## UNIT STANDARD 9015

#### Unit Standard Title

Apply knowledge of statistics and probability to critically interrogate and effectively communicate findings on life related problems

#### NQF Level

4

#### Credits

6

#### Purpose

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

There is open access to this unit standard. Learners should be competent in Communication and Mathematical Literacy 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.

#### Specific Outcomes and Assessment Criteria

**Specific Outcome 1**: Critique and use techniques for collecting, organising and representing data

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.

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 (Liked 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 centre 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**

* Situations or issues that can be dealt with through statistical methods are identified correctly
* Appropriate methods for collecting, recording and organising data are used so as to maximise efficiency and ensure the resolution of a problem or issue
* 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

**Specific Outcome 2**: Use theoretical and experimental probability to develop models, make predictions and study problems

Performance in this specific outcome includes the requirement to:

* Use the laws governing independent, complementary and mutually exclusive events.
* 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**

* Experiments and simulations are chosen and/or designed appropriately in terms of the situation to be modelled.
* Predictions are based on validated experimental or theoretical probabilities
* The results of experiments and simulations are interpreted correctly in terms of the real context
* The outcomes of experiments and simulations are communicated clearly

**Specific Outcome 3:** Critically interrogate and use probability and statistical models in problem solving and decision making in real world situations

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**

* Statistics generated from the data are interpreted meaningfully and interpretations are justified or critiqued.
* Assumptions made in the collection or generation of data and statistics are defined or critiqued appropriately.
* Tables, diagrams, charts and graphs are used or critiqued appropriately in the analysis and representation of data, statistics and probability values
* Predictions, conclusions and judgements are made on the basis of valid arguments and supporting data, statistics and probability models
* Evaluations of the statistics identify potential sources of bias, errors in measurement, potential uses and misuses and their effects: Effects on arguments, judgements, conclusions and ultimately the audience

#### Unit Standard Essential Embedded Knowledge

* Methods for collecting, organising and analysing data
* Measures of centre and spread
* Techniques for representing and evaluating statistics
* Randomness, probability and association.

#### Critical Cross-field Outcomes (CCFO)

* Identify and solve problems using critical and creative thinking: Solve a variety of problems based on data, statistics and probability.
* Collect, analyse, organise and critically evaluate information: Gather, organise, evaluate and critically interpret data and statistics to make sense of situations.
* Communicate effectively: Use everyday language and mathematical language to represent data, statistics and probability and effectively communicate or critique conclusions
* 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

|  |
| --- |
| Programme Structure And Strategy 9015 |
| **UNIT STANDARD 9016 Level 4** |
| **Credits 6: notional hours 60, classroom training 18 hours, workplace hours 42** |
| **Programme Outcomes** | **Page2** | **Outcomes**  | **Time Allocation** | **Delivery strategy** | **EEK** | **CCFO** | **Formative Assessment** | **Summative Assessm: total time 42 hours****Questionnaire and workplace assignments** |
| INTRODUCTION | 10 |   | Introductory activities: 60 minutes | Discussion  |  |  |  |  |
| Programme Overview | 10 |   |   |   |   |   |
| Personal objectives and expectations | 11 |   |   |   |   |   |
| **UNIT STANDARD 9015** | 165 |   |   |   |   |   |
| **Section 1: COLLECT ORGANISE AND REPRESENT DATA** | 169 | SO1 | 155 minutes | Discussion & slide show |   | Identify and solve problems  |   | QuestionnaireWorkplace: Do a survey about people’s feelings about the 2010 Soccer World Cup that was hosted by South Africa. Analyse the questionnaires and display your findings in a graph  |
| The Role of Statistics when processing data | 170 | AC1, 2, 3 |   |   |   |   |
| The Use Of Statistics In Work Or Every Day Life | 171 | AC4, 5 |   |   | Collect, analyseorganise  |   |
| Research | 171 | AC6, 7 |   |   |   |   |
| Sources of information | 173 | AC8, 9 |   |   |   |   |   |
| Formative Assessment 1 | 176 |   | 15 minutes |   |   | Individual activity: Secondary information |   |
| Formative Assessment 2 | 177 |   | 15 minutes |   |   | Individual activity: Evaluate Secondary information |   |
| Conducting A Survey (Research) | 177 |   |   |   |   |   |   |
| Formative assessment 3 | 178 |   | 15 minutes |   |   | Group activity: Research topic |   |
| Formative Assessment 4 | 182 |   | 15 minutes |   |   | Individual activity: Segment customers |   |
| Formative Assessment 5 | 187 |   | 15 minutes |   |   | Individual activity: Explain terms |   |
| Formative Assessment 6 | 188 |   | 15 minutes |   |   | Individual activity: Random digits |   |
| Formative Assessment 7 | 191 |   | 15 minutes |   |   | Individual activity: Statistics |   |
| Formative assessment 8 | 200 |   | 15 minutes |   |   | Individual activity: Survey |   |
| Formative Assessment 9 | 202 |   | 15 minutes |   |   | Individual activity: Likert Scale |   |
| Formative assessment 10 | 204 |   | 15 minutes | N/A |   | Group activity: Discuss research |   |
| Formative assessment 11 | 208 |   | 15 minutes |   |   | Group activity: List of questions |   |
| Formative assessment 12 | 209 |   | 15 minutes |   |   | Group activity: Pilot survey |   |
| Formative assessment 13 | 210 |   | 15 minutes |   |   | Group activity: Replies from survey |   |
| **Section 2: MODELS PREDICTIONS PROBLEMS** | 211 | SO2 | 155 minutes | Discussion & slide show |   |   |   |   |
| Drawing Conclusions from Data | 212 | AC1, 2, 3, 4 |   |   | Communicate |   |  You have to analyse and interpret staff attendance reports generated by your department / section. Draw conclusions about staff absenteeism based on the reports. Make predictions and recommendations about staff absenteeism based on the reports. Draw up graphs to illustrate the current situation and to illustrate your predictions and recommendations |
| Formative assessment 14 | 216 |   | 15 minutes |   |   | Group activity: Predictions |   |
| Formative assessment 15 | 219 |   | 15 minutes |   |  | Group activity: Probabilities |   |
| Formative assessment 16 | 219 |   | 15 minutes |   |   | Group activity: Probable outcomes |   |
| **Section 3: PROBABILITY AND STATISTICAL MODELS** | 221 | SO3 | 155 minutes | Discussion & slide show |   |   |   |  QuestionnaireWorkplace: Find out what US$ exchange rate, or the gold price or the price of oil or the platinum price was for the last six months. Use one price or exchange rate per week and plot the fluctuations on a line chart, indicating only monthly fluctuations  |
| Analyse The Data | 222 | AC1  |   |   | Use mathe-matics |   |
| Formative assessment 17 | 222 |   | 15 minutes |   |   | Individual activity: Analyses |  QuestionnaireWorkplace: You have to do research about the tyres one of the vehicle uses in a period of 6 months. You may also do research about uniform requisitionWrite a report to your supervisor, using the format approved in your organisation |
| Formative assessment 18 | 223 |   | 15 minutes | N/A |   | Group activity: Analyses |
| Determining Trends Using Statistics | 223 | AC2 |   |   |   |   |
| Describing data | 224 | AC3 |   |   |   |   |
| Formative assessment 19 | 227 |   | 15 minutes |   |   | Individual activity: Mean, Median & Mode |   |
| Formative assessment 20 | 229 |   | 15 minutes |   |   | Individual activity: Table |   |
| Graphical Representation Of Data | 230 | AC4  |   |   |   |   |  QuestionnaireWorkplace: from first activity - Analyse the questionnaires and display your findings in a graph |
| Formative assessment 21 | 232 |   | 15 minutes |   |   | Individual activity: Calculations |
| Graphing data | 235 | AC5 |   |   |   |   |
| Formative assessment 22 | 237 |   | 15 minutes |   |   | Individual activity: Chart |
| Formative assessment 23 | 240 |   | 15 minutes |   |   | Individual activity: Column Chart |
| Formative assessment 24 | 244 |   | 15 minutes |   |   | Individual activity: Chart |
| Formative assessment 25 | 245 |   | 15 minutes |   |   | Individual activity: draw chart |
| Comparing (Correlation) Of Data | 248 |   |   |   |   |   | QuestionnaireWorkplace: Compare the actual and budgeted expenses from the table in a graphShow the expenses in this table in a pie chartAttach the scatter plot you did in your summative assessment to this assessment |
| Formative assessment 26 | 249 |   | 15 minutes |   |   | Group activity: Correlation plot |   |
| Summarising Data | 249 |   |   |   |   |   |   |
| Formative assessment 27 | 250 |   | 15 minutes |   |   | Group activity: Histogram |   |
| **Revision**  |  |  | **60 minutes** |   |   |   |   |   |
| **Summative questionnaire** |  |  | **90 minutes** |   |   |   |   |   |
| **Notional Hours** |  |  | **18** |   |   |   |   | **42** |

# COLLECT ORGANISE AND REPRESENT DATA

#### Specific outcome

Critique and use techniques for collecting, organising and representing data

#### Assessment criteria

* Situations or issues that can be dealt with through statistical methods are identified correctly
* Appropriate methods for collecting, recording and organising data are used so as to maximise efficiency and ensure the resolution of a problem or issue
* 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

## The Role of Statistics when processing data

What is statistics? Most of us seem to become startled or frightened when the word ‘statistics’ is mentioned. This should not be so. We all use statistics in our daily lives. It is a natural part of ourselves that has allowed us to survive from the days of living in the dark caves of our ancestors to exploring our moon, other planets as well as the galaxy.

*‘But statistics means working with numbers and I am not good with numbers’* you may say. It is true that statisticians work with numerical facts that they call data. However, these numerical facts are simply ways of

* describing,
* grouping and
* summarizing information so that it is more easily understood.

Keep in mind that all numbers used in this module are fictitious, that is, they are numbers I invented myself. If I use numbers that are really true, I will let you know. So if I say something has risen or fallen by 20%, I made that value up just as an example.

Here are a few examples that you undoubtedly have heard and may have some idea of what they mean. (Remember to ignore the actual numbers!)

* Think of our sportsmen and sportswomen. We discuss their average runs at the cricket crease, bowling rates and the athlete’s best, worst and average times on the track. We know the average number of goals a striker makes in a season. How do we obtain this information? Why do we say one player is better than another? Can we compare a gymnast with a cyclist?
* We hear the weather forecaster say that there is a 20% probability of rain today. Perhaps based on this information we decide not to take an umbrella with us. If the weather forecaster said that the probability of rain today is 80%, would you leave the umbrella behind?
* The government announces that the rate of crime is down by 20%, inflation is up by 10%, unemployment is down to 28% and the CPI (Consumer Price Index) has risen 3 percentage points on a year-to-year basis. What does all this mean? How do they get this information? Are these figures accurate?
* You have heard that smoking is linked to lung cancer and that HIV is linked to AIDS and that the evidence for these statements is ‘statistical’. What kind of evidence is ‘statistical’ evidence?
* A medical researcher claims that taking a certain tonic reduces the risk of heart attack. How can an experiment be designed to prove or disprove this statement? What is risk?
* You use statistics when you cross the street or drive a vehicle. When crossing the street you look both ways, see vehicles approaching, estimate their speeds, estimate your chance of making it across the street without being run over then you move or stay. If you move you have already decided how fast you will move based on the information you have just gathered and analysed. You behave similarly when in a vehicle. Our ancestors behaved similarly when looking for food.
* If you are a hunter or a sport-shooting enthusiast, you know how to aim and shoot your firearm. You also know that you rarely, if ever, hit the target in exactly the same place with every shot you take. Why is there a slight variation where the bullet hits the target? Why are the bullet holes on my target scattered, or spread, more than on another shooters target?

At this point you may have a few questions that you would like answered.

* What exactly do these figures mean?

?

* How did they (sport announcer, weather forecaster, government, researcher) obtain this information?
* How accurate is the information?
* What does it all mean?

We humans have survived for many thousands of years simply because we are born with the ability to take in information, organise and analyse it and then draw conclusions based on our analysis.

Information

This is all that statistics is about:

* obtain information,
* organize and analyse the information
* and then draw conclusions from this information

The goal, however, of statistics is to provide insight by using numbers. In fact, the information usually contains some uncertainty but statistical thinking can deal with it.

Every discipline has developed its own language or terminology over time. Statistics is no different and many words and terms are used with specific meanings attached to them. Some of the same terms are also used daily by non-statisticians but in a loose, semi-defined manner. Statisticians use words to express specific ideas.

Information

For instance, statisticians use the word ‘data’ where ‘information’ has been used in this introduction. The word ‘data’ is almost always used in the plural for statistical work and this module uses ‘data’ as a result.

## The Use Of Statistics In Work Or Every Day Life

Statistics is the collection and analysis of numerical data in large quantities. This means that you gather information about a subject and then you analyse the information or data, so that you can distinguish trends. It is a very useful and easy way to “see” the story the numbers are telling.

Every time before an election, one of the organisations, Markinor, who collect and analyse data, will tell us before the election which political party will win the election and by how big a margin they will win the election. This is an example of gathering information and then analysing the information in order to find out what the trends are.

In the workplace, you can gather information about absenteeism at the workplace, how much fuel your vehicles use, how many employees are off sick during winter, how much stationery is used by the administration department, etc. Once you have the information, you can analyse it to find out what the trend is.

## Research

Research is defined as all activities that provide information to guide business, societal and life decisions. Research is an information gathering activity that is intended to guide strategic or operational business, societal and life decisions about target groups, competitive strategies, etc.

### Why would you want to do research?

Issues/questions that are commonly addressed by research in a business:

4 C’s analysis:

customers

competitors

company

climate

* Customers (Customer analysis)
* Competitors (Competitive analysis)
* Company (Operational analysis)
* Climate (Environmental analysis)

#### Customer analysis

The following questions are common in customer analysis:

* How big is the existing market?
* How big is the potential market?
* How fast is the market growing?

?

* What are the buyer’s background characteristics?
* How and why do buyers use the product?
* How and where is the product bought?
* How brand loyal are buyers?
* What market segments exists, and how large are the various segments?

#### Competitive analysis

In addition to analysing customers, market research may be used to describe your organisation’s competitive position in the market. Relevant questions may include:

* What market share do you and various competitors hold?
* What future sales do you forecast?
* What are the awareness levels?
* How do buyers look at the different brands?
* What are the repurchase rates for the various brands?
* How satisfied are the customers with the various brands?
* What are your competitors’ resources and strategies?

#### Operational analysis

Questions that arise under the heading of operational analysis include:

* How effective is your distribution?
* How effective is your advertising?
* How effective are your sales promotions?
* How effective are your sales people?
* How effective are your pricing strategies?
* How might consumers respond to product changes?
* How might buyers respond to a new product?

### The value of research

Research derives its value from helping managers to make better decisions. It does not change the outcomes of those decisions: it simply helps managers know which course of action is best. Therefore, the value of research in any given situation depends on the importance of the decision at issue, the level of uncertainty about the proper course of action and the ability of the research to reduce that uncertainty.

