Table of Contents

INTRODUCTION .......................................................................................................................... 2

SECTION 1: INTRODUCTION TO STATISTICS ................................................................. 7

SECTION 2: THE NATURE OF DATA, DATA COLLECTION AND SOURCES .................. 11

SECTION 3: PRESENTATION OF DATA ................................................................................. 27

SECTION: 4 MANAGEMENT STATISTICS ............................................................................. 45

SECTION 5: PROBABILITY DISTRIBUTION FUNCTIONS .................................................. 71

SECTION 6: PREDICTION (CORRELATION AND REGRESSION) .................................... 101

SECTION 7: FORECASTING METHODS USING TIME SERIES ANALYSIS ...................... 118

BIBLIOGRAPHY ....................................................................................................................... 136
AIMS OF THIS MODULE

The broad aims of this module are to:

• provide a working knowledge of quantitative analyses which are essential for managers
• give the student a proper understanding of when to use appropriate techniques
• help the student interpret and evaluate the results from using these techniques
• perform calculations that aid in decision making

LEARNING OUTCOMES

At the end of the module the student should be able to:

• explain why quantitative techniques and calculations are important to a manager
• perform statistical analyses in practice and extract additional information from business data
• manipulate gathered (grouped and ungrouped) data through various statistical methods to generate useful information to support management decisions
• prepare and interpret reports expressed in statistical terms
• assess the validity of statistical findings and the relevance and reliability of results
INTRODUCTION
Introduction

Statistics are all around us. In fact it would be difficult to go through a full week without using statistics. So, what is statistics?

Statistics is the science of designing studies, gathering data, and then classifying, summarising, interpreting, and presenting these data to support the decisions that are needed.

The word "statistics" is used in several different senses. In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing data, interpreting data, displaying data, and making decisions based on data. This is what courses in "statistics" generally cover.

In a second usage, a "statistic" is defined as a numerical quantity (such as the mean) calculated in a sample. Such statistics are used to estimate parameters. A parameter is a characteristic of a population eg. the mean. A statistic is a measure based on sample data.

Without statistics we couldn't plan our budgets, pay our taxes, enjoy games to their fullest, evaluate classroom performance. Are you beginning to get the picture? We need statistics.

Let's take a look at the most basic form of statistics, known as descriptive statistics. This branch of statistics lays the foundation for all statistical knowledge (pretty important, huh?), but it is not something that you should learn simply so you can use it in the distant future. Descriptive statistics can be used NOW, in English class, in physics class, in history, at the football stadium, in the grocery store. You probably already know more about these statistics than you think.

Another form of statistics is called inferential statistics, which deals with inferences (decision making, predictions, etc) about the process or population being studied.

This module has been divided into five sections and at the beginning of each section you will find a list of Learning Outcomes. These provide you with an outline of what you should have learned by the time you have completed the section, and can also be used to focus your study.

The concepts are best learned and understood through practice, and we thus suggest that you do all the exercises recommended throughout the guide, as well as carefully going through all the worked examples. A good suggestion is trying to work out the given examples without looking at the solution, then comparing your answers with the model answers. In this way, you can keep track of your progress.
Contents and Structure

Section 1: Introduction

In the introduction, we explain why it is important to be able to manipulate numbers and have a feel for magnitude. We also discuss how best to carry out the exercises in the text, and what basic pre-work is needed.

Section 2: The Nature of Data and Data Collection

This section deals with “data”, which is raw information. We discuss the different types of data that is available to us, and where to source it, as well as techniques for data collection.

Section 3: Presentation of Data

We discuss how to structure and present data in an understandable format, and which formats are appropriate to use. In particular, the use of tables and graphs are discussed.

Section 4: Management Statistics

In this section, statistical measures and behaviours of random variables are presented and analysed.

Section 5: Probability distributions

This section of the module deals with the Binomial (discrete) and Normal (continuous) distribution functions. The difference between each distribution is discussed and probabilities are calculated using these two distribution functions.

Section 6: Prediction (Correlation and Regression)

Here we discuss the least squares method which is essentially deriving a straight line graph to fit a set of data points and use the regression line to make predictions. In addition, we discuss the strength of the relationship between two variables using the Pearson correlation co-efficient.
Section 7: Forecasting Methods using Time Series Analysis

Different forecasting techniques and their merits are demonstrated, as well as moving averages.

How to use the Manual

Don’t try to complete the manual in a few long sessions. You will study more effectively if you divide your study into two-hour sessions.

If you want to take a break it would be a good idea to stop at the end of a section.

As you work through the manual you will come across Activities and Self Assessment Exercises. These are designed to help you study and prepare for the examinations.

Access to a Computer

The exercises given to you in this module will be arduous to do without a personal computer. Before doing the exercises on a computer, make sure you are able to apply the techniques manually, to help you better understand the calculations carried out by the computer.

ACTIVITY

Activities ask you to carry out specific tasks. In most cases there are no right or wrong answers to the Activities. The aim of the Activities is to give you an opportunity to apply what you have learned.

SELF CHECK ACTIVITY

Occasionally you will be required to assess your grasp of concepts by applying concepts to specific situations. Suggested answers to these activities are provided at the end of the specific unit.
Module Assessment

Assignment

You will be required to complete and submit an assignment. This assignment is assessed as part of your coursework. Therefore, it is very important that you complete it.

Examination

An examination will be written at the end of each semester. The assessment strategy will focus on application of theory to practice.

READING

Prescribed Reading:

This manual has been designed to be read in conjunction with the following textbook:

SECTION 1: INTRODUCTION TO STATISTICS
A statistic is an algebraic expression combining scores into a single number. Statistics serve two functions: they estimate parameters in population models and they describe the data.

**Statistics consists of the principles and methods for Designing studies**

1. Collecting data
2. Presenting and analysing data
3. Interpreting the results

**Statistics has been described as**

- Turning data into information
- Data-based decision making
- The technology of the "Scientific Method"

**The scientific approach to decision making can be summarised as:**

Hypothesis ➔ Data ➔ Conclusion

You are probably asking yourself the question, "When and where will I use statistics?". If you read any newspaper or watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a news program on television, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."
Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques to analyze the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Think back to some of the business conversations you have had during the last week or two, and the type of information that has been offered during these conversations. More than likely, you will have heard expressions like:

- The cost of that material is R1,500-00
- 80% of those that responded thought it was a good idea
- it took him two hours to complete the task
- 4,500 Toyota Corollas were sold in August

You would not be blamed for feeling a little frustrated with these expressions, because they don’t really give you good information. In each case, we are missing:

- a basis of comparison
- the source of the data
- how it was collected
- why it was collected in the first instance

If these criteria were in place, we would have some information to assist us in our decision-making.

For example, knowing that 4,500 Toyota Corollas were sold in August means nothing to me if I am selling Ford Escorts. But if I see that in July, 3,000 Toyota Corollas were sold, and the sales for Ford Escorts dropped from 2,000 to 500 in August, I now have a better understanding of the significance of that data. The reason for the drop in sales of the Ford Escorts may be due to the increase in Toyota sales. I now have information on which to make a decision on what to do next.

In this module, we will look at some of the techniques that can be applied by business managers to collect, analyse and interpret quantitative information to make informed decisions.
Numerical And Computer Literacy

Any manager who does not have numerate proficiency and computer literacy will be at a disadvantage in today’s business environment. It is important that you are able to do basic arithmetic manipulation of numbers and that you have a feel for magnitude. What do we mean by a feel for magnitude? It means you are able to recognise a number that is obviously wrong. Imagine if through a finger error, you calculated that your budget to replace ten new personal computers in your department was R8,000-00 instead of R80,000-00. If you didn’t notice the one zero missing before submitting your budget, you will be very embarrassed when the time to effect the purchase arrives.

ACTIVITY

Study Chapter 2, “Tools of the Trade” in the prescribed text, Wisniewski and Stead (1996) and carry out the self-assessment exercises. Ensure you are able to do these exercises before continuing.

ACTIVITY

Make sure that you have access to and are familiar with one of the commonly used spreadsheets (Excel, Quattro or Lotus 1-2-3), and learn how to use it before continuing with this module.
SECTION 2: THE NATURE OF DATA, DATA COLLECTION AND SOURCES
CONTENTS

- Data Sources
- Data Types
- Data Collection Methods

LEARNING OUTCOMES

The objectives of this chapter are:

- To convince the reader of the benefits of beginning any marketing research with a thorough search of secondary sources of data
- To articulate the advantages of secondary data
- To highlight the potential errors which can be hidden within secondary data
- To outline some of the main internal and external sources of data available to commercial enterprises, and
- To help the reader to recognize the transition, in marketing research, from a dependence upon published sources of secondary data to electronically stored secondary data.

The learner should be able to:

1. Describe various data collection techniques and state their uses and limitations.
2. Advantageously use a combination of different data collection techniques.
3. Identify various sources of bias in data collection and ways of preventing bias.
4. Identify ethical issues involved in the implementation of research and ways of ensuring that your research informants or subjects are not harmed by your study.
5. Establish a data source and determine the reliability of the data source.
6. Recognise the type of data being returned or presented.
7. Use the appropriate technique for data collection and understand the advantages and disadvantages of each data collection technique.
Data Sources

When one is confronted with a situation in which you need data, it is often surprising how much is actually available to us. Newspapers, magazines and the Internet, as well as internal company records have a wealth of data that can be put to good use.

Wegner (1999) refers to the following data sources:

- Internal
- External
- Primary
- Secondary

Internal Data Sources

Within our own organisations, internal data is generated during the course of normal business activities, for example:

Financial data – sales vouchers, credit notes, accounts receivable.
Production data – monthly production, defect rates, WIP levels.
Human Resource data – time sheets, staff demographics, wage and salary schedules.
Marketing data – monthly sales, advertising expenditure, customer profiles.
External Data Sources

Sources for data external to our own organisation may be private institutions, trade/ employer/ employee associations, profit motivated organisations and government bodies.

The cost of the external data depends on the source, but you may be surprised how much information is freely available, either on the Internet or in business publications. A detailed study of the economic indicators in financial publications will tell you a surprising amount. Statistics SA, the government’s source of statistical data, has virtually all its data available on its home page on the internet.

Many regard new motor vehicle sales as a good indicator of the economy – and these are published for all NAAMSA members monthly.

Private sources of information include:

- South African Chamber of Business (SACOB)
- Business Partners (previously Small Business Development Corporation)
- Industrial Development Corporation (IDC)
- Bureau of Economic Research
- Bureau of Market Research
- Bureau of Financial Analysis
- SA Labour Development Research Unit

Public Domain sources include:

- Newspapers, journals, trade magazines
- Reference libraries
- Bank economic reports
- Human Sciences Research Council (HSRC)
- Council for Scientific and Industrial Research (CSIR)

Primary Data Sources

Primary data - collected by the researcher himself, and is captured at the point where it is generated for the first time, normally with a specific purpose in mind. Examples include salary surveys and market research surveys.
The advantages of primary data:

- Primary data is directly relevant to the problem at hand.
- Primary data generally offers greater control over data accuracy.

The disadvantages of primary data:

- Primary data could be time consuming to collect.
- Primary data is generally more expensive to collect (ask a market research company for a quote!)

Secondary Data Sources

Secondary data is data which has been collected by individuals or agencies for purposes other than those of our particular research study. For example, if a government department has conducted a survey of, say, family food expenditures, then a food manufacturer might use this data in the organization's evaluations of the total potential market for a new product. Similarly, statistics prepared by a ministry on agricultural production will prove useful to a whole host of people and organizations, including those marketing agricultural supplies. Such data is already in existence either within or outside an organisation.

Some examples are:

- “Aged” market research figures.
- Previous financial statements.
- An industry market research from which you are extracting data for your company.

Advantages of secondary data:

- The data is already in existence.
- Access time is relatively short.
- The data is generally less expensive to acquire.

Disadvantages of secondary data:

- Data may not be problem specific or entirely relevant to your situation.
- Data may be dated and hence inappropriate.
- It may be difficult to determine the data accuracy.
• The data may not be suitable for further manipulation.
• Secondary data has often been manipulated by the time you receive it; you have no way of knowing how reliable the data is, what has been omitted or what has been extrapolated, which could lead you to wrong conclusions. i.e. has the data been "massaged"?

Data Types

Wegner (1999) suggests two reasons why an understanding of the nature of data is necessary:

• to assess data quality, and
• to select the appropriate statistical method to use to analyse the data.

The type of data gathered determines the type of analysis which can be performed; an incorrect application of a statistical method to a particular data type can render the findings invalid.

Data type is determined by the nature of the random variable that the data represents.

Wegner (1999) identifies two types of random variables; qualitative and quantitative.

**Quantitative** data measures either **how much** or **how many** of something, i.e. a set of observations where any single observation is a number that represents an **amount** or a **count**.

**Qualitative** data provide **labels**, or **names**, for categories of like items, i.e. a set of observations where any single observation is a word or code that represents a **class** or **category**

**Qualitative** random variables yield categorical or non-numeric responses. The data generated are classified into one of a number of categories.

Categories are usually represented by codes, which cannot be manipulated arithmetically. These codes are merely used as labels.
Figure 2.1 shows an example of data codes.

<table>
<thead>
<tr>
<th>Random variable</th>
<th>Response categories</th>
<th>Data codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management level</td>
<td>Supervisor</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Section Head</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Department Head</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>General Manager</td>
<td>4</td>
</tr>
<tr>
<td>Wine Preference</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>(Do you like red wine?)</td>
<td>No</td>
<td>1</td>
</tr>
</tbody>
</table>

*Figure 2.1: Data Codes for Qualitative Data*

Quantitative random variables yield numeric responses, and can be meaningfully manipulated using conventional arithmetic operations. Examples are age, distance, number of items, monetary amount, etc.

Each of these random variable categories can be associated with a different type of data classification. Wegner (1999) defines two data classification types:

- **Data type 1**
  - Nominal–scaled
  - Ordinal–scaled
  - Interval–scaled
  - Ratio

- **Data type 2**
  - Discrete
  - Continuous

Qualitative data can be divided into:

**Nominal variables**: Variables with no inherent order or ranking sequence, e.g. numbers used as names (group 1, group 2...), gender, etc.

**Ordinal variables**: Variables with an ordered series, e.g. "greatly dislike, moderately dislike, indifferent, moderately like, greatly like". Numbers assigned to such variables indicate rank order only - the "distance" between the numbers has no meaning.
**Interval variables:** Equally spaced variables, e.g. temperature. The difference between a temperature of 66 degrees and 67 degrees is taken to be the same as the difference between 76 degrees and 77 degrees. Interval variables do not have a true zero, e.g. 88 degrees is not necessarily double the temperature of 44 degrees.

**Nominal–scaled data** is associated mainly with qualitative random variables. There is no implied ordering between groups of the random variable, and each category is of equal importance. *Figure 2.1* is a good example.

**Ordinal-scaled data** is also associated mainly with qualitative random variables. Like nominal-scaled data, it is also assigned to one of a number of coded categories, but there is now a ranking implied between the categories in terms of being better, bigger, longer, older, taller or stronger, etc.

An example is shown in *Figure 2.2*.

<table>
<thead>
<tr>
<th>Qualitative Random Response</th>
<th>Response Categories</th>
<th>Data codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>T-shirt size</td>
<td>Small / medium / large</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>Turnover</td>
<td>&lt;5m / 5-10m / &gt;10m</td>
<td>1, 2, 3</td>
</tr>
</tbody>
</table>

*Figure 2.2: Example of ordinal-scaled data*

**Interval-scaled Data** is associated with quantitative random variables; differences can be measured between values. Interval-scaled data possesses both order (implied ranking) and distance properties. An example is shown in *Figure 2.3*.

<table>
<thead>
<tr>
<th>Indicate your response to the statement “Shopping is a social experience for me.”</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Tick a value from the Likert Rating Scale.)</td>
</tr>
<tr>
<td>Strongly Disagree</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

*Figure 2.3: Example of Interval-scaled data.*

In social research studies, such as market research, the Likert Rating Scale is often used for respondents to indicate a preference or a perception on a scale; interval-scaled properties are created for the study.

The data does not contain an absolute original, so the ratio of values cannot be meaningfully compared. A rating of 4 in the example in *Figure 2.3* would reflect a stronger perception than a 2 (this is a property of order); it is not,
however, possible to conclude that a rating of 4 is twice as important as a rating of 2. But it is possible to conclude that the difference in perception (or preference) between 3 and 4 is the same as between 1 and 2. This is a property of distance.

**Ratio-scaled Data** is mainly associated with quantitative random variables; it is numeric data with a zero origin. Examples are age, distance, time, mass, sales, units and income.

