 12 months. The data are listed below.

   1; 2; 6; 0; 3; 2; 3; 1; 0; 5; 4; 0; 3; 2; 4; 0; 3; 1; 0; 2; 0; 2; 1; 0; 1; 0; 3; 4; 1; 1; 0; 1; 2; 2; 5

   a) Arrange the data in order
   b) Calculate the mean
   c) Find the median
   d) What is the mode?
   e) Calculate the range

2. The frequency distribution of the amounts of bonus pay earned during the last year for 65 salespersons follows:

<table>
<thead>
<tr>
<th>Amount of Bonus Pay</th>
<th>Number of Salespersons</th>
<th>Total amount paid to this salary bracket</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 – under 100</td>
<td>4</td>
<td>50x4 = 200</td>
</tr>
<tr>
<td>100 – under 200</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>200 – under 400</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>400 – under 600</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>600 – under 1000</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1000 – under 2000</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td></td>
</tr>
</tbody>
</table>

   a) Complete the table above
   b) Calculate the mean amount of bonus pay per employee.
   c) What is the approximate total amount of bonuses paid to all salespersons? Why is this an approximation?
   d) Draw a histogram of the data.

### 2.6 How spread-out is data?

Sometimes data are not spread equally (i.e. they are skewed), making the mean senseless. In this case the median is a better representation of the central tendency. To get an idea of the spread of data, we calculate quartiles and percentiles.
Quartiles and the 5-number summary
Consider the following list of data: 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11
The median is 6 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11
The median is the middle most value. The first quartile is the value that lies in the middle of the group of data below the median.
The first quartile is 3 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11
The third quartile is the value that lies in the middle of the group of data above the median.
The third quartile is 9 1; 2; 3; 4; 5; 6; 7; 8; 9; 10; 11
You can quite clearly see that the median is the same as the second quartile.
The inter-quartile range is the 3rd quartile – 2nd quartile. Here it is 9-3 = 6. The Inter-quartile range tells us that 50% of data lie between the values of 3 and 9.
We can summarise the data set in a 5-number system:

- Minimum value of the data set
- 1st quartile
- 2nd quartile = median
- 3rd quartile
- Maximum value of the data set

Let us prepare a 5-number summary for the data set above:
- Minimum value of the data set: 1
- 1st quartile: 3
- 2nd quartile = median: 6
- 3rd quartile: 9
- Maximum value of the data set: 11

Below is a list of test marks of students, arranged in order. Calculate the 1st, 2nd and 3rd quartiles.
2; 5; 5; 5; 12; 13; 13; 14; 15; 15; 15; 17; 19; 19; 19; 20; 22; 24; 25; 26; 26; 30
First we calculate the 2nd quartile or median:
2; 5; 5; 5; 12; 13; 13; 14; 15; 15; 15; 17; 19; 19; 19; 20; 22; 24; 25; 26; 26; 30
Median = (15+17)/2 = 16
Now we calculate the 1st quartile:
2; 5; 5; 5; 12; 13; 13; 14; 15; 15; 15; 17; 19; 19; 19; 20; 22; 24; 25; 26; 26; 30
1st quartile = (12+13)/2 = 12,5
Now we calculate the 3rd quartile:
2; 5; 5; 5; 12; 13; 13; 14; 15; 15; 17; 19; 19; 19; 20; 22; 24; 25; 26; 26; 30
3rd quartile = (22+24)/2 = 23
Now comes the 5-number summary:
- Minimum value of the data set: 2
- 1st quartile: 12,5
- 2nd quartile = median: 16
- 3rd quartile: 23
- Maximum value of the data set: 30
- Inter-quartile range: 10,5
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture
NQF Level 4
Unit Standard No: 9015

To see how valuable the 5-number summary is, let us look at an example.

Mrs. Naidoo has set 3 Maths tests for her class. She wants to compare how the learners have fared in the three tests. Below is the summary of her results.

<table>
<thead>
<tr>
<th>Test</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>4</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>1st quartile</td>
<td>11</td>
<td>17</td>
<td>26.5</td>
</tr>
<tr>
<td>2nd quartile/median</td>
<td>26</td>
<td>30</td>
<td>39</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>39</td>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td>Maximum</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Comparing tests 1 and 2, we can see that the minimum and maximum values are the same. The quartiles for test 2 are much higher than for test 1, showing that pupils fared better in test 2. In Test 3 the minimum value is higher, showing that the weakest pupil improved. The quartiles are even higher than test 2, indicating that the class improved overall. Mrs. Naidoo is happy.

Please complete Activity 7
1. Below are the same data gathered regarding vet visits to farms. This time the data have been arranged in order:
   0; 0; 0; 0; 0; 0; 0; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 2; 2; 2; 2; 2; 2; 2; 3; 3; 4; 4; 4; 4; 5; 5; 6
   Find the following:
   a) the second quartile
   b) the first quartile
   c) the third quartile
   d) the inter-quartile range

Cumulative frequency tables:
You saw in activity 7 that the calculations of large data sets is very cumbersome. There is a quicker way, namely by using cumulative frequency tables. First, let us revise frequency tables.

In a certain rural area a survey was done to see how many dogs people had. The results are listed below. Arrange these data in a frequency table:
3; 5; 1; 3; 7; 5; 6; 5; 9; 5; 2; 4; 4; 5; 5; 8
First, a tally table is drawn up. Secondly, the tallies are counted and recorded in the frequency column.

<table>
<thead>
<tr>
<th>Number</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>II</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>II</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>I+I+I</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>I</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>16</strong></td>
</tr>
</tbody>
</table>
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

<table>
<thead>
<tr>
<th>Number</th>
<th>Tally</th>
<th>Frequency</th>
<th>Cummulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 1 dog</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>1</td>
<td>1+1=2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 2 dogs or less</td>
</tr>
<tr>
<td>3</td>
<td>II</td>
<td>2</td>
<td>2+2=4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 3 dogs or less</td>
</tr>
<tr>
<td>4</td>
<td>II</td>
<td>2</td>
<td>4+2=6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 4 dogs or less</td>
</tr>
<tr>
<td>5</td>
<td>I I I</td>
<td>6</td>
<td>6+6=12</td>
</tr>
<tr>
<td></td>
<td>I</td>
<td></td>
<td>Total of people having 5 dogs or less</td>
</tr>
<tr>
<td>6</td>
<td>I</td>
<td>1</td>
<td>12+1=13</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 6 dogs or less</td>
</tr>
<tr>
<td>7</td>
<td>I</td>
<td>1</td>
<td>13+1=14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 7 dogs or less</td>
</tr>
<tr>
<td>8</td>
<td>I</td>
<td>1</td>
<td>14+1=15</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 8 dogs or less</td>
</tr>
<tr>
<td>9</td>
<td>I</td>
<td>1</td>
<td>15+1=16</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Total of people having 9 dogs or less</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>16</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number</th>
<th>Frequency</th>
<th>Cummulative Frequency</th>
<th>Which data items lie here?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Item 1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>Item 2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>Items 3-4</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6</td>
<td>Items 5 - 6</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>12</td>
<td>Items 7 - 12</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Median lies between item 8 and 9. This means that the median is 5</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>13</td>
<td>Item 13</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>14</td>
<td>Item 14</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>15</td>
<td>Item 15</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>16</td>
<td>Item 16</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

The median is 5.

First quartile calculation: 1\textsuperscript{st} quartile lies between \((16+1)/4 = 4,25\) i.e between items 4 and 5. From the table we can see that item 4 is 3 and item 5 is 4. Thus the 1\textsuperscript{st} quartile is \((3+4)/2 = 3,5\)

3\textsuperscript{rd} quartile calculation: The 3\textsuperscript{rd} quartile lies between \(3(16+1)/4 = 12,75\), i.e. between items 12 and 13. The 3\textsuperscript{rd} quartile is 5,5.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture
NQF Level 4
Unit Standard No: 9015

**Cumulative frequency graphs (Ogives)**

Let us plot the ogive of the data in the table above.

<table>
<thead>
<tr>
<th>Number</th>
<th>Cumulative Frequency</th>
<th>Points to plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>(0; 0)</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>(1; 1)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(2; 2)</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>(3; 4)</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>(4; 6)</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>(5; 12)</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>(6; 13)</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>(7; 14)</td>
</tr>
<tr>
<td>8</td>
<td>15</td>
<td>(8; 15)</td>
</tr>
<tr>
<td>9</td>
<td>16</td>
<td>(9; 16)</td>
</tr>
</tbody>
</table>

**Ogive for vet visits in the past year**

Why Ogives are useful:
- An ogive usually has the shape of a stretched out S. The better the S shape, the more the data are centered around the middle scores.
- Just by looking at the graph you can tell that the number of farmers interviewed was 16.
- We can use the ogive to **estimate** quartiles:
  - **Median**: take \((16+1)/2 = 8,5\).
  - Draw a horizontal line on the graph at 8,5.
  - Draw a vertical line down from the point where your line cuts the graph's line.
  - Read off the x-value: it is approximately 4,5.
  - **1st quartile**: Take \((16+1)/4 = 4,25\)
  - Draw a horizontal line on the graph at 4,25.
  - Draw a vertical line down from the point where your line cuts the graph's line.
  - Read off the x-value: it is approximately 3.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems.

Primary Agriculture

NQF Level 4

Unit Standard No: 9015

• 3rd quartile: Take \(3(16+1)/4 = 12.75\)
• Draw a horizontal line on the graph at 12.75.
• Draw a vertical line down from the point where your line cuts the graph’s line.
• Read off the x-value: it is approximately 6

**Variance and standard deviation**
The measure of how measures are scattered around the mean are variance and standard deviation.

A farmer wanted to see which plot of land was better for growing trees, plot A or plot B. He measured the height of 7 young trees on each plot.

Plot A: 364cm, 372cm, 364cm, 368cm, 370cm, 368cm, 370cm
Plot B: 304cm, 388cm, 332cm, 432cm, 400cm, 352cm, 368cm

We calculate the mean of each group:
Plot A: \(2576/7 = 368\)cm
Plot B: \(2576/7 = 368\)cm

The means for both plots were the same, so we cannot gain much insight that way.

We now calculate the deviation from the mean for each data point:

<table>
<thead>
<tr>
<th>Plot A</th>
<th>Plot B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (cm)</td>
<td>Deviation from the mean</td>
</tr>
<tr>
<td>364</td>
<td>368 – 364 = 4</td>
</tr>
<tr>
<td>372</td>
<td>368 – 372 = -4</td>
</tr>
<tr>
<td>364</td>
<td>368 – 364 = 4</td>
</tr>
<tr>
<td>368</td>
<td>368 – 368 = 0</td>
</tr>
<tr>
<td>370</td>
<td>368 – 370 = -2</td>
</tr>
<tr>
<td>368</td>
<td>368 – 368 = 0</td>
</tr>
<tr>
<td>370</td>
<td>368 – 370 = -2</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
</tr>
<tr>
<td>Mean (Variance)</td>
<td>56/7 = 8</td>
</tr>
</tbody>
</table>

Add up all the deviations for Plot A: \(-4 + 4 + 0 - 2 + 0 - 2 = 0\)
Add up all the deviations for Plot B: \(-65 - 20 + 36 - 64 - 32 + 16 + 0 = 0\)
This is also not helping so we need to go one step further.

