

Master of Business Administration

Quantitative Methods

STUDY GUIDE

Copyright © 2016

MANAGEMENT COLLEGE OF SOUTHERN AFRICA

All rights reserved; no part of this book may be reproduced in any form or by any means, including photocopying machines, without the written permission of the publisher.

Please report all errors and omissions to the following email address: modulefeedback@mancosa.co.za

TABLE OF CONTENTS		
UNIT	TITLE OF SECTION	PAGE
	General outcomes	2
	Prescribed reading	3
1	Summarising data: summary tables and graphs	4
2	Measures of central location	23
3	Measures of dispersion (variability)	49
4	Probability	68
5	Probability distributions	86
6	Hypothesis testing	103
7	Simple linear regression and correlation analysis	134
8	Forecasting: Time series analysis	147
9	Decision analysis: Decision trees and payoff tables	163
	Solutions to unit exercises	184
	References	214
	Tables	215

General outcomes

Studying this module will enable the student to:

- Apply simple statistical tools and analysis to solve business-related problems.
- Interpret and analyse business data for production, planning, forecasting and other decision-making functions.
- Communicate effectively with statistical analysts.
- Apply quantitative methods and techniques to other management disciplines – economics, accounting, financial management, marketing and research.

Syllabus: The syllabus for the module is as follows:

Topic 1: Descriptive statistics:

- a) Summarising data: summary tables and graphs
- b) Measures of central location
- c) Measures of dispersion (variability)
- d) Probability and probability distributions

Topic 2: Inferential statistics:

- a) Hypothesis testing
- b) Simple linear regression and correlation analysis

Topic 3: Forecasting – time series analysis

Topic 4: Decision analysis – decision trees and payoff tables



Prescribed textbook:

Wegner, T (2012). Applied Business Statistics: Methods and Excel-based Applications (3rd edition), Juta & Co, Ltd: Cape Town

Recommended textbook:

Lind, Marchal and Wathen (2005). Statistical Techniques in Business and Economics (12th edition), New York: McGraw-Hill. Chapter 1

The purpose of this module

Statistics as a subject has been included in the MBA curriculum because it is needed in two main areas:

1. Descriptive statistics are used in subjects like finance and operations to describe business phenomena.
2. Descriptive and inferential statistics are widely used in marketing research. Business leaders need to understand how to research and test their markets and customer bases effectively in order to market to their customers strategically.
3. It is a requirement for an MBA degree that you complete a research project. In this research project you will have to collect data. In processing the data to make decisions you will need inference. Inference (hypothesis testing) is covered in the latter part of this module.

UNIT 1
SUMMARISING DATA: SUMMARY TABLES AND
GRAPHS

UNIT 1: SUMMARISING DATA; SUMMARY TABLES AND GRAPHS

OBJECTIVES

By the end of this study unit, you should be able to:

1. Recognise whether the type of data under consideration is quantitative or qualitative.
2. Summarise a set of quantitative data by means of a frequency distribution.
3. Graphically represent a set of data using a histogram, frequency polygon and cumulative frequency ogive.
4. Summarise a set of qualitative data by means of pie and bar charts.

CONTENTS

- 1.1 Introduction
 - 1.2 Types of data
 - 1.3 Graphical techniques for quantitative data
 - 1.4 Pie charts, bar charts and line charts
- Exercises



Prescribed textbook: Chapters 1 and 2

1.1 Introduction

Types of data and methods of summarising data are described in this unit.

1.2 Types of data

Statistics is the science of collecting and analysing data. Data are obtained by measuring the values of one or more variables. Data can be classified as either **quantitative** data or **qualitative** data.

1.2.1 Quantitative data

Quantitative data are measurements recorded on a naturally occurring numerical scale.

Some examples of quantitative data are:

- The time you have to wait for the next bus.
- Your height or weight.

1.2.2 Qualitative data

Qualitative data are non-numeric. Some examples of qualitative data are:

- The political party you support.
- Your gender.

Sometimes arbitrary numerical values are assigned to qualitative data, eg for gender, male is assigned a 1 and female is assigned a 2.

The appropriate graphical method to be used in presenting data depends, in part, on the type of data under consideration. Later in the guide, when statistical inference is covered, the data type will help to identify the appropriate statistical technique to be used in solving a problem.



SELF-ASSESSMENT ACTIVITY

How do I identify quantitative data?

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Quantitative data are real numbers. They are not numbers arbitrarily assigned to represent qualitative data. An experiment that produces qualitative data always asks for verbal, non-numerical responses (eg, yes and no; defective and non-defective; Catholic, Protestant and other).

1.2.3 Data classification

Numerical data can also be classified as **discrete** (when only specific values occur – like the number of students in an interval) or **continuous** (when you can have intermediate or fractional values – like height or distance).

Continuous data are sometimes summarised in tables giving the number of data items in each interval.

EXAMPLE

Mass (kg)	Frequency
45 – 50	6
50 – 55	14
55 – 60	25
60 – 65	11



SELF-ASSESSMENT ACTIVITY

For each of the following examples of data, determine whether the data type is quantitative or qualitative:

- The weekly level of the prime interest rate during the past year.
- The make of car driven by each of a sample of executives.
- The number of contacts made by each of a company's salespeople during a week.
- The rating (excellent, good, fair or poor) given to a particular television programme by each of a sample of viewers.
- The number of shares traded on the New York Stock Exchange each week throughout 2012.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

- Quantitative, if the interest rate level is expressed as a percentage. If the level is simply observed as being high, moderate or low, then the data type is qualitative.
- Qualitative.
- Quantitative.
- Qualitative.
- Quantitative.

1.3 Graphical techniques for quantitative data

This section introduces basic descriptive statistics methods used for organising a set of numerical data in tabular form and presenting it graphically. The presentation of the grouped data enables the user to quickly grasp the general shape of the distribution of the data.

1.3.1 Frequency distributions

A frequency distribution is a table with data summarised into groups known as *intervals*.

The steps in creating a frequency distribution are given on **page 35 of the prescribed textbook**.

EXAMPLE

The weights in pounds of a group of workers are:

173	165	171	175	188
183	177	160	151	169
162	179	145	171	175
168	158	186	182	162
154	180	164	166	157

Step 1. Determine the data range.

$$\text{Range} = \text{maximum data value} - \text{minimum data value} = 188 - 145 = 43$$

Step 2. Choose the number of intervals.

Choose five intervals for a small sample size.

Step 3. Determine the interval width.

$$\text{Interval width} = \frac{\text{data range}}{\text{number of intervals}} = \frac{43}{5} = 8,6$$

Using this calculation as a guide, grouping the data into intervals of width 10 pounds makes practical sense.

Step 4. Set up the interval limits.

Lower limit (weight in lbs)	Upper limit
140	< 150
150	< 160
160	< 170
170	< 180
180	< 190

Note: the upper limit is defined as 'up to but not including'. Alternatively for discrete data, ie where no fractional values are encountered, the upper limits can be defined as 149, 159, 169 etc.



TIP

Decide on the upper limit approach you want to use (< 150 or 149) and then be consistent with your approach whenever you need to design a frequency table. This study guide and the textbook have standardised on the 'less than but not including', i.e. < 150, approach.

Step 5. Tabulate the data values.

Interval (weight in lbs)	Frequency
140 – 150	1
150 – 160	4
160 – 170	8
170 – 180	7
180 – 190	5

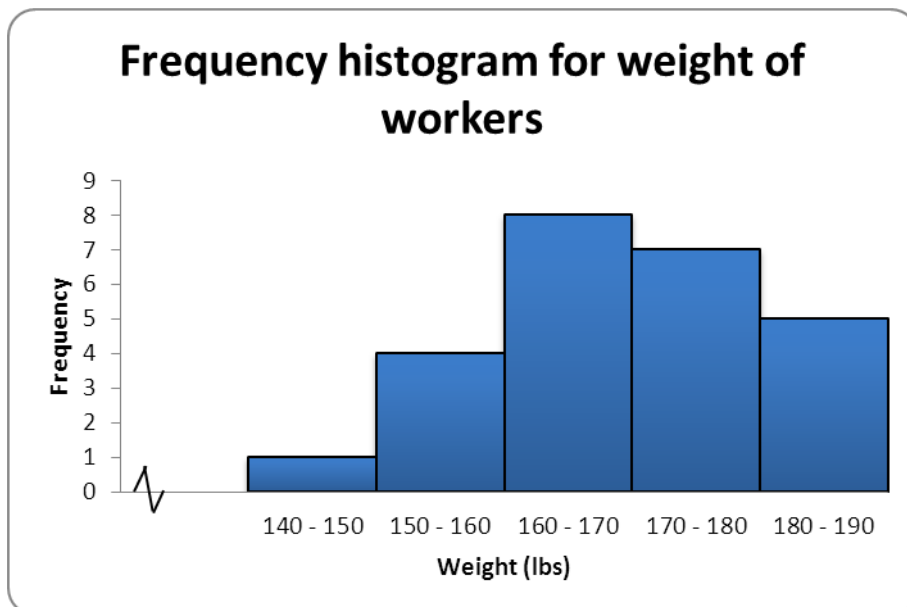
1.3.2 Histogram

A frequency distribution can be graphically depicted as a *histogram*.

The steps in creating a histogram are given on *page 36 of the prescribed textbook*.

EXAMPLE

Using the data from the example in section 1.3.1, the histogram can be depicted as:



TIP

Remember that histograms are similar to bar charts, but the bars touch each other.



TIP

If you graph data and part of an axis is not to scale (in example the x-axis from 0 to 140 is not to scale), show a 'broken' axis.

1.3.3 Frequency polygon

A frequency polygon is constructed by plotting the frequency of each interval above the **midpoint** of that interval and then joining the points with straight lines. The polygon is closed by considering one additional interval (with zero frequency) at each end of the distribution and extending a straight line to the midpoint of each of these intervals.

Before constructing a frequency polygon, calculate the midpoints for each interval.

EXAMPLE

Using the data from the example in section 1.3.1, the midpoints are calculated as:

Interval (weight in lbs)	Frequency	Midpoint
140 – 150	1	145
150 – 160	4	155
160 – 170	8	165
170 – 180	7	175
180 – 190	5	185



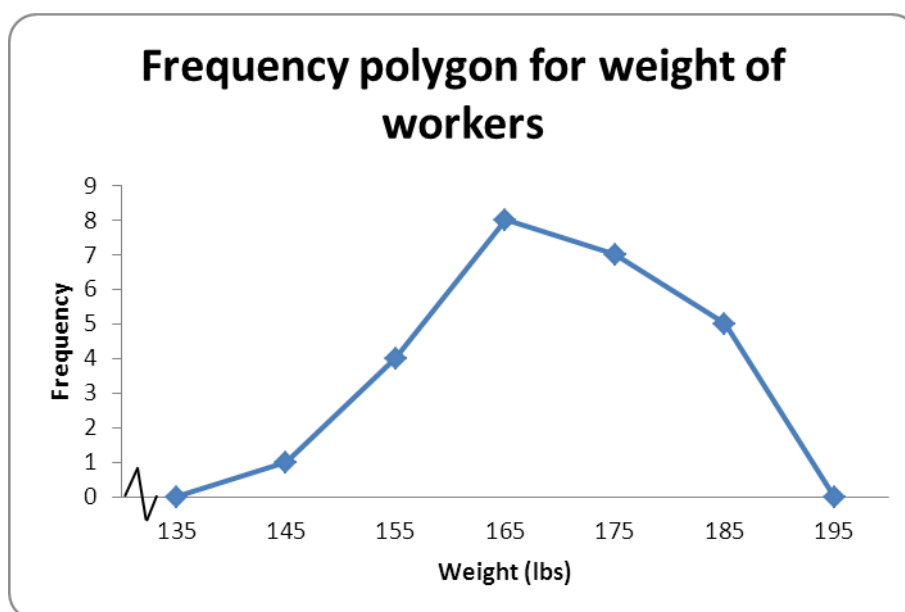
TIP

An easy way to calculate midpoint of an interval is to halve sum of the lower and upper limits.

In the case of example, for the first interval:

$$\frac{\text{lower limit} + \text{upper limit}}{2} = \frac{140 + 150}{2} = 145$$

The frequency polygon can be depicted as:



1.3.4 Cumulative frequency distribution

A cumulative frequency distribution summarises the cumulative frequency of a dataset. It results in a 'running total' of frequencies.

The steps in creating a cumulative frequency distribution are given on **page 39 of the prescribed textbook**.

EXAMPLE

Using the data from the example in section 1.3.1:

For each interval, calculate the cumulative frequency by adding the frequency count of the interval in question to the cumulative frequency of the interval before.

Interval (weight in lbs)	Frequency	Cumulative frequency
140 – 150	1	1
150 – 160	4	5
160 – 170	8	13
170 – 180	7	20
180 – 190	5	25

1.3.5 Ogive

An ogive is a graph of the cumulative frequency distribution.

To construct the ogive, the cumulative relative frequency of each interval is plotted above the upper limit of that interval and the points representing the cumulative frequencies are then joined by straight lines.

The ogive is closed at the lower end by extending a straight line to the lower limit of the first interval.

EXAMPLE

Using the data from the example in section 1.3.1:



1.3.6 Relative distributions

For each of the frequency distribution and the cumulative frequency distribution, relative distributions can be calculated.

A relative frequency distribution includes the percentage of sample size or relative frequency (frequency relative to the total sample size) for each interval.

EXAMPLE

Using the data from the example in section 1.3.1:

Interval (weight in lbs)	Frequency	Relative frequency (factor)	Relative frequency (percentage)
140 – 150	1	0,04	4%
150 – 160	4	0,16	16%
160 – 170	8	0,32	32%
170 – 180	7	0,28	28%
180 – 190	5	0,20	20%

A relative cumulative frequency distribution includes the cumulative percentage of sample size or relative cumulative frequency (cumulative frequency relative to the total sample size) for each interval.

EXAMPLE

Using the data from the example in section 1.3.1:

Interval (weight in lbs)	Frequency	Cumulative frequency	Relative frequency (factor)	Relative frequency (percentage)
140 – 150	1	1	0,04	4%
150 – 160	4	5	0,20	20%
160 – 170	8	13	0,52	52%
170 – 180	7	20	0,80	80%
180 – 190	5	25	1,00	100%

1.4 Pie charts, bar charts and line charts

The methods described in the previous section are appropriate for summarising quantitative data. But we should also be able to describe data that are qualitative or categorical. These data consist of attributes or names of the categories into which the observations are sorted.

1.4.1 Pie chart

A **pie chart** is a useful method for displaying the percentage of observations that fall into each category of qualitative data.

A pie chart is an effective method of showing the percentage breakdown of a whole entity.

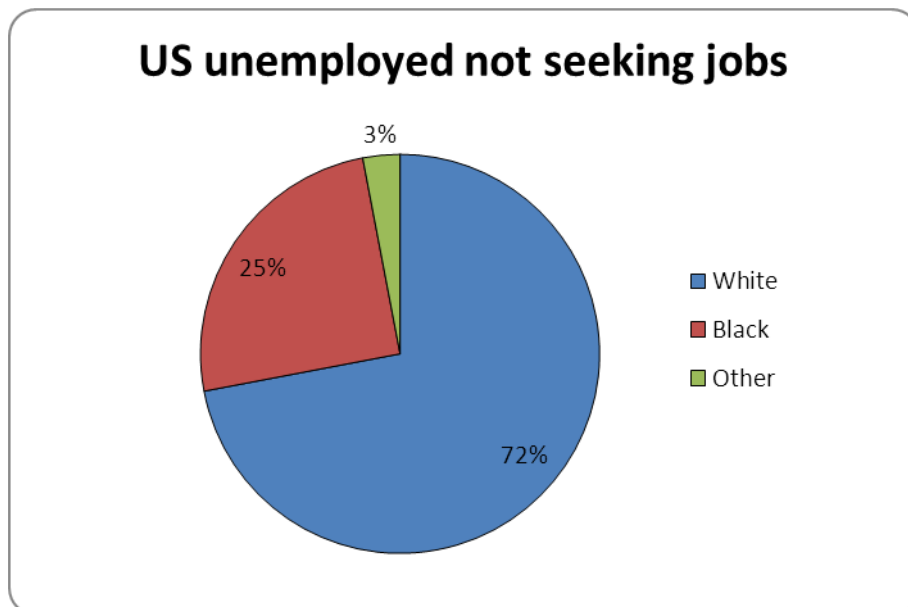
EXAMPLE

A *New York Times* article reports that “6 million Americans who say they want work are not even seeking jobs”. These 6 million Americans are broken down by race:

Race	Frequency
White	4 320 000
Black	1 500 000
Other	180 000

We need to first determine the percentage of the 6 million Americans belonging to each of the three racial categories: 72% white, 25% black and 3% other. Each category is represented by a slice of the pie (a circle) that is proportional in size to the percentage (or relative frequency) corresponding to that category.

Since the entire circle corresponds to 360° , the angle between the lines demarcating the white sector is therefore $(0,72)(360) = 259,2^\circ$. In a similar manner, we can determine the angles for the black and other sectors as 90° and $10,8^\circ$, respectively.

**1.4.2 Bar charts**

Bar charts are a quick and easy way of showing variation in or between variables.

Rectangles of equal width are drawn so that the area enclosed by each rectangle is proportional to the size of the variable it represents. This type of graph not only illustrates a general trend, but also allows a quick and accurate comparison of one period with another or the illustration of a situation a particular time.

When drawing up bar charts take care to:

- Make the bars reasonably wide so that they can be clearly seen.
- Draw them neatly and professionally.
- Ensure that the bars all have the same width.
- Ensure that the gaps between the bars have the same width.

We can produce a variety of bar charts to provide an overview of the data.

1.4.3 Simple bar chart

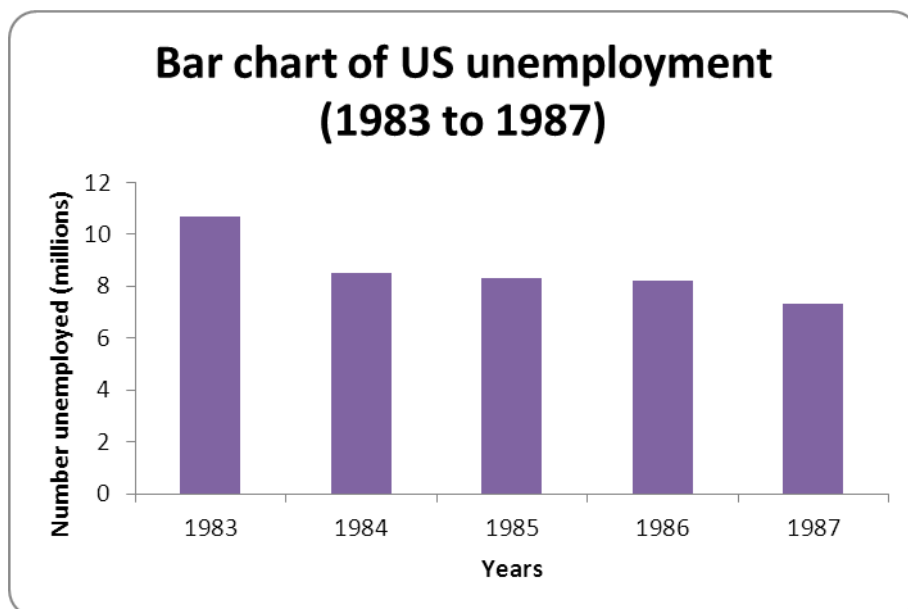
A simple bar chart comprises bars representing each variable drawn either vertically or horizontally. While a **bar chart** can be used to display the frequency of observations that fall into each category, if the categories consist of points in time and the objective is to focus on the trend in frequencies over time, a **line chart** is useful.

EXAMPLE

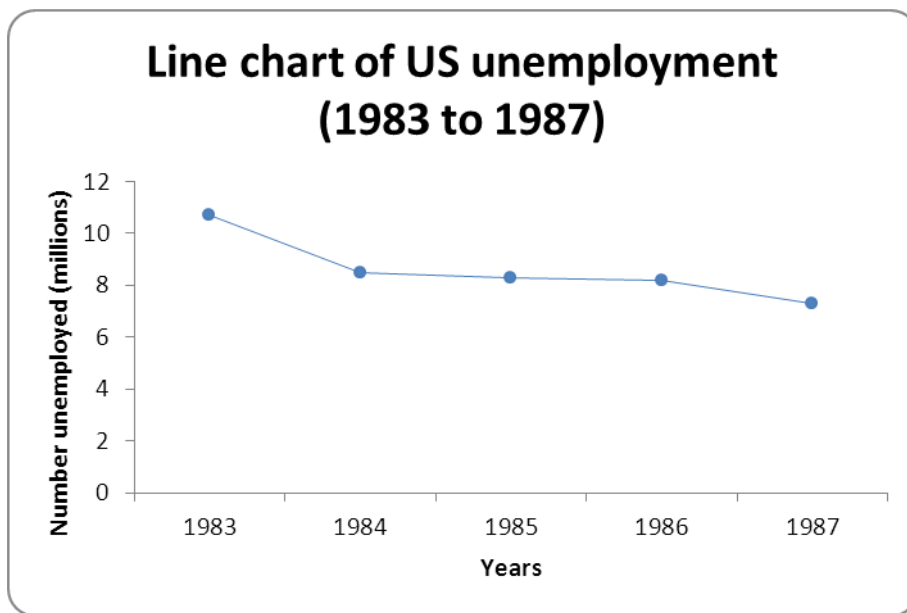
According to the *New York Times* (27 September 1987), the June levels of unemployment in the United States for five years are:

Year	Unemployed (millions)
1983	10,7
1984	8,5
1985	8,3
1986	8,2
1987	7,3

For the bar chart, the five years or categories are represented by intervals of equal width on the horizontal axis. The height of the vertical bar erected above any year is proportional to the frequency (number of unemployed) corresponding to that year.

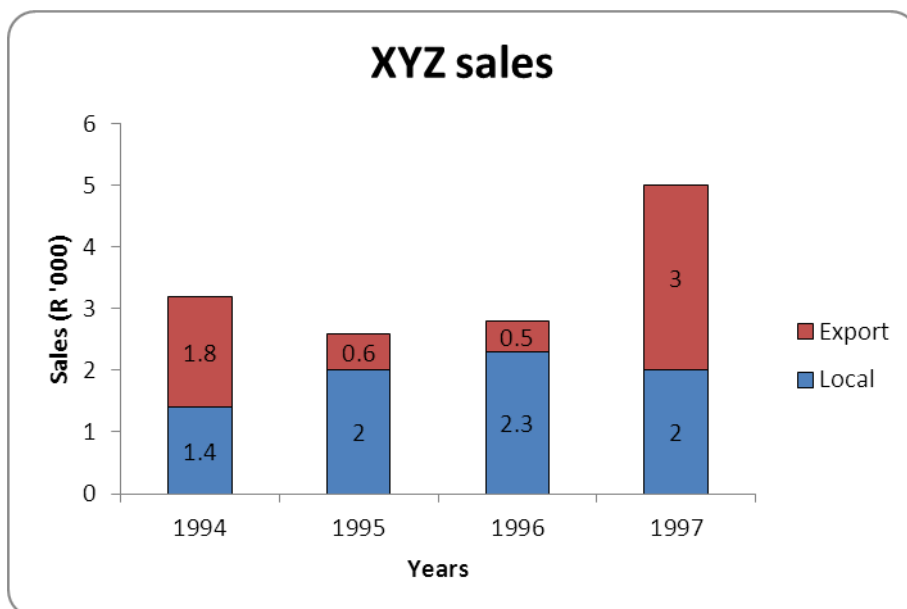


A line chart is obtained by plotting the frequency of a category above the point on the horizontal axis representing that category and then joining the points with straight lines.



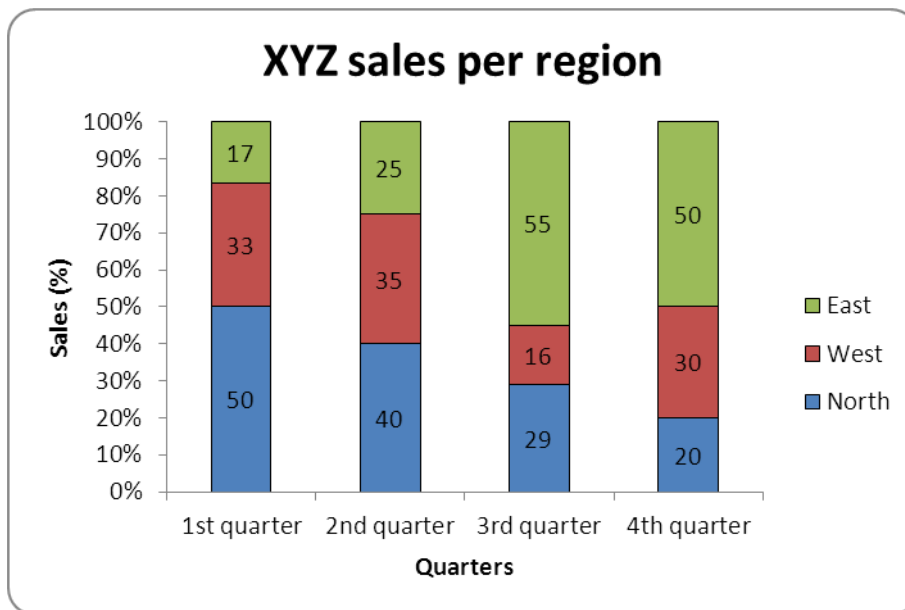
1.4.4 Component or stacked bar chart

In a **component or stacked bar chart**, a single bar is drawn for each variable, with the heights of the bars representing the totals of the categories. Each bar is then subdivided to show the components that make up the total bar. These components may be identified by colouring or shading, accompanied by an explanatory key to show what each component represents.



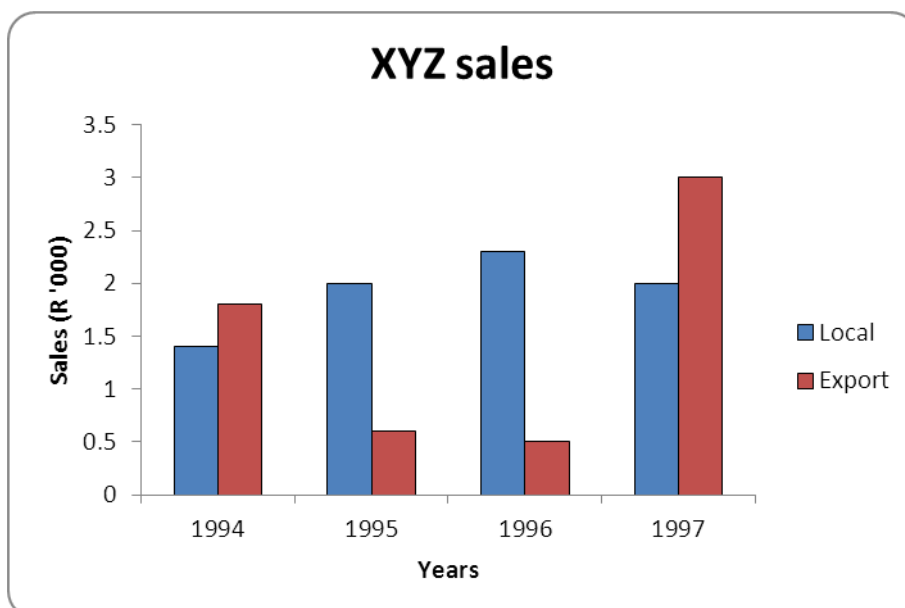
1.4.5 Percentage component bar chart

A percentage component comprises components converted to percentages of the total with the bars divided in proportion to these percentages. The scale is a percentage scale and the height of each bar is therefore 100%.



1.4.6 Multiple bar chart

For multiple or cluster bar charts, two or more bars are grouped together in each category. The use of a key helps to distinguish between the categories.



1.4.7 Scatter diagrams

The relationship between two *quantitative* variables can be depicted in a scatter diagram. Economists, for example, are interested in the relationship between inflation rates and unemployment rates. Business owners are interested in many variables, including the relationship between their advertising expenditures and sales levels.

A scatter diagram is a plot of all pairs of values (x, y) for the variables x and y.

EXAMPLE

An educational economist wants to establish the relationship between an individual's income and education. She takes a random sample of 10 individuals and asks for their income (in R '000s) and education (in years).

x (years of education)	y (income in R '000)
11	25
12	33
11	22
15	41
8	18
10	28
11	32
11	24
17	53
11	26

If we feel the value of one variable (such as income) depends to some degree on the value of the other variable (such as years of education), the first variable (income) is called the **dependent variable** and is plotted on the vertical or y-axis. The second variable is the independent variable and is plotted on the x-axis.

**TIP**

Think of the independent variable (x-axis) as the 'cause' and the dependent variable (y-axis) as the 'effect'.

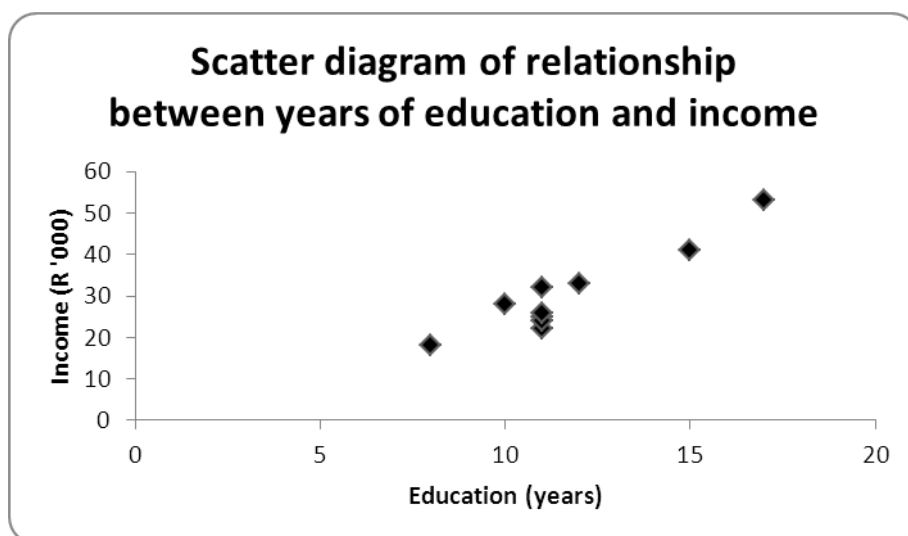


Figure 1

The scatter diagram allows us to observe two characteristics about the relationship between education (x) and income (y):

1. Because these two variables move together, ie their values tend to increase together and decrease together, there is a positive relationship between the two variables.
2. The relationship between income and years of education appears to be linear, since we can imagine drawing a straight line (as opposed to a curved line) through the scatter diagram that approximates the positive relationship between the two variables.

The pattern of a scatter diagram provides us with information about the relationship between two variables. Figure 1 depicts a positive linear relationship.

If two variables move in opposite directions and the scatter diagram consists of points that appear to cluster around a straight line, then the variables have a negative linear relationship (see Figure 2).

It is possible to have nonlinear relationships (see Figure 3 and Figure 4), as well as situations in which the two variables are unrelated (see Figure 5). In Unit 7, we will compute numerical measures of the strength of the linear relationship between two variables.

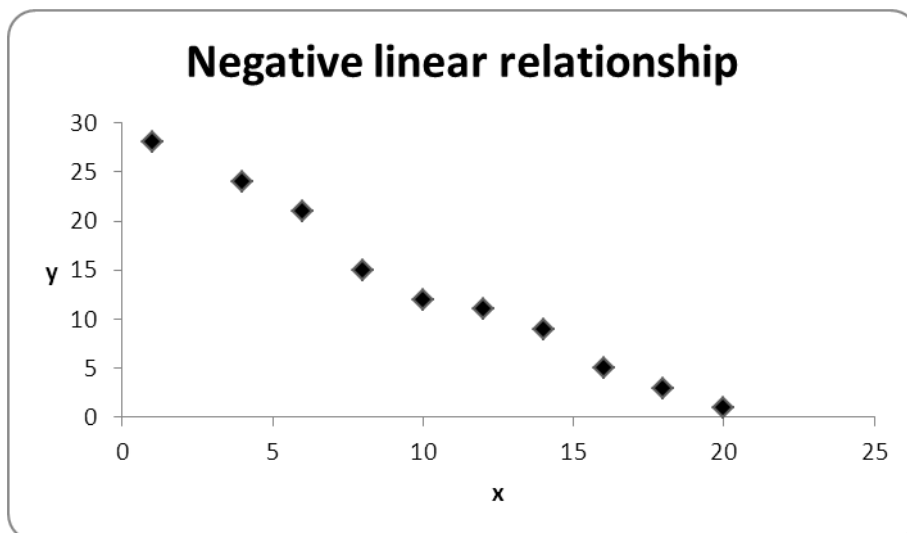


Figure 2

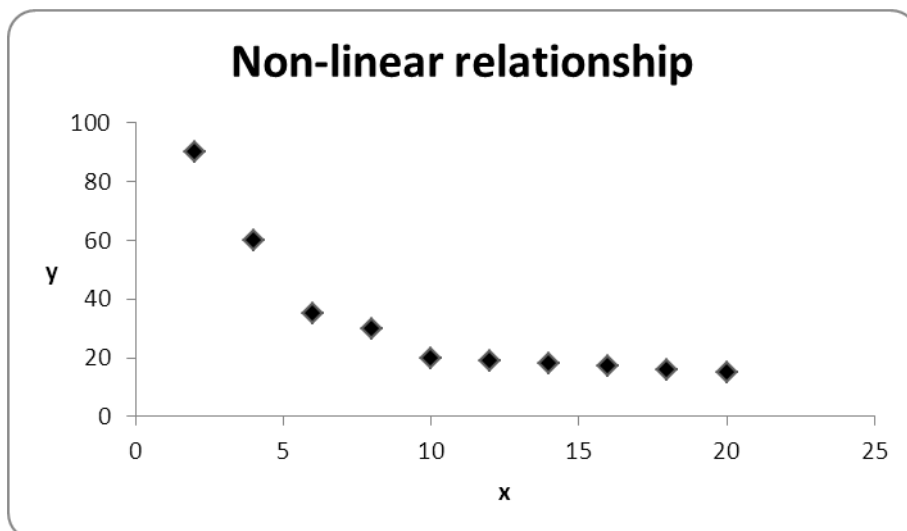


Figure 3

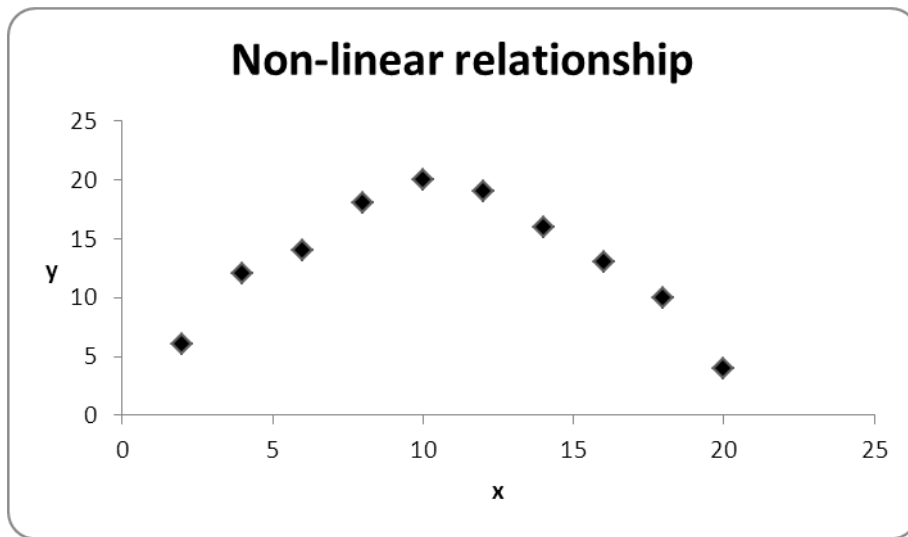


Figure 4

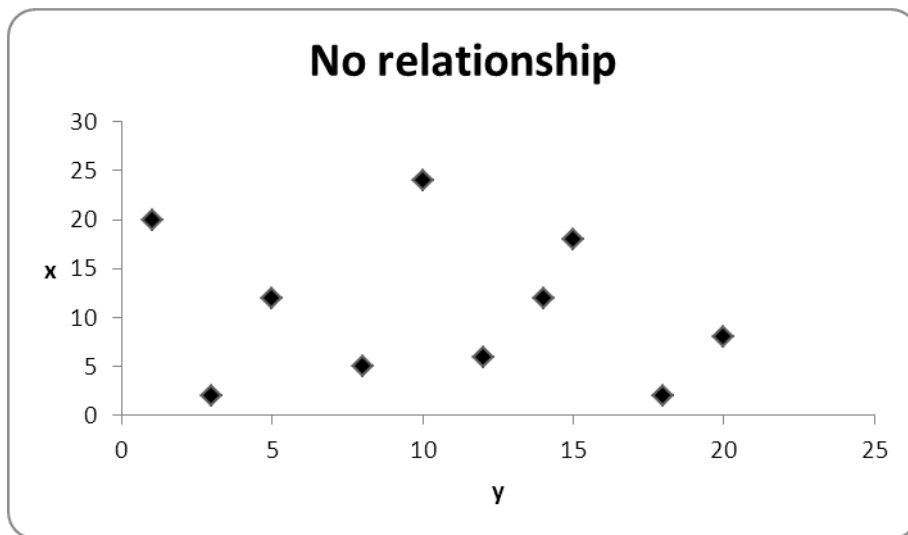


Figure 5

Unit 1 Exercises: (Solutions are found at the end of the module guide)**Exercise 1.1**

Produce a pie chart showing the percentage market share of the passenger car market held by each of South Africa's car manufacturers.

Manufacturer	1991 Sales (units)
Toyota	51 653
Nissan	20 793
Volkswagen	39 757
Delta	20 949
Ford	18 631
Mercedes Benz	15 756
BMW	15 431
MMI	14 731
Total 1991 sales	197 701

Exercise 1.2

Produce a component bar chart showing the breakdown of car sales for Toyota, Nissan and Ford between the first and second half of 1991.

Manufacturer	1991 Sales (units)		
	Total units	First half of 1991	Second half of 1991
Toyota	51 653	19 629	32 024
Nissan	20 793	9 565	11 228
Ford	18 631	9 875	8 756
Totals	91 077	39 069	52 008

Exercise 1.3

Produce a line graph showing the trend in market share for Volkswagen and Nissan from 1982 to 1991.

Year	Volkswagen	Nissan
1982	13,4	9,9
1983	11,6	9,6
1984	9,8	8,2
1985	14,4	6,8
1986	17,4	7,8
1987	19,9	9,7
1988	21,3	11,7
1989	22,2	10,2
1990	19,6	10,6
1991	20,1	10,5

Comment on the findings.

Exercise 1.4

Areas of continents of the world.

Continents	Area in millions of square kilometres
Africa	30,3
Asia	26,3
Europe	4,9
North America	24,3
Oceania	8,5
South America	17,9
Russia	20,5

- a) Draw a bar chart of the above information.
- b) Construct a pie chart to represent the total area.

Exercise 1.5

The distance travelled (in kilometres) by a courier service motorcycle on 30 trips is recorded by the driver.

24	19	21	27	20	17	17	32	22	26
18	13	23	30	10	13	18	22	34	16
18	23	15	19	28	25	25	20	17	15

- a) Define the random variable and the data type.
- b) From the dataset, prepare:
 - An absolute frequency distribution.
 - A relative frequency distribution.
 - A cumulative frequency distribution.
- c) Construct the following graphs:
 - A histogram.
 - An ogive.
 - The relative ogive.
- d) From the graphs, read off:
 - What percentage of trips are between 25 and 30 km long?
 - What percentage of trips are less than 25 km long?
 - What percentage of trips are 22 km or more?
 - Below which distance are 55% of the trips made?
 - Above which distance are 20% of the trips made?

Exercise 1.6

Tourists seeking holiday accommodation in a self-catering complex in the ABC resort in Namibia can make either a one- or a two-week booking. The booking received last season are:

Tourist's home country	Type of booking	
	One-week	Two-week
France	13	44
Germany	29	36
Holland	17	21
Ireland	8	5

- Produce a simple bar chart to show the total number of bookings by home country.
- Produce a component bar chart to show the number of bookings by home country and type of booking.
- Produce a cluster bar chart to show the number of bookings by home country and type of booking.

Exercise 1.7

A roadside breakdown assistance service answer 37 calls in Cape Town on one day. The response times taken to deal with these calls are noted and arranged in a grouped frequency distribution.

Response time (minutes)	Number of calls
20 - 30	4
30 - 40	8
40 - 50	17
50 - 60	6
60 - 70	2

- Produce a histogram to portray this distribution and describe the shape of the distribution.
- Find the cumulative frequency for each interval.
- Produce an ogive of the distribution.