## Sources of information

### Internal data sources

#### Sales and expense records

Sales and expense records take two forms. The first is the traditional accounting compilations used to prepare income statements for an organisation’s operating units. The second is sales and expense information organized by customer groups and not by operating units. The uses of traditional sales and expense data for operating units include:

* Sales data for a product or business unit can be analysed to measure seasonal fluctuations and make short term sales forecasts
* Sales for a product can be correlated with prices to estimate to estimate the price elasticity of demand.
* It can also be correlated with advertising expenditure to estimate the advertising response function
* Sales can be compared before, during and after a promotion to measure the effects of promotion.

Sales and expense data based on customers rather than operating units are less traditional but lie at the heart of database business, societal and life, which involves using purchase records and background data on individual customers to tailor what is offered to them or to develop target profiles for potential new customers. Example of using internal information in this way include:

* The profitability of an individual customer or a group of customers can be calculated to determine whether price concessions or business, societal and life expenditures are justified.
* Heavy purchasers can be compared with light purchasers to develop a profile of key target customers
* A customer purchase records can be analysed to estimate that customers buying cycle, so that purchase reminders can be sent when a customer is due for a purchase. This also enables companies to send promotional incentives to these customers.

The problem with using sales and expense records is the difference between what has happened and what is yet to come. They provide information on what happed in the past, but future conditions may be different. The past is not always the best prediction for the future.

#### Sales peoples reports

Four types of reports can be useful:

* Request and information reports show customer reports that can not be fulfilled and customer complaints. Documenting these, helps the company recognize problems and opportunities.
* Lost sales reports provide information on lost sales opportunities. This can alert management to trends and patterns
* Call reports show the date and time and dates of sales call, the company and person visited, the issues discussed, and the out come of the visit.
* Activity reports summarise a sales person’s activities over some time period, how many calls were made, to whom and on what dates. The biggest limitation on using sales peoples reports is the workload they put on the sales for a. On the other hand imposing sales reports on sales people can cause resentment.



#### Street news

Street news about customers and competitors activities is another source of internal information. It should be made a regular part of a company’s business, societal and life information system. This can be stabled by:

* Considering the types of information required
* Communicating these guidelines to people in the organization and establishing a reporting system.
* Regularly analysing and reporting this information

### Sources Of Primary Information

#### Observation

Data obtain through observation is called observational data. It involves observing people, objects, or events. Observations can be executed by human observers or mechanical devices e.g.

* manufacturers invite children to play in a room containing various toys and observers record which toys are used.
* Service shoppers visit stores or restaurants and record information about the service they receive

Information can also be gathered by way of self-reporting methods. One advantage of observation is that it does not suffer from respondents forgetting what happened or distorting their answers to make a good impression.

Two terms relevant to distortion is obstructive versus unobstructive and reactive versus unreactive measurement.

Obstructive means that the object being measured is aware of the measurement; reactive means that the objective reacts in changing in some way. The major problem with observation in regard to self-report measures is its limited applicability.

#### Surveys

When you do a survey you gather data by means of structured interviews. People who conduct surveys usually use a standardised questionnaire, which has certain advantages and disadvantages over structured interviews. The advantages of surveys are:

* The use of structured questionnaires makes it easy to analyse the data since all respondents (people who complete the questionnaires) are asked the same question in the same order.
* It allows the researcher to control the interview without being present
* It allows structured questionnaires as well as survey interviews to be done by telephone or through mail, which means they can be cheaper than interviews which require personal interaction.
* The use of telephone or mail and the lower costs per interview makes it possible to conduct a large number of interviews with a broader cross-section of the market.

The disadvantages of surveys are:

* Structured interviews reduce flexibility
* Deep feeling and hidden ones can not be probed very well
* Question are limited to those that provide short answers.

In conclusion, surveys are good for measuring facts but less so for in-depth studies or profiles of individual respondents.

#### Personal surveys, telephone surveys, and mail surveys

* Personal surveys offer maximum questionnaire flexibility.
* Personal surveys are used where telephone surveys are not appropriate.
* Intercept surveys, usually conducted in shopping malls, allow objects to be shown to respondents at a lower cost than in personal interviews.
* Sample quality is low.
* Save time with short questionnaire
* Telephone surveys offers a good sample quality
* The cost of telephone surveys is low.
* Mail surveys offer the advantage of lower costs
* Low response rate with mail surveys

### Sources Of Secondary Information

Secondary information is less costly to acquire.

#### Library Sources

They contain four types of information, which may be used in research projects.

* Books; they are not useful as sources of statistical information and market forecasts. The other problem with books is timeliness.
* Government documents
* Periodicals
* Computerized data base

#### Non-Library Sources

* Trade associations
* Government agencies
* Media companies
* Local private sources
* Syndicated data services
* Personal networking
* Internet

### How Do We Evaluate Secondary Information?

Secondary information is obtained from a source outside the organisation. Because we have no control over how secondary information was collected, analysed and presented, we always have to make sure that the integrity of the information is intact.

When you evaluate secondary information, ask the following questions:

* Who sponsored the research?
* Who conducted the research?
* Who provided the information?
* Who reported the information?
* What information was gathered?
* When was the information gathered?
* Where was the information gathered?
* How was the information gathered?

## Conducting A Survey (Research)

The basic process of conducting a market survey is as follows:

* Determine the aim of the survey (also called research)
* Identify the population and sample: from which population group will you gather the information?
* Decide how to collect replies: how will you get their replies to your questionnaire: by phone, mail, verbally, etc.?
* Design your questionnaire: the questions you will ask in order to collect the information
* Run a pilot survey: a test survey to check the process, the questions, the aim and the information
* Carry out the main survey: the actual research
* Analyse the data: now you will analyse the data in order to find out what it tells you.

It is important that the questions in your questionnaire asks questions that relate to the aim of the research in the first place.

This is why you must first determine the aim and identify the population sample and also decide how you will collect the replies before you draw up the questionnaire.

You must also carefully think about how you will analyse the data before you start collecting the data. You should actually know before you even do the questionnaire how you will analyse the data.

### Step 1:Determine Your Research Aims

****Start your survey by setting sown the aims for the survey. Why are you doing research and what do you want to achieve? What do you want to know? If we use Markinor as an example, they want to determine before the election who is going to win in which area, and how the other political parties will do during the elections.

In the workplace it can be that you want to find out:

* How many passengers you transport per route
* Why customers use your organisation rather than one of the opponents
* How much fuel your vehicles use
* How many man hours are lost every year during the winter due to illness of staff members
* How you can improve your customer service
* What other services customers require from your organisation

Doing a survey does not have to be difficult and complicated or expensive.

There is a story about a construction company that was trying to find a competitive edge. The management decided to do market research and so it asked customers about the worst habits of the competitors. Of course, the customers talked about the bad habits of constructions companies:

* Being impolite
* Don’t care about the dirt that workers bring into the home
* Staff and equipment that looked shoddy

So, what did this construction company do to be better than their competitors? They

* Bought new equipment and kept it in good condition
* Trained their workers to be polite
* Dressed the workers well in order to project a good image

Did the company benefit from the market research and the changes that were brought in as a result of the market research? Yes. In less than two years the company increased its yearly sales FIVE TIMES!

### Market segmentation

Within the total population group there are groups of people who have similar characteristics, such as age, income, type of work, hobbies, etc. This is called market segmentation. The needs and desires of teenagers are different to that of middle aged people: teenagers like to hang around in malls playing games and socialising with their friends, while middle aged people prefer to visit friends at their homes and have a braai or something similar.

When we collect information we will divide the population into segments, which gives us the opportunity to make deductions and weighed assumptions on target groups. Market segmentation is usually done according to the following factors:

#### Demographically

Demographic factors describe the population (or your customers):

* Sex
* Age
* social class ‑ A B C D
* occupation
* family size

#### Geographically

Where to they live and work?

People who live in different areas have different wants and needs. A good example is health care – people in cities have more access to hospitals than people who live in rural areas.

In the same way do people who live in different regions have different wants and needs and experience life differently.

Some examples of geographic factors combined with demographic factors:

* Mothers with babies under 2 years of age, BCD, throughout South Africa
* All pool owners in Mpumalanga
* Farmers in the Free State, and so on

#### Sociologically

This has to do with the lifestyle of the population:

* 'Under‑privileged young marrieds'
* or ‘Upwardly mobile young executives'
* political views: conservative, liberal
* hobbies and interests: sport, movies, etc.

#### Product-Wise

How does the customer behave regarding your product or service:

* When do they buy the product?
* Why?
* How much do they buy?
* Who uses your products?
* What other products do they use?
* Who used the products of the competition?
* All those who have not yet tried the product

Sometimes the target market can be simply all housewives or all current consumers.

### Example Of Research Outcome based on market segmentation

SA’s 10 living standards measure.

Sunday Times, May 15 2005

|  |
| --- |
| **LSM 1: RURAL DWELLERS** |
| * Average income R879 a month (mostly social grants)
* Traditional huts
* Very high unemployment – 83%
* High illiteracy – one in five has no formal schooling
* No tap water
* No fridge, no insurance
* Biggest representation in Kwa-Zulu Natal and Eastern Cape
 |  |
| **LSM 2: FARM WORKER LEVEL** |
| * 91% live in rural areas – one third in traditional huts – with some in formal settlements
* Average income R1 068 a month.
* Most employed are farm workers and labourers
* One third have access to running water – mostly outside
* First level have TV’s, at 30%; and cell phones, at 13%
 | N0269 |
| **LSM 3: MATCHBOX HOUSE/INFORMAL SETTLER LEVEL** |
| * Over a third live in urban areas
* Mostly “matchbox housing” and informal settlements
* Average income R1 048 a month
* Almost half own a TV set and a fridge
 |  |
| **LSM 4: THE URBAN POOR** |
| * High number of backyard and poor township dwellers.
* Average income R1 774 a month
* Highest number live in Gauteng
* Three-quarters own a TV set
* Low interest in pets
 |  |
| **LSM 5: “SOMETHING TO LOSE”-** |
| * First level to have significant levels of insurance
* Average income R2 427 a month.
* Almost two-thirds employed.
* Almost 90% have a TV set; a VCR and/or a fridge, and entertainment centre.
* Microwaves appear, and almost half have at least a kitchen sink
* Rates of clothing purchase close to middle-class levels
 |  |
| **LSM 6: FOLKS IN THE FLATS** |
| * Very urban , with the highest level of flat living
* Average income around R4000 a month
* Virtually everyone has a TV, a fridge, and entertainment centre
* Employment includes “professional/technical”, but service and clerical/sales is still high
* Many have washing machines and freezers
* Hiring of DVDs and Videos is common
 |  |
| **LSM 7: TOEHOLD IN THE MIDDLECLASSES** |
| * High number in small houses and cluster homes
* Average income R6 455 a month.
* Some can afford domestic help
* Most have cell phones
* Most have ATM cards, few have a credit card
* A third own a vehicle
 | hh01335_ |
| **LSM 8: TOWNHOUSE AND RETIREMENT GENERATION** |
| * Highest level of townhouse dwelling
* Older age profile than middle LSMs
* Average income R8 471 a month
* A quarter own a PC and DVD player
* Job types include administrative/managerial
 |  |
| **LSM 9: HOUSEWIVES AND HOLIDAY SPENDERS** |
| * Highest level of “housewives”, most have domestic help
* Average income R11 560 usually dominated by males
* English and Afrikaans dominate as home languages
* 100% ownership of “everyday appliances”.
* Many spend on timeshare holidays and travel.
 |  |
| **LSM 10: SWIMMING POOLS, SELF-EMPLOYMENT AND SUBURBIA: WELCOME TO EASY STREET** |
| * Medical insurance, stock exchange investment and home loans peak.
* Most have DSTV; and pay for private security services.
* Swimming-pool ownership almost 300% over LSM9, and almost everyone has a PC
* Most are English-speaking
* Average income R18 649 a month
* One in five have gym contracts
* Only category more likely to eat in restaurants than takeaways.
* Many have second homes, recreational vehicles.
* Pet heaven – 73% own dogs, 26% keep cats.
 |  |

*The South African Advertising Research Foundation.*

### Step 2: Identify The Population And Sample

Sampling is a basic concept used in statistics and is used to try to estimate what the parameters of a population are. For example, I may want to estimate how tall South African men are on the average. Perhaps, I measure a few hundred men and calculate their average. The men chosen for measurement are the sample. The parameter I am looking for is the average height of South African men.

One sip of milk is sufficient to let you know that the milk is sour. This is what sampling is all about. In order gain information about the whole you only need to examine a part. On the other hand, sipping a glass of water tells me nothing about the milk.

Here are a few definitions that statisticians use when dealing with samples:

* A ‘**population’** is the entire group of objects about which information is wanted.
* A ‘**unit’** is any individual member of the population.
* A ‘**sample’** is a part or subset of the population used to gain information about the whole.
* A ‘**sampling frame’** is the list or units from which the sample is chosen.
* A ‘**variable’** is a characteristic of a unit that is to be measured for those units in the sample.

The distinction between population and sample is extremely important to statistics. First, we look at a population then some examples to make the distinction clear.

A population is defined in terms of our desire for information about that population. For example, if I want information about all high school students in South Africa, then the population is all high school students in South Africa. Even if I can only choose one high school and its students, the population remains all high school students in South Africa.

**It is extremely important to define clearly the population of interest.**

If you want to determine how many South Africans are in favour of gun control laws, you must define the population precisely. Are all South African residents included, or only citizens? What is the minimum age you require? Are individuals imprisoned for violent crimes allowed to be included, or just those with minor offences, or none at all?

The following examples are presented to give an understanding of population, sample and variables. Data (numbers) used in most examples are completely fictitious as you have been warned.

#### Example 1

The national census attempts to collect basic detailed information from each household in the country.

* Population: all South African households
* Sample: the entire population, as far as possible.
* Variables: the number of occupants, age, race, gender, family relationships, access to electricity, water and telephone services.

#### Example 2

You want to do market research to find out your customers’preference and usage of various products or services. Undoubtedly you have heard radio or TV announcers state that their station is listened to or watched by 80% of South African while their competitors only make up the remainder.

* Population: all South Africans who have either a radio or a TV.
* Sample: about 1500 residents of South Africa
* Variables: gender, racial group, age, residential area, income category and the answers to the specific question or questions.

When Markinor does a survey before an election, they do not ask every person in the country for their views or opinions, they choose a number of people from the various population groups. The number of people they choose is called a **sample**: a sub-set of the population, while the **population** is the members of the group you are interested in.

When you choose the sample for a countrywide survey, you have to make sure that your sample represents the entire population. Usually the **sample** will then be chosen from a list that contains all the members of the population, such a list is called a **sampling frame**.

This is probably what Markinor does when they do surveys before elections. Luckily, for most of the research we want to do we do not have to go countrywide: we can usually choose from our community or customers or the customers of our competitors.

### Sampling methods

When we sample we want to estimate the variables, or parameters, of the population. Therefore, the samples must be taken in a manner that ensures that they are representative of the population. When our sampling methods produce results that consistently and repeatedly differ from the truth about the population, we are doing something wrong.

Our sampling must consistently produce results that have high precision in estimating the parameters of the population and these estimates must have low bias. A lot of statistics-talk seems contrary to the way people talk. A statistician may use ‘high precision’ but he may also refer to lack of precision. Sometimes this practice produces some interesting double negatives.