We re-write the table and add an extra column where we square the deviations from the mean. For the first value Plot A (364) the deviation from the mean was 4. We square this number to obtain 16. We do this for every single value. See table below.

<table>
<thead>
<tr>
<th>Plot A</th>
<th>Plot B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (cm)</td>
<td>Deviation from the mean</td>
</tr>
<tr>
<td>364</td>
<td>4</td>
</tr>
<tr>
<td>372</td>
<td>-4</td>
</tr>
<tr>
<td>364</td>
<td>4</td>
</tr>
<tr>
<td>368</td>
<td>0</td>
</tr>
<tr>
<td>370</td>
<td>-2</td>
</tr>
<tr>
<td>368</td>
<td>0</td>
</tr>
<tr>
<td>370</td>
<td>-2</td>
</tr>
<tr>
<td>Total</td>
<td>0</td>
</tr>
<tr>
<td>Mean (Variance)</td>
<td>56/7 = 8</td>
</tr>
</tbody>
</table>
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  NQF Level 4  Unit Standard No: 9015

| mean = std deviation | 8 = 2.83 | 1410 = 37.55 |

Now we add up all the squares of the deviation from the mean to give us a total. For Plot A the total is 56. Now we calculate the mean of the square of the deviations i.e. we take 56 and divide by 7 (there were 7 measurements in the sample). This gives an answer of 8. This is called the variance.

Variance is defined as the mean of the square of the deviations from the mean.

If we take the square root of the variance, we get the standard deviation. The standard deviation for Plot A is 2.83.

The standard deviation is defined as the square root of the mean of the square of the deviations from the mean. Quite a mouth full!!!

Let us summarise the steps:

i) Add up all the data values
ii) Find the mean.
iii) Calculate the deviations from the mean for each value.
iv) Square the deviations from the mean.
v) Add up all the square deviations from the mean and divide by the number of values. This gives you the variance.
vi) Take the square root of the variance to obtain the standard deviation.

Let us take a closer look at the farmer's result for his trees.

<table>
<thead>
<tr>
<th>Plot A</th>
<th>Plot B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>368cm</td>
</tr>
<tr>
<td>Variance</td>
<td>8</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>2.83</td>
</tr>
</tbody>
</table>

Both the variance and the standard deviation show us that there was very little variation in the trees from Plot A. The trees from Plot B, however, had hugely different growths. If the farmer needs to sell fairly uniform trees to logging companies, he would be better off to plant on Plot A. He could do further analysis to see why the trees on Plot B are so very different to one another, or he could use the land for another purpose.
Please complete Activity 8

1. Farmer Black has breeds his own cattle for meat. He would like to build an abattoir on his farm, so that he can slaughter the meat himself. There is only one butcher shops in the area, and farmer Black needs to find out if it is worth selling his meat there. He did a survey to see how many residents visit either shop. The results are set out below as frequency tables.

<table>
<thead>
<tr>
<th>No. of visits</th>
<th>Frequency</th>
<th>Cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Butcher shop A

a. Complete the table above by working out the cumulative frequencies.

b. Calculate the following values: (make sure that you show all your working.
   b. median.
   c. 1st quartile
   d. 3rd quartile
   e. Range
   f. Inter-quartile range

2. Average weekly wages and salaries (in Rand s) in 60 rural areas of South Africa are:

<table>
<thead>
<tr>
<th>300</th>
<th>285</th>
<th>270</th>
<th>302</th>
<th>308</th>
<th>295</th>
<th>304</th>
<th>312</th>
<th>270</th>
<th>280</th>
</tr>
</thead>
<tbody>
<tr>
<td>310</td>
<td>305</td>
<td>306</td>
<td>312</td>
<td>285</td>
<td>306</td>
<td>316</td>
<td>302</td>
<td>290</td>
<td></td>
</tr>
<tr>
<td>290</td>
<td>310</td>
<td>320</td>
<td>280</td>
<td>306</td>
<td>275</td>
<td>294</td>
<td>306</td>
<td>300</td>
<td></td>
</tr>
<tr>
<td>298</td>
<td>320</td>
<td>280</td>
<td>320</td>
<td>302</td>
<td>285</td>
<td>298</td>
<td>302</td>
<td>270</td>
<td></td>
</tr>
<tr>
<td>296</td>
<td>266</td>
<td>285</td>
<td>325</td>
<td>302</td>
<td>262</td>
<td>300</td>
<td>294</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td>306</td>
<td>268</td>
<td>286</td>
<td>315</td>
<td>294</td>
<td>296</td>
<td>310</td>
<td>290</td>
<td>325</td>
<td></td>
</tr>
<tr>
<td>305</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Construct a frequency distribution of these data using equal class intervals of 40.

Frequency table:

<table>
<thead>
<tr>
<th>Classes Mark</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>290 – 329</td>
<td>310</td>
</tr>
<tr>
<td>330 – 369</td>
<td>350</td>
</tr>
<tr>
<td>370 – 409</td>
<td>390</td>
</tr>
<tr>
<td>410 – 449</td>
<td>430</td>
</tr>
<tr>
<td>450 – 489</td>
<td>470</td>
</tr>
<tr>
<td>490 – 529</td>
<td>510</td>
</tr>
</tbody>
</table>

(b) What is the mean weekly wage?
(c) What is the median weekly wage?
(d) What is the mode of the weekly wage?
(e) What is the value of the first quartile of weekly wages?
Please complete Activity 9
1. A lecturer wanted to see if his Maths examination was well structured. He analysed the students results and drew up a frequency table.

<table>
<thead>
<tr>
<th>Percentage</th>
<th>Number of learners</th>
<th>Cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-10</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>11-20</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>21-30</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>31-40</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>41-50</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>51-60</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>61-70</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>71-80</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>81-90</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>91-100</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

a) Complete the table.
b) Draw an ogive on the grid provided.
c) Use the graph to find the median.
d) Find the 1st and 3rd quartile using the graph
e) Is the examination well structured? Give a reason for your answer.

Please complete Activity 10
1. A farmer plants sugar cane and mieres on his farm. He has collected data regarding the yield that he has obtained each year for the past 10 years. He has calculated the variance and the standard deviation for each of the data sets:

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sugar cane</td>
<td>14</td>
<td>3,74</td>
</tr>
<tr>
<td>Mielies</td>
<td>28</td>
<td>5,29</td>
</tr>
</tbody>
</table>

a. What do these numbers tell you about the spread of the data?
b. Which is the more reliable crop? Justify your answer.
2. A soccer team has two excellent players: Seboko and Surprise. The coach has placed Seboko on the team and has left Surprise off. Naturally, Surprise is annoyed. His father advises him to analyse the goals scored by both players in the matches of the last season. Surprise collects the data:
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  NQF Level 4  |  Unit Standard No: 9015

 Surprise | Seboko
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Goals scored</td>
<td>Deviation from the mean</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

**a.** Complete the table above to calculate the variance and standard deviation for each player.

**b.** Both players scored the same mean number of goals over the last 10 matches, yet Surprise is the better player. Explain why this is so.

**c.** Suggest a reason why the coach chose Seboko over Surprise.

**d.** Suggest a way in which Surprise could handle the situation when he speaks to the coach.

### 2.7 Scatterplots

Sometimes you have more than one variable to consider at one time. For example, a big company has to decide whether they really get more income if they spend more on advertising. They are trying to compare two variables. We say the data are bivariate.

To see whether income does increase with increased cost of advertising, the company would collect as many data as possible regarding both variables. They would pair the values and write them as (x;y) ordered pairs. They would then plot all these points on a graph. The horizontal axis is the explanatory variable X and the vertical axis is the response variable Y. The result will be scattered points all over the graph. This is known as a scatter plot. Below is an example of a scatter plot.
The purpose of collecting bivariate observations is to answer such questions as:
Are the variables related?
If so, does an increase in one variable cause an increase or decrease in the other variable?
What is the nature of the relationship indicated by the data?
Can we quantify the strength of the relation?
Can we make predictions?
Studying the $x$ measurements by themselves or the $y$ measurements by themselves would not help us answer these questions.
The scatterplot provides a visual impression of the relation between the $X$ and $Y$ variables. If the points cluster along a line, a linear relation is indicated. Points that band around a curve indicate a curvilinear relation. If the points form a patternless cluster, no relation among the variables is indicated.

Constructing and interpreting a scatterplot
Total revenue (millions of Rands) and operating income (millions of Rands) for the $n = 26$ teams in the Provincial Rugby League for the 2003 – 2004 season can be determined from the data given in the table below. Suppose we believe that total revenue determines or explains, to a large extent, operating income. Plot these pairs of observations as a scatterplot.

<table>
<thead>
<tr>
<th>Total Revenue</th>
<th>Operating Income</th>
<th>Total Revenue</th>
<th>Operating Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>30.4</td>
<td>49</td>
<td>10.5</td>
</tr>
<tr>
<td>79.3</td>
<td>20</td>
<td>48.4</td>
<td>9.9</td>
</tr>
<tr>
<td>91.1</td>
<td>27</td>
<td>46</td>
<td>9</td>
</tr>
<tr>
<td>88.7</td>
<td>26.3</td>
<td>45.8</td>
<td>8.9</td>
</tr>
<tr>
<td>81.5</td>
<td>10.7</td>
<td>52</td>
<td>9</td>
</tr>
<tr>
<td>50.6</td>
<td>11.1</td>
<td>40.3</td>
<td>7</td>
</tr>
<tr>
<td>78</td>
<td>18</td>
<td>44</td>
<td>7.5</td>
</tr>
<tr>
<td>59.1</td>
<td>12.7</td>
<td>44.7</td>
<td>7.7</td>
</tr>
<tr>
<td>64.5</td>
<td>16</td>
<td>38</td>
<td>6.2</td>
</tr>
<tr>
<td>66</td>
<td>13.9</td>
<td>42</td>
<td>6.9</td>
</tr>
<tr>
<td>58</td>
<td>4.8</td>
<td>30</td>
<td>6</td>
</tr>
<tr>
<td>54</td>
<td>12</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>49</td>
<td>10</td>
<td>25</td>
<td>8</td>
</tr>
</tbody>
</table>

Scatterplot of Operating Income and Total Revenue
Solution and Discussion

The scatterplot of total revenue and operating income for the Rugby League teams is shown in the figure above. Total revenue is the explanatory variable (X variable), and operating income as the response (Y variable).