UNIT 2
MEASURES OF CENTRAL TENDENCY

UNIT 2: MEASURES OF CENTRAL TENDENCY

OBJECTIVES

By the end of this study unit, you should be able to:

1. Determine the mean, median and mode for grouped and ungrouped data.
2. Describe the symmetry/skewness of a set of data in terms of the mean, median and mode.
3. Calculate the range, standard deviation, variance, quartiles and inter-quartile range for grouped as well as ungrouped data.

CONTENTS

- 2.1 Introduction
- 2.2 Ungrouped data
 - 2.2.1. Mean
 - 2.2.2. Median
 - 2.2.3. Mode
- 2.3 Grouped data
 - 2.3.1. Mean for grouped data
 - 2.3.2. Median for grouped data
 - 2.3.3. Mode for grouped data
- 2.4 The best measure for central location
- 2.5 Skewness
- 2.6 Non-central location measures – quartiles and percentiles
 - 2.6.1. Quartiles and percentiles for ungrouped data
 - 2.6.2. Quartiles and percentiles for grouped data
- Exercises



READING

Prescribed textbook: Chapter 3

2.1 Introduction

This unit discusses numerical descriptive measures used to summarise and describe sets of data.

There are three commonly used numerical measures of central tendency or central location of a dataset: the **mean**, the **median** and the **mode**. You are expected to know how to compute each of these measures for a given dataset. Moreover, you are expected to know the advantages and disadvantages of each of these measures, as well as the type of data for which each is an appropriate measure.

Data may be ungrouped (sometimes called 'raw' data) or grouped into intervals as covered in the previous unit.

2.2 Ungrouped data

2.2.1 Arithmetic mean

The first and most important measure of central location is the arithmetic mean (average), often just referred to as the mean.

To calculate the mean of ungrouped data we merely add the numbers together and divide the total by the number of values.

Definition: Mean:

The **arithmetic mean** of a dataset is obtained by adding each value in the dataset and dividing the total by the number of variables in the dataset. It is referred to simply as the mean.

Formula: Mean:

$$\bar{x} = \frac{\sum x}{n}$$

Where,

Σ denotes summation of a set of values

x is the variable used to represent raw scores

n represents the number of scores being considered

The result can be denoted by \bar{x} for the mean of a sample from a larger population

The computed mean of all values of a population is denoted by the Greek letter μ (pronounced mu)

EXAMPLE

Find the mean of the dataset:

2	3	6	7	12
---	---	---	---	----

The mean is:

$$\bar{x} = \frac{\sum x}{n} = \frac{2+3+6+7+12}{5} = 6$$

2.2.2 Median

The median is the middle value of an ordered set of numbers.

Note: It is important that the values are in sequential order before you choose the middle value.

Definition: Median:

The median of a dataset is the middle value when the values are arranged in order of increasing (or decreasing) magnitude.

After first arranging the original values in increasing (or decreasing) order, the median will be either of the following:

- If the number of values is odd, the median is the number that is exactly in the middle of the list.
- If the number of values is even, the median is found by computing the mean of the two middle numbers.

EXAMPLE

Over a seven-day period, the number of customers (per day) purchasing at Hides Leather Shop is:

4	80	50	10	60	12	5
---	----	----	----	----	----	---

Array – arranged in order of increasing magnitude:

4	5	10	12	50	60	80
---	---	----	----	----	----	----

The number of values is odd, therefore the median is the middle number of the list:

$$\text{median} = 12$$

EXAMPLE

Over an eight-day period, the number of customers observed at the shop per day is:

21	5	11	7	12	15	20	5
----	---	----	---	----	----	----	---

Array – arranged in order of increasing magnitude:

5	5	7	11	12	15	20	21
---	---	---	----	----	----	----	----

The number of values is even, therefore the median is the mean of the middle two numbers in the list.

$$\text{median} = \frac{11 + 12}{2} = 11,5$$

**SELF-ASSESSMENT ACTIVITY**

The time taken to complete an assembly task has been measured for a group of six employees:

8	2	7	3	6	9
---	---	---	---	---	---

Find the median time taken.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Begin by arranging the scores in increasing order.

2	3	6	7	8		9
---	---	---	---	---	--	---

We note that the numbers 6 and 7 share the middle position thus the median is the average of the 3rd and 4th values.

$$\text{median} = \frac{6 + 7}{2} = 6,5$$

2.2.3 Mode

The mode is the most common value. If we look at the set of numbers:

3	4	5	6	6	6	7
---	---	---	---	---	---	---

The mode is 6 because it is the number that appears most often.

Definition – Mode:

The mode of a dataset is the value that occurs most frequently.

Where no score is repeated there is no mode. Where two scores occur with the same highest frequency, the dataset is **bimodal**. If more than two scores occur with the same highest frequency, each is a mode and the dataset is **multimodal**.

EXAMPLE

The commission earnings of five salespeople are:

R5 000	R5 200	R5 200	R5 700	R8 600
--------	--------	--------	--------	--------

The modal commission is R5 200.

The lengths of stay (in days) for a sample of nine patients in a hospital are:

17	19	19	4	19	26	4	21	4
----	----	----	---	----	----	---	----	---

The dataset is bimodal with two modes, 19 and 4 days.

EXAMPLE

There are 40 buck, 25 elephant and 20 smaller animals at a water hole.

The modal category is buck since it has the highest frequency.

The mode is the only central measure that can be used with data at the **nominal level** of measurement.

EXAMPLE

The hourly income rates (in \$) of five students are:

4	9	7	16	10
---	---	---	----	----

There is no mode.

2.3 Grouped data

Once data is grouped into intervals, the original or raw data is no longer of relevance or may not be known and the frequency distribution data needs to be used for measuring central location.



TIP

Formulae can be presented in different ways. In this text we have wherever possible, used the formulae from the textbook.

Remember if a lecturer uses a formula that looks slightly different, it is up to you as a masters' level student to check that it is still the same formula.

2.3.1 Mean for grouped data

Because the original or raw data is no longer available or of relevance, each dataset observation is assumed to take on the value of the midpoint of its interval. In order to calculate the mean, the total of all values (i.e. midpoint values) is used.

Formula: Mean for grouped data:

$$\text{mean, } \bar{x} = \frac{\sum fx}{n} \text{ or } \bar{x} = \frac{\sum fx}{\sum f}$$

Where,

f is the frequency

x is the midpoint of the interval

n is the number of observations in the dataset

Steps in calculating the mean of grouped data:

Step 1. Extend the frequency distribution to add the further columns needed.

Step 2. Calculate the midpoint of each interval in the frequency distribution and include in a new column.

$$\text{midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

Each observation is then 'allocated' the midpoint as its value.

Step 3. Multiply the frequency of each interval by the midpoint and include in a new column.

$$fx = \text{frequency} \times \text{midpoint}$$

Step 4. The total of this column provides the total of all observations (using their 'allocated' midpoint values).

Step 5. This total is divided by the total number of observations in the dataset to obtain the mean.

EXAMPLE

Using the data from the example in section 1.3.1, calculate the mean of the grouped data:

Interval (weight in lbs)	Frequency
140 – 150	1
150 – 160	4
160 – 170	8
170 – 180	7
180 – 190	5

Step 1. Extend the frequency distribution to add the further columns needed.

Interval (weight in lbs)	Frequency	Midpoint x	fx
140 – 150	1		
150 – 160	4		
160 – 170	8		
170 – 180	7		
180 – 190	5		
Total			

Step 2. Calculate the midpoint of each interval in the frequency distribution and include in a new column.

$$\text{midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2} = \frac{140 + 150}{2} = 145 \text{ etc}$$

Each observation is then 'allocated' the midpoint as its value.

Interval (weight in lbs)	Frequency	Midpoint x	fx
140 – 150	1	145	
150 – 160	4	155	
160 – 170	8	165	
170 – 180	7	175	
180 – 190	5	185	
Total			

Step 3. Multiply the frequency of each interval by the midpoint and include in a new column.

$$fx = \text{frequency} \times \text{midpoint} = 1 \times 145 = 145 \text{ etc}$$

Interval (weight in lbs)	Frequency	Midpoint x	fx
140 – 150	1	145	145
150 – 160	4	155	620
160 – 170	8	165	1 320
170 – 180	7	175	1 225
180 – 190	5	185	925
Total	25		4 235

Step 4. The total of this column provides the total of all observations (using their 'allocated' midpoint values).

Step 5. This total is divided by the total number of observations in the dataset to obtain the mean.

$$\text{mean, } \bar{x} = \frac{\sum fx}{n} = \frac{4\,235}{25} = 169,4$$



SELF-ASSESSMENT ACTIVITY

The number of times per week that a particular photocopy machine breaks down is recorded over a period of 60 weeks.

Number of breakdowns	0	1	2	3	4	5
Number of weeks	15	12	16	10	5	2

Find the mean number of breakdowns per week over the 60-week period.

SOLUTION TO SELF-ASSESSMENT ACTIVITY



TIP

If the value against the frequency is not an interval, as in this case, the actual value is used and a midpoint does not need to be calculated.

Frequency (number of breakdowns)	Value x	fx
0	15	0
1	12	12
2	16	32
3	10	30
4	5	20
5	2	10
Total	60	104

$$\text{mean, } \bar{x} = \frac{\sum fx}{\sum f} = \frac{104}{60} = 1,73$$

EXAMPLE

The times taken to complete a particular assembling task have been measured for 250 employees.

Time (minutes)	Number of people (f)	x	fx
0 - 5	2	2,5	5,0
5 - 10	2	7,5	15,0
10 - 15	3	12,5	37,5
15 - 20	5	17,5	87,5
20 - 25	5	22,5	112,5
25 - 30	18	27,5	495,5
30 - 35	85	32,5	2 762,5
35 - 40	92	37,5	3 450,0
40 - 45	37	42,5	1 572,5
45 - 50	1	47,5	47,5
Total	250		8 585,0

$$\text{mean, } \bar{x} = \frac{\sum fx}{\sum f} = \frac{8\,585}{250} = 34,34 \text{ minutes}$$

**SELF-ASSESSMENT ACTIVITY**

The durations of 100 machine breakdowns are recorded and summarised. Find the mean of the distribution.

Time (minutes)	Frequency
0 – 10	3
10 – 20	13
20 – 30	30
30 – 40	25
40 – 50	14
50 – 60	8
60 – 70	4
70 – 80	2
80 – 90	1
Total	100

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Time (minutes)	Frequency	x	fx
0 – 10	3	5	15
10 – 20	13	15	195
20 – 30	30	25	750
30 – 40	25	35	875
40 – 50	14	45	630
50 – 60	8	55	440
60 – 70	4	65	260
70 – 80	2	75	150
80 – 90	1	85	85
Total	100		3 400

$$\text{mean, } \bar{x} = \frac{\sum fx}{\sum f} = \frac{3\,400}{100} = 34 \text{ minutes}$$

2.3.2 Median for grouped data

The cumulative frequency is used to find the median for grouped data.

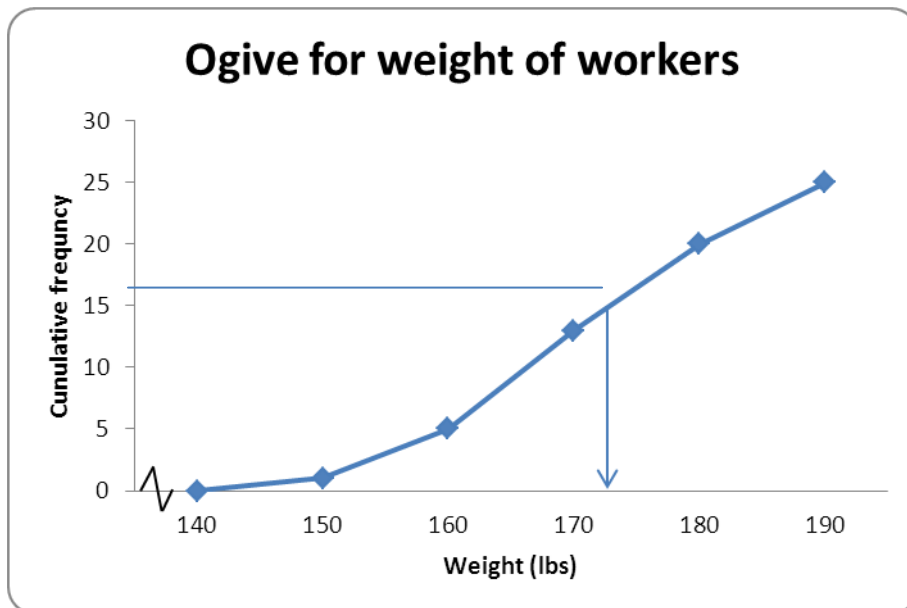
Graphical approach: Median for grouped data:

The ogive is used to determine the median value by reading off the x-value associated with the 50% cumulative frequency on the y-axis.

EXAMPLE

Using the data from the example in section 1.3.1:

$$\text{median frequency} = \frac{n}{2} = \frac{30}{2} = 15$$



The median value reading from the ogive is 173 lbs.

Formula: Median for grouped data:

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

Where,

O_{me} is the lower limit of the median interval

c is the interval width

n is the number of observations in the dataset

$f(<)$ is the cumulative frequency count of all intervals before the median interval

f_{me} is the frequency count of the median interval

The formula uses the median interval and calculates how far into the median interval, the median value lies.

Steps in calculating the median of grouped data:

Step 1. Extend the frequency distribution to be a cumulative frequency distribution.

Step 2. Establish the median interval from the cumulative frequency. Establish at which point the cumulative frequency exceeds the median point for the first time. This is the median interval.

Step 3. Substitute the required values into the formula and calculate the median.

EXAMPLE

Using the data from the example in section 1.3.1:

Step 1. Extend the frequency distribution to be a cumulative frequency distribution.

Interval (weight in lbs)	Frequency	Cumulative frequency
140 – 150	1	1
150 – 160	4	5
160 – 170	8	13
170 – 180	7	20
180 – 190	5	25

Step 2. Establish the median interval from the cumulative frequency. The cumulative frequency for the interval 170 – 180 is 20; this is the point at which the cumulative frequency exceeds the median frequency of 15 for the first time.

$$\text{median frequency} = \frac{n}{2} = \frac{30}{2} = 15$$

Therefore *median interval* = 170 – 180

Step 3. Substitute the required values into the formula and calculate the median.

$$M_{\varepsilon} = O_{m\varepsilon} + \frac{c \left[\frac{n}{2} - f(<) \right]}{f_{m\varepsilon}}$$

$$M_{\varepsilon} = 170 + \frac{10 \left[\frac{30}{2} - 13 \right]}{7} = 172,86 \text{ lbs}$$



TIP

Sense test that the calculated median falls within the median interval.

As with the mean, the value for **the median f ungrouped data is more accurate**. If the data is available (e.g. when you do your research project) it is better to use the ungrouped data to get the median.

**SELF-ASSESSMENT ACTIVITY**

The time taken to complete an assembling task has been measured for 250 employees :

Time taken (minutes)	Number of people (f)	Cumulative frequency $f(<)$
0 – 5	2	2
5 – 10	2	4
10 – 15	3	7
15 – 20	5	12
20 – 25	5	17
25 – 30	18	35
30 – 35	85	120
35 – 40	92	212
40 – 45	37	249
45 – 50	1	250
Total	250	

What is the position of the median?

SOLUTION TO SELF-ASSESSMENT ACTIVITY

$$\text{median frequency} = \frac{n}{2} = \frac{250}{2} = 125$$

Therefore *median interval* = 35 – 40

This is the point at which the cumulative frequency (212) exceeds the median frequency (125) for the first time.

$$M_s = O_{m\epsilon} + \frac{c \left[\frac{n}{2} - f(<) \right]}{f_{m\epsilon}}$$

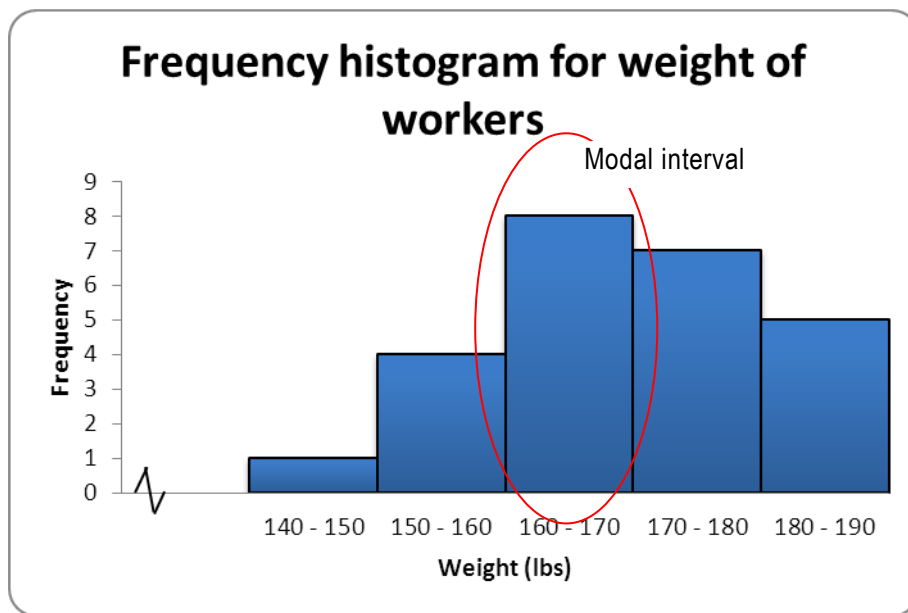
$$M_s = 35 + \frac{5 \left[\frac{250}{2} - 120 \right]}{92} = 35,27 \text{ minutes}$$

2.3.3 Mode for grouped data

The mode is the most common value. In the case of grouped data, the mode is the maximum value of the histogram.

EXAMPLE

Using the data from the example in section 1.3.1:



Formula: Mode for grouped data:

$$M_o = O_{mo} + \frac{c(f_m - f_{m-1})}{2f_m - f_{m-1} - f_{m+1}}$$

Where,

O_{mo} is the lower limit of the modal interval

c is the interval width

f_m is the frequency of the modal interval

f_{m-1} is the frequency of the interval immediately preceding the modal interval

f_{m+1} is the frequency of the interval immediately following the modal interval

This formula uses interpolation to *pull* the modal value within the modal interval towards the interval with the highest frequency.

The modal interval is the interval with the highest frequency.

EXAMPLE

Using the data from the example in section 1.3.1:

Interval (weight in lbs)	Frequency
140 – 150	1
150 – 160	4
160 – 170	8
170 – 180	7
180 – 190	5

The modal interval is 160 – 170 with the highest frequency of 8.

$$M_o = O_{mo} + \frac{c(f_m - f_{m-1})}{2f_m - f_{m-1} - f_{m+1}}$$

$$M_o = 160 + \frac{10(8 - 4)}{2 \times 8 - 4 - 7} = 168 \text{ lbs}$$

Unlike the median and the mean, the value for the mode is more accurate from grouped data. So **whenever possible calculate the mode from the grouped data.**

Calculation of the mode from a grouped frequency distribution

It is not possible to calculate the exact value of the mode of the original data in a grouped frequency distribution, since information is lost when the data are grouped. However, it is possible to make an estimate of the mode. The interval with the highest frequency is called the modal interval.



SELF-ASSESSMENT ACTIVITY

The durations of 100 machine breakdowns are recorded and summarised. Find the mode of the distribution.

Time (minutes)	Frequency
0 – 10	3
10 – 20	13
20 – 30	30
30 – 40	25
40 – 50	14
50 – 60	8
60 – 70	4
70 – 80	2
80 – 90	1
Total	100

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Time (minutes)	Frequency
0 – 10	3
10 – 20	13
20 – 30	30
30 – 40	25
40 – 50	14
50 – 60	8
60 – 70	4
70 – 80	2
80 – 90	1
Total	100

The interval having the highest frequency of 30 is the interval 20 – 30.

$$M_o = O_{mo} + \frac{c(f_m - f_{m-1})}{2f_m - f_{m-1} - f_{m+1}}$$

$$M_o = O_{mo} + \frac{c(f_m - f_{m-1})}{2f_m - f_{m-1} - f_{m+1}} = 20 + \frac{10(30 - 13)}{2 \times 30 - 13 - 25} = 27,73 \text{ minutes}$$

2.4 The best measure for central location

The different measures of central location have different advantages and disadvantages and there are no objective criteria to determine the most representative average of all datasets. Each researcher has to use his/her own discretion on a set of data.

Measure	Usage	Advantages	Disadvantages
Mean	Most familiar average.	Exists for each dataset. Takes every score into account. Works well with many statistical methods.	Is affected by extreme scores.
Median	Commonly used.	Always exists. Is not affected by extreme scores. Is often a good choice if there are some extreme scores in the dataset.	Does not take every score into account.
Mode	Sometimes used.	Is not affected by extreme scores. Is appropriate for data at the nominal level.	It might not exist or there may be more than one mode. It does not take every score into account.

The best measure for central location

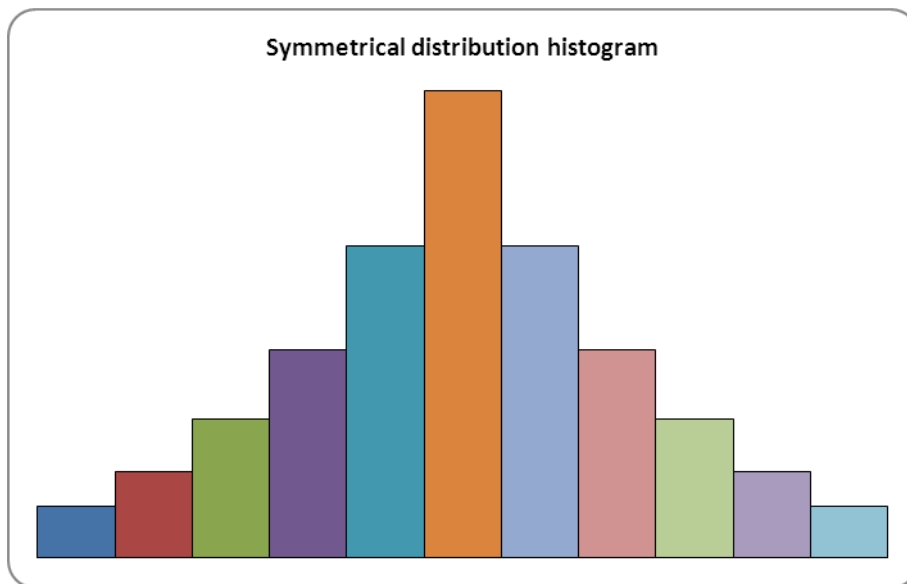
The arithmetic mean is more affected by extreme values. If the data have some values that are very large or small (relative to the other values) then it is better to use the median. When we get to the normal distribution in a later unit, you will see why the arithmetic mean is important.

2.5 Skewness

If there are large extreme values in the data the mean is pulled to the right or left and we say that the distribution exhibits skewness or kurtosis.

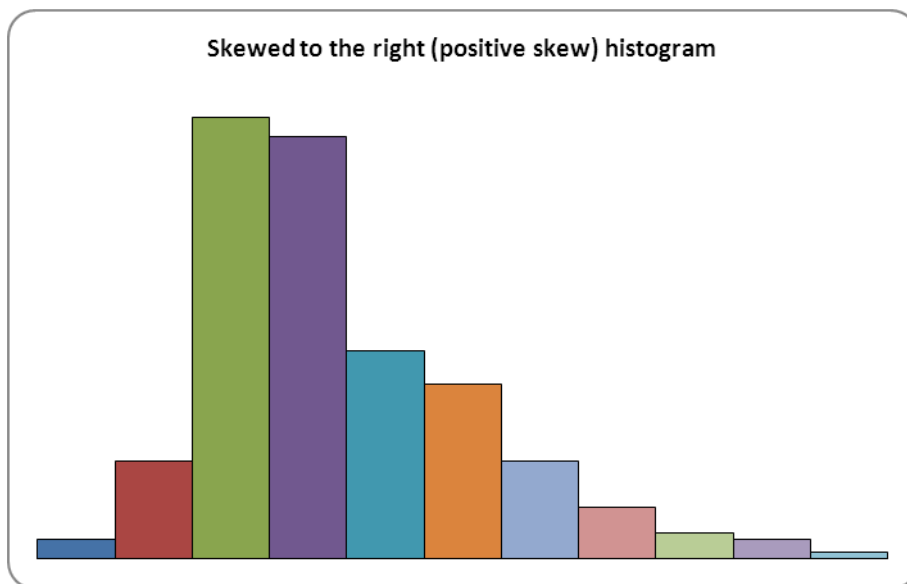
For a **symmetrical distribution** or **normal distribution** the mean, median and mode will be about the same.

$$\text{mean} \approx \text{median} \approx \text{mode}$$



For a distribution that is **skewed to the right** the mode will be less than the median and the median will be less than the mean.

$$mode < median < mean$$



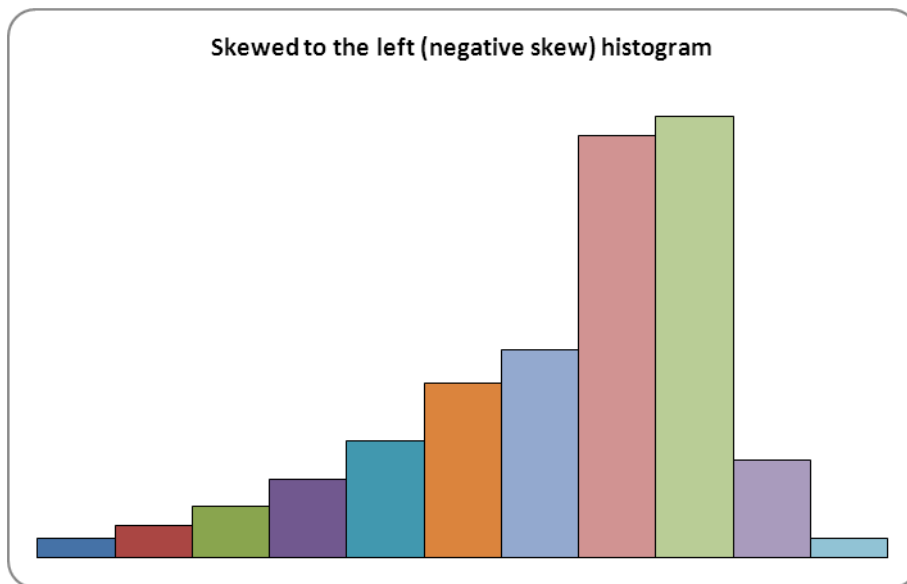
For a distribution that is **skewed to the left** the mean is the smallest, followed by the median, while the mode is the largest.

$$mean < median < mode$$



TIP

A negatively skewed distribution (skewed to the left) has the mean, median and mode in alphabetical order.



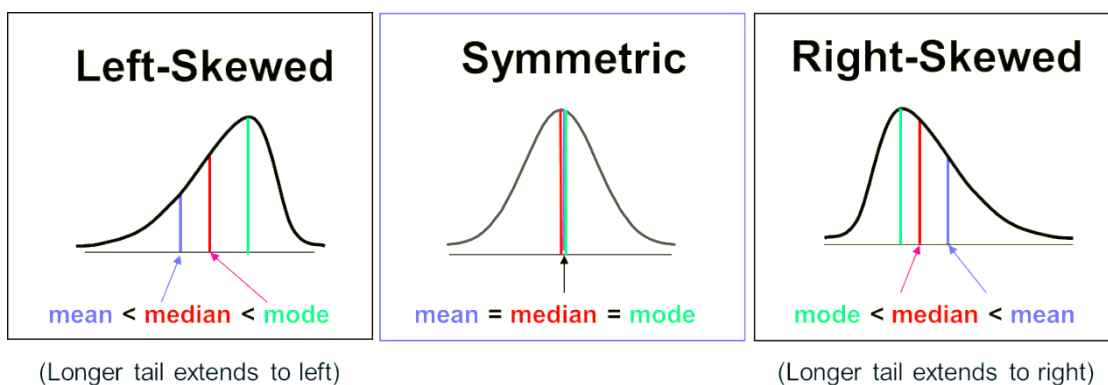
As a general rule the difference between the median and the mode is about twice the difference between the mean and the median.

If the data are skewed to the left there are some outliers on the left (small values). If the data are skewed to the right then there are some large outliers.

Revision:

For a dataset that is approximately *symmetrical* with one mode, the *mean, median and mode tend to have about the same value*. For a dataset that is obviously asymmetrical, it is preferable to report both the mean and median. The mean is relatively reliable; that is, when samples are drawn from the same population, the sample means tend to be more consistent than other averages.

A comparison of the mean and median can reveal information about skewness. Data can be identified as skewed to the left, symmetrical or skewed to the right. Data skewed to the left will have the mean and median to the left of the mode:



Source: Groebner et al (2011)

Which measure to use?

If the data are qualitative, the only appropriate measure of central location is the mode.

If the data are ranked, the most appropriate measure of central location is the median.

For quantitative data, however, it is possible to compute all three measures.

Which measure you should use depends on your objective. The mean is most popular because it is easy to compute and to interpret (in particular, the mean is generally the best measure of central location for purposes of statistical inference, as you will see in later units). It has the disadvantage that it may be unduly influenced by a few very small or very large observations. To avoid this influence, you might choose to use the median. It could be that the data consists, e.g., of salaries or house prices. The mode, representing the value occurring most frequently (or the midpoint of the interval with the largest frequency) should be used when the objective is to find the value (such as shirt size or house price) that is most popular with consumers.

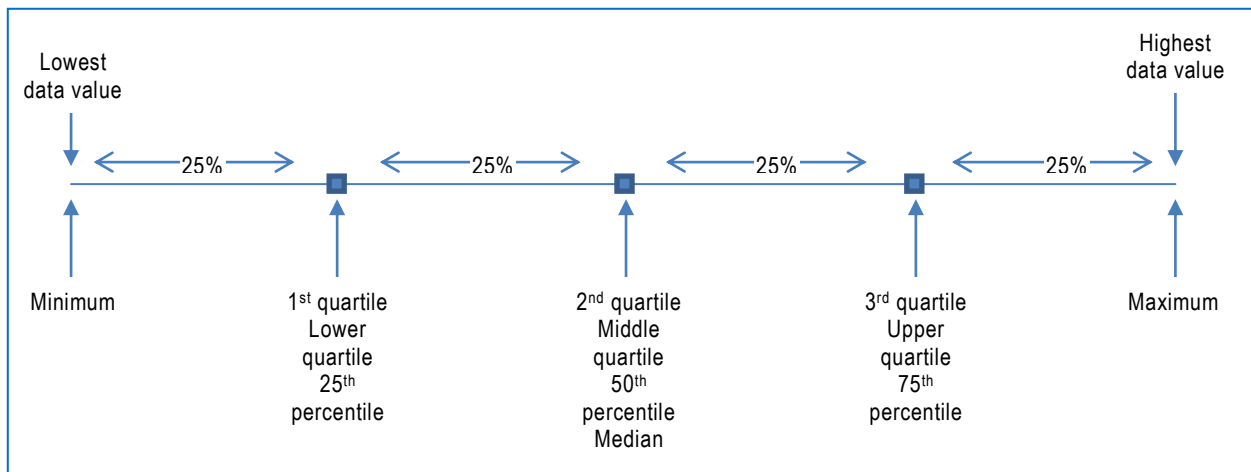
2.6 Non-central location measures – quartiles and percentiles

Non-central location measures or measures of relative position include quartiles and percentiles.

Quartiles are measures that divide the dataset into quarters.

Percentiles indicate specific percent positions of data values within the dataset.

In order to pinpoint quartiles or percentiles, the data need to be **ordered in sequence** from lowest to highest rank.



2.6.1 Quartiles and percentiles for ungrouped data

For ungrouped data, quartiles and percentiles are found by calculating the position. This is similar to the calculations used for determining the position of the median for ungrouped data.

Formula: Quartiles and percentiles for ungrouped data:

$$\text{quartile position} = \frac{q(n+1)}{4} \text{ (round down)}$$

$$\text{percentile position} = \frac{p(n+1)}{100} \text{ (round down)}$$

Where,

q is the required quartile, i.e. 1,2 or 3

n is the sample size

p is the required percentile, i.e. 45 for 45% or 70 for 70%

Note: These formulae will suffice for the purposes of this discussion and the required level of understanding. More precise values for continuous data can be calculated by interpolating the fractional values rather than rounding down. Your textbook describes this process in detail with an example.

EXAMPLE

40 pay TV subscribers have rated their experience (out of 100):

42	43	44	45	45	47	50	51
51	51	51	52	53	54	54	56
57	58	58	59	62	62	62	63
63	64	65	67	67	69	69	72
73	74	75	78	78	83	84	87

Calculate all quartiles as well as the 60th and 70th percentiles of the dataset.

The data need to be ranked in order to determine these values. This dataset has already been sequenced.

The positions for the required measures are then calculated, after which the position is pinpointed and the value determined:

$$\text{first quartile position} = \frac{1(n+1)}{4} = \frac{1(40+1)}{4} = 10,25 = 10 \text{ (rounded)}$$

$$\text{value} = 51 \text{ rating}$$

$$\text{second quartile position} = \frac{2(n+1)}{4} = \frac{2(40+1)}{4} = 20,5 = 20 \text{ (rounded)}$$

$$\text{value} = 59 \text{ rating}$$

$$\text{third quartile position} = \frac{3(n+1)}{4} = \frac{3(40+1)}{4} = 30,75 = 30 \text{ (rounded)}$$

$$\text{value} = 69 \text{ rating}$$

$$\text{60th percentile position} = \frac{60(n+1)}{100} = \frac{60(40+1)}{100} = 24,6 = 24 \text{ (rounded)}$$

value = 63 rating

$$70\text{th percentile position} = \frac{70(n+1)}{100} = \frac{70(40+1)}{100} = 28,7 = 28 \text{ (rounded)}$$

value = 67 rating

42	43	44	45	45	47	50	51
51	Q1 51	51	52	53	54	54	56
57	58	58	Q2 59	62	62	62	P60 63
63	64	65	P70 67	67	Q3 69	69	72
73	74	75	78	78	83	84	87

2.6.2 Quartiles and percentiles for grouped data

For grouped data formulae similar to that of the median are used to calculate quartiles and percentiles.

Formula: Quartiles for grouped data:

$$Q_j = O_{Q_j} + \frac{c \left[\frac{j \times n}{4} - f(<) \right]}{f_{Q_j}}$$

Where,

j is the required quartile, i.e. 1,2 or 3

O_{Q_j} is the lower limit of the quartile interval

c is the interval width

n is the number of observations in the dataset

$f(<)$ is the cumulative frequency count of all intervals before the quartile interval

f_{Q_j} is the frequency count of the quartile interval

The formula uses the quartile interval and calculates how far into the interval the quartile value lies.

It is important to first calculate the quartile position in order to be able to establish the quartile interval:

$$\text{quartile position} = \frac{j \times n}{4}$$

Formula: Percentiles for grouped data:

$$P_j = O_{P_j} + \frac{c \left[\frac{j \times n}{100} - f(<) \right]}{f_{P_j}}$$

Where,

j is the required percentile, i.e. 45%, 70% etc.

O_{P_j} is the lower limit of the percentile interval

c is the interval width

n is the number of observations in the dataset

$f(<)$ is the cumulative frequency count of all intervals before the percentile interval

f_{pj} is the frequency count of the percentile interval

The formula uses the percentile interval and calculates how far into the percentile interval, the percentile value lies.

It is important to first calculate the percentile position in order to be able to establish the percentile interval:

$$\text{percentile position} = \frac{j \times n}{100}$$

EXAMPLE

Using the data from the example in section 1.3.1 calculate all quartiles and the 60th and 70th percentiles:

Ensure you have a cumulative frequency distribution.

Interval (weight in lbs)	Frequency	Cumulative frequency
140 – 150	1	1
150 – 160	4	5
160 – 170	8	13
170 – 180	7	20
180 – 190	5	25

Calculate the required positions within the dataset:

$$\text{first quartile position} = \frac{j \times n}{4} = \frac{1 \times 25}{4} = 6,25$$

$$\text{second quartile position} = \frac{j \times n}{4} = \frac{2 \times 25}{4} = 12,5$$

$$\text{third quartile position} = \frac{j \times n}{4} = \frac{3 \times 25}{4} = 18,75$$

$$\text{60th percentile position} = \frac{j \times n}{100} = \frac{60 \times 25}{100} = 15$$

$$\text{70th percentile position} = \frac{j \times n}{100} = \frac{70 \times 25}{100} = 17,5$$

Calculate the values associated with the positions.

$$Q_1 = O_{Q_1} + \frac{c \left[\frac{1 \times n}{4} - f(<) \right]}{f_{Q_1}} = 160 + \frac{10(6,25 - 5)}{8} = 161,56 \text{ lbs}$$

$$Q_2 = O_{Q_2} + \frac{c \left[\frac{2 \times n}{4} - f(<) \right]}{f_{Q_2}} = 160 + \frac{10(12,5 - 5)}{8} = 169,38 \text{ lbs}$$

$$Q_3 = O_{Q_3} + \frac{c \left[\frac{3 \times n}{4} - f(<) \right]}{f_{Q_3}} = 170 + \frac{10(18,75 - 13)}{7} = 178,21 \text{ lbs}$$

$$P_{60} = O_{P_{60}} + \frac{c \left[\frac{60 \times n}{100} - f(<) \right]}{f_{P_{60}}} = 170 + \frac{10(15 - 13)}{7} = 172,86 \text{ lbs}$$

$$P_{70} = O_{P_{70}} + \frac{c \left[\frac{70 \times n}{100} - f(<) \right]}{f_{P_{70}}} = 170 + \frac{10(17,5 - 13)}{7} = 176,43 \text{ lbs}$$

**TIP**

Sense test that the calculated values fall within the associated intervals.

**SELF-ASSESSMENT ACTIVITY**

The time taken to complete an assembling task has been measured for 250 employees:

Time taken (minutes)	Number of people (f)	Cumulative frequency $f(<)$
0 – 5	2	2
5 – 10	2	4
10 – 15	3	7
15 – 20	5	12
20 – 25	5	17
25 – 30	18	35
30 – 35	85	120
35 – 40	92	212
40 – 45	37	249
45 – 50	1	250
Total	250	

Calculate the first and third quartiles and the 40th percentile.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Calculate the required positions within the dataset:

$$\text{first quartile position} = \frac{j \times n}{4} = \frac{1 \times 250}{4} = 62,5$$

$$\text{third quartile position} = \frac{j \times n}{4} = \frac{3 \times 250}{4} = 187,5$$

$$\text{40th percentile position} = \frac{j \times n}{100} = \frac{40 \times 250}{100} = 100$$

Calculate the values associated with the positions.

$$Q_1 = O_{Q_1} + \frac{c \left[\frac{1 \times n}{4} - f(<) \right]}{f_{Q_1}} = 30 + \frac{5(62,5 - 35)}{85} = 31,62 \text{ minutes}$$

$$Q_3 = O_{Q_3} + \frac{c \left[\frac{3 \times n}{4} - f(<) \right]}{f_{Q_3}} = 35 + \frac{5(187,5 - 120)}{92} = 38,67 \text{ minutes}$$

$$P_{40} = O_{P_{40}} + \frac{c \left[\frac{40 \times n}{100} - f(<) \right]}{f_{P_{40}}} = 30 + \frac{5(100 - 35)}{85} = 33,82 \text{ minutes}$$

Unit 2 Exercises: (Solutions are found at the end of the module guide)

Exercise 2.1

A supermarket sells kilogram-bags of pears. The numbers of pears in 21 bags are:

7	9	8	8	10	9	8	10	10	8	9
10	7	9	9	9	7	8	7	8	9	

- Find the mode, median and mean for these data.
- Compare each of the measures and comment on the likely shape of the distribution.
- Plot a simple bar chart to portray the data.

Exercise 2.2

The number of credit cards carried by 25 shoppers is:

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

- Determine the mode and median of this distribution.
- Calculate the mean of the distribution and compare it to the mode and median.
What can you conclude about the shape of the distribution?
- Draw a bar chart to represent the distribution and confirm your conclusions in (b).

Exercise 2.3

A supermarket has one checkout counter for customers who wish to purchase 10 items or less.

The numbers of items presented at this checkout by 19 customers are:

10	8	7	7	6	11	10	8	9	9
9	6	10	9	8	9	10	10	10	

- Find the mode, median and mean for these data.
- What do your results for (a) tell you about the shape of the distribution?
- Plot a simple bar chart to portray the distribution.

Exercise 2.4

The numbers of driving tests attempted before passing by 28 clients of a driving school are:

Tests taken	Number of clients
1	10
2	8
3	4
4	3
5	3

- Obtain the mode, median and mean from this frequency distribution and compare their value.
- Plot a simple bar chart of the distribution.