Such data is the strongest form of statistical data that can be gathered and lends itself to the widest range of statistical methods.

Ratio-scaled data is gathered through a measurement process, and can be manipulated meaningfully through normal arithmetic operations. If ratio-scaled data is grouped into categories, that data becomes ordinal-scaled; an example is the data for “Turnover” in *Figure 2.2.*

**Discrete Data**

A random variable whose observation can take on only specific values, usually only integer (whole numbers) values, is referred to as discrete.

Discrete variables are usually obtained by counting. There are a finite or countable number of choices available with discrete data. You can't have 2.63 people in the room.

**Examples are:**

- The number of students in a class
- The number of cars sold in a month by a dealer.

**Continuous Data**

A random variable whose observation can take on any value in an interval is said to generate continuous data, and that any value between a lower and upper limit is valid.

Continuous variables are usually obtained by measuring. Length, weight, and time are all examples of continuous variables. Since continuous variables are real numbers, we usually round them. This implies a boundary depending on the number of decimal places. For example: 64 is really anything $63.5 \leq x < 64.5$. 
Likewise, if there are two decimal places, then 64.03 is really anything 63.025 ≤ x < 63.035. Boundaries always have one more decimal place than the data and end in a 5.

**Examples are:**

- The time taken to travel to work daily.
- The tensile strength of steel.
- The speed of an aircraft.

**Data Collection Methods**

Wegner (1999) suggests three approaches to gathering data for statistical analyses:

- Observation
- Interview
- Experimentation.

**Observation Methods**

Use of observation as a measurement procedure, assigning numerals to human behavioral acts, is discussed. Observation has important advantages which makes it best suited for certain kinds of studies, and some limitations which preclude its use in others. The central problems in the use of observation are: (1) the effect of the observer on the observed, which is usually not severe and can be minimized; (2) observer inference, which is a crucial strength and a crucial weakness; and (3) the unit of behavior to be used, which involves the molar-
molecular problem. The considerations in planning both unstructured and structured observation studies are discussed, including what to observe, how to record it, how to maximize validity and reliability, and how to handle the relationship between the observer and the observed. Behavior is usually sampled using event sampling or time sampling.

Primary data can be collected by direct observation of the respondent or item in action. Examples are:

- Vehicle traffic surveys
- Observing the purchase behaviour of brands in a store.
- Quality control inspection.

An advantage of direct observation is that the respondent is generally not aware of being observed and therefore behaves in a natural way. This reduces the likelihood of gathering biased data.

A disadvantage is that it is a passive form of data collection, and there is little opportunity to probe for reasons or investigate behaviour further.

Secondary data can be obtained through desk research (abstraction), from a variety of source documents. A wide variety of organisations and individuals continually consult and use secondary data for decision making or opinion forming.

**Observation** is a technique that involves systematically selecting, watching and recording behaviour and characteristics of living beings, objects or phenomena.

**Observation of human behaviour** is a much-used data collection technique. It can be undertaken in different ways:

- **Participant observation**: The observer takes part in the situation he or she observes.
  
  (For example, a doctor hospitalised with a broken hip, who now observes hospital procedures ‘from within’.)

- **Non-participant observation**: The observer watches the situation, openly or concealed, but does not participate.
Observations can be open (e.g., ‘shadowing’ a health worker with his/her permission during routine activities) or concealed (e.g., ‘mystery clients’ trying to obtain antibiotics without medical prescription). They may serve different purposes. Observations can give additional, more accurate information on behaviour of people than interviews or questionnaires. They can also check on the information collected through interviews especially on sensitive topics such as alcohol or drug use, or stigmatising diseases. For example, whether community members share drinks or food with patients suffering from feared diseases (leprosy, TB, AIDS) are essential observations in a study on stigma.

**Observations of human behaviour** can form part of any type of study, but as they are time consuming they are most often used in small-scale studies.

**Observations** can also be made on objects. For example, the presence or absence of a latrine and its state of cleanliness may be observed. Here observation would be the major research technique.

If observations are made using a defined scale they may be called **measurements**. Measurements usually require additional tools. For example, in nutritional surveillance we measure weight and height by using weighing scales and a measuring board. We use thermometers for measuring body temperature.

**Interview Methods**

An **INTERVIEW** is a data-collection technique that involves oral questioning of respondents, either individually or as a group.

Interviews can be conducted through direct questioning or a questionnaire. Interview data can be gathered through personal (face-to-face) interviews, postal surveys and telephone surveys.

Answers to the questions posed during an interview can be recorded by writing them down (either during the interview itself or immediately after the interview) or by tape-recording the responses, or by a combination of both.

Interviews can be conducted with varying degrees of flexibility. The two extremes, high and low degree of flexibility, are described below:

- **High degree of flexibility:**
For example:

When studying sensitive issues such as teenage pregnancy and abortions, the investigator may use a list of topics rather than fixed questions. These may, e.g., include how teenagers started sexual intercourse, the responsibility girls and their partners take to prevent pregnancy (if at all), and the actions they take in the event of unwanted pregnancies. The investigator should have an additional list of topics ready when the respondent falls silent, (e.g., when asked about abortion methods used, who made the decision and who paid). The sequence of topics should be determined by the flow of discussion. It is often possible to come back to a topic discussed earlier in a later stage of the interview.

The unstructured or loosely structured method of asking questions can be used for interviewing individuals as well as groups of key informants.

A flexible method of interviewing is useful if a researcher has as yet little understanding of the problem or situation he is investigating, or if the topic is sensitive. It is frequently applied in exploratory studies. The instrument used may be called an interview guide or interview schedule.

• Low degree of flexibility:

Less flexible methods of interviewing are useful when the researcher is relatively knowledgeable about expected answers or when the number of respondents being interviewed is relatively large. Then questionnaires may be used with a fixed list of questions in a standard sequence, which have mainly fixed or pre-categorised answers.

For example:

After a number of observations on the (hygienic) behaviour of women drawing water at a well and some key informant interviews on the use and maintenance of the wells, one may conduct a larger survey on water use and satisfaction with the quantity and quality of the water.

Though in principle one may speak of loosely structured questionnaires, in practice the term questionnaire appears to be so hooked to tools with pre-categorised answers that we have decided to use the term interview guide for loosely structured tools. However, in reality there is often a mixture of open and pre-categorised answers. In that case we will still use the term questionnaire.

Administering written questionnaires

A WRITTEN QUESTIONNAIRE (also referred to as self-administered questionnaire) is a data collection tool in which written questions are presented that are to be answered by the respondents in written form.
A written questionnaire can be administered in different ways, such as by:

- Sending questionnaires by mail with clear instructions on how to answer the questions and asking for mailed responses.
- Gathering all or part of the respondents in one place at one time, giving oral or written instructions, and letting the respondents fill out the questionnaires; or
- Hand-delivering questionnaires to respondents and collecting them later.

Personal interviews have the advantage that accurate data can be obtained immediately, and qualitative data can be obtained by probing for reasons and observing non-verbal responses. They are, however, time consuming, and expensive if trained interviewers are required.

Telephone interviews allow more flexibility, in that call-backs can be made if a respondent is not available initially, and people are more willing to talk on the telephone from the security of their home or an office. It is more cost effective, as a larger sample of respondents can be reached in a relatively short time. The main disadvantage of telephone interviewing is that non-verbal responses cannot be observed.

Postal surveys (they can be conducted by fax or by e-mail as well) are best used when the target population is large and/or geographically dispersed.

A larger sample of respondents can be reached, making them more cost effective. Because respondents can answer anonymously, more honest, considered responses would be given. However, questions have to be shorter and simpler, and the possibility of probing is limited. Data collection can take a long time, and there is no control over who answers the questionnaire, or the possibility of check-backs on the validity of responses.

According to Wegner (1999) the response rates of postal surveys are very low (5% - 15%). The questionnaire is the data collection instrument used to gather data in all interview situations. The design of the questionnaire is critical to ensure that the correct research questions are addressed and that accurate and appropriate data are collected.

**Experimentation**

Primary data can also be generated through the manipulation of variables under controlled conditions. Data on the primary variable under study can be monitored and recorded while conscious efforts are made by the researcher to control the effects of a number of influencing factors.
ACTIVITY 1

(1) For each of the following variables, indicate:

(i) The data type

(ii) The measurement scale (i.e. nominal, ordinal, interval or ratio)

- The shelf life of milk
- The number of life policies issued per day
- The area of a shop floor
- The number of pages in a text book
- The flavours available in Dogmor food chunks
- The wood types that can be used to make a desk
- The size categories for shoes
- The voltage produced by a generator
- The car types in the Mercedes range
- The Yes/No/Sometimes response to “Do you drink Gin?”
- The number of loaves of bread sold daily by a bakery
- The income per day of a bakery
- The monthly birth-rate at a maternity hospital
- The mass of babies at birth
- The daily distance travelled by a courier service truck
- The names of teams in a cricket league.
ACTIVITY 2

Study the statistics printed in newspapers, magazines and on the Internet. The more you study them, the more information you will start obtaining from them. You will also start picking up trends.
SECTION 3: PRESENTATION OF DATA
CONTENTS

- Tables in Business
- Line Graphs
- Bar Chart
- Pie Chart
- Grouped Frequency distributions

LEARNING OUTCOMES

The learner should be able to:

- Construct suitable tables for a data set
- Construct a line or time series graph
- Construct a pie chart
- Construct a bar chart
- Construct a frequency distribution, histogram and frequency polygon
- Explain when and how to use the different types of diagrams
- Interpret the information when presented with these diagrams.
**Tables in Business**

The major difference between data and information is that the latter is presented in an ordered format. One such compact and efficient way of presenting data is in the form of a table.

Normally, the first stage of ordering is to present the data in a table. Once data is in the form of a table, we can:

- Make comparisons, within the table and with other data
- Perform additional calculations
- Examine the component structure of the data.

Using a table to list information according to category is often much clearer than writing out all the information in paragraph form. Let's look at an example of some data first written up in a paragraph. Try to think what information is being given and what sort of trends one could find from the information.

**Example:** During the 1995-1996 academic year, a survey of the holdings of university research libraries and rank was done in the United States and Canada. It was found that Syracuse University, in New York, had 2,692,147 holdings, and was figured to rank eighty-first. Harvard University ranked first with 13,369,855 holdings. The University of Connecticut was ranked fiftieth place, and reported 2,626,066 holdings. The Massachusetts Institute of Technology reported 2,448,647 holdings, and was ranked in seventy-third place. (Source: Association of Research Libraries).
As you can see, the paragraph above contains a lot of numbers and is not always easy to follow. The information given in the paragraph would be easier to decipher if it was presented in a table. To create a table, you need to determine the following things:

- Title of the table.
- Label of each row and/or column.
- Number of rows and columns necessary.
- Data entry for each cell.


<table>
<thead>
<tr>
<th>Institution</th>
<th>Rank</th>
<th>Holdings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harvard Univ.</td>
<td>1</td>
<td>13,369,855</td>
</tr>
<tr>
<td>U. Connecticut</td>
<td>50</td>
<td>2,626,066</td>
</tr>
<tr>
<td>Mass. Inst. Tech.</td>
<td>73</td>
<td>2,448,647</td>
</tr>
<tr>
<td>Syracuse Univ.</td>
<td>81</td>
<td>2,692,147</td>
</tr>
</tbody>
</table>

*Figure 3.1: Table depicting rank of University research libraries*

Now the statistics are talking to us!

Notice that the order of the universities in the table is different from the order they are listed in the paragraph. When moving from the paragraph to the table, it is best to order the instances by any numerical data. In this case, we ordered them by rank.

From the information in the table, we can see that the number one ranked university contains a lot more holdings than the other three institutions.

We carry out these processes to convert the data into information with which we are able to make good decisions.

**Exercise:** Let's take a look at how busy South African ports were during 1998. We are interested in the type of vessels that arrived at the ports, the number of vessels and the tonnage of each.
Imagine if the information were given to you in the following fashion (Statistics in this example are taken from “Port of Durban Statistics 1998” published by Portnet):

Richards Bay received 1,602 ocean going vessels, with a tonnage of 162,217,630. 96 coasters with a tonnage of 800,184 arrived at port. But she received no fishing vessels. 6 trawlers with a tonnage of 2,682 and 18 miscellaneous vessels with a tonnage of 1,035,454 arrived. Thus a total of 1,722 vessels with a tonnage of 164,055,950 was received by Richards Bay.

Durban received 4,127 ocean going vessels with a tonnage of 192,240,641 and 181 coasters with a tonnage of etc., etc. Portray this information in a table with appropriate headings.

A commonly used table in business is a **frequency table**, which shows the number of occurrences of a variable falling into a specific range or category.

**Example:** Consider a group of 47 males of various ages. 12 are between 20 and 29 years of age, 13 are between 30 and 39 years of age, 7 are between 40 and 49 years of age, 8 are between 50 and 59 years of age while the rest are between 60 and 69 years of age.

This data can be presented in a frequency table as follows:

<table>
<thead>
<tr>
<th>Age Interval (years) :</th>
<th>20-29:</th>
<th>30-39:</th>
<th>40-49:</th>
<th>50-59:</th>
<th>60-69:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of males</td>
<td>12</td>
<td>13</td>
<td>7</td>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

*Figure 3.2: Grouped frequency distribution showing male ages*

Immediately we can make the following conclusions:

1. Majority of the males are “young”, i.e. below 40 years of age.
2. Most lie between 30 and 39 years of age.

**NOTE:** The sum of the number of men in each interval (i.e. 12+13+7+8+7) must equal to the total number in the group which is 47 in this case.
Example: Consider a class consisting of 100 students. Suppose the teacher gives the entire class a statistics test which has a maximum mark of 100. Upon marking the scripts (which are in no order whatsoever), he puts the marks into a table as shown below. This represents raw data since there is no set order of the marks.

RAW DATA

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>60</td>
<td>58</td>
<td>63</td>
<td>72</td>
<td>52</td>
<td>63</td>
<td>82</td>
<td>65</td>
<td>58</td>
</tr>
<tr>
<td>75</td>
<td>59</td>
<td>26</td>
<td>52</td>
<td>71</td>
<td>55</td>
<td>55</td>
<td>55</td>
<td>52</td>
<td>62</td>
</tr>
<tr>
<td>90</td>
<td>58</td>
<td>25</td>
<td>51</td>
<td>68</td>
<td>47</td>
<td>52</td>
<td>56</td>
<td>55</td>
<td>35</td>
</tr>
<tr>
<td>80</td>
<td>41</td>
<td>47</td>
<td>49</td>
<td>52</td>
<td>48</td>
<td>44</td>
<td>64</td>
<td>56</td>
<td>18</td>
</tr>
<tr>
<td>35</td>
<td>42</td>
<td>38</td>
<td>48</td>
<td>53</td>
<td>45</td>
<td>48</td>
<td>62</td>
<td>57</td>
<td>52</td>
</tr>
<tr>
<td>12</td>
<td>44</td>
<td>85</td>
<td>47</td>
<td>45</td>
<td>41</td>
<td>75</td>
<td>51</td>
<td>51</td>
<td>48</td>
</tr>
<tr>
<td>65</td>
<td>46</td>
<td>76</td>
<td>46</td>
<td>46</td>
<td>40</td>
<td>25</td>
<td>52</td>
<td>48</td>
<td>56</td>
</tr>
<tr>
<td>50</td>
<td>54</td>
<td>74</td>
<td>36</td>
<td>32</td>
<td>50</td>
<td>66</td>
<td>53</td>
<td>46</td>
<td>48</td>
</tr>
<tr>
<td>55</td>
<td>51</td>
<td>72</td>
<td>24</td>
<td>8</td>
<td>51</td>
<td>50</td>
<td>48</td>
<td>42</td>
<td>47</td>
</tr>
<tr>
<td>45</td>
<td>35</td>
<td>65</td>
<td>56</td>
<td>44</td>
<td>60</td>
<td>55</td>
<td>49</td>
<td>40</td>
<td>50</td>
</tr>
</tbody>
</table>

*Figure 3.3: Student Marks*

On examining the data, we see the following:

Maximum value = 90

Minimum value = 8

Range (max – min) = 82
Since the highest possible mark in this case is 100, and the lowest is 0, an interval size (width) of 10 is easy to work with. Hence, we may choose to use the following intervals:

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>From</th>
<th>To</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&lt; 10</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>&lt; 20</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>&lt; 30</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>&lt; 40</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>&lt; 50</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>&lt; 60</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>&lt; 70</td>
<td></td>
</tr>
<tr>
<td>70</td>
<td>&lt; 80</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>&lt; 90</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>&lt; 100</td>
<td></td>
</tr>
</tbody>
</table>

Now that we have decided on the intervals, we can do a frequency count, which we do by registering each mark in the correct interval. For example, we would register the first value, 28, in the interval 20 < 30; we would tally the rest as shown in figure 3.4.