The points in the scatterplot look as though they lie along a straight line. Low revenue is paired with low operating income. This is what we would expect. When this happens, we say there is a positive association or positive relation between the two variables.

In addition to providing a graphical description of the association between two variables, scatterplots often reveal information that is not evident from looking at the numbers themselves. The next example illustrates this point.

## Example

### Scatterplot illustrating a linear relation with increasing variation

The following table contains estimated and actual costs (millions of Rands) for 26 construction projects. Plot the bivariate construction cost data as a scatterplot.

<table>
<thead>
<tr>
<th>Estimated Cost</th>
<th>Actual Cost</th>
<th>Estimated Cost</th>
<th>Actual Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>575</td>
<td>.918</td>
<td>8.947</td>
<td>13.371</td>
</tr>
<tr>
<td>6.127</td>
<td>7.214</td>
<td>3.157</td>
<td>1.553</td>
</tr>
<tr>
<td>11.215</td>
<td>14.577</td>
<td>37.400</td>
<td>27.973</td>
</tr>
<tr>
<td>28.195</td>
<td>30.028</td>
<td>7.650</td>
<td>7.642</td>
</tr>
<tr>
<td>30.100</td>
<td>38.173</td>
<td>13.700</td>
<td>3.692</td>
</tr>
<tr>
<td>21.091</td>
<td>15.320</td>
<td>29.003</td>
<td>29.522</td>
</tr>
<tr>
<td>40.630</td>
<td>34.100</td>
<td>5.292</td>
<td>5.292</td>
</tr>
<tr>
<td>1.803</td>
<td>2.003</td>
<td>.960</td>
<td>.707</td>
</tr>
<tr>
<td>18.048</td>
<td>20.099</td>
<td>1.240</td>
<td>1.246</td>
</tr>
<tr>
<td>8.102</td>
<td>4.324</td>
<td>1.419</td>
<td>1.143</td>
</tr>
<tr>
<td>10.730</td>
<td>10.523</td>
<td>38.936</td>
<td>21.571</td>
</tr>
</tbody>
</table>

### Solution and Discussion

Let the explanatory variable, X, be estimated cost and the response variable, Y, be actual cost. The (x,y) values are graphed as a scatterplot in the figure below, with the horizontal axis representing estimated cost and the vertical axis representing actual cost.

The first point to be plotted is (x1,y1) = (.575,.918). Again, the southwest to northeast pattern of the points indicates a positive association between estimated cost and actual cost; that is, (relatively) high estimated costs tend to occur with (relatively) high actual costs, and (relatively) low estimated costs with (relatively) low actual costs.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture
NQF Level 4
Unit Standard No: 9015

If the engineers could explain (predict) construction costs with no error, the estimated cost would equal the actual cost for each project and all the points would lie along the diagonal line through the origin. Notice that the points in the scatterplot “fan out” as the values increase. The points corresponding to small projects are closer together than the points corresponding to big (expensive) projects. The deviation (difference) of actual cost from estimated cost appears to increase with the size of the project. The engineers typically come close to determining the costs of smaller projects. They are less successful with the larger ones.

2.8 The correlation coefficient – a measure of linear relation

The scatterplot gives a visual impression of the relation between the x and y values in a bivariate data set. In many cases, the points appear to cluster around a straight line.
A numerical measure of the closeness of the scatter to a straight line is provided by the sample correlation coefficient.
The sample correlation coefficient, denoted by r, is a measure of the strength of the linear relation between the X and Y variables. The manner in which the correlation coefficient assesses the strength of the linear relation is summarized as follows:

♦ The value of r is always between -1 and +1.
♦ The magnitude of r indicates the strength of the linear relation, and its sign indicates the direction. In particular,
  • $r > 0$ if the pattern of $(x, y)$ values is a band that runs from lower left to upper right
  • $r < 0$ if the pattern of $(x, y)$ values is a band that runs from upper left to lower right
  • $r = +1$ if all $(x, y)$ values lie exactly on a straight line with a positive slope (perfect positive linear relation)
  • $r = -1$ if all $(x, y)$ values lie exactly on a straight line with a negative slope (perfect negative linear relation)
A value of r close to -1 or +1 represents a strong linear relation. A value of r close to 0 means that the linear relation is very weak.
It is a good idea to interpret r in conjunction with a scatterplot of the bivariate data. If there is no visible relation, that is, if the y values do not change in any direction as the x values change, then r will be close to 0.
Also, a value of r near 0 can occur if the scatterplot points band around a curve that is far from linear. These situations, and others, are illustrated in the figure below. Keep in mind that the correlation coefficient is a measure of a linear or straight-line relation. A value of r close to 0 indicates the absence of a linear relation, but it does not necessarily mean that there is no relationship.
The figure below illustrates the correspondence between scatter diagram patterns and the value of \( r \). Notice that (e) and (f) correspond to situations where \( r = 0 \). The zero correlation in (e) is due to an absence of any relation between \( X \) and \( Y \). The zero correlation in (f) is due to a relation that is quite strong but far from linear.

A commodities trader contacted a supplier of sunflower to learn about demand. The supplier provided the average price per ton (in R'000) and the number of tons sold in '000s of tons for an eight-year period. The data are as follows:

<table>
<thead>
<tr>
<th>Price per ton (R'000)</th>
<th>Tons sold (000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.25</td>
<td>125</td>
</tr>
<tr>
<td>1.75</td>
<td>105</td>
</tr>
<tr>
<td>2.25</td>
<td>65</td>
</tr>
<tr>
<td>2.00</td>
<td>85</td>
</tr>
<tr>
<td>2.50</td>
<td>75</td>
</tr>
<tr>
<td>2.25</td>
<td>80</td>
</tr>
<tr>
<td>2.70</td>
<td>50</td>
</tr>
<tr>
<td>2.50</td>
<td>55</td>
</tr>
</tbody>
</table>

a. Draw a scatterplot of the data.
b. Do the points follow a pattern? Explain

c. Which of the following options is correct? 
   i) r>0  
   ii) r<0  
   iii) r=0

d. Describe the relationship in words.

<table>
<thead>
<tr>
<th>Concept (SO 2)</th>
<th>I understand this concept</th>
<th>Questions that I still would like to ask</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiments and simulations are chosen and/or designed appropriately in terms of the situation to be modelled.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predictions are based on validated experimental or theoretical probabilities.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The results of experiments and simulations are interpreted correctly in terms of the real context.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The outcomes of experiments and simulations are communicated clearly.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

My Notes ...

.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
.................................................................
Appendix 1

A More information about structuring questionnaires.

♦ Yes/no questions

Any question on a survey that has yes or no as a possible response is nominal, and so binomial statistics will be applied whenever a single yes/no question serves as the dependent variable or one of the dependent variables in an analysis.

♦ Likert scales

A special kind of survey question uses a set of responses that are ordered so that one response is greater than another. The term Likert scale is named after the inventor, Rensis Likert, whose name is pronounced “Lickert.” Generally, this term is used for any question that has about 5 or more possible options. An example might be: "How would you rate your department administrator?" 1=very incompetent, 2=somewhat incompetent, 3=neither competent, 4=somewhat competent, or 5=very competent.

Likert scales are either ordinal or interval, and many psychometricians would argue that they are interval scales because, when well constructed, there is equal distance between each value. So if a Likert scale is used as a dependent variable in an analysis, normal theory statistics are used such as regression would be used.

♦ Physical measures

Most physical measures, such as height, weight, systolic blood pressure, distance etc., are interval or ratio scales, so they fall into the general "continuous" category. Therefore, normal theory type statistics are also used when such a measure serves as the dependent variable in an analysis.

♦ Counts

Counts are tricky. If a variable is measured by counting, such as the case if a researcher is counting the number of days a hospital patient has been hospitalized, the variable is on a ratio scale and is treated as a continuous variable. So normal theory statistics, like correlation, is used. If a researcher is counting the number of subjects in an experiment (or number of cases in the data set), a continuous type measure is not really being used. Counting in this instance is really examining the frequency that some value of a variable occurs. For example, counting the number of subjects in the data set that report having been hospitalized in the last year, relies on a dichotomous variable in the data set that stands for being hospitalized or not being hospitalized (e.g., from a question such as “have you been hospitalized in the last year?”). Even if one were to count the number of cases based on the question "how many days in the past year have you been hospitalized," which is a continuous measure, the variable being used in the analysis is really not this continuous variable.

Instead, the researcher would actually be analyzing a dichotomous variable by counting the number of people who had not been hospitalized in the past year (0 days) vs. those that had been (1 or more days).
B Survey Question and Answer Types

So you've decided that you need a better understanding of the characteristics of people who visit your website, or of some other business-related question. Developing a focused and effective questionnaire will help you to efficiently and accurately pinpoint the information that will help you make more informed decisions.

Developing a questionnaire is as much an art as it is a science. And just as an artist has a variety of different colours to choose from in the palette, you have a variety of different question formats with which to question an accurate picture of your customers, clients and issues that are important to them.

♦ The Dichotomous question

The dichotomous question is generally a "yes/no" question. An example of the dichotomous question is:
Have you ever purchased a product or service from our website?
Yes
No

If you want information only about product users, you may want to ask this type of question to "screen out" those who haven't purchased your products or services. Researchers use "screening" questions to make sure that only those people they are interested in participate in the survey.

You may also want to use yes/no questions to separate people or branch into groups of those who "have purchased" and those who "have not yet purchased" your products or services. Once separated, different questions can be asked of each of these groups.

You may want to ask the "have purchased" group how satisfied they are with your products and services, and you may want to ask the "have not purchased" group what the primary reasons are for not purchasing. In essence, your questionnaire branches to become two different sets of questions.

♦ The multiple-choice questions

The multiple-choice question consists of three or more exhaustive, mutually exclusive categories. Multiple-choice questions can ask for single or multiple answers. In the following example, we could ask the respondent to select exactly one answer from the 7 possible, exactly 3 of the 7, or as many as 3 of the 7 (1,2, or 3 answers can be selected).

Example

A multiple-choice question to find out how a person first heard about your website is:
How did you first hear about our web site?
Television
Radio
Newspaper
Magazine
Word-of-mouth
Internet
Other: Please Specify _______________

For this type of question it is important to consider including an "other" category because there may be other avenues by which the person first heard about your site that you might have overlooked.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

♦ Rank order scaling

Rank order scaling questions allow a certain set of brands or products to be ranked based upon a specific attribute or characteristic. Perhaps we know that Toyota, Honda, Mazda, and Ford are most likely to be purchased. You may request that the options be ranked based upon a particular attribute. Ties may or may not be allowed. If you allow ties, several options will have the same scores.

Based upon what you have seen, heard, and experienced, please rank the following brands according to their reliability. Place a "1" next to the brand that is most reliable, a "2" next to the brand that is next most reliable, and so on. Remember, no two cars can have the same ranking.