Exercise 2.5

Spina Software Solutions operates an on-line help and advice service for PC owners. The numbers of calls made to them by subscribers in a month are:

Number of subscribers		
Calls made	Female	Male
1	31	47
2	44	42
3	19	24
4	6	15
5	1	4

Find the mode, median and mean for both distributions and use them to compare the two distributions.

Exercise 2.6

Toofley the chemists own 29 pharmacies. The number of packets of a new skin medication sold in each of their shops in a week is:

7	22	17	13	11	20	15	18	5	22
6	18	10	13	33	13	9	8	9	19
19	8	12	12	21	20	12	13	22	

- Find the mode and range of the data.
- Identify the median of the data.
- Find the lower and upper quartile values.

Exercise 2.7

The kilocalories per portion in a sample of 32 different breakfast cereals are recorded and collated:

Kcal per portion	Frequency
80 – 120	3
120 – 160	11
160 – 200	9
200 – 240	7
240 - 280	2

Obtain an approximate value for the median of the distribution.

UNIT 3
MEASURES OF DISPERSION (VARIABILITY)

UNIT 3: MEASURES OF DISPERSION (VARIABILITY)

OBJECTIVES

By the end of this study unit, you should be able to:

1. Define the various measures of dispersion.
2. Compute each dispersion measure for both grouped and ungrouped sets of data.
3. Interpret each measure of dispersion.

CONTENT

3.1 Introduction

3.2 Range

3.3 Variance

3.4 Standard deviation

3.5 Coefficient of variation

3.6 Interquartile and interpercentile ranges

3.7 Quartile deviation

Exercises



Prescribed textbook: Chapter 3

3.1 Introduction

“Dispersion (or spread) refers to the extent to which the data values of a numeric random variable are scattered about their central location value” Wegner (2012).

In unit 2, the concept of central location was introduced. The variability among data is one characteristic to which averages are not sensitive. It is possible to have two datasets with identical measures of central location but with wider spreads of data.

EXAMPLE

Consider two groups of data:

Dataset A	Dataset B
65	42
66	54
67	58
68	62
71	67
73	77
74	77
77	85
77	93
77	100
Computed measures of central location	
Mean = 71,5	Mean = 71,5
Median = 72	Median = 72
Mode = 77	Mode = 77

Although there is no difference in the computed central measures between the two groups, the scores of dataset B are much more widely scattered than those of dataset A.

The measures that are used to measure dispersion are:

- Range
- Variance
- Standard deviation
- Interquartile range
- Quartile deviation

3.2 Range

The range measures the difference between the highest and lowest values in a dataset. It is considered a rough measure of spread as it depends on only two values. It is affected by outliers and gives no indication of the clustering of the data.

Formula: Range for ungrouped data:

$$\text{range} = \text{highest value} - \text{lowest value}$$

EXAMPLE

For the data in a previous example:

Dataset A	Dataset B
$\text{range} = 77 - 45 = 32$	$\text{range} = 100 - 42 = 58$

The ranges indicate that the data in dataset B are more widely spread than that in dataset A.

Formula: Range for grouped data:

$$\text{range} = \text{upper limit of highest interval} - \text{lower limit of lowest interval}$$



SELF-ASSESSMENT ACTIVITY

The merchandising manager for a retail clothing chain has recorded 30 observations on the number of days between re-orders for a particular range of woman's clothing.

18	26	15	17	7	27	24	17	10	17
23	29	28	18	10	23	16	9	12	26
5	12	23	22	24	14	16	26	19	22

Find the range of the number of days between re-orders.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

$$\text{range} = \text{highest value} - \text{lowest value} + 1 = 29 - 5 + 1 = 24 \text{ days}$$

Interpretation

24 days separates the shortest time between successive re-orders from the longest time between successive re-orders for a particular range of women's clothing.

3.3 Variance

The variance (s^2) measure the average squared deviation from the mean for a dataset.

Formula: Variance for ungrouped data:

$$\text{variance } (s^2) = \frac{\sum(x - \bar{x})^2}{n - 1}$$

or

$$\text{variance } (s^2) = \frac{\sum x^2 - n\bar{x}^2}{n - 1}$$

where,

x is each value of the dataset

\bar{x} is the mean of the dataset

n is the sample size

For grouped data, the original dataset values are changed to the interval midpoints.

Formula: Variance for grouped data:

$$\text{variance } (s^2) = \frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n - 1}$$

where,

f is the interval frequency

x is the interval midpoint

n is the sample size



TIP

Whenever you see the \sum sign in an equation, you will need a column in the table for the expression immediately following the \sum sign. Consider also having columns for each of the components of the expression. This is illustrated in the next example.

EXAMPLE

Calculate the variance of the sample scores: 2, 3, 5, 6, 9, 17.

Both variance formulae are used in this example, with all the necessary table columns included for both formulae.

First it is necessary to calculate the mean:

$$\text{mean, } \bar{x} = \frac{\sum x}{n} = \frac{42}{6} = 7$$

$$\bar{x}^2 = 49$$

x	$x - \bar{x}$	$(x - \bar{x})^2$	x^2
2	-5	25	4
3	-4	16	9
5	-2	4	25
6	-1	1	36
9	2	4	81
17	10	100	289
\sum 42	0	150	444

$$\text{variance } (s^2) = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{150}{6 - 1} = 30$$

or

$$\text{variance } (s^2) = \frac{\sum x^2 - n\bar{x}^2}{n - 1} = \frac{444 - 6 \times 49}{6 - 1} = \frac{150}{5} = 30$$

3.4 Standard deviation

The standard deviation is the square root of the variance. It offers a measure of the average deviation from the mean.

Formula: Standard deviation for ungrouped data:

$$\text{standard deviation, } s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

or

$$\text{standard deviation, } s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n - 1}}$$

where,

x is each value of the dataset

\bar{x} is the mean of the dataset

EXAMPLE

Find the standard deviation of the sample scores in a previous example.

$$s = \sqrt{s^2} = \sqrt{30} = 5,48$$

For grouped data, the original dataset values have been changed to the interval midpoints.

Formula: Standard deviation for grouped data:

$$\text{standard deviation, } s = \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}}$$

where,

f is the interval frequency

x is the interval midpoint

n is the sample size

Notation:

s = standard deviation of a set of sample scores.

σ = standard deviation of a set of population scores.

s^2 = variance of a set of sample scores.

σ^2 = variance of a set of population scores.

Note: Articles in professional journals and reports often use SD for standard deviation and Var for variance.

EXAMPLE

Using the data from the example in section 1.3.1 and including the additional columns required for the standard deviation formula:

Interval (weight in lbs)	Frequency, f	Midpoint, x	x^2	fx	fx^2
140 – 150	1	145	21 025	145	21 025
150 – 160	4	155	24 025	620	96 100
160 – 170	8	165	27 225	1 320	217 800
170 – 180	7	175	30 625	1 225	214 375
180 – 190	5	185	34 225	925	171 125
Σ	25	825	137 125	4 235	720 425

$$\begin{aligned} \text{standard deviation, } s &= \sqrt{\frac{\sum fx^2 - \frac{(\sum fx)^2}{n}}{n-1}} = \sqrt{\frac{720\,425 - \frac{4\,235^2}{25}}{25-1}} = \sqrt{\frac{720\,425 - 717\,409}{24}} \\ &= \sqrt{125,67} = 11,21 \text{ lbs} \end{aligned}$$

**SELF-ASSESSMENT ACTIVITY**

The errors in seven invoices are recorded as follows: 120, 30, 40, 8, 5, 20, 29
Calculate the standard deviation.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

$$\text{mean, } \bar{x} = \frac{\sum x}{n} = \frac{252}{7} = 36$$

Number of errors, x	$x - \bar{x}$	$(x - \bar{x})^2$
120	84	7 056
30	-6	36
40	4	16
8	-28	784
5	-31	961
20	-16	256
29	-7	49
Σ 252	0	9 158

$$\text{standard deviation, } s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{9\,158}{7 - 1}} = 39,07 \text{ errors}$$

**SELF-ASSESSMENT ACTIVITY**

The time (in hours per week) that 50 office staff members spend using personal computers are:

Time (hours per week)	Frequency, f
0 – 3	14
3 – 6	6
6 – 9	6
9 – 12	7
12 – 15	14
15 – 18	3
Σ	50

Calculate the standard deviation.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Interval (time in hours per week)	Frequency, f	Midpoint, x	x^2	fx	fx^2
0 – 3	14	1,5	2,25	21,00	31,50
3 – 6	6	4,5	20,25	27,00	121,50
6 – 9	6	7,5	56,25	45,00	337,50
9 – 12	7	10,5	110,25	73,50	771,75
12 – 15	14	13,5	182,25	189,00	2 551,50
15 – 18	3	16,5	272,25	49,50	816,75
Σ	50	54,0	643,50	405,00	4 630,50

$$\begin{aligned} \text{standard deviation, } s &= \sqrt{\frac{\Sigma fx^2 - \frac{(\Sigma fx)^2}{n}}{n-1}} = \sqrt{\frac{4\,630,5 - \frac{405^2}{50}}{50-1}} = \sqrt{\frac{4\,630,5 - 3\,280,5}{49}} \\ &= \sqrt{27,55} = 5,25 \text{ hours} \end{aligned}$$

3.5 Coefficient of variation

The coefficient of variation offers a measure of the dispersion relative to the mean. This enables comparison between datasets with different means.

Formula: Coefficient of variation:

$$\text{coefficient of variation, } CV = \frac{s}{\bar{x}} \%$$

where,

s is the standard deviation

\bar{x} is the mean

**TIP**

In order to express a result as a percentage %, multiply the expression by 100.

The coefficient of variation is therefore:

$$\text{coefficient of variation, } CV = \frac{s \times 100}{\bar{x}}$$

EXAMPLE

Using the data from a previous example, calculate the coefficient of variation.

$$\text{mean, } \bar{x} = \frac{\Sigma fx}{n} = \frac{825}{25} = 33$$

standard deviation, $s = 11,21$

$$\text{coefficient of variation, } CV = \frac{s}{\bar{x}} \% = \frac{11,21}{33} \% = 33,97\%$$

Interpretation: the data are moderately dispersed around the mean.

All the measures of dispersion described so far have dealt with a single set of data. In practice, it is often important to compare two or more sets of data with different means, sample sizes or measurement units and the coefficient of variation can be used to do this.

The higher the coefficient of variation result, the more variability there is in a set of data.

EXAMPLE

A manufacturing company produces a product in two sizes, a 1 000 ml bottle and a 500 ml bottle. Different filling equipment is used for each size. Because of mechanical variability in the filling equipment, there is a standard deviation of 5 ml and 4 ml respectively.

Calculate the coefficient of variation for each filling machine and determine which machine is more consistent.

$$\text{coefficient of variation for 1 000 ml product, } CV = \frac{s}{\bar{x}} \% = \frac{5}{1\,000} \% = 0,5\%$$

$$\text{coefficient of variation for 500 ml product, } CV = \frac{s}{\bar{x}} \% = \frac{4}{500} \% = 0,8\%$$

Interpretation

Although the machine filling the smaller bottle has a lower standard deviation, the CVs indicate that the machine filling the larger bottle is relatively more consistent.



SELF-ASSESSMENT ACTIVITY

Two growers of grapefruit have obtained statistics regarding the mass of their current crops :

Grower A: $\bar{x} = 300$ g with $s = 20$ g

Grower B: $\bar{x} = 280$ g with $s = 40$ g

Which grower's grapefruit are more uniform in mass?

SOLUTION TO SELF-ASSESSMENT ACTIVITY

$$\text{coefficient of variation for grower A, } CV = \frac{s}{\bar{x}} \% = \frac{20}{300} \% = 6,67\%$$

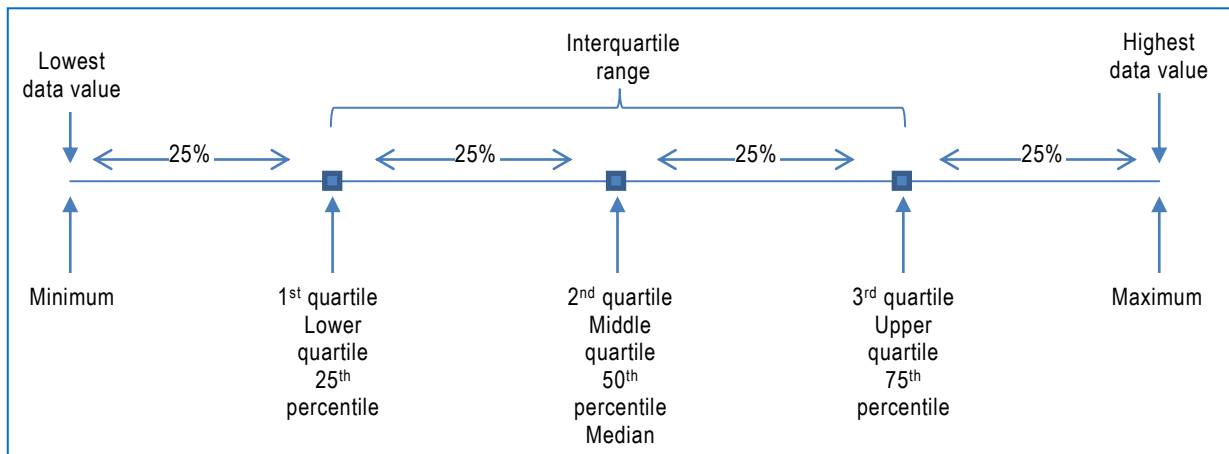
$$\text{coefficient of variation for grower B, } CV = \frac{s}{\bar{x}} \% = \frac{40}{280} \% = 14,29\%$$

Grower A's grapefruit has the lower CV and therefore is more consistent.

3.6 Interquartile and interpercentile ranges

In order to eliminate outliers (very low and very high values) and their effect on measures of central location and dispersion, ranges of a dataset to include mid values are often used:

- The interquartile range excludes the highest and lowest quarters of values.
- An interpercentile or mid-percentile range excludes a certain percentage of values at the lowest and highest ends of the dataset.



Formula: Interquartile range:

$$\text{interquartile range} = Q_3 - Q_1$$

where,

Q_3 is the third or upper quartile

Q_1 is the first or lower quartile

Refer unit 2 for calculation of quartiles.

EXAMPLE

Using the data from the example in section 1.3.1 calculate the interquartile range:

Interval (weight in lbs)	Frequency	Cumulative frequency
140 – 150	1	1
150 – 160	4	5
160 – 170	8	13
170 – 180	7	20
180 – 190	5	25

The quartiles are calculated in a previous example

$$Q_1 = 161,56 \text{ lbs}$$

$$Q_3 = 178,21 \text{ lbs}$$

$$\text{interquartile range} = Q_3 - Q_1 = 178,21 - 161,56 = 16,65 \text{ lbs}$$



SELF-ASSESSMENT ACTIVITY

The time taken to complete an assembling task has been measured for 250 employees:

Time taken (minutes)	Number of people (f)	Cumulative frequency $f(<)$
0 – 5	2	2
5 – 10	2	4
10 – 15	3	7
15 – 20	5	12
20 – 25	5	17
25 – 30	18	35
30 – 35	85	120
35 – 40	92	212
40 – 45	37	249
45 – 50	1	250
Total	250	

Calculate the interquartile range (the first and third quartiles were calculated in a previous self-assessment exercise).

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Quartiles already calculated in a previous self-assessment activity:

$$Q_1 = O_{Q_1} + \frac{c \left[\frac{1 \times n}{4} - f(<) \right]}{f_{Q_1}} = 30 + \frac{5(62,5 - 35)}{85} = 31,62 \text{ minutes}$$

$$Q_3 = O_{Q_3} + \frac{c \left[\frac{3 \times n}{4} - f(<) \right]}{f_{Q_3}} = 35 + \frac{5(187,5 - 120)}{92} = 38,67 \text{ minutes}$$

$$\text{interquartile range} = Q_3 - Q_1 = 38,67 - 31,62 = 7,05 \text{ minutes}$$

Formula: Interpercentile or mid-percentile range:

The mid-percentile range is the percentage of the range exactly in the middle of the dataset.

To calculate the upper and lower percentiles required for the upper and lower limits of the range:

$$\text{lower percentile of range} = \frac{100\% - \text{required range percentile}}{2}$$

$$\text{upper percentile of range} = \text{lower percentile of range} + \text{required range percentile}$$

Calculate the required positions and values for these percentiles

interpercentile or mid percentile range

= value of upper percentile of range

- value of lower percentile of range

EXAMPLE

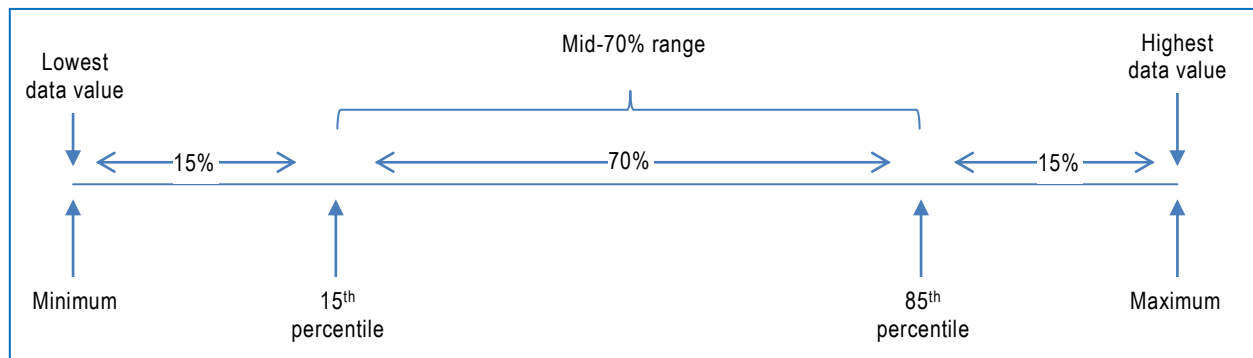
Using the data from the example in section 1.3.1 calculate the mid-70% range:

Interval (weight in lbs)	Frequency	Cumulative frequency
140 – 150	1	1
150 – 160	4	5
160 – 170	8	13
170 – 180	7	20
180 – 190	5	25

Calculate the upper and lower percentiles required for the upper and lower limits of the range:

$$\text{lower percentile of range} = \frac{100\% - \text{required range percentile}}{2} = \frac{100\% - 70\%}{2} = 15\%$$

$$\begin{aligned} \text{upper percentile of range} &= \text{lower percentile of range} + \text{required range percentile} \\ &= 15\% + 70\% = 85\% \end{aligned}$$



Calculate the required positions and values for these percentiles

$$\text{85th percentile position} = \frac{j \times n}{100} = \frac{85 \times 25}{100} = 21,25$$

$$\text{15th percentile position} = \frac{j \times n}{100} = \frac{15 \times 25}{100} = 3,75$$

$$\text{upper percentile, } P_{85} = O_{P_{85}} + \frac{c \left[\frac{85 \times n}{100} - f(<) \right]}{f_{P_{85}}} = 180 + \frac{10(21,25 - 20)}{5} = 182,5 \text{ lbs}$$

$$\text{lower percentile, } P_{15} = O_{P_{15}} + \frac{c \left[\frac{15 \times n}{100} - f(<) \right]}{f_{P_{15}}} = 150 + \frac{10(3,75 - 1)}{4} = 156,88 \text{ lbs}$$

interpercentile or mid percentile range

= value of upper percentile of range

- value of lower percentile of range = 182,5 - 156,88 = 25,62 lbs



SELF-ASSESSMENT ACTIVITY

The time taken to complete an assembling task has been measured for 250 employees:

Time taken (minutes)	Number of people (f)	Cumulative frequency $f(<)$
0 – 5	2	2
5 – 10	2	4
10 – 15	3	7
15 – 20	5	12
20 – 25	5	17
25 – 30	18	35
30 – 35	85	120
35 – 40	92	212
40 – 45	37	249
45 – 50	1	250
Total	250	

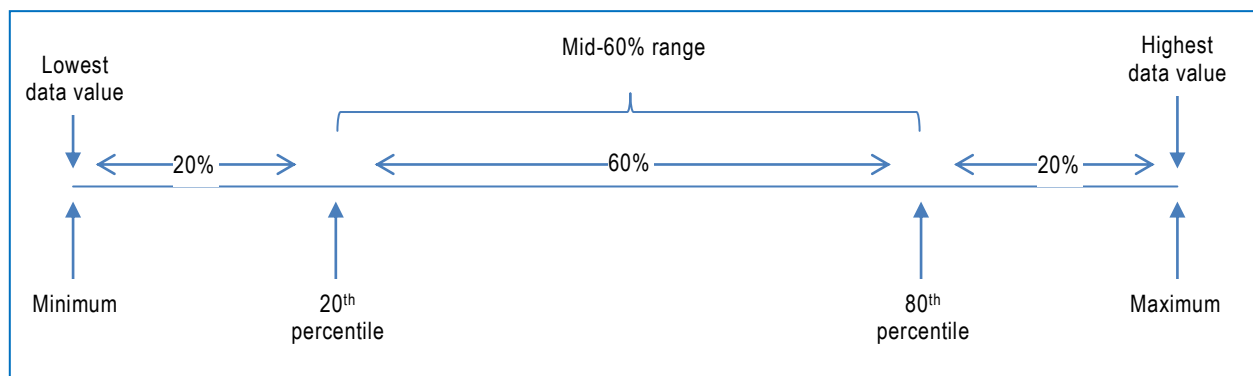
Calculate the mid-60% range.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Calculate the upper and lower percentiles required for the upper and lower limits of the range:

$$\text{lower percentile of range} = \frac{100\% - \text{required range percentile}}{2} = \frac{100\% - 60\%}{2} = 20\%$$

$$\begin{aligned} \text{upper percentile of range} &= \text{lower percentile of range} + \text{required range percentile} \\ &= 20\% + 60\% = 80\% \end{aligned}$$



Calculate the required positions and values for these percentiles

$$80\text{th percentile position} = \frac{j \times n}{100} = \frac{80 \times 250}{100} = 200$$

$$20\text{th percentile position} = \frac{j \times n}{100} = \frac{20 \times 250}{100} = 50$$

$$\text{upper percentile, } P_{80} = O_{P_{80}} + \frac{c \left[\frac{80 \times n}{100} - f(<) \right]}{f_{P_{80}}} = 35 + \frac{5(200 - 120)}{92} = 39,35 \text{ minutes}$$

$$\text{lower percentile, } P_{20} = O_{P_{20}} + \frac{c \left[\frac{20 \times n}{100} - f(<) \right]}{f_{P_{20}}} = 30 + \frac{5(50 - 35)}{85} = 30,88 \text{ minutes}$$

interpercentile or mid percentile range

= value of upper percentile of range

– value of lower percentile of range = 39,35 – 30,88 = 8,47 minutes

3.7 Quartile deviation

The quartile deviation is half the interquartile range of a dataset. It is a measure of the spread through the middle half of the dataset. It can be useful because it is not influenced by extremely high or extremely low values.

Formula: quartile deviation:

$$\text{quartile deviation} = \frac{IQR}{2} = \frac{Q_3 - Q_1}{2}$$

where,

IQR is the interquartile range

Q_3 is the third or upper quartile

Q_1 is the first or lower quartile

Refer unit 2 for calculation of quartiles and formula for the calculation of the interquartile range.

EXAMPLE

Using the data from the example in section 1.3.1 calculate the quartile deviation:

Interval (weight in lbs)	Frequency	Cumulative frequency
140 – 150	1	1
150 – 160	4	5
160 – 170	8	13
170 – 180	7	20
180 – 190	5	25

The interquartile range was calculated in example

$$IQR = 16,65 \text{ lbs}$$

$$\text{quartile deviation} = \frac{IQR}{2} = \frac{16,65}{2} = 8,33 \text{ lbs}$$



SELF-ASSESSMENT ACTIVITY

The time taken to complete an assembling task has been measured for 250 employees:

Time taken (minutes)	Number of people (f)	Cumulative frequency f(<)
0 – 5	2	2
5 – 10	2	4
10 – 15	3	7
15 – 20	5	12
20 – 25	5	17
25 – 30	18	35
30 – 35	85	120
35 – 40	92	212
40 – 45	37	249
45 – 50	1	250
Total	250	

Calculate the quartile deviation (the interquartile range was calculated in a previous self-assessment exercise).

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Interquartile deviation already calculated in the previous self-assessment activity:

$$\text{interquartile range} = 7,05 \text{ minutes}$$

$$\text{quartile deviation} = \frac{IQR}{2} = \frac{7,05}{2} = 3,53 \text{ minutes}$$

Unit 3 Exercises: (Solutions are found at the end of the module guide)**Exercise 3.1**

The percentage of family income allocated to groceries for a sample of 50 shoppers is:

Percentage	Frequency
10 – 20	6
20 – 30	14
30 – 40	16
40 – 50	11
50 – 60	3

Calculate the mean and standard deviation.

Exercise 3.2

The monthly rent paid by 150 employees in a company is:

Rent (Rand)	Number of employees
100 – 160	20
160 – 220	35
220 – 280	39
280 – 340	22
340 – 400	19
400 – 460	15

Calculate:

- The median monthly rent.
- The standard deviation of the monthly rent.
- Find the lower and upper quartile monthly rents.
- Determine the quartile deviation of monthly rent.
- Interpret the meanings of these results.

Exercise 3.3

Results obtained from a frequency distribution are:

$$\sum f = 100$$

$$\sum fx = 3\,460$$

$$\sum fx^2 = 124\,690$$

Find the mean and standard deviation of the x variable.

Exercise 3.4

Consider the following two sets of data, each with 5 observations:

Set 1	30	40	50	60	70
Set 2	130	140	150	160	170

Which of the following descriptive measures is the same for both sets of data? Give reasons for your answers. No calculations are necessary.

- Mean.
- Variance.
- Mode.

Exercise 3.5

12 female clients who purchase a particular brand of perfume give the following ratings (based on the 5-point Likert scale) to each of two statements concerning brand preference.

The Likert rating scale is as follows:

- 1 – Strongly disagree.
- 2 – Disagree.
- 3 – Neutral.
- 4 – Agree.
- 5 – Strongly agree.

Response ratings to each statement.

Statement 1	2	2	3	4	2	3	2	1	3	3	2	4
Statement 2	3	3	2	4	3	3	2	4	4	5	3	4

- Determine the mean response rating per statement.
- Find the standard deviation of response ratings per statement.
- Compute the coefficient of variation for responses to each statement. On which statement is there less consensus? Explain.

Exercise 3.6

Employee bonuses earned by workers at a furniture factory in a recent month (in Rands) are:

47	31	42	33	58	51
25	28	62	29	65	91
51	30	43	72	73	37
29	39	53	61	52	35

- Find the mean and standard deviation of bonuses.
- Find the interquartile range and quartile deviation.

Exercise 3.7

Voditel International owns a large fleet of company cars. The mileages, in thousands of kilometres, of a sample of 17 of their cars over the last financial year is:

11	31	27	26	27	35	23	19	28
25	15	36	29	27	26	22	20	

Calculate the mean and standard deviation of these mileage figures.

Exercise 3.8

The kilocalories per portion in a sample of 32 different breakfast cereals are recorded as:

Kcal per portion	Frequency
80 – 120	3
120 – 160	11
160 – 200	9
200 – 240	7
240 - 280	2

Calculate approximate values for the mean and standard deviation of the distribution.

UNIT 4 PROBABILITY

UNIT 4: PROBABILITY

OBJECTIVES

By the end of this study unit, you should be able to:

1. Describe what is meant by probability.
2. Apply the properties and rules of probabilities to solve problems.
3. Describe the general approaches for assigning probabilities.
4. Identify a sample space for a random experiment.
5. Describe conditional probability and independent events.
6. Construct and use a probability tree.

CONTENT

- 4.1 Introduction
 - 4.2 Establishing probabilities
 - 4.3 Probability symbols
 - 4.4 Properties of a probability
 - 4.5 Probability concepts
 - 4.6 Calculating probabilities
 - 4.7 Probability rules
 - 4.7.1. Addition rule
 - 4.7.2. Multiplication rule
- Exercises



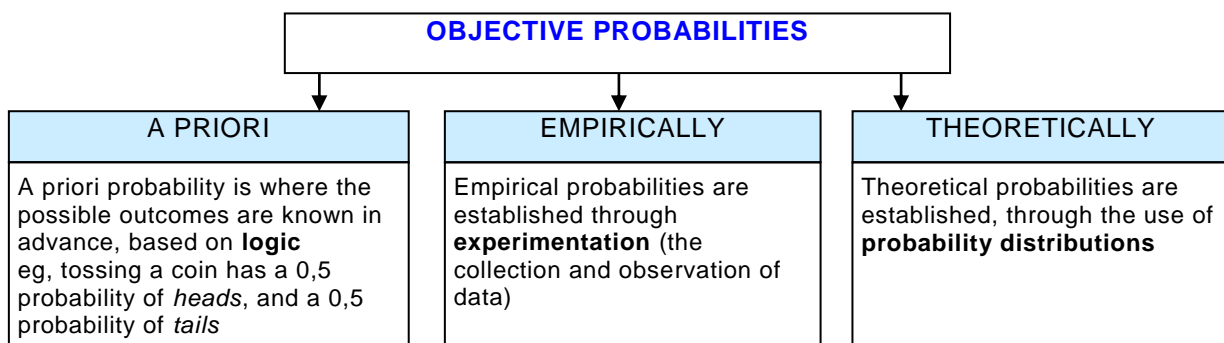
Prescribed textbook: Chapter 4

4.1 Introduction

This unit introduces the basic concepts of probability. It outlines rules and techniques for assigning probabilities to events. Probability plays a critical role in statistics. All of us form simple probability conclusions in our daily lives. Sometimes these determinations are based on fact, while others are subjective. If the probability of an event is high, one expects the event to occur rather than not to occur. If the probability of rain is more than 50%, it is more likely to rain than not.

4.2 Establishing probabilities

Probabilities can be based on educated guesswork (subjective probabilities) or they can be objectively established.



Definition: Probability:

Probability can be defined as a measure of **how likely** it is that an event will occur.

The probability of an event occurring is **represented by a number that lies between 0 (impossible) and 1 (certain)**.

Probabilities are expressed as proportions or factors, e.g. 0, 5, rather than as percentages, e.g. 50%, although in speaking, we tend to use percentages.

If you conduct a trial, e.g. toss a coin, there will be an outcome (either a head or a tail). Because there exists only 2 possible outcomes in your trial, there is an even chance (0, 5 probability) of the outcome being heads, and a 0, 5 probability of the outcome being tails with all possible outcomes adding up to 1.

4.3 Probability symbols

Commonly used symbols when describing probability relationships are:

Symbol	Meaning
$P(\dots)$	Probability of whatever event is stated between brackets.
$P(A)$	Probability of an event of a specific type (or specific properties).
$P(\bar{A})$	Probability that the event will not occur.
$P(A \cap B)$	The symbol is read as 'and'; this is the probability that event A and B will occur.
$P(A \cup B)$	The symbol is read as 'or'; this is the probability that event A will occur or event B will occur.
$P(A B)$	Probability of A, given that B happens/does something.

4.4 Properties of a probability

Five basic probability requirements apply:

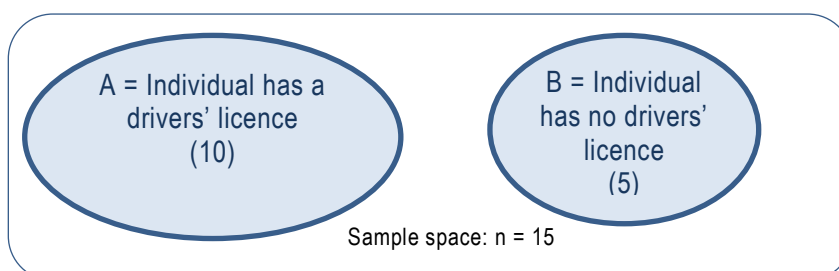
1.	Each even probability lies between 0 and 1 inclusive.
2.	If an event cannot occur, then $P(A) = 0$
3.	If an event is certain to occur, then $P(A) = 1$
4.	The sum of the probabilities of all possible events in the sample space equals 1, i.e. $P(A_1) + P(A_2) + P(A_3) + P(A_4) + \dots + P(A_n) = 1$ for n events.
5.	If $P(A)$ is the probability of an event occurring, then the probability of the event not occurring is $P(\bar{A}) = 1 - P(A)$

4.5 Probability concepts

Events are said to be **mutually exclusive** if they cannot occur together in a single trial, ie at the same time.

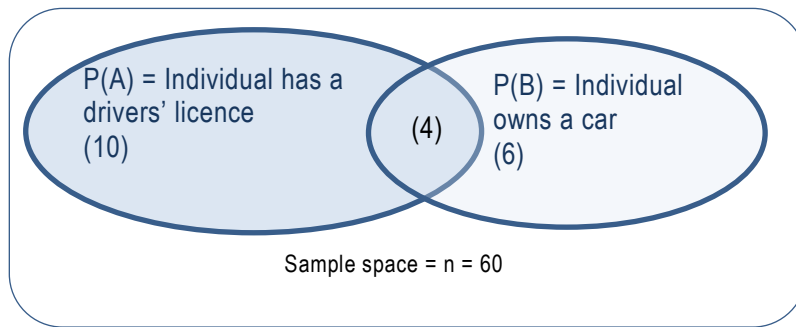
EXAMPLE

An individual either has a drivers' licence or he does not.



These events **are mutually exclusive** as there is no possibility that an individual both has a drivers' licence and does not have one (there is no overlap in the Venn diagrams).

However, an individual can both own a car and have a drivers' licence:



In this case the events **are not mutually exclusive**.

4.6 Calculating probabilities

Definition: Marginal probability $P(A)$

A **marginal probability** is the probability of a single event occurring.

Marginal probability is denoted by $P(A)$.

EXAMPLE

A sample of university students indicates degree courses studied by gender:

Course	Gender		Total
	Male	Female	
BCom	20	18	38
BA	14	18	32
BSc	23	27	50
Total	57	63	120

Note: this table is known as a **contingency table**.

What is the probability of a student being female?

$$P(\text{female}) = \frac{63}{120} = 0,525 \text{ or } 52,5\%$$

There is a 0,525 probability or a 52, 5% chance that a student is female.



TIP

It is recommended that you round all probability calculations to 4 decimal places. That way if you do need to present your results as a percentage, the percentage value will be to 2 decimal places.

Definition: Joint probability $P(A \cap B)$

Joint probability is the probability that event A and event B will occur at the same time in a single trial.

Joint probability is denoted by $P(A \cap B)$.

Example

Using the table in the previous example, calculate the probability that a BCom student will be male.

$$P(\text{BCom} \cap \text{male}) = \frac{20}{120} = 0,1667 \text{ or } 16,67\%$$

There is a 0,1667 probability or 16,67% chance that a BCom student will be male.

If events are mutually exclusive, there is no joint probability.

**TIP**

In sense-testing answers to probability questions, remember, probability can never be greater than 1.

Definition: Conditional probability $P(A|B)$

Conditional probability is the probability that event A will occur given that event B has already occurred.

Conditional probability is denoted by $P(A|B)$.

Formula: Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Important point about conditional probability: when probability is conditional, the sample space is reduced to only those outcomes associated with the given event.

**TIP**

The key to recognising conditional probability is to look for the words "given that" or "knowing that" or their equivalent.

Example

Using the table in the previous example, calculate the probability that a student is studying BCom given that the student is male.

Instead of dividing by the total sample space of 120, the division is by only the male students, i.e. the sample space has been reduced to only males, i.e. 57 students.

$$P(BCom|male) = \frac{P(BCom \cap male)}{P(male)} = \frac{20}{57} = 0,3509 \text{ or } 35,09\%$$

There is a 0,3509 probability or 35,09% chance that a student is studying BCom given that the student is male.



SELF-ASSESSMENT ACTIVITY

A company's employees are classified according to age and salary:

Age	Salary			Total
	< \$25 000	\$25 000 - \$45 000	> \$45 000	
< 30	32	3	0	35
30 – 45	10	18	21	49
> 45	1	10	5	16
Total	43	31	26	100

One employee is selected at random and two events are defined as:

- A** The employee is under 30
- B** The employee's salary is below \$25 000.

Calculate each probability and express in words:

- a) $P(A|B)$
- b) $P(B|A)$
- c) $P(A|\bar{B})$
- d) $P(\bar{A}|B)$

SOLUTION TO SELF-ASSESSMENT ACTIVITY

- a) Given that the employee's salary is below \$25 000, the probability that the employee is under 30:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\frac{32}{100}}{\frac{32 + 10 + 1}{100}} = \frac{32}{43} = 0,7442 \text{ or } 74,42\%$$

Alternatively, rather than using the formula directly, we can find $P(A|B)$ by simply consulting the contingency table. Given that the selected employee's salary is below \$25 000, the sample space of possible outcomes is reduced to the 43 employees who fall into this category. Of these employees, 32 are under the age of 30.

$$P(A|B) = \frac{32}{43} = 0,7442 \text{ or } 74,42\%$$

b) Given that the employee is under 30, the probability that the employee's salary is below \$25 000:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{\frac{32}{100}}{\frac{32+3+1}{100}} = \frac{32}{35} = 0,9143 \text{ or } 91,43\%$$

Or directly from the table, by reducing the sample size to the 35 under age 30 employees.

c) Given that the employee's salary is at least \$25 000, the probability that the employee is under 30:

$$P(A|\bar{B}) = \frac{P(A \text{ and } \bar{B})}{P(\bar{B})} = \frac{\frac{3+0}{100}}{\frac{31+26}{100}} = \frac{3}{57} = 0,0526 \text{ or } 5,26\%$$

Or directly from the table, by reducing the sample size to the 57 employees with salaries above \$25 000 (31+26).

d) Given that the employee's salary is below \$25 000, the probability that the employee is at least 30:

$$P(\bar{A}|B) = \frac{P(\bar{A} \text{ and } B)}{P(B)} = \frac{\frac{10+1}{100}}{\frac{43}{100}} = \frac{11}{43} = 0,2558 \text{ or } 25,58\%$$

Or directly from the table, by reducing the sample size to the 43 employees with salaries below \$25 000.

4.7 Probability rules

4.7.1 Addition rule

The addition rule is used to find the probability of either of one or more events occurring. It includes the probability that events occur simultaneously if they are mutually exclusive events.

Formula/rule: Addition rule for non-mutually exclusive events

The probability of either event A occurring or event B occurring (or more) or both events occurring in a single trial:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The reason for deducting the joint probability is to ensure the 'overlap' is not counted twice.

**TIP**

The 'or' in probability denotes addition, or +.

EXAMPLE

Using the student sample in the previous example, calculate the probability that the student is male or is studying BCom or both.

$$P(\text{male} \cup \text{BCom}) = P(\text{male}) + P(\text{BCom}) - P(\text{male} \cap \text{BCom}) = \frac{57}{120} + \frac{38}{120} - \frac{20}{120} = \frac{75}{120} \\ = 0,625 \text{ or } 62,5\%$$

There is a 0,625 probability or 62, 5% chance that a student is male or is studying BCom or both.

Formula/rule: Addition rule for mutually exclusive events

The probability of either event A occurring or event B occurring in a single trial (by definition both events cannot occur simultaneously):

$$P(A \cup B) = P(A) + P(B)$$

EXAMPLE

Using the student sample in the previous example, calculate the probability that the student is studying BSc or studying BCom.

$$P(\text{BSc} \cup \text{BCom}) = P(\text{BSc}) + P(\text{BCom}) = \frac{50}{120} + \frac{38}{120} = \frac{88}{120} = 0,7333 \text{ or } 73,33\%$$

There is a 0, 7333 probability or 73, 33% chance that a student is studying BSc or BCom.

EXAMPLE

The distribution of blood types in a certain country is roughly:

A: 41% B: 9% AB: 4% O: 46%

An individual is brought into the emergency room after an automobile accident. What is the probability that he will be of type A or B or AB?

$$P(\text{A or B or AB}) = P(A) + P(B) + P(AB) = 0,41 + 0,09 + 0,04 = 0,554$$

Since it is impossible for one individual to have two different blood types, these events are mutually exclusive.

4.7.2 Multiplication rule

The multiplication rule is used to establish the joint probability of two events occurring at the same time in a single trial.

Formula/rule: Multiplication rule for statistically dependent events

The probability of both event A occurring and event B occurring in a single trial. The events are associated (ie, they are statistically dependent).

$$P(A \cap B) = P(A|B) \times P(B)$$

EXAMPLE

Using the student sample in the previous example, calculate the probability that the student is studying BCom and is male.

$$P(BCom \cap male) = P(BCom|male) \times P(male) = \frac{20}{57} \times \frac{57}{120} = \frac{20}{120} = 0,1667 \text{ or } 16,67\%$$

There is a 0,1667 probability or 16,67% chance that a student is studying BCom and is male. (You'll note this is the same as the answer in example, which proves that the events are dependent).