‘**Bias’** is consistent, repeated divergence (difference) of the sample statistic from the population parameter in the same direction.

‘**Lack of precision’** means that in repeated sampling the values of the sample statistic are spread out or scattered. Lack of precision means that the result of sampling is not repeatable.

The easiest way to visualize bias and precision, or lack of precision, is to refer to the following targets. Of course, we want to hit the centre, the bull’s-eye! The figure on the next page shows four targets that represent

* high bias and high precision,
* low bias and low precision,
* high bias and low precision, and
* low bias and high precision.

When the bias is high the pattern is further away from the centre of the target. High precision means that the pattern is clustered close to its neighbours.

With statistics, we are aiming at low bias and high precision. This means that the values we obtain should be consistently close together and that they are good estimates of the value we are trying to estimate. Just like trying to hit the bull’s eye.

It is now time to introduce a few more terms.

* A ‘**probability sample’** is a sample where every element in the population has a known probability of being included in the sample.
* A ‘**non-probability** sample’ is one where elements in the population do not have a known probability of being included in the sample.

There are times in a probability sample where certain sections of the population may have a probability of zero. That is, there are times when they are intentionally excluded. The idea of probabilities and the details of sampling are discussed in other sections.



### Probability sample

We will discuss 2 probability sampling techniques:

#### Simple random sampling (SRS)

The simple random sampling (SRS) technique is extremely simple and forms the basis for other probability sampling techniques. An SRS of size n is a sample of n units chosen in such a way that every collection of n units from the sampling frame has the same chance of being chosen. This technique ensures that all units in the population have the same probability of being chosen.

As the SRS is so important, a considerable amount of time is spent on it in this module. In addition, many of the concepts used in this module are first introduced when dealing with the SRS.

The disadvantages of an SRS are:

* The numbering of all elements in the population can be time consuming, especially when the population is large.
* A complete list of the population elements is often not available.
* If the population is spread out over a large geographical area, contact with the elements in remote locations could be costly and time consuming.
* Since all possible samples have the same chance of being selected, sampling from a heterogeneous population (one that varies considerably) could lead to large sampling errors.

#### Systematic Sampling

The systematic sampling technique is very easy to set up and is a simplified version of the SRS.

The advantages of systematic sampling are:

* It may be more convenient and easier to administer than other sampling techniques. This is especially true if the information is kept in some filing system (electronic or manual) like the cards used in a library system or invoices.
* A complete list of the population is not always necessary. In quality control only each nth (5th, 22nd) element coming from the production line is inspected.
* This technique may ensure a more representative sample by covering the whole range of population elements.
* If the population is homogeneous, systematic random sampling is just as reliable as an SRS. If the population is heterogeneous, systematic random sampling is more reliable than an SRS. (This means that you need to take fewer samples to obtain the same result.)

The disadvantages of systematic sampling are:

* If a cycle exists in the records systematic sampling may coincide with this cycle and produce a biased result.

### Non-probability sampling

Two non-probability sampling methods are discussed and some examples given. These techniques are convenience sampling and quota sampling. Convenience sampling is highly subjective and unreliable and should never be used for statistical purposes. Quota sampling is sometimes used in practice. In non-probability sampling the units are not chosen randomly.

#### Convenience sampling

The name gives the game away here. This technique consists of selecting the population elements with the greatest convenience.

The greengrocer who samples the top layer of apples in the first crate is performing convenience sampling. Many of us have been approached while shopping at a large shopping centre, on a Saturday morning.

It may be convenient for the one asking the questions, but who shops there? If the shopping centre is in Sandton I would expect quite a different range of clientele from those shopping in Soweto.

Convenience sampling, although used by many companies because they are cheap and quick, is usually a waste of time, cost and effort.

Results from convenience sampling usually systematically over-represent some parts of the population and under-represent other. (Who shops in Sandton on a Saturday morning for basic necessities?)

Another example of convenience sampling is when a radio or TV station ‘asks for your comment on this issue so we know what South Africans think’. This example has another element buried in it that is almost guaranteed to produce a strong bias: voluntary response.

A voluntary response sample chooses itself by responding to questions of this nature. People who feel strongly about an issue are more likely to take the time and trouble to respond.

This is especially true if the person responding has strong negative feelings about the issue being discussed. Opinions from convenience sampling, especially coupled with a voluntary response rarely represent the population as a whole.

Convenience sampling is not discussed further in this module however it is used in many places as a ‘good example of what not to do’.

#### Quota Sampling

A quota sample is a selection made on the basis of specific guidelines about which item should be selected. Quota sampling may be viewed as a simplified version of stratified random sampling. In quota sampling, homogenous (similar) cells are defined by the ‘researcher’ and the field worker is then expected to fill the quota from the population. For example, shopping behaviour may be required and the cells (strata) are, men, women and children. The field worker must then obtain responses from the number of people specified in each cell.

Quota sampling is more reliable than convenience sampling or judgement sampling but has no guaranteed or measurable advantages over probability sampling.

Another example of quota sampling is a real estate agent wishing to estimate the average value of houses in a population of cluster houses. The real estate agent may want to know which percentage are one-storey, two-storey, multi-storey or split-level, the percentage that has air conditioning, the percentage that has access to swimming pools and the current market value of each dwelling. The agent would use his or her subjective judgement to decide which dwelling to include in the sample. Quota sampling is not discussed further in this module.

### Using sampling methods

The problem with non-probability sampling techniques is due to human choice. The human element may be removed by allowing an impersonal chance to choose the sample. The SRS (simple random sample) is the basis of probability sampling techniques so I concentrate on this method in more detail than on the other methods. In fact, I introduce a number of concepts and definitions when discussing the SRS that will be very useful in other section of this manual.

### Simple Random Sampling (SRS)

A simple random sample of size n is a sample of n units chosen in such a way that every collection of n units from the sampling frame has the same chance of being chosen.

Note that this method gives every possible sample of size n the chance of being the sample actually chosen. What I mean here is that every time I take n units as my sample, every possible sample of n units has the same chance of being selected. Let’s do an SRS and see how it’s done.

There are ten of us in a room and three of us are to be chosen to represent the group for some purpose. Because we can’t decide who should go, we decide to choose the ‘volunteers’ randomly. Each name is put on a piece of paper and all pieces of paper are identical. The papers are then put in a box and thoroughly mixed. Draw one name (piece of paper) from the box and read it out or put it aside. Mix the remaining names again and draw the second one. Do the same for the third name as well. You have just performed an SRS of size 3 from a population of 10.

Physical mixing and drawing of pieces of paper certainly is simple and this is all that is behind an SRS.

Another way to accomplish the same result without the hassle of handling bits and pieces of paper is to use a table of random numbers. I created such a table for this module (handout 1). This table consists of 12,500 random digits.

Choosing a random digit is like spinning a wheel with the ten digits (0 through 9) on it. such a wheel is shown below. When you need a random digit you spin the wheel. When you need another one you spin it again and continue doing so until you have the number of digits you require.



#### Select from 10

We now want to choose three of the ten of us as we just did by putting our names on slips of paper and drawing them randomly. This time, however, we will use the table of random digits to do it for us. Because there are ten of us we label each person 0 through 9. We then enter the table somewhere and systematically read through it. You may read left to right, right to left, upward or downward. When you run out of a row or column, continue systematically in the next one. It doesn’t matter which way you read the number, just be consistent each time you use the table.

I decided to start on line 123, the first column and read the numbers left to right. The first three digits are: 0, 9 and 5. These are the three people chosen from our group.

Let’s do this again but this time start in row 127. The first three digits are: 9, 9 and 4. We have a problem because person number 9 appears twice. Ignore duplicates and continue until you have three numbers that are all different. The fourth number is 8. In this case those labelled 9, 4 and 8 are chosen.

#### Select from 100

I have been asked to test 5 rounds of ammunition out of a box containing 100 cartridges. In order to randomly choose the 5 rounds, I label each cartridge 00 through 99. (Actually, I’m smarter than that. The cartridges are in a box packed 10 by 10 so I decide that the first row of cartridges is labelled 00 to 09, the next row is labelled 10 to 19 and so forth.)

I decide to start drawing my SRS from row 17 but begin with column 6: 33699, 23672 and so on. I take these numbers and group them in pairs: 33, 69, 92, 36 and 72. Notice that I read across the break in the column, no duplicates appeared so I remove and test cartridges numbered 33, 69, 92, 36 and 72 for testing.

#### Select any number

You are working for a hospital (hotel, restaurant, prison) that has 300 beds, (rooms, tables, cells) and need to select 7 of these randomly to check on its condition.

You label each from 000 to 299 and start in row 101, column 1 and begin to read off the following numbers, 3 digits at a time. The first three digits are 550. This number is outside the range of 000 to 299, so you ignore it and continue. The sequence you have and the actions you take follow:

* 550, ignore
* 348, ignore
* 121, keep
* 790, ignore
* 564, ignore
* 819, ignore
* 431, ignore.

At this point you realize that you are going to ignore about 70% of all three-digit numbers because you only have 300 items out of 1000 possibilities. So, being clever, you give each item three numbers. To make it easy for yourself, you assign number 000 to 299 to the items, as well as 300 to 599 and 600 to 899. So now the first item is labelled 000, 300 and 600. You may find duplicates, which you ignore, but you should only find about 10% that are not in the range 000 to 899. We now repeat the exercise with the following results:

* 550, keep (item 250)
* 348, keep (item 48)
* 121, keep
* 790, keep (item 190)
* 564, keep (item 264)
* 819, keep (item 219)
* 431, keep (item 131).

Just to make sure you are still with me and before we get involved in more interesting matters, I’m going to summarize how to select numbers from a table of random digits.

* If no more than 10 units must be labelled, one-digit labels should be used.
* If 11 to 100 units must be labelled, two-digit labels should be used.
* If 101 to 1000 must be labelled, three-digit numbers should be used.
* Always use as few digits as possible for labels.
* It is also good practice to begin with zero rather than one even when not all the labels are required.
* Ignore unused labels.

It is more efficient to give several labels to each unit in the same sampling frame. If you use multiple labels ensure that each unit has the same number of labels.

#### A worked example

The figure on the next page contains 100 circles of five different sizes: 5mm, 10mm, 15mm, 20mm and 25mm. Perhaps the circles represent something of importance to you. You will find a slightly larger copy of this figure on the next page in handout 2 although the legend is not shown. Make copies of this handout rather than working directly in your manual.

* The circles may represent the size and distribution of bubbles that you may find in a vat when converting coal to petrol, brewing mampoer or making cheese. If the bubbles are too large, too small, too many or too few, the product may be damaged. For example, if your cheese has too many large bubbles it may no longer fit in the containers you have ordered. To complicate matters, if you put the cheese into the same containers, your customers will complain that the mass of cheese in the container is too low. Of course, if the bubbles are too small, you may deliver a higher mass of cheese in the same container and lose money.
* The circles may represent the size of trout and their distribution on your fish farm. If you have too many large trout they may start eating your small trout. You may have too many small trout and will not be able to fill the orders you have for large trout.
* The circles may represent tumours or cultures that are being studied if you are a medical or research person.
* The circles may represent the size and distribution of trees in a plantation if you are a forester.

In statistics, we don’t usually know what the population parameters are so we must be able to estimate them. In this example, I am going to give you the values (answers) and then show you how statistics may be used to estimate the population parameters.

For this example the population consists of the 100 circles. I am interested in finding out the average size of the circles as well as some indication of how many circles of each size there are in the population. The true values could be calculated by taking a census and you would find the following:

* There are 10 circles with a diameter of 25mm.
* There are 15 circles with a diameter of 20mm.
* There are 20 circles with a diameter of 15mm.
* There are 25 circles with a diameter of 10mm.
* There are 30 circles with a diameter of 5mm.

You could easily calculate the average diameter by adding up all the sizes (10x25 + 15x20 + 20x15 + 25x10 + 30x5 = 1250) and then divide the result by 100. The average diameter of all the circles is 12.5mm.

Because there are 100 circles, the percentage that each circle appears is 30%, 25%, 20%, 15% and 10% for the 5 millimetres, 10 millimetres, 15 millimetres, 20 millimetres and 25 millimetres, respectively.

We can now say that we expect the real mean (statistics-talk for average) is likely to be between 6.25 and 23.75. This is an excellent example of **low precision** if you think back to the target example. This example probably has low bias but we can’t be sure of this right now.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sample** | **Average 1** | **Average 2** | **Average 3** | **Average 4** |
| 1 | 6.25 | 11.25 | 15.00 | 11.25 |
| 2 | 13.75 | 15.00 | 12.50 | 17.50 |
| 3 | 11.25 | 13.75 | 13.75 | 16.25 |
| 4 | 12.50 | 23.75 | 12.50 | 13.75 |
| 5 | 8.75 | 8.75 | 12.50 | 12.50 |
| 6 | 10.00 | 13.75 | 17.50 | 10.00 |
| 7 | 8.75 | 13.75 | 6.25 | 13.75 |
| 8 | 10.00 | 13.75 | 8.75 | 15.00 |
| 9 | 7.50 | 18.75 | 13.75 | 10.00 |
| 10 | 12.50 | 12.50 | 8.75 | 8.75 |

The results of using an SRS of size 16 appear in

|  |  |
| --- | --- |
| **Sample** | **Average** |
| 1 | 12.8125 |
| 2 | 11.5625 |
| 3 | 9.0625 |
| 4 | 11.8750 |
| 5 | 11.2500 |
| 6 | 11.8750 |
| 7 | 11.5625 |
| 8 | 11.8750 |
| 9 | 12.1875 |
| 10 | 11.8750 |

‘This looks better’, you will say, and I agree.

The minimum and maximum average diameters are 9.06250mm and 12.81250mm, respectively. This gives us a range of 3.7500mm. All of a sudden the range became smaller and the precision increased! We may now expect the real mean to be somewhere between 9.06mm and 12.81mm. (It’s not necessary to use so many digits after the decimal. After all, we are only approximating the real mean.)

The more samples we take, the better the estimate of the parameter we are estimating.

The average, or mean, is only one of several methods we may use to determine where the centre of the data is located. Likewise, the range is only one of the methods we may use to determine the spread, or scatter, of the data around the average, or mean.

I’m now going to average the averages to see what happens. I can do this only if I am careful to do it correctly. Because all the averages obtained from an SRS of size 4 are the same size (4), I can average all the averages. If I do this, I obtain 12.40625mm, or 12.40mm for simplicity. Averaging the averages of the SRS of size 16 produces 11.59375mm, or 11.59mm.

You might think that using more samples of size 4 is better than less of size 16, but you would be wrong. Both have the same ‘accuracy’. It just happens that in this example, the results turned out this way.

I also counted how many times a circle appeared as a result of the SRS and calculated its percentage of occurrence. The graph on the next page shows the results.

Originally we knew the percentage that each circle appeared was 30%, 25%, 20%, 15% and 10% for the 5, 10, 15, 20 and 25mm circles, respectively.

From the graph and the table we can see the frequency of the exact sizes and the estimate of the frequency of occurrence of each circle. We may estimate the occurrence of the circles as 35%, 22.8%, 17.8% 15.9% and 8.4%. The estimates are not exactly what we started with but they aren’t too bad considering I used a pretty small sample.