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>FROM</th>
<th>TO</th>
<th>TALLY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&lt; 10</td>
<td></td>
<td>i</td>
</tr>
<tr>
<td>10</td>
<td>&lt; 20</td>
<td></td>
<td>ii</td>
</tr>
<tr>
<td>20</td>
<td>&lt; 30</td>
<td></td>
<td>iii</td>
</tr>
<tr>
<td>30</td>
<td>&lt; 40</td>
<td></td>
<td>iiiii</td>
</tr>
<tr>
<td>40</td>
<td>&lt; 50</td>
<td></td>
<td>ii</td>
</tr>
<tr>
<td>50</td>
<td>&lt; 60</td>
<td></td>
<td>i</td>
</tr>
<tr>
<td>60</td>
<td>&lt; 70</td>
<td></td>
<td>i</td>
</tr>
<tr>
<td>70</td>
<td>&lt; 80</td>
<td></td>
<td>i</td>
</tr>
<tr>
<td>80</td>
<td>&lt; 90</td>
<td></td>
<td>iii</td>
</tr>
<tr>
<td>90</td>
<td>&lt; 100</td>
<td></td>
<td>i</td>
</tr>
</tbody>
</table>

**Figure 3.4: Tally of Student Marks**

**Note:** This function is easily done on a spreadsheet – if you are unsure how to do it, look up “frequency distribution” in the HELP facility.
When the tallying is complete, you can construct a frequency table as shown in *figure 3.5*.

The column *Cumulative Occurrences* refers to the total occurrences encountered up until a particular interval.

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>Occurrences</th>
<th>Cumulative Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>&lt; 10</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>&lt; 20</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>&lt; 30</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>&lt; 40</td>
<td>6</td>
</tr>
<tr>
<td>40</td>
<td>&lt; 50</td>
<td>29</td>
</tr>
<tr>
<td>50</td>
<td>&lt; 60</td>
<td>34</td>
</tr>
<tr>
<td>60</td>
<td>&lt; 70</td>
<td>12</td>
</tr>
<tr>
<td>70</td>
<td>&lt; 80</td>
<td>7</td>
</tr>
<tr>
<td>80</td>
<td>&lt; 90</td>
<td>3</td>
</tr>
<tr>
<td>90</td>
<td>&lt; 100</td>
<td>1</td>
</tr>
</tbody>
</table>

*Figure 3.5: Frequency table of Student Marks*

This frequency table could be modified to show ratios as percentages if this is preferred.

**ACTIVITY**

Study chapter 4, *Tables in Business*, of the prescribed text. Thereafter, answer the self-review questions and do the student exercises at the end of the chapter. Make sure you work through the computer exercises.
A table is a useful way to present detailed data, but a picture or diagram is more powerful if we want to focus attention on a particular aspect, such as a dominant feature or a trend. There are many different types of graphs we can use to portray data. The important graphs we will discuss are the line graph, bar graph and pie graph.

**Line Graphs**

One method of showing trends or comparative trends is to use a line graph. A good example is one that is used to show the growth in terms of profits of a particular organization.

**Example:**

The table in *figure 3.6* shows the total profit made by Kings Plastics for six consecutive years, from 1997 – 2002.

<table>
<thead>
<tr>
<th>Year</th>
<th>Profit (R 000's)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>78</td>
</tr>
<tr>
<td>1998</td>
<td>65</td>
</tr>
<tr>
<td>1999</td>
<td>53</td>
</tr>
<tr>
<td>2000</td>
<td>49</td>
</tr>
<tr>
<td>2001</td>
<td>38</td>
</tr>
<tr>
<td>2002</td>
<td>16</td>
</tr>
</tbody>
</table>

*Figure 3.6: Profit of Kings Plastic from 1997-2002.*

A study of the table shows that the profits have dropped considerably. This trend can be better depicted using a line graph (*Figure 3.7*) as shown on the following page:
Figure 3.7: Line graph depicting profits of Kings Plastics from 1997-2002.

The line graph clearly shows the drop in profits from 1997 to 2002.

Note the features of a line graph:

1. The vertical (y) and horizontal (x) axes are perpendicular to each other.
2. An appropriate scale is used such that the data points are reasonably spaced.
3. The data points are clearly marked (using small square points in this case).
4. The points are joined by lines (usually straight), which indicate a clear trend of the variables concerned.

Note: In this case, the yearly rate at which the profits decrease changes, hence the slope (or gradient) of the graph changes from year to year.
Bar Chart

The two most commonly used charts for business presentations are bar charts and pie charts – both of these very clearly and simply convey a large amount of information. We will look at the bar chart first.
A bar chart consists of a series of bars, the length of each bar representing the value of the variable being plotted. The bars can be either drawn vertically or horizontally.

Example:
If we take the previous example of the profits of Kings Plastics, we may plot the bar chart as follows:

*Figure 3.8: Bar Graph (vertical) depicting yearly profit of Kings Plastics*
The corresponding horizontal bar chart would look like:

![Bar Graph (horizontal) depicting yearly profit of Kings Plastics](image)

*Figure 3.9: Bar Graph (horizontal) depicting yearly profit of Kings Plastics*

Note the features of a bar chart:

1. The width of each bar must be the same (we use length to represent magnitude of the value, not width).
2. An appropriate scale must be used on each axis such that the lengths of the bars are reasonable.
3. The distance between each bar must be kept constant to give the bar chart uniformity.
4. The bars may or may not be coloured (this is purely up to the person drawing the bar chart).

**Pie Charts**

A pie chart is usually used when proportions are to be depicted relative to a whole. In essence, it is a circle divided into segments, with the size of each segment proportional to the value of the variable, relative to the whole, and is usually expressed in percentage terms. To illustrate the use of a pie chart, consider the following example.

**Example:** Consider a father who gives spending money to each of his three sons. Josh, the eldest gets R 120, Matt gets R 80 and David, the youngest gets R 50. This data can be expressed in a pie chart as follows:

Firstly, we calculate the percentage (in terms of the total amount the father gave out) that each son receives. The total in this case is R 120 + R 80 + R 50 = R 250. The percentages are:
The advantage of pie charts and bar charts is the visual impact that they have in conveying information. Pie charts are limited to a relatively small amount of data; with more data you will need to resort to a bar chart.

**Exercises:**
1. Consider the number of ocean going vessels that arrived in South Africa.
If we extracted the data for each port, we would arrive at the table shown in Figure 3.11.

<table>
<thead>
<tr>
<th>Port</th>
<th>Arrivals</th>
</tr>
</thead>
<tbody>
<tr>
<td>RICHARDS BAY</td>
<td>1,602</td>
</tr>
<tr>
<td>DURBAN</td>
<td>4,127</td>
</tr>
<tr>
<td>EAST LONDON</td>
<td>109</td>
</tr>
<tr>
<td>PORT ELIZABETH</td>
<td>802</td>
</tr>
<tr>
<td>MOSSEL BAY</td>
<td>14</td>
</tr>
<tr>
<td>CAPE TOWN</td>
<td>2,346</td>
</tr>
<tr>
<td>SALDANHA BAY</td>
<td>306</td>
</tr>
<tr>
<td>TOTAL</td>
<td>9,306</td>
</tr>
</tbody>
</table>

*Figure 3.11: Arrivals of Ocean-going Vessels*

Draw a pie chart depicting these values.

2. Express the data in the table by means of a bar chart.

3. Consider the following raw data:

\[
12 \quad 9 \quad 18 \quad 22 \quad 5 \quad 13 \quad 32 \quad 49 \quad 25 \quad 28
\]

Portray the data in the form of a frequency table. Use a suitable class width.

4. Draw a line graph depicting the following data:

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>18</td>
</tr>
</tbody>
</table>
Frequency Distributions

A histogram is a graphic display of a frequency distribution, using a bar graph. Earlier, we constructed a frequency table for the ages of 47 males. We can illustrate this information on a histogram as follows:

![Histogram of ages of 47 males](image)

*Figure 3.12: Histogram of ages of 47 males*

The advantage of representing information in a histogram is the visual impact it has – it is quicker and easier to see how ages are distributed.

**Exercise**: Use the grouped frequency distribution corresponding to the student marks and draw the corresponding histogram. What conclusions can you draw from the histogram?

**Cumulative Frequency Distribution (OGIVE)**

Cumulative frequencies are useful for determining the portion or percentage of observations that fall below or above a given value.
In our example of student marks, we could, for example, ask what percentage of students achieved more than 50%. A table showing the cumulative frequencies is shown in figure 3.13, while the information is shown graphically in figure 3.14.

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>FREQUENCY</th>
<th>CUMULATIVE FREQUENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>&lt; 10</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>&lt; 20</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>&lt; 30</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>&lt; 40</td>
<td>6</td>
</tr>
<tr>
<td>40</td>
<td>&lt; 50</td>
<td>29</td>
</tr>
<tr>
<td>50</td>
<td>&lt; 60</td>
<td>34</td>
</tr>
<tr>
<td>60</td>
<td>&lt; 70</td>
<td>12</td>
</tr>
<tr>
<td>70</td>
<td>&lt; 80</td>
<td>7</td>
</tr>
<tr>
<td>80</td>
<td>&lt; 90</td>
<td>3</td>
</tr>
<tr>
<td>90</td>
<td>&lt; 100</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 3.13: Cumulative Frequencies for Student Marks

Note: The width (interval size) of all intervals is 10. This is one of the main features of a grouped frequency distribution. Also, intervals must never overlap. Once the grouped frequency table has been constructed, an OGIVE curve can be drawn. An OGIVE is simply a line graph depicting the upper limit of each interval on the horizontal (x) axis (for example, the upper limit of the interval 40 < 50 is just 50) against the cumulative frequency on the vertical (y) axis.
We can see from the table, or more easily on the graph, that 43 students (or 43%, since we have exactly 100 observations) had marks of 50% or less – so 57% of the class managed to achieve more than 50%.

Evident from Figure 3.14 is the very steep line between marks of 40 and 60; 16% of students had a mark of 40 or less; 79% of students had a mark of 60 or less (63% of students marks thus fall in the interval of 40 to 60). Again, one may question whether this is a good distribution of marks.

Another term associated with this type of analysis is percentile, which, in our example, would refer to a mark that a percentage of students have not achieved. For example, one can read from figure 3.14:

- The 90th percentile is a mark of 70; thus 90% of students received a mark of 70% or less; or, 10% of students received a mark of more than 70.
- The 25th percentile is 45; 25% of students received 45 or less; or 75% of students received a mark of more than 45.

Also associated with this type of analysis is quartiles, which divide an ordered data-set into quarters.
The **lower quartile (or 25th percentile)** is that observation which separates the lower 25 percent of observation from the top 75 percent of ordered observation.

The **middle quartile (or 50th percentile)** is the median. It divides an ordered data set into two equal halves.

The **upper quartile (or 75th percentile)** is that observation which separates the top 25 percent of observations from the bottom 75 percent of ordered observations.

**ACTIVITY 2**

Study Chapter 6, “Business Diagrams: Histograms”, of the prescribed text.

Answer the self-review questions and carry out the exercises from page 196 to 199.
SECTION: 4 MANAGEMENT STATISTICS
CONTENTS

• Measures of Central Location
  Calculation of mean, median and mode.

• Measures of Dispersion
  Calculation of range, variance and standard deviation.

LEARNING OUTCOMES

The learner should be able to:

• Understand, calculate and interpret measures of central location.
• Understand the concept of “skewness” and interpret the reliability of a measure of central location.
• Understand and calculate measures of dispersion for a set of data.

READING

Prescribed Reading:

This manual has been designed to be read in conjunction with the following textbook:

We saw in the previous chapter that graphical displays of statistical data are useful as a means of communicating broad overviews of the behaviour of a random variable. However, there is a need for numerical measures (statistics) about the behaviour pattern of a random variable.

The behaviour pattern of any random variable can be described by:

- A measure of **central location**, and
- A measure of **spread** of observations about this central value.

After discussing these behaviour patterns, we will look at frequently used probability distribution functions in business, the binomial and normal distributions.

**Commonly used index numbers will be demonstrated, after which we will look at Sampling and Sampling Methods.**

**MEASURES OF CENTRAL LOCATION/TENDENCY**

**Measures of Central Location**

Observations of a random variable tend to group about some central value. The statistical measures that quantify where the majority of observations are concentrated are referred to as **measures of central location/tendency**.

Central tendency is a typical or representative score. If the mayor is asked to provide a single value which best describes the income level of the city, he or she would answer with a measure of central tendency.

A central location statistic represents a typical value or middle data point of a set of observations and is useful for comparing data sets.

**There are three main measures of central location:**

- Arithmetic mean (or average)
- Mode and
- Median (also called second quartile or the 50th percentile).
The computation for each measure differs slightly for ungrouped (or raw) data and grouped data (data summarised into a frequency distribution). The latter is of utmost importance.

**Arithmetic Mean**

The arithmetic mean for raw (ungrouped) data is calculated as follows:

\[ \bar{x} = \frac{\text{the sum of all the observations}}{\text{the number of observations}} \]

In equation form:

\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]

where:

- \( n \) = the number of observations in the sample
- \( x_i \) = the value of the \( i \)th observation of random variable \( x \)
- \( \bar{x} \) = symbol for a sample arithmetic mean

\[ \sum_{i=1}^{n} = \text{shorthand notation for the sum of n individual observations} \]

i.e. \[ \sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots \]
Example: Calculate the mean for the following data:

32 35 36 37 38 38 39 39 39 40 40 42 45

Solution:

\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]

\[ = \frac{500}{13} \]

\[ = 38.46 \]

If we take our previous example of student marks (the raw data is shown in Figure 3.3), we can calculate the mean using the equation shown above. We obtain a value of

\[ \bar{x} = 51.34 \]

Show the full calculation and verify the value above.
An easier method of calculating the mean is to use the grouped data shown in Figure 4.1 below, where the midpoint and frequency of observations for each interval is tabled.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Occurrences or Frequency (f)</th>
<th>Midpoint (x)</th>
<th>fx</th>
</tr>
</thead>
<tbody>
<tr>
<td>From 0</td>
<td>To 10</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>&lt; 20</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>&lt; 30</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>30</td>
<td>&lt; 40</td>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>40</td>
<td>&lt; 50</td>
<td>29</td>
<td>45</td>
</tr>
<tr>
<td>50</td>
<td>&lt; 60</td>
<td>34</td>
<td>55</td>
</tr>
<tr>
<td>60</td>
<td>&lt; 70</td>
<td>12</td>
<td>65</td>
</tr>
<tr>
<td>70</td>
<td>&lt; 80</td>
<td>7</td>
<td>75</td>
</tr>
<tr>
<td>80</td>
<td>&lt; 90</td>
<td>3</td>
<td>85</td>
</tr>
<tr>
<td>90</td>
<td>&lt; 100</td>
<td>1</td>
<td>95</td>
</tr>
</tbody>
</table>

\[ \sum = 5200 \]

**Figure 4.1: Grouped Data for Student Marks**

With this grouped data, we calculate the mean using the formula:

\[ \bar{x} = \frac{\sum fx}{n} \]

where:

\( \bar{x} \) is the midpoint of the interval and \( f \) is the frequency of the interval.

Using this formula, we can calculate the average as:

\[ \bar{x} = \frac{5200}{100} = 52 \]
If we compare this to the value calculated from the raw data (51.34), we see that this method can give a very close approximation.

**Mode**

The mode is the most frequently occurring value in a set of data. It is seldom computed for ungrouped data, since it is simply the value occurring the most number of times.

**Example**: Determine the mode for the following data:

32 35 36 37 38 38 39 39 39 40 40 42 45

**Solution**: The value appearing the most times (three times in this case) is 39. Hence the mode is 39.

We can also calculate the mode for grouped data, using the formula:

\[
M_o = O_{mo} + \frac{c[f_m - f_{m-1}]}{2f_m - f_{m-1} - f_{m+1}}
\]

Where:

- \(O_{mo}\) = the lower limit of the modal class.
- \(c\) = the class width.
- \(f_m\) = the frequency of the modal class.
- \(f_{m-1}\) = the frequency of the class before (above) the modal class.
- \(f_{m+1}\) = the frequency of the class after (below) the modal class.
In our example, the interval $50 < 60$ has the most observations (34), and thus qualifies as the "modal class interval". Therefore:

$$M_o = 50 + \frac{10(5)}{27} = 51.85$$

**Median**

The median is the value of a random variable that divides an ordered data-set into two equal parts, i.e. half the observations will fall below the median value and the other half above it.