Formula/rule: Multiplication rule for statistically independent events

The probability of both event A occurring and event B occurring in a single trial. The events are independent.

$$P(A \cap B) = P(A) \times P(B)$$



TIP

The 'and' in probability denotes multiplication or X

EXAMPLE

In a pack of 52 cards, there are 13 cards in each of the four suits, spades, hearts, diamonds and clubs. There are 4 kings in the pack, one of each suit.

What is the probability of drawing a king of diamonds from the pack?

$$P(king \cap diamond) = P(king) \times P(diamond) = \frac{4}{52} \times \frac{13}{52} = \frac{1}{13} \times \frac{13}{52} = \frac{1}{52} = 0,0192 \text{ or } 1,92\%$$

There is a 0,0192 probability or 1,92% chance that a card drawn from the pack is the king of diamonds.

EXAMPLE

The probability that a certain plant will flower during the first summer is 0,6. If five plants are planted, calculate the probability that all of them will have flowers during the first summer.

$$\begin{aligned}
 P(\text{all flowers}) &= P(\text{1st plant flowering and 2nd and 3rd and 4th and 5th and 6th}) \\
 &= P(\text{1st}) \times P(\text{2nd}) \times P(\text{3rd}) \times P(\text{4th}) \times P(\text{5th}) \times P(\text{6}) \\
 &= 0,6 \times 0,6 \times 0,6 \times 0,6 \times 0,6 \times 0,6 = 0,078
 \end{aligned}$$

**SELF-ASSESSMENT ACTIVITY**

A manufacturing plant conducted a survey to determine its employees' reactions toward a proposed change in working hours. A breakdown of the responses is:

Division	Agree	Disagree
Production	17	23
Office	8	2

Suppose an employee is chosen at random, with the relevant events being defined as:

- A: The employee works in production.
 B: The employee agrees with the proposed change.

Express each of the following events in words:

- \bar{A}
- $(A \text{ or } B)$
- $(A \text{ and } B)$
- $(A \text{ or } \bar{B})$

SOLUTION TO SELF-ASSESSMENT ACTIVITY

- \bar{A} The employee works in the office.
- $(A \text{ or } B)$ The employee either works in production or agrees with the proposed change or both.
- $(A \text{ and } B)$ The employee works in production and agrees with the proposed change.
- $(A \text{ or } \bar{B})$ The employee either works in production or disagrees with the proposed change or both.

**SELF-ASSESSMENT ACTIVITY**

Refer to the previous self-assessment activity and find:

- $P(\bar{A})$
- $P(A \text{ or } B)$
- $P(A \text{ and } B)$

SOLUTION TO SELF-ASSESSMENT ACTIVITY

- $P(\bar{A}) = \frac{8}{50} + \frac{2}{50} = \frac{10}{50} = 0,2$
- $P(A \text{ or } B) = \frac{17}{50} + \frac{23}{50} = \frac{48}{50} = 0,96$
- $P(A \text{ and } B) = \frac{17}{50} = 0,34$

**SELF-ASSESSMENT ACTIVITY**

Wild azaleas are classified by colour and by the presence and absence of fragrance.

Fragrance	White	Pink	Orange	Total
Yes	12	60	58	130
No	50	10	10	70
Total	62	70	68	200

If an azalea is randomly selected from the group, calculate:

- $P(\text{fragrance})$
- $P(\text{orange})$
- $P(\text{orange with fragrance})$
- $P(\text{orange knowing it has fragrance})$
- $P(\text{has fragrance given that it is orange})$

SOLUTION TO SELF-ASSESSMENT ACTIVITY

- $P(\text{fragrance}) = \frac{130}{200} = 0,65$
- $P(\text{orange}) = \frac{68}{200} = 0,34$

c) $P(\text{orange with fragrance}) = \frac{58}{200} = 0,29$

d) $P(\text{orange knowing it has fragrance}) = \frac{58}{130} = 0,4462$

e) $P(\text{has fragrance given that it is orange}) = \frac{58}{68} = 0,8529$

Unit 4 Exercises: (Solutions are found at the end of the module guide)**Exercise 4.1**

The personnel department of an insurance firm analysed the qualifications profile of their 129 managers, noting the highest qualification achieved by each.

Qualification	Section head	Department head	Division head
Matric	28	14	?
Diploma	20	24	6
Degree	?	10	14
Total	53	?	28

- Define the two random variables, their measurement scale and data type.
- Complete the contingency table.
- What is the probability of a person selected at random:
 - Having only a matric?
 - Being a section head and having a degree?
 - Being a department head given that they have a diploma?
 - Being a division head?
 - Being a division head or a section head?
 - Having matric or a diploma or a degree?
 - Having matric given that the person is a department head?
 - Being a division head or having a diploma?
- For each probability computed in (d), state:
 - the type of probability (i.e. marginal, joint, conditional)
 - which probability rule, if any, was applied (i.e. addition rule; multiplication rule)
- Are the events in (v) and (viii) mutually exclusive?

Exercise 4.2

The incomplete relative frequency table for events X_1, X_2, X_3, X_4 and Y_1, Y_2 and Y_3 , is:

	X_1	X_2	X_3	X_4	Total
Y_1		0,03	0,12	0,03	0,25
Y_2	0,05		0,10		
Y_3	0,08	0,12	0,18	0,10	
Total		0,22			

- Find $P(X_1)$
- Find $P(Y_1)$
- Find $P(Y_1 \text{ and } X_4)$
- Find $P(X_1 \text{ or } Y_2 \text{ or } Y_3)$
- Find $P(X_1 \text{ or } Y_2)$

Exercise 4.3

The following probability distribution refers to two characteristics: age and traffic offences over the past 12 months of residents in Bloemfontein.

Age (years)	No offence F_1	One offence F_2	Two or more offences F_3
< 18 E_1	0,23	0,12	0,05
\geq 18 E_2	0,45	0,14	0,01

- What is the probability that a randomly selected resident had no traffic offences in the last 12 months, given that he/she is 18 or older?
- What is the probability that a randomly selected resident had two or more traffic offences in the last 12 months?
- Calculate $P(E_1 \cup F_2)$.
- If event A is a randomly selected resident under 18 with fewer than two offences, find $P(\bar{A})$

Exercise 4.4

2 500 employees of a large corporation are classified by gender and by opinion on a proposal to emphasise fringe benefits rather than wage increases in pending wage discussions:

Gender	In favour F_1	Neutral F_2	Opposed F_3	Total
Male E_1	900	150	450	1 500
Female E_2	300	100	600	1 000
Total	1 200	250	1050	2 500

- Given that an employee chosen at random is a male, find the probability that the employee is opposed to the proposal.
- Find the joint probability that an employee picked at random is a female opposed to the proposal.

Exercise 4.5

A wine dealer has classified the last 200 customers according to age and type of wine purchased:

Type of wine purchased	Age of customer			Total
	Under 30	30 to < 50	50 and over	
Namibian	100	30	20	150
French	2	2	16	20
German	2	16	2	20
Other	4	6	0	10
Total	108	54	38	200

Find the following probabilities:

- $P(< \text{age } 30)$.
- $P(\text{Namibian})$.
- $P(\text{Namibian or French})$.
- $P(< \text{age } 30 \text{ and Namibian})$.
- $P(< \text{age } 30 \text{ or Namibian})$.

Exercise 4.6

A company has 1 000 credit customers. They are classified according to the size of their account balances and the timeliness of their payments.

Last payment	< R100	R100 – R500	> R500	Total
On time		0,45		0,85
Late			0,03	
Total	0,20	0,50		

- Complete the missing probabilities.
- How many customers have a balance of less than R100 and made their last payment late?

Exercise 4.7

A business person carries fire insurance on home and store. During a given year the probability is 0, 01 of fire at the home (event A); the chance of a store fire (event B) is 0, 06. If A and B are independent, find:

- $P(A \text{ and } B)$
- $P(\bar{A})$

Exercise 4.8

300 employees of a small manufacturing company are cross-classified on the basis of age and work category.

Age	Production	Sales	Office	Total
<25	50	2	50	102
25 – 40	70	24	50	144
> 40	40	4	10	54
Total	160	30	110	300

An employee is selected at random from this population. Calculate the probability that the employee is:

- under 25 years of age
- a production workers
- a sales person and between 25 and 40 years of age
- over 40, given that he/she is an office worker
- A production worker or under 25 or both.

Exercise 4.9

1 000 adult shoppers are cross-classified on the basis of lifestyle (L) and the shop (S) from which they make most of their clothing purchases.

Lifestyle	S ₁	S ₂	S ₃	S ₄	Total
L ₁	110	214	16	58	398
L ₂	152	170	22	52	396
L ₃	52	36	16	102	206
Total	314	420	54	212	1 000

Calculate:

- $P(L_2 \text{ or } S_4)$
- $P(L_3)P(S_1 | L_3)$. Using probability notation, what is $P(L_3)P(S_1 | L_3)$ equivalent to?

Exercise 4.10

Explain the meaning of the following terms used in the theory of probability and give an example of each:

- Joint events.
- Independent events.
- Mutually exclusive events.

Exercise 4.11

In a group of people on holiday it is established that there are:

- 10 males under the age of 21.
- 8 females under the age of 21.
- 6 males aged between 21 and 30.
- 5 females aged between 21 and 30.
- 7 males over the age of 30.

a) Calculate the probability that, if one person is selected at random from the group this person will be:

(i) A male under the age of 30.

(ii) A female.

(iii) A female over the age of 30.

(iv) A male over the age of 21.

(v) A person not older than 21.

(vi) A male given that this person is over 30 years.

(vii) A female or between 21 and 30 years.

(viii) A male or a female.

b) For each question in (a), state the probability type and the probability rule used, if appropriate.

UNIT 5
PROBABILITY DISTRIBUTIONS

UNIT 5: PROBABILITY DISTRIBUTIONS

OBJECTIVES

By the end of this study unit, you should be able to:

1. Define a probability distribution or function.
2. Distinguish between discrete and continuous random variables.
3. Use the binomial, Poisson and normal distribution models to calculate probabilities.

CONTENT

5.1 Introduction

5.2 Discrete probability distributions

5.3 The binomial distribution

5.4 The Poisson distribution

5.5 Continuous probability distributions

5.5.1. The normal probability distribution

5.5.2. The standard normal distribution or z-curve

Exercises

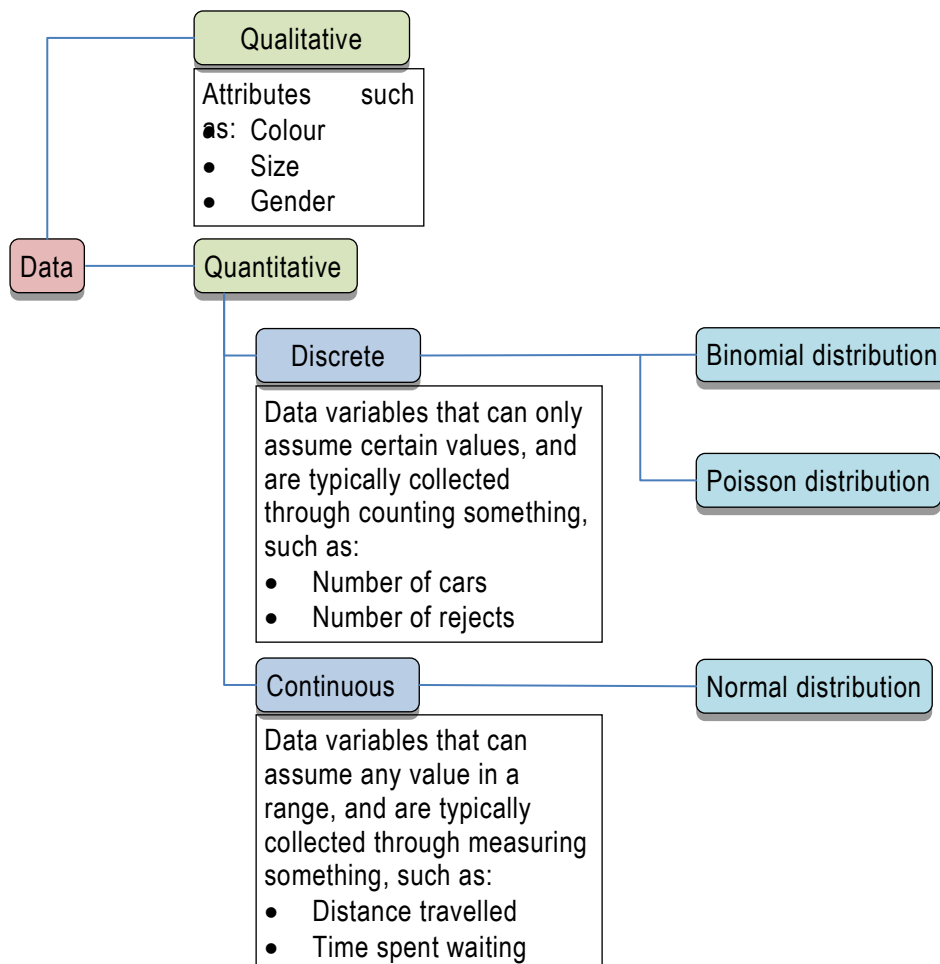


Prescribed textbook: Chapter 5

5.1 Introduction

Probability distributions are classified as either discrete or continuous, depending on the random variable.

- A random variable is discrete if it can assume only a countable number of possible values (0, 1, 2, 3, etc.), e.g., number of people.
- A continuous random variable has an uncountable number of possible values; it can take on any value in an interval, e.g. distance.



Definition: Probability distribution

A probability distribution is all possible outcomes of a random variable and their associated probabilities of occurrence.

Definition: Probability function

A probability function is a function of a discrete random variable that gives the probability that the outcome associated with that variable will occur.

A probability function is denoted by $P(x)$, where

- $0 \leq P(x) \leq 1$.
- $\sum P(x) = 1$, the sum of the probabilities for all possible outcomes, equals one.

5.2 Discrete probability distributions

Discrete probability distributions determine the outcomes of discrete random variables.

5.2.1 The binomial distribution

A discrete random variable follows a **binomial distribution** if:

- The experiment consists of n identical trials.
- Each trial has 1 of 2 possible mutually exclusive outcomes: success or failure, eg a person is male or not, a part is defective or not.
- The probability (π) that the trial results in a success remains the same from trial to trial. The probability of failure is $1 - \pi$.
- The trials are independent of each other (the outcome of a trial does not affect the outcome of any other trial).

Formula: Binomial distribution probability

The probability of x successes in a randomly drawn sample of n trials of a binomial experiment is:

$$P(x) = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}$$

Where

n sample size or number of trials

π probability of success of each trial

$(1 - \pi)$ the probability of failure of each trial

x the number of successful outcomes

$n!$ the factorial of n , ie $n \times (n - 1) \times (n - 2) \dots \times 1$, eg $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$



TIP

The success outcome with its probability of π relates to the binomial probability $P(x)$. It's important to always ensure the π relates to the probability being calculated.

For example if we want to assess the probability of a student being a vegetarian, then π will be the sample proportion or probability of being a vegetarian and $1 - \pi$ will be the probability of not being a vegetarian.



TIP

The first part of the function can be easily computed on a statistics calculator or in Excel using the 'combination' function:

Part of binomial formula	Function on statistics calculator
$P(x) = \frac{n!}{x!(n-x)!}$	${}_n C_r$ <p>Where n sample size r required probability, i.e. x</p>

EXAMPLE

A textile firm has found from experience that only 20% of people applying for certain stitching-machine jobs are qualified for the work. If 5 people are interviewed, what is the probability of finding 3 qualified people?

$$P(3) = \frac{5!}{3!(5-3)!} 0,2^3 (1 - 0,2)^{5-3} = 0,0512 \text{ or } 5,12\%$$

There is a 0,0512 probability or 5,12% chance of finding 3 qualified people.



TIP

Sense test the 'layout' of the formula, success probability π will be to the power of the number you're trying to calculate and failure probability $1 - \pi$ will be to the power of the rest of the sample.

**TIP**

Words such as 'at least', 'less than', 'more than', 'no more than', 'less than', 'no less than' and similar need the calculation of multiple probabilities.

Remember that all possible options add up to 1, with all probabilities being $0, 1, 2, \dots, n$

EXAMPLE

A textile firm has found from experience that only 20% of people applying for certain stitching-machine jobs are qualified for the work. If 5 people are interviewed, what is the probability of finding at least two qualified people?

$$\pi = 0,2$$

$$\text{probability (at least 2)} = P(2) \text{ or } P(3) \text{ or } P(4) \text{ or } P(5) = P(2) + P(3) + P(4) + P(5)$$

$$P(2) = \frac{5!}{2!(5-2)!} 0,2^2(1-0,2)^{5-2} = 0,2048$$

$$P(3) = 0,0512 \text{ (see previous example)}$$

$$P(4) = \frac{5!}{4!(5-4)!} 0,2^4(1-0,2)^{5-4} = 0,0064$$

$$P(5) = \frac{5!}{5!(5-5)!} 0,2^5(1-0,2)^{5-5} = 0,0003$$

$$\begin{aligned} \text{probability } (\geq 2) &= P(2) + P(3) + P(4) + P(5) = 0,2048 + 0,0512 + 0,0064 + 0,0003 \\ &= 0,2627 \text{ or } 26,27\% \end{aligned}$$

Alternatively there is a simpler approach because only two probabilities need to be calculated:

$$\text{probability (at least 2)} = P(\geq 2) = 1 - P(< 2) = 1 - (P(0) + P(1))$$

$$P(0) = \frac{5!}{0!(5-0)!} 0,2^0(1-0,2)^{5-0} = 0,3277$$

$$P(1) = \frac{5!}{1!(5-1)!} 0,2^1(1-0,2)^{5-1} = 0,4096$$

$$P(\geq 2) = 1 - (P(0) + P(1)) = 1 - (0,3277 + 0,4096) = 0,2627 \text{ or } 26,27\%$$

There is a 0,2627 probability or 26,27% chance of finding at least two qualified people.

Important note:

$$\begin{aligned} P(0) \text{ or } P(1) \text{ or } P(2) \text{ or } P(3) \text{ or } P(4) \text{ or } P(5) \\ &= P(0) + P(1) + P(2) + P(3) + P(4) + P(5) \\ &= 0,3277 + 0,4096 + 0,2048 + 0,0512 + 0,0064 + 0,0003 = 1 \end{aligned}$$



SELF-ASSESSMENT ACTIVITY

A shoe store's records show that 30% of customers use a credit card to make payment. On a particular morning, 7 customers purchase shoes from the store. Determine the probability that:

- 3 customers will pay by credit card.
- At least one customer will pay by credit card.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

$$\pi = 0,3$$

$$P(3) = \frac{7!}{3!(7-3)!} 0,3^3 (1 - 0,3)^{7-3} = 0,2269 \text{ or } 22,69\%$$

There is a 0,2269 probability or 22,69% chance that 3 customers will pay by credit card.

$$P(\text{at least } 1) = P(\geq 1) = 1 - P(0)$$

$$P(0) = \frac{7!}{0!(7-0)!} 0,3^0 (1 - 0,3)^7 = 0,0824 \text{ or } 8,24\%$$

There is a 0,0824 probability or 8,24% chance that at least 1 customer will pay by credit card.

5.2.2 The Poisson distribution

The Poisson distribution is a discrete distribution that measures the number of occurrences of an event when the average occurrences is given.

- The number of successes that occur in a specified interval is independent of the number of successes that occur in any other interval.
- The probability that a success will occur in an interval is the same for all intervals of equal size and is proportional to the size of the interval.
- x is the count of the number of successes that occur in a given interval and may take on any value from 0 to ∞ (infinity).
- The specified interval can be a period of time, space or volume.
- Examples: the number of defects in a length of fabric, the number of calls in a time period, the number of customers who enter a mall in a specified time period etc.

Formula: Poisson distribution probability

The probability of x occurrences in an interval of a Poisson experiment is:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Where

λ mean number of occurrences for the specified interval

e a mathematical constant, the base of the natural logarithm, 2,71828 (can be found as a function on a statistics calculator)

x the number of successful outcomes of the experiment

**TIP**

Ensure the λ (average occurrences) and the probability you're trying to establish are in the same measurement unit. For example, if you're trying to establish probability in an hour's timeframe and the λ is in a 30 minute timeframe, adjust the λ to be for an hour.

EXAMPLE

If a company receives an average of 3 calls per 5 minute period in a working day, the probability of receiving:

No calls during a randomly selected 5 minute is:

$$\lambda = 3 \text{ in } 5 \text{ minutes}$$

$$P(0) = \frac{3^0 e^{-3}}{0!} = 0,0498 \text{ or } 4,98\%$$

5 calls during the next 10 minutes is:

$$\lambda = 3 \text{ in } 5 \text{ minutes} = 6 \text{ in } 10 \text{ minutes}$$

$$P(5) = \frac{6^5 e^{-6}}{5!} = 0,1606 \text{ or } 16,06\%$$

At least 2 calls during the next 2.5 minutes is:

$$\lambda = 3 \text{ in } 5 \text{ minutes} = 1,5 \text{ in } 2,5 \text{ minutes}$$

$$P(\text{at least } 2) = P(\geq 2) = 1 - (P(0) + P(1))$$

$$P(0) = \frac{1,5^0 e^{-1,5}}{0!} = 0,2231$$

$$P(1) = \frac{1,5^1 e^{-1,5}}{1!} = 0,3347$$

$$P(\text{at least } 2) = P(\geq 2) = 1 - (P(0) + P(1)) = 1 - (0,2231 + 0,3347) = 0,4422 \text{ or } 44,22\%$$



SELF-ASSESSMENT ACTIVITY

A tollgate operator has observed that cars arrive randomly at an average of 60 cars per hour. Calculate the probability that:

- only 2 cars will arrive during a specified 1-minute period;
- At least 3 cars will arrive during a specified 2-minute period.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

$\lambda = 60$ in 60 minutes = 1 in 1 minute

$$P(2) = \frac{1^2 e^{-1}}{2!} = 0,1839 \text{ or } 18,39\%$$

There is a 0,1839 probability or 18,39% chance that only 2 cars will arrive in a 1-minute period.

$\lambda = 1$ in 1 minute = 2 in 2 minutes

$$P(\text{at least } 3) = P(\geq 3) = 1 - (P(0) + P(1) + P(2))$$

$$P(0) = \frac{2^0 e^{-2}}{0!} = 0,1353$$

$$P(1) = \frac{2^1 e^{-2}}{1!} = 0,2707$$

$$P(2) = \frac{2^2 e^{-2}}{2!} = 0,2707$$

$$P(\text{at least } 3) = P(\geq 3) = 1 - (P(0) + P(1) + P(2)) = 1 - (0,1353 + 0,2707 + 0,2707) = 0,3233 \text{ or } 32,33\%$$

There is a 0,3233 probability or 32,33% chance that at least 3 cars will arrive in a 2-minute period.

5.3 Continuous probability distributions

A continuous distribution is a distribution in which the random variable may assume any value within a given range or interval. The most frequently used continuous probability distribution is the normal distribution.

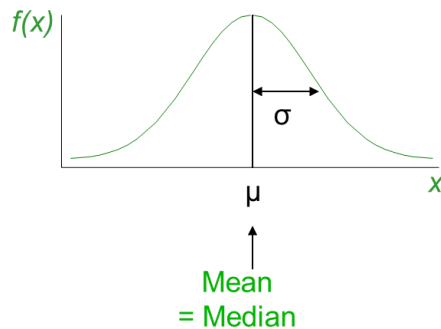
The scope of this module is limited to the normal distribution with respect to continuous probability distributions. You are encouraged to explore uniform and exponential continuous probabilities.

5.3.1 The Normal probability distribution

Normal probability distributions have the following characteristics:

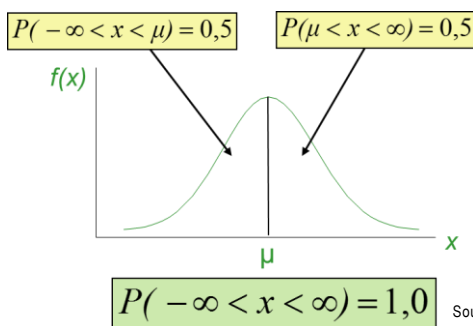
- The distribution is bell-shaped in appearance.
- The distribution is symmetrical around a central mean, μ .

The x-axis represents the possible values of the random variable, which are infinite.



- The left and right hand tails of the distribution approach the x-axis but never touch it. This means it is asymptotic, i.e. all values will have a non-zero probability.

- Two parameters are necessary to construct the normal distribution: the mean, μ and standard deviation σ . The centre of the distribution is determined by μ and the spread by σ .



- Probabilities for continuous variables correspond to areas under the normal curve.

- The total area under the normal curve is equal to 1 (or 100%). This means that the symmetric areas to the left and the right of the mean will each comprise 50% of the total area, each with a probability of occurring of 0,5.

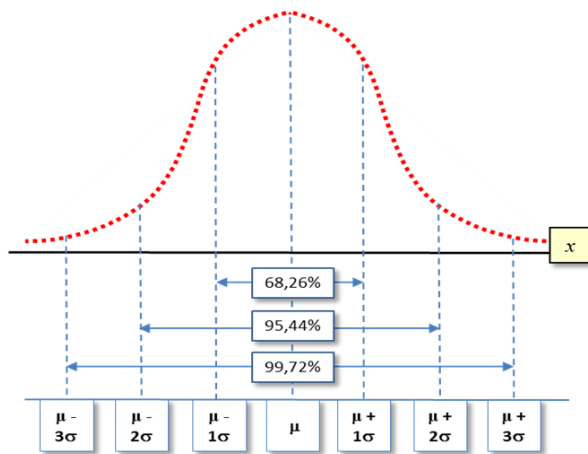
Source: Donnelly (2013)

- It is possible to convert any normal distribution into a standard form, z by expressing the difference between the value of interest x and the mean μ in units of standard deviation. This z will show the number of standard deviations that a particular value lies to the right or left of the mean. Any x -value greater than the mean will have a positive z -value. The areas under this standard normal distribution are contained in published tables.
- In the normal distribution table (Table 2.1 at the end of this module), a value of z , to the first decimal place, is found in the left-hand column; the second decimal place is located across the top row. The corresponding area will be from the middle of the curve, the μ , to a specified x -value.
- Probability is defined by the area under the curve, either from $-\infty$ to the specified x -value, from the specified x -value to $+\infty$, or between two x -values.

5.3.2 The standard normal distribution or z-curve

The standard normal distribution or z-curve has *mean, $\mu = 0$* and *standard deviation, $\sigma = 1$* and can be used as a 'template' for determining probabilities for variables in all normal distributions.

All normal distributions exhibit the same 'profile' with respect to the probabilities relating to multiples of the standard deviation:



By using a standardising formula, we can convert any normal distribution to the standard normal distribution in order to determine probabilities of the random variable. For convenience, probability tables for the z-curve are provided to the end of the module. Converting a normal distribution to the standard normal distribution tells us how many standard deviations away from the mean each value is.

Formula: normal distribution standardising formula

The standardising formula I used to convert an x value to a z -value for use with z -tables:

$$z = \frac{x - \mu}{\sigma}$$

Where:

z is the value of z corresponding to the value of x

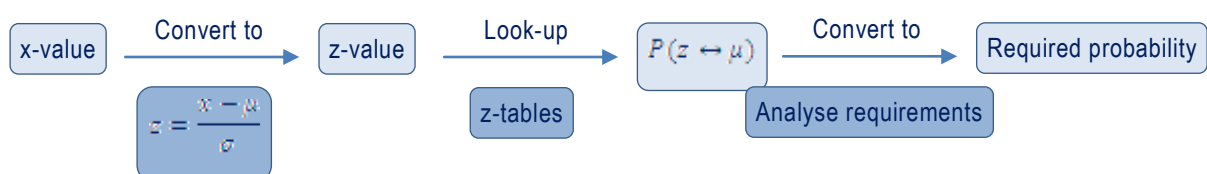
x is the value of the random variable of the normal distribution under consideration

μ is the mean of the normal distribution under consideration

σ is the standard deviation of the normal distribution under consideration

The steps for determining probabilities are:

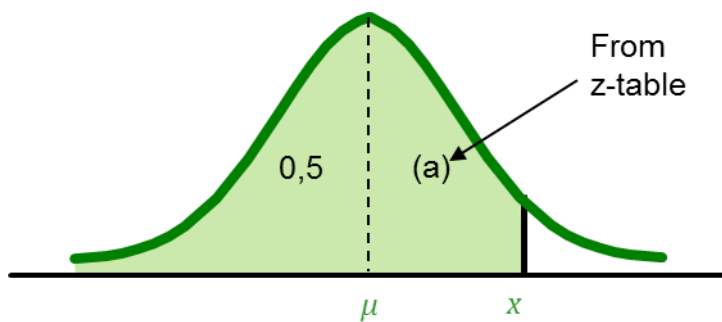
Determining probability from x -value:



For the last step to convert to the required probability, the analysis will depend on the requirements.

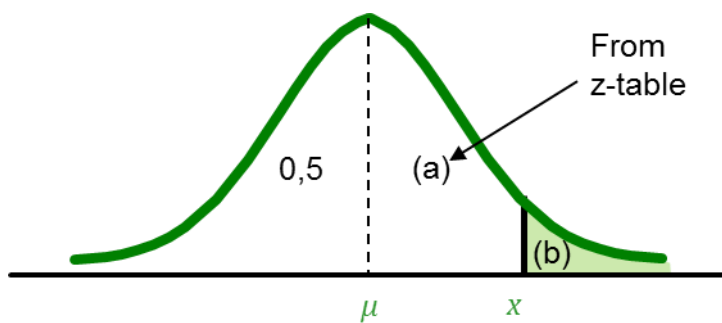
Important note – remember the z-tables provide the probability for the area between z and the mean:

To establish $P(< x)$, the z-table value (a), needs to be added to 0,5:



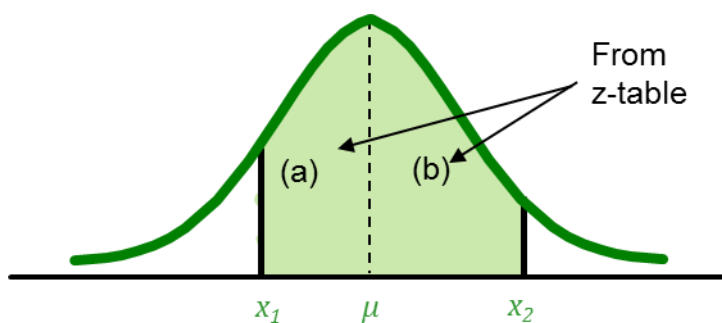
To establish $P(> x)$, the z-table value (a), needs to be subtracted from 0,5, in order to find (b),

because $(a) + (b) = 0,5$:



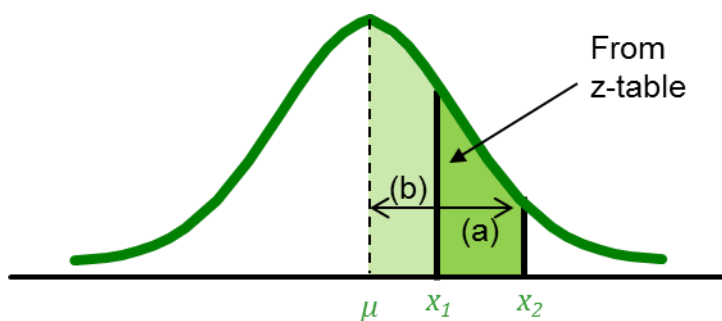
To establish $P(x_1 - x_2)$, if the two values are on either side of the mean, the z-table values (a) and (b)

for each respective x value need to be added together.

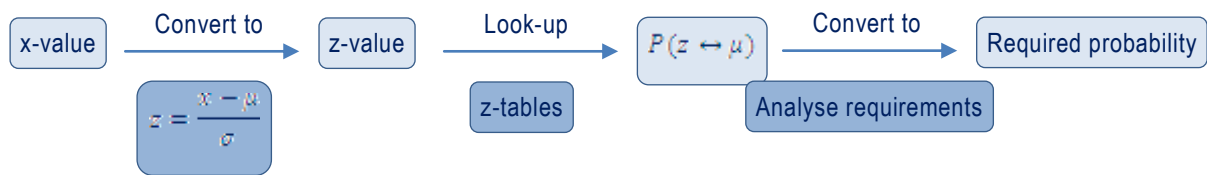


To establish $P(x_1 - x_2)$, if the two values are both on the same side of the mean, the smaller z-table

value (a) needs to be subtracted from the larger z-table value (b).



Reverse the steps to determine the x-value from the probability (reading from right to left in the diagram):



For the last step to convert to the required probability, the analysis will depend on the requirements. Important note – remember the z-tables provide the probability for the area between z and the mean:

EXAMPLE

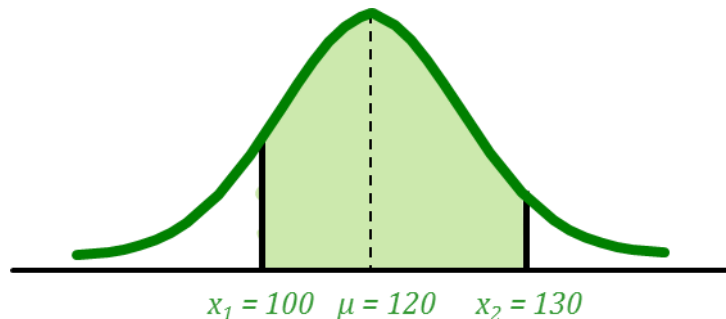
The time it takes a randomly selected job applicant to perform a certain task is normally distributed with a mean value of 120 seconds and a standard deviation of 20 seconds.

Determine the probability that:

- A randomly selected candidate will complete the task between 100 and 130 seconds.
- A randomly selected candidate will complete the task between 75 and 100 seconds.
- A randomly selected candidate will complete the task within 75 seconds.
- If the slowest 10% are to be given advanced training, what task times will qualify individuals for such training?

$$\mu = 120 \text{ seconds}, \sigma = 20 \text{ seconds}$$

- A randomly selected candidate will complete the task between 100 and 130 seconds.

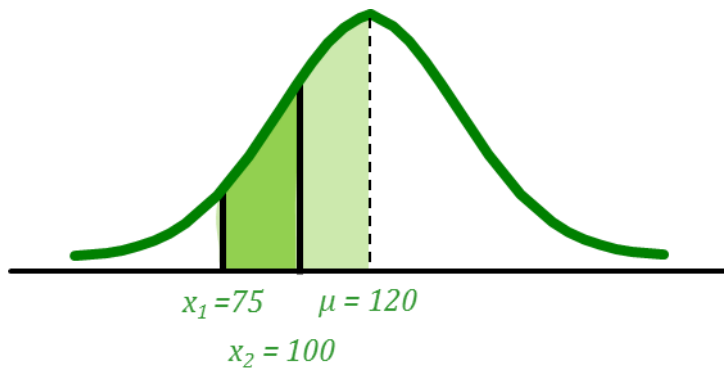


$$\text{for } x_1, z = \frac{100 - 120}{20} = -20; \text{ from } z \text{ table, area} = 0,3413$$

$$\text{for } x_2, z = \frac{130 - 120}{20} = 0,5; \text{ from } z \text{ table, area} = 0,1915$$

$$P(100 < x < 130) = 0,3413 + 0,1915 = 0,5328 \text{ or } 53,28\%$$

- b) A randomly selected candidate will complete the task between 75 and 100 seconds.

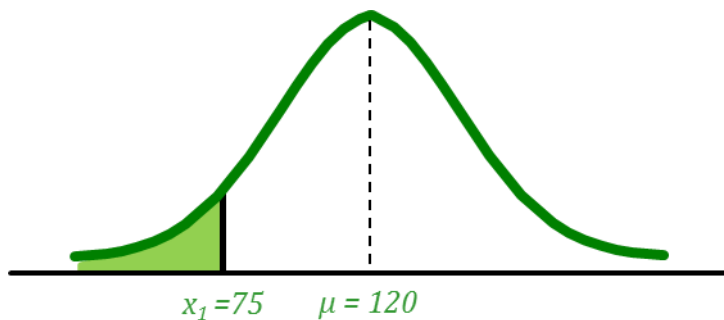


for $x_1, z = \frac{75 - 120}{20} = -2,25$; from z table, area = 0,4878

for x_2 , area = 0,3413 (from previous question)

$P(75 < x < 100) = 0,4878 - 0,3413 = 0,1465$ or 14,65%

- c) A randomly selected candidate will complete the task within 75 seconds.

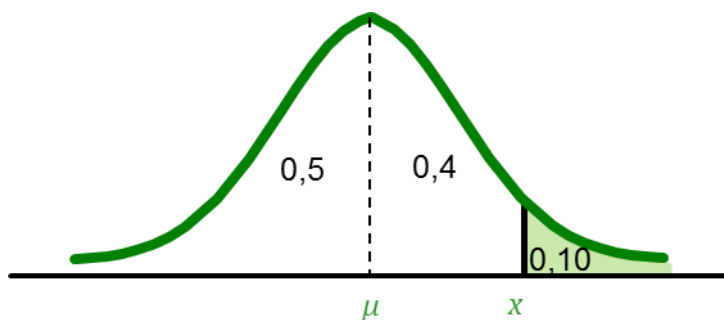


for x_1 , area = 0,4878 (from previous question)

$P(x < 75) = 0,5 - 0,4878 = 0,0122$ or 1,22%

- d) If the slowest 10% are to be given advanced training, what task times will qualify individuals for such training?

Note: be careful of a question like this. Where will the slowest 10% lie? We're talking about task times, so the slowest job applicants will take the longest, falling in other right hand tail of the distribution.



Looking up 0, 4 in the z-table, the nearest z-value is 1, 28. We can now calculate x from the formula:

$$z = \frac{x - \mu}{\sigma} \therefore x = \sigma z + \mu = 20 \times 1,28 + 120 = 145,6 \text{ seconds}$$



TIP

It's useful to sketch a diagram of the distribution and the x-values before calculation. That way you can visually establish which side of the mean each of the values is and what probabilities need to be calculated.

Unit 5 Exercises: (Solutions are found at the end of the module guide)

Exercise 5.1

A telephone salesman has established that 10% of his calls lead to a sale and that each call is independent of all other calls. Calculate the probability that he:

- a) Makes no sales in 12 calls.
- b) Makes fewer than 3 sales in 15 calls.

Exercise 5.2

Bases on past experience, 10% of the accounts of a large wholesale company are incorrect. If a random sample of 3 accounts is selected, what is the probability that:

- a) Exactly one account is incorrect?
- b) At least one account is incorrect?

Exercise 5.3

A company which supplies ready-mix concrete receives, on average, 6 orders per day.

- a) What is the probability that, on a given day:
 - (i) No orders are received?
 - (ii) No more than 2 orders are received?
 - (iii) At least 3 orders are received?
- b) What is the probability that, on a given half-day, no orders are received?
- c) What is the mean and standard deviation of orders received per day?

Exercise 5.4

The average number of calls coming into a switchboard during the busiest part of the day for a small firm is 5 calls per minute. If the number of incoming calls follows a Poisson distribution, what is the probability that for any given minute there will be exactly 2 calls?

Exercise 5.5

The average number of claims per hour made to a certain insurance company is 1, 2. What is the probability that in any given hour either two or three claims will be received?

Exercise 5.6

Given that x follows a normal distribution with a mean (μ) of 64 and a standard deviation (σ) of 0, 5 find:

- a) $P(x < 63)$
- b) $P(x > 63,7)$
- c) $P(62,9 < x < 64,3)$
- d) $P(x >?) = 0,1026$

e) $P(x >?) = 0,9772$

Exercise 5.7

Find the following probabilities for the standard normal distribution, z:

- a) $P(z > 1,5)$
- b) $P(z < -0,68)$
- c) $P(0 < z < 1,5)$

Exercise 5.8

Customers are known to arrive at a muffler shop on a random basis, with an average of 2 customers per hour arriving at the facility.

- a) What is the probability that more than 3 customers will require service during a particular hour?
- b) What is the probability that fewer than 4 customers will require service during a 4 hour period in the morning on a particular day?

UNIT 6
HYPOTHESIS TESTING

UNIT 6: HYPOTHESIS TESTING

OBJECTIVES

By the end of this study unit, you should be able to:

1. Describe and use the steps in a general hypothesis-testing procedure.
2. Distinguish between a one-tailed and a two-tailed test.
3. Conduct tests of hypothesis concerning values of the following parameters.
 - Population mean (large and small samples).
 - Population proportion.
 - The difference between 2 population means.
 - The difference between 2 population proportions.
4. Conduct tests of hypothesis using the chi-square distribution.
5. Perform independence of association hypothesis tests.
6. Perform equality of multiple proportions hypothesis tests.
7. Perform goodness-of-fit hypothesis tests.
8. Interpret the results of hypothesis tests.