When you complete your results, I doubt you will obtain the same values as I did. Always keep in mind that we are estimating the true value of the parameter from a sample and not calculating the parameter from a census.

The lesson you should learn in this example is that an estimate of a parameter becomes more accurate as the sample size grows.

### Sampling distributions

We have just learned that random sampling produces different results each time it is used. Does this mean that the results are invalid? Can I really believe them?

There is another property of random sampling that is even more important than the lack of bias. A sample statistic from an SRS has a predictable pattern with repeated sampling. This pattern is called the ‘**sampling distribution’** of the statistic. Knowledge of the sampling distribution allows us to make statements about how far the sample proportion, for example, is likely to wander from the population parameter.

I’m going to make the assumption that I know the proportion of defects in a certain population as p = 0.05. In other words, I know that 5% of the items (ball bearings, sweets, computers) in the population are defective. Ask yourself a question: If I took 100 items, how many should I find defective? The answer is obviously 5.

But would I always get 5 bad items? Would there be times when I found no defective items or 10 defective items. Maybe! But I doubt that I would get exactly 5 on each sample of 100 items. And if I did, I would suspect foul play! How many defective items do you expect me to find if I take a sample of 25 items? Do you really think that I would find  defective items (0.05×25 = 1.25)?

I took the output from an automated system that is designed to test for a specific defect in a product and reproduced it in the graph below. The sample size in this case is 25 and the defective items are 20% (p = 0.20). This is pretty high for a mass-produced item but fortunately the defect was a slight colour variation that did not affect the use of the product. It did, however, cause items on a shelf to appear different and this worried the manufacturer.

Let’s explain the graph in the figure before I start analysing it. The horizontal axis represents the number of defects that were found (1 to 9) in each sample. The vertical axis represents how many times that number of defects occurred. Looking at the table at the bottom of the graph, you can see the following:

1 defective item occurred 1 time (0.04%)

2 defective items occurred 12 times (0.48%)

3 defective items occurred 158 times (6.32%)

4 defective items occurred 601 times (24.04%)

5 defective items occurred 966 times (38.64%)

6 defective items occurred 594 times (23.76%)

7 defective items occurred 142 times (5.68%)

8 defective items occurred 25 times (1.00%)

9 defective items occurred 1 time (0.04%).

The percentage of occurrence is shown in parentheses. Note that about 39% of the cases had 5 defects, 86.44% had 4, 5 or 6 defects (601 + 966 + 594 = 2161), and 98.44% of the cases had 3, 4, 5, 6 or 7 defects. I think you now see that  = 0.20, approximately.

If we look just at the defects, they range from 1 to 9. That is, in each sample of 25 units, the proportion of defects ranged from 4% to 36% ( = 0.04 to  = 0.36). We know (actually assume) that the proportion of defects is 20% (p = 0.20). Why does one sample give us an estimate of 0.04 and another one give an estimate of 0.36? One sample on its own is not a reliable estimator of the true proportion. However, many samples taken together produce a reliable estimator of the true proportion.



I’m going to summarize the concept involved with sampling variability before we continue because it’s extremely important.

* Despite the sampling variability of statistics from an SRS, the values of those statistics have a known distribution (pattern) in repeated sampling.
* When the sampling frame lists the entire population, simple random sampling produces unbiased estimates. This means that the values of a statistic computed from an SRS neither consistently overestimates nor consistently underestimates the value of the population parameter.
* The precision of a statistic from an SRS depends on the size of the sample and can be made as high as desired by taking a large enough sample.

To be completely honest with you, the examples with the proportions and the one with the defects were not the result of an SRS. However, the sampling methods used in these two cases are based on random selection and also share the three basic properties listed above.

### Problems that may occur

OK, I have convinced you that sampling is the answer to all our problems! Right?

Things may and do go wrong. There are two separate types of errors than can affect our results. Sometimes the effects are so drastic that ‘we back the wrong horse’, so to speak. (Sampling won’t help you with horse betting, by the way!). These errors are called ‘sampling errors’ and ‘non-sampling errors’.

* Sampling errors are errors caused by the act of taking a sample. These errors cause sample results to differ from the results of a census.
* Non-sampling errors are errors that are not related to the act of selecting a sample and may even be present in a census.

When I’m counting defects in products, measuring the diameters of ball bearings or obtaining results from laboratory tests, sampling errors are my biggest problem.

This is easy to fix, however: use an SRS or another type of probability (statistical) sampling method.

If I am concerned with public opinion, what a consumer really wants to buy or personal matters, non-sampling errors become very serious concerns in almost every sampling technique.

#### Sampling errors

We know that differences in sampling will occur just as the day follows the night. In fact, the whole idea of sampling is to use the long-term effects of probability to be able to predict how precise our outcome will be. Sampling errors are different. They are the result of introducing bias into the sampling process. Here are two examples that have caused the uninitiated to have red faces.

Do you have a telephone at your place of residence? I don’t mean a cell phone I mean a telephone. If so, is the telephone in your name and is your number in the telephone directory? Or do you have an unlisted number? Telephone interviewing is very popular but if I use a telephone book to sample households my sampling frame omits a large portion of the populace.

Statistics SA, in its 2001 census reports, indicates that 10.2% of all South Africans have a telephone in their dwelling and no cell phone and 14.2% have both. Stats SA, as it is known, also state that 18.0% of South Africans have a cell phone and no telephone. (There is no indication in their data of how many telephones are unlisted.) So if I use a telephone directory to select my sample, I am excluding 75.6% of the population from my sample and only including 24.4%.

Those that I am excluding have no access to a telephone at all, or have limited access either nearby or far away. Some have access only to public telephones or through neighbours and friends.

Stats SA also state that 12.0% of black Africans, 43.2% coloureds, 74.8% Indians or Asians and 78.6% whites have telephones in their dwellings or both telephone and cell phone access.

So you have a phone in your dwelling. Who answers the phone when it rings?

Telephone surveys are often conducted with equipment that dials digits randomly. This certainly gets around the problems associated with using telephone books, but it also has its share of problems.

Usually, the survey team will speak to the first person that answers the phone. If a child, domestic servant or relative answers, the information obtained may be biased. If there is no answer at a particular number, the equipment moves on to the next number until someone answers. This, of course, biases the results towards those who are more available than others. The time of day the phone calls are made also has an impact.

Telephone surveys, although important, should not be the only way a survey is conducted. If it is, then the results will most likely be biased.

#### Non-sampling errors

The biggest problem to deal with when dealing with non-sampling errors is missing data. Missing data usually results in a refusal to answer a question.

Another problem concerns incorrect answers. Many people just lie when it concerns personal details like age or income. In addition, people just have bad memories. In any case, all people are inaccurate when it comes to time based events. ‘How often did you visit a doctor in the last year’ or ‘How many cigarettes did you smoke last week’ usually produce inaccurate results.

Processing errors also occur but are usually related to the human side of coding. If I must key in a ‘1’ for male and a ‘2’ for female, I just get tired and make mistakes. If I must key in ‘1’ for single, ‘2’ for married, ‘3’ for divorced and ‘4’ for widow/widower, I’ll tire quicker and will probably enter ‘5’!

Perhaps the trickiest problems to detect and fix are those that result directly from the method used to collect the data. This may include the question itself. If I am trying to obtain information on financial investments, and ask a farmer in the Free State: ‘Do you own stock?’ or ‘How much is your stock worth?’ I might just get a few surprises. He knows that sheep and cattle are stock but may have to think about including ostrich, trout or game.

The timing of a questionnaire or survey may be very important. If your survey happens just after an event, will it affect your response? If you asked a typical American on 10 September 2001 ‘Do you feel safe from terrorist attacks?’ you would probably obtain quite a different response if you asked the same question on 12 September 2001. (The World Trade Centre was destroyed on 11 September 2001.)

Questions on questionnaires must never hint at an answer or try to obtain a certain response. A few typical questions of this nature are: ‘Do you favour the banning of guns in South Africa in order to reduce the rate of violent crimes?’ or ‘Do you favour re-imposing the death penalty in order to reduce the rate of crime?’ These are loaded questions but we see them all the time.

Here are two questions on the same topic. How would you answer them?

* Do you believe that there should be a law prohibiting abortions?
* Do you believe there should be a law to protect the life of the unborn child?

Often the race of the interviewer affects the answer especially if dealing with racial issues. The interviewer might just take the attitude that ‘I am just doing my job’ but the person being interviewed might just think (subconsciously) ‘I am giving the answers he/she wants to hear’. Setting up a questionnaire is not an easy matter!

### Answers you need to know

In order to develop critical thinking on all issues and become a participating citizen contributing toward the future of South Africa, you must ask the following questions when confronted with statistical results.

* What was the population? That is, whose opinions were being sought?
* How was the sample selected? Look for mention of random sampling or probability sampling.
* What was the size of the sample?
* How were the subjects contacted?
* When was the survey conducted? Was the survey conducted just after some event that may have influenced opinion and responses?
* What were the exact questions asked?

### Systematic Sampling

In systematic sampling, sampling begins by randomly selecting the first element. Thereafter, subsequent observations are selected at a uniform interval. Think of systematic sampling as a circle. You choose a random place on the circle and then keep going around the circle until the desired number of elements is selected.

For example, I want to choose 6 books randomly from a selection of 20. The number of total elements in my population is 20 and I number them from 0 to 19. I want six books so I can calculate the stepping rate around the circle as . I could also choose the stepping rate from a table of random digits as well.

I then choose a random number to start. Using the table of random digits, I look at column 1 row 1 and obtain 53. If I divide 53 by 20 and keep the remainder (13) I have my starting point.

The figure on the next page shows a circle with the starting point, 13, marked by an arrow. The subsequent books are selected by counting systematically around the circle. I went in a clockwise manner but could have gone counter clockwise if I chose.

The books numbered 13, 16, 19, 2, 5 and 8 are selected from the 20 available.

This sampling scheme has resulted in some biased results when applied blindly. If you are using this technique and there is a cycle in your population, just ensure you don’t end up using the same cycle to sample.

For example, if I choose to select every third unit and my population consists of row upon row of flats that are three storeys tall, I could end up sampling the same floor in all buildings and never obtain results from tenants on other floors. (This actually happened!)

Another natural cycle consists of the days of the week. Let’s assume that I want to estimate the amount of traffic passing through an intersection. If I choose 7 as my sampling interval and my randomly chosen day is a Sunday, I will be only obtaining data from Sunday. The traffic on a Sunday will probably provide a different result than a Monday or a Friday.



### Identify the population and sample for your survey

In order to find out how big your sample size should be (how many people you should question) you must first determine how many responses (completed questionnaires) you will need for the analysis.

A general rule is to look for about 20-30 responses in each of the major sub-categories for the sample. For example, if a key aspect of your research is to compare male and female then you should look for about 30 females and 30 males in your responses.

Once you know how many responses you want, 60 per the above example, you have to find out how many questionnaires you have to send out. Usually, about 20% of the people will reply to the questionnaire. This means that you then have to send out about 300 questionnaires to get 60 responses (people replying to your questionnaire)

On the other hand, if you are going to interview your customers by telephone or face-to-face, you will need less questionnaires.

### Design the survey

There are three general forms of surveys: questionnaires, interviews and network analysis. This manual introduces questionnaires and only briefly describes interviews.

In addition to surveying with questionnaires and interviews researchers may make direct observations.

Sometimes the researcher can use special technologies such as screening rooms equipped with ‘wiggle-meters’ (that measure how attentive respondents are by tracking how often they shift weight, or wiggle, during a session).

In some sociological studies, the behaviour of the subjects may be monitored and a questionnaire or interviewer may just be used to distract the subject from the real purpose of the research.

### Questionnaires

In this section I’m going to introduce you to two types of questionnaires: questionnaires that measure your judgment or feeling about something and questionnaires that measure your achievement in something.

* Questionnaires that measure your judgement are usually very difficult to prepare and use.
* Questionnaires that measure your achievements are usually fairly easy to prepare.

### Methods to measure judgments

Questionnaires are often constructed to ask people to report their attitudes or rate some behaviour. These questionnaires are trying to estimate opinions and not obtain facts concerning achievement. A judgment is something you or I perceive, prefer, like or dislike. A fact would be something like ‘What is 2 plus 2?’

When we deal with preparing questionnaires we leave the field of statistics and enter the fields of sociology, psychology or communications. We then have to learn a whole new set of words and our terminology changes as well. When we deal with analysing questionnaires, we again enter the field of statistics and change our vocabulary back to that of a statistician.

The person who sets up the questionnaire is usually called the researcher and the ones responding to the questionnaire are usually called the respondents or subjects. Some researchers make a distinction between subject and respondent. These researchers use ‘subject’ when referring to the people used in designing the questionnaire or those approached to take part in the survey. They then use ‘respondent’ for those people who complete the questionnaire.

Usually several cycles of questions and responses are required in order to obtain the final set of questions and to determine how to rate the answers when the questionnaire is finally used. A questionnaire that is to be used to measure judgments is normally called a ‘tool’ or ‘measurement instrument’. A questionnaire that is used to measure achievement is normally called a ‘test’ or ‘assessment’.

A ‘scale’ usually refers to a question as well as the value that is used to rate the question in a questionnaire.

Judgments and opinions may be measured in many ways but the most popular ones are Thurstone scales, Likert scales, Guttman scales, and semantic differential-type scales. Don’t be frightened by their names! It’s not that difficult to understand what they do but they are difficult and time consuming to prepare properly. We will only discuss Likert scales.

#### Likert scales

This technique consists of statements that reflect clear positions on an issue. The respondents are asked to indicate their responses on 1 to 5 point scale: Strongly Agree, Agree, Neither Agree nor Disagree, Disagree, Disagree Strongly.

A few researchers use four point scales and leave out the middle one entirely. Using an even number places stress on the respondent to either agree or disagree with the statement and not just select the middle item for all answers. Occasionally researchers add additional scale positions (Disagree Slightly and Agree Slightly) but the basic system is not defined by the number of points used.

You have probably filled out Likert scales at one time or another. These scales are very popular because they are easy to develop, use and analyse. In addition, the researcher can determine reliability at the same time the data are collected but this is not covered in this module.

Despite the ease of creating and using Likert scales the total score sometimes hides important details from the researcher. In addition, if researchers change their topics it may mean that the Likert scale will have to be created again.

#### Personal Report of Communication Apprehension (PRCA-24)

Directions: This instrument is composed of 24 statements concerning your feelings about communication with other people. Please indicate in the space provided the degree to which each statement applies to you by marking whether you (1) Strongly Agree, (2) Agree, (3) Are Undecided, (4) Disagree, or (5) Strongly Disagree with each statement. There are no right or wrong answers. Many of the statements are similar to other statements. Do not be concerned about this. Work quickly, just record your first impression.

[ ] 1. I dislike participating in group discussions.

[ ] 2. Generally, I am comfortable while participating in a group discussion.

[ ] 3. I am tense and nervous while participating in group discussions.

[ ] 4. I like to get involved in group discussions.

[ ] 5. Engaging in a group discussion with new people makes me tense and nervous.

[ ] 6. I am calm and relaxed while participating in group discussions.

[ ] 7. Generally, I am nervous when I have to participate in a meeting.