- Honda
- Toyota
- Mazda
- Ford

♦ The rating scale

A rating scale question requires a person to rate a product or brand along a well-defined, evenly spaced continuum. Rating scales are often used to measure the direction and intensity of attitudes. The following is an example of a comparative rating scale question:

Which of the following categories best describes your last experience purchasing a product or service on our website? Would you say that your experience was:

- Very pleasant
- Somewhat pleasant
- Neither pleasant nor unpleasant
- Somewhat unpleasant
- Very unpleasant

♦ The semantic differential scale

The semantic differential scale asks a person to rate a product, brand, or company based upon a seven-point rating scale that has two bi-polar adjectives at each end. The following is an example of a semantic differential scale question.

Would you say our website is:

(7) Very Attractive
(6)
(5)
(4)
(3)
(2)
(1) Very Unattractive

Notice that unlike the rating scale, the semantic differential scale does not have a neutral or middle selection. A person must choose, to a certain extent, one or the other adjective.
The staple scale

The staple scale asks a person to rate a brand, product, or service according to a certain characteristic on a scale from +5 to -5, indicating how well the characteristic describes the product or service. The following is an example of a staple scale question:

When thinking about Data Mining Technologies, Inc. (DMT), do you believe that the word "innovative" aptly describes or poorly describes the company? On a scale of +5 to -5 with +5 being "very good description of DMT" and -5 being "poor description of DMT," how do you rank DMT according to the word "innovative"?

- (+5) Describes very well
- (+4)
- (+3)
- (+2)
- (+1)
- Innovative
- (-1)
- (-2)
- (-3)
- (-4)
- (-5) Poorly Describes

The constant sum question

A constant sum question permits collection of "ratio" data, meaning that the data is able to express the relative value or importance of the options (option A is twice as important as option B).

The following question asks you to divide 100 points between a set of options to show the value or importance you place on each option. Distribute the 100 points giving the more important reasons a greater number of points. The computer will prompt you if your total does not equal exactly 100 points.

When thinking about the reasons you purchased our TargetFind data mining software, please rate the following reasons according to their relative importance.

- Seamless integration with other software
- User friendliness of software
- Ability to manipulate algorithms
- Level of pre- and post-purchase service
- Level of value for the price
- Convenience of purchase/quick delivery

Total 100 points

This type of question is used when you are relatively sure of the reasons for purchase, or you want input on a limited number of reasons you feel are important. Questions must sum to 100 points and point totals are checked by javascript.

The open-ended question

The open-ended question seeks to explore the qualitative, in-depth aspects of a particular topic or issue. It gives a person the chance to respond in detail.
Although open-ended questions are important, they are time-consuming and should not be over-used.

An example of an open-ended question might be:

(If the respondent indicates they did not find what they were looking for...)
What products of services were you looking for that were not found on our website?
If you want to add an "Other" answer to a multiple choice question, you would use branching instructions to come to an open ended question to find out What Other....

The demographic question

Demographic questions are an integral part of any questionnaire. They are used to identify characteristics such as age, gender, income, race, geographic place of residence, number of children, and so forth. For example demographic questions will help you to classify the difference between product users and non-users. Perhaps most of your customers come from the Northeast, are between the ages of 50 and 65, and have incomes between R50,000 and R75,000.
And by better understanding the type of people who use or are likely to use your product, you can allocate promotional resources to reach these people, in a more cost effective manner.
Psychographic or life style questions are also included in the template files. These questions provide an in-depth psychological profile and look at activities, interests and opinions of respondents.

Appendix 2

Sampling methods

From the food you eat to the television you watch, from political elections to school board actions, much of your life is regulated by the results of sample surveys.
A sample is a group of units selected from a larger group (the population). By studying the sample, one hopes to draw valid conclusions about the larger group.
A sample is generally selected for study because the population is too large to study in its entirety. The sample should be representative of the general population. This is often best achieved by random sampling. Also, before collecting the sample, it is important that one carefully and completely defines the population, including a description of the members to be included.
A common problem in business statistical decision-making arises when we need information about a collection called a population but find that the cost of obtaining the information is prohibitive. For instance, suppose we need to know the average shelf life of current inventory. If the inventory is large, the cost of checking records for each item might be high enough to cancel the benefit of having the information.
On the other hand, a hunch about the average shelf life might not be good enough for decision-making purposes. This means we must arrive at a compromise that involves selecting a small number of items and calculating an average shelf life as an estimate of the average shelf life of all items in inventory.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture
NQF Level 4
Unit Standard No: 9015

This is a compromise, since the measurements for a sample from the inventory will produce only an estimate of the value we want, but at substantial savings. What we would like to know is how "good" the estimate is and how much more will it cost to make it "better". Information of this type is intimately related to sampling techniques. This section provides a short discussion on the common methods of business statistical sampling.

- **Cluster sampling**
  Cluster sampling can be used whenever the population is homogeneous but can be partitioned. In many applications the partitioning is a result of physical distance. For instance, in the insurance industry, there are small "clusters" of employees in field offices scattered about the country. In such a case, a random sampling of employee work habits might not require travel to many of the "clusters" or field offices in order to get the data. Totally sampling each one of a small number of clusters chosen at random can eliminate much of the cost associated with the data requirements of management.

- **Random sampling**
  Random sampling is probably the most popular sampling method used in decision making today. Many decisions are made, for instance, by choosing a number out of a hat or a numbered bead from a barrel, and both of these methods are attempts to achieve a random choice from a set of items. But true random sampling must be achieved with the aid of a computer or a random number table whose values are generated by computer random number generators.

- **Cross-Sectional Sampling**
  Cross-Sectional study the observation of a defined population at a single point in time or time interval. Exposure and outcome are determined simultaneously. What is a statistical instrument? A statistical instrument is any process that aim at describing a phenomena by using any instrument or device, however the results may be used as a control tool. Examples of statistical instruments are questionnaire and surveys sampling.

  What is grab sampling technique? The grab sampling technique is to take a relatively small sample over a very short period of time, the results obtained are usually instantaneous. However, the Passive Sampling is a technique where a sampling device is used for an extended time under similar conditions. Depending on the desirable statistical investigation, the passive sampling may be a useful alternative or even more appropriate than grab sampling. However, a passive sampling technique needs to be developed and tested in the field.

---

My Notes ...

---
Session 3

Probability

After completing this session, you should be able to:
SO 3: Critically interrogate and use probability and statistical models.

In this session we explore the following concepts:
- What is probability?
- The probability of a single event
- Compound events
- Probability tree diagrams
- Qualities of a Good Estimator
- What Is the Margin of Error?
- Bias Reduction Techniques: Bootstrapping and Jackknifing
- Sample Size Determination
- Bayesian Statistical Inference: An Introduction

3.1 What is probability?

Probability is, in essence, the mathematics of chance. How sure can we be that a particular event is going to happen?

When we go to bed at night, we can be sure that the sun will rise the next morning. If today is Monday, then we can be certain that tomorrow is Tuesday. The chance that these events will occur is 100%. The probability of these events is 1.

Some events could possibly happen, but it is unlikely. For example: The chance that you win the lotto if you have bought a ticket is very slim, but it could happen. The chance that a lion is seen in the middle of town is very low, but it could happen if the circus was in town. The probability of these events lies between 0 and 1. A probability of ½ means a 50-50 chance of it happening.

Then you get some events that will never happen. For example: the chance of throwing a 10 on a normal die is 0%. A die only goes up to 6. The probability of this is 0.
3.2 The probability of a single event

Probability of a single event = Number of ways in which an event can occur
Number of possible outcomes

Sipho’s house has 1 lounge, 1 bathroom, 1 kitchen and 2 bedrooms.
What is the probability that Sipho is in the kitchen?
Probability = Number of ways in which an event can occur
Number of possible outcomes
= 1/5

What is the probability that Sipho is in a bedroom?
Probability = Number of ways in which an event can occur
Number of possible outcomes
= 2/5

Please complete Activity 12
1. A vending machine at a college has the following items on sale: cool drink, chocolates, crisps and chewing gum. A survey showed that customers would have the following preferences:
   • Crisps: ¼
   • Cool drink: ½
   • Chewing gum: 1/20
   • Chocolate: 1/5
   a. Convert all fractions to a percentage.
   b. Which is the most popular item?
   c. Which is the least popular item?
2. Clive has 2 pairs of yellow socks, 3 pairs of red socks, 4 pairs of black socks and 6 pairs of white socks in his sock drawer. There is a power failure while he is getting dressed and he cannot see which colour socks he is putting on.
   a. What is the probability that he will put on red socks?
   b. What is the probability that he will put on yellow socks?
   c. What is the probability that he will put on white socks?
   d. What is the probability that he will put on black socks?

3.3 Compound events

Two or more occurrences happen independently of each other
Sometimes two or more events happen at the same time. Let us take the example of Clive and his coloured socks again. This time, Clive is also trying to put on a T-shirt.
He has the following items:

<table>
<thead>
<tr>
<th>Socks</th>
<th>T-shirts</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 pairs yellow</td>
<td>3 yellow</td>
</tr>
<tr>
<td>3 pairs red</td>
<td>6 red</td>
</tr>
<tr>
<td>4 pairs black</td>
<td>6 black</td>
</tr>
<tr>
<td>6 pairs white</td>
<td>4 white</td>
</tr>
<tr>
<td>Total</td>
<td>15</td>
</tr>
</tbody>
</table>
What is the probability of Clive putting on a yellow shirt and yellow socks?

We work out each probability separately and then multiply them to get the total probability.

\[
P(\text{yellow socks}) = \frac{2}{15} \\
P(\text{yellow T-shirt}) = \frac{3}{20} \\
P(\text{socks and shirt}) = \frac{2}{15} \times \frac{3}{20} = \frac{1}{50}
\]

What is the probability that he will wear black socks and a red shirt?

\[
P(\text{black socks}) = \frac{4}{15} \\
P(\text{red T-shirt}) = \frac{6}{20} \\
P(\text{socks and shirt}) = \frac{4}{15} \times \frac{6}{20} = \frac{2}{25}
\]

A drawer contains 3 red paperclips, 4 green paperclips, and 5 blue paperclips. One paperclip is taken from the drawer and then replaced. Another paperclip is taken from the drawer. What is the probability that the first paperclip is red and the second paperclip is blue?

Because the first paper clip is replaced, the sample space of 12 paperclips does not change from the first event to the second event. The events are independent.

\[
P(\text{red then blue}) = P(\text{red}) \times P(\text{blue}) = \frac{3}{12} \times \frac{5}{12} = \frac{15}{144} = \frac{5}{48}
\]

When two or more occurrences happen independently of each other, then we multiply the probabilities.

### Complementary events

In a parking lot there are 20 vehicles. 16 of the vehicles are cars.