**For ungrouped or raw data, there are one of two possibilities :**

1. If the number of observations is odd, then simply arrange the observations in ascending order and the median will be the middle value. i.e. if there are $n$ observations (where $n$ is odd), then the $[(n+1)/2]$ value is the median.

2. If the number of observations is even, then we simply arrange the observations in ascending order and take the average of the middle two values. i.e. if there are $n$ observations (where $n$ is even), then the median will be the average of the $[n/2]$ and $[(n/2)+1]$ value.

**Example:** Determine the median for the following data :

32 35 36 37 38 39 39 39 40 40 42 45

**Solution:** The data is already arranged in ascending order. In this case there are 13 values. Hence the median is the middle value which is 39.
For grouped data, the median can be calculated using the following equation:

\[ M_e = O_{me} + \frac{c[(n/2) - f(<)]}{f_{me}} \]

where

- \( O_{me} \) = lower limit of the median interval
- \( c \) = class width
- \( n \) = sample size (number of observations)
- \( f_{me} \) = absolute frequency of the median interval
- \( f(<) \) = cumulative absolute frequency of the interval before the median interval

The median interval is that class interval into which the \((n/2)\)th observation falls, using the less than ogive values.

For the student marks in figure 4.1, with \( n = 100 \), we can see that the 50th observation \((n/2 = 100/2 = 50)\) lies in the interval 50 < 60.

We can calculate the median value as follows:

\[ M_e = 50 + \frac{10(50 - 29)}{34} = 56.18 \]

Comparing the Mean, Median and Mode.
Skewed Distributions and Measures of Central Tendency

Skewness refers to the asymmetry of the distribution, such that a symmetrical distribution exhibits no skewness. In a symmetrical distribution the mean, median, and mode all fall at the same point, as in the following distribution ($M_d$ has been used to denote the median).

![Figure 4.3: A symmetrical frequency distribution](image)

An exception to this is the case of a bi-modal symmetrical distribution. In this case the mean and the median fall at the same point, while the two modes correspond to the two highest points of the distribution. An example follows:

![Figure 4.4: A bimodal frequency distribution](image)

A positively skewed distribution is asymmetrical and points in the positive direction. If a test was very difficult and almost everyone in the class did very poorly on it, the resulting distribution would most likely be positively skewed.
In the case of a positively skewed distribution, the mode is smaller than the median, which is smaller than the mean. This relationship exists because the mode is the point on the x-axis corresponding to the highest point, that is the score with greatest value, or frequency. The median is the point on the x-axis that cuts the distribution in half, such that 50% of the area falls on each side.

The mean is the balance point of the distribution. Because points further away from the balance point change the center of balance, the mean is pulled in the direction the distribution is skewed. For example, if the distribution is positively skewed, the mean would be pulled in the direction of the skewness, or be pulled toward larger numbers.
One way to remember the order of the mean, median, and mode in a skewed distribution is to remember that the mean is pulled in the direction of the extreme scores. In a positively skewed distribution, the extreme scores are larger, thus the mean is larger than the median.

A negatively skewed distribution is asymmetrical and points in the negative direction, such as would result with a very easy test. On an easy test, almost all students would perform well and only a few would do poorly.
The order of the measures of central tendency would be the opposite of the positively skewed distribution, with the mean being smaller than the median, which is smaller than the mode.

The choice of a representative central location value depends on the shape of the frequency distribution.

If a distribution is distorted by extreme values (i.e. skewed), then the median or the mode is more representative of the distribution than the mean.

For a skewed distribution, the median may be the best measure of central location as it is not pulled by extreme values (as the mean is), nor is it as highly influenced by the frequency of occurrence of a single value (as the mode is).
Other Measures of Central Location

Other measures you will come across, but not used as frequently as the mean, median and mode are:

**Geometric mean** – used for percentage changes or growth rates

**Harmonic mean** – used when a data set represents rates of change

**Weighted arithmetic mean** – used if the importance (weight) of each observation is different.

Measures of Dispersion / Variability

Variability refers to the spread or dispersion of scores. A distribution of scores is said to be highly variable if the scores differ widely from one another, so the classical definition of dispersion (or spread) is the extent by which the observations of random variable are scattered about the central value.

Measures of dispersion provide useful information with which the reliability of the central value may be judged. Widely dispersed observations indicate that the central value has low reliability, and does not represent the observations very well. Conversely, a high concentration of observations about the central value indicates higher reliability, with the central value being more representative.

The measures that are used to describe dispersion are:

- Range
- Inter-quartile range
- Quartile deviation
- Variance
- Standard deviation

**Range**

The range is the difference between the highest and lowest observed values in a data set.

It is simply the largest score minus the smallest score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores.

\[
\text{Range} = \text{Maximum value} - \text{Minimum value}, \quad \text{for ungrouped data}
\]

\[
= \text{Upper limit (highest class)} - \text{Lower limit (lowest class)}, \quad \text{for grouped data}
\]
Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

**Distribution 1**
32 35 36 36 37 38 40 42 42 43 43 45

**Distribution 2**
32 32 33 33 33 34 34 34 34 34 35 45

For this reason, among others, the range is not the most important measure of variability.

Referring again to the example of student exam marks (see figure 3.3), we can use the ungrouped data to calculate the range:

- Maximum value = 90
- Minimum value = 8
- Range = 90 – 8 = 82

If we use the grouped data from *Figure 4.2*:

- Upper limit (highest class) = 100
- Lower limit (lowest class) = 0
- Range = 100 – 0 = 100

Obviously the range calculated from the grouped data is not as accurate a measure as that calculated with the raw data; in this case, it is also a poor estimate. For larger sets of data, it is normally a much closer estimate.

The range is a crude estimate of spread. It is easily calculated, but is distorted by extreme values (“outliers”). An “outlier” would be the minimum or maximum value. It is thus a volatile and unstable measure of dispersion as it can vary greatly between samples taken from the same population. It also provides no information on the clustering of observations within the data set about a central value as it uses only two observations (i.e. the maximum and minimum) in its computation.
Inter-Quartile Range

Because the range can be distorted by extreme values (“out-liers”), a modified range that excludes these is often calculated. The inter-quartile range considers the viability shown by only the middle 50 percent of observations, and is the difference between the upper and lower quartiles.

\[
\text{Inter-quartile range} = Q_3 - Q_1
\]

OR

\[
= 75^{th} \text{ percentile} - 25^{th} \text{ percentile}
\]

Figure 3.14 showed the cumulative frequency polygon (or OGIVE) for our student marks as an example.

From the figure, we can read off the 75\text{th} and 25\text{th} percentile.

\[
75^{th} \text{ percentile} = 58 \\
25^{th} \text{ percentile} = 45 \\
\text{Inter-quartile range} = 58 - 45 = 13
\]

This measure of dispersion removes much of the instability inherent in the range by excluding “out-liers”, but it excludes 50 percent of all observations from further analysis. It also provides no information on the clustering of observations within the data set as it uses only two observations (Q₁ & Q₃) in its calculation.

Quartile Deviation

This measure of variation is simply the inter-quartile range divided by 2:

Quartile deviation (Q.D.) = \( \frac{Q_3 - Q_1}{2} \)

Continuing with our student marks example,

Q.D. = \( \frac{13}{2} \) = 6.5
Remembering the median of 51.84, we interpret this as follows:

50 percent of all observations are expected to lie within 6.5 marks either side of 51.84, i.e. between 45.34 and 58.34.

Alternatively, 25 percent of the marks are expected to be within 6.5 marks below the median (45.34 to 51.84), and 25 percent of the marks are expected to lie within 6.5 marks above the median, (51.84 to 58.34).

The quartile deviation is useful as a measure of dispersion if the sample of observations contains excessive “outliers”, as it ignores the top 25 percent and bottom 25 percent of the ranked observations.

As with the inter-quartile range, the quartile deviation does not use all the observations and therefore gives no indication of the spread of values between the upper and lower quartiles.

### Variance

The variance has become the most used measure of dispersion, because it:

- Takes every observation into account, and
- Is based on an average deviation from a central value.

The calculation depends on whether the data is ungrouped or grouped.

\[
\text{Variance} = \frac{\text{Sum of squared deviations}}{(\text{Sample size} - 1)}
\]

\[
S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}
\]

**Calculation for ungrouped data:**

Note that the variance could almost be the average squared deviation around the mean if the expression were divided by \( n \) rather than \( n-1 \). It is divided by \( n-1 \), called the degrees of freedom (df), for theoretical reasons. If the
mean is known, as it must be to compute the numerator of the expression, then only n-1 scores that are free to vary. That is if the mean and n-1 scores are known, then it is possible to figure out the nth score.

The formula for the variance presented above is a definitional formula, it defines what the variance means. The variance may be computed from this formula, but in practice this is rarely done. It is done here to better describe what the formula means. The computation is performed in a number of steps, which are presented below:

**Step One** - Find the mean of the scores.

**Step Two** - Subtract the mean from every score.

**Step three** - Square the results of step two.

**Step Four** - Sum the results of step three.

**Step Five** - Divide the results of step four by n-1.

**Step Six** - Take the square root of step five.

**Example:**

Consider the following simple example showing the ages of 7 second-hand cars:

| 13 | 7 | 10 | 15 | 12 | 18 | 9 |
The calculation of the squared deviation of each observation from the sample mean is shown in Figure 4.7:

<table>
<thead>
<tr>
<th>Car Ages</th>
<th>Mean</th>
<th>Deviation</th>
<th>Squared deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>12</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>0</td>
<td>84</td>
</tr>
</tbody>
</table>

*Figure 4.7: Calculation of Variance*

**Note:** The sum of deviations (3rd column) must be zero.

Applying the formula on page 56, we can calculate variance = 84/6 = 14 years.

An alternative formula that is easier to use is:

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n - 1}$$
Continuing with the car ages example, we can calculate as follows:

<table>
<thead>
<tr>
<th>Car age $x_i$</th>
<th>$x_i^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>169</td>
</tr>
<tr>
<td>7</td>
<td>49</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>15</td>
<td>225</td>
</tr>
<tr>
<td>12</td>
<td>144</td>
</tr>
<tr>
<td>18</td>
<td>324</td>
</tr>
<tr>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>Tot = 84</td>
<td>Tot = 1092</td>
</tr>
</tbody>
</table>

*Figure 4.8: Car ages – variance calculation*

Using the alternate formula:

$$Variance = \frac{(1092 - 7(12)^2)}{6} = 14$$

In our example of student marks, the variance is 210.1 (check this figure using either formula). The variance can also be calculated for grouped data.

**Calculation for Grouped Data**

For grouped data, the variance can be calculated as:

$$S_x^2 = \frac{\sum_{i=1}^{n} f_i x_i^2 - n \bar{x}^2}{n - 1}$$

- $f_i = \text{frequency in interval}$
- $x_i = \text{interval midpoint}$
If we use the example of student marks, *Figure 4.9* shows the calculations of the variance.

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>Frequency</th>
<th>Midpoint</th>
<th>( f_i \cdot x_i )</th>
<th>( x_i^{**2} )</th>
<th>( f_i \cdot x_i^{**2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>From</td>
<td>To</td>
<td></td>
<td>( f_i )</td>
<td>( x_i )</td>
<td>( f_i \cdot x_i )</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>5.5</td>
<td>5.5</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>2</td>
<td>2</td>
<td>15.5</td>
<td>31.0</td>
</tr>
<tr>
<td>21</td>
<td>30</td>
<td>5</td>
<td>5</td>
<td>25.5</td>
<td>127.5</td>
</tr>
<tr>
<td>31</td>
<td>40</td>
<td>8</td>
<td>8</td>
<td>35.5</td>
<td>284.0</td>
</tr>
<tr>
<td>41</td>
<td>50</td>
<td>31</td>
<td>31</td>
<td>45.5</td>
<td>1410.5</td>
</tr>
<tr>
<td>51</td>
<td>60</td>
<td>32</td>
<td>32</td>
<td>55.5</td>
<td>1776.0</td>
</tr>
<tr>
<td>61</td>
<td>70</td>
<td>10</td>
<td>10</td>
<td>65.5</td>
<td>655.0</td>
</tr>
<tr>
<td>71</td>
<td>80</td>
<td>8</td>
<td>8</td>
<td>75.5</td>
<td>604.0</td>
</tr>
<tr>
<td>81</td>
<td>90</td>
<td>3</td>
<td>3</td>
<td>85.5</td>
<td>256.5</td>
</tr>
<tr>
<td>91</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>95.5</td>
<td>0.0</td>
</tr>
<tr>
<td>SUM</td>
<td>100</td>
<td></td>
<td>5150</td>
<td></td>
<td>287025.2</td>
</tr>
</tbody>
</table>

*Figure 4.9: Variance Calculations for Grouped Data*

Using the formula above, we can calculate the variance

\[
\sigma^2 = \frac{287025.2 - 100(51.5)^2}{99} = 220.2
\]

This is close to the value of 210.10 calculated for the raw data earlier.

The variance is a measure of average squared deviation about the arithmetic mean. It is expressed in squared units. Consequently, its meaning in a practical sense is obscure. To provide meaning, the dispersion measure should be expressed in the original unit of measure of the random variable.
Standard Deviation

A standard deviation is a statistical measure, which expresses the average deviation about the mean in the original units of the random variable. It is the square root of the variance and is written as

\[ S_x = \sqrt{S_x^2} \]

The standard deviation measures variability in units of measurement, while the variance does so in units of measurement squared. For example, if one measured height in inches, then the standard deviation would be in inches, while the variance would be in inches squared. For this reason, the standard deviation is usually the preferred measure when describing the variability of distributions.

In our examples of exam marks, we can easily calculate the standard deviation:

Ungrouped data

\[ S_x = \sqrt{210.1} = 14.49 \]

Grouped data

\[ S_x = \sqrt{220.2} = 14.84 \]

The standard deviation is a relatively stable measure of dispersion across different samples of the same random variable. It is therefore a powerful statistic, which describes how the observations are spread across the mean.

Coefficient of Variation

It is sometimes necessary to compare samples of data from different random variables to establish which sample data shows greater variability. A direct comparison of their respective standard deviations would be misleading as the random variables may be measured in different units.

The comparison would be more meaningful if the measures of variability were expressed in the same units. This can be achieved by producing a measure of relative variability, i.e. relative to their mean, expressed in percentage terms.
A statistic that shows this relative dispersion about a mean for a random variable is called the coefficient of variation, and is defined as:

\[ CV = \frac{S_x}{\bar{X}} \times 100\% \]

A coefficient close to zero indicates low variability and a tight clustering of observations about the mean. Conversely, a large coefficient of variation value indicates that observations are more spread about their mean value.

We can calculate the coefficient of variation for the student marks, using the ungrouped data:

\[ CV = \frac{16.49}{51.34} \times 100\% = 28.2\% \]

This low value of the coefficient indicates that in spite of the large range of the data, the marks are generally tightly clustered around the mean.

**SUMMARY**

Statistics serve to estimate model parameters and describe the data. Two categories of statistics were described in this chapter: measures of central tendency and measures of variability. In the former category were the mean, median, and mode. In the latter were the range, standard deviation, and variance. Measures of central tendency describe a typical or representative score, while measures of variability describe the spread or dispersion of scores.
Fauly ATMs are a major problem plaguing STANDARD BANK SOUTH AFRICA. For 20 consecutive days, the number of faulty ATMs reported around the country were recorded and presented in the table below.