CONTENT

6.1 Introduction

6.2 Hypothesis testing concepts

6.2.1. The Null hypothesis (H_0) and the alternative hypothesis (H_1)

6.2.2. Rejection region and significance level

6.3 Steps in hypothesis testing

6.4 The t-distribution

6.5 Chi-squared hypothesis tests – hypothesis testing for multiple comparisons

6.5.1. Test of independence between variables

6.5.2. Test for the difference between proportions

6.5.3. Goodness-of-fit test

6.6 The logic behind the expected frequency calculation

Exercises



Prescribed textbook: Chapters 8, 9 and 10

6.1 Introduction

Summary or descriptive statistics use various methods to summarise/describe sample data collected in order to facilitate interpretation. Summary statistics has been covered in sections 1 to 5 of this study guide.

Inferential statistics allow us to use the same sample data to infer features of the population from which it is drawn. The most important applications of inferential statistics are confidence interval estimating and hypothesis testing.

Confidence interval testing is not covered in this course.

Hypothesis testing uses sample evidence to statistically test whether a claim made about a population is valid. The results of the sample are used to make an inference about the population as a whole.

Example of a claim: a product manufacturer claims more people use their product than any other.

The hypotheses covered in the course relate to:

- A single population mean, μ , e.g. on average 200 customers pass a store every hour.
- A single population proportion, π , e.g. 4 out of 10 students study part-time.
- A comparison or the difference between two population means, $\mu_1 - \mu_2$, e.g. on average people under 25 spend more on clothing than those older than 50.
- A comparison or the difference between two population proportions, $\pi_1 - \pi_2$, e.g. 60% of Durban students are under 35 whereas only 40% of Johannesburg students fall into that age category.
- Multiple comparisons, χ^2 , e.g. the accident profiles of three market segments are all the same.

6.2 Hypothesis testing concepts

In this section, we will present the basic concepts of hypothesis testing. The next section takes you through the step by step process of testing a hypothesis.

6.2.1 The null hypothesis (H_0) and the alternative hypothesis (H_1)

For every hypothesis statement there needs to be an opposite statement, i.e. claim and counterclaim, e.g.:

Statement: "The defendant is guilty"; alternative: "The defendant is not guilty".

The first statement we refer to as the null hypothesis (H_0).

The second statement we refer to as the alternative hypothesis (H_1).

Important note - when formulating hypothesis statements:

- The null hypothesis always contains an equality sign (= or \geq or \leq).
- The alternative hypothesis always contains an inequality sign counter to that of the null hypothesis (< or > or \neq).

EXAMPLE

Statement: Our sales average R1 million per month.

Null hypothesis: $H_0: \mu = R1 \text{ million}$

Alternative hypothesis: $H_1: \mu \neq R1 \text{ million or } \mu <> R1 \text{ million}$

6.2.2 Rejection region and significance level

Except in the case of multiple comparisons, a normal distribution is used to assess whether the null hypothesis is accepted or rejected.

The two possible errors encountered in hypothesis testing are:

A **type I error** where a valid null hypothesis is rejected, with probability of α .

A **type II error** where a false null hypothesis is accepted, with probability of β .

The error probabilities of α and β are inversely related which simply means that any attempt to reduce the one will increase the other one (Keller: 2005, p326).

Reducing the probability of one type of error results in the probability of the other error type increasing.

The limits of acceptance and rejection are set (using the tails of the normal distribution) based on a level of significance, α (the probability of a type I error).

These limits, or cut-off points, reflect the level of risk considered acceptable in drawing a wrong conclusion.

The level of significance defines the likelihood of rejecting the null hypothesis when in fact it is true (significance levels are related to confidence levels used in probability sampling).

If the statistical test used to test the hypothesis finds in favour of the alternative hypothesis, we reject the null hypothesis. If not, the null hypothesis is not rejected.

Note: we **never** say that we accept the null hypothesis. We either reject the null hypothesis in favour of the alternative or not.

6.3 Steps in hypothesis testing

Once the type of hypothesis test has been identified, five process steps are performed:

- Formulate the null and alternative hypotheses.
- Decide on the level of significance (usually provided in example questions).

- Determine the rejection region and formulate the rejection rule. This step includes establishing the critical value about which the hypothesis value is tested.
- Calculate the sample test statistic.
- Apply the rejection rule to the sample test statistic.
- Draw a conclusion, both statistically and from a management perspective (ie in English).

Each step is now examined further.

Step 1. Formulate the null and alternative hypotheses

The null and alternative hypothesis are formulated from the management question.

Important note: the null hypothesis is not always the claim; the hypotheses may be formulated in such a way that the alternative hypothesis is the claim, in which case you need to be very careful when drawing conclusions at the end of the test.

The null and therefore alternative hypotheses are formulated based on the identified approach:

Approach	Null hypothesis formats	Alternative hypothesis formats
Single population mean	$H_0: \mu =$	$H_1: \mu \neq$
	$H_0: \mu \leq$	$H_1: \mu >$
	$H_0: \mu \geq$	$H_1: \mu <$
Single population proportion	$H_0: \pi =$	$H_1: \pi \neq$
	$H_0: \pi \leq$	$H_1: \pi >$
	$H_0: \pi \geq$	$H_1: \pi <$
Comparison of two population means	$H_0: \mu_1 = \mu_2$ or $H_0: \mu_1 - \mu_2 = 0$ or a specified value	$H_1: \mu_1 \neq \mu_2$ or $H_1: \mu_1 - \mu_2 \neq 0$ or a specified value
	$H_0: \mu_1 \leq \mu_2$	$H_1: \mu_1 > \mu_2$
	$H_0: \mu_1 \geq \mu_2$	$H_1: \mu_1 < \mu_2$
Comparison of two population proportions	$H_0: \pi_1 = \pi_2$ or $H_0: \pi_1 - \pi_2 = 0$ or a specified value	$H_1: \pi_1 \neq \pi_2$ or $H_1: \pi_1 - \pi_2 \neq 0$ or a specified value
	$H_0: \pi_1 \leq \pi_2$	$H_1: \pi_1 > \pi_2$
	$H_0: \pi_1 \geq \pi_2$	$H_1: \pi_1 < \pi_2$
Multiple comparisons – see section 6.5	$H_0: \pi_1 = \pi_2 = \pi_3 = \pi_4$ etc or H_0 : The groups being compared all have the same proportions	$H_1: \pi_1 \neq \pi_2 \neq \pi_3 \neq \pi_4$ etc or H_1 : The groups being compared have different proportions

EXAMPLE

Management question: A store owner believes that on average at least 200 people pass by his store every hour.

$$H_0: \mu \geq 200$$

$$H_1: \mu < 200$$

Management question: A store owner believes that fewer than 4 out of 10 of his customers are female.

$$H_0: \pi \leq 0,4$$

Where π is the proportion of female customers.

$$H_1: \pi < 200$$

Management question: A store owner believes that more customers enter his store than his neighbouring store.

$$H_0: \mu_1 \leq \mu_2$$

Where μ_1 is the average number of customers entering the store and μ_2 is the average number of customers entering the neighbouring store.

$$H_1: \mu_1 > \mu_2$$

Note: in this case the store owner's claim becomes the alternative hypothesis, because it doesn't contain an equality sign.

Management question: A store owner believes that he has fewer female customers than his neighbouring store.

$$H_0: \pi_1 \geq \pi_2$$

Where π_1 is the proportion of female customers and π_2 is the proportion of the neighbouring store's female customers.

$$H_1: \pi_1 < \pi_2$$

Note: again the store owner's claim becomes the alternative hypothesis, because it doesn't contain an equality sign.

Management question: A store owner believes that customers in each of three different age categories have different spending patterns across a group of products.

$$H_0: \pi_1 = \pi_2 = \pi_3$$

Where π_1 is the spending proportion displayed by the first age category, π_2 is the spending proportion displayed by the second age category and π_3 is the spending proportion displayed by the third age category.

$$H_1: \pi_1 \neq \pi_2 \neq \pi_3$$

In the case of multiple proportions it is often easier (and easier to understand) to state the hypotheses in English:

H_0 : The age categories all have the same spending patterns

H_1 : The age categories have different spending patterns

Note, an English hypothesis statement is only permissible for multiple comparisons.

Step 2. Decide on the level of significance (usually provided in example questions)

The probability that the test statistic falls into the rejection region (ie a type I error) is given by the level of significance, α .



TIP

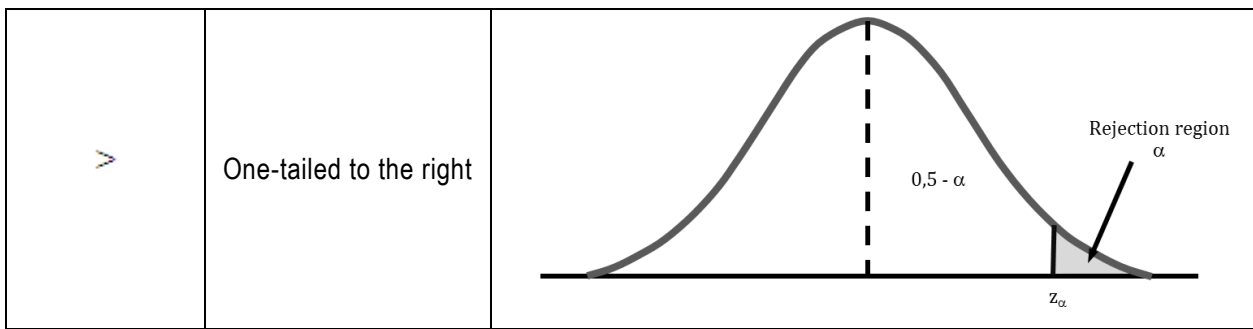
The level of significance is usually given as a percentage, e.g. 5%. A good idea is to immediately change this to a probability, e.g. 0, 05, so as to make it easier to calculate or look up critical values.

Step 3. Determine the rejection region and formulate the rejection rule

(This step includes establishing the critical value about which the hypothesis value is tested.)

The rejection region is formulated from the alternative hypothesis (if the alternative hypothesis holds true, then the null hypothesis is rejected), using the level of significance.

Alternative hypothesis sign	Type of test	Rejection region
$<>, ie \neq$	Two-tailed Important note: the level of significance is divided between the two tails	
$<$	One-tailed to the left	



The critical z-value is determined from the standard normal tables, e.g. if the test is one-tailed to the right and $\alpha = 0,05$:

$$P(0 - z) = 0,45 \text{ therefore } z = 1,645$$

The **rejection rule** is the decision rule for rejecting the null hypothesis and is formulated in line with the rejection region. The rejection rule will read as:

Reject H_0 if $z_{calc} > 1,645$ where z_{calc} is the calculated sample test statistic

Step 4. Calculate the sample test statistic

The sample test statistic is calculated using the sample data gathered. The formula used depends on the hypothesis approach:

Approach	Formula
Single population mean	$z_{calc} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
Single population proportion	$z_{calc} = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$
Comparison of two population means	$z_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Comparison of two population proportions	$z_{calc} = \frac{p_a - p_b}{\sqrt{(p \times q) \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}}$ <p>where:</p> $p = \frac{n_a p_a + n_b p_b}{n_a + n_b}$ $q = 1 - p$

Approach	Formula
Multiple comparisons – see section 6.5	$\chi^2_{calc} = \sum \frac{(f_o - f_e)^2}{f_e}$

**TIP**

The numerator of the hypothesis test statistic formula aligns with the hypothesis approach and the null hypothesis.

Step 5. Apply the rejection rule to the sample test statistic

Once the test statistic has been calculated, the rejection rule can be applied to see whether the test statistic value relative to the critical value will result in the null hypothesis being rejected or not.

Step 6. Draw a conclusion, both statistically and from a management perspective (i.e. in English)

The results of the test are interpreted and a conclusion relating to the claim is drawn.

EXAMPLE

A statistical analyst who works for a large insurance company is in the process of examining several pension plans. Because the length of life of pension plan holders is critical to the plans' integrity, the analyst needs to know if the average age has changed. In the last census (2011), suppose the average age of retirees is 67, 5 years. To determine whether the average age has increased, the analyst selects a random sample of 100 retirees and finds that $\bar{x} = 68,2$. If we assume that the population standard deviation is $\sigma = 3,1$ can we conclude at a 5% level of significance that the average of retirees has increased since 2011?

The question asks if there is sufficient evidence to conclude that μ (the mean age at present) is greater than the mean age in 2011 (67, 5).

Step 1. Formulate the null and alternative hypotheses

We are dealing with a single population mean.

$$H_0: \mu \leq 67,5$$

$$H_1: \mu > 67,5$$

Step 2. Decide on the level of significance

$$\alpha = 0,05$$

Step 3. Determine the rejection region and formulate the rejection rule

This is a one-tailed test to the right with $z_{crit} = 1,645$

Reject H_0 if $z_{calc} > 1,645$

Step 4. Calculate the sample test statistic

$$z_{calc} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{68,2 - 67,5}{\frac{3,1}{\sqrt{100}}} = 2,26$$

Step 5. Apply the rejection rule to the sample test statistic.

2,26 is $> 1,645$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that the average age is greater than that of 2011 (at a 5% level of significance).



SELF-ASSESSMENT ACTIVITY

It is important for airlines to know the approximate total weight of baggage carried on each plane. An airline researcher believes that the average baggage weight for each adult is 60 kg. To test his belief, he draws a random sample of 50 adult passengers and weighs their baggage. He finds the sample mean to be 57,1 kg. If he knows that the population standard deviation is 10 kg, can he conclude at a 5% significance level that his belief is incorrect?

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Step 1. Formulate the null and alternative hypotheses

We are dealing with a single population mean.

$$H_0: \mu = 60$$

$$H_1: \mu \neq 60$$

Step 2. Decide on the level of significance

$$\alpha = 0,05$$

Step 3. Determine the rejection region and formulate the rejection rule

This is a two-tailed test with $z_{crit} = -1,96$ and $+1,96$

Reject H_0 if $z_{calc} < -1,96$ or $> 1,96$

or

Reject H_0 if $-1,96 > z_{calc} > 1,96$

Step 4. Calculate the sample test statistic

$$z_{calc} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{57,1 - 60}{\frac{10}{\sqrt{50}}} = -2,05$$

Step 5. Apply the rejection rule to the sample test statistic.

$$-2,05 \text{ is } < -1,96$$

Step 6. Draw a conclusion, both statistically and from a management perspective (i.e. in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that the average baggage weight is not 60 kgs (at a 5% level of significance).

6.4 The t-distribution

For comparisons involving a population mean or the comparison of two population means where the population standard deviation is unknown, i.e. the sample size is small $n \leq 30$ or the sample sizes together are small $n_1 + n_2 \leq 30$, the t-distribution using t_{crit} and t_{calc} is used. (Theoretically the t-distribution should be used any time the population standard deviation is unknown, but the use is usually limited to the smaller sample sizes, because the z-distribution offers a good approximation for larger sample sizes).

Using the t-distribution affects two steps of the hypothesis testing process:

Step 7. Determine the rejection region and formulate the rejection rule

From the t-distribution, the critical value is determined using the level of significance, α , and the degrees of freedom, df , where for a single population mean,

$$df = n - 1$$

And for the comparison of two population means,

$$(n_1 - 1) + (n_2 - 1)$$

The value of t_{crit} can then be read from the t-table (see appendix 2)

Step 8. Calculate the sample test statistic

To calculate the test statistic for the t-distribution, the formula for a single population mean is:

$$t_{calc} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

The formula for the comparison of two means is:

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

EXAMPLE

The desired percentage of silicon dioxide in a certain type of cement is 5. A random sample of 27 specimens give a sample average percentage of 5,21% and a sample standard deviation of 0,38%. Use a 1% level of significance and test whether the sample result indicates a change in the percentage average.

Step 1. Formulate the null and alternative hypotheses

We are dealing with a single population mean with a sample size smaller than 30.

$$H_0: \mu = 5$$

$$H_1: \mu \neq 5$$

Step 2. Decide on the level of significance

$$\alpha = 0,01 \text{ and } \frac{\alpha}{2} = 0,005$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = 27 - 1 = 26$$

This is a two-tailed test with $t_{crit} = -2,78 \text{ and } +2,78$

Reject H_0 if $t_{calc} < -2,78 \text{ or } > 2,78$

or

Reject H_0 if $-2,78 > t_{calc} > 2,78$

Step 4. Calculate the sample test statistic

$$t_{calc} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{5,21 - 5}{\frac{0,38}{\sqrt{26}}} = 2,82$$

Step 5. Apply the rejection rule to the sample test statistic.

2,82 is $> 2,78$

Step 6. Draw a conclusion, both statistically and from a management perspective (i.e. in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that there is a change in the average percentage of silicone dioxide in a certain type of cement (at a 1% level of significance).



SELF-ASSESSMENT ACTIVITY

A machine is supposed to be adjusted to produce components to a dimension of 2,0 cm. In a sample of 50 components, the mean is found to be 2,001 cm and the standard deviation 0,003 cm. Is there evidence to suggest that the machine is set too high? Use a 5% level of significance.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Step 1. Formulate the null and alternative hypotheses

We are dealing with a single population mean with a sample size greater than 30.

$$H_0: \mu = 2,0$$

$$H_1: \mu > 2,0$$

Step 2. Decide on the level of significance

$$\alpha = 0,05$$

Step 3. Determine the rejection region and formulate the rejection rule

This is a one-tailed test to the right with $z_{crit} = 1,645$

$$\text{Reject } H_0 \text{ if } z_{calc} > 1,645$$

Step 4. Calculate the sample test statistic

$$z_{calc} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2,001 - 2,0}{\frac{0,003}{\sqrt{50}}} = 2,36$$

Step 5. Apply the rejection rule to the sample test statistic.

$$2,36 \text{ is } > 1,645$$

Step 6. Draw a conclusion, both statistically and from a management perspective (i.e. in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that the machine is set too high (at a 5% level of significance).

EXAMPLE

The management of a mine wishes to investigate the effect of a 4-day workweek on absenteeism. 2 random samples each of size 40 are selected. Employees of group 1 work 10-hour days (4-day week) and group 2 work 8-hour days (5-day week). If group 1 averages 4 hours of absenteeism per week, with a standard deviation of 1, 2 and group 2 averages 4, 4 hours of absenteeism per week, with a standard deviation of 1, 5, should we conclude that the shorter workweek reduces absenteeism at a 5% level of significance?

Step 1. Formulate the null and alternative hypotheses

We are dealing with the comparison of two population means which together have sample size greater than 30.

$$H_0: \mu_1 \geq \mu_2$$

$$H_1: \mu_1 < \mu_2$$

Where μ_1 is the mean for group 1 and μ_2 is the mean for group 2.

Step 2. Decide on the level of significance

$$\alpha = 0,05$$

Step 3. Determine the rejection region and formulate the rejection rule

This is a one-tailed test to the left with $z_{crit} = -1,645$

$$\text{Reject } H_0 \text{ if } z_{calc} < -1,645$$

Step 4. Calculate the sample test statistic

$$z_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{4 - 4,4}{\sqrt{\frac{1,2^2}{40} + \frac{1,5^2}{40}}} = -1,32$$

Step 5. Apply the rejection rule to the sample test statistic.

$$-1,32 \text{ is not } < -1,645$$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is not rejected in favour of the alternative hypothesis.

The conclusion then is that the shorter week does not reduce absenteeism (at a 5% level of significance).



SELF-ASSESSMENT ACTIVITY

In order to determine whether there is a difference in the performance of 2 training methods, samples of individuals from each of the methods are checked. For the 6 individuals from method 1, the mean efficiency score is 35, with a standard deviation of 6. For the 8 individuals from method 2, the mean efficiency score is 27, with a standard deviation of 7. Use a 2% level of significance.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Step 1. Formulate the null and alternative hypotheses

We are dealing with the comparison of two population means which together have sample size of less than 30.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Where μ_1 is the mean for method 1 and μ_2 is the mean for method 2.

Step 2. Decide on the level of significance

$$\alpha = 0,02 \text{ and } \frac{\alpha}{2} = 0,01$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = (6 - 1) + (8 - 1) = 12$$

This is a two-tailed test with $t_{crit} = -2,68$

Reject H_0 if $t_{calc} < -2,68$ or $> 2,68$

Step 4. Calculate the sample test statistic

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{35 - 27}{\sqrt{\frac{(6 - 1)6^2 + (8 - 1)7^2}{6 + 8 - 2} \left(\frac{1}{6} + \frac{1}{8}\right)}} = 2,24$$

Step 5. Apply the rejection rule to the sample test statistic.

2,24 is not $> 2,68$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is not rejected in favour of the alternative hypothesis.

The conclusion then is that there is no difference between the two training methods (at a 2% level of significance).



SELF-ASSESSMENT ACTIVITY

Workers in 2 different mine groups are asked what they consider to be the most important labour-management problem. In group A, 200 out of a random sample of 400 workers feel that fair adjustment of grievances is the most important problem. In group B, 60 out of a random sample of 100 workers feel that this is the most important problem. Would you conclude that these 2 groups differ with respect to the proportion of workers who believe that fair adjustment of grievances is the most important problem? Test at a 10% level of significance.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Step 1. Formulate the null and alternative hypotheses

We are dealing with the comparison of two population proportions.

$$H_0: \pi_a = \pi_b$$

$$H_1: \pi_a \neq \pi_b$$

Where π_a is the mean for group A and π_b is the mean for group B.

Step 2. Decide on the level of significance

$$\alpha = 0,10 \text{ and } \frac{\alpha}{2} = 0,05$$

Step 3. Determine the rejection region and formulate the rejection rule

This is a two-tailed test with $z_{crit} = -1,645 \text{ and } 1,645$

Reject H_0 if $z_{calc} < -1,645 \text{ or } > 1,645$

Step 4. Calculate the sample test statistic

$$z_{calc} = \frac{p_a - p_b}{\sqrt{(p \times q) \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}}$$

where:

$$p = \frac{n_a p_a + n_b p_b}{n_a + n_b}$$

$$q = 1 - p$$

Calculating:

$$p_a = \frac{200}{400} = 0,5 \text{ and } p_b = \frac{60}{100} = 0,6$$

$$p = \frac{n_a p_a + n_b p_b}{n_a + n_b} = \frac{400 \times 0,5 + 100 \times 0,6}{400 + 100} = 0,52$$

$$q = 1 - p = 1 - 0,52 = 0,48$$

$$z_{calc} = \frac{p_a - p_b}{\sqrt{(p \times q) \left(\frac{1}{n_a} + \frac{1}{n_b} \right)}} = \frac{0,5 - 0,6}{\sqrt{(0,52 \times 0,48) \left(\frac{1}{400} + \frac{1}{100} \right)}} = -1,79$$

Step 5. Apply the rejection rule to the sample test statistic.

$$-1,79 \text{ is } < -1,645$$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that the two groups do differ in their beliefs (at a 5% level of significance).

6.5 Chi-squared hypothesis tests – hypothesis testing for multiple comparisons

The z and t hypothesis tests use measures of central location. Chi-squared hypothesis tests test for patterns of outcomes using frequency counts where observed frequencies are compared to expected frequencies.

The chi-squared test is used to test for:

- The dependence or independence of association between two categorical variables, eg are spending amounts related to gender?

- Comparisons of proportions across more than two populations.
- Conformance by data to a pattern.

The same hypothesis testing steps are used as in other hypothesis testing with the following differences:

Step 3. Formulate the null and alternative hypotheses

When determining associating between variables, the chi-squared hypotheses can be stated in English rather than using a formula.

Step 3. Determine the rejection region and formulate the rejection rule

The chi-squared hypothesis test is formulated using a contingency table.

From the chi-squared distribution, the critical value is determined using the level of significance, α , and the degrees of freedom, df , calculated as

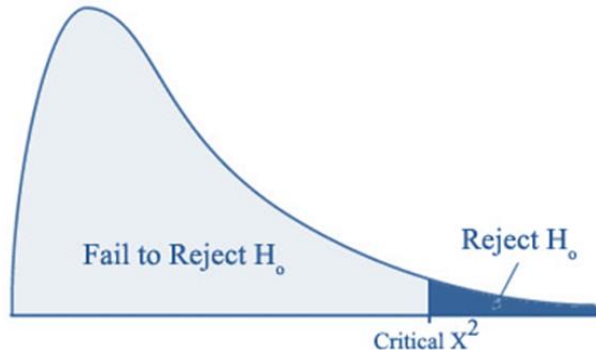
$$df = (r - 1)(c - 1)$$

Where

r = number of rows in the contingency table

c = number of columns in the contingency table

The value of χ^2_{crit} can then be read from the chi-squared table (see appendix 3).



Step 4. Calculate the sample test statistic

To calculate the test statistic for the chi-squared distribution, the formula is:

$$\chi^2_{calc} = \sum \frac{(f_o - f_e)^2}{f_e}$$

The sample statistic elements are calculated from the contingency table comprising sample values. The observed frequencies, f_o are the actual sample values. The expected frequencies, f_e are the frequencies that apply if each sample follows the same probability as the total and are calculated as:

$$f_e = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

The examples in each section provide step by step instructions. Work through each example carefully to aid your understanding of the process.

6.5.1 Test of independence between variables

It is often important to determine whether relationships exist between different variables or whether the variables may be considered independent of each other.

EXAMPLE

A random sample of adults is selected from each of 4 ethnic groups in Cape Town. Respondents are asked to specify their primary source of news. The results are:

News source	Ethnic group				Total
	A	B	C	D	
TV	30	20	25	20	95
Radio	25	25	20	20	90
Newspaper	10	10	5	30	55
Total	65	55	50	70	240

Is there a relationship between ethnic group and source of news, at a 2, 5% level of significance?

Step 1. Formulate the null and alternative hypotheses

We are dealing with multiple comparisons – a test of relationship.

H_0 : There is no relationship between ethnic group and source of news

H_1 : There is a relationship between ethnic group and source of news

Important note: For a chi-squared hypothesis test of association, words are permissible for the null and alternative hypotheses. Note, however, that the hypotheses still follow the 'rules'. The null hypothesis is still one of 'equality', indicating there is no difference between the groups; they are all the same or equal with respect to their source of news.

Step 2. Decide on the level of significance

$$\alpha = 0,025$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = (r - 1)(c - 1) = (3 - 1)(4 - 1) = 6$$

Important note: remember for the degrees of freedom to only include the data rows and columns and not the total rows and columns.

Using the degrees of freedom and the level of significance, look up the critical value in the chi-squared table and formulate the rejection rule:

$$\chi_{crit}^2 = 14,45$$

Reject H_0 if $\chi_{calc}^2 > 14,45$

Important note: For the chi-squared hypothesis test, the rejection rule is always $>$. The calculated statistic calculates the extent to which the samples are different from each other. If the test statistic is sufficiently different from expectation, i.e. it exceeds the critical value, the null hypothesis is rejected.

Step 4. Calculate the sample test statistic

$$\chi^2_{\text{calc}} = \sum \frac{(f_o - f_e)^2}{f_e}$$

Each element of the formula needs to be in a column in order to calculate the required \sum value:

Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
30	$f_e = \frac{95 \times 65}{240} = 25,7292$	4,2708	18,2397	0,7089
25	$f_e = \frac{90 \times 65}{240} = 24,375$	0,625	0,3906	0,016
10	$f_e = \frac{55 \times 65}{240} = 14,8958$	-4,8958	23,9689	1,6091
20	$f_e = \frac{95 \times 55}{240} = 21,7708$	-1,7708	3,1357	0,144
25	$f_e = \frac{90 \times 55}{240} = 20,625$	4,375	19,1406	0,928
10	$f_e = \frac{55 \times 55}{240} = 12,6042$	-2,6042	6,7819	0,5381
25	$f_e = \frac{95 \times 50}{240} = 19,7917$	5,2083	27,1264	1,3706
20	$f_e = \frac{90 \times 50}{240} = 18,75$	1,25	1,5625	0,0833
5	$f_e = \frac{55 \times 50}{240} = 11,4583$	-6,4583	41,7096	3,6401
20	$f_e = \frac{95 \times 70}{240} = 27,7083$	-7,7083	59,4179	2,1444
20	$f_e = \frac{90 \times 70}{240} = 26,25$	-6,25	39,0625	1,4881
30	$f_e = \frac{55 \times 70}{240} = 16,0417$	13,9583	194,8341	12,1455
Σ	240			24,8161

The **observed frequency column** lists the actual values of the samples. These can be in any sequence from the original contingency table. In this case, the first ethnic group is listed, followed by the second etc.

The **expected frequency column** applies the proportions of all groups to each group (the point is to establish whether these proportions hold true). The formula for the expected frequency is:

$$f_e = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

**TIP**

A good sense test is to ensure that the totals for the observed and expected frequency columns are equal.

Step 5. Apply the rejection rule to the sample test statistic.

$$24,82 \text{ is } > 14,45$$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that there is a relationship between ethnic group and news source (at a 2,5% level of significance).

**SELF-ASSESSMENT ACTIVITY**

A manufacturer of women's clothing is interested to know if age is a factor in whether women would buy a particular garment depending on its quality. A researcher samples 3 age groups and each woman is asked to rate the garment as excellent, average or poor. Test the hypothesis, at a 5% level of significance, that rating is not related to age group.

Rating	Age group		
	15 – 20	21 – 30	31 – 60
Excellent	40	47	46
Average	51	74	57
Poor	29	19	37

SELF-ASSESSMENT ACTIVITY SOLUTION

Step 1. Formulate the null and alternative hypotheses

We are dealing with multiple comparisons – a test of relationship.

H_0 : *There is no relationship between age group and quality assessment*

H_1 : *There is a relationship between age group and quality assessment*

Step 2. Decide on the level of significance

$$\alpha = 0,05$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = (r - 1)(c - 1) = (3 - 1)(3 - 1) = 4$$

$$\chi_{crit}^2 = 9,49$$

Reject H_0 if $\chi_{calc}^2 > 9,49$

Step 4. Calculate the sample test statistic

In this case, the totals for the contingency table need to first be calculated:

Rating	Age group			Totals
	15 – 20	21 – 30	31 – 60	
Excellent	40	47	46	133
Average	51	74	57	182
Poor	29	19	37	85
Totals	120	140	140	400

Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
40	39,9	0,1	0,01	0,0003
51	54,6	-3,6	12,96	0,2374
29	25,5	3,5	12,25	0,4804
47	46,55	0,45	0,2025	0,0044
74	63,7	10,3	106,09	1,6655
19	29,75	-10,75	115,5625	3,8845
46	46,55	-0,55	0,3025	0,0065
57	63,7	-6,7	44,89	0,7047
37	29,75	7,25	52,5625	1,7668
Σ	400			8,7505

Step 5. Apply the rejection rule to the sample test statistic.

$$8,75 \text{ is } < 9,49$$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is not rejected in favour of the alternative hypothesis.

The conclusion then is that there is a no relationship between age group and quality assessment (at a 5% level of significance).

6.5.2 Test for the difference between proportions

The chi-squared hypothesis test is also used for testing whether two or more proportions are statistically equal.

EXAMPLE

Consider the process of assembling television sets. Management may be interested in testing the hypothesis that the proportion of defective units produced is the same for each of 6 possible assembly-line speeds. 6 samples of 100 each are recorded. Use a 1% level of significance.

	Assembly-line speed (units per hour)						
Quality	A = 60	B = 70	C = 80	D = 90	E = 100	F = 110	Total
Defective	6	4	5	5	6	4	30
Acceptable	94	96	95	95	94	96	570
Total	100	100	100	100	100	100	600

Step 1. Formulate the null and alternative hypotheses

We are dealing with multiple comparisons of proportions.

$$H_0: \pi_A = \pi_B = \pi_C = \pi_D = \pi_E = \pi_F$$

$$H_1: \pi_A \neq \pi_B \neq \pi_C \neq \pi_D \neq \pi_E \neq \pi_F$$

Step 2. Decide on the level of significance

$$\alpha = 0,01$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = (r - 1)(c - 1) = (2 - 1)(6 - 1) = 5$$

$$\chi_{crit}^2 = 15,09$$

$$\text{Reject } H_0 \text{ if } \chi_{calc}^2 > 15,09$$

Step 4. Calculate the sample test statistic

Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
6	5	1	1	0,2
94	95	-1	1	0,0105
4	5	-1	1	0,2
96	95	1	1	0,0105

Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
5	5	0	0	0
95	95	0	0	0
5	5	0	0	0
95	95	0	0	0
6	5	1	1	0,2
94	95	-1	1	0,0105
4	5	-1	1	0,2
96	95	1	1	0,0105
Σ	600			0,842

Step 5. Apply the rejection rule to the sample test statistic.

$$0,842 \text{ is } < 15,09$$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is not rejected in favour of the alternative hypothesis.

The conclusion then is that the population proportion of defectives is the same for each assembly-line speed tested (at a 1% level of significance).



SELF-ASSESSMENT ACTIVITY

A manufacturer of car batteries conducts a study to determine whether there are any differences in recall with respect to an advertisement when different media are used.

	Media		
Recall	Magazine	TV	Radio
Remember	25	10	7
Don't remember	73	93	108

At the 10% level of significance, determine whether there is evidence of a difference in recall of an advertisement for different media.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Step 1. Formulate the null and alternative hypotheses

We are dealing with multiple comparisons of proportions.

$$H_0: \pi_M = \pi_T = \pi_R$$

$$H_1: \pi_M \neq \pi_T \neq \pi_R$$

Step 2. Decide on the level of significance

$$\alpha = 0,10$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$$

$$\chi_{crit}^2 = 4,61$$

Reject H_0 if $\chi_{calc}^2 > 4,61$

Step 4. Calculate the sample test statistic

First total the contingency table:

	Media			
Recall	Magazine	TV	Radio	Total
Remember	25	10	7	42
Don't remember	73	93	108	274
Total	98	103	115	316

Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
25	13,0253	11,9747	143,3934	11,0088
73	84,9747	-11,9747	143,3934	1,6875
10	13,6899	-3,6899	13,6154	0,9946
93	89,3101	3,6899	13,6154	0,1525
7	15,2848	-8,2848	68,6379	4,4906
108	99,7152	8,2848	68,6379	0,6883
Σ	316			19,0223

Step 5. Apply the rejection rule to the sample test statistic.

$$19,02 \text{ is } < 4,61$$

Step 6. Draw a conclusion, both statistically and from a management perspective (i.e. in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that there is a difference in recall for different media (at a 10% level of significance).

6.5.3 Goodness-of-fit test

The chi-squared goodness-of-fit test is used to determine whether a set of sample data differs significantly from what is expected.

EXAMPLE

A manufacturer of soap wishes to know if consumers have a preference for bath soap fragrances. To answer the question, a random sample of 200 adult shoppers is offered a free bar of soap. The recipients choose from 4 fragrances. Determine at a 1% level of significance.

Rose	Lavender	Sandalwood	Lemon
66	53	45	36

Step 1. Formulate the null and alternative hypotheses

We are dealing with multiple comparisons to decide goodness-of-fit.

H_0 : There is no fragrance preference, ie all fragrances are equal

H_1 : There is fragrance preference

Step 2. Decide on the level of significance

$$\alpha = 0,01$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = c - 1 = 4 - 1 = 3$$

Important note: with only one row, only the columns are used to calculate the degrees of freedom.

$$\chi_{crit}^2 = 11,345$$

Reject H_0 if $\chi_{calc}^2 > 11,345$

Step 4. Calculate the sample test statistic

Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
66	50	16	256	5,12
53	50	3	9	0,18
45	50	-5	25	0,5
36	50	-14	196	3,92
Σ	200			9,72

Note: if no preference is expected, all fragrances should have an equal weighting for the 400 women surveyed. For this type of chi-squared hypothesis test, the expected frequencies need to be carefully thought through.

Step 5. Apply the rejection rule to the sample test statistic.

$$9,72 \text{ is } < 11,345$$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is not rejected in favour of the alternative hypothesis.

The conclusion then is that there is no fragrance preference (at a 1% level of significance).

EXAMPLE

Car manufacturers' national market shares are:

Manufacturer	National market share %
Volkswagen	37%
Toyota	30%
Delta	15%
BMW	10%
Mercedes	8%

The ownership pattern of a random sample of 2 000 car owners in Pretoria is: Volkswagen: – 758, Toyota – 680, Delta – 300, BMW – 162 and Mercedes – 100. Does the ownership pattern in Pretoria differ significantly from the national pattern? Use a 5% level of significance.

Step 1. Formulate the null and alternative hypotheses

We are dealing with multiple comparisons of proportions.

H_0 : The Pretoria ownership pattern is the same as the national pattern

H_1 : The Pretoria ownership pattern is different to the national pattern

Step 2. Decide on the level of significance

$$\alpha = 0,05$$

Step 3. Determine the rejection region and formulate the rejection rule

$$df = c - 1 = 5 - 1 = 4$$

$$\chi_{crit}^2 = 9,49$$

Reject H_0 if $\chi_{calc}^2 > 9,49$

Step 4. Calculate the sample test statistic

Observed frequency f_o	Expected frequency f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
758	$2\,000 \times 37\% = 740$	18	324	0,4378
680	$2\,000 \times 30\% = 600$	80	6 400	10,6667
300	$2\,000 \times 15\% = 300$	0	0	0
162	$2\,000 \times 10\% = 200$	-38	1 444	7,22
100	$2\,000 \times 8\% = 160$	-60	3 600	22,5
Σ	2 000			40,82

Step 5. Apply the rejection rule to the sample test statistic.

$$40,82 \text{ is } > 9,49$$

Step 6. Draw a conclusion, both statistically and from a management perspective (ie in English).

The null hypothesis is rejected in favour of the alternative hypothesis.

The conclusion then is that the Pretoria ownership pattern is different to the national pattern (at a 5% level of significance).

6.5.4 The logic behind the expected frequency calculation

Some students like to understand the logic behind calculations and this can assist them in performing the tests.

The expected frequency column applies the proportions of all groups to each group (the point of the formula is to establish whether these proportions hold true).

The logic for the expected frequency calculation is to calculate the proportions that apply to the whole, and then apply each to the individual group to see what is expected if all groups are the same. In this table we leave out the individual group observations and then calculate the expected frequencies.

News source	Ethnic group					Proportions for the whole
	A	B	C	D	Total	
TV					95	$\frac{95}{240} = 0,3958$
Radio					90	$\frac{90}{240} = 0,375$
Newspaper					55	$\frac{55}{240} = 0,2292$
Total	65	55	50	70	240	1

The expected frequencies are then calculated by applying the total proportions to each group (some rounding difference may occur using this method, but they're generally not material):

News source	Ethnic group					Proportions for the whole
	A	B	C	D	Total	
TV	$0,3958 \times 65 = 25,727$	$0,3958 \times 55 = 21,769$	$0,3958 \times 50 = 19,79$	$0,3958 \times 70 = 27,706$	95	0,3958
Radio	$0,375 \times 65 = 24,375$	$0,375 \times 55 = 20,625$	$0,375 \times 50 = 18,75$	$0,375 \times 70 = 26,25$	90	0,375
Newspaper	$0,2292 \times 65 = 14,928$	$0,2292 \times 55 = 12,606$	$0,2292 \times 50 = 11,46$	$0,2292 \times 70 = 16,044$	55	0,2292
Total	65	55	50	70	240	1

Unit 6 Exercises: (Solutions are found at the end of the module guide)**Exercise 6.1**

A hairdresser finds that the time taken to cut the men's hair follows a bimodal distribution with a mean of 30 minutes and a standard deviation of 10 minutes. She notes that 36 males from a particular college take an average of 27 minutes to have their hair cut. Does this group take less time than usual to have their hair cut? Use a one-sided test at a 5% level of significance and draw appropriate conclusions.

Exercise 6.2

For exercise 6.1, use a two-sided test at a 5% level of significance and draw appropriate conclusions.

Exercise 6.3

For exercise 6.1, use a one-sided test at a 1% level of significance and draw appropriate conclusions.

Exercise 6.4

A certain brand of cooking oil is advertised as containing an "average of 10% saturated fats". A random sample of 100 bottles reveals that their percentage of saturated fats has a sample mean of 10, 9% with a standard deviation of 3, 6%.

Based on this evidence, do you believe the manufacturer's claim? Use a one-sided test at a 5% level of significance.

Exercise 6.5

From exercise 6.4, if H_0 is rejected at a 5% level of significance, is it true that H_0 will *automatically* be rejected at a 1% level of significance?

Explain.

Exercise 6.6

The number of French fries in medium-size packet sold at a fast food outlet is usually known to follow a normal distribution with a mean of 18 fries. An inspector is investigating complaints that the number of fries in recent packets has been fewer than expected. To test this he selects a random sample of 9 packets and counts the number of fries in each. The results are:

14	12	20	16	15	15	17	18	13
----	----	----	----	----	----	----	----	----

Are the complaints justified? Use a one-sided test at a 5% significance level.

Exercise 6.7

A charcoal chicken establishment advertises its chickens as having an average weight of 1, 80 kg.