What is the probability of the first vehicle leaving the parking lot being a car?

\[
P(\text{car}) = \frac{16}{20} = \frac{4}{5}
\]

What is the probability of the first vehicle leaving the parking lot NOT being a car?

\[
P(\text{Not car}) = \frac{4}{20} = \frac{1}{5}
\]

Notice that \(\frac{4}{5} + \frac{1}{5} = \frac{5}{5} = 1\)

\[
P(\text{event}) + P(\text{not event}) = 1
\]

### Either one or the other event happening

You have a deck of 52 playing cards. The jokers have been removed. A card is drawn from the deck. What is the probability of the card being red or a jack.

Quite clearly, \(P(\text{red}) = \frac{13}{52}\) and \(P(\text{Jack}) = \frac{4}{52}\)

But there are two red jacks in the deck!! You cannot count those twice. Their probability is \(\frac{2}{52}\).

Thus we use the formula

\[
P(\text{red or Jack}) = P(\text{red}) + P(\text{Jack}) - P(\text{red and Jack}) \\
= \frac{13}{52} + \frac{4}{52} - \frac{2}{52} \\
= \frac{15}{52}
\]

If you need to calculate the probability of either one or another event, then you use the formula: \(P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y)\)
Putting it all together

A pair of dice is rolled. What is the probability that the sum of the numbers rolled is either 7 or 11?

- Six outcomes have a sum of 7:
  - (1,6), (2,5), (3,4), (4,3), (5,2), (6,1)
  - \( P(7) = \frac{6}{36} \)
- Two outcomes have a sum of 11:
  - (5,6), (6,5)
  - \( P(11) = \frac{2}{36} \)

The sum of the numbers cannot be 7 and 11 at the same time, so these events are mutually exclusive.

\[ P(7 \text{ or } 11) = P(7) + P(11) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9} \]

A pair of dice is rolled. What is the probability that the sum of the numbers rolled is either an even number or a multiple of 3?

- Of the 36 possible outcomes, 18 are even sums.
  - \( P(\text{even}) = \frac{18}{36} = \frac{1}{2} \)
- Sums of 3, 6, 9, and 12 are multiples of 3.
  - There are 12 sums that are multiples of 3.
  - \( P(\text{multiple of 3}) = \frac{12}{36} = \frac{1}{3} \)
  - However, some of these outcomes appear in both events.
- The sums that are even and a multiple of 3 are 6 and 12.
- There are 6 ordered pairs with these sums.
- \( P(\text{even AND a multiple of 3}) = \frac{6}{36} = \frac{1}{6} \)
- \( P(\text{even OR a multiple of 3}) = \frac{18}{36} + \frac{12}{36} - \frac{6}{36} = \frac{24}{36} = \frac{2}{3} \)

Please complete Activity 13

1. Sandile has 3 plastic forks, 5 metal forks, 4 plastic knives and 6 metal knives in his cutlery drawer. What is the probability that he will grab a plastic fork and a plastic knife from the drawer if he is not looking?
2. The probability of a woman having a baby boy is 50% and that of having a girl is also 50%. What is the probability that a woman who has three children will have three boys?
3. Lebo has 3 blue pens, 2 red pens, 5 black pens and 2 pencils in his pencil case.
   a. What is the probability that he takes a black pen?
   b. What is the probability that he takes an item that is not a black pen?
4. One card is drawn from a deck of 52 cards. What is the probability that the card will be
   a. red or an ace?
   b. a king of hearts?

My Notes ...

.............................................................
.............................................................
.............................................................
.............................................................
3.4 **Probability tree diagrams**

Suppose you toss a coin. It has a 50% chance of showing heads and a 50% chance of showing tails.

![Tree Diagram for Coin Toss]

**Calculating the probability of certain outcomes**

Suppose that you roll a die at the same time as tossing the coin. Now you end up with a large number of possibilities: You could have heads and 1, heads and 2 etc OR tails and 1, tails and 2 etc. To visualise this we expand the diagram above:

![Expanded Tree Diagram]

The probability of tossing heads AND rolling a six is:

\[ P(\text{heads AND } 6) = P(\text{heads}) \times P(6) \]

\[ = \frac{1}{2} \times \frac{1}{6} \]

\[ = \frac{1}{12} \]

To calculate the probability using a tree diagrams:

if you move along the arms, then you multiply the probabilities.

**Calculating the probability of multiple outcomes**

You want to know what the probability is of calculating the probability of tossing tails and an odd number

\[ P(\text{tails and odd number}) = \frac{1}{12} + \frac{1}{12} + \frac{1}{12} = \frac{3}{12} = \frac{1}{4} \]

To calculate the probability using a tree diagrams:

if you move down a column, then you add the probabilities.
Independent events

Two events are said to be independent if the result of the second event is not affected by the result of the first event.

If A and B are independent events, the probability of both events occurring is the product of the probabilities of the individual events

If A and B are independent events,

$$P(A \text{ and } B) = P(A) \times P(B).$$

A drawer contains 3 red paperclips, 4 green paperclips, and 5 blue paperclips. One paperclip is taken from the drawer and then replaced. Another paperclip is taken from the drawer. What is the probability that the first paperclip is red and the second paperclip is blue?

Because the first paper clip is replaced, the sample space of 12 paperclips does not change from the first event to the second event. The events are independent.

$$P(\text{red then blue}) = P(\text{red}) \times P(\text{blue}) = \frac{3}{12} \times \frac{5}{12} = \frac{15}{144} = \frac{5}{48}.$$ 

If the result of one event IS affected by the result of another event, the events are said to be dependent.

If A and B are dependent events, the probability of both events occurring is the product of the probability of the first event and the probability of the second event once the first event has occurred.

If A and B are dependent events,

and A occurs first,

$$P(\text{A and B}) = P(A) \times P(\text{B, once A has occurred}).$$

A drawer contains 3 red paperclips, 4 green paperclips, and 5 blue paperclips. One paperclip is taken from the drawer and is NOT replaced. Another paperclip is taken from the drawer. What is the probability that the first paperclip is red and the second paperclip is blue?

Because the first paper clip is NOT replaced, the sample space of the second event is changed. The sample space of the first event is 12 paperclips, but the sample space of the second event is now 11 paperclips. The events are dependent.

$$P(\text{red then blue}) = P(\text{red}) \times P(\text{blue}) = \frac{3}{12} \times \frac{5}{11} = \frac{15}{132} = \frac{5}{44}.$$ 

3.5 Qualities of a good estimator

A "Good" estimator is the one, which provides an estimate with the following qualities:

- **Unbiasedness**

An estimate is said to be an unbiased estimate of a given parameter when the expected value of that estimator can be shown to be equal to the parameter being
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture
NQF Level 4
Unit Standard No: 9015

estimated. For example, the mean of a sample is an unbiased estimate of the mean of the population from which the sample was drawn. Unbiasedness is a good quality for an estimate, since, in such a case, using weighted average of several estimates provides a better estimate than each one of those estimates. Therefore, unbiasedness allows us to upgrade our estimates. For example, if your estimates of the population mean \( \mu \) are say, 10, and 11.2 from two independent samples of sizes 20, and 30 respectively, then a better estimate of the population mean \( \mu \) based on both samples is \([20 (10) + 30 (11.2)] (20 + 30) = 10.75\).

- **Consistency**

  The standard deviation of an estimate is called the standard error of that estimate. The larger the standard error the more error in your estimate. The standard deviation of an estimate is a commonly used index of the error entailed in estimating a population parameter based on the information in a random sample of size \( n \) from the entire population.

  An estimator is said to be "consistent" if increasing the sample size produces an estimate with smaller standard error. Therefore, your estimate is "consistent" with the sample size. That is, spending more money to obtain a larger sample produces a better estimate.

- **Efficiency**

  An efficient estimate is one, which has the smallest standard error among all unbiased estimators.

  The "best" estimator is the one, which is the closest to the population parameter being estimated.

The above figure illustrates the concept of closeness by means of aiming at the centre for unbiased with minimum variance. Each dartboard has several samples: The first one has all its shots clustered tightly together, but none of them hit the centre. The second one has a large spread, but around the centre. The third one
is worse than the first two. Only the last one has a tight cluster around the centre, therefore has good efficiency.

If an estimator is unbiased, then its variability will determine its reliability. If an estimator is extremely variable, then the estimates it produces may not on average be as close to the population parameter as a biased estimator with small variance.

The following chart depicts the quality of a few popular estimators for the population mean $\mu$:

The Good Qualities of an Estimator for $\mu$

The widely used estimator of the population mean $\mu$ is $\bar{x} = \Sigma x_i/n$, where $n$ is the size of the sample and $x_1, x_2, x_3, \ldots, x_n$ are the values of the sample that have all of the above good properties. Therefore, it is a "good" estimator.

If you want an estimate of central tendency as a parameter for a test or for comparison, then small sample sizes are unlikely to yield any stable estimate. The mean is sensible in a symmetrical distribution as a measure of central tendency; but, e.g., with ten cases, you will not be able to judge whether you have a symmetrical distribution. However, the mean estimate is useful if you are trying to estimate the population sum, or some other function of the expected value of the distribution. Would the median be a better measure? In some distributions (e.g., shirt size) the mode may be better. If there are outliers, the median is better than the mean as a measure of central tendency."
3.6 What is the margin of error?

Estimation is the process by which sample data are used to indicate the value of an unknown quantity in a population. Results of estimation can be expressed as a single value, known as a point estimate; or a range of values, referred to as a confidence interval. Whenever we use point estimation, we calculate the margin of error associated with that point estimate. For example, for the estimation of the population proportion, by the means of sample proportion (p), the margin of error is calculated often as follows:

$$\pm 1.96 \left[ p(1-p)/n \right]^{1/2}$$

In newspapers and television reports on public opinion polls, the margin of error often appears in a small font at the bottom of a table or screen. However, reporting the amount of error only, is not informative enough by itself, what is missing is the degree of the confidence in the findings. The more important missing piece of information is the sample size n; that is, how many people participated in the survey, 100 or 100000? By now, you know well that the larger the sample size the more accurate is the finding, right? The reported margin of error is the margin of "sampling error". There are many non-sampling errors that can and do affect the accuracy of polls. Here we talk about sampling error. The fact that sub-groups might have sampling error larger than the group, one must include the following statement in the report: "Other sources of error include, but are not limited to, individuals refusing to participate in the interview and inability to connect with the selected number. Every feasible effort was made to obtain a response and reduce the error, but the reader (or the viewer) should be aware that some error is inherent in all research."

If you have a yes/no question in a survey, you probably want to calculate a proportion P of Yes's (or No's). In a simple random sample survey, the variance of p is p(1-p)/n, ignoring the finite population correction, for large n, say over 30. Now a 95% confidence interval is

$$p - 1.96 \left[ p(1-p)/n \right]^{1/2}, \quad p + 1.96 \left[ p(1-p)/n \right]^{1/2}.$$  

A conservative interval can be calculated, since p(1-p) takes its maximum value when p = 1/2. Replace 1.96 by 2, put p = 1/2 and you have a 95% conservative confidence interval of 1/\sqrt{n}. This approximation works well as long as p is not too close to 0 or 1. This useful approximation allows you to calculate approximate 95% confidence intervals.