<table>
<thead>
<tr>
<th>DAY</th>
<th>No. of faulty ATMs</th>
<th>DAY</th>
<th>No. of faulty ATMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38</td>
<td>11</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>12</td>
<td>45</td>
</tr>
<tr>
<td>3</td>
<td>33</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>41</td>
<td>14</td>
<td>51</td>
</tr>
<tr>
<td>5</td>
<td>56</td>
<td>15</td>
<td>28</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>19</td>
<td>17</td>
<td>46</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>18</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>32</td>
<td>19</td>
<td>44</td>
</tr>
<tr>
<td>10</td>
<td>52</td>
<td>20</td>
<td>22</td>
</tr>
</tbody>
</table>

1. Determine the range from the data in the table.
2. Group the data in a frequency distribution with a lowest class lower limit of 10 faulty ATMs and a class width of 10 faulty ATMs.
3. From the grouped frequency distribution, determine the following (number of faulty ATMs over the 20 day period):
   (i) mean.
   (ii) median.
   (iii) mode.
4. Determine the standard deviation and interpret its value.
5. Draw an ogive curve and use it to estimate the median.
6. What type of data (discrete or continuous) is portrayed in the table? Explain.
Solution:

1. Range = \( x_{\text{max}} - x_{\text{min}} = 56 - 17 = 39 \) faulty ATMs

2. |
<table>
<thead>
<tr>
<th>Interval</th>
<th>Frequency, ( f )</th>
<th>Midpoint, ( x )</th>
<th>( fx )</th>
<th>( fx^2 )</th>
<th>Cum. Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 – 19</td>
<td>2</td>
<td>14.5</td>
<td>29</td>
<td>420.5</td>
<td>2</td>
</tr>
<tr>
<td>20 – 29</td>
<td>5</td>
<td>24.5</td>
<td>122.5</td>
<td>3001.25</td>
<td>7</td>
</tr>
<tr>
<td>30 – 39</td>
<td>6</td>
<td>34.5</td>
<td>207</td>
<td>7141.5</td>
<td>13</td>
</tr>
<tr>
<td>40 – 49</td>
<td>4</td>
<td>44.5</td>
<td>178</td>
<td>7921</td>
<td>17</td>
</tr>
<tr>
<td>50 – 59</td>
<td>3</td>
<td>54.5</td>
<td>163.5</td>
<td>8910.75</td>
<td>20</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>20</td>
<td></td>
<td>700</td>
<td>27395</td>
<td></td>
</tr>
</tbody>
</table>

3. (i) **Mean** = \( \frac{\sum fx}{n} \)

\[
= \frac{700}{20} \\
= 35 \text{ faulty ATMs}
\]

(ii) **Median** = \( O_{\text{me}} + c[\frac{n}{2} - f(<)] / f_{\text{me}} \)

\[
= 30 + 9[10 - 7] / 6 \\
= 34.5 \text{ faulty ATMs}
\]

(iii) **Mode** = \( O_{\text{mo}} + c[f_m - f_{m-1}] / (2f_m - f_{m-1} - f_{m+1}) \)

\[
= 30 + 9[6 - 5] / (2 \times 6 - 5 - 4) \\
= 33 \text{ faulty ATMs}
\]
4. Variance = \[ \Sigma fx^2 - n \text{mean}^2 \]
\[ \frac{\Sigma fx^2 - n \text{mean}^2}{n - 1} \]
\[ = \frac{27395 - 20 (35)^2}{20 - 1} \]
\[ = 152.37 \]

Standard deviation = Variance^{1/2} = 152.37^{1/2} = 12.34 faulty ATMs.

This value is fairly small, showing consistency of the ATMs.

5.

6. Discrete data, since number of faulty ATMs has to be a whole number.
SECTION 5: PROBABILITY DISTRIBUTION FUNCTIONS
CONTENTS

• Binomial and Normal Probability Distributions

• Index Numbers

• Sampling and Sampling Distributions

LEARNING OUTCOMES

• Calculate probabilities and distinguish between binomial and normal distributions

• Calculate and use index numbers

• Understand appropriate sampling techniques to obtain statistical data
Binomial and Normal Probability Distributions

A probability distribution is a list of all the possible outcomes of a random variable and their associated probabilities of occurrence. There are numerous problem situations in practice where the outcomes of a specific random variable follow known probability patterns. If the behaviour of a random variable can be matched to a known probability pattern, then probabilities for the random variable can be found directly by applying an appropriate theoretical probability distribution function.

Probability distribution functions can be classified as:

- **Discrete**, which assumes that the outcomes of a random variable can take on only specific (usually integer) values. An example is the Binomial Probability Distribution.
- **Continuous**, where the variable can take on any value (as opposed to only discrete values) in an interval. They are used to find probabilities associated with intervals of x values. An example is the Normal Distribution.

Binomial Probability Distribution

A discrete random variable can be described by the Binomial distribution if it satisfies the following four conditions:

(i) There are only two mutually exclusive and collectively exhaustive outcomes of the random variable.

   Generally, these two outcomes are referred to as **success** or **failure**. Each outcome has an associated probability:

   - The probability of the **success** outcome is denoted by p
   - The probability of the **failure** is denoted by q
   - \( p + q = 1 \); hence \( q = (1 - p) \)

(ii) The random variable is observed n times. Each observation of the random variable in its problem setting is called a trial. Each trial generates either a success or failure outcome. Thus n outcomes are observed.
(iii) The trials are assumed to be independent of one another. Thus the outcome on any trial is in no way influenced by the outcome on any other trial. This means that p and q remain constant for each trial of the process under study.

(iv) The binomial question is “What is the probability that r successes will occur in n trials of the process under study?”

These can be summarized as: An experiment with a fixed number of independent trials, each of which can only have two possible outcomes.

The fact that each trial is independent actually means that the probabilities remain constant.

**Examples of binomial experiments**

- Tossing a coin 20 times to see how many tails occur.
- Asking 200 people if they watch ABC news.
- Rolling a die to see if a 5 appears.

**Examples which aren’t binomial experiments**

- Rolling a die until a 6 appears (not a fixed number of trials)
- Asking 20 people how old they are (not two outcomes)
- Drawing 5 cards from a deck for a poker hand (done without replacement, so not independent)

The binomial formula calculates the probability of r successes, and is stated as follows:

\[
P(r) = \frac{n!}{r!(n-r)!} p^r q^{(n-r)}
\]

for \( r = 0,1,2,3,\ldots \)
Where \( n = \) the number of trials (observations)

\[
\begin{align*}
n & = \text{the number of trials (observations)} \\
r & = \text{the number of success outcomes in n trials} \\
p & = \text{probability of a success outcome} \\
q & = \text{probability of a failure outcome} \\
! & = \text{the mathematical symbol for factorial, where } n! = n \times (n - 1) \times (n - 2) \times \ldots \times 1
\end{align*}
\]

**Example:**

**What is the probability of rolling exactly two sixes in 6 rolls of a die?**

There are five things you need to do to work a binomial story problem.

1. Define Success first. Success must be for a single trial. Success = "Rolling a 6 on a single die"
2. Define the probability of success \( (p): p = 1/6 \)
3. Find the probability of failure: \( q = 5/6 \)
4. Define the number of trials: \( n = 6 \)
5. Define the number of successes out of those trials: \( r = 2 \)

Anytime a six appears, it is a success (denoted S) and anytime something else appears, it is a failure (denoted F). The ways you can get exactly 2 successes in 6 trials are given below. The probability of each is written to the right of the way it could occur. Because the trials are independent, the probability of the event (all six dice) is the product of each probability of each outcome (die)
Notice that each of the 15 probabilities are exactly the same: \((1/6)^2 \times (5/6)^4 = 0.0134\)

Also, note that the 1/6 is the probability of success and you needed 2 successes. The 5/6 is the probability of failure, and if 2 of the 6 trials were success, then 4 of the 6 must be failures.

Note that 2 is the value of \(r\) and 4 is the value of \(n-r\).

Further note that there are **fifteen** ways this can occur. This is the number of ways 2 successes can be occur in 6 trials without repetition and order not being important, or a combination of 6 things, 2 at a time. Hence, the probability is \(15 \times 0.0134 = 0.201\) (20.1 %).

**Example: The car hire problem**

A car hire firm rents out only BMW and Mazda cars. Experience has shown that one in four clients choose a BMW. If 5 reservations are randomly selected from today's bookings, what is the probability that 2 will have requested a BMW?
One in four clients hires a BMW, so:

\[ p = \frac{1}{4} = 0.25 \]  
(A BMW is hired)

\[ q = \frac{3}{4} = 0.75 \]  
(A BMW is not hired)

We want to know the probability of 2 success outcomes, i.e. we require \( P(2) \) and we have 5 observations, thus \( n = 5 \).

**Using the binomial formula:**

\[
P(2) = \frac{5!}{2!(5-2)!} (0.25)^2 (0.75)^{(5-2)}
\]

\[= 0.2637\]

With larger values of \( n \) and \( r \), these calculations can become more elaborate.

Fortunately, most spreadsheets have the formula built in; for the above example, we could use the following formula in Microsoft Excel:

\[
\text{BINOMDIST}(2,5,0.25, \text{FALSE}) = 0.2637
\]

**The Normal Probability Distribution**

A normal probability distribution function finds the probabilities for a continuous random variable and has the following characteristics:

- It is **bell-shaped**
- It is **symmetrical** about a central value
- The tails of the distribution never touch the x-axis, i.e. **asymptotic**
- A normally distributed random variable is described by two parameters – the **mean** (\( \mu \)) and **standard deviation** (\( \sigma \)).
- The area under the curve equals one.
The probability associated with a particular range of \( x \) values is described by the area under the curve between the limits of the \( x \) range; for example:

\[ x_1 < x < x_2 \]

**Figure 5.1** below illustrates the features of the normal curve.

**Symmetric:** Each side is the mirror image of the other

**Asymptotic:** Tail approaches the \( x \) axis but never reaches it

Area = 0.5  Area = 0.5

Mean, median, and mode

---

**Figure 5.1: A typical normal distribution**

There are three areas on a standard normal curve that all statistics students should know. The first is that the total area below 0.0 is 0.50, as the standard normal curve is symmetrical like all normal curves. This result generalizes to all normal curves in that the total area below the value of \( \mu \) is 0.50 on any member of the family of normal curves.
The second area that should be memorized is between Z-scores of -1.00 and +1.00. It is 0.68 or 68%.

The total area between plus and minus one σ unit on any member of the family of normal curves is also 0.68.

The third area is between Z-scores of -2.00 and +2.00 and is 0.95 or 95%.

This area (0.95) also generalizes to plus and minus two σ units on any normal curve.

Knowing these areas allow computation of additional areas. For example, the area between a Z-score of 0.0 and 1.0 may be found by taking 1/2 the area between Z-scores of -1.0 and 1.0, because the distribution is symmetrical between those two points. The answer in this case is 0.34 or 34%. A similar logic and answer is found for the area between 0.0 and -1.0 because the standard normal distribution is symmetrical around the value of 0.0.

The area below a Z-score of 1.0 may be computed by adding 0.34 and 0.50 to get .84. The area above a Z-score of 1.0 may now be computed by subtracting the area just obtained from the total area under the distribution (1.00), giving a result of 1.00 - 0.84 or 0.16 or 16%. 
The area between -2.0 and -1.0 requires additional computation. First, the area between 0.0 and -2.0 is 1/2 of 0.95 or 0.475. Because the 0.475 includes too much area, the area between 0.0 and -1.0 (0.34) must be subtracted in order to obtain the desired result. The correct answer is 0.475 - 0.34 or 0.135.

![Diagram showing normal distribution with areas shaded]

Using a similar kind of logic to find the area between Z-scores of .5 and 1.0 will result in an incorrect answer because the curve is not symmetrical around 0.5. The correct answer must be something less than 0.17, because the desired area is on the smaller side of the total divided area.

If a data set follows a normal distribution, we can predict the frequency of data in intervals defined by the mean and standard deviation, as follows:

\[
\mu - \sigma < x < \mu + \sigma : \, 68.26\%
\]

\[
\mu - 2\sigma < x < \mu + 2\sigma : \, 95.50\%
\]

\[
\mu - 3\sigma < x < \mu + 3\sigma : \, 99.73\%
\]

\[
\mu = \text{population mean} \quad \sigma = \text{standard deviation}
\]

If we look at our student marks, we can calculate the intervals and compare the count with the normal distribution calculation.

Remember the mean was 51.34 and the standard deviation was 14.49 (calculated using the ungrouped data).
The table in *figure 5.2* shows how the values predicted by the normal distribution compares with the actuals.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Actual Frequency</th>
<th>%</th>
<th>% by Normal Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower limit</td>
<td>Upper limit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36.85</td>
<td>65.83</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>22.36</td>
<td>80.32</td>
<td>94</td>
<td>94</td>
</tr>
<tr>
<td>7.87</td>
<td>94.81</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

*Figure 5.2: Actual vs. Predicted Frequency for Student Marks*

Except for the one standard deviation range, the predictions are quite accurate; certainly accurate enough for practical application.

So the normal distribution can, within limitations, be used for predicting frequencies.

**Standard Normal Distribution**

It is more common to use a Standard Normal Distribution, as probabilities (equal to the area under the curve) are worked out for it, and presented in tables. The random variable is defined in terms of a parameter \( z \), and hence it is often called the *z-distribution*.

The Standard Normal Probability Distribution function, with random variable \( z \), has the following characteristics:

- A mean equal to zero.
- A standard deviation equal to one.

The areas under the standard normal distribution can be read off from standard normal tables. The table gives the area between 0 and \( z \). It is given on the following page (*Figure 5.3*):
<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.0000</td>
<td>0.0040</td>
<td>0.0080</td>
<td>0.0120</td>
<td>0.0160</td>
<td>0.0199</td>
<td>0.0239</td>
<td>0.0279</td>
<td>0.0319</td>
<td>0.0359</td>
</tr>
<tr>
<td>0.1</td>
<td>0.0398</td>
<td>0.0438</td>
<td>0.0478</td>
<td>0.0517</td>
<td>0.0557</td>
<td>0.0596</td>
<td>0.0636</td>
<td>0.0675</td>
<td>0.0714</td>
<td>0.0753</td>
</tr>
<tr>
<td>0.2</td>
<td>0.0793</td>
<td>0.0832</td>
<td>0.0871</td>
<td>0.0910</td>
<td>0.0948</td>
<td>0.0987</td>
<td>0.1026</td>
<td>0.1064</td>
<td>0.1103</td>
<td>0.1141</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1179</td>
<td>0.1217</td>
<td>0.1255</td>
<td>0.1293</td>
<td>0.1331</td>
<td>0.1368</td>
<td>0.1406</td>
<td>0.1443</td>
<td>0.1480</td>
<td>0.1517</td>
</tr>
<tr>
<td>0.4</td>
<td>0.1554</td>
<td>0.1591</td>
<td>0.1628</td>
<td>0.1664</td>
<td>0.1700</td>
<td>0.1736</td>
<td>0.1772</td>
<td>0.1808</td>
<td>0.1844</td>
<td>0.1879</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1915</td>
<td>0.1950</td>
<td>0.1985</td>
<td>0.2019</td>
<td>0.2054</td>
<td>0.2088</td>
<td>0.2123</td>
<td>0.2157</td>
<td>0.2190</td>
<td>0.2224</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2257</td>
<td>0.2291</td>
<td>0.2324</td>
<td>0.2357</td>
<td>0.2389</td>
<td>0.2422</td>
<td>0.2454</td>
<td>0.2486</td>
<td>0.2517</td>
<td>0.2549</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2580</td>
<td>0.2611</td>
<td>0.2642</td>
<td>0.2673</td>
<td>0.2704</td>
<td>0.2734</td>
<td>0.2764</td>
<td>0.2794</td>
<td>0.2823</td>
<td>0.2852</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2881</td>
<td>0.2910</td>
<td>0.2939</td>
<td>0.2967</td>
<td>0.2995</td>
<td>0.3023</td>
<td>0.3051</td>
<td>0.3078</td>
<td>0.3106</td>
<td>0.3133</td>
</tr>
<tr>
<td>0.9</td>
<td>0.3159</td>
<td>0.3186</td>
<td>0.3212</td>
<td>0.3238</td>
<td>0.3264</td>
<td>0.3289</td>
<td>0.3315</td>
<td>0.3340</td>
<td>0.3365</td>
<td>0.3389</td>
</tr>
<tr>
<td>1.0</td>
<td>0.3413</td>
<td>0.3438</td>
<td>0.3461</td>
<td>0.3485</td>
<td>0.3508</td>
<td>0.3531</td>
<td>0.3554</td>
<td>0.3577</td>
<td>0.3599</td>
<td>0.3621</td>
</tr>
<tr>
<td>1.1</td>
<td>0.3643</td>
<td>0.3665</td>
<td>0.3686</td>
<td>0.3708</td>
<td>0.3729</td>
<td>0.3749</td>
<td>0.3770</td>
<td>0.3790</td>
<td>0.3810</td>
<td>0.3830</td>
</tr>
<tr>
<td>1.2</td>
<td>0.3849</td>
<td>0.3869</td>
<td>0.3888</td>
<td>0.3907</td>
<td>0.3925</td>
<td>0.3944</td>
<td>0.3962</td>
<td>0.3980</td>
<td>0.3997</td>
<td>0.4015</td>
</tr>
<tr>
<td>1.3</td>
<td>0.4032</td>
<td>0.4049</td>
<td>0.4066</td>
<td>0.4082</td>
<td>0.4099</td>
<td>0.4115</td>
<td>0.4131</td>
<td>0.4147</td>
<td>0.4162</td>
<td>0.4177</td>
</tr>
<tr>
<td>1.4</td>
<td>0.4192</td>
<td>0.4207</td>
<td>0.4222</td>
<td>0.4236</td>
<td>0.4251</td>
<td>0.4265</td>
<td>0.4279</td>
<td>0.4292</td>
<td>0.4306</td>
<td>0.4319</td>
</tr>
<tr>
<td>1.5</td>
<td>0.4332</td>
<td>0.4345</td>
<td>0.4357</td>
<td>0.4370</td>
<td>0.4382</td>
<td>0.4394</td>
<td>0.4406</td>
<td>0.4418</td>
<td>0.4429</td>
<td>0.4441</td>
</tr>
<tr>
<td>1.6</td>
<td>0.4452</td>
<td>0.4463</td>
<td>0.4474</td>
<td>0.4484</td>
<td>0.4495</td>
<td>0.4505</td>
<td>0.4515</td>
<td>0.4525</td>
<td>0.4535</td>
<td>0.4545</td>
</tr>
<tr>
<td>1.7</td>
<td>0.4554</td>
<td>0.4564</td>
<td>0.4573</td>
<td>0.4582</td>
<td>0.4591</td>
<td>0.4599</td>
<td>0.4608</td>
<td>0.4616</td>
<td>0.4625</td>
<td>0.4633</td>
</tr>
<tr>
<td>1.8</td>
<td>0.4641</td>
<td>0.4649</td>
<td>0.4656</td>
<td>0.4664</td>
<td>0.4671</td>
<td>0.4678</td>
<td>0.4686</td>
<td>0.4693</td>
<td>0.4699</td>
<td>0.4706</td>
</tr>
<tr>
<td>1.9</td>
<td>0.4713</td>
<td>0.4719</td>
<td>0.4726</td>
<td>0.4732</td>
<td>0.4738</td>
<td>0.4744</td>
<td>0.4750</td>
<td>0.4756</td>
<td>0.4761</td>
<td>0.4767</td>
</tr>
<tr>
<td>2.0</td>
<td>0.4772</td>
<td>0.4778</td>
<td>0.4783</td>
<td>0.4788</td>
<td>0.4793</td>
<td>0.4798</td>
<td>0.4803</td>
<td>0.4808</td>
<td>0.4812</td>
<td>0.4817</td>
</tr>
<tr>
<td>2.1</td>
<td>0.4821</td>
<td>0.4826</td>
<td>0.4830</td>
<td>0.4834</td>
<td>0.4838</td>
<td>0.4842</td>
<td>0.4846</td>
<td>0.4850</td>
<td>0.4854</td>
<td>0.4857</td>
</tr>
<tr>
<td>2.2</td>
<td>0.4861</td>
<td>0.4864</td>
<td>0.4868</td>
<td>0.4871</td>
<td>0.4875</td>
<td>0.4878</td>
<td>0.4881</td>
<td>0.4884</td>
<td>0.4887</td>
<td>0.4890</td>
</tr>
<tr>
<td>2.3</td>
<td>0.4893</td>
<td>0.4896</td>
<td>0.4898</td>
<td>0.4901</td>
<td>0.4904</td>
<td>0.4906</td>
<td>0.4909</td>
<td>0.4911</td>
<td>0.4913</td>
<td>0.4916</td>
</tr>
</tbody>
</table>
Figure 5.3: A standard normal distribution table