[ ] 8. Usually I am calm and relaxed while participating in meetings.

[ ] 9. I am very calm and relaxed when I am called upon to express an opinion at a meeting.

[ ] 10. I am afraid to express myself at meetings.

[ ] 11. Communicating at meetings usually makes me uncomfortable.

[ ] 12. I am very relaxed when answering questions at a meeting.

[ ] 13. While participating in a conversation with a new acquaintance, I feel very nervous.

[ ] 14. I have no fear of speaking up in conversations.

[ ] 15. Ordinarily I am very tense and nervous in conversations.

[ ] 16. Ordinarily I am very calm and relaxed in conversations.

[ ] 17. While conversing with a new acquaintance, I feel very relaxed.

[ ] 18. I’m afraid to speak up in conversations.

[ ] 19. I have no fear of giving a speech.

[ ] 20. Certain parts of my body feel very tense and rigid while giving a speech.

[ ] 21. I feel relaxed while giving a speech.

[ ] 22. My thoughts become confused and jumbled when I am giving a speech.

[ ] 23. I face the prospect of giving a speech with confidence.

[ ] 24. While giving a speech I get so nervous, I forget facts I really know.

The Likert scale is often presented with the question in one column and the scores ‘SA’, ‘A’, ‘U’ or ‘N’, ‘D’ and ‘SD’ in another column which allows the respondent to circle the answer instead of entering a number.

The items scored on this Likert scale are ‘Group’ (your score in a group), ‘Meeting’ (your score in a meeting), ‘Dyadic’ (your score in a one-on-one encounter), ‘Public’ (your score in a public meeting) and ‘Overall (your overall score). The scoring is done as follows where the numbers in parentheses represent the numbers the respondent entered.

Group = 18 - (1) + (2) - (3) + (4) - (5) + (6)

Meeting = 18 - (7) + (8) + (9) - (10) - (11) + (12)

Dyadic = 18 - (13) + (14) - (15) + (16) + (17) - (18)

Public = 18 + (19) - (20) + (21) - (22) + (23) - (24)

Overall CA = Group + Meeting + Dyadic + Public

(Dyadic means person-to-person or one-to-one.)

### Interviews

Questionnaires are handy but researchers sometimes find interview methods more useful for a couple of reasons. It is easy for many people to ignore a cold questionnaire but it may be difficult for them to ignore a live person who asks questions.

Of course, the interviewer may arouse some suspicions but this is part of the job to involve respondents in the task. Interviews are conducted to increase the willingness of the respondent to participate and to obtain information that may be lost with a questionnaire.

The interviewer may record information (such as a respondent’s manner and body language) that would be absent with the questionnaire method. Of course, there is some art involved in completing interview studies. So when you read studies that use the interview method, imagine the give and take involved and the pressures the researcher and respondent felt.

### Step 3: Decide How To Collect Replies

Now you have to decide the following:

* Are you going to mail the questionnaire
* Are you going to send the questionnaire by e-mail
* Are you going to phone customers
* Are you going to question customers face-to-face

If you are going to conduct the survey by means of telephone or face-to-face interviews, you have to decide who will do the interviews and how you will control the quality of the interviews. Remember, if you do not do this yourself, people can lie about how many interviews they conducted and what the response of the people they interviewed were.

It will also be important for the interviewer to explain to the potential respondent why they should answer the questions. They should persuade people to take part in the survey, not force them.

### Step 4: Questionnaire Design

Most researchers make the mistake of asking too many questions. Your greatest enemy in survey research may well be poor response rate. Clear and concise questionnaires can help get the best response.

Design of the questionnaire can be split into three elements:

* Determine the questions to be asked
* Select they question type for each question type and specify the wording
* Design the question sequence and overall questionnaire layout

#### Determine the Questions to be Asked

Obviously, your questions should relate directly to the aim of the survey and to the specific information that you will require.

#### Decide on layout and sequence

* Do not clutter up the form with unnecessary headings and numbers.
* Include the contact and return information on the questionnaire, irrespective of whether addressed return envelopes are provided, these can easily become separated.
* Identify individual questions for reference purposes.
* Be careful not to overfill the page.
* Avoid using lots of lines, borders and boxes since these can make the page look too ‘dense’.
* Small fonts may put people off, especially people with bad eyesight.
* Use a good legible font.
* Make good use of italics and bold types, think of using italics consistently to give instructions.
* Consider using bold for the questions themselves or for headings. Symbol fonts may also be useful.
* Begin with questions that will raise interest.
* You should try to keep the flow through the questionnaire logical and very simple, i.e. avoid any complex branching.

#### Question types

**Open-ended Questions**: E.g. Do you think football hooliganism is caused by: (*tick if appropriate*)

|  |  |
| --- | --- |
| Lack of discipline at home |  |
| Players’ behaviour on pitch |  |
| Family breakdown | ?????? |
| Youth unemployment |  |
| Poor schooling |  |
| Violence on T.V. |  |
| Other (please specify) |  |

**Single vs. Multiple Response:** E.g. What is your most usual means of travelling to

college?

|  |  |
| --- | --- |
| Bus |  |
| Car |  |
| Bike  |  |

**Rated Response:** A popular approach in the social science is to use Likert scales such as the example below.

Please state how often you use the following: (*Please circle the numbers as appropriate)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Very often | Often | Occasionally | Never |
| Newspapers | 1 | 2 | 3 | 4 |
| Books | 1 | 2 | 3 | 4 |
| Periodicals | 1 | 2 | 3 | 4 |

#### Questionnaire instructions

Questionnaire instructions should help respondents by explaining how they should answer questions. Instructions must be clear and should explain to people how they are to complete the questionnaire. As a matter of routine, many researchers find it useful to provide a category where the respondent can indicate ‘no response’ or ‘not applicable.’ This strategy is useful since researchers can separate incomplete responses from thoughtful refusals to respond. The table on the next page shows some typical methods used for questions.

#### Wording of questions

When you compile a questionnaire, think carefully how you phrase the questions that you are going to ask the people out there. You have to phrase the questions in such a manner that the people who complete them must:

* Be able to understand the question
* Be able to answer the question
* Be willing to give you the information you need

You also have to ensure that the questions and the way you ask them cannot be constituted as biased in any way, e.g. biased based on race, gender, age, religion, culture, etc.

#### Formatting

Most survey forms begin with a brief statement of introduction to announce the survey, request participation, assure confidentiality (if appropriate), and indicate how to return the survey form. A letter of introduction is included with mailed survey forms although sometimes introductory letters are used in other settings as well.

The important part of the survey form usually includes the questions or statements to which respondents must react. In practice, most questionnaires place demographic questions (gender, age, racial group) first but some researchers argue that demographics should be placed last to avoid boring people with dull background questions at the beginning of the survey.

Next, questions that rely on the same sort of response mode (multiple choice, true and false) are grouped together. Each section should be preceded by instructions for completing the items. Some researchers recommend that questions that deal with the same issue should be grouped together. However, this ordering sometimes creates bias in response as people may try to respond in ways that are consistent with early statements on the survey. For example, suppose I ask you to evaluate the quality of teaching in statistics classes and at a later point ask you to state what you believe are the most important needs at your place of work. It is quite likely that you will mention the quality of teaching as being one of the important needs.

Although you can’t avoid the effect of question order, you should attempt to estimate what the effect will be. This will allow you to interpret results in a meaningful fashion. If the order of questions seems an especially important issue, you might prepare more than one version of the questionnaire where each contains different ordering of questions.

The correct length for a questionnaire almost always poses a problem for researchers. In general, brief questionnaires are preferred to lengthy ones but sometimes you may have to prepare long questionnaires.

### Determining reliability and validity

Researchers must evaluate questionnaires in order to determine whether they are valid measures of the concepts or opinions being sought. ‘Validity’ is the degree to which a measure actually measures what it claims to measure.

In addition to formal methods to establish validity, researchers often include other controls in their surveys. Sometimes they try to detect when people are not responding consistently or accurately. Three types of controls are check questions, test taking measures and alternating polarity of items. A check question involves asking the same question twice at different locations in the questionnaire, usually once positively worded and once negatively worded, such as:

The physical working conditions are good. [ SA | A | N | D | DS ]

The physical working conditions are poor. [ SA | A | N | D | DS ]

If people pay attention, they should answer the items consistently. If they answer in very different ways, then they are not really alert when completing the survey. Check questions help identify people who are not responding consistently so that they can be discarded from the final sample.

Another control method is the use of measures that test the behaviour of the respondent. Scales have been developed to reveal whether an individual is responding in unrealistic ways to questions.

#### The MMPI LIE (L) scale

The MMPI Lie (L) scale is a scale used to identify respondents who are attempting to avoid being honest in their responses. With these types of questions any answer of false is scored as telling a lie.

Circle T if the statement is true for you and circle F if the statement is false for you.\*

[ T | F ] At times I feel like swearing.

[ T | F ] I get angry sometimes.

[ T | F ] Sometimes when I am not feeling well I am cross.

This scale consists of 15 items that are usually included in general personality tests in a random order. If a respondent scores high on these items they may be deleted from the sample. Some believe that the lie scale can be faked by testwise respondents but this hasn’t been conclusively proven.

#### Polarity rotation of items

Another method of control is the use of polarity rotation of items. This process means: avoid phrasing all items positively; and, avoid placing all positive adjectives on the same side of the measurement items.

Researchers often are concerned when respondents follow predictable patterns in selecting their responses. If a person gets in the habit of responding positively to items on a questionnaire, he or she may not be paying attention to the content. It is rumoured that when right-handed people get tired they tend to check positions toward the right sides of questionnaires, whereas when left-handed people get tired they tend to check positions toward the left sides of questionnaires.

### Increase Participation Rates In Surveys

* In face to face and telephone interviews, participation rates depend mainly on:
* The number of call backs used to reach unavailable respondents
* The experience of the interviewer in gaining co-operation.
* In mail surveys, using pre-notification postcards, follow-ups, monetary incentives, self-addressed stamped envelopes, persuasive cover letters, and professional questionnaires can increase response.

### Step 5: Run A Pilot Survey

Test the questionnaire on a small sample of your subjects first. If this is not possible at least test it on some colleagues or friends. The aim here is to detect any flaws in your questioning and correct these prior to the main survey.

Having done your pilot survey, you can make amendments that will help to maximise your response rate and minimise your error rate on answers.

### Step 6: Carry Out The Main Survey

The purpose of doing a pilot survey is to find out if you have to change anything in the questionnaires or in your population sample or even the aim of your survey.

If you are using fieldworkers, you have to ensure that they are well trained to minimise errors in the collecting of data.

* Errors when choosing respondents
* Interviewer dishonesty
* Misinterpreting or misreporting of information
* Non responses: where people are not at home or refuse to answer questions

For the process of actually doing the market research, you also have to

* Set deadlines: start on a specific day and end on a specific day
* Determine the number of questionnaires you want to complete by that day
* If you employ field workers, how many questionnaires every day and how many at the end of the period
* You have to put an administrative process in place: who is going to collect the completed questionnaires?
* You must also have a quality and cost control system in place to prevent dishonesty and prevent fieldworkers from charging too much and wasting too much time. You could, for example, pay per correctly completed questionnaire.

### Interviews

Questionnaires are handy but researchers sometimes find interview methods more useful for a couple of reasons. It is easy for many people to ignore a cold questionnaire but it may be difficult for them to ignore a live person who asks questions. Of course, the interviewer may arouse some suspicions but this is part of the job to involve respondents in the task. Interviews are conducted to increase the willingness of the respondent to participate and to obtain information that may be lost with a questionnaire.

The interviewer may record information (such as a respondent’s manner and body language) that would be absent with the questionnaire method. Of course, there is some art involved in completing interview studies. So when you read studies that use the interview method, imagine the give and take involved and the pressures the researcher and respondent felt.

* The interviewer should be warm and professional.
* Establish rapport to make respondents feel at ease.
* Professional appearance is important



# MODELS PREDICTIONS PROBLEMS

#### Specific outcome

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

#### Assessment criteria

* Experiments and simulations are chosen and/or designed appropriately in terms of the situation to be modelled.
* Predictions are based on validated experimental or theoretical probabilities
* The results of experiments and simulations are interpreted correctly in terms of the real context
* The outcomes of experiments and simulations are communicated clearly

## Drawing Conclusions from Data

In order to draw conclusions from our data we have to understand it and to look at it to see what the data is telling us. In this section we are looking at simple probability theory and how to make confidence statements about our data.

### The Theory Of Probability

The theory of probability plays a significant role in facilitating decision making during uncertainty.

Various techniques are used to determine probabilities:

* A priori probabilities
* Subjective probabilities
* Objective probabilities

Priori and objective probabilities are those probabilities which can be determined when all possible outcomes are known.

Probability statements such as these are made without taking into account the actual frequency of the occurrence. The estimates of probability are made on the basis of deductive reasoning and by assuming that the occurrence of the event is entirely due to chance. Therefore "a priori" probabilities of outcomes are determined for problems associated with games of chance, for example the roll of a dice, the spinning of a coin, roulette wheel, etc.

Most probabilities cannot be estimated by deduction. Probabilities must be estimated only after experimentation and observation. This is also true of future possibilities of the outcome of events. It is possible to explain the outcome of a future event by observing its frequency of occurrence in the past. This data is then projected into the future.

Such probability statements are referred to as statistical or inductive statements. These estimates are based on underlying assumptions:

* That enough past events have been observed.
* That casual influences will be unchanged and remain as they were in the past.

Statistical techniques have been developed to measure the probable range of deviation of the actual from the expected outcomes. They can be used as a measure of risk if they measure the variability of an outcome. Therefore, statistical probabilities can alter either because of changes in a person's knowledge of the causes, which affects deductive reasoning, or because of changes in underlying forces, which affect inductive probabilities. However, in order to keep explanations as simple as possible, probability is regarded as the relative frequency with which an event occurs in the long run.

In order to draw conclusions from our data we have to understand it and to look at it to see what the data is telling us. In this section we are looking at simple probability theory and how to make confidence statements about our data.

#### Probability

Even the rules of football agree that tossing a coin avoids favouritism. Probability is nothing more than tossing a coin!

The main idea of statistical data collection is the deliberate introduction of randomness into choice or assignment of units. Remember the SRS?

Tossing a coin and choosing an SRS are random in the sense that the exact outcome is not predictable in advance, however, a predictable long-term pattern exists and can be expressed as a relative frequency distribution of the outcomes after many trials. Randomness is involved in all games of chance: dice, cards, roulette and the lotto.

The term ‘random’ does not mean ‘haphazard’ because a pattern emerges after a long time. However, we never see enough repetitions to observe the long-term regularity that probability describes.

In the Introduction, I said that we all have an innate tendency to understand and use statistics. And this is true. What is also true is that we **humans do not have an innate ability to understand probability**. We get confused and do silly things like gamble our money away or have another child to break the run of six boys (girls).

Because we have a memory we automatically think that chance events also have a memory. They don’t! Whether we are looking at tossing a coin, rolling a die or two dice, playing the roulette wheel, having children or observing the offspring of animals, the outcome is purely random but totally predictable in the long run.