For continuous random variables, such as the estimation of the population mean \( \mu \), the margin of error is calculated often as follows:

$$\pm 1.96 \frac{S}{\sqrt{n}}.$$  

The margin of error can be reduced by one or a combination of the following strategies:

1. Decreasing the confidence in the estimate – an undesirable strategy since confidence relates to the chance of drawing the wrong conclusion (i.e., increases the Type II error).
2. Reducing the standard deviation – something we cannot do since it is usually a static property of the population.
3. Increasing the sample size – this provides more information for a better decision.

### 3.7 Bias reduction techniques: bootstrapping and jackknifing

Some inferential statistical techniques do not require distributional assumptions about the statistics involved. These modern non-parametric methods use large amounts of computation to explore the empirical variability of a statistic, rather than making a priori assumptions about this variability, as is done in the traditional parametric t- and z- tests.

**Bootstrapping**

Bootstrapping method is to obtain an estimate by combining estimators to each of many sub-samples of a data set. Often M randomly drawn samples of T observations are drawn from the original data set of size n with replacement, where T is less than n.

**Jackknife estimator:**

A jackknife estimator creates a series of estimate, from a single data set by generating that statistic repeatedly on the data set leaving one data value out each time. This produces a mean estimate of the parameter and a standard deviation of the estimates of the parameter.

- **Monte Carlo simulation**

Monte Carlo simulation allows for the evaluation of the behaviour of a statistic when its mathematical analysis is intractable. Bootstrapping and jackknifing allow inferences to be made from a sample when traditional parametric inference fails. These techniques are especially useful to deal with statistical problems, such as small sample size, statistics with no well-developed distributional theory, and parametric inference condition violations. Both are computer intensive. Bootstrapping means you take repeated samples from a sample and then make statements about a population. Bootstrapping entails sampling-with-replacement from a sample. Jackknifing involves systematically doing n steps, of omitting 1 case from a sample at a time, or, more generally, n/k steps of omitting k cases; computations that compare "included" vs. "omitted" can be used (especially) to reduce the bias of estimation. Both have applications in reducing bias in estimations.

- **Re-sampling**

Re-sampling including the bootstrap, permutation, and other non-parametric tests – is a method for hypothesis testing, confidence limits, and other applied problems in statistics and probability. It involves no formulas or tables. Following the first publication of the general technique (and the bootstrap) in 1969 by Julian Simon and subsequent independent development by Bradley Efron, re-sampling has become an alternative approach for testing hypotheses. There are other findings: "The bootstrap started out as a good notion in that it presented, in theory, an elegant statistical procedure that was free of distributional conditions. In practice the bootstrap technique doesn't work very
well, and the attempts to modify it make it more complicated and more confusing than the parametric procedures that it was meant to replace." While re-sampling techniques may reduce the bias, they achieve this at the expense of increase in variance. The two major concerns are:

- The loss in accuracy of the estimate as measured by variance can be very large.
- The dimension of the data affects drastically the quality of the samples and therefore the estimates.

## 3.8 Sample size determination

At the planning stage of a statistical investigation, the question of sample size (n) is critical. This is an important question therefore it should not be taken lightly. To take a larger sample than is needed to achieve the desired results is wasteful of resources, whereas very small samples often lead to what are no practical use of making good decisions. The main objective is to obtain both a desirable accuracy and a desirable confidence level with minimum cost.

Students sometimes ask me, what fraction of the population do you need for good estimation? I answer, "It's irrelevant; accuracy is determined by sample size." This answer has to be modified if the sample is a sizable fraction of the population. The confidence level of conclusions drawn from a set of data depends on the size of the data set. The larger the sample, the higher is the associated confidence. However, larger samples also require more effort and resources. Thus, your goal must be to find the smallest sample size that will provide the desirable confidence.

For an item scored 0 or 1, for no or yes, the standard error (SE) of the estimated proportion p, based on your random sample observations, is given by:

\[
SE = \left[ \frac{p(1-p)}{n} \right]^{1/2}
\]

where p is the proportion obtaining a score of 1, and n is the sample size. This SE is the standard deviation of the range of possible estimate values.

The SE is at its maximum when p = 0.5, therefore the worst-case scenario occurs when 50% are yes, and 50% are no.

Under this extreme condition, the sample size, n, can then be expressed as the largest integer less than or equal to:

\[
n = \frac{0.25}{SE^2}
\]

To have some notion of the sample size, for example for SE to be 0.01 (i.e. 1%), a sample size of 2500 will be needed; 2%, 625; 3%, 278; 4%, 156, 5%, 100.

Note, incidentally, that as long as the sample is a small fraction of the total population, the actual size of the population is entirely irrelevant for the purposes of this calculation.

### Pilot Studies

When the needed estimates for sample size calculation is not available from an existing database, a pilot study is needed for adequate estimation with a given precision. A pilot, or preliminary, sample must be drawn from the population, and the statistics computed from this sample are used in determination of the sample size. Observations used in the pilot sample may be counted as part of the final sample, so that the computed sample size minus the pilot sample size is the number of observations needed to satisfy the total sample size requirement.
**Sample Size with Acceptable Absolute Precision:**
The following presents the widely used method for determining the sample size required for estimating a population mean and proportion. Let us suppose we want an interval that extends unit on either side of the estimator. We can write

\[ \text{Absolute Precision} = (\text{reliability coefficient}) \times (\text{standard error}) = \frac{Z}{2} \left( \frac{S}{n^{1/2}} \right) \]

Suppose, based on a pilot sample of size \( n \), the estimated proportion is \( p \), then the required sample size with the absolute error size not exceeding \( \Delta \), with 1-confidence is:

\[ \left[ t^2 \frac{n}{2} \left( p(1-p) \right) \right] / \left[ t^2 \frac{p(1-p)}{2} - 2(n-1) \right] \]

where \( t = t \) being the value taken from the t-table with parameter d.f. = \( n-1 \), corresponding to the desired 1-confidence interval.

For large pilot sample sizes \( n \), say over 30, the simplest sample size determinate is:

\[ \left[ \left( \frac{Z}{2} \right)^2 S^2 \right] / 2 \quad \text{for the Mean} \]
\[ \left[ \left( \frac{Z}{2} \right)^2 p(1-p) \right] / 2 \quad \text{for the proportion} \]

where \( \Delta \) is the desirable margin of error (i.e., the absolute error), which is the half-length of the confidence interval with 100(1-)\% confidence interval.

**Sample Size with Acceptable Type I and Type II Errors:**
One may use the following sample size determinate, which is based on the size of type I and Type II errors:

\[ 2(Z_{\alpha/2} + Z_{\beta/2})^2 S^2 / \delta^2 \]

where \( \alpha \) and \( \beta \) are the desirable type I, and type II errors, respectively. \( S^2 \) is the variance obtained from the pilot run, and \( \delta \) is the difference between the null and alternative \((\mu_0 - \mu_a)\).

**Sample Size with Acceptable Relative Precision:**
You may use the following sample size determinate for a desirable relative error \( \Delta \) in \%, which requires an estimate of the coefficient of variation (CV in \%) from a pilot sample with size over 30:

\[ \left[ (Z_{\alpha/2})^2 CV^2 \right] / \Delta^2 \]

**Sample Size Based on the Null and an Alternative:**
One may use power of the test to determine the sample size. The functional relation of the power and the sample size is known as the operating characteristic curve. On this curve, as sample size increases, the power function increases rapidly. Let \( \delta \) be such that:

\( \mu_0 = \mu_0 + \delta \)

is an alternative to represent departure from the null hypothesis. We wish to be reasonably confident to find evidence against the null, if in fact the particular alternative holds. That is, the type error \( \beta \), is the probability of failing to find evidence at least at level of \( \alpha \), when the alternative holds. This implies

Required sample size = \( \left( z_1 + z_2 \right) S^2 / \delta^2 \)

Where: \( z_1 = \left| \text{mean} - \mu_0 \right| / SE \), \( z_2 = \left| \text{mean} - \mu_a \right| / SE \), the mean is the current estimate for \( \mu \), and \( S \) is the current estimate for \( \sigma \).

All of the above sample size determinates could also be used for estimating the mean of any unimodal population, with discrete or continuous random variables, provided the pilot run size \( n \) is larger than (say) 30.

In estimating the sample size, when the standard deviation is not known, instead of \( S^2 \) one may use 1/4 of the range for sample size over 30 as a "good" estimate.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life-related problems. 

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

for the standard deviation. It is a good practice to compare the result with IQR/1.349.

One may extend the sample size determination to other useful statistics, such as correlation coefficient (r) based on acceptable Type I and Type II errors:

\[2 + \left(\frac{(Z_{\alpha/2} + Z_{\beta/2}(1 - r^2))r}{1 - r^2}\right)^2\]

provided r is not equal to -1, 0, or 1.

The aim of applying any one of the above sample size determinates is at improving your pilot estimates at feasible costs.

### 3.9 Bayesian statistical inference: An introduction

Statistical inference describes the procedures by which we use the observed data to draw conclusions about the population from which the data came or about the process by which the data were generated. Our assumption is that there is an unknown process that generates the data we have and that this process can be described by a probability distribution, which, in turn, can be characterized by some unknown parameters. For instance, for a normal distribution the unknown parameters are \( \mu \) and \( \sigma^2 \).

Broadly speaking, statistical inference can be classified under two headings: classical inference and Bayesian inference. Classical statistical inference is based on two premises:

1. The sample data constitute the only relevant information.
2. The construction and assessment of the different procedures for inference are based on long-run behaviour under essentially similar circumstances.

In Bayesian inference we combine sample information with prior information. Suppose that we draw a random sample \( x_1, x_2, \ldots, x_n \) of size n from a normal population.

In statistical inference we take the sample mean \( \bar{x} \) as our estimate of \( \mu \). Its variance is \( \sigma^2 / n \). The inverse of this variance is known as the sample precision. Thus the sample precision is \( n / \sigma^2 \).