A random sample of 25 chickens purchased by a customer over the past few weeks has an average weight of 1, 65 kg with a standard deviation of 0, 6 kg.

Is it possible that the establishment is being truthful? Test at a 5% significance level.

Exercise 6.8

The scores of learner drivers over the years on a written driving test follow a distribution that is skewed to the right with a mean of 65 and a standard deviation of 12,5. A random sample of 30 learner drivers from TAFE score an average of 71, 0.

Is this evidence that these TAFE students perform differently from other learner drivers? Test at $\alpha = 0.01$.

Exercise 6.9

A manufacturer claims that his market share is 60%. However a random sample of 500 customers reveals that only 275 are users of his product. Test the manufacturer's claim at the 1% level of significance.

Exercise 6.10

In a training process, the average time taken is 6, 4 hours with a standard deviation of 1, 1 hours. 8 employees are trained using a new method and they have an average training time of 6, 2 hours. Use a 5% level of significance to determine if the new process reduces the average training time.

Exercise 6.11

A professional claims that at least 40% of all salesmen employed by firms switch jobs within 3 years of being hired. The alternative hypothesis is that the rate of job changing is below 40%. At a significance level of 1%, should the claim be accepted or rejected if sample results show that 25 out of 100 salesmen change jobs within 3 years?

Exercise 6.12

A supermarket chain believes that its store customers spend half an hour or more shopping.

A consumer body wants to verify this claim. They observe the entry and departure times from chain supermarkets of 86 randomly selected customers. The sample average shopping time is 23,4 minutes with a standard deviation of 14,1 minutes. Test the validity of the supermarket's belief at the 5 % level of significance.

Exercise 6.13

The marketing manager of Mores Desserts which launched a new flavoured pudding one month ago, wishes to assess the product's success in the market place. If average sales per week are less than R3 500 over this period, the product will be withdrawn. The results from a sample of 16 supermarkets countrywide indicate that average sales per week are R3 140 with a sample standard deviation of R648. Should the new pudding flavour be withdrawn? Advise the marketing manager by performing an appropriate hypothesis test at the 5% level of significance.

Exercise 6.14

Unemployment is said to be currently standing at not less than 15% of the economically active population. A random sample of 300 households in the Johannesburg area establishes that 34 of these households have at least 1 unemployed job-seeker. Is this claim about the % of unemployed job-seekers correct? Test at the 1% level of significance.

Exercise 6.15

A consumer testing service compares gas ovens to electric ovens by baking one type of bread in 5 ovens of each type. Assume the baking times are normally distributed. The gas ovens have an average baking time of 0,9 hours with a standard deviation of 0,09 hours and the electric ovens have an average baking time of 0,7 hours with a standard deviation of 0,16 hours. Test the hypothesis of identical mean baking times for the two kinds of ovens at the 5% level of significance. Assume identical variances.

Exercise 6.16

A small opinion poll of 200 Windhoek residents and 100 Rundu residents indicates that 48% and 52% respectively will purchase an electric car if the price is less than R16 000. Is the proportion of residents of Rundu who will purchase an electric vehicle different from that of Windhoek? Test at the 5% level of significance.

Exercise 6.17

Cartons of milk are advertised to contain 1 litre of milk. To test this claim, you measure every carton you buy for a month. Of the 20 litres you measures, you find that the average fill is 0,982 litres with a sample standard deviation of 0,068 litres. Test the claim at the 5% significance level that 1 litre cartons of milk are being under-filled.

Exercise 6.18

A town planning sub-committee in Windhoek wants to know if there is any difference in the average traveling time to work of car and train commuters. They carry out a survey amongst car and train commuters:

Car commuters	Train commuters
$\bar{x}_1 = 29,6 \text{ minutes}$	$\bar{x}_2 = 29,6 \text{ minutes}$
$s_1 = 5,2 \text{ minutes}$	$s_2 = 5,2 \text{ minutes}$
$n_1 = 22 \text{ drivers}$	$n_2 = 36 \text{ passengers}$

Test the hypothesis at the 5% significance level that it takes car commuters less time to get to work than train commuters.

Exercise 6.19

A sample of 500 respondents is selected in a large metropolitan area in order to determine various information concerning consumer behaviour. Among the questions asked is "Do you enjoy shopping for clothing?" Of the 240 males, 136 answer positively; of the 260 females, 224 answer positively.

- a) Is there evidence of a difference in the proportion of males and females who enjoy shopping for clothing? Use a 5% level of significance.
- b) Suppose that instead of determining whether there is a difference between the two groups in (a), we wish to know if there is evidence that the proportion of females who enjoy shopping for clothing is higher than the proportion of males? At the 5% level of significance, what conclusion do you reach?

UNIT 7
SIMPLE LINEAR REGRESSION AND CORRELATION
ANALYSIS

UNIT 7: SIMPLE LINEAR REGRESSION AND CORRELATION ANALYSIS

OBJECTIVES

By the end of this study unit, you should be able to:

1. Explain the purpose of regression and correlation analysis.
2. Compute and explain the meaning of the coefficients of correlation and determination.
3. Compute the regression equation and use this measure to do estimates.
4. Compute correlation by making use of ranking.

CONTENT

- 7.1 Introduction
 - 7.2 Simple linear regression analysis
 - 7.3 Correlation analysis
 - 7.4 The coefficient of determination
 - 7.5 Multiple regression
- Exercises



Prescribed textbook: Chapter 12

7.1 Introduction

Regression and correlation analysis are statistical tools used to study the relationship between two variables, one of which is dependent and the other independent.

This type of analysis is used to determine:

- Whether there is a relationship between the variables.
- How good that relationship is.
- How the relationship can be used to make estimates.

We will restrict ourselves to cases involving only two variables and will assume that the relationship between the two variables approaches a straight line graphically.

7.2 Simple linear regression analysis

Linear regression analysis finds the straight line equation representing the relationship between two numeric variables, the independent variable, x and the dependent variable, y .



TIP

In using regression analysis it is very important to understand which variable is which.

It can be helpful to think of x as the 'cause' and y as the 'effect', because x determines or influences y .

Linear regression is conducted using the following steps:

Step 1. Identify the dependent and independent variables

Identify the predictor or independent variable, x and the dependent or variable which needs to be estimated, y .

It can be useful to construct a table containing the known variable values.

Step 2. Construct a scatter plot

A scatter plot is a graph of the variable values and can be useful in visually determining the relationship between the variables:

- Whether the relationship approximates a straight line.
- How strong this relationship is.

- Whether the relationship is direct (as the value of x increases, so does the value of y) or inverse (as the value of x increase, the value of y decreases).

See also section 1.4.7 Scatter diagrams.

Step 3. Calculate the linear regression equation

Formula Straight-line graph:

$$y = a + bx$$

where:

x is the values of the independent variable

y is the values of the dependent variable

a is the y axis intercept

b is the slope of the straight line

n is number of observations

Regression analysis uses the method of least squares to find the straight line equation of best fit:

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

$$\text{intercept } a = \frac{\sum y}{n} - b \left(\frac{\sum x}{n} \right)$$

Step 4. Make estimations from the regression equation

Use the straight line equation to estimate values of y from values of x and establish the validity of these estimations.

EXAMPLE

The number of sales made by sales people in an organisation are listed against the number of sales calls made:

Sales	32	24	40	22	32
Sales calls made	14	10	16	8	12

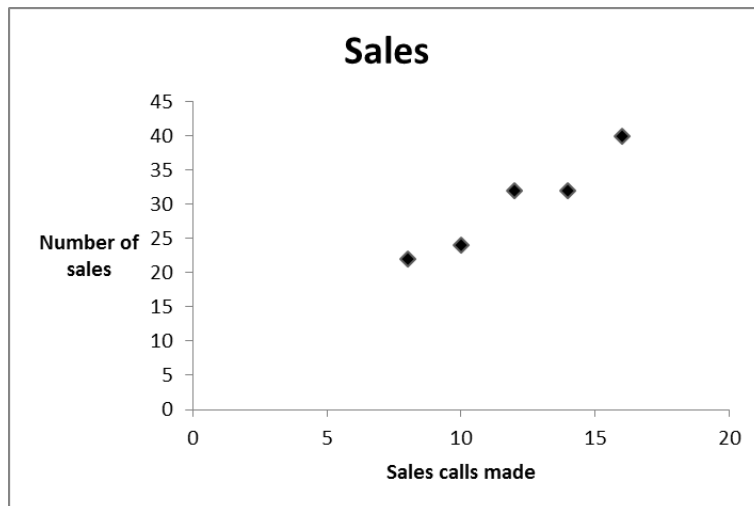
Using a scatter plot, comment on the relationship between sales called and sales made. Calculate straight line equation which best fits the data and estimate the sales which will be concluded from 15 sales calls.

Step 1. Identify the dependent and independent variables

Sales calls made, x is used to estimate sales, y .

Sales calls made, x	Sales, y
14	32
10	24
16	40
8	22
12	32

Step 2. Construct a scatter plot



A visual review of the scatterplot indicates that there is a strong positive linear relationship between sales calls made and number of sales concluded.

Step 3. Calculate the linear regression equation

Start by extending the data table to include all the required values for the formula:

Sales calls made, x	Sales, y	xy	x^2
14	32	448	196
10	24	240	100
16	40	640	256
8	22	176	64
12	32	384	144
Σ 60	150	1 888	760

$$\text{slope } b = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{5 \times 1\,888 - 60 \times 150}{5 \times 760 - (60)^2} = 2,2$$

$$\text{intercept } a = \frac{\Sigma y}{n} - b \left(\frac{\Sigma x}{n} \right) = \frac{150}{5} - 2,2 \left(\frac{60}{5} \right) = 3,6$$

$$y = 3,6 + 2,2x$$

Step 4. Make estimations from the regression equation

For 15 sales calls:

$$x = 15$$

$$y = 3,6 + 2,2 \times 15 = 36,6 \approx 37$$

15 sales calls is estimated to result in 37 sales.



SELF-ASSESSMENT ACTIVITY

You have observed 10 workers on the shop floor and have timed how long it takes each to produce an item. You have been able to match these times with the length of the workers' experience. The results are:

Person	Experience (months)	Time taken (minutes)
A	2	27
B	5	26
C	3	30
D	8	20
E	5	22
F	9	20
G	12	16
H	16	15
I	1	30
J	6	19

Draw a scatterplot for the data and comment. Determine the straight line equation and estimate how long a workers with 4 months', 10 months' and 24 months' experience will take to produce an item. If the company would like an item to be produced within 22 minutes, advise how much experience workers need.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Step 1. Identify the dependent and independent variables

Experience, x is used to estimate time taken, y .

Step 2. Construct a scatter plot



A visual review of the scatterplot indicates that there is a strong inverse (or negative) linear relationship between worker experience and time taken to produce one item.

**TIP**

If the relationship between the variables is inverse, except the value of b , the slope of the straight line equation, to be negative.

Step 3. Calculate the linear regression equation

Person	x , Experience (months)	y , Time taken (minutes)	xy	x^2
A	2	27	54	4
B	5	26	130	25
C	3	30	90	9
D	8	20	160	64
E	5	22	110	25
F	9	20	180	81
G	12	16	192	144
H	16	15	240	256
I	1	30	30	1
J	6	19	114	36
Σ	67	225	1 300	645

$$\text{slope } b = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{10 \times 1\,300 - 67 \times 225}{10 \times 645 - (67)^2} = -1,06$$

$$\text{intercept } a = \frac{\Sigma y}{n} - b \left(\frac{\Sigma x}{n} \right) = \frac{225}{10} + 1,06 \left(\frac{67}{10} \right) = 29,6$$

$$y = 29,6 - 1,06x$$

Step 4. Make estimations from the regression equation

With 4 months' experience:

$$x = 4$$

$$y = 29,6 - 1,06 \times 4 = 25,36 \approx 25$$

A worker with 4 months' experience is estimated to take 25 minutes to produce one item.

With 10 months' experience:

$$x = 10$$

$$y = 29,6 - 1,06 \times 10 = 19$$

A worker with 10 months' experience is estimated to take 19 minutes to produce one item.

To produce an item within 22 minutes:

With 24 months' experience:

$$x = 24$$

$$y = 29,6 - 1,06 \times 24 = 4,16 \approx 4$$

A worker with 24 months' experience is estimated to take 4 minutes to produce one item.

Important note: this estimation is **outside the range of the sample data and is therefore considered unreliable**. Because the sample doesn't extend to include this experience time, we can't be sure what will happen. It's always possible that very experienced workers may even display a decline in productivity!

To produce an item within 22 minutes:

$$y = 22$$

$$22 = 29,6 - 1,06x \therefore x = 7,17$$

A worker will need at least 7 months' experience to produce one item within 22 minutes.

7.3 Correlation analysis

In section the best-fit straight line equation is calculated for given data.

Correlation analysis determines how strong the linear relationship is between the x and y variables.

Formula Pearson's correlation coefficient:

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

where:

r is the sample correlation coefficient

x is the values of the independent variable

y is the values of the dependent variable

a is the y axis intercept

b is the slope of the straight line

n is number of observations

Correlation analysis again uses the method of least squares.



TIP

You'll notice that the numerator for the correlation coefficient is the same as that for the slope of the regression equation. Additionally the first part of the denominator is also the same as the denominator for the slope. If you have already calculated the slope, you can therefore re-use that part of the calculation when calculating the correlation coefficient.



TIP

The correlation coefficient gives the factor for strength of the linear relationship. If you convert the factor, say from 0,84 to 84%, you can think of the correlation coefficient as giving a percentage strength. The important point is that the correlation coefficient can never exceed 1 if positive or be below -1 if negative.

EXAMPLE

Calculate and interpret the correlation coefficient for the data in example

Start by extending the data table to include all the further required values for the formula:

	Sales calls made, x	Sales, y	xy	x^2	y^2
	14	32	448	196	1 024
	10	24	240	100	576
	16	40	640	256	1 600
	8	22	176	64	484
	12	32	384	144	1 024
Σ	60	150	1 888	760	4 708

$$r = \frac{5 \times 1\,888 - 60 \times 150}{\sqrt{[5 \times 760 - 60^2] \times [5 \times 4\,708 - 150^2]}} = 0,96$$

Interpretation: there is a strong (0, 96) positive correlation between sales calls made and sales concluded.



SELF-ASSESSMENT ACTIVITY

Calculate the correlation coefficient for the data in the previous self-assessment activity.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Person	x , Experience (months)	y , Time taken (minutes)	xy	x^2	y^2
A	2	27	54	4	729
B	5	26	130	25	676
C	3	30	90	9	900
D	8	20	160	64	400
E	5	22	110	25	484
F	9	20	180	81	400
G	12	16	192	144	256
H	16	15	240	256	225
I	1	30	30	1	900
J	6	19	114	36	361
Σ	67	225	1 300	645	5 331

$$r = \frac{10 \times 1\,300 - 67 \times 225}{\sqrt{[10 \times 645 - 67^2] \times [10 \times 5\,331 - 225^2]}} = -0,90$$

Interpretation: there is a strong inverse relationship between months of experience and the time taken to produce one item.



TIP

If the slope of the regression equation is negative, this indicates an inverse relationship between the variables. The correlation coefficient will then also be negative.

7.4 The coefficient of determination

The coefficient of determination measures the extent to which the independent variable determines the dependent variable.

Formula: Coefficient of determination:

$$\text{coefficient of determination} = r^2$$

where:

r is the sample correlation coefficient

EXAMPLE

Calculate and interpret the coefficient of determination for the data in example.

$$\text{coefficient of determination} = 0,96^2 = 0,92$$

Interpretation: 92% of the sales concluded are determined by the sales calls made. The remaining 8% are determined by other factors.

7.5 Multiple regression

When more than two variables are used in regression analysis, it is called multiple regression. This topic falls outside this course, but do be aware that more techniques exist.

Unit 7 Exercises: (Solutions are found at the end of the module guide)**Exercise 7.1**

Outstanding balances on the monthly bills of 9 credit card accounts and the household income of the account holders are:

Balance (\$)	250	1 630	970	2 190	410	830	0	550	0
Income (\$'000)	15	23	26	28	31	35	37	38	42

- Plot these figures on a scatter diagram.
- Calculate the Pearson correlation coefficient and comment on its value.

Exercise 7.2

The weekly turnover and total display area, in m², of 8 late night grocery stores are:

Turnover (\$)	23	37	33	41	47	86	72	95
Display area (m²)	15	21	30	45	61	77	79	92

- Identify which variable is the dependent variable.
- Plot a scatter diagram to portray these figures.
- Calculate the Pearson correlation coefficient and discuss its value.

Exercise 7.3

The cost of placing a full-page colour advertisement and circulation figures of 9 magazines are:

Cost (\$)	9	43	16	17	19	13	20	44	35
Circulation	135	2 100	680	470	450	105	275	2 250	695

- Which of these variables should be dependent variables and why?
- Plot a scatter diagram to portray the data.

Exercise 7.4

A consumer group has tested 9 makes of personal stereo. The prices and the scores (out of 100) awarded by a panel of experts the group commissioned to test the products are:

Price (\$)	95	69	18	32	27	70	49	35	50
Score	74	63	28	33	37	58	38	43	50

- Plot a scatter diagram to portray the data.
- Find the equation of the line of best fit using simple linear regression.
- Plot the line of best fit on the scatter diagram produced for (a).
- Use the regression equation from (c) to predict the score that the panel of experts would award to a personal stereo priced \$45.

Exercise 7.5

The annual turnovers (in R) and numbers of employees of 12 major retail companies are:

Turnover (R)	20,1	14	10,7	10,6	8,6	8,1	5,5	4,9	4,6	4,5	4,3	4,1
Number of employees	126	141	107	101	92	70	52	34	57	32	47	26

- Which variable should be the independent variable?
- Portray these data in the form of a scatter diagram.
- Find the equation of the line of best fit and plot the line on your scatter diagram.
- Work out the coefficient of determination and outline what it tells you about the relationship between the two variables.

Exercise 7.6

An economist studying the market for designer watches has produced a regression model to describe the relationship between sales of different brands of watch (in thousands of units) and the advertising expenditure used to promote them (R).

$$\text{Sales} = 4,32 + 6,69 \times \text{advertising expenditure}$$

$$r^2 = 64,8\%$$

- If there is no advertising expenditure to promote a brand of watch, what sales can be expected?
- By how many units are sales expected to increase for every extra R1m spent on advertising?
- What is the value of the Pearson correlation coefficient?

UNIT 8
FORECASTING: TIME-SERIES ANALYSIS

UNIT 8: FORECASTING: TIME-SERIES ANALYSIS

OBJECTIVES

After completion of this unit, the learner will be able to:

1. State the four possible components that make up a time series.
2. Use linear models to analyse and project secular trends.
3. Measure the cyclical effect.
4. Measure the seasonal effect by computing the seasonal indices.
5. Use time series in forecasting.

CONTENT

8.1 Introduction

8.2 Components of a time series

8.3 Histogram

8.4 Time-series decomposition

8.5 Introduction

8.6 Plotting time series data

8.7 Components of a time series

8.8 Trend analysis

8.8.1. Method of moving averages

8.8.2. Seasonal variations

8.8.3. Regression analysis – method of least squares

8.8.4. Regression analysis – method of zero-sum

8.9 Forecasting

Exercises



Prescribed textbook: Chapter 14

8.1 Introduction

Any variable measured over time in sequential order is called a **time series**.

The objective of time series analysis is to analyse how observed data changes over time, in order to detect recurring patterns that enable us to forecast the future behaviour of the data. The assumption underlying time – series analysis is that those factors that have influenced patterns of economic activity in the past and present will continue to do so in more or less the same manner in the future. Thus, time-series analysis helps us cope with uncertainty about the future.

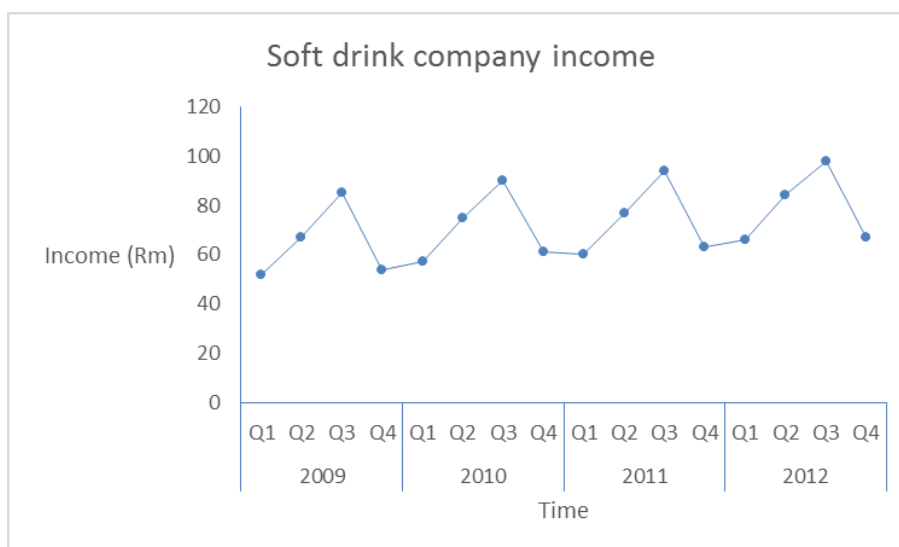
8.2 Plotting time series data

A line graph is used to present time series data. This allows for visual determination of trends and recurring patterns in the data.

EXAMPLE

The quarterly income (in Rm) of a soft drink company has been recorded for 4 years:

	2009	2010	2011	2012
Quarter 1 (January to March)	52	57	60	66
Quarter 2 (April to June)	67	75	77	84
Quarter 3 (July to September)	85	90	94	98
Quarter 4 (October to December)	54	61	63	67

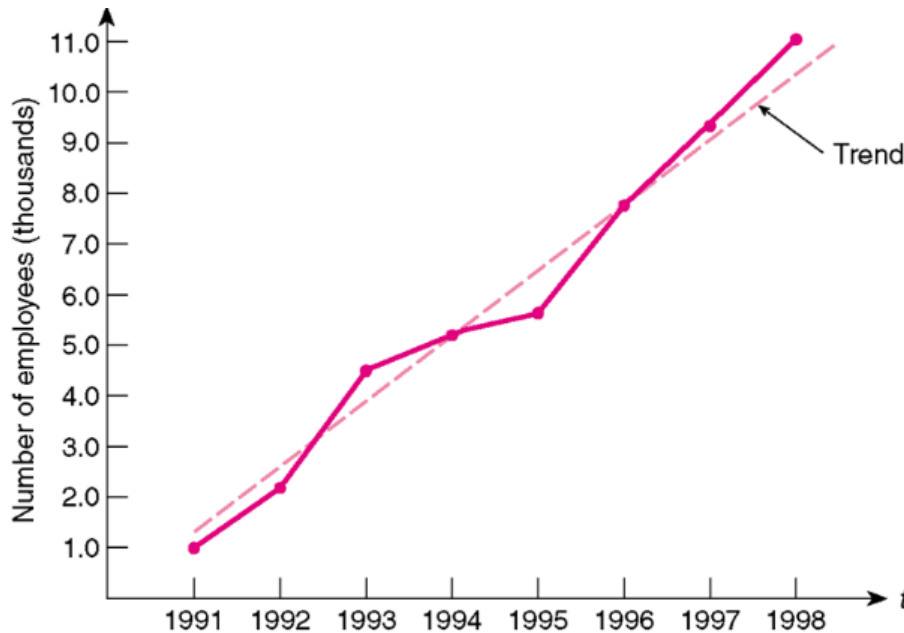


Interpretation: the graph shows a recurring pattern with sales at their lowest in the 1st quarter of every years and their highest in the 3rd quarter. There is a slight upward rise over time.

8.3 Components of a time series

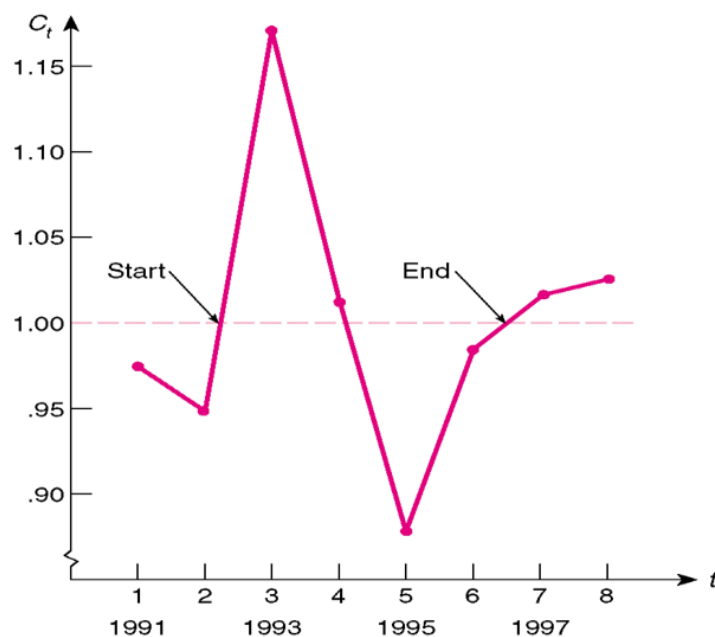
The major goal of time-series analysis is to identify and isolate the influencing factors in the time series for forecasting purposes. These factors are known as the components of the time series.

Trend (T): This is the general increase or decrease in the time series over an extended period of time caused by long-term trends e.g. population growth.



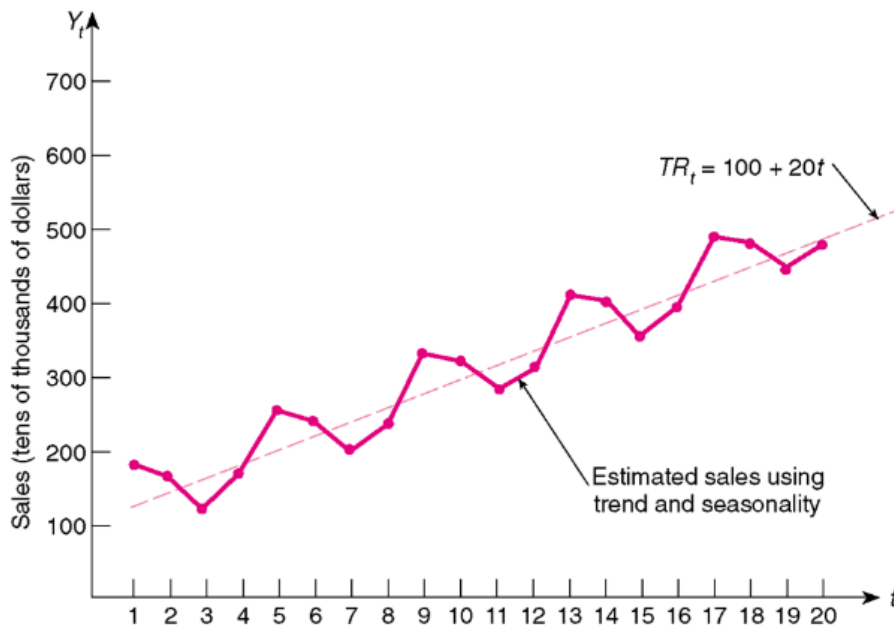
Source: Kvanli (2000)

Cycles (C): Cycles are variations that occur because of upward and downward swings in the general cycle of prosperity, recession, depression and recovery. This wavelike pattern, with the periods of expansion and contraction not of equal length, describes a long-term trend that is generally apparent over a number of years.



Source: Kvanli (2000)

Seasonality (S): Seasonal variations occur over short repetitive calendar periods and have a duration of less than a year and represent predictable deviations from the trend, e.g. annual agricultural crop yield.



Source: Kvanli (2000)

Random or irregular variations (I): Random variations occur over short intervals and are unpredictable, with no pattern to their behaviour. They tend to hide the existence of the other more predictable components. Examples of what can cause these movements are unexpected changes in the weather, political unrest, theft and war.

8.4 Trend analysis

The objectives behind trend analysis are:

- To describe the general underlying movement of a series as a straight line passing through the data with a positive or negative slope.
- To eliminate the trend in order to bring into focus the other movements in the series.

The two trend analysis methods studied in this course are:

- The moving average method with seasonal analysis.
- Regression analysis.

8.4.1 Method of moving averages

The moving average method smooths out peaks and valleys in a set of observations. The objective is to bring out the trend by eliminating any obscuring seasonal, cyclical or random fluctuations. One of its drawbacks is that value for some years are lost at the beginning and end of the series.

The method is explained by way of an example.

EXAMPLE

Repeating the information from the example in section 8.2:

The quarterly income (in Rm) of a soft drink company has been recorded for 4 years:

	2009	2010	2011	2012
Quarter 1 (January to March)	52	57	60	66
Quarter 2 (April to June)	67	75	77	84
Quarter 3 (July to September)	85	90	94	98
Quarter 4 (October to December)	54	61	63	67

The data is listed chronologically by time period:

Calculating a 3-quarter moving average:

Year	Quarter	Income (Rm)	3-quarter moving total (3QMT)	3-quarter moving average (3QMA)
2009	Q1	52		
	Q2	67	204	68,00
	Q3	85	206	68,67
	Q4	54	196	65,33
2010	Q1	57	186	62,00
	Q2	75	222	74,00
	Q3	90	226	75,33
	Q4	61	211	70,33
2011	Q1	60	198	66,00
	Q2	77	231	77,00
	Q3	94	234	78,00
	Q4	63	223	74,33
2012	Q1	66	213	71,00
	Q2	84	248	82,67
	Q3	98	249	83,00
	Q4	67		

Steps in calculating the 3-quarter moving average:

Step 1. Sum the 1st 3 values and record the total against the middle value of the 3 in the 3-quarter moving total column:

$$2009\ Q1 + 2009\ Q2 + 2009\ Q3 = 52 + 67 + 85 = 204$$

Step 2. Average this total value and record in the 3-quarterly moving average column:

$$\frac{204}{3} = 68$$

Step 3. Move one quarter down and sum the next 3 values and record the total against the middle value of the 3 in the 3-quarter moving total column:

$$2009\ Q2 + 2009\ Q3 + 2009\ Q4 = 67 + 85 + 54 = 206$$

Step 4. Average this total value and record in the 3-quarterly moving average column:

$$\frac{206}{3} = 68,7$$

Step 5. Continue calculating 3-quarter averages by moving down 1 quarter at a time.

Important note: There will be no values against the 1st and last quarters in the series, because values are entered against the middle value of 3 quarters. In the case of the 1st quarter, there is no previous quarter and in the case of the last quarter there is no next quarter, so it isn't possible to calculate a 3-quarter total or average.

Calculating a 4-quarter moving average:

Year	Quarter	Income (Rm)	4-quarter moving total (4QMT)	4-quarter moving average (4QMA)	Centred 4-quarter moving average (C4QMA)
2009	Q1	52			
	Q2	67			
			258	64,50	
	Q3	85			65,125
			263	65,75	
	Q4	54			66,750
			271	67,75	
2010	Q1	57			68,375
			276	69,00	
	Q2	75			69,875
			283	70,75	
	Q3	90			71,125
			286	71,50	
	Q4	61			71,750
			288	72,00	
2011	Q1	60			72,500
			292	73,00	
	Q2	77			73,250
			294	73,50	
	Q3	94			74,250
			300	75,00	
	Q4	63			75,875
			307	76,750	
2012	Q1	66			77,250
			311	77,75	
	Q2	84			78,250
			315	78,75	
	Q3	98			
	Q4	67			

Steps in calculating the 4-quarter moving average:

The principle is the same for that of the 3-quarter moving average.

Step 1. Sum the 1st four values and record the total half-way between the 2nd and 3rd values of the 4 in the 4-quarter moving total column:

$$2009\ Q1 + 2009\ Q2 + 2009\ Q3 + 2009\ Q4 = 52 + 67 + 85 + 54 = 258$$

**TIP**

When calculating 4-quarter moving averages, double space the data table so that the 4-quarter moving total and average values correctly line up **between** the 2nd and 3rd values and don't incorrectly line up against an existing value.

Step 2. Average this total value and record in the 4-quarterly moving average column:

$$\frac{258}{4} = 64,5$$

Step 3. Move 1 quarter down and sum the next 4 values and record the total between the 2nd and 3rd values of this 2nd group of 4 in the 4-quarter moving total column:

$$2009\ Q2 + 2009\ Q3 + 2009\ Q4 + 2010\ Q1 = 67 + 85 + 54 + 57 = 263$$

Step 4. Average this total value and record in the 4-quarterly moving average column:

$$\frac{263}{4} = 65,75$$

Step 5. Continue calculating 4-quarter averages by moving down 1 quarter at a time.

Important note: There will be no values against the 1st, 2nd, last and 2nd last quarters in the series, because values are entered against the 'middle' value of 4 quarters. In the case of the 1st and 2nd quarters, there is no previous quarter and in the case of the 2nd last and last quarter there is no next quarter, so it isn't possible to calculate a 4-quarter total or average.

Step 6. We now find ourselves with values in the 4-quarter moving average column which don't line up with the x-values. In order to rectify this, the values need to be centred. This is achieved by averaging 2 values at a time in a moving sequence:

$$\frac{64,50 + 65,75}{2} = 65,125$$

$$\frac{65,75 + 67,75}{2} = 66,75$$

As you can see from the final centred column, the values have been smoothed significantly from the original fluctuating values.

**TIP**

Sense test your average calculations. The average of a sequence of numbers cannot be higher than the highest number in the sequence or lower than the lowest number in the sequence. This will apply to the moving average and the centred moving average values.

**8.4.2 Seasonal variations**

Seasonal variations occur within a period of 1 year or less. Therefore, period data (weekly, monthly, quarter, daily, etc.) is required. Seasonal variation is generally expressed as an index number or percentage and can be identified using the ratio-to-moving-average-method. A requirement for this method is that we have a time series long enough to allow us to observe the variable over several seasons.

EXAMPLE**Calculate seasonal indices**

Using the data table in the example in section 8.2, we add a seasonal index or percentage column onto the table, after which a summary table is constructed in order to calculate the seasonal indices across all years.

Year	Quarter	Income (Rm)	4-quarter moving total (4QMT)	4-quarter moving average (4QMA)	Centred 4-quarter moving average (C4QMA)	%
2009	Q1	52				
	Q2	67				
			258	64,50		
	Q3	85			65,125	130,518%
			263	65,75		
	Q4	54			66,750	80,899%
		271	67,75			
2010	Q1	57			68,375	83,364%
			276	69,00		
	Q2	75			69,875	107,335%
			283	70,75		
	Q3	90			71,125	126,538%
			286	71,50		
	Q4	61			71,750	85,017%
		288	72,00			
2011	Q1	60			72,500	82,759%
			292	73,00		
	Q2	77			73,250	105,119%
			294	73,50		
	Q3	94			74,250	126,599%
			300	75,00		
	Q4	63			75,875	83,031%
		307	76,750			
2012	Q1	66			77,250	85,437%
			311	77,75		
	Q2	84			78,250	107,348%
			315	78,75		
	Q3	98				
	Q4	67				

Construct summary table:

Year	Q1	Q2	Q3	Q4	
2009			130,518%	80,899%	
2010	83,364%	107,335%	126,538%	85,017%	
2011	82,759%	105,119%	126,599%	83,031%	
2012	85,437%	107,348%			
Modified mean	82,759%	105,119%	126,599%	83,031%	397,509%
Adjustment factor	1,0063	1,0063	1,0063	1,0063	1,0063
Seasonal index	83,277%	105,778%	127,393%	83,552%	400%

Interpretation of seasonal index:

Values fluctuate seasonally showing a 17% drop in quarter 1, a 6% increase in quarter 2, a 27% increase in quarter 3 and a 16% drop in quarter 4.

Steps to calculate seasonal indices:

Step 1. Calculate 4-quarterly centred moving averages per example.

Step 2. Establish the percentage extent to which the actual value deviates from the 4-quarter centred average and enter each value in the % column.

$$\text{percentage deviation} = \frac{x}{C4QMA} = \frac{85}{65,125}$$

Step 3. Transcribe all percentage values into a summary table

Step 4. Calculate the modified mean for each quarter. The modified mean is calculated by removing the highest and lowest values and averaging the remaining values. If only one value remains after eliminating the highest and lowest values, then that remaining value becomes the modified mean. If there are only 2 values to start with, then the two values are averaged to give the mean.

Step 5. Total the modified mean percentages for all quarters. These should add up to 400% (there are 4 quarters). If they don't add up to 400%, use an adjustment factor to increase or decrease each value to achieve a total of 400% to finalise the seasonal indices. The adjustment factor is calculated as:

$$\text{adjustment factor} = \frac{400\%}{\text{modified mean total}} = \frac{400\%}{397,509\%} = 1,0063$$

Step 6. Adjust each modified mean value by the adjustment factor to achieve the seasonal index for that quarter.

8.4.3 Regression analysis – method of least squares

This method is useful to identify the trend as a series of values. The method of least squares was covered in unit 7.

EXAMPLE

Using the data in the example in section 8.2, calculate the trend line equation using the method of least squares.

The information has been given along a timeline of quarters and years. In order to enable the calculations, each quarter is coded with an x-value (the independent variable) from 1 extending to the last quarter which becomes $x = 16$. Income is the independent y variable. The additional columns required to calculate the straight line equation are then added and completed.

Year	Quarter	x	y	xy	x^2
2009	Q1	1	52	52	1
	Q2	2	67	134	4
	Q3	3	85	255	9
	Q4	4	54	216	16
2010	Q1	5	57	285	25
	Q2	6	75	450	36
	Q3	7	90	630	49
	Q4	8	61	488	64
2011	Q1	9	60	540	81
	Q2	10	77	770	100
	Q3	11	94	1034	121
	Q4	12	63	756	144
2012	Q1	13	66	858	169
	Q2	14	84	1176	196
	Q3	15	98	1470	225
	Q4	16	67	1072	256
Σ		136	1 150	10 186	1 496

$$\text{slope } b = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2} = \frac{16 \times 10\,186 - 136 \times 1\,150}{16 \times 1\,496 - (136)^2} = 1,2088$$

$$\text{intercept } a = \frac{\Sigma y}{n} - b \left(\frac{\Sigma x}{n} \right) = \frac{1\,150}{16} - 1,2088 \left(\frac{136}{16} \right) = 61,60$$

$$y = a + bx = 61,6 + 1,21x$$

8.4.4 Regression analysis – method of zero-sum

This method is a useful alternative to the method of least squares and requires fewer calculations.

The x-values are coded in such a way as to ensure $\Sigma x = 0$.

EXAMPLE

Using the data in the example in section 8.2, calculate the trend line equation using the zero-sum method.

The zero-sum method uses an x-code which ensures the total of x is zero. With an uneven number of entries, the zero-sum x-codes are sequential numbers; with an even number of entries (as in this case), the x-codes are alternating odd numbers worked from the middle.

Year	Quarter	<i>x</i>	<i>y</i>	<i>xy</i>	<i>x</i> ²
2009	Q1	-15	52	-780	225
	Q2	-13	67	-871	169
	Q3	-11	85	-935	121
	Q4	-9	54	-486	81
2010	Q1	-7	57	-399	49
	Q2	-5	75	-375	25
	Q3	-3	90	-270	9
	Q4	-1	61	-61	1
2011	Q1	1	60	60	1
	Q2	3	77	231	9
	Q3	5	94	470	25
	Q4	7	63	441	49
2012	Q1	9	66	594	81
	Q2	11	84	924	121
	Q3	13	98	1 274	169
	Q4	15	67	1 005	225
Σ		0	1 150	822	1 360

The formulae are simpler, because $\sum x = 0$.

$$\text{slope } b = \frac{\sum xy}{\sum x^2} = \frac{822}{1\,360} = 0,6044$$

$$\text{intercept } a = \frac{\sum y}{n} = \frac{1\,150}{16} = 71,875$$

$$y = a + bx = 71,88 + 0,6x$$

8.4.5 Forecasting

The regression straight line equation, combined with the calculated seasonal indices can be used to forecast future values.

Important note: when values fluctuate seasonally as in the case of our example, using just the trend line equation for forecasting is of no value. The straight line (trend line) values need to be adjusted by the seasonal indices in order to offer more likely forecasts (if the trend continues).

EXAMPLE

Using the data from the previous example, forecast the expected income for each quarter in 2013.

Using the method of least squares equation, $y = 61,6 + 1,21x$:

The x-codes continue (the last quarter of 2012 was 16, therefore the 1st quarter of 2013 is 17, the 2nd quarter 18 etc.).

Year	Quarter	x-code	Straight line equation	Trend line value	Seasonal index	Seasonalised trend
2013	Q1	17	$y = 61,6 + 1,21(17)$	82,17	83,277%	68,43
	Q2	18	$y = 61,6 + 1,21(18)$	83,38	105,778%	88,20
	Q3	19	$y = 61,6 + 1,21(19)$	84,59	127,393%	107,76
	Q4	20	$y = 61,6 + 1,21(20)$	85,8	83,552%	71,69

Using the zero-sum method equation, $y = 71,88 + 0,6x$:

The x-codes continue (the last quarter of 2012 was 15, therefore the 1st quarter of 2013 is 17, the 2nd quarter 19, etc).