The ‘Law of Averages’ doesn’t exist with random events and we humans have difficulty understanding this. We think that after having tossed 10 heads in a row the ‘Law of Averages’ will turn up a tail next. It might, but the next toss still has a 50-50 chance of being either a head or a tail. A coin has no memory and it doesn’t know that it turned up heads 10 times in a row.

When we gamble in games of chance, for example the roll of a dice, the spinning of a coin, roulette wheel, etc we consider the probability of winning: in the case of dice 1 out of 6 or 2 out of 12, flipping a coin (heads or tails) 1 out of 2 and, in the case of roulette, 1 out of 75 or higher. These probabilities are called “a priori” probabilities.

#### Simple probability

Probability may be defined as long-term relative frequency. If in a long sequence of repetitions of a random event the relative frequency approaches a fixed number, that number is the probability of the outcome. It can also be said that probability varies between 0 and 1. If the probability is 0, the outcome never occurs. If the probability is 1, the outcome always occurs.

**Experiment and observe**

Most probabilities cannot be estimated by deduction.

Probabilities must be estimated only after experimentation and observation.

This is also true of future possibilities of the outcome of events.

**It is possible to explain the outcome of a future event by observing its frequency of occurrence in the past.** This data is then projected into the future.

You know that if you toss a coin enough times you will obtain about 50% heads and 50% tails. (Usually there is a little more metal on the ‘head’ side of a coin and it turns up slightly more frequently than tails.)

But how do we work out probabilities? We **observe**. Only by observation can we be reasonably sure of the approximate value of the probability of an outcome

If you have the time, try tossing a coin several hundred times and record the outcomes. You won’t get exactly 50% head and 50% tails but you will see that 50% is the value you are heading towards. Now take the same coin but instead of tossing it, hold it on end on a hard table surface with the index finger of one hand and snap it with the index finger of the other hand. The coin will spin for a while before falling with either heads or tails showing. Do not count the spins that fall off the table or the spins that bump into anything else. What do you expect to find: 50% heads and 50% tails? A long series of trials reveals that the probability of a spinning coin is not .

**Definition**



Probability =

The probability of an event is a measure of how likely the event is to occur. Often we can work this out by ourselves with little thought or difficulty. We must just remember that the probability of any event must be a number and this number must be between 0 and 1.

In addition, if we assign a probability to every possible outcome of a random event, the sum of these probabilities must be 1. The probability of obtaining a head from a tossed coin is 0.5 and the probability of obtaining a tail from a tossed coin is 0.5. Their sum is 1.

**Handout 3** has a figure that shows a method to calculate probabilities. This is called a probability tree and shows the probabilities for tossing a coin four times. To see the probability of obtaining four consecutive heads, follow the links for H, H, H and H and you obtain 0.0625 as the probability of four heads in a row.

Let’s get off the coin for now and look at one die (this is the singular of dice). I’m also assuming the die is not ‘loaded’, that is, manipulated in some way to cause the relative frequency of a side to appear more often than the others.

If I roll one die can I get a 0 or a 7? No. The die only has numbers of 1 to 6. How many sides does a die have? 6. So every roll of one die has the probability  of appearing. And .

Now let’s look at two dice. I know that only the numbers between 2 and 12 can appear, but they do not all appear with the same probability. How many combinations of the two dice give me a 2?

The figure below shows the 36 possible combinations there are with 2 dice.



Because there are 36 combinations each possible outcome has a probability of . But I can obtain a particular sum of the spots on the dice in different ways. How many ways can I get a 6? There are five different combinations that produce a 6: 5+1, 4+2, 3+3, 2+4, and 5+1. Because there are 5 possible ways to obtain a 6, its probability is .

The table in handout 2 lists the totals, the probability and the combinations that the outcome of rolling two dice has.

To add a little more to this gambling game, what is the probability of rolling a 7 immediately followed by an 11? It’s simply the product of the two probabilities: .

That is, I expect to see this combination in less than 1% of the time.

What is the probability of rolling either a 7 or an 11? It’s simply the addition of the two probabilities: .

#### Probability and odds

Probability originated in gambling and in gambling the chance of an event is often stated in terms of odds rather than probability. You are rolling two dice and want to get a 7 but you have heard that the odds against this happening is 5 to 1. The probability of a 7 is 

because there are 6 ways you may roll a 7 with two dice (1+6, 2+5, 3+4, 4+3, 5+2 and 6+1).

Odds of 5 to 1 means that failure to roll a 7 happens five times as often as success. In the long run five of every six tries will fail and one will succeed. Therefore, odds of A to B against an outcome means that the probability of that outcome is:



If the odds against the favourite in a horse race are 3 to 2 this is equivalent to the horse having a probability of  of winning.

### The National Lottery or Lotto

I want to say a bit about the Lotto since we are dealing with probabilities. What is the probability of winning the Lotto if you purchase one ticket? The balls drawn for the Lotto are drawn on a random basis and if you match the six numbers drawn, you win. In order to calculate the probability of winning with one ticket I’m going to first calculate how many combinations exist.

There are 49 balls numbered from 1 to 49. On the first draw any one of the 49 balls could be drawn, on the second draw any of the remaining 48 balls could be drawn and so forth.

Since you don’t have to get the balls in the order drawn any of those balls drawn could have first, second and so forth, so we have to compensate for this. (If you had to get the balls in the order drawn your probability of winning would be much lower.)

To calculate the number of combinations, we multiply the number of balls that could have been drawn as follows:



But as I said we must compensate for the six balls being in any order. The first ball could have been in any one of the six positions, the second ball could have been in any one of the five remaining positions and so forth. To compensate we simply divide by these possibilities:



Therefore there are about 14 million possible combinations of numbers appearing in the Lotto. You just bought one Lotto ticket so your chances of winning are 1 in 14 million! The probability of winning is something like 0.00000007. Not very good, is it? But then you want to increase your chances of winning so you spend R2,500 and buy 1000 tickets. Your probability of winning has increased to 0.00007. This is still 1 in 13,984 and that is not very promising.

You could win one of the lower amounts and cover your losses but in the long run you will probably run out of money!

Lotteries are a poor bet to ‘get rich quick’. Professional gamblers avoid them because they don’t want to waste their money on such a bad bargain. Politicians usually avoid them because ‘they’re afraid they might win and somebody might find out and think it was rigged’. But why is such a poor bet so popular?

Here are a few reasons for its popularity:

* General lack of knowledge of how bad the odds are of winning plays a part but so does good advertising. Many people like myself play the Lotto for different reasons:
* R2.50 is very little to spend on a lot of entertainment
* The proceeds are specifically designed to help the needy through the organizations supported by the Lotto
* If we didn’t have the Lotto the government would have to raise taxes.

The major attraction of the Lotto is probably the possibility of wealth no matter how unlikely the top prize is to win. After all, my chances of winning a million are a lot better than my chances of earning a million!

Let’s see how the Lotto works. The Lotto has seven divisions of winnings:

* Division 1: 6 numbers correct (10%)
* Division 2: 5 numbers correct plus the bonus ball (2%)
* Division 3: 5 numbers correct (4%)
* Division 4: 4 numbers correct plus the bonus ball (2.5%)
* Division 5: 4 numbers correct (7.5%)
* Division 6: 3 numbers correct plus the bonus ball (5%)
* Division 7: 3 numbers correct (16%).

The percentage payout has been very stable since the Lotto began, probably because it was designed that way. The percentages in brackets represent the amount of money that is paid out in each category. Remember if you and 99 other people win a particular category, all 100 split the earnings equally!

So for every million Rand (R1,000,000) that is paid to purchase tickets the payout is going to be:

* Division 1: R100,000
* Division 2: R20,000
* Division 3: R40,000
* Division 4: R25,000
* Division 5: R75,000
* Division 6: R50,000
* Division 7: R160,000.

In total, for each R1,000,000 received by ticket sales, R470,000 is paid out in winnings and R530,000 is retained by the National Lottery. For each Lotto ticket sold at R2.50, Lotto pays out R1.175 and retains R1.325. Now that’s a good return!

From the amount that Lotto retains, it has to pay for its own expenses including salaries, advertising and all the entertainment they provide with the actual and special draws that contestants just can’t resist.

### Mutually Exclusive Events

In simple terms, two events are mutually exclusive if they cannot occur at the same time (i.e. they have no outcomes in common). In short, mutual exclusivity implies that at ***most*** one of the events may occur.

In logic, two **mutually exclusive** propositions are propositions that logically cannot both be true. If an event X means that another event Y does not take place, X and Y are called mutually exclusive.

For example, when tossing a coin, it can only land on either head or tails, not both. The two events are mutually exclusive.

The probability of occurrence of two or more mutually exclusive events is obtained by adding the probabilities of the individual events. If the sum is equal to 1, we call them complementary events.

### Independent Events

In probability theory, to say that two events are **independent,** intuitively means that the occurrence of one event makes it neither more nor less probable that the other occurs. For example:

* The event of getting a 6 the first time a die is rolled and the event of getting a 6 the second time are **independent**.
* By contrast, the event of getting a 6 the first time a die is rolled and the event that the sum of the numbers seen on the first and second rolls is 8 are **dependent**.
* If two cards are drawn *with* replacement from a deck of cards, the event of drawing a red card on the first trial and that of drawing a red card on the second trial are *independent*.
* By contrast, if two cards are drawn *without* replacement from a deck of cards, the event of drawing a red card on the first trial and that of drawing a red card on the second trial are *dependent*.

When **two** coins are tossed, the probability that one will land on heads is ½ and the probability that the other one will land on heads is also ½.

These two events are **not mutually exclusive**. They are **independent**. The probability of the occurrence together of two or more independent events is obtained by multiplying the probabilities of the individual events.

Formative assessment

Make a list of all the possible outcomes if two coins A and B are tossed.

|  |  |
| --- | --- |
| If two coins are tossed 50 times, how many times do you expect to get two heads |  |
| If two coins are tossed 50 times, how many times do you expect to get one tail and one head? |  |

Make a list of all possible outcomes if two dices X and Y are rolled simultaneously.

|  |  |
| --- | --- |
| Make a list of the different outcomes of which the total is 6. | 5: 5+1; 1+5; 4+2; 2+4; 3+3 |
| What is the probability that, if two dice are rolled simultaneously, the total is 6? | 5/12 |

### Summary

This module dealt with statistical and probability theory in simple terms and used very little mathematics. The mathematics and equations that are used are reduced as much as possible while still retaining the message of statistics. The three major areas of statistics covered are producing data, organizing and analysing data and drawing conclusions from data.

In the section on producing data, various techniques and examples are discussed. These include several sampling techniques that should be used and others that should be avoided. Surveys and questionnaires were discussed as well as the necessary topic of ethics when dealing with data and people.

In the section on organizing and analysing data, techniques for validity and accuracy of data are discussed along with the different measurement scales used to record data. This section also shows how to display data and distribution as well as giving the formulas and techniques to display data and distributions.

In the section of drawing conclusions from data, simple probability theory is covered and includes a few comments on the topics close to us all. This section also indicates how to be confident about our results.

Learn to use numbers and statistics to ensure you are a responsible and participating citizen in our society.

# PROBABILITY AND STATISTICAL MODELS

#### Specific outcome

Critically interrogate and use probability and statistical models in problem solving and decision making in real world situations

#### Assessment criteria

* Statistics generated from the data are interpreted meaningfully and interpretations are justified or critiqued.
* Assumptions made in the collection or generation of data and statistics are defined or critiqued appropriately.
* Tables, diagrams, charts and graphs are used or critiqued appropriately in the analysis and representation of data, statistics and probability values
* Predictions, conclusions and judgements are made on the basis of valid arguments and supporting data, statistics and probability models
* Evaluations of the statistics identify potential sources of bias, errors in measurement, potential uses and misuses and their effects on arguments, judgements, conclusions and ultimately the audience

## Analyse The Data

Now you have all the information, you have to do something with it: analyse the information. The only reason you went to all the trouble of conducting market research is that you want information, so that you can analyse the information and use it for the benefit of your new venture.

Once you have all the information, you have to organise it, summarise it and simplify the information so that you can make sense of it.

The easiest would be to prepare a document that lists all the questions. You then count all the answers and add the totals to the document. If we use the questions from a previous Formative Assessment:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Do you play computer games? | Yes | 1450 | No | 550 |

Now you can calculate a percentage of the sample: who plays computer games and therefore might be interested in buying a new game.

* Total questionnaires received: 2000
* Total Yes 1450 percentage of sample: 72%
* Total No 550 percentage of sample 18%

This means that, of the people who took part in the market research, 72% do play computer games.

Formative assessment

Analyse the responses for the following questions:

|  |  |  |
| --- | --- | --- |
| Why do you use a taxi to and from work | Cheap | 1631 |
| Fast | 1091 |
| Safe | 312 |
| Convenient | 1849 |

If you sent out 2000 questionnaires, what percentage of the sample use taxis because they are cheap, what percentage use taxis because they are fast, what percentage use taxis because they are safe and what percentage use taxis because they are convenient?

|  |  |
| --- | --- |
| Cheap  | 81% |
| Fast  | 54% |
| Safe | 16% |
| Convenient  | 92% |

|  |  |  |
| --- | --- | --- |
| Which taxi route do you use every day? | Route A | 755 |
| Route B | 830 |
| Route C | 415 |

What percentage uses Route A, Route B and Route C?

|  |  |
| --- | --- |
| Route A | 38% |
| Route B | 42% |
| Route C | 20% |

You would do this for all the questions.

On the basis of the above information, you can now make a decision as to whether there is actually place in the market for a new taxi service, or whether your improved service will satisfy a need with the customers.

Formative assessment

In a group, analyse the responses to the questions that you received from the other group during your pilot survey. Write down your conclusions.

|  |
| --- |
|  |
|  |
|  |
|  |
|  |
|  |
|  |
|  |

## Determining Trends Using Statistics

We determine and identify trends in regarding issues that affect our society, such as

* crime and health;
* relevant characteristics of targets such as age, range, gender, socio-economic , cultural belief and performance;
* and the attitudes or opinions of people on issues by doing research as indicated and then by interpreting this information statistically.

Once we have the data we look at common grounds and averages to determine these trends.

### Mean, median, and mode

Mean, median, and mode are three kinds of "averages". There are many "averages" in statistics, but these are, I think, the three most common, and are certainly the three you are most likely to encounter in your pre-statistics courses, if the topic comes up at all.

* The "mean" is the "average" you're used to, where you add up all the numbers and then divide by the number of numbers.
* The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in numerical order, so you may have to rewrite your list first.
* The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.

## Describing data

Tables organize data and graphs present a clear overall picture of the distribution and spread of the data. However, I want to be able to describe the data in specific detail. I want to say that the centre of the data is ‘here’ and it has a spread ‘this large’.

This section shows us how to calculate the centre of any distribution of values and to tell how spread the values are from the centre.

In general, the set of numbers used in this section is kept to a minimum. The techniques described, however, work on any number of items. A smaller set of numbers allows the principles to be demonstrated by hand or with a calculator. A large set of numbers would just become an exercise in concentration and patience, unless of course, you are using a computer to do the work for you.

### Measuring centre or average

I have a lot of data that I’ve collected and I would like to represent that data by one number: its centre. I want to know the average value of my data.