In the Bayesian inference we have prior information on \( \mu \). This is expressed in terms of a probability distribution known as the prior distribution. Suppose that the prior distribution is normal with mean \( \mu_0 \) and variance \( \sigma_0^2 \), that is, precision \( 1 / \sigma_0^2 \). We now combine this with the sample information to obtain what is known as the posterior distribution of \( \mu \). This distribution can be shown to be normal. Its mean is a weighted average of the sample mean and the prior mean, weighted by the sample precision and prior precision, respectively. Thus,

\[
\text{Posterior mean} = \frac{(W_1 \bar{x} + W_2 \mu_0)}{(W_1 + W_2)}
\]

\[
\text{Posterior variance} = \frac{1}{W_1 + W_2}
\]

where

\[W_1 = \text{Sample precision} = n/S^2, \text{ and } W_2 = \text{Prior precision} = n/\sigma_0^2\]

Also, the precision (or inverse of the variance) of the posterior distribution of \( \mu \) is \( W_1 + W_2 \), that is, the sum of the sample precision and prior precision. The posterior mean will lie between the sample mean and the prior mean. The posterior variance will be less than both the sample and prior variances.
In this course we do not discuss Bayesian inference because this would take us into a lot more detail than we intend to cover. However, the basic notion of combining the sample mean and prior mean in inverse proportion to their variances will be of interest while being useful.

Please complete Activity 14

1. This exercise lets you simulate playing up to three games of chance at once. The games are organized into three columns. You do not have to input values for all three games every time. If you do not set the probabilities for a game, the applet will ignore that game when it runs.

   - Two items at the top of the applet let you describe each of three games that can be simulated. You can choose the type of game: throwing a die, picking cards from a deck, spinning a spinner, flipping a coin, or another game. In the text boxes below the type of game, describe what wins. For example, if the game is flipping a coin, what wins could be "heads."

<table>
<thead>
<tr>
<th>Type of Game:</th>
<th>Die</th>
<th>Coin</th>
<th>Cards</th>
</tr>
</thead>
<tbody>
<tr>
<td>What Wins:</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: these boxes have NOTHING to do with the actual calculation! They are for you to record your notes about what is being simulated. In the next section, you yourself have to enter the correct theoretical probabilities based on your analysis.

Under Theoretical Probabilities, you must enter the probability as a fraction of winning each game you have described. For instance, if the game was flipping a coin with heads winning, you would expect to win one time for every two tries. So you would enter "1" in the upper box (numerator) and "2" in the lower box (denominator).
To simulate playing the game, choose running the game 20, 40, and then 60 times each. Calculate how many times you could win theoretically.

The results of playing the game are displayed in the **Experimental Probabilities** section.

**Experimental Probabilities**

<table>
<thead>
<tr>
<th>Wins</th>
<th>Tries</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**a.** How did your experimental data compare to your theoretical data?
**b.** What can you conclude between running the experiment 20, 40, and 60 times?
**c.** List the three games in terms of winning a particular game. Explain why you say so.
**d.** Draw a frequency distribution of your data (use data of 40 times).
**e.** Draw a histogram of your data (use data of 40 times).

2. Discuss the terms:
   - **Outcome**,  
   - **Probability**,  
   - **Theoretical probability**,  
   - **Experimental probability**,

3. Discuss the terms:
   - **Mutually Exclusive versus Independent Events**  
   - **Qualities of a Good Estimat**  
   - **What Is the Margin of Error?**  
   - **Bootstrapping**  
   - **Jackknifing**  
   - **Bayesian Statistical Inference**

4. Discuss the different Sample Size Determination techniques

---

**Concept (SO 3)**

<table>
<thead>
<tr>
<th>Statistics generated from the data are interpreted meaningfully and interpretations are justified or critiqued.</th>
<th>I understand this concept</th>
<th>Questions that I still would like to ask</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics generated from the data are interpreted meaningfully and interpretations are justified or critiqued.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Assumptions made in the collection or generation of data and statistics are defined or critiqued appropriately. |                          |                                        |
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

<table>
<thead>
<tr>
<th>Tables, diagrams, charts and graphs are used or critiqued appropriately in the analysis and representation of data, statistics and probability values.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictions, conclusions and judgements are made on the basis of valid arguments and supporting data, statistics and probability models.</td>
</tr>
<tr>
<td>Evaluations of the statistics identify potential sources of bias, errors in measurement, potential uses and misuses and their effects.</td>
</tr>
</tbody>
</table>

My Notes ...

.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
.................................................................................................................................................................
## Glossary

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Array</td>
<td>Representing data when it has been sorted from the smallest to largest values. The data set is also referred to as being ranked.</td>
</tr>
<tr>
<td>Bias</td>
<td>A statistic is biased if, in the long run, it consistently over or underestimates the parameter it is estimating.</td>
</tr>
<tr>
<td>Biased sample</td>
<td>A biased sample is one in which the method used to create the sample results in samples that are systematically different from the population.</td>
</tr>
<tr>
<td>Census</td>
<td>The gathering of data on all the possible observations in a population.</td>
</tr>
<tr>
<td>Correlation</td>
<td>A single number that measures the strength of the linear association between two variables.</td>
</tr>
<tr>
<td>Data</td>
<td>The raw material used for statistical analyses. (see raw data)</td>
</tr>
<tr>
<td>Data set / Dataset</td>
<td>The collection or listing of all the data, either on paper or in electronic form e.g. in a spreadsheet.</td>
</tr>
<tr>
<td>Experimental probability</td>
<td>The chances of something happening, based on repeated testing and observing results. It is the ratio of the number of times an event occurred to the number of times tested. For example, to find the experimental probability of winning a game, one must play the game many times, then divide the number of games won by the total number of games played.</td>
</tr>
<tr>
<td>Grouped Data</td>
<td>Data that has been summarised into a frequency distribution.</td>
</tr>
<tr>
<td>Measurement</td>
<td>The process of representing attributes or aspects of items (events, things, people etc.) by numbers in a systematic way.</td>
</tr>
<tr>
<td>Outcome</td>
<td>Any one of the possible results of an experiment.</td>
</tr>
<tr>
<td>Outlier</td>
<td>A measurement not close to a measure of central tendency i.e. an atypical.</td>
</tr>
</tbody>
</table>
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

<table>
<thead>
<tr>
<th><strong>Parameter</strong></th>
<th>A measure describing a population attribute.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Percentile Rank</strong></td>
<td>The proportion of scores in a distribution that a specific score is greater than or equal to. For instance, if you received a score of 95 on a math test and this score was greater than or equal to the scores of 88% of the students taking the test, then your percentile rank would be 88. You would be in the 88th percentile.</td>
</tr>
<tr>
<td><strong>Population</strong></td>
<td>The entire collection of measurements that is the focus of concern.</td>
</tr>
<tr>
<td><strong>Probability</strong></td>
<td>The measure of how likely it is for an event to occur. The probability of an event is always a number between zero and 100%. The meaning (interpretation) of probability is the subject of theories of probability. However, any rule for assigning probabilities to events has to satisfy the axioms of probability.</td>
</tr>
<tr>
<td><strong>Random sample</strong></td>
<td>A set of items that have been drawn from a population in such a way that each time an item was selected, every item in the population had an equal opportunity to appear in the sample. I.e. a representative sample from a population. In practical terms, it is not so easy to draw a random sample. First, the only factor operating when a given item is selected, must be chance.</td>
</tr>
<tr>
<td><strong>Random Variable</strong></td>
<td>The different (random) values that a variable can take on at each measurement or observation.</td>
</tr>
<tr>
<td><strong>Range</strong></td>
<td>The highest value minus the lowest value in a dataset.</td>
</tr>
<tr>
<td><strong>Raw data</strong></td>
<td>The actual data values, observations or measurements that have been taken without any alterations, other than corrections.</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td>A subset of measurements taken from a population.</td>
</tr>
<tr>
<td><strong>Sampling</strong></td>
<td>The process of selecting a representative subset of observations from a population to determine the characteristics (i.e. the population parameters) of the random variable under study. Sampling produces data concerning a...</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>population</td>
<td>population by observing characteristics within a sample taken from the population.</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>The square root of the variance.</td>
</tr>
<tr>
<td>Statistics</td>
<td>The study of how to collect, organise, analyse, and interpret numerical information. The word statistic is also used to refer to a measure found from analysing a sample, such as the mean.</td>
</tr>
<tr>
<td>Surveys</td>
<td>Used to gather information concerning people.</td>
</tr>
<tr>
<td>Theoretical probability</td>
<td>The chances of events happening as determined by calculating results that would occur under ideal circumstances. For example, the theoretical probability of rolling a 4 on a four-sided die is 1/4 or 25%, because there is one chance in four to roll a 4, and under ideal circumstances one out of every four rolls would be a 4.</td>
</tr>
<tr>
<td>Ungrouped Data</td>
<td>Raw data values that have not been summarised in any way.</td>
</tr>
<tr>
<td>Variability</td>
<td>The spread of dispersion of test scores, most often expressed as a standard deviation. (See Standard Deviation)</td>
</tr>
<tr>
<td>Variable</td>
<td>Any characteristic being measured or observed.</td>
</tr>
<tr>
<td>Variance</td>
<td>The variance is a measure of how spread out a distribution is. It is computed as the average squared deviation of each number from its mean. The square of the standard deviation.</td>
</tr>
</tbody>
</table>
Checklist for practical assessment ...

Use the checklist below to help you prepare for the part of the practical assessment when you are observed on the attitudes and attributes that you need to have to be found competent for this learning module.

<table>
<thead>
<tr>
<th>Observations</th>
<th>Answer Yes or No</th>
<th>Motivate your Answer (Give examples, reasons, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can you identify problems and deficiencies correctly?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are you able to work well in a team?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do you work in an organised and systematic way while performing all tasks and tests?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are you able to collect the correct and appropriate information and / or samples as per the instructions and procedures that you were taught?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are you able to communicate your knowledge orally and in writing, in such a way that you show what knowledge you have gained?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Can you base your tasks and answers on scientific knowledge that you have learnt?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are you able to show and perform the tasks required correctly?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are you able to link the knowledge, skills and attitudes that you have learnt in this module of learning to specific duties in your job or in the community where you live?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

♦ The assessor will complete a checklist that gives details of the points that are checked and assessed by the assessor.
♦ The assessor will write commentary and feedback on that checklist. They will discuss all commentary and feedback with you.
♦ You will be asked to give your own feedback and to sign this document.
♦ **It will be placed together with this completed guide in a file as part of you portfolio of evidence.**
♦ The assessor will give you feedback on the test and guide you if there are areas in which you still need further development.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

```

Paperwork to be done ...

Please assist the assessor by filling in this form and then sign as instructed.

<table>
<thead>
<tr>
<th>Learner Information Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit Standard</td>
</tr>
<tr>
<td>Program Date(s)</td>
</tr>
<tr>
<td>Assessment Date(s)</td>
</tr>
<tr>
<td>Surname</td>
</tr>
<tr>
<td>First Name</td>
</tr>
<tr>
<td>Learner ID / SETA Registration Number</td>
</tr>
<tr>
<td>Job / Role Title</td>
</tr>
<tr>
<td>Home Language</td>
</tr>
<tr>
<td>Gender: Male: Female:</td>
</tr>
<tr>
<td>Race: African: Coloured: Indian/Asian: White:</td>
</tr>
<tr>
<td>Employment: Permanent: Non-permanent:</td>
</tr>
<tr>
<td>Disabled Yes: No:</td>
</tr>
<tr>
<td>Date of Birth</td>
</tr>
<tr>
<td>ID Number</td>
</tr>
<tr>
<td>Contact Telephone Numbers</td>
</tr>
<tr>
<td>Email Address</td>
</tr>
<tr>
<td>Postal Address</td>
</tr>
</tbody>
</table>
```

agriculture REPUBLIC OF SOUTH AFRICA

Version: 01  
Version Date: July 2006
Terms & Conditions

This material was developed with public funding and for that reason this material is available at no charge from the AgriSETA website (www.agriseta.co.za).