Year	Quarter	x-code	Straight line equation	Trend line value	Seasonal index	Seasonalised trend
2013	Q1	17	$y = 71,88 + 0,6(17)$	82,08	83,277%	68,35
	Q2	19	$y = 71,88 + 0,6(18)$	82,68	105,778%	87,46
	Q3	21	$y = 71,88 + 0,6(19)$	83,28	127,393%	106,09
	Q4	23	$y = 71,88 + 0,6(20)$	83,88	83,552%	70,08

Unit 8 Exercises: (Solutions are found at the end of the module guide)**Exercise 8.1**

Consider the quarterly demand levels for electricity (in 1 000 megawatts) in Cape Town from 2009 to 2012:

	2009	2010	2011	2012
January to March	21	35	39	78
April to June	42	54	82	114
July to September	60	91	136	160
October to December	12	14	28	40

- Find the trend line for electricity demand using $\sum x = 0$.
- Find the seasonal index for each quarter.
- Estimate the demand for electricity for quarter 4, 2013.

Exercise 8.2

Use the following table to calculate the seasonal index using the ratio-to-moving average method:

Quarter	2010	2011	2012
1	54	59	63
2	55	59	65
3	56	60	67
4	55	60	69

Exercise 8.3

The management of an office building is studying a plan to reduce energy costs in the building. They have assembled quarterly data on electricity costs for the past three years (in R1 000).

Year	Quarter 1	Quarter 2	Quarter 3	Quarter 4
1	2,4	3,8	4,0	3,1
2	2,6	4,1	4,1	3,2
3	2,6	4,5	4,3	3,3

Compute seasonal indices for the building's electricity usage by the ratio-to-moving average method.

Exercise 8.4

The fitted linear trend equation for sales (in millions of Rand) for a company is:

$$y = 10 + 0.8x$$

where $x = 1$ in 1998 (x in one year units)

Find the trend value for 2013.

Exercise 8.5

The gross domestic product (GDP) for a certain country from 1980 to 1990 is:

Year	GDP (\$m)
1980	2 255
1981	2 650
1982	3 224
1983	4 049
1984	4 657
1985	5 432
1986	5 649
1987	6 227
1988	6 957
1989	7 271
1990	8 295

- a) Find the straight line trend $y = a + bx$, for this data if $\sum x = 0$.
- b) Use this equation to predict the GDP for 1992.

UNIT 9
DECISION ANALYSIS: DECISION TREES AND PAYOFF
TABLES

UNIT 9: DECISION ANALYSIS: DECISION TREES AND PAYOFF TABLES

OBJECTIVES

After completion of this unit, the learner will be able to:

1. Make decisions under conditions of certainty, risk and uncertainty.
2. Construct payoff and opportunity-loss tables.
3. Make decisions with or without probabilities.

CONTENT:

9.1 Introduction

9.2 Problem formulation

9.3 Classification of decision problems

9.3.1. Decisions under conditions of uncertainty

9.3.2. Decisions under conditions of certainty

9.3.3. Decisions under conditions of risk

9.4 Basic concepts

9.4.1. Constructing a payoff table

9.4.2. Constructing an opportunity-loss table

9.4.3. Determining the best action

9.5 Decision-making without probabilities

9.5.1. Maximin criterion (pessimistic approach)

9.5.2. Maximax criterion (optimistic approach)

9.5.3. Minimax regret criterion

9.6 Decision-making with probabilities

9.6.1. Prior analysis to determine the best action: discrete probability distributions

9.6.2. Expected monetary value criterion (EMV)

9.6.3. Minimum expected opportunity-loss (EOL) criterion

9.6.4. The expected value of perfect information (EVPI)

9.7 Decision trees

9.7.1. Guidelines for drawing up decision trees

Exercises

9.1 Introduction

The purpose of hypothesis testing is to help us reach a decision about a population by examining sample data from the population. This type of decision-making is known as **classical statistical inference**. Only 2 states of the parameter are tested and economic consequences are not considered.

Statistical decision theory is an alternative to classical statistical inference. Statistical data can be applied to achieve optimal or best decisions in an economic sense. This involves a logical and quantitative analysis of all factors and possibilities that can influence a decision problem and assist in arriving at an appropriate action to solve the problem. It is based upon the decision-maker's subjective or personal preferences and perceptions regarding evaluation of the probabilities to be used in a decision framework. Such probabilities express the strength of the decision-maker's belief regarding the uncertainties that are involved when there is little or no direct information available.

The following elements are contained in a decision involving several alternative choices:

- Environmental or outside influences affect the decision-making process but are not under the control of the decision-maker. Such influences are referred to as **events or states of nature**. These influences can be defined in terms of **probabilities**.
- The decision-maker usually has alternative strategies that can be employed in making a decision. These strategies are known as **alternative courses of action**.
- An outcome, usually a profit or loss, is the result of every outcome-event combination, and is shown in a payoff table.

The steps to follow are:

Step 1. clearly define the problem at hand

Step 2. list the possible alternatives

Step 3. identify the possible outcomes

Step 4. list the payoff (or profit) of the combination of alternatives and outcomes in a decision table

Step 5. select one of the mathematical decision theory models

Step 6. apply the model and make your decision

Two assumptions underlying decision analysis are:

- There is either an economic gain or an economic loss associated with each possible action.
- The decision-maker selects the solution that maximises the expected gain or minimises the expected loss.

9.2 Problem formulation

A decision can be seen as a choice among alternative courses of action. These courses of action replace the null and alternative hypotheses. You can be said to have a problem if you do not know what course of action is best and/or you are in doubt about the solution.

To formulate a problem situation, the following conditions need to be met:

- The decision-maker needs several possible options to evaluate prior to selecting the course of action.
- The decision-maker needs to be able to list all possible conditions (states of nature or events), which will affect the outcome of each action.
- It should be possible to calculate the monetary worth of the various outcomes, if a specific action is taken. These values can be positive (profit) or negative (loss) and are called **payoffs**.
- The decision-maker needs to be able to determine how to select the best course of action which results in the largest profit.

9.3 Classification of decision problems

9.3.1 Decisions under conditions of uncertainty

We are said to make a decision under conditions of uncertainty in situations where the outcome of each action depends on the prevailing event. It cannot be predicted with certainty which event will prevail because the conditions and the probability of occurrence are beyond the control of the decision-maker. However, the decision-maker may assign **subjective probabilities** to the event and these are based upon the best information that can be found.

Once probabilities have been assigned to possible outcomes, regardless of how these probabilities are derived, the solution procedure is the same. Therefore, calculations involving risk and uncertainty are handled in the same way and both are referred to in this unit as **conditions of uncertainty**.

9.3.2 Decisions under conditions of certainty

The decision-maker knows with certainty, the influencing factors (events) that will occur. Naturally, the alternative that will result in **the highest expected monetary value (EMV) will be chosen**. The most common method used is the method of linear programming, which is not covered in this module.

9.3.3 Decisions under conditions of risk

Risk is a condition where the occurrence of the possible events can be assigned probabilities or described with some probability.

9.4 Basic concepts

9.4.1 Constructing a payoff table

- **List all courses of action** you might consider in order to solve the problem. At least 3 alternatives need to be available so that a choice exists.
- **List all the states of nature (events)** that can occur for each alternative course of action.
- The events (E_1, E_2, E_3, \dots) each indicate a separate column and the actions (A_1, A_2, A_3, \dots) each indicate a separate row of the payoff table.

- A **probability value**, based either on historical information or managerial judgement, is included for each event.
- A payoff value is entered in each of the cells of the table. If probabilities are assigned to the events, an expected value can be calculated by multiplying each payoff by the respective probability of the event.
- Due to the events in the payoff table being mutually exclusive as well as exhaustive, the sum of the probabilities for all possible events must equal 1.

EXAMPLE

Tim Day, hospital administrator for Johannesburg General Hospital, is trying to determine whether to build a large wing onto the existing hospital, to build a small wing or to build no wing at all. If the population of Johannesburg continues to grow, a large wing could return R150 000 to the hospital each year. If a small wing is built, it will return R60 000 to the hospital each year. If the population remains the same, the hospital will suffer a loss of R85 000 if the large wing is built and a loss of R45 000 if the small wing is built.

Payoff table:

Action	Population growth	No population growth
Large wing	R150 000	(R85 000)
Small wing	R60 000	(R45 000)
No wing	R0	R0

9.4.2 Constructing an opportunity-loss table

From a payoff table we can construct an **opportunity-loss table**, which shows the opportunity-loss for every action-event combination.

Opportunity-loss is simply the difference between the payoff actually realised for a selected action and the payoff which could have been obtained had the best or optimal action been selected.

- Find the highest payoff in each column of the payoff table.
- Subtract each payoff from the highest payoff in that column.
- The best action for each event will have a zero entry.
- Entries in the opportunity-loss table will always be positive.
- Entries represent losses, therefore the smaller the value the better.

EXAMPLE

Using the data from the example in section 9.4.1:

Opportunity-loss table:

Action	Population growth	No population growth
Large wing	$R0$	$R0 - (R85\ 000) = R85\ 000$
Small wing	$R150\ 000 - R60\ 000 = R90\ 000$	$R0 - (R45\ 000) = R45\ 000$
No wing	$R0$	$R0$
Most opportunity lost	$R150\ 000$	$R0$

9.4.3 Determining the best action

There are several methods that can be used to help in making a decision about the best action. We distinguish between methods without probabilities and methods with probabilities.

9.5 Decision-making without probabilities

A prime consideration in choosing a method for determining the best action, when probabilities about the events are unknown, is the decision-maker's general attitude towards possible losses and gains.

9.5.1 Maximin criterion (pessimistic approach)

This procedure guarantees that the decision-maker can do no worse than to achieve the best of the poorest outcomes possible and will be the typical choice of a risk-averter (pessimist).

- Determine the lowest payoff associated with each action, ie the smallest value of each row.
- Comparing these minimum payoffs, choose that action (row) with the largest payoff, thus maximising the minimum payoffs.
- If the payoffs are quantities such as losses or costs, which are to be minimised, select the highest value for each action (row) and choose the action that minimises the maximum payoff.

9.5.2 Maximax criterion (optimistic approach)

Here, the decision-maker is concerned only with the best that can happen with respect to each action and will tend to choose the action to generate the highest possible payoff. It is the typical choice of a risk-seeker (optimist).

- For each possible action, identify the maximum possible payoff, ie the highest value in each row.
- Comparing these payoffs, select that action with the highest maximum payoff, thus maximising the maximum payoffs.
- If the payoffs are quantities such as losses or costs, which are to be minimised, select the lowest possible cost associated with each action and choose the action that minimises the minimum payoff.

9.5.3 Minimax regret criterion

This method is applied using an opportunity-loss table and ensures the decision-maker will have the least regret (i.e. the least opportunity lost).

- Find the highest (maximum) possible regret value associated with each possible action (row).
- Identify the lowest (minimum) value from these maxima.
- Choose that for which the maximum regret is the smallest, thus minimising the maximum regret.

EXAMPLE

You plan to print souvenir T-shirts to sell during a football match at FNB stadium. A decision needs to be made whether to print 10 000, 15 000 or 20 000 T-shirts. The demand for T-shirts will depend on attendance, which may be high, medium or low. The following payoff tables show the expected profits (in R'000) for different attendance levels together with each possible quantity of T-shirts produced. Determine the best action without probabilities.

Payoff table

Action (number of T-shirts)	High attendance	Medium attendance	Low attendance
10 000	12	10	9,6
15 000	20	18	6
20 000	30	16	4

Opportunity-loss table

Action (number of T-shirts)	High attendance	Medium attendance	Low attendance
10 000	$30 - 12 = 18$	$18 - 10 = 8$	$9,6 - 9,6 = 0$
15 000	$30 - 20 = 10$	$18 - 18 = 0$	$9,6 - 6 = 3,6$
20 000	$30 - 30 = 0$	$18 - 16 = 2$	$9,6 - 4 = 5,6$
Most opportunity lost	30	18	9,6

Maximin criterion:

Action	Lowest payoff
10 000	9,6
15 000	6
20 000	4

The action with the highest minimum payoff is to produce 10 000 T-shirts.

Maximax criterion:

Action	Highest payoff
10 000	12
15 000	20
20 000	30

The action with the highest maximum payoff is to produce 20 000 T-shirts.

Minimum regret criterion:

Action	Maximum regret from opportunity-loss table
10 000	18
15 000	10
20 000	5,6

The action with the smallest maximum regret is to produce 20 000 T-shirts.



SELF-ASSESSMENT ACTIVITY

The manager of a small grocery store needs to decide how many loaves of bread to stock each day. The store has never sold fewer than 11 or more than 14 loaves. The cost per loaf is R2 and the retail price is R3. At the end of each day, any unsold bread is given away to the street children's home. Determine the best action without probabilities using events 11, 12, 13 and 14 and actions 11, 12, 13 and 14.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Payoff table

Calculations reflect total profit = R1 profit for each loaf sold less R2 cost for each loaf given away.

Action (number of loaves stocked)	Events (number of loaves sold)			
	11	12	13	14
11	R_{11}	R_{11}	R_{11}	R_{11}
12	$R_{11} - R_2 = R_9$	R_{12}	R_{12}	R_{12}
13	$R_{11} - R_4 = R_7$	$R_{12} - R_2 = R_{10}$	R_{13}	R_{13}
14	$R_{11} - R_6 = R_5$	$R_{12} - R_4 = R_8$	$R_{13} - R_2 = R_{11}$	R_{14}

Opportunity-loss table

Action (number of loaves stocked)	Events (number of loaves sold)			
	11	12	13	14
11	$R_{11} - R_{11} = R_0$	$R_{12} - R_{11} = R_1$	$R_{13} - R_{11} = R_2$	$R_{14} - R_{11} = R_3$
12	$R_{11} - R_9 = R_3$	$R_{12} - R_{12} = R_0$	$R_{13} - R_{12} = R_1$	$R_{14} - R_{12} = R_2$
13	$R_{11} - R_7 = R_4$	$R_{12} - R_{10} = R_2$	$R_{13} - R_{13} = R_0$	$R_{14} - R_{13} = R_1$
14	$R_{11} - R_5 = R_6$	$R_{12} - R_8 = R_4$	$R_{13} - R_{11} = R_2$	$R_{14} - R_{14} = R_0$

Most opportunity lost	<i>R11</i>	<i>R12</i>	<i>R13</i>	<i>R14</i>
------------------------------	------------	------------	------------	------------

Maximin criterion:

Action	Lowest payoff
11	R11
12	R9
13	R7
14	R5

The action with the highest minimum payoff is to stock 11 loaves.

Maximax criterion:

Action	Highest payoff
11	R11
12	R12
13	R13
14	R14

The action with the highest maximum payoff is to stock 14 loaves.

Minimum regret criterion:

Action	Maximum regret from opportunity-loss table
11	R3
12	R3
13	R4
14	R6

The action with the smallest maximum regret is to stock 11 loaves.

9.6 Decision-making with probabilities

If the decision problem falls within the risk uncertainty category, probabilities can be used to measure the likelihood of the occurrences of the various events.

- If the possible number of alternatives is limited, payoff tables can be used to solve problems and discrete probability distributions are applicable.
- If the number of possibilities becomes too large, it is more realistic to analyse a decision in terms of continuous probability distributions, such as the normal distribution.

9.6.1 Prior analysis to determine the best action: discrete probability distributions

Prior probabilities are determined by using information available from past experience, judgemental (subjective) information about the events or a combination of available and subjective information. No investigation, sampling or other form of experiment has been carried out in order to determine these probabilities.

9.6.2 Expected monetary value criterion (EMV)

This method is applied to select that action which has the highest expected payoff or monetary value.

- Specify the possible courses of action.
- Specify the possible events.
- Set up a payoff table.
- Assume probabilities representing the likelihood of occurrence of events and include them in the payoff table.
- Compute the expected payoff for each event by multiplying the payoff in each cell of the payoff table by the probability of the event for that column.
- The resulting expected payoffs are then summarised for each action to determine the EMV for the action.
- Make a decision by choosing the highest expected payoff (EMV).

9.6.3 Minimum expected opportunity-loss (EOL) criterion

The minimum expected opportunity-loss method measures the expected cost of uncertainty, due to uncertain knowledge about which event will prevail. The purpose is to select that action which yields the minimum expected opportunity-loss (EOL). An opportunity-loss is defined as the difference between the payoff of the best action that could have been selected for an event and the payoff of any other action for the same event (ie it is the profit forgone due to the failure to take the best action). The EOL criterion will always lead to the same decision as the EMV criterion and the EOL for each action is calculated in the same way as the EMV.

9.6.4 The expected value of perfect information (EVPI)

Before actually selecting a decision action, the decision-maker needs to decide whether to stop the analysis after using only prior information or to postpone the decision until additional sample information about the events can be obtained in order to reduce uncertainty. By eliminating uncertainty, the decision-maker will know which event is going to occur and can make a decision that is best for that event. To help in this decision, the decision-maker should be concerned about the cost of additional information and its potential value.

A measure to calculate this cost is the EVPI, which indicates the maximum expected gain in profit if additional information does provide certainty about the events. The maximum amount spent to be certain of the event that will occur should not exceed this maximum expected gain.

To calculate the EVPI, the opportunity-loss table and the EOL calculation can be used. The EOL of an action is the difference between the expected profit when perfect information is available and the expected profit under uncertain conditions. The minimum EOL is thus the EVPI

If additional information is not available at a cost equal to or less than the EVPI, the decision-maker cannot improve profit by obtaining additional information.

EXAMPLE

Using the data from the previous example, with the demand for T-shirts of high, medium and low having the respective probabilities of 0,4, 0,35 and 0,25:

Expected payoff table

Action (number of T-shirts)	High attendance	Medium attendance	Low attendance	Total
	0,4	0,35	0,25	
10 000	$0,4 \times 12 = 4,8$	$0,35 \times 10 = 3,5$	$0,25 \times 9,6 = 10,7$	10,7
15 000	$0,4 \times 20 = 8,0$	$0,35 \times 8 = 6,3$	$0,25 \times 6 = 1,5$	15,8
20 000	$0,4 \times 30 = 12$	$0,35 \times 6 = 5,6$	$0,25 \times 4 = 1,0$	18,6

The maximum expected monetary value (EMV) is R18 600 and the decision will be to produce 20 000 T-shirts.

Expected opportunity-loss table

Action (number of T-shirts)	High attendance	Medium attendance	Low attendance	Total
	0,4	0,35	0,25	
10 000	7,2	2,8	0,0	10
15 000	6,0	0,0	0,9	4,9
20 000	0,0	0,7	1,4	2,1

Since we are dealing with losses, select the action that minimises the expected opportunity loss, namely to produce 20 000 T-shirts with an expected loss of R2 100.

Expected value of perfect information (EVPI)

It can be mathematically proved that the $EVPI = EOL$; therefore the maximum amount you should be willing to pay for additional sample information is R2 100, because that is the maximum expected gain in profits that could result if more accurate information is available.



SELF-ASSESSMENT ACTIVITY

Refer to the first example. The following probabilities of selling bread are known:

$$P(11) = 0,3; P(12) = 0,4; P(13) = 0,2; P(14) = 0,1$$

Calculate the EMV and the EVPI.

SOLUTION TO SELF-ASSESSMENT ACTIVITY

Expected payoff table:

Action (number of loaves stocked)	Events (number of loaves sold)				Total
	11 0,3	12 0,4	13 0,2	14 0,1	
11	R3,30	R4,40	R2,20	R4,00	R13,90
12	R2,70	R4,80	R2,60	R1,20	R11,30
13	R2,10	R4,00	R2,60	R1,30	R10,00
14	R1,50	R3,20	R2,20	R1,40	R8,30

The maximum expected monetary value (EMV) is R13, 90 and the decision will be to stock 11 loaves.


Expected opportunity-loss table


Action (number of loaves stocked)	Events (number of loaves sold)				Total
	11 0,3	12 0,4	13 0,2	14 0,1	
11	R0	R,40	R,40	R0,3	R1,10
12	R,90	R0	R,20	R0,20	R1,30
13	R1,20	R,80	R0	R0,10	R2,10
14	R1,80	R1,60	R,40	R0	R3,80

Since we are dealing with losses, select the action that minimises the expected opportunity loss, namely to stock 11 loaves with an expected loss of R1, 10.

9.7 Decision trees

In Section 9.4 to 9.6, problems where a decision needs to be made between alternative actions with various outcomes is illustrated by means of payoff tables. A payoff table can also be represented diagrammatically in the form of a decision tree. A decision tree shows alternative actions that a decision maker can follow as well as events that can arise, by means of branches. Decision trees incorporate points where a decision needs to be made and points indicating the events that follow on decisions. These points are shown by means of symbols, known as nodes:

 **A decision node** shows that at those points a choice needs to be made between one or more alternatives.

 **An event node** where one or more uncertain events can occur.

EXAMPLE

The application of a decision tree is explained by means of this example.

Fresh Market Supplies buys a highly perishable product at R3 per box and sells it at R5 a box. If the product is not sold within 1 day, it is destroyed at a cost of R0, 25 per box. The purchasing manager has to deal with the problem of deciding how many boxes to order to satisfy the next day's demand.

The sales records show that the sales for the past 150 days are:

Boxes sold per day	Number of days on which that quantity has been sold
100	15
200	45
300	60
400	30

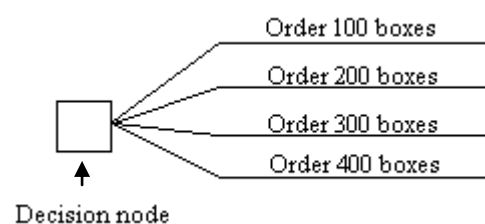
An analysis of the problem brings to light the fact that there are 4 decision-making alternatives. Fresh Market Supplies can order 100, 200, 300 or 400 boxes per day. In addition, there are four possible outcomes for each alternative, namely that the demand for the product can be 100, 200, 300 or 400 boxes per day.

Required: use a decision tree to analyse the data.

The problem of the purchasing manager of Fresh Market Supplies in example above can be represented by means of a decision tree. The daily demand probabilities are:

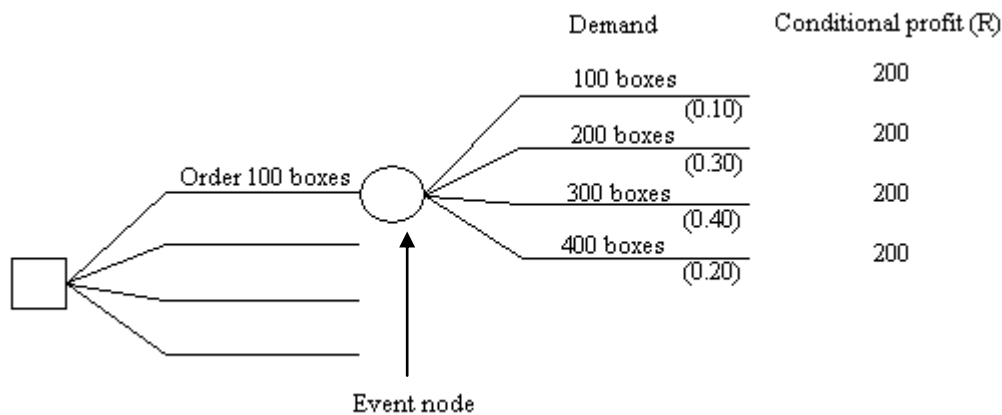
$$\frac{15}{150} = 0,10; \frac{45}{150} = 0,30; \frac{60}{150} = 0,40; \frac{30}{150} = 0,20$$

Alternative order quantities



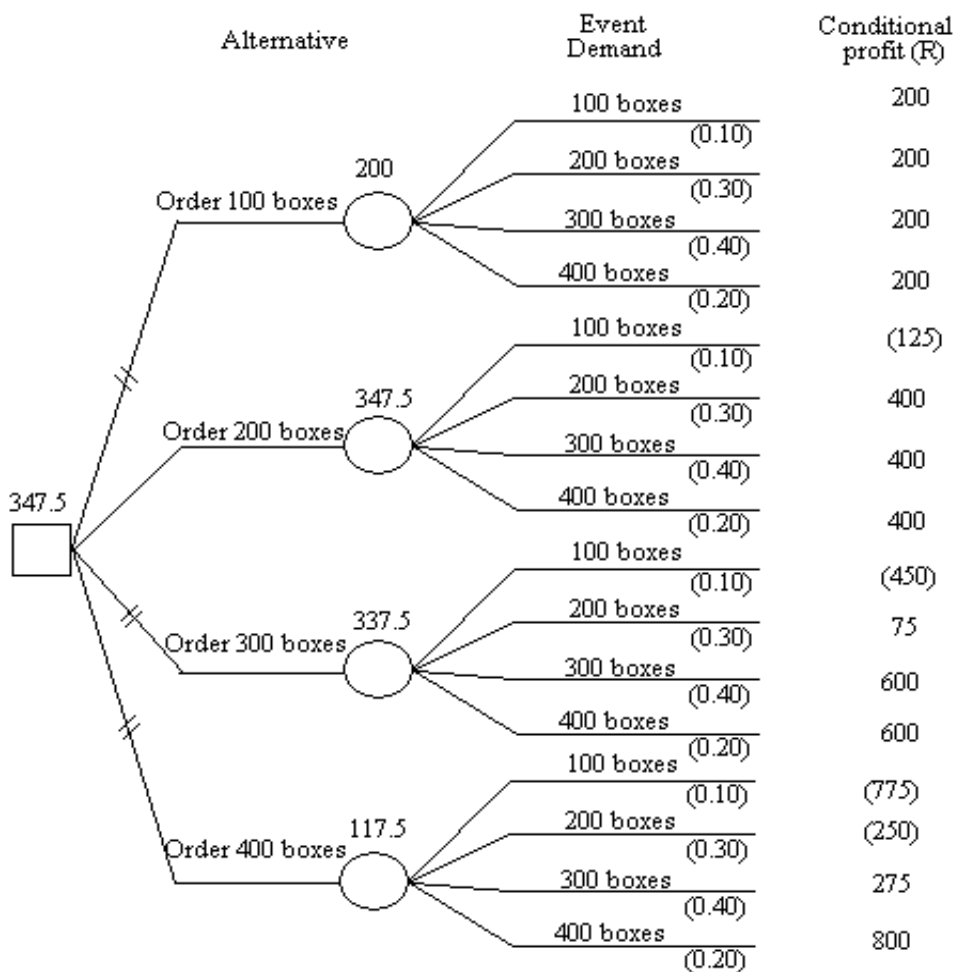
If the purchasing manager decides to order a specific quantity, for example 100 boxes, various uncertain events can happen as the demand for the product can vary from 100 boxes to 400 boxes.

The events for the alternative of 100 boxes ordered



Note that the outcome of events following on the event node is uncertain and that the probabilities allocated to them are entered below each branch. The conditional profit in respect of each event is calculated and entered at the end of the branch. By multiplying the conditional profit of each branch by the probability applying to that branch and adding the results, the expected value of the alternative of ordering 100 boxes can be calculated.

Complete decision tree



Once the expected value has been calculated for each alternative, the best alternative is chosen. The other alternatives are cancelled by drawing two parallel lines through the branch and entering the expected monetary value of the optimal alternative above the decision node.

Steps to be followed with decision trees

- Step 1. The problem is defined, the various decisions to be made are identified and the events that can result from each decision are determined.
- Step 2. The decision tree is drawn, making use of decision and event nodes.
- Step 3. Probabilities are allocated to the events.
- Step 4. The conditional value (usually profit or loss) is determined in respect of each event.
- Step 5. The expected monetary value is calculated for each alternative.
- Step 6. A choice is made.

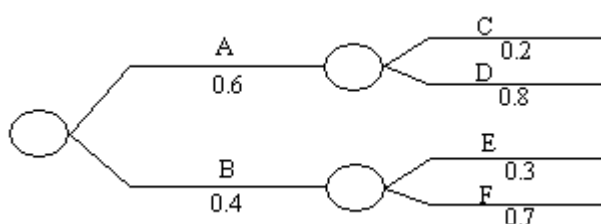
The choices between the various alternatives are made by evaluating the decision tree backwards, ie from right to left on the diagram. An alternative that is not chosen is shown by two parallel lines (//) on the decision tree.

Always bear in mind that a decision tree does not provide a final answer, but rather highlights a particular strategy.

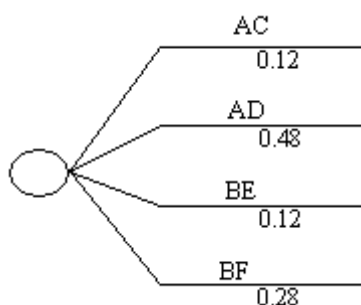
9.7.1 Guidelines for drawing up decision trees

The following guidelines can be followed when drawing up a decision tree:

- (a) Branches originating from a connecting node need to all be of the same type.
 - (b) A decision node can be followed by another decision node or by an event node.
 - (c) Alternatives originating from a decision node need to include all alternative actions.
 - (d) An event node is usually followed by a decision node or by allocating provisional values to the event.
- If two event nodes follow each other, the second case represents a conditional probability.



One of the event nodes can be eliminated by calculating the joint probability of the two events.



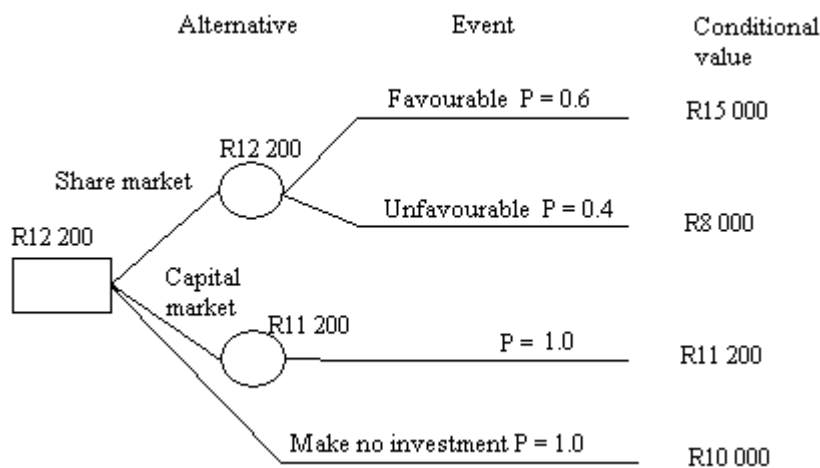
- (e) Events originating from an event node need to be of equal value and mutually independent. The outcome of an event node always has a total probability of 1, irrespective of the number of outcomes.

The events frequently show a favourable result as opposed to an unfavourable result, for example success as opposed to failure.

EXAMPLE

Mr Rykaard has R10 000 available which he wishes to invest for a year. He approaches his friend, the bank manager, for advice. The bank manager informs him that he can invest in either the share market or the capital market. By investing in the capital market, he can earn a fixed interest rate of 12%. If the share market reacts favourably, he will be able to make a profit of R5 000 on his capital, without taking dividends into account. If the share market is unfavourable however, he will lose R2 000 of his capital. The probability of a favourable share market, i.e. rising share prices, is 0,6 while the probability of an unfavourable share market is 0,4.

Choice of investment



*The value of investment after one year

Calculation of the expected value:

(a) Investment on the share market:

$$= R15\,000 \times 0,6 + R8\,000 \times 0,4 = R9\,000 + R3\,200 = R12\,200$$

(b) Investment on the capital market = $R11\,200 \times 1,0 = R11\,200$

Investment on the capital market is shown as a single outcome with a probability of 1,0 because it is certain that interest of R1 200 will be earned. The alternative choice, namely to make no investment, should also be shown. From the example it is clear that an amount of more than R10 000 will indeed be recovered at the end of the period. In most problems it is not so clear or certain that the investment amount will be recovered and as a result this alternative, namely to **do nothing** so as to ensure that no loss is made, should always be borne in mind.

Decision trees are of increasing value when more complex problems are being considered. The assumptions made in this example can all be changed and this can give rise to complex problems. The following assumptions applied in respect of example

- a) Outcomes have a discrete distribution that can be continuous.

- b) Only two alternative investment are considered. This can be extended to any number.
- c) A duration of 1 year is assumed. Any period can be considered, in which case the amounts receivable during the various periods are discounted to a present value
- d) A conversion from one form of investment to another and extension of the investment and change in return, for example interest during the investment period is not considered. In practice all these factors need to be taken into consideration.

Unit 9 Exercises: (Solutions are found at the end of the module guide)

Use the following information the case study below to solve Exercise 9.1, 9.2, 9.3 and 9.4.

Case study

Following the success of their ABC fast food restaurant in Windhoek, the proprietors, John and Peter, are thinking of expanding the business. They can do this by investing in new sites or by franchising the operation to aspiring fast food entrepreneurs who will pay a fee to John and Peter. The estimated profits for each strategy depend on the future demand for healthy fast food, which could increase, remain stable or decline. Another possibility for John and Peter is to accept the offer of R20m that a major international fast food company has made for their business.

Expected profits (in Rm) for John and Peter

Decision alternative	Increasing demand	Steady demand	Decreasing demand
Invest	100	40	(30)
Franchise	60	50	0
Sell	20	20	20

Exercise 9.1

Which strategy should be selected in the case study according to the maximax decision rule?

Exercise 9.2

Which strategy should be selected in the case study according to the maximin decision rule?

Exercise 9.3

Which strategy should be selected in the case study according to the minimax regret decision rule?

Exercise 9.4

Which strategy should be selected in the case study according to the equal likelihood decision rule?

Exercise 9.5

Ivan Loyer claims she has been unfairly dismissed by her employers. She consults the law firm of Zackon and Vorovat, who agree to take up her case. They advise her that if she wins her case she can expect compensation of R15 000 but if she loses she will receive nothing. They estimate their fee to be R1 500, which she will have to pay whether she wins or loses. Under the rules of the relevant tribunal she cannot be asked to pay her employer's costs. As an alternative they offer her a 'no win no fee' deal under which she pays no fee but if she wins her case Zackon and Vorovat take one-third of the compensation she receives. She can decide against bringing the case, which will incur no cost and result in no compensation.

Advise Ivana what to do:

- a) Using the maximax decision rule.

- b) Using the maximin decision rule.
- c) Using the maximax regret decision rule.
- d) Using the equal likelihood decision rule.

Exercise 9.6

Zak 'the snack' Cusker rents out a pitch for his stall at a music festival. The night before the festival he has to decide whether to load his van with ice-cream products or burgers or a mix of burgers and ice-cream products. The takings he can expect in R depend on the weather:

Action	Sun	Showers
Ice cream	2 800	1 300
Mix	2 100	2 200
Burgers	1 500	2 500

Recommends which load Zak should take using:

- a) The maximax decision rule.
- b) The maximin decision rule.
- c) The maximax regret decision rule.
- d) The equal likelihood decision rule.

Exercise 9.7

V Nimanía Plc builds water treatment facilities throughout Namibia. One contract has raised concerns about an installation in an area prone to outbreaks of a dangerous disease. The company has to decide whether or not to vaccinate the employees who will be working there. Vaccination will cost N\$200 000, which will be deducted from the profit it makes from the venture. The company expects a profit of N\$1,2m from the contract but if there is an outbreak of the disease and the workforce has not been vaccinated, delays will result in the profit being reduced to N\$0,5m. If the workforce has been vaccinated and there is an outbreak of the disease, the work will progress as planned but disruption to infrastructure will result in profit being reduced by N\$0,2 million.

Advise the company using:

- a) The maximax decision rule.
- b) The maximin decision rule.
- c) The minimax regret decision rule.
- d) The equal likelihood decision rule.

Exercise 9.8

Pashley Package Holidays in country ABC has to decide whether to discount its holidays to overseas destinations next summer in response to poor consumer confidence in international travel following recent flood events. If they do not discount their prices and consumer confidence in air travel remains low the company expects to sell 1 300 holidays at a profit of R60 per holiday. However, if they discount their prices and confidence remains low they expect to sell 2 500 holidays at a profit R35 per holiday. If they do not discount their prices and consumer confidence in air travel recovers they expect to sell 4 200 holidays at a profit of R50 each. If they do discount their prices and consumer confidence recovers, they expect to sell 5 000 holidays at a profit of R20 each.

Recommend which course of action the company should take with the aid of:

- a) The maximax decision rule
- b) The maximin decision rule
- c) The minimax regret decision rule
- d) The equal likelihood decision rule

Exercise 9.9

Cloppock Cotton is a farming collective in a central ABC republic. Their operations have been reliant on a government subsidy paid out to cotton farmers to support the production since cotton is a key export commodity. There are rumours that government will reduce the subsidy for the next crop. The Cloppock farmers have to decide whether to increase or decrease the number of hectares farmed or to keep it the same. The payoffs (in \$, the international currency) for these strategies under the same subsidy regime and under reduced subsidies are:

Area	Same subsidy	Reduced subsidy
Increase	\$80 000	(\$40 000)
The same	\$40 000	\$15 000
Decreased	\$20 000	\$17 000

Suggest what the farmers should do using:

- (a) The maximax decision rule.
- (b) The maximin decision rule.
- (c) The minimax regret decision rule.
- (d) The equal likelihood decision rule.

Exercise 9.10

Op-en-Wakker Limiter is a company in the clothing industry with a number of chain stores. At present the company is very inactive in the Western Cape and would like to expand to this marketing area. The company needs to take a decision on the size of their first shop in the Western Cape. Once the shop has been erected an extensive advertising campaign will be instituted. This will cost R50 000 and will have an 85% chance of favourable results.

The management calculates that there is a 0, 4 chance of the company dominating a large share of the market with their new shop and a 0, 6 chance that an average market penetration will be achieved. If a large share of the market is captured, a large shop will be needed. This will cost R8 000 000 to erect and will yield an annual estimated profit of R13 000 000. If an average share of the market is obtained, a smaller shop costing R5 000 000 will be required which will yield an annual estimated profit of R8 000 000.

The company can also build a small shop, wait for the result of the advertising campaign and then if necessary enlarge the shop at a cost of R4 500 000.

- a) Draw up a decision tree for the company.
- b) Determine what action the company should take.

Exercise 9.11

The following decision table gives the payoffs (in \$m) for 3 decision alternatives for 4 different possible states of demand for a company's product (less demand, same demand, moderate increase in demand and large increase in demand).

- a) Draw a decision tree to represent this payoff table.
- b) Compute the expected monetary value for this investment.

Decision alternative	Less demand (0,10)	Same demand (0,25)	Moderate increase (0,40)	Large increase (0,25)
No expansion	(\$3m)	\$2m	\$3m	\$6m
Add on	(\$40m)	(\$28m)	\$10m	\$20m
Build a new facility	(\$210m)	(\$145m)	(\$5m)	\$55m

Exercise 9.12

The dilemma of how to invest \$10 000

Suppose we determine that there is a 0, 25 probability of having a stagnant economy, a 0, 45 probability of having a slow-growth economy and a 0, 30 probability of having rapid-growth economy. The investment is \$10 000 for the four possibilities (Stock, Bonds, CDs and Mixtures) is:

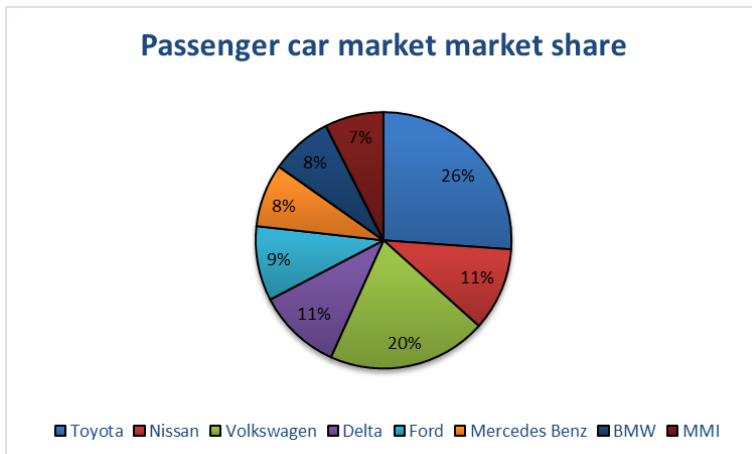
Decision alternative	Stagnant (0,10)	Slow growth (0,45)	Rapid growth (0,30)
Stocks	(\$500)	\$700	\$2 200
Bonds	(\$100)	\$600	\$900
CDs	\$300	\$500	\$750
Mixture	(\$200)	\$650	\$1 300

- a) Draw a decision tree for the investment.
- b) Compute the expected monetary value EMV for the \$10 000 investment.