The values in the table are the areas between zero and the z-score. That is, $P(0 < Z < z)$

Some examples that can be read off from the tables are:

For $z=0$, $P(z > 0) = 0.5$
For $z=1$, $P(z > 1) = 0.1587$

What happens if the z-value is negative? Because the table is symmetrical, the area to the left of the negative value will be the same as that to the right of the positive value. Thus

For $z=-1$, $P(z < -1) = P(z > 1) = 0.1587$

How do we determine the area between the two values of $z$?

Using the above example, we want to find $P(-1 < z < 1)$.

The total area under the curve is 1 (one), so we simply subtract the tail areas.

Thus $P(-1 < z < 1) = 1 - (0.1587 \times 2) = 0.6826$
Using the z-distribution to find probabilities for ranges of x-values

Values of x associated with any normal distributed random variable can be converted into corresponding z-values by using the transformation formula:

\[ z = \frac{x - \mu_x}{\sigma_x} \]

Where:

\( \mu_x \) is the mean, and

\( \sigma_x \) is the standard deviation

Let’s take another look at our student marks; the mean calculated was 51.34, and the standard deviation 14.49. Any value of z in this example can thus be calculated using the following formula:

\[ z = \frac{x - 51.34}{14.49} \]

Using the tables, we can determine the frequencies we found in Figure 5.2.

For example, if we wanted to calculate

\[ 36.85 < x < 65.83 \]

We would convert this expression to one that would read:

\[ \frac{36.85 - 51.34}{14.49} < z < \frac{65.83 - 51.34}{14.49} \]

OR

\[ -1 < z < 1 \]

As we saw previously, this area is 68.26%.
Similarly, we can read off the areas for two and three standard deviations on both sides of the mean.

**Two deviations:**

\[22.36 < x < 80.32\]

Or \(-2 < z < 2\)

This gives us an area of \(1 - 2 \times 0.0228 = 0.9544\).

**Three deviations:**

\[7.87 < x < 94.81\]

Or \(-3 < z < 3\)

This gives us an area of \(1 - 2 \times 0.0014 = 0.9972\).

The Standard Normal Distribution tables are most useful for determining the probable frequency of a variable between two limits, and can be used for any data set that follows a normal distribution.

*Figure 5.4: % values on a normal distribution*
Example:

If a test is normally distributed with a mean of 60 and a standard deviation of 10, what proportion of the scores is above 85? This problem is very similar to figuring out the percentile rank of a person scoring 85. The first step is to figure out the proportion of scores less than or equal to 85. This is done by figuring out how many standard deviations above the mean 85 is. Since 85 is 85-60 = 25 points above the mean and since the standard deviation is 10, a score of 85 is 25/10 = 2.5 standard deviations above the mean. Or, in terms of the formula,

\[ z = \frac{X - \mu}{\sigma} \]

\[ = \frac{85 - 60}{10} = 2.5 \]

A z table can be used to calculate that 0.9938 of the scores are less than or equal to a score 2.5 standard deviations above the mean. It follows that only 1-0.9938 = .0062 of the scores are above a score 2.5 standard deviations above the mean. Therefore, only 0.0062 of the scores are above 85.

Computing Normal Probabilities

There are several different situations that can arise when asked to find normal probabilities.
### Situation | Instructions
--- | ---
Between zero and any number | Look up the area in the table
Between two positives, or Between two negatives | Look up both areas in the table and subtract the smaller from the larger.
Between a negative and a positive | Look up both areas in the table and add them together
Less than a negative, or Greater than a positive | Look up the area in the table and subtract from 0.5000
Greater than a negative, or Less than a positive | Look up the area in the table and add to 0.5000

**This can be shortened into two rules.**

1. If there is only one z-score given, use 0.5000 for the second area, otherwise look up both z-scores in the table
2. If the two numbers are the same sign, then subtract; if they are different signs, then add. If there is only one z-score, then use the inequality to determine the second sign (≤ is negative, and ≥ is positive).
Finding z-scores from probabilities

This is more difficult, and requires you to use the table inversely. You must look up the area between zero and the value on the inside part of the table, and then read the z-score from the outside. Finally, decide if the z-score should be positive or negative, based on whether it was on the left side or the right side of the mean. Remember, z-scores can be negative, but areas or probabilities cannot be.

<table>
<thead>
<tr>
<th>Situation</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area between 0 and a value</td>
<td>Look up the area in the table&lt;br&gt;Make negative if on the left side</td>
</tr>
<tr>
<td>Area in one tail</td>
<td>Subtract the area from 0.5000&lt;br&gt;Look up the difference in the table&lt;br&gt;Make negative if in the left tail</td>
</tr>
<tr>
<td>Area including one complete half&lt;br&gt;(Less than a positive or greater than a negative)</td>
<td>Subtract 0.5000 from the area. Look up the difference in the table. Make negative if on the left side.</td>
</tr>
<tr>
<td>Within z units of the mean</td>
<td>Divide the area by 2&lt;br&gt;Look up the quotient in the table&lt;br&gt;Use both the positive and negative z-scores</td>
</tr>
<tr>
<td>Two tails with equal area&lt;br&gt;(More than z units from the mean)</td>
<td>Subtract the area from 1.000&lt;br&gt;Divide the area by 2&lt;br&gt;Look up the quotient in the table&lt;br&gt;Use both the positive and negative z-scores</td>
</tr>
</tbody>
</table>

**ACTIVITY**

Study Chapter 8 of the prescribed text, answer the self-review questions and do the exercises from page 288 to 291.
Index Numbers

An index number is a summary measure of the change in the level of activity of a single item or collection (often referred to as basket) of related items from one time period to another.

It is constructed by expressing the value of an item in the current period as a ratio of its value in the base period.

In percentage terms

\[
\text{Index Number} = \frac{\text{Current Period Value}}{\text{Base Period Value}} \times 100\%
\]

The base period is normally given a value of 100.

Some of the better known index numbers in South Africa are:

- JSE Actuaries Indices – all share index, gold index, industrial index.
- CPI – Consumer Price Index (1985 = 100)
- PPI – Production Price Index (1980 = 100)

There are two major categories of index numbers – price and quantity. In both cases, a single or composite index may be used.

A **price index** measures the percentage change in price between any two periods of time.

For a single item, the relative price change from one time period to another is found by computing its price relative:
A **quantity** index measures the percentage change in consumption level of either an individual item or a basket of items from one time period to another.

For a single item, the relative quantity change from one time period to another is found by computing its quantity relative.

\[
\text{Quantity Relative} = \frac{q_1}{q_0} \times 100%
\]

where

\[
q_1 = \text{quantity in current period} \quad q_0 = \text{quantity in base period}
\]

A **composite index** combines the relative prices and quantities.

A commonly used composite index is the Laspeyres index.

\[
\text{Laspeyres price index} = \frac{\sum (p_1 \times q_0)}{\sum (p_0 \times q_0)} \times 100%
\]

Quantities at **base** period levels are held constant.
Example:

In the following share portfolio problem, the Laspeyres composite index is calculated for price and quantity. The base year in *Figure 5.5* is 1986.

<table>
<thead>
<tr>
<th>Share</th>
<th>Base Year</th>
<th>1992</th>
<th>Base</th>
<th>Price</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p</td>
<td>q</td>
<td>p</td>
<td>q</td>
<td>*</td>
</tr>
<tr>
<td>A</td>
<td>65</td>
<td>350</td>
<td>115</td>
<td>300</td>
<td>22750</td>
</tr>
<tr>
<td>B</td>
<td>200</td>
<td>240</td>
<td>120</td>
<td>60</td>
<td>48000</td>
</tr>
<tr>
<td>C</td>
<td>1260</td>
<td>50</td>
<td>1890</td>
<td>100</td>
<td>63000</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>133750</strong></td>
</tr>
</tbody>
</table>

* $p_0 \times q_0$

** $p_1 \times q_0$

*** $p_0 \times q_1$

*Figure 5.5: Share portfolio performance*

**Laspeyres price index**

\[
\frac{163550}{133750} \times 100 = 122.3\%
\]

This shows that the value of the share increased, on aggregate by 22.3%.

**Laspeyres quantity index**

\[
\frac{157500}{133750} \times 100 = 117.8\%
\]

This result shows us that the number of units of shares held has increased on average by 17.8%.

The price index indicates the increase in the value of the portfolio if all quantities of shares remain the same. Conversely, the quantity index indicates the increase in the shares bought, since all prices have been kept the same in the calculation.
Index numbers are generally based on samples of items. Hence sampling errors are introduced. Furthermore, technological changes, product quality changes and changes in consumer purchasing patterns can individually and collectively make comparisons over time unreliable.

**ACTIVITY**

Study Chapter 7 of the prescribed text, answer the self-review questions and do the student exercises from page 241 to 243.

**Sampling and Sampling Distributions**

It is seldom possible to gather all the data on a random variable under study for analysis purposes; usually only a subset, or sample, is collected.

Any statistical analysis performed on the sample is valid for that sample. But the behaviour of the whole population can be inferred from the behaviour of the sample.

It is important to distinguish between a population and a sample.

A population consists of all possible observations of the random variable under study. Examples are:

- All the residents in a suburb, town or city under study,
- The entire population of cell phone owners in the country.

Gathering data on all possible observations in a population is called a census.
It is seldom practical or necessary to gather data on every possible observation in the population. More commonly, a subset of all observations, or a sample, is gathered on the random variable, analysed and used as the basis for decision making. Sampling is generally preferred, for the following reasons:

- It is more cost effective to gather sample data
- Sample data can be collected more timeously
- Some data requires destructive testing (e.g. battery life, impact resistance of items, shelf life), and a census would not be appropriate
- The data collection for a sample is easier to control, and will thus be more accurate.

According to Wegner (1999), a measure that is found from analysing sample data is called a statistic, while a measure describing a population is called a parameter. The various notations used for these measures are shown in figure 5.6.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Sample</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>( \bar{x} )</td>
<td>( \mu_x )</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>( S_x )</td>
<td>( \sigma_x )</td>
</tr>
<tr>
<td>Size</td>
<td>( n )</td>
<td>( N )</td>
</tr>
<tr>
<td>Proportion</td>
<td>( p )</td>
<td>( \pi )</td>
</tr>
</tbody>
</table>

*Figure 5.6: Notations for Samples and Populations*

Most of the data used in managerial decision making is derived from a sample of observations. However, a manager’s need is to know about the population parameter values of a random variable, not its sample statistics. For example, a quality controller in a beer bottling process is more interested to know the population mean volume of all bottles filled, rather than the sample mean volume of the few filled bottles drawn regularly from the production line and tested.

Inferential statistics is that area of statistics that aims to estimate the true population parameters, with the following process:

- Draw a sample of observations on the random variable under study
- Produce the appropriate sample statistics
- Derive estimates of the values of the corresponding population parameters based on these simple statistics.
Statistical inference is performed in two ways:

- Through estimation where the sample statistic is used to estimate likely values of the corresponding population parameter. Point estimation and confidence interval estimation are techniques used for this.
- Through hypothesis testing, where a claim, statement or hypothesis is made about the population parameter, and tested using sample evidence.

Sampling and Sampling Methods

Sampling is the process of selecting a representative subset of observations from a population to determine characteristics (i.e. the population parameters) of the random variable under study.

There are two basic methods of sampling:

- Non-probability sampling methods
- Probability sampling methods.

**Non-probability sampling methods**

Any sampling method in which the observations are not selected randomly is called non-probability sampling. There are 3 types:

- Convenience sampling, in which a sample is drawn to suit the convenience of the researcher;
- Judgement sampling, where the researcher uses his or her judgement to select the sample;
- Quota sampling, in which the population is divided into segments, and a quota of observations is collected from each segment.

The major disadvantage of non-probability sampling methods is the unrepresentative nature of the sample with respect to the population from which it is drawn. Consequently, results from any statistical inference would probably be invalid.

However, non-probability samples can be used in exploratory research to obtain initial impressions of the characteristics of a random variable under study.
### Probability Sampling Methods

Probability sampling includes all selection methods where the observations to be included in a sample have been selected on a purely random basis from the population.

There are **five types** of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

- **Random sampling** is analogous to putting everyone’s name into a hat and drawing out several names. Each element in the population has an equal chance of occurring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling.

- **Systematic sampling** is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every \( k \)th element is taken. This is similar to lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.

- **Convenience sampling** is very easy to do, but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first people the surveyor runs into.

- **Cluster sampling** is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.

- **Stratified sampling** also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

### The Sampling Distribution

A sampling distribution shows the relationship between a sample statistic and its corresponding population parameter. It describes how a particular sample statistic varies about the true population parameter. From this relationship, the level of confidence in estimating the population parameter from a single sample statistic can be established.
Measures of sample statistics whose behaviour is generally described with respect to their corresponding population parameters are:

- Mean
- Proportion
- Difference between two means
- Difference between two proportions

Let us look at the sampling distribution of a single sample mean through an example from Wegner (1999). We will find the probability that a single sample mean lies within a certain distance of its unknown population.

**Example:** Assume that typing speed, measured in words per minute, is normally distributed. A random sample of 100 typists is selected and their typing speeds measured. Assume that the population deviation of typing speed is 8 words per minute.

What is the probability that the sample mean differs from the unknown population mean of typing speeds by no more than one word per minute in either direction?