Users are free to produce and adapt this material to the maximum benefit of the learner.

No user is allowed to sell this material whatsoever.
Acknowledgements

- **Project Management:**
  M H Chalken Consulting
  IMPETUS Consulting and Skills Development

- **Donors:**
  Citrus Academy
  Boland College
  Weskus College

- **Developer:**
  Dr M Roets
  Mr J H P van der Merwe
  Cabeton Consulting

- **Authenticator:**
  Rural Integrated Engineering
  Prof P J Robbertse
  Ms D Naidoo
  Le Toit Management Consultants cc
  Mr M M Ratsaka
  Mr D N Cronje

- **Technical Editing:**
  Mr R H Meinhardt

- **OBE Formatting:**
  Ms P Prinsloo
  Ms B Enslin
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

- **Design:**
  Didacsa Design SA (Pty) Ltd

- **Layout:**
  Ms A du Plessis
  Ms N Matloa
**Purpose of the Unit Standard**

This Unit Standard is designed to provide credits towards the mathematical literacy requirement of the NQF at Level 4. The essential purposes of the mathematical literacy requirement are that, as the learner progresses with confidence through the levels, the learner will grow in:

- A confident, insightful use of mathematics in the management of the needs of everyday living to become a self-managing person.
- An understanding of mathematical applications that provides insight into the learner’s present and future occupational experiences and so develop into a contributing worker.
- The ability to voice a critical sensitivity to the role of mathematics in a democratic society and so become a participating citizen.

People credited with this unit standard are able to:

- Critique and use techniques for collecting, organising and representing data.
- Use theoretical and experimental probability to develop models, make predictions and study problems.
Critically interrogate and use probability and statistical models in problem solving and decision making in real-world situations.

**LEARNING ASSUMED TO BE IN PLACE AND RECOGNITION OF PRIOR LEARNING**

The credit value is based on the assumption that people starting to learn towards this unit standard are competent in Mathematical Literacy and Communications at NQF level 3.

**UNIT STANDARD RANGE**

This unit standard includes the requirement to:

Critique the selection of samples in terms of size and representativeness.

Identify features of distributions: symmetry and asymmetry, clusters and gaps, and possible outliers in data and consider their effects on the interpretation of the data. Critique the use of data from samples to estimate population statistics.

Apply an understanding of random phenomena to critique and interpret real life and work related situations.

Critique arguments based on probability in terms of an understanding of random behaviour and the law of large numbers (e.g. lottery `hot` numbers).

Demonstrate understanding of and determine probabilities for independent, disjoint and complementary events.

Judge or critique probability values.

Further range statements are provided for specific outcomes and assessment criteria as needed.

**Specific Outcomes and Assessment Criteria:**

**SPECIFIC OUTCOME 1**

Critique and use techniques for collecting, organising and representing data.

**OUTCOME NOTES**

Specific purposes include:

Determining trends in societal issues such as crime and health;

Identifying relevant characteristics of target groups such as age range, gender, socio-economic group, cultural belief, and performance;

Considering the attitudes or opinions of people on issues.

**OUTCOME RANGE**

Techniques include:

The formulation of questions in surveys to obtain data;

The methods and devices (e.g. tables of random numbers, calculators or computers) used to select random samples;

Different instruments and scales such as yes/no (dichotomous) and 5 point (Likert scales) and discrete and continuous variables;

Evaluation of data gathering techniques and of data collected so that faults and inconsistencies are identified;
Calculating measures of center and spread such as mean, median, mode, range; and variance; Using scatter plots and lines of best fit to represent the association between two variables; Correlation.

ASSESSMENT CRITERIA

ASSESSMENT CRITERION 1
1 Situations or issues that can be dealt with through statistical methods are identified correctly.

ASSESSMENT CRITERION 2
2. Appropriate methods for collecting, recording and organising (data are used so as to maximise efficiency and ensure the resolution of a problem or issue.

ASSESSMENT CRITERION 3
3. Data sources and databases are selected in a manner that ensures the representativeness of the sample and the validity of resolutions.

ASSESSMENT CRITERION 4
4. Activities that could result in contamination of data are identified and explanations are provided of the effects of contaminated data.

ASSESSMENT CRITERION 5
5. Data is gathered using methods appropriate to the data type and purpose for gathering the data.

ASSESSMENT CRITERION 6
6. Data collection methods are used correctly.

ASSESSMENT CRITERION 7
7 Calculations and the use of statistics are correct.

ASSESSMENT CRITERION 8
8. Graphical representations and numerical summaries are consistent with the data, are clear and appropriate to the situation and target audience.

ASSESSMENT CRITERION 9
9. Resolutions for the situation or issue are supported by the data and are validated in terms of the context.

SPECIFIC OUTCOME 2
Use theoretical and experimental probability to develop models.

OUTCOME NOTES
Use theoretical and experimental probability to develop models, make predictions and study problems.

OUTCOME RANGE
Performance in this specific outcome includes the requirement to:

Use the laws governing independent, complementary and mutually exclusive events.
Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

Primary Agriculture  
NQF Level 4  
Unit Standard No: 9015

Determine theoretical and experimental probabilities.

Use simulations (e.g. six sided spinners, random number generators in calculators or computers) for comparing experimental results (e.g. the rolling of a die) with mathematical expectations.

Compare experimental results with mathematical expectations using probability models.

**ASSESSMENT CRITERIA**

**ASSESSMENT CRITERION 1**
1. Experiments and simulations are chosen and/or designed appropriately in terms of the situation to be modelled.

**ASSESSMENT CRITERION 2**
2. Predictions are based on validated experimental or theoretical probabilities.

**ASSESSMENT CRITERION 3**
3. The results of experiments and simulations are interpreted correctly in terms of the real context.

**ASSESSMENT CRITERION 4**
4. The outcomes of experiments and simulations are communicated clearly.

**SPECIFIC OUTCOME 3**
Critically interrogate and use probability and statistical models.

**OUTCOME NOTES**
Critically interrogate and use probability and statistical models in problem solving and decision making in real world situations.

**OUTCOME RANGE**
Performance in this specific outcome includes, the requirement to:

Source and interpret information from a variety of sources including databases.

Manipulate data in different ways to support opposing conclusions.

Evaluate statistically based arguments and make recommendations and describe the use and misuse of statistics in society.

Make inferences about a population on the basis of a sample selected from it.

Make comparisons between predictions and actual occurrences.

**ASSESSMENT CRITERIA**

**ASSESSMENT CRITERION 1**
1. Statistics generated from the data are interpreted meaningfully and interpretations are justified or critiqued.

**ASSESSMENT CRITERION 2**
2. Assumptions made in the collection or generation of data and statistics are defined or critiqued appropriately.

**ASSESSMENT CRITERION 3**

3. Tables, diagrams, charts and graphs are used or critiqued appropriately in the analysis and representation of data, statistics and probability values.

**ASSESSMENT CRITERION 4**

4. Predictions, conclusions and judgements are made on the basis of valid arguments and supporting data, statistics and probability models.

**ASSESSMENT CRITERION 5**

5. Evaluations of the statistics identify potential sources of bias, errors in measurement, potential uses and misuses and their effects.

**ASSESSMENT CRITERION RANGE**

Effects on arguments, judgements, conclusions and ultimately the audience.

**UNIT STANDARD ACCREDITATION AND MODERATION OPTIONS**

Providers of learning towards this unit standard will need to meet the accreditation requirements of the GENFETQA.

Moderation Option:
The moderation requirements of the GENFETQA must be met in order to award credit to learners for this unit standard.

**UNIT STANDARD ESSENTIAL EMBEDDED KNOWLEDGE**

The following essential embedded knowledge will be assessed through assessment of the specific outcomes in terms of the stipulated assessment criteria. Candidates are unlikely to achieve all the specific outcomes, to the standards described in the assessment criteria, without knowledge of the listed embedded knowledge. This means that the possession or lack of the knowledge can be inferred directly from the quality of the candidate’s performance against the standards.

Methods for collecting, organising and analysing data

Measures of center and spread

Techniques for representing and evaluating statistics

Randomness, probability and association.

**Critical Cross-field Outcomes (CCFO):**

**UNIT STANDARD CCFO IDENTIFYING**

Identify and solve problems using critical and creative thinking:
Solve a variety of problems based on data, statistics and probability.

**UNIT STANDARD CCFO COLLECTING**

Collect, analyse, organise and critically evaluate information:
Gather, organise, evaluate and critically interpret data and statistics to make sense of situations.

UNIT STANDARD CCFO COMMUNICATING
Communicate effectively:
Use everyday language and mathematical language to represent data, statistics and probability and effectively communicate or critique conclusions.

UNIT STANDARD CCFO CONTRIBUTING
Use mathematics:
Use mathematics to critically analyse, describe and represent situations and to solve problems related to the life or work situations of the adult with increasing responsibilities.

UNIT STANDARD ASSESSOR CRITERIA
Assessors should keep the following general principles in mind when designing and conducting assessments against this unit standard:

Focus the assessment activities on gathering evidence in terms of the main outcome expressed in the title to ensure assessment is integrated rather than fragmented. Remember we want to declare the person competent in terms of the title. Where assessment at title level is unmanageable, then focus assessment around each specific outcome, or groups of specific outcomes.

Make sure evidence is gathered across the entire range, wherever it applies. Assessment activities should be as close to the real performance as possible, and where simulations or role-plays are used, there should be supporting evidence to show the candidate is able to perform in the real situation.

Do not focus the assessment activities on each assessment criterion. Rather make sure the assessment activities focus on outcomes and are sufficient to enable evidence to be gathered around all the assessment criteria.

The assessment criteria provide the specifications against which assessment judgements should be made. In most cases, knowledge can be inferred from the quality of the performances, but in other cases, knowledge and understanding will have to be tested through questioning techniques. Where this is required, there will be assessment criteria to specify the standard required.

The task of the assessor is to gather sufficient evidence, of the prescribed type and quality, as specified in this unit standard, that the candidate can achieve the outcomes again and again and again. This means assessors will have to judge how many repeat performances are required before they believe the performance is reproducible.

All assessments should be conducted in line with the following well documented principles of assessment: appropriateness, fairness, manageability, integration into work or learning, validity, direct, authentic, sufficient, systematic, open and consistent.

UNIT STANDARD NOTES
N/A