It turns out that if there isn’t just one way of measuring the centre of my data but there are many ways to do so. All are correct and all are used for different reasons. But the most commonly used methods are the mean, the median and the mode. In statistics-talk these are three ‘measurements of central tendency’.

#### Calculating centres and averages

When calculating centres and averages we are trying to determine where the middle of the data are. There are three main methods used for calculating the centre of data: mean or average, mode and median.

#### Calculating the mean or average

The arithmetic mean is the average we all know and use: add up all the numbers and divide by their count (how many there are).

The mean of the following 5 numbers: 8, 3, 12, 5 and 10 is:



The ‘5’ under the line (in the denominator) is the number of items or the count of items and has nothing to do with the value ‘5’ above the line (in the numerator).

If I had 50 numbers I would add all 50 numbers together and divide by 50.

**When do we use it?** Suppose, for example, that at a party there are ten people aged 14, 15, 16, 14, 15, 16, 16, 15, 60, and 65 respectively. The mean of these ages is 24.6 which are not at all typical of the people at the party. A better statistic would be a median.

#### Calculating the mode

**The mode is simply the number that occurs most frequently.**

The mode of this sequence of numbers (2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12 and 18) is 9. The number ‘9’ occurs the most number of times. This set of numbers is call ‘unimodal’ because it has one mode. (The set of numbers used to calculate the mean has no mode because all numbers occur the same number of times.)

The following sequence of numbers (2, 3, 4, 4, 4, 5, 5, 7, 7, 7 and 9) has two modes and is called ‘bimodal’.

If a set of numbers has more than two modes the set is called ‘multimodal’.

**Mode is used if you have the same number occurring so frequently in a set of data that it can be regarded as the typical item**. Suppose, for example that members of a group are asked to contribute to a gift for another person, and the contributions are

R1, R1, R2, R100, R100, R0.50, R2, R1, R1, R100, R100, R100.

This set of data can be described well by saying that the majority of people each contributed R100.

The mode is useful when dealing with nominal data (grouped data) like eye colour or ordinal data (ordered data) like shoe sizes. The mode is the value occurring most frequently. The shoe salesperson might have the following range of shoes with the associated number of sales during one month as shown in

**Shoe sizes and quantity sold**

The mode is the most useful measure of the centre of this data: size 10½. The salesperson now has an idea of the most common shoe size. To look at it another way: if I averaged the shoe sizes and told the salesperson the average shoe size you sold this month is 9¼, do you think this ‘statistic’ is meaningful? I don’t!

As an exercise in visualizing a table of data, imagine the previous data plotted as a distribution or histogram. Can you visualize the distribution? Frequency or relative frequency distributions are most commonly displayed in histograms. The ages that were displayed in the pyramid diagram are two histograms placed back to back

Histograms look like bar charts but they are different.

* Histograms may only be construction of data on an interval/ratio scale (age, length, height). Divide the range of data into classes of equal width. In the example dealing with ages, groups were formed of people with ages in five-year intervals (0-4, 5-9, 10-14). You must ensure that the classes are precisely specified and no data can fall into more than one class.
* Count the number of observations in each class. These counts are the frequencies of the classes.
* Draw the histogram remembering to keep the data scale horizontal and the frequency scale vertical. Each bar represents a class. The base of the bar covers the class and the bar height is the class frequency. A histogram is drawn with no horizontal space between the bars unless a class is empty (has zero height).
* Bars in a histogram are vertical and the base scale is marked off in equal units. Bars in bar charts may be vertical or horizontal and there is no base scale. In other words, I may use any type of scale in a bar chart although nominal or ordinal scales are the most commonly used.
* The widths of the bars in a histogram have meaning. The base of each bar covers a class (interval) of values and the height represents the class frequency. The widths of bar charts have no meaning but they should all be the same so that they do not confuse or deceive you.
* The bars in a histogram touch each other unless there is a class with no entries. The bars cover an entire range of values. Even when the values of a variable have gaps between them, we extend the bases to meet halfway.

Remember that our eyes and brains respond to the area of the bars in a histogram. Therefore, all the widths must be equal, just like in a bar chart and for the same reason: so that they do not confuse or deceive you.

#### Calculating the median

The median is the middle value of the set of sorted values. Note that the data must be sorted before you can calculate the median by hand. A calculator or computer automatically sorts the set of numbers before calculating the mean.

When we calculate the median we are not concerned with the values of the set of data but rather the position of each number in the sorted order. This is important so pay attention.

There are two separate problems with calculating the median: an odd number of data values and an even number of data values.

For an odd number of data values (3, 4, 4, 5, 6, 8, 8, 8 and 10) select the middle value: 6. If you look at this data set you will see that there are four values below (smaller value) and four values above (larger value) 6. Therefore, 6 is the median.

For an even number of data values (5, 5, 7, 9, 11, 12, 15 and 18) take the average of the middle two values. In this case the median is 10. That is ½(9 + 11) = 10. The values 9 and 11 are the middle two values because they each have three values below and three values above them. Just average the middle two numbers.

Formative assessment

Follow the steps as detailed

mean = average

**Find the mean, median, and mode for the following list of values: 13, 18, 13, 14, 13, 16, 14, 21, 13**

The **mean** is the usual average, so:

(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15

Note that the mean isn't a value from the original list. This is a common result. You should not assume that your mean will be one of your original numbers.

The **median** is the middle value, so I'll have to rewrite the list in order:

13, 13, 13, 13, 14, 14, 16, 18, 21

There are nine numbers in the list, so the middle one will be the (9 + 1) ÷ 2 = 10 ÷ 2 = 5th number:

median = middle

13, 13, 13, 13, 14, 14, 16, 18, 21

So the median is 14.

The **mode** is the number that is repeated more often than any other, so 13 is the mode.

mean: 15

mode = most often

median: 14

mode: 13

Find the mean, median, and mode for the following list of values: 1, 2, 4, 7

The mean is the usual average: (1 + 2 + 4 + 7) ÷ 4 = 14 ÷ 4 = 3.5

The median is the middle number. In this example, the numbers are already listed in numerical order, so I don't have to rewrite the list. But there is no "middle" number, because there are an even number of numbers. In this case, the median is the mean (the usual average) of the middle two values: (2 + 4) ÷ 2 = 6 ÷ 2 = 3

The mode is the number that is repeated most often, but all the numbers appear only once. Then there is no mode.

mean: 3.5

median: 3

mode: none

The list values were whole numbers, but the mean was a decimal value. Getting a decimal value for the mean (or for the median, if you have an even number of data points) is perfectly okay; don't round your answers to try to match the format of the other numbers.

Find the mean, median, and mode for the following list of values:  8, 9, 10, 10, 10, 11, 11, 11, 12, 13

The mean is the usual average: (8 + 9 + 10 + 10 + 10 + 11 + 11 + 11 + 12 + 13) ÷ 10 = 105 ÷ 10 = 10.5

The median is the middle value. In a list of ten values, that will be the (10 + 1) ÷ 2 = 5.5th value; that is, I'll need to average the fifth and sixth numbers to find the median: (10 + 11) ÷ 2 = 21 ÷ 2 = 10.5

The mode is the number repeated most often. This list has two values that are repeated three times.

About the only hard part of finding the mean, median, and mode is keeping straight which "average" is which. Just remember the following:

* mean: regular meaning
* median: middle
* mode: most often

### Frequency distribution and range

When you are drawing a chart or graph, you will need the frequency distribution and range of the information in order for the graph or chart to make sense.

The range is defined as the difference between the largest and smallest values in the data set.

range = largest - smallest

The range is one measure of the spread of a set of data. If the range is very large we may expect the values in the data set to be spread widely.

#### Definitions

* **Frequency distribution**: where you arrange (distribute) data in some kind of order. A frequency distribution tells you how often certain numbers or values occur.
* **Population**: The objects we are busy investigating (in the example on the next page, the learners in this class)
* **Range**: The difference between the lowest and highest items in a set of data is called the range of the data set.

Formative assessment

Use the names of learners in your class and the number of children they have to complete the table on the following page:

|  |  |
| --- | --- |
| Names of Learners in your class | Number of Children |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 1 |   |   |   |   |   |   |
| 2 |   |   |   |   |   |   |
| 3 |   |   |   |   |   |   |
| 4 |   |   |   |   |   |   |
| 5 |   |   |   |   |   |   |
| 6 |   |   |   |   |   |   |
| 7 |   |   |   |   |   |   |
| 8 |   |   |   |   |   |   |
| 9 |   |   |   |   |   |   |
| 10 |   |   |   |   |   |   |
| 11 |   |   |   |   |   |   |
| 12 |   |   |   |   |   |   |
| 13 |  |  |  |  |  |  |
| 14 |  |  |  |  |  |  |
| 15 |  |  |  |  |  |  |
| 16 |  |  |  |  |  |  |
| 17 |  |  |  |  |  |  |
| 18 |  |  |  |  |  |  |
| 19 |  |  |  |  |  |  |
| 20 |  |  |  |  |  |  |
| 21 |  |  |  |  |  |  |
| 22 |  |  |  |  |  |  |

We would like to indicate how many learners (number of cases) fall within the class intervals.

The class intervals range from 0 – 1children, 2 – 3 children etc.

|  |  |  |
| --- | --- | --- |
| **Class Interval** | **Tally (number of learners)** | **Number of children** |
| **0 - 1** |   |   |
| **2 - 3** |   |   |
| **4 - 5** |   |   |
| **6 and more** |   |   |
| **TOTAL:** |   |   |

What is the range of the data set?

|  |
| --- |
|  |
|  |

## Graphical Representation Of Data

The aim of statistics is to provide insight by means of numbers. In order to achieve this we must collect numbers that are valid in the sense of being both correct and relevant to the issue at hand.

Once you have analysed the information, you will want to present it in such a way that everyone understands the information.

Data are presented in tables and graphs. This section covers some simple ways to represent data in graphs and to begin seeing how these numbers are distributed. The ‘distribution’ of a variable simply describes the values the variable takes on and how often each value occurs.

We want to use our knowledge of statistics to communicate facts and to support decisions. Data, like words, need to be organized in order to tell us anything useful. Tables of data are usually very large and I believe the best place for them is in a computer storage device or in an archive somewhere. I may need to get at these data for some or other reason but I don’t want their volume to confuse me. I like simple things that I can understand.

### Displaying data

I am going to start of with a table of data, taken from the 2001 census and published by Stats SA.

|  |  |  |  |
| --- | --- | --- | --- |
| **Language** | **Male** | **Female** | **Total** |
| **Afrikaans** | 2,900,214 | 3,083,212 | 5,983,426 |
| **English** | 1,772,483 | 1,900,720 | 3,673,203 |
| **IsiNdebele** | 342,366 | 369,455 | 711,821 |
| **IsiXhosa** | 3,726,376 | 4,180,777 | 7,907,153 |
| **IsiZulu** | 5,045,450 | 5,631,855 | 10,677,305 |
| **Sepedi** | 1,987,170 | 2,221,810 | 4,208,980 |
| **Sesotho** | 1,704,071 | 1,851,115 | 3,555,186 |
| **Setswana** | 1,774,785 | 1,902,231 | 3,677,016 |
| **SiSwati** | 571,429 | 623,002 | 1,194,431 |
| **Tshivenda** | 482,134 | 539,623 | 1,021,757 |
| **Xitsonga** | 1,001,446 | 990,761 | 1,992,207 |
| **Other** | 126,117 | 91,175 | 217,292 |
| **Total** | 21,434,041 | 23,385,736 | 44,819,777 |

*First home language by gender: South Africa, 2001*

This table lists the languages along with the number of people, grouped by gender, who use this language as their home language. Look at the table. It contains a few very important details.

Firstly, a table is made up of rows and columns. The rows extend horizontally across the page and the columns extend vertically down the page. Secondly, the table has a caption that describes the contents of the table. In this case the caption is below the table while in other cases the caption may be above the table. This table has a total column as its last row that may or may not appear in other tables. Thirdly, each column is labelled (Language, Male, Female and Total) so that you know what is in each column of data.

A table is pretty simple. It lists data in rows and columns and you can find the information you are looking for by going to either the row or columns of interest and looking either across or down. The intersection of the row and column (where they cross) is the data you require. In this example, if I want to find out how many people reported that they speak a home language that is not one of the official nine languages, I would look in the row called ‘Other’ and read the number to its right in the ‘Total’ column: 217,292.

### Frequency tables

One of the first things to do when organizing a set of data is usually to count how often each value occurs. Stats SA kindly did this for us and presented the data in the form of a table

Because rates or proportions are often more useful than totals, we calculate these and display them per the table on the next page.

The technique used to create a table of proportions or ratios is simple but can become a little tedious without a calculator or a computer program to do it for you. If you are doing this by hand, that is, actually dividing the numbers, you will probably make a few mistakes. So check your work again.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Male** | **Female** | **Total** |
| **Afrikaans** | 0.14 | 0.13 | 0.13 |
| **English** | 0.08 | 0.08 | 0.08 |
| **IsiNdebele** | 0.02 | 0.02 | 0.02 |
| **IsiXhosa** | 0.17 | 0.18 | 0.18 |
| **IsiZulu** | 0.24 | 0.24 | 0.24 |
| **Sepedi** | 0.09 | 0.10 | 0.09 |
| **Sesotho** | 0.08 | 0.08 | 0.08 |
| **Setswana** | 0.08 | 0.08 | 0.08 |
| **SiSwati** | 0.03 | 0.03 | 0.03 |
| **Tshivenda** | 0.02 | 0.02 | 0.02 |
| **Xitsonga** | 0.05 | 0.04 | 0.04 |
| **Other** | 0.01 | 0.00 | 0.00 |
| **Total** | 1.00 | 1.00 | 1.00  |

The basic idea of creating a proportion of a ratio is to divide the individual value by the total value. In the case of our example, the totals are already given, so you probably won’t have to calculate them again. Or should you? Ensuring your data is correct is called ‘internal consistency’.

Formative assessment

Do the calculations that follow

I’ll show you the calculations for the first three languages for ‘Male’, ‘Female’ and ‘Total’ in order to show you how easy it is to do.

Male:



Female:



Total:



If you check the results shown in the table from Stats SA, by summing each column you will find that the sums are 1.01, 1.00 and 0.99 for the ‘Male’, ‘Female’ and ‘Total’ columns. What happened? The arithmetic is correct but when I rounded the fractions to two decimal places, a little precision was lost. These errors are called ‘round-off errors’ or ‘rounding errors’ and they will be with us any time we round numbers. You’ll get used to it!

* The ‘frequency’ of any value of a variable is the number of times that value occurs in the data. A frequency is a count.
* The ‘relative frequency’ of any value is the proportion or fraction or per cent of all observations that have that value.

In the example above, the total number of South Africans who use Afrikaans as a first language is 5,983,426. This is a frequency because it’s a count of something: first language Afrikaans speakers.

The relative frequency is usually expressed in decimal form and in this case is 0.13. However, we could just as well express this value as a percentage (13%). Remember that 1% is  or 0.01. A number in decimal form can be changed to a percentage by moving the decimal point two places to the right. (This is the same as multiplying the decimal form by 100 and putting a per cent sign, %, behind the result. ‘Per cent’ means ‘by 100’ or ‘per 100’,)

Frequencies and relative frequencies are a very common way of summarizing data when a nominal scale is used (gender, responses on questionnaires, eye colour).