Expressed mathematically, we need to find:

\[ P(-1 < \bar{x} - \mu_x < +1) \]

We are studying the behaviour of the sample mean with respect to its population mean. We can use the sampling distribution of the sample means to find the required probability.

The standard deviation of sample means, also called the standard error, is calculated using the formula:

\[ \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} \]

Given \( \sigma_x = 8 \) and \( n = 100 \)

we can calculate \( \sigma_{\bar{x}} = 0.80 \)
Irrespective of the population distribution, the distribution of sample means will always be normal, so we can use the properties of the normal distribution to predict behaviour. The sampling distribution of the sample mean is related to the standard normal probability distribution (the z-distribution) through the following transformation formula:

\[ z = \frac{\bar{x} - \mu_x}{\sigma_{\bar{x}}} \]

or we could write

\[ \bar{x} - \mu_x = z\sigma_{\bar{x}} \]

The original equation can therefore be written as

\[ P(-1 < z\sigma_{\bar{x}} < 1) \]

or

\[ P(-1 < z \times 0.8 < 1) \]

Dividing throughout by 0.8

\[ P(-1.25 < z < 1.25) \]

Reading between the tails from the table, the area between the two values of z is 0.7888. Thus there is a 78.9% chance that a single sample mean of typing speeds will lie within 1 word per minute of the true (but unknown) population mean of typing speeds. This is based on a sample size of 100 typists, and drawn from a normal population of typing speeds, with a standard deviation of 8 words per minute. Or, there is a 21.1% chance that it will be outside one word per minute.

**ACTIVITY**

Study Chapter 9 of the prescribed text, answer the self-review questions and do the student exercises at the end of the chapter.
Worked example (solution on next page)

1. According to the Natal Mercury, 34% of Standard Bank employees resign because of poor pay, 28% resign because of a career change, while the remainder resign due to family commitments. Determine the following probabilities (for a group of seven resigned employees):

   (i) Three resigned due to poor pay.
   (ii) Four resigned due to family commitments.
   (iii) At least one resigned due to career change.

2. Explain why the binomial distribution is relevant in this situation.

3. An executive usually replies to his e-mails fairly quickly. The mean time he takes to reply to his e-mails is 30 minutes, with a standard deviation of 6 minutes. Determine the following probabilities using the z-function:

   (i) he takes between 18 and 42 minutes to answer an e-mail.
   (ii) he takes between 24 and 36 minutes to answer an e-mail.
Solution:

1.

\[ P(r) = \frac{n! \cdot p^r \cdot q^{n-r}}{r! \cdot (n-r)!} \]

(i) \( n = 7, r = 3, (n - r) = 4, p = (0.34) \) and \( q = (0.66) \)

\[ P(3) = \frac{7! \cdot (0.34)^3 \cdot (0.66)^4}{(3!) \cdot (4!)} \]

\[ = 0.261 = 26.1\% \]

(ii) \( n = 7, r = 4, (n - r) = 3, p = (0.38) \) and \( q = (0.62) \)

\[ P(4) = \frac{7! \cdot (0.38)^4 \cdot (0.62)^3}{(4!) \cdot (3!)} \]

\[ = 0.174 = 17.4\% \]

(iii) \( P(0) = \frac{7! \cdot (0.28)^0 \cdot (0.72)^7}{(0!) \cdot (7!)} \)

\[ = 0.101 = 10.1\% \]

\[ P(\text{at least 1}) = 1 - 0.101 = 0.899 = 89.9\% \]

2. The outcomes are exhaustive (even though there are three possible outcomes, the chance of a
successful outcome (p) for one option plus that for an unsuccessful outcome is just the sum of the probabilities of the failure outcomes, q). Thus, \( p + q = 1 \).

3. \( \mu = 30 \) minutes
   \( \sigma = 6 \) minutes
   \[ z = \frac{x - \mu}{\sigma} \]

   (i) \( 18 < x < 42 \) implies \( (18 - 30) / 6 < z < (42 - 30) / 6 \) 
   i.e. \(-2 < z < 2\)
   Probability is 95.50%.

   (ii) \( 24 < x < 36 \) implies \( (24 - 30) / 6 < z < (36 - 30) / 6 \) 
   i.e. \(-1 < z < 1\)
   Probability is 68.26%.
SECTION 6: PREDICTION (CORRELATION AND REGRESSION)
CONTENTS

• Relationships between variables

• Scatterplots

• Simple Linear Regression and Correlation Analysis

• Reliability of prediction results

LEARNING OUTCOMES

• Identify independent and dependent variables.
• Carry out a simple linear regression and correlation analysis
• Recognize reliability of prediction results.
When two variables are related, it is possible to predict a person's score on one variable from their score on the second variable with better than chance accuracy. This section describes how these predictions are made and what can be learned about the relationship between the variables by developing a prediction equation. It will be assumed that the relationship between the two variables is linear. Although there are methods for making predictions when the relationship is nonlinear, these methods are beyond the scope of this module.

Regression and correlation analyses are statistical methods that attempt to quantify and describe possible relationships between variables. This relationship can assist with the prediction of unknown values of certain variables from known values of the related variables.

Regression analysis quantifies the underlying structural relationship between variables.

Correlation analysis determines the strength of this identified association.

Simple linear regression analysis aims to find a linear relationship between the values of two random variables only.

One of the random variables is termed the independent variable (x). It is the variable for which values are known or easily determined, and in certain instances can be controlled.

The other random variable is termed the dependent variable (y). Values are not readily known and need to be estimated from values of the independent variable (x).

Given that the relationship is linear, the prediction problem becomes one of finding the straight line that best fits the data. Since the terms "regression" and "prediction" are synonymous, this line is called the regression line.
The table in figure 6.1 shows pairs of random variables, between which possible relationships exist.

<table>
<thead>
<tr>
<th>Independent Variable (x); Potential predictors of y.</th>
<th>Dependant Variable (y); Variable to be estimated.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertising</td>
<td>Company Turnover</td>
</tr>
<tr>
<td>Training</td>
<td>Labour Productivity</td>
</tr>
<tr>
<td>Speed</td>
<td>Fuel Consumption</td>
</tr>
<tr>
<td>Hours Worked</td>
<td>Machine Output</td>
</tr>
<tr>
<td>Daily Temperature</td>
<td>Electricity Demand</td>
</tr>
<tr>
<td>Hours Studied</td>
<td>Examination Results</td>
</tr>
<tr>
<td>Product price</td>
<td>Product Sales Level</td>
</tr>
<tr>
<td>Bond Interest Rate</td>
<td>Number of Bond Defaulters</td>
</tr>
<tr>
<td>Cost of living</td>
<td>Poverty</td>
</tr>
</tbody>
</table>

*Figure 6.1: Relationships*

Regression analysis aims to find a linear function i.e., a straight line that best fits the actual observations.

**A straight-line graph is defined as follows:**

\[
\hat{y} = a + bx
\]

where

- \( x \) = values of the independent variable
- \( \hat{y} \) = estimated value of the dependent variable
- \( a \) = the y intercept (where the regression line cuts the y axis)
- \( b \) = the slope of the regression line (for every unit change in \( x \), \( y \) changes by \( b \) units)

Graphically the straight line (\( y = a + bx \)) may look as shown in *figure 6.2*:
It is however uncommon to find such a perfect straight-line relationship shown in figure 6.2. We usually talk about the "best-fit" straight line, i.e. a line passing through as many of the data points as possible.
Scatterplot

A scatterplot is a graphical plot of the values of the independent and dependent variables. The independent variables $x$ are recorded along the horizontal axis and the dependent values $y$ along the vertical axis. Pairs of $x$ and $y$ observations are plotted in space.

A visual inspection of the likely relationship between the two variables $x$ and $y$, as provided by a scatterplot, will provide an initial insight into the likely regression and correlation analysis results.

If for example the data points are widely scattered and the range of $y$ values is large for any given $x$ value, then a linear regression function will be of little value as an estimation function for $y$, and the correlation measure will show almost no association.

Examples of various scatterplots are shown in figure 6.3 below:

![Graph showing direct linear relationship with small dispersion]
Inverse linear relationship
With small dispersion

Direct linear relationship
With greater dispersion
When we do regression analysis we try to find the "best fit" line. The strength of the fit is indicated by the correlation.
The regression line is that line which minimises the sum of the squared deviations of the observations from the fitted line. Without providing the derivation, the coefficients $a$ and $b$ that result from this “method of least squares” are as follows:

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$a = \frac{(\sum y) - b(\sum x)}{n}$$

The correlation coefficient most commonly used is Pearson’s correlation coefficient ($r$), which is calculated as follows:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2}(n\sum y^2 - (\sum y)^2)}$$

Here are some properties of $r$

- $r$ only measures the strength of a linear relationship. There are other kinds of relationships besides linear.
- $r$ is always between -1 and 1 inclusive. -1 means perfect negative linear correlation and +1 means perfect positive linear correlation. 0 means a poor (or no) correlation.
- $r$ has the same sign as the slope of the regression (best fit) line.
- $r$ does not change if the independent ($x$) and dependent ($y$) variables are interchanged.
- $r$ does not change if the scale on either variable is changed. You may multiply, divide, add, or subtract a value to/from all the $x$-values or $y$-values without changing the value of $r$.

The correlation coefficient is a dimensionless number since it is a proportion.

A low correlation does not necessarily imply that the variables are unrelated, but simply that a straight line poorly describes the relationship. A nonlinear relationship may well exist. Pearson’s correlation coefficient does not identify non-linear association.
A correlation does not necessarily imply a cause and effect relationship, merely an observed association.

Example:

Most of South Africa’s power stations are coal fired. Assume a random sample of 10 power stations was selected and their coal usage and electricity generated for 1992 was obtained.

The data are shown in figure 6.4.

<table>
<thead>
<tr>
<th>Coal Usage in 1992 (in million tons)</th>
<th>Electricity Generated (million kilowatt hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>35</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>18</td>
<td>32</td>
</tr>
<tr>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
</tr>
<tr>
<td>14</td>
<td>32</td>
</tr>
<tr>
<td>11</td>
<td>29</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
</tr>
</tbody>
</table>

*Figure 6.4: Coal Usage for Electricity generated*

In this case, electricity generated is the dependent variable $y$ and the coal usage the independent variable $x$.

A scatterplot of the data is shown below in figure 6.5, along with the best fit line (dotted).
From the scatterplot, we can already see a strong linear (direct) relationship between coal usage $x$ and electricity generated.

There is little dispersion, since the points lie near the best line fit.

When carrying out linear regression calculations it is useful to construct the table shown in figure 6.6.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$x^2$</th>
<th>$xy$</th>
<th>$y^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>35</td>
<td>225</td>
<td>525</td>
<td>1225</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>36</td>
<td>108</td>
<td>324</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
<td>100</td>
<td>240</td>
<td>576</td>
</tr>
<tr>
<td>18</td>
<td>32</td>
<td>324</td>
<td>576</td>
<td>1024</td>
</tr>
<tr>
<td>9</td>
<td>24</td>
<td>81</td>
<td>216</td>
<td>576</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>49</td>
<td>140</td>
<td>400</td>
</tr>
<tr>
<td>14</td>
<td>32</td>
<td>196</td>
<td>448</td>
<td>1024</td>
</tr>
<tr>
<td>11</td>
<td>29</td>
<td>121</td>
<td>319</td>
<td>841</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>25</td>
<td>70</td>
<td>196</td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>64</td>
<td>176</td>
<td>484</td>
</tr>
<tr>
<td>Total</td>
<td>103</td>
<td>250</td>
<td>1221</td>
<td>2818</td>
</tr>
</tbody>
</table>

Figure 6.5: Scatterplot

Figure 6.6: Calculations for Linear Regression
We can now calculate $b$ and $a$ using the above formulae:

$$
b = \frac{10(2818) - (103)(250)}{10(1221) - (103)^2}
= 1.518
$$

$$
a = \frac{250 - (1.518)(103)}{10}
= 9.367
$$

We can therefore define the estimated regression line as:

$$
\hat{y} = 9.367 + 1.518x
for \ 5 \leq x \leq 18.
$$

Pearson's correlation coefficient can be calculated as follows:

$$
r = \frac{10(2818) - (103)(250)}{\sqrt{(10)(1221) - (103)^2)((10)(6670) - (250)^2)}
= .9371
$$

This correlation coefficient is close to +1, hence the association between $x$ and $y$ is very strong and positive. Values of $x$ can therefore confidently be used to estimate values of $y$.

The regression line can be used to estimate values of $y$ from known values of $x$, by substituting the given $x$ value into the regression equation.

For example, estimate the level of electricity that would be generated for 12 million tons of coal:

$$
\hat{y} = 9.367 + 1.518x
= 9.367 + 1.518(12)
= 27.58
$$

Thus with 12 million tons of coal, 27.58 million kilowatt hours of electricity can be expected to be generated.
Notes on extrapolation

Extrapolation is the process of estimating values of y, using values of x which lie outside the range of x values used in the construction of the estimated regression line. In our example, valid estimates of y are produced only from values within the domain of x between (and including) the values of 5 and 18.

If values of y are estimated outside the limits of the domain of x, the estimates can be unreliable as the relationship between x and y outside these limits is unknown and may in fact be quite different to that which is defined within the domain.

For example, if we substitute x = 0 in our regression equation,

\[ y = 9.367 + 1.518 (0) \]
\[ = 9.367 \]

We could interpret this as meaning that 9,367 million kilowatt-hours of electricity will be generated if no coal is used – this is clearly nonsense!

ACTIVITY

1. Study Chapter 12 of the prescribed text, “Business Forecasting: Simple Linear Regression”.
   . Answer the self-review questions.

2. Study the spreadsheet programme you are using, and make sure you are able to carry out the linear regression options given.

3. Carry out the exercises from page 414 to 416.
Worked example (solution on next page)

STANDARD BANK cashiers get rewarded with annual bonuses, depending on the number of years of service given to the bank. Below is a table showing the bonus (in R 000s) and the number of years of service of the employees.

<table>
<thead>
<tr>
<th>Bonus (R 000's)</th>
<th>Number of years of service</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>8</td>
</tr>
<tr>
<td>23</td>
<td>6</td>
</tr>
<tr>
<td>37</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>60</td>
<td>14</td>
</tr>
<tr>
<td>45</td>
<td>11</td>
</tr>
</tbody>
</table>

1. Portray the above information in a scatter-graph.
2. Determine the co-efficient of correlation and interpret its value.
3. Calculate the linear regression equation and use it to predict the number of years of service of an employee who received a bonus of R 70 000.
4. Is this value reliable? Explain.
Solution:

1.

2.

<table>
<thead>
<tr>
<th>X</th>
<th>y</th>
<th>xy</th>
<th>x²</th>
<th>y²</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>32</td>
<td>256</td>
<td>64</td>
<td>1024</td>
</tr>
<tr>
<td>6</td>
<td>23</td>
<td>138</td>
<td>36</td>
<td>529</td>
</tr>
<tr>
<td>9</td>
<td>37</td>
<td>333</td>
<td>81</td>
<td>1369</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>33</td>
<td>9</td>
<td>121</td>
</tr>
<tr>
<td>14</td>
<td>60</td>
<td>840</td>
<td>196</td>
<td>3600</td>
</tr>
<tr>
<td>11</td>
<td>45</td>
<td>495</td>
<td>121</td>
<td>2025</td>
</tr>
<tr>
<td>Σ</td>
<td>51</td>
<td>208</td>
<td>2095</td>
<td>507</td>
</tr>
</tbody>
</table>
\[ r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left(\frac{n\sum x^2 - (\sum x)^2}{\sum y^2 - (\sum y)^2}\right)^2}} \]

\[ = \frac{6 (2095) - (51)(208)}{\sqrt{\left(\frac{6 x 507 - 51^2}{6 x 8668 - 208^2}\right)^2}} \]

\[ = 1962 \]

\[ ----- \]

\[ = 1963.70 \]

\[ = + 0.999 \]

This value of \( r \) indicates a strong, direct linear relationship between number of years of service and bonus.