In fact, it is such a handy way of summarizing facts that it is often used with an interval/ratio scale. In this case, we artificially group items and then count how many items fall into each group.

In this example, again taken from the census figures of Stats SA, we are looking at the age of a person. Age is measured on an interval/ratio and is a continuous value.

The table below shows the actual values of the age of all South Africans grouped into five years age groups. The table also displays the percentage of each age group. Note that the column totals do not sum to 100%. I only used three decimal figures to calculate the results in order to demonstrate the potential problems with round-off errors.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age** | **Male** | **%** | **Female** | **%** | **Total** | **%** |
| 0-4 | 2,223,731 | 10.3% | 2,226,085 | 9.5% | 4,449,816 | 9.9% |
| 5-9 | 2,425,804 | 11.3% | 2,427,751 | 10.3% | 4,853,555 | 10.8% |
| 10-14 | 2,518,956 | 11.7% | 2,542,961 | 10.8% | 5,061,917 | 11.2% |
| 15-19 | 2,453,079 | 11.4% | 2,528,642 | 10.8% | 4,981,721 | 11.1% |
| 20-24 | 2,099,293 | 9.7% | 2,195,230 | 9.3% | 4,294,523 | 9.5% |
| 25-29 | 1,899,124 | 8.8% | 2,035,814 | 8.7% | 3,934,938 | 8.7% |
| 30-34 | 1,594,488 | 7.4% | 1,746,412 | 7.4% | 3,340,900 | 7.4% |
| 35-39 | 1,441,507 | 6.7% | 1,630,264 | 6.9% | 3,071,771 | 6.8% |
| 40-44 | 1,233,632 | 5.7% | 1,385,832 | 5.9% | 2,619,464 | 5.8% |
| 45-49 | 967,604 | 4.5% | 1,119,776 | 4.7% | 2,087,380 | 4.6% |
| 50-54 | 769,499 | 3.5% | 868,521 | 3.7% | 1,638,020 | 3.6% |
| 55-59 | 552,323 | 2.5% | 652,943 | 2.7% | 1,205,266 | 2.6% |
| 60-64 | 444,510 | 2.0% | 620,784 | 2.6% | 1,065,294 | 2.3% |
| 65-69 | 304,763 | 1.4% | 483,164 | 2.0% | 787,927 | 1.7% |
| 70-74 | 232,547 | 1.0% | 398,922 | 1.7% | 631,469 | 1.4% |
| 75-79 | 136,436 | 0.6% | 231,101 | 0.9% | 367,537 | 0.8% |
| 80-84 | 90,835 | 0.4% | 180,111 | 0.7% | 270,946 | 0.6% |
| 85+ | 45,907 | 0.2% | 111,425 | 0.4% | 157,332 | 0.3% |
| Total | 21,434,038 | 99.1% | 23,385,738 | 99.0% | 44,819,777 | 99.1% |

*Age distribution and ratio in five-year intervals by gender: South Africa 2001*

## Graphing data

Graphs, like tables, should be clearly labelled to show the variables that are being presented and the units being used. There are three things to remember when putting data in a graph:

* Make your data stand out
* Avoid clutter on the graph
* Use visual perception to get the facts to others.

An excellent example of a graph is shown in one the next page. This graph is the one Stats SA distributed with the age information in the previous example.

The impact of this graph along with its clarity of presentation is striking. This type of graph is called a pyramid or butterfly graph because of its shape and it belongs in the category of bar charts.

The vertical axis displays the age categories and the horizontal axis displays the percentage of each age group according to male and female. The data for the male groups is in green and is shown on the left while that for the female groups is shown on the right. This type of graph allows you to compare two groups of items at the same time.

I’m giving a bit of interpretation of this graph so that you can get the feel of looking at a picture and still see data. The very top of the graph shows that there are very few of us in the 85+ year group. This makes sense because as the population ages its population declines. In other words, the further the green or yellow lines are from the centre the more people there are in that age group. Conversely, the closer the green or yellow lines are to the centre the fewer people there are in that age group.

In addition, the female population slightly outnumbers the male population in every category with the exception of the bottom two where they are approximately equal. This makes sense from two viewpoints. Firstly, the female population is slightly greater than the male population (52% to 48%, respectively). Secondly, females tend to live about five years longer than males. This may also be seen in the top end of the graph (above the 35-39 year group) where the male to female ration starts declining. At the 70-74 year groups and up the number of females is approximately double the number of males.

.

**Age distribution of males and females in the total population, Census 2001**



### Bar graphs

Bar graphs compare measurements at intervals, the bars run horizontally. Column charts compare measurements at intervals and provide a view of data at a specific time. The bars run vertically

The example below shows a column chart indicating how many ice creams were sold from January to May.

If you use a bar chart, the bars will run horizontally and not vertically as with a column chart.

**Number Of Ice Creams Sold**



In a bar chart, the heights of the bars are important.

When you draw a bar graph, state clearly what you are representing on the two axes. This means that you have to label the axes. Also insert it on the graph above.

* Draw the axes at right angles to each other
* Choose a scale for the vertical axis and write in the units.
* Use a ruler to help you read off the height of a bar.

Formative assessment

Study the chart on the next page and answer the questions below:

|  |  |
| --- | --- |
| How many ice creams were sold during January? | 92 |
| During which month were the most ice creams sold? | January and February |
| In which month were 70 ice creams sold? | March |
| What was the total number of ice creams sold for the period? | 350 |
| What is the range of the data set? | 90-30=60 |



The bars of the chart may be vertical or horizontal. They may touch each other as shown in the figure or they may overlap each other or be separated from each other. But be careful when creating your own bar chart! Each bar must have the same width because our eyes and mind respond to the area of the bars. When the bars have the same width and a height that varies with the variable then the area (height times width) also varies and our eyes and brains receive the correct impression.



### Line Graphs or Curves

Show the changes in data or trends over a given period of time. They are used to emphasize rather than compare. We use a dot to show the height of each bar. If we join the dots, we get a line graph.

0

10

20

30

40

50

60

70

80

90

100

Jan

Feb

March

April

May

Formative assessment

The example above is a column chart indicating how many ice creams were sold from January to May. Study the column chart and answer the questions below

|  |  |
| --- | --- |
| What tendencies do you pick up from this graph? | More ice cream is sold during hot summer months |
| When would be a good time to start a new ice cream business? | During the spring or summer |

Line graphs show the behaviour of a variable over time. Time is placed on the horizontal axis and the variable being plotted is shown on the vertical axis.

A good example of a time based line graph is the food index from Stats SA. You will find this chart on the page following the ice cream chart.



The graph makes it clear that the food index rose sharply between January 2002 and January 2003 after which its rise slowed somewhat.

As you can see in the graph, time (years and months) is displayed on the horizontal axis and the index value is displayed on the vertical axis. All time periods are of a fixed length and the length is the same for each month. I am pointing this out because I have seen graphs that try to distort facts by altering the graph in some respects.

A very handy feature with the line chart is that you may have several graphs on the same chart. This makes them a bit easier to compare. The line chart shows the Consumer Price Index for the historical metropolitan areas (CPI), the Consumer Price Index excluding interest rates on mortgage bonds (CPIX) and the Food Price Index (FPI). The FPI was shown on the last two figures.

In this graph we can see the same as before with the FPI, however, the CPI and CPIX are also shown in comparison. The interpretation of these index figures is not discussed here, however, you can see that more than one graph may be represented and comparisons may be made. The FPI represents the price of food, as you probably have already guessed. In comparison to the CPI and the CPIX note how sharply the FPI rose from about July 2002. If you remember, there was a big outcry concerning the cost of food and, in particular, basic food prices near the end of 2002. This graph shows you that rise.

### Pie Charts

The pie chart shows how a whole is divided into parts. The home language distribution is shown on the next page.

The pie chart is a good option to choose when you want to show the relationships that parts have to the total. In this example, all the home languages are compared. They have been ordered from largest (IsiZulu) to the smallest (Other) and are displayed in this manner in the pie chart. The legend on the right as well as the labels and colouring of each section of the pie make the pie chart easy to understand.

Pie charts show us the parts that make up the whole but humans don’t see angles as clearly as we see length.

For this reason, the pie chart is not a good choice to compare sizes of various parts with the whole. In addition, the divisions used in the example of the pie chart are causing the graph to become a bit crowded.

If I tried to do a pie chart of the age groups, it would probably look pretty messy! If not messy, it certainly would look crowded. I also think the 85+year group would be hard to see. There are alternatives to the pie chart!



A pie chart shows the breakdown of a total. A pie chart is a good way to show how a fixed number is divided. The whole circle (360ْ) represents the total number (or 100%) and we express each part as a fraction or percentage of the whole. A pie chart is constructed by converting the share of each component into a percentage of 360 degrees.

In the example below, a total of 2000 cars were sold in 1996 and the pie chart shows the breakdown of the 2000 cars: which manufacturer sold how many cars.



Ideally, your charts should look as follows:

Your information as a column chart



Your information as a bar chart



Pie Chart



## Comparing (Correlation) Of Data

When you do a survey, it is because you want to obtain certain information (data) and then you want to compare it to

* what it should be,
* what the rest of the world does,
* what the ideal should be, etc.

Only once you have analysed and compared the data can you come to conclusions about the aim of your survey:

* how many ice creams were sold
* how many cars were sold
* how many employees become ill during the winter
* how many passengers do you transport on a specific route
* how much fuel does your bus or taxi use, etc.

Of course, all graphs and charts serve to compare information (data) and help you to reach a conclusion, but a correlation plot is a graph that shows the relationship between two sets of scores or values. In the example below, between the number of people who are HIV/AIDS positive and the access to primary health care

This correlation plot is shown as a scatter graph (a line chart without lines)



## Summarising Data

The purpose of this section is to revise the calculation of averages by using mean, mode and median and also to give examples of when each average is appropriate.

There are often situations in which it is useful to summarise a whole set of data by describing it with a single number. The sum of a set of numbers can be used to summarise data (e.g. the total mass of the pack of rugby forwards is an indication of the possible power of a rugby team).

The total is not always a useful summarizing number, however. Suppose we wish to compare the heights of men and women in a group. The totals cannot be used for the data given because there are different numbers of men and women.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Height of 10 year olds(m) | 1.76 | 1.77 | 1.8 | 1.66 | 1.6 | 1.79 | 1.8 |  |  |
| Height of 9 year olds(m) | 1.69 | 1.7 | 1.5 | 1.42 | 1.42 | 1.75 | 1.67 | 1.62 | 1.6 |

Although the total for the 9 year olds (14.37 m) is higher than the total for the10 year olds (11.12 m), when we consider individual heights it seems that 10 year olds are typically taller than 9 year olds.

What we actually need is a single number which is typical or representative of the heights of individual 9 year and 10 year olds. One such a number is the arithmetic mean or average.

### Mean

**Arithmetic mean** = (sum of cases)/total number of cases

* Mean height of men = 11.12/7 = 1.59
* Mean height of women = 14.37/9 = 1.6

### Median

**Median** is the middle value in a spread of values arranged in order from the lowest to the highest.

The median of an even number of items is the mean of the two middle items when the items are arranged according to size.

The median of an odd number of items is the middle item when the items are arranged according to size.

When do we use it? Suppose, for example, that at a party there are ten people aged 14, 15, 16, 14, 15, 16, 16, 15, 60, and 65 respectively. The mean of these ages is 24.6 which are not at all typical of the people at the party. A better statistic would be a median.

### Mode

When numbers occur frequently in a set of data, the number occurring most frequently is the **mode.** This is used if you have the same number occurring so frequently in a set of data that it can be regarded as the typical item. Suppose, for example that members of a group are asked to contribute to a gift for another person, and the contributions are

R1, R1, R2, R100, R100, R0.50, R2, R1, R1, R100, R100, R100.

This set of data can be described well by saying that the majority of people each contributed R100.

For each set of data (information) that you have collected, you will have to decide which of the three statistics (mean, mode, and median) will give you the best description of the data.

### Using centres and averages

For each set of data (information) that you have collected, you will have to decide which of the three statistics (mean, mode, and median) will give you the best description of the data.

The **mode** is useful when dealing with nominal data (grouped data) like eye colour or ordinal data (ordered data) like shoe sizes

We now know how to calculate the mean, mode and average but we must still learn how and when to use them. Although the mode is very important in many real life situations it is not used that often in statistical calculations. In many situations the mode is not useful at all because there is no mode. The median is used more frequently than the mode because it is able to describe the data set with more flexibility. The median is also easily understood. The mean, or average, is the most commonly used measure of the centre of a set of data because it is backed by statistical theory. Let’s now see these different ways of measuring the centre of data in action.

Assume that I am the owner of a sporting body and I employ twelve sportsmen (or sportswomen). The annual salaries of each sportsman (sportswoman) are as follows:

* six receive R200,000
* four receive R400,000
* one receives R800,000
* the superstar receives R2,400,000
* What is their average salary?
* If I used the mode to calculate the average salary it would be R200,000 because this value occurs with the highest frequency (6 times).
* If I use the median I must first order the salaries as shown in the table below and then determine the middle value:

|  |  |
| --- | --- |
| Order | Salary |
| 1 | R200,000  |
| 2 | R200,000  |
| 3 | R200,000  |
| 4 | R200,000  |
| 5 | R200,000  |
| 6 | R200,000  |
| 7 | R400,000  |
| 8 | R400,000  |
| 9 | R400,000  |
| 10 | R400,000  |
| 11 | R800,000  |
| 12 | R2,400,000. |

There are an even number of entries (12) so I must obtain the average of the middle two values (R200,000 and R400,000) and indicate that the median is R300,000.

If I use the average (mean) I must sum all the salaries and divide by the number of players which is R500,000 (R60,000,000 divided into 12 players).

Which is the best number to use for the average salary?

The mode indicates that the highest frequency and most players receive R200,000. This number just doesn’t seem right to me even though half the players receive this salary and the other half receive more.

The average produces R500,000. No player receives this amount but then it is an average. However, 10 players receive less that R500,000 and only two receive more. Again, this just doesn’t look right to me.

The median also produces a value that no player receives but at least half the players receive less than this amount and half receive more than this amount. This is the middle of the road average so I would agree with it. As a matter of fact, if the superstar was paid R24,000,000, this measurement of the centre (R300,000) would not change and I would still have half of the players being less and half being paid more. However, if I used the average (mean) I would find that the average salary would be R2,300,000!

The median is a good choice to use for the central value when the distributions are skewed to the right or to the left. Salaries are almost always skewed to the right (very few people obtain very high salaries while most of the workers receive salaries that are on the lower end.) Note that the median value always has half the values on one side and half the values on the other side. There is always a middle value that exists or is the average of the two centre values. The values don’t matter, only their positions when sorted matter.

When the distribution of values is more or less symmetrical and there are no outliers, then the average or mean is the best value to use for the centre of the data set.

Values that skew the distribution of values of the data affect the mean or average, sometimes dramatically. If I did pay my superstar R24,000,000 and advertised that my average salary bill is R2,300,000, the other 11 players would be at my door asking why their salary is so low!