3. \[ b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \]

\[ = \frac{(6 x 2095) - (51 x 208)}{6 x 507 - 51^2} \]

\[ = 1962 \]

\[ ----- \]

\[ = 441 \]

\[ = 4.45 \]
\[ a = \frac{(\Sigma y) - b(\Sigma x)}{n} \]

\[ = \frac{208 - 4.45 \times 51}{6} \]

\[ = -3.16 \]

Therefore, the regression line is: \[ y = a + bx \]

\[ = -3.16 + 4.45x \]

For \( y = 70 \), we can use the regression equation to find \( x \).

\[ x = \frac{(3.16 + 70)}{4.45} = 16.4 \approx 16 \text{ years.} \]

4. No. The line is only valid for \( x \) values between 3 and 14 years. Using it to predict values outside this range is extrapolation.
SECTION 7: FORECASTING METHODS
USING TIME SERIES ANALYSIS
CONTENTS

• Components of a Time Series

• Trend Analysis

• Moving Average

• Exponential Smoothing

• Seasonal Analysis

LEARNING OUTCOMES

The learner should be able to:

• Understand and state the principles of forecasting and time series

• State the principles of seasonality and trend

• Calculate a trend using moving averages, and illustrate it on a graph

• Calculate seasonal components and interpret these

READING

Prescribed Reading:

This manual has been designed to be read in conjunction with the following textbook:

Forecasting is an integral part of business management. The better the forecast, the better management will be able to plan for the future.

Although there are many methods for making forecasts, some are better suited than others for particular situations. Forecasting is a critical function that needs to be done by businesses. It is needed to assist us in financial planning determining staff levels and ordinary raw materials for production and other business functions.

The most common tool used for forecasting is time series analysis. It assumes that the actual values of a random variable in a time series are influenced by a variety of environmental forces operating over time. Time series analysis attempts to isolate and quantify the influence of these different environmental forces operating on the time series into a number of different components.

**Components of a Time Series**

Time series analysis assumes that four underlying forces individually and collectively determine the random variables value in a time series in any time period. They are

- Trend (T)
- Cyclical Variations (C)
- Seasonal Variations (S)
- Random (irregular) variation (R)

**Trend (T)**

Trend is defined as a long-term smooth underlying movement in time series. It describes the effect that long-term factors have on the series. These long-term factors tend to operate fairly gradually and in one direction for a long period of time. Thus, a smooth curve or a straight line, such as shown in figure 7.1 usually describes the trend component:
Examples of long-term trends are

- Population growth
- Urbanisation
- Technological improvements
- Shifts in habits and attitudes

**Cyclical Fluctuation (C)**

Cycles are medium to long term deviations from the trend. They reflect alternating periods of relative expansion and contraction. They are wave like movements in a time series that can vary greatly in duration and amplitude. They are difficult to measure statistically and their use in statistical forecasting is limited.

The most common form of cycle is the business cycle between periods of relatively good economic activity to poor economic activity. The causes of these are difficult to determine. Action by government, trade unions and world organisations induce levels of pessimism and optimism into the economy which are reflected in changes in the time series levels. Index numbers are used to describe cyclical fluctuations. An illustration of cycles is shown in figure 7.2.
Seasonal Variations (S)

Seasonal variations are fluctuations that are repeated periodically, usually within a year (i.e. daily, weekly, monthly or quarterly). They are readily isolated through statistical analysis. Seasonal fluctuations are caused by re-occurring events such as climatic conditions, special occurring events (e.g. Easter, Christmas) and religious, public and school holidays. An example is shown in figure 7.3. The regular patterns of seasonal fluctuation are measured by seasonal indices.
Random Fluctuation (I)

These are caused by unpredictable occurrences, which may be evident or sometimes not so evident. Examples of evident events are natural disasters such as floods, droughts or fires and man-made disasters such as strikes or boycotts.

These variations follow no specific pattern, and cannot be analysed statistically, and thus cannot be incorporated into forecasts.

Decomposition of a Time Series

By using time series analysis, we try to isolate the influence of each of the four components on the series. We do this through the Multiplicative Time Series Model, which states that the actual values of a time series, \( y \), can be found by multiplying the trend component by each of the following:

- Cyclical index \( C \)
- Seasonal index \( S \)
- Irregular measure \( I \)

The trend component is expressed in the active units of the variable we are looking at. The seasonal and cyclical indices are, by definition, index numbers and expressed relative to the trend.

Mathematically, this is expressed as:

\[
y = T \times C \times S \times I
\]

Statistical analysis can be used effectively to isolate the trend (T) and the seasonal (S) components, but is of less value in quantifying the cyclical movements, and of no value in isolating irregular components.

We will examine statistical approaches to quantify Trend and Seasonal variation only. More sophisticated models would be needed to isolate the other two components.
Analysing Trend Using Moving Averages

The most common methods used for trend isolation are:

- The moving average which produce a smooth curve
- Regression analysis which involves fitting a straight line

Regression analysis was discussed in the last section.

The moving average removes the short-term fluctuations in a time series by taking successive averages of groups of observations.

To illustrate the method, let’s say we sold 30 widgets during the month of June. We want to estimate what our sales will be for July. Our best guess might be that we will sell 30 widgets during July – basically we have used a “one month moving average” as our forecast.

When we want to forecast for August, we may want to take into account what happened during June and July. Let’s say we had sales of 40 during July. If we took a two-month moving average, our forecast for August would be

\[
(F/C)_{August} = \frac{(Actual)_{June} + (Actual)_{July}}{2}
\]

\[
= \frac{30 + 40}{2}
\]

\[
= 35
\]

What do we do for September? Let’s say the sales for August were 30. We now have a choice between 3 forecasts:

1-Month moving average:

\[
(F/C)_{Sept} = (Actual)_{Aug} = 30
\]
2-Month moving average:

\[
(F / C)_{\text{Sept}} = \frac{(Actual)_{\text{Aug}} + (Actual)_{\text{July}}}{2} \n= \frac{30 + 40}{2} \n= 35
\]

3-Month moving average:

\[
(F / C)_{\text{Sept}} = \frac{(Actual)_{\text{Aug}} + (Actual)_{\text{July}} + (Actual)_{\text{June}}}{3} \n= \frac{30 + 40 + 30}{3} \n= 33.3
\]

The table in the figure 7.4 below demonstrates the calculation for each forecast, while figure 7.5 shows the results graphically.

<table>
<thead>
<tr>
<th>MONTH</th>
<th>ACTUAL</th>
<th>MOVING AVERAGE FORECAST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 MONTH</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>30</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>35</td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>52</td>
<td>45</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>52</td>
</tr>
<tr>
<td>9</td>
<td>36</td>
<td>35</td>
</tr>
<tr>
<td>10</td>
<td>38</td>
<td>36</td>
</tr>
<tr>
<td>11</td>
<td>60</td>
<td>38</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>13</td>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>14</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>15</td>
<td>55</td>
<td>50</td>
</tr>
</tbody>
</table>
From the graph plots, we can see that

- No plot is available in the early months for the 3-month and the 5-month moving averages, as enough data is not available. For example, the first month in which a 5 month moving average forecast can be used is month 6.

- The higher the number of months used in a moving average, the smoother the curve – compare the 5 month to the 3 month and the 1 month curves. Another way of seeing it is that the higher the number of months, the less the curve is influenced by variations in trend.

- The 1-month moving average replicates the actual exactly, but “lags” by one month. This is characteristic of all simple moving average curves – they will lag the actual figures, i.e. upward or downward shifts in trend will only be detected after the event. The more months used to calculate the moving average, the longer it takes for the change to register.

**Exponential Smoothing**

The moving average technique has the advantage that it is simple to use and easy to understand. Two of its major disadvantages are:
The forecast always lags the actual, as discussed in the last chapter
No account is taken of the error in previous forecasts

The exponential smoothing technique allows us to calculate a “smoothed average” which consists of two parts:

- The most recent demand (new information) and
- The historical smoothed average (old information)

The smoothed average can be calculated by using the following formula:

Choosing a value for the smoothing coefficient can be problematic; the best way to carry this out is to test different values and see which yields the lowest error.

\[ F_{t+1} = F_t + \alpha (D_t - F_t) \]

where

- \( F_{t+1} \) = forecast for the next period represented by \( t + 1 \)
- \( F_t \) = forecast for the latest period, represented by \( t \)
- \( D_t \) = actual demand for period \( t \)
- \( \alpha \) = smoothing coefficient

The error to test would be the mean absolute deviation or MAD, calculated with the following formula:

\[
MAD = \frac{\sum_{t=1}^{n} |D_t - F_t|}{n}
\]

where

- \( |D_t - F_t| \) = absolute value of the error
- \( n \) = number of periods being reviewed

It is common to use a smoothing value in the ranges of 0.1 to 0.3.

Example:
Assume the last forecast was for 100, but only 90 was actually sold.
Set the smoothing coefficient at 0.2.

Using the above formula, we can calculate the exponentially smoothed forecast:

\[
F_r = 90 \\
D_r = 100 \\
\alpha = 0.2 \\
F_{r+1} = 100 + 0.2(90 - 100) \\
= 98
\]

Example:

The table in the figure 7.6 below shows calculations for exponential smoothing. We need a “start up” forecast value; it is common to use a moving average for this. The moving average calculated for the previous three weeks was 70; this value is used as “start up”.

<table>
<thead>
<tr>
<th>Week</th>
<th>Actual Demand</th>
<th>Forecast Alpha = .2</th>
<th>Forecast Alpha = .5</th>
<th>Forecast Alpha = 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Demand</td>
<td>Forecast</td>
<td>Absolute Error</td>
<td>Forecast</td>
</tr>
<tr>
<td>1</td>
<td>80</td>
<td>70</td>
<td>10</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>85</td>
<td>72</td>
<td>13</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>90</td>
<td>75</td>
<td>15</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
<td>78</td>
<td>22</td>
<td>85</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
<td>82</td>
<td>23</td>
<td>93</td>
</tr>
<tr>
<td>6</td>
<td>115</td>
<td>87</td>
<td>28</td>
<td>99</td>
</tr>
<tr>
<td>7</td>
<td>110</td>
<td>92</td>
<td>18</td>
<td>107</td>
</tr>
<tr>
<td>8</td>
<td>120</td>
<td>96</td>
<td>24</td>
<td>108</td>
</tr>
<tr>
<td>9</td>
<td>125</td>
<td>101</td>
<td>24</td>
<td>114</td>
</tr>
<tr>
<td>10</td>
<td>115</td>
<td>106</td>
<td>9</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>187</td>
<td>104</td>
<td>104</td>
</tr>
</tbody>
</table>

*Figure 7.6: Calculations for Exponential Smoothing*
Seasonal Analysis

A seasonal index is the ratio of the demand for a particular season to the demand for an average season. Thus, if demand is for 100 units in an average season and demand for the summer season is 80, the summer season index is $80 / 100 = 0.8$. Some sort of averaging process is used to arrive at the average value of 100 per season.

Example: Metro Movers and Seasonal Indices

There is no need to look at actual demand data to know that a moving company has a seasonal demand – it is common sense. Even so, a good starting point in seasonal analysis is scrutiny of the demand graph. The graph of past quarterly demand for Metro Movers is shown in figure 7.7. It is clear that summer demand is by far the highest in every year and autumn demand is generally the lowest. The seasonal index measures how much higher and how much lower. Figure 7.8 shows calculations of seasonal indices for the 16 available past demands. (Note. Besides seasonality, it looks like there is a slight upward trend over the 16 quarters. We shall ignore the trend for now.)

![Figure 7.7: Seasonal Demand history for Metro Movers](image-url)
The mean seasonal demand is calculated using a four-period moving average, and is centred on the middle of a given season, that is a month and a half into the season. It includes demands going back six months and forward six months from that point. Thus, the first figure in column 3 is based on demands for the last one and a half months of spring 1997; and all of summer, autumn and winter 1997; and the first one and a half months of spring 1998. So,

This is a bit cumbersome, but it ensures that no one season is weighted more heavily than any other.

\[
\frac{(90/2) + 160 + 70 + 120 + (130/2)}{4} = 115
\]

The seasonal indices are shown rearranged by year and season in figure 7.9. The three values for each season need to somehow be reduced to a single index. The index for autumn is steadily rising, from 0.61 to 0.73 to 0.96. That is not sufficient reason to expect it to continue to rise, however, especially since the other seasons do not show trends. Thus, the projections of the seasonal indices for 2001 are the means of each column.
Metro movers may now use the seasonal indices in fine-tuning its demand forecasts for each coming season. For example, suppose that they expect to move 480 vans of goods next year based on projection of the mean of past years’ demands. It would be naïve to divide 480 by 4 and project 120 vans in each season. Instead,

Divide 480 by 4 = 120 vans in an average season
Multiply the average by the index for each season:

Spring 2001 : $120 \times 0.85 = 102$ vans
Summer 2001: $120 \times 1.46 = 175$ vans
Autumn 2001 : $120 \times 0.76 = 91$ vans
Winter 2001 : $120 \times 0.93 = 112$ vans

Yearly total = 480

**Example:** The manager of Black Belt tracks the department’s weekly output of pallets. Each pallet holds a constant number of cases of product and the manager uses a simple, four-week moving average in a spreadsheet. The table *(Figure 7.10)* below shows a sample, from the end of a 52-week cycle, of the department’s production of pallets.
Table: Pallet Production by Week

<table>
<thead>
<tr>
<th>Week</th>
<th>Pallets</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>192</td>
</tr>
<tr>
<td>49</td>
<td>178</td>
</tr>
<tr>
<td>50</td>
<td>193</td>
</tr>
<tr>
<td>51</td>
<td>205</td>
</tr>
<tr>
<td>52</td>
<td>218</td>
</tr>
</tbody>
</table>

Figure 7.10: Black Belt Pallet production

The manager has the two basic ingredients needed for generating any forecasts: production data and a forecasting period. The period, or divisor, in this case is weeks. With this information, she can execute both the short-term and long-term forecasting methods.

**Short Term: Looking for Trends in Moving Average Plots**

Statistical software can provide Black Belts with several options for completing forecasts. In this case, for a short-term prediction, the manager chooses to plot the moving average by using a time series.

The figure 7.11 below shows the manufacturing manager's four-week moving average from the past year as it would appear in a software program.
Although the visual representation of the analysis is helpful, the true focus here is the accuracy measures, which represent the differences between the actual and the forecasted pallet quantities. One of these accuracy measures is Mean Absolute Deviation (MAD). It gauges the accuracy of the fitted time series values and expresses the deviation in the same units as the data, which makes it easier to understand the amount of error.

**The formula for MAD:**

$$MAD = \sum \left| y - \hat{y} \right| / n$$

Where $y$ is the actual value at a time, $\hat{y}$ is the fitted value and $n$ is the number of observations.

<table>
<thead>
<tr>
<th>Length of Moving Average</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 weeks</td>
<td>15.56</td>
</tr>
<tr>
<td>4 weeks</td>
<td>14.66</td>
</tr>
<tr>
<td>5 weeks</td>
<td>13.28</td>
</tr>
<tr>
<td>6 weeks</td>
<td>13.72</td>
</tr>
<tr>
<td>7 weeks</td>
<td>14.06</td>
</tr>
</tbody>
</table>

Because the manager is looking for a forecast with the least amount of prediction error, it is best to iterate through different lengths of the moving average in order to find lower values of MAD. The table, left, shows the results for five different moving-average iterations.

The table illustrates that the manager would have a slightly more accurate forecast with a five- or six-week moving average.

When examining the graph in Figure 7.11, the manager may also notice that there are extreme values at points 40 and 45 and that the predicted values were essentially pulled down around these points. This should create interest for further review.

One way to the manager can conduct this review and assess the effects of the two extreme points is to place the data into an individuals control chart, as shown in the figure 7.13 below, and see if there is deviation outside of the 3-sigma control limits.
Points 40 and 45 do exceed the control limits. Of course, production output is not a single process and cannot be controlled simply by applying statistical process control, but the individuals chart is a familiar tool for Black Belts and may provide valuable insight for the manager’s forecast.

Upon review of the points outside the control limits, the manager finds a probable explanation: They occurred at two holidays, Thanksgiving and Christmas, when the department was shut down for several days. Knowing this, the manager removes the two points from the data set and reruns the moving averages to see if the MAD decreases.

The manager finds that the MAD does decrease after removing the two extreme points; the updated data is shown in the table (Figure 7.14) below.

<table>
<thead>
<tr>
<th>Length of Moving Average</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 weeks</td>
<td>11.88</td>
</tr>
<tr>
<td>4 weeks</td>
<td>11.63</td>
</tr>
<tr>
<td>5 weeks</td>
<td>11.03</td>
</tr>
<tr>
<td>6 weeks</td>
<td>11.29</td>
</tr>
<tr>
<td>7 weeks</td>
<td>11.05</td>
</tr>
</tbody>
</table>

The manager can now expect better short-term forecasts using a five-week period. Operations are dynamic, however, and it would be best to review the forecast periodically and adjust as necessary.
Study Chapter 13, Business Forecasting: Time Series Analysis, of the prescribed text, answer the self-review questions and carry out the exercises from page 442 to 447.
BIBLIOGRAPHY
BIBLIOGRAPHY