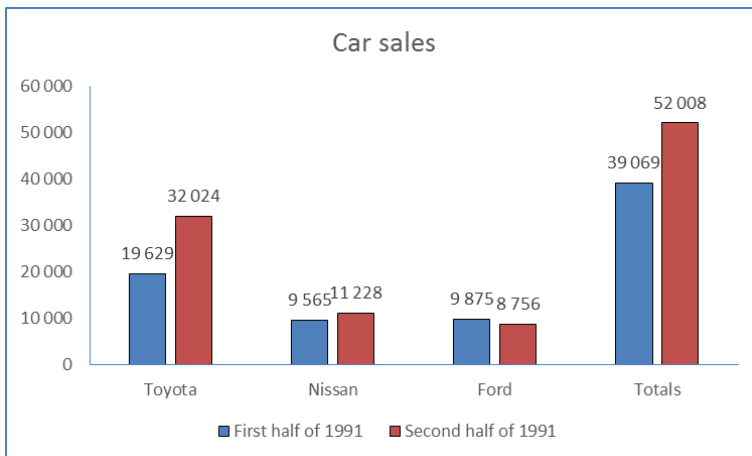
SOLUTIONS TO UNIT EXERCISES

SOLUTIONS TO UNIT 1 EXERCISES

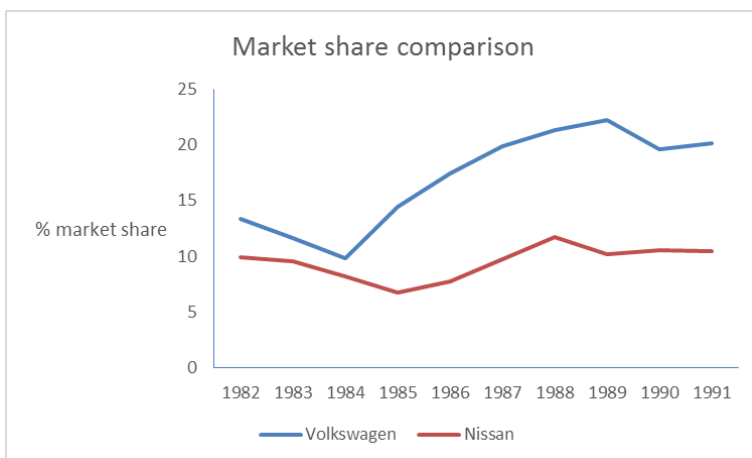
Exercise 1.1



Exercise 1.2

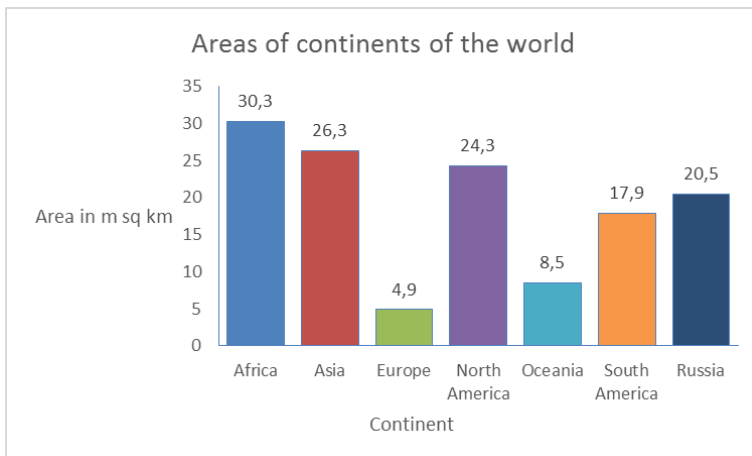


Exercise 1.3

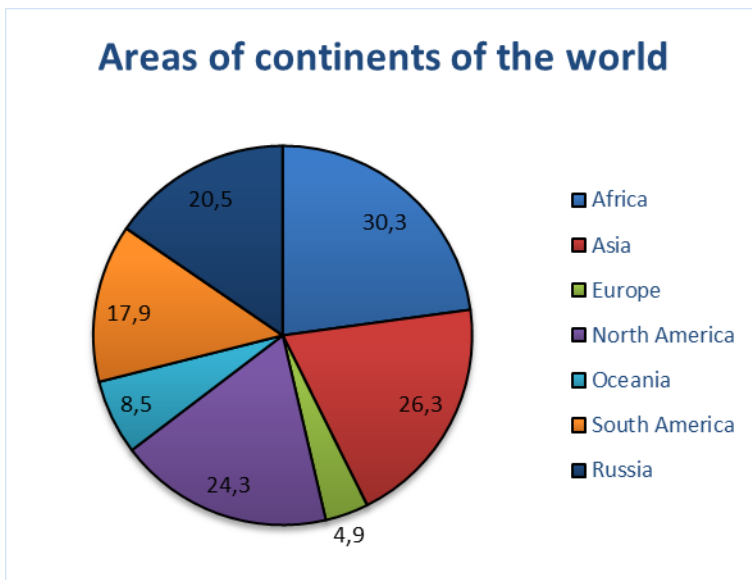


Exercise 1.4

a)



b)



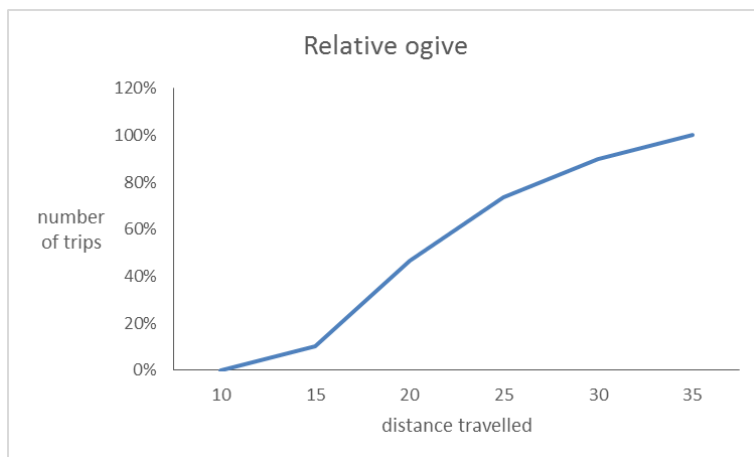
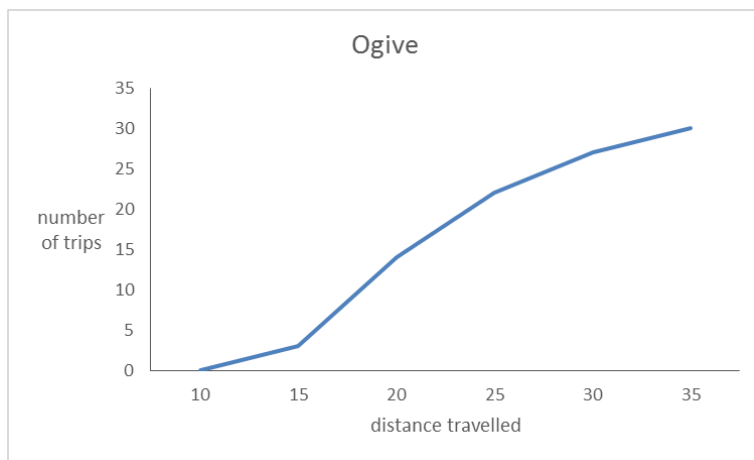
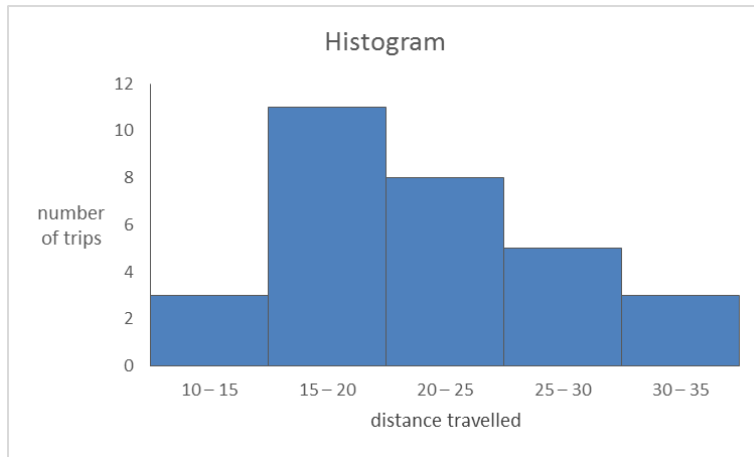
Exercise 1.5

a) Distance travelled (in km); continuous.

b) Frequency distributions:

Distance	Absolute frequency	Relative frequency	Cumulative frequency	Relative cumulative frequency
10 – 15	3	0,1 or 10%	3	0,1 or 10%
15 – 20	11	0,367 or 36,7%	14	0,467 or 46,7%
20 – 25	8	0,267 or 26,7%	22	0,734 or 73,4%
25 – 30	5	0,167 or 16,7%	27	0,9 or 90%
30 – 35	3	0,1 or 10%	30	1 or 100%

c)



d)

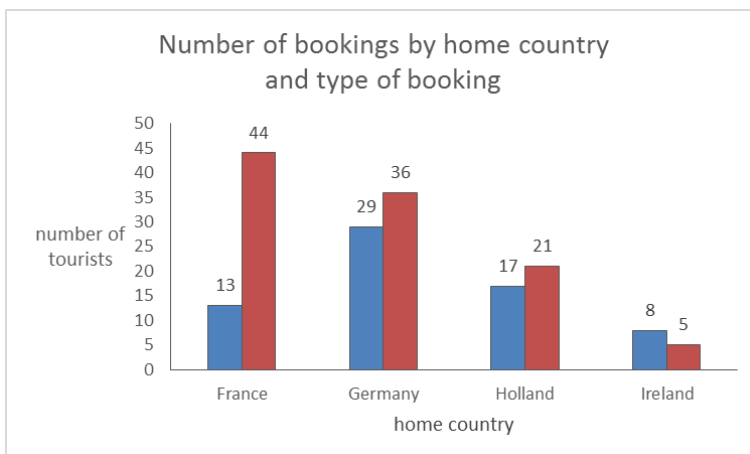
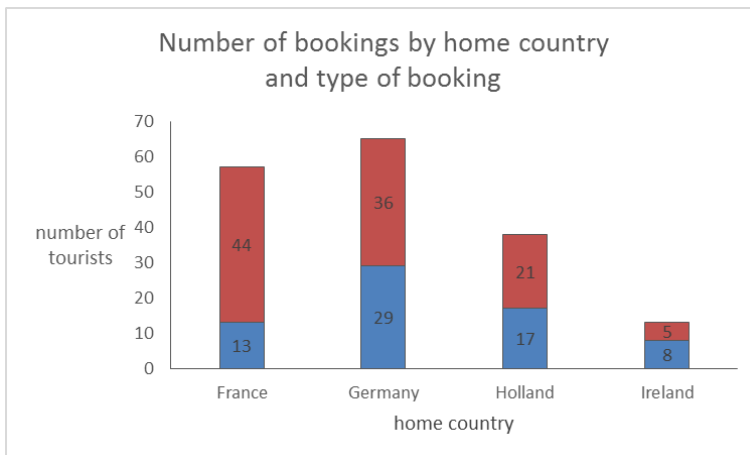
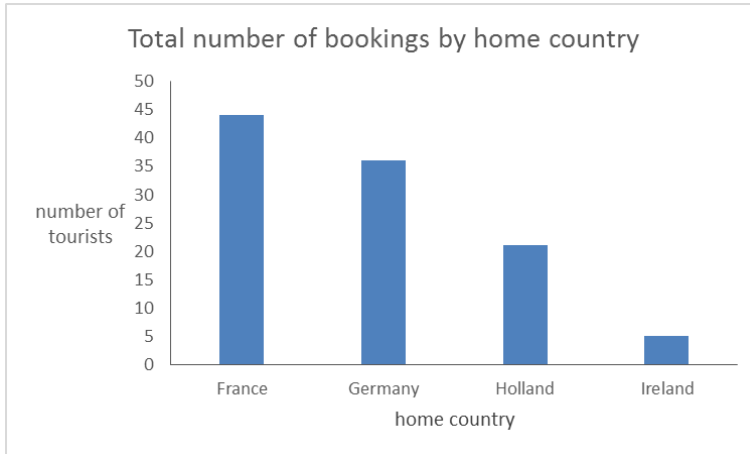
- approx 17%
- approx 73%
- approx 42%
- under 21 km
- above 27 km

Exercise 1.6

a)

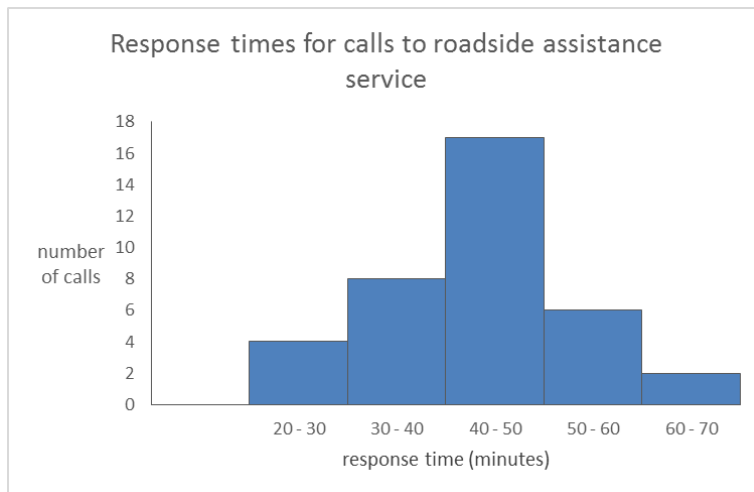
For the simple bar chart, the table needs to be extended to include a total column.

Tourist's home country	Type of booking		Total
	One-week	Two-week	
France	13	44	57
Germany	29	36	65
Holland	17	21	38
Ireland	8	5	13



Exercise 1.7

a)

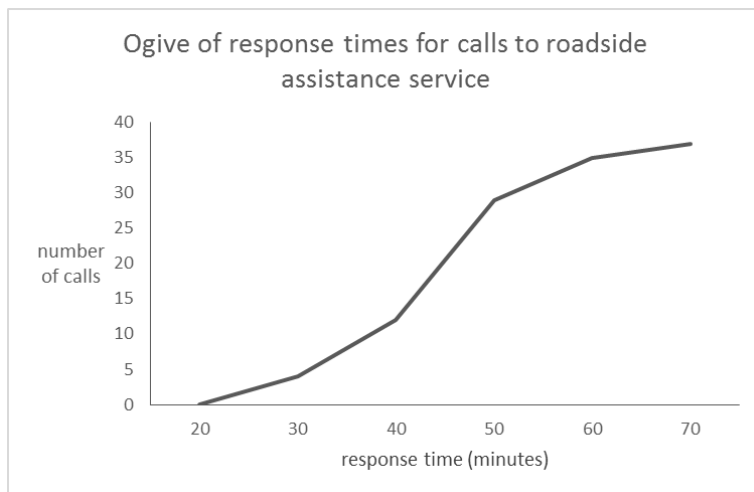


The distribution is broadly symmetrical

(b)

Response time (minutes)	Cumulative frequency
20 – 30	4
30 – 40	12
40 – 50	29
50 – 60	35
60 – 70	37

(c)



SOLUTIONS TO UNIT 2 EXERCISES

Exercise 2.1

(a) Array 7 7 7 7 8 8 8 8 8 8 9 9 9 9 9 9 10 10 10 10

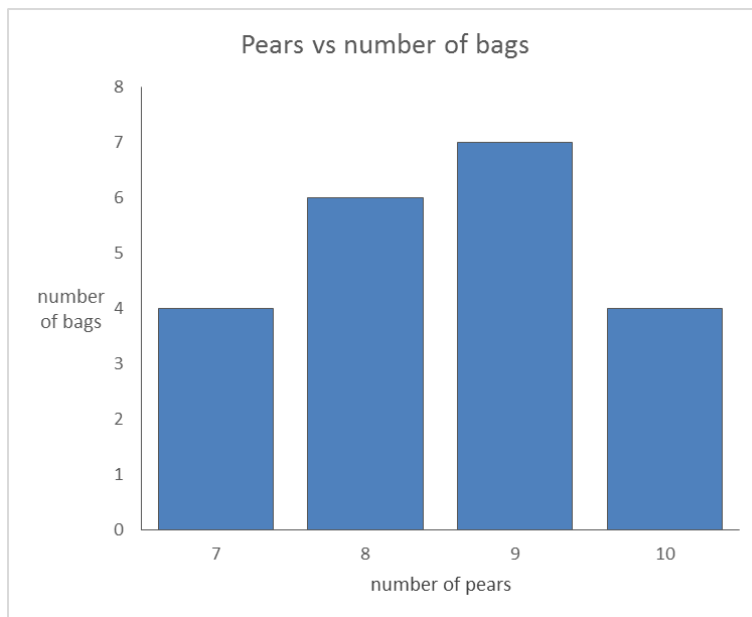
Mode = 9

$$\text{Median} = \frac{21 + 1}{2} = 11\text{th value, } 9 \text{ (in bold)}$$

$$\text{Mean} = \frac{(7 + 7 + 7 + \dots + 10)}{21} = \frac{179}{21} = 8,524$$

(b) Close so fairly symmetrical distribution

(c)



Exercise 2.2 and 2.3

The same format as for exercise 2.1.

Exercise 2.4

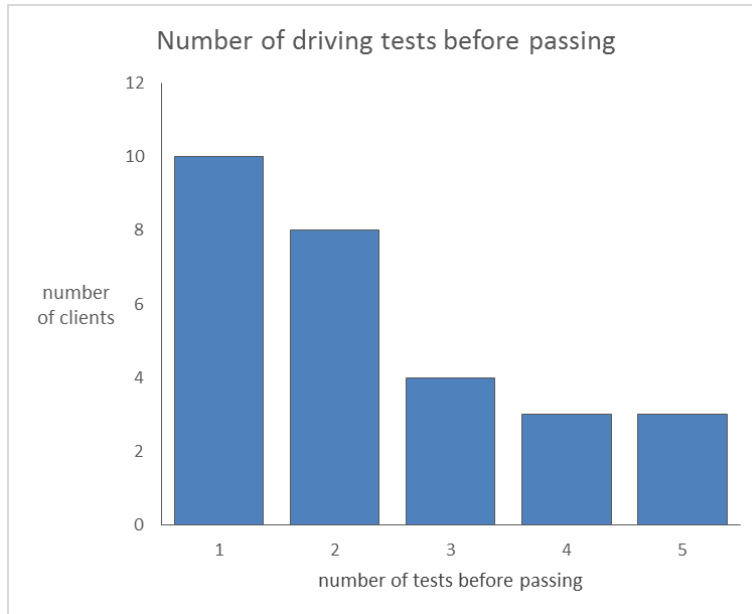
(a) Mode = 1,

Median position = $\frac{28+1}{2} = 14,5\text{th position}$, midway between the 14th and 15th values. 1st to 10th

values are 1, 11th to 18th are 2, so the median is 2

$$\text{Mean} = \frac{[(10 \times 1) + (8 \times 2) + (4 \times 3) + (3 \times 4) + (3 \times 5)]}{28} = \frac{10 + 16 + 12 + 12 + 15}{28} = 20 \text{ } 321$$

(b)



Exercise 2.5

F: Mode = 2, Median = 2, Mean = 2,03

M: Mode= 1, Median = 2, Mean = 2,144

Exercise 2.6

a) Mode = 13, Range = 33-5 = 28

Array: 5 6 7 8 8 9 9 10 11 12 12 12 13 13 **13** 13 15 17 18 18 19 19 20 20 21 22 22 22 23

b) Median = (29 + 1)/ 2 = 15th value, 13 (in bold)

c) Quartile position = (15 + 1) / 2 = 8th value, lower quartile = 10, upper quartile =19

Exercise 2.7

Interval	Midpoint (x)	Frequency (f)	fx	x^2	fx^2
80 – 120	100	3	300	10 000	30 000
120 – 160	140	11	1 540	19 600	215 600
160 – 200	180	9	1 620	32 400	291 600
200 – 240	220	7	1 540	48 400	338 800
240 – 280	260	2	520	67 600	135 200
Σ		32	5 520		1 011 200

$median\ position = (32 + 1)/2 = 16,5th$

median interval is 160 - 200

$median = 160 + \frac{16,5 - 14}{9} \times 40 = 171,111$

SOLUTIONS TO UNIT 3 EXERCISES**Exercise 3.1**

Mean = 33, 2%;

Standard deviation = 11, 0083%

Exercise 3.2

(a) Median (Q_2) = R250, 77

Standard deviation = R90, 90

(b) Q_1 = R190, 00; Q_3 = R330, 46

(c) Quartile deviation = R70, 23

Exercise 3.3

Mean = 34, 60

Standard deviation = 7,088

Exercise 3.4

Variance

Exercise 3.5

a) Mean (#1) = 2,583; Mean (#2) = 3,333

b) Standard deviation (#1) = 0,9014; Standard deviation (#2) = 0,8890

c) CV (#1) = 34,897%; CV (#2) = 26,673%

Exercise 3.6

a) Mean = R45,5; Standard deviation = R14,431

b) Interquartile range = R23,5; QD = R11,75

Exercise 3.7

Mean = $(11 + 31 + 27 + \dots + 20) / 17 = 427 / 17 = 25,118$

$$\begin{aligned} \text{standard deviation, } s &= \sqrt{\frac{1}{n-1} \left(\sum x^2 - \frac{(\sum x)^2}{n} \right)} = \sqrt{\frac{1}{16} \left[(11^2 + 31^2 + \dots + 20^2) - \frac{(427)^2}{17} \right]} \\ &= \sqrt{\frac{1}{16} (11\,391 - 10\,725,235)} = \sqrt{\frac{1}{16} \times 665,765} = 6,451 \end{aligned}$$

Exercise 3.8

Interval	Midpoint (x)	Frequency (f)	fx	x^2	fx^2
80 – 120	100	3	300	10 000	30 000
120 – 160	140	11	1 540	19 600	215 600
160 – 200	180	9	1 620	32 400	291 600
200 – 240	220	7	1 540	48 400	338 800
240 – 280	260	2	520	67 600	135 200
Σ		32	5 520		1 011 200

$$\text{approximate mean} = \frac{5520}{32} = 172,5$$

$$\text{standard deviation} = \sqrt{\frac{1}{31} \times \left(1\,011\,200 - \frac{5520^2}{32} \right)} = 43,626$$

SOLUTIONS TO UNIT 4 EXERCISES**Exercise 4.1**

a)

“Qualifications”; ordinal-scaled; discrete

“Management level”; ordinal-scaled; discrete

b)

Qualification	Section head	Department head	Division head
Matric	28	14	8
Diploma	20	24	6
Degree	5	10	14
Total	53	48	28

- c) (i) $P(\text{matric}) = 0,3876$
(ii) $P(\text{section head} \cap \text{degree}) = 0,0388$
(iii) $P(\text{dept. head} / \text{diploma}) = 0,48$
(iv) $P(\text{division head}) = 0,2171$
(v) $P(\text{division head} \cap \text{section head}) = 0,6279$
(vi) $P(\text{matric} \cup \text{diploma} \cup \text{degree}) = 1,00$
(vii) $P(\text{matric}/\text{dept. head}) = 0,2917$
(viii) $P(\text{division head} \cup \text{diploma}) = 0,5581$
- d) Events in (v): mutually exclusive
Events in (viii): not mutually exclusive

Exercise 4.2

- a) $P(X^1) = 0,20$
b) $P(Y^1) = 0,25$
c) $P(Y^2 \cap X^4) = 0,05$
d) $P(Y^1 \cup Y^2 \cup Y^3) = 1,00$
e) $P(Y^1 \cup Y^2) = 0,52$

Exercise 4.3

- a) $P(F_1/E_2) = 0,75$
b) $P(F_3) = 0,06$
c) $P(E_1 \cup F_2) = 0,54$
d) $P(\bar{A}) = 0,65$

Exercise 4.4

- a) $P(F_3/E_1) = 0,30$
- b) $P(E_2 \cap F_3) = 0,24$

Exercise 4.5

- (a)
 - (i) $P(<age\ 30) = 0,54$
 - (ii) $P(\text{Namibian}) = 0,75$
 - (iii) $P(\text{Namibian} \cup \text{French}) = 0,85$
 - (iv) $P(<age\ 30 \cap \text{Namibian}) = 0,50$
 - (v) $P(<age\ 30 \cup \text{Namibian}) = 0,79$
- (b) $P(E_1 \cup E_2) = 0,18$
- (c)
 - (i) ${}_3P_3 = 6$
 - (ii) 3

Exercise 4.6

- (a)
 - ii) 70 customers
- (b)
 - i) $P(A \cap B) = 0,0006$
 - ii) $P(A \cup B) = 0,0694$
 - iii) $P(\bar{A}) = 0,99$

Exercise 4.7

- (a)
 - i) $P(<25\ years) = 0,34$
 - ii) $P(\text{Production Worker}) = 0,533$
 - iii) $P(\text{Sales} \cap 25-40\ years) = 0,08$
 - iv) $P(>40\ years / \text{Office Worker}) = 0,09$
 - v) $P(\text{Production Worker} \cup <25\ years) = 0,707$
- (b)
 - i) $P(L2 \cup S4) = 0,556$
 - ii) $P(L3)P(S1/L3) = 0,05199$

Exercise 4.8

Refer to study guide

Exercise 4.9

(a)

i) $P(\text{male} \cap < 36 \text{ years}) = 0,4444$

ii) $P(\text{female}) = 0,3611$

iii) $P(\text{female} \cap > 30 \text{ years}) = 0,0$

iv) $P(\text{male} \cap > 21 \text{ years}) = 0,3611$

v) $P(< 21 \text{ years}) = 0,50$

vi) $P(\text{male} / > 30 \text{ years}) = 1,0$

vii) $P(\text{female} \cup 21-30 \text{ years}) = 0,5278$

viii) $P(\text{female} \cup \text{male}) = 1,0$

SOLUTIONS TO UNIT 5 EXERCISES

Exercise 5.1

- (i) $P(r = 0) = 0,2824$
- (ii) $P(r < 3) = 0,8160$

Exercise 5.2

- (i) $P(r = 1) = 0,243$
- (ii) $P(r \geq 1) = 0,271$

Exercise 5.3

- (a)
 - (i) $P(x = 0) = 0,002479$
 - (ii) $P(x \leq 2) = 0,06197$
 - (iii) $P(x \geq 3) = 0,93803$
- (b) $P(x = 0) = 0,04979$
- (c) $\mu = 6$ orders per day; $\sigma = 2,45$ orders per day

Exercise 5.4

$P(x = 2) = 0,0838$

Exercise 5.5

$P(x = 2) + P(x = 3) = 0,3034$

Exercise 5.6

- a) $P(x < 63) = 0,0228$
- b) $P(x > 63,7) = 0,7257$
- c) $P(62,9 < x < 64,3) = 0,7118$
- d) $x = 64,635$
- e) $x = 63$

Exercise 5.7

- a) $P(z > 1.5) = 0,0668$
- b) $P(z < -0.68) = 0,2482$
- c) $P(0 < z < 1.5) = 0,4332$

Exercise 5.8

a) $P(x > 3) = 0,14287$

b) $P(x < 4) = 0,04238$

SOLUTIONS TO UNIT 6 EXERCISES

Exercise 6.1

$$H_0: \mu = 30$$

$$H_1: \mu < 30$$

$$n = 36, \bar{x} = 27, \sigma = 10$$

Test statistic is $z = -1,80$. Critical value at $\alpha = 0,05$ is $-1,645$.

Since $z = -1,80 < -1,645$ H_0 is *rejected* at $\alpha = 0,05$.

We therefore conclude that this group did take *less* time.

Exercise 6.2

$$H_0: \mu = 30$$

$$H_1: \mu < 30$$

$$n = 36, \bar{x} = 27, \sigma = 10$$

Test statistic is $z = -1,80$. Critical value at $\alpha = 0,05$ is $1,96$.

Since $z = -1,80 > -1,96$ H_0 is *not rejected* at $\alpha = 0,05$.

We therefore conclude that this group takes the *same* time as other males.

Exercise 6.3

$$H_0: \mu = 30$$

$$H_1: \mu < 30$$

$$n = 36, \bar{x} = 27, \sigma = 10$$

Test statistic is $z = -1,80$. Critical value at $\alpha = 0,01$ is $-2,33$.

Since $z = -1,80 > -2,33$ H_0 is *not rejected* at $\alpha = 0,01$.

We therefore conclude that this group takes the *same* time as other males.

Exercise 6.4

$$H_0: \mu = 10$$

$$H_1: \mu > 10$$

$$n = 100 \text{ (large)}$$

$$\bar{x} = 10,9$$

$$s = 3,6$$

Test statistic is $z = 2,5$

Since $z = 2,5$ is $> 1,645$ reject H_0 at $\alpha = 0,05$.

No, do *not* believe the claim. Conclude that the average fat content is *more than* 10%.

Exercise 6.5

No.

Two-sided: If $|z| > 1,96$, it does *not* always follow that $|z| > 2$, 58.

One-sided: If $z > 1,645$, it does *not* always follow that $z > 2$, 33.

Exercise 6.6

$$H_0: \mu = 18$$

$$H_1: \mu > 18$$

$$n = 9$$

$$\bar{x} = 15,56$$

$$s = 2,51$$

Test statistic is $t = 2,92$ with $v = 8$.

One-sided critical value at $\alpha = 0,05$ is 1,860.

Since $t = 2,92 > 1,860$, *reject* H_0 at $\alpha = 0,05$.

The complaints *are* justified. Conclude that the mean number of fries is *less than* 18.

Exercise 6.7

Use a *two-sided* test.

$$H_0: \mu = 1,8$$

$$H_1: \mu \neq 1,8$$

$$n = 25$$

$$\bar{x} = 1,65$$

$$\sigma = 0,6$$

Test statistic is $z = 1,25$. Critical value at $\alpha = 0,05$ is 1,96.

Since $|z| = 1,25 < 1,96$, H_0 is not rejected at $\alpha = 0,05$.

Yes, it is *possible* that the establishment is being truthful.

Exercise 6.8 Use a two-sided test.

$$H_0: \mu = 65$$

$$H_1: \mu \neq 65$$

$$n = 30$$

$$\bar{x} = 71$$

$$\sigma = 12,5$$

Test statistic is $z = 2,63$.

Since $|z| = 2,63 > 2,58$, reject H_0 at $\alpha = 0,01$.

Yes, it appears that these TAFE students *did* perform differently from other learner drivers.

Exercise 6.9

$H_0: \pi = 0,60$; $Z_{\text{calc}} = -2,282$; do not reject H_0

Exercise 6.10

$H_0: \mu_x \geq 6,4$; $Z_{\text{calc}} = -0,5143$; do not reject H_0

Exercise 6.11

$H_0: \pi \geq 0,40$; $Z_{\text{calc}} = -3,0618$; Reject H_0

Exercise 6.12

$H_0: \mu_x \geq 30$; $Z_{\text{calc}} = -4,34085$; Reject H_0

Exercise 6.13

$H_0: \mu_x \geq 3500$; $t_{\text{calc}} = -2,222$; Reject H_0

Exercise 6.14

$H_0: \pi \geq 0,15$; $Z_{\text{calc}} = -2,006$; do not reject H_0

Exercise 6.15

$$H_0: \mu_1 - \mu_2 = 0$$

$Z_{\text{calc}} = 2,436$ Reject H_0

Exercise 6.16

$H_0: \pi_1 - \pi_2 = 0$; $Z_{\text{calc}} = 0,65372$; do not reject H_0

Exercise 6.17

$H_0: \mu_x \geq 1$; $t_{\text{calc}} = -1,1838$; do not reject H_0

Exercise 6.18

$H_0: \mu_1 - \mu_2 \leq 0$; $Z_{\text{calc}} = 3,65795$ Reject H_0

Exercise 6.19

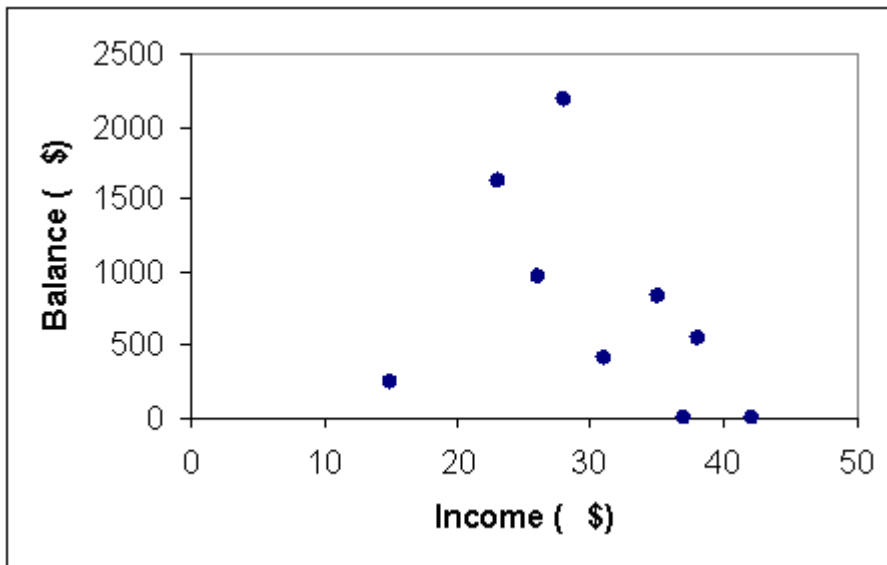
(i) $H_0: \pi_1 - \pi_2 = 0$; $Z_{\text{calc}} = -7,3369$; Reject H_0

(ii) $H_0: \pi_1 - \pi_2 \geq 0$; $Z_{\text{calc}} = -7,3369$; Reject H_0

SOLUTIONS TO UNIT 7 EXERCISES

Exercise 7.1

(a)



(b)

Income (<i>x</i>)	Balance (<i>y</i>)	<i>x</i> ²	<i>y</i> ²	<i>xy</i>	
15	250	225	62500	3750	
23	1630	529	2656900	37490	
26	970	676	940900	25220	
28	2190	784	4796100	61320	
31	410	961	168100	12710	
35	830	1225	688900	29050	
37	0	1369	0	0	
38	550	1444	302500	20900	
42	0	1764	0	0	
275	6830	8977	9615900	190440	n = 9

$$\begin{aligned}
 r &= \frac{9 \times 190440 - 275 \times 6830}{\sqrt{(9 \times 8977 - 275^2) \times (9 \times 9615900 - 6830^2)}} \\
 &= \frac{-164290}{\sqrt{5138 \times 39894200}} = \frac{-164290}{\sqrt{204976399600}} \\
 &= \frac{-164290}{452743.194} \\
 &= -0.363 \qquad \text{weak, negative}
 \end{aligned}$$

Exercise 7.2

- (a) Turnover
- (b) 0,943 strong, positive

Exercise 7.3

- (a) Cost
- (b) 0,907 strong, positive

Exercise 7.4

Critics' rank	Takings rank	d	d ²
1	2	-1	1
2	5	-3	9
3	1	2	4
4	3	1	1
5	4	1	1
6	6	0	0

$$r_s = 1 - \frac{6 \times 16}{6(6^2 - 1)} = 1 - \frac{96}{6 \times 35}$$

$$= 1 - \frac{96}{210} = 1 - 0.457 = 0.543$$

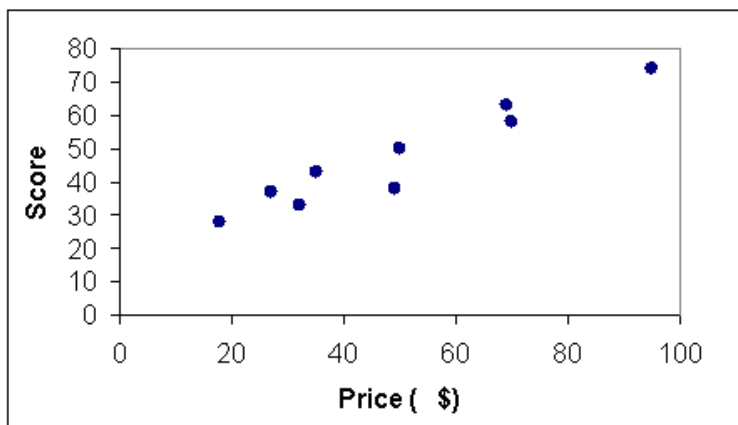
Weak, positive

Exercise 7.5

- (a) 0,952

Exercise 7.6

- (a)



(b)

Price (\$)	Score	x ²	xy	
95	74	9025	7030	
69	63	4761	4347	
18	28	324	504	
32	33	1024	1056	
27	37	729	999	
70	58	4900	4060	
49	38	20401	1862	
35	43	1225	1505	
50	50	2500	2500	
445	424	26889	23863	n = 9

$$b = \frac{23863 - (445 \times 424) / 9}{26889 - (445)^2 / 9} = \frac{23863 - 188680 / 9}{26889 - 198025 / 9}$$

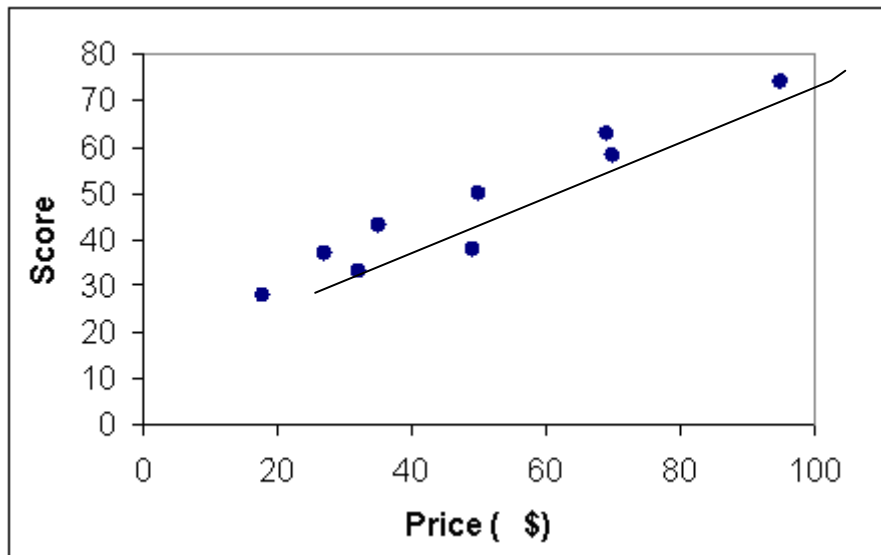
$$= \frac{23863 - 20964.444}{26889 - 22002.778} = \frac{2898.556}{4886.222} = 0.593$$

$$a = (424 - 0.593 \times 445) / 9 = (424 - 263.885) / 9$$

$$= 160.115 / 9 = 17.791$$

$$\text{Score} = 17.791 + 0.593 \text{ price}$$

(c)



$$(d) \text{ Score} = 17,791 + 0,593 (45) = 17,791 + 26,685 = 44,476$$

Exercise 7.7

- (a) Turnover
- (c) Turnover = $-0,027 + 0,113$ Employees
- (d) 0,804 (80.4 %)

Exercise 7.8

- (a) 4 320
- (b) 6 690
- (c) 0,805

SOLUTIONS TO UNIT 8 EXERCISES

Exercise 8.1

(i) $T = 62,875 + 2475x$ where $x = -15$ Q1 1988
 $= -13$ Q2 1988

$= -11$ Q3 1988

(ii) Jan – March 77, 97

Apr – June 115, 98

July – Sept 175, 61

Oct – Dec 30, 44

(iii) Estimated $y = 36, 47$ megawatts

Exercise 8.2

Quarter 1 101, 39

Quarter 2 100, 41

Quarter 3 100, 55

Quarter 4 97, 64

Exercise 8.3

Quarter 1 72, 97

Quarter 2 119, 51

Quarter 3 117, 57

Quarter 4 89, 95

Exercise 8.4

$T = 22, 8$

Exercise 8.5

(i) $T = 5151,454 + 593,009x$ where $x = -1$ in 1984
 $= 0$ in 1985

$= +1$ in 1986

(ii) Trend (1992) = R9 302,513 million

Exercise 8.6

(i) $y = 53,4 + 7,7x$ where $x = -1$ for Bulk Sale 2

$= 0$ for Bulk Sale 3

$= +1$ for Bulk Sale 4

(ii) Expected sales volume = 107,3 units

SOLUTIONS TO UNIT 9 EXERCISES**Exercise 9.1**

The best pay-off available from investing is R100m, from franchising, R50m and from selling, R20m, so according to the maximax rule they should invest.

Exercise 9.2

The worst pay-off available from investing is -R30m, from franchising, R0m and from selling, R20m, so according to the maximin rule they should sell.

Exercise 9.3

If they knew that demand would increase in the future they should choose to invest, but if instead they had chosen to franchise they would be R 40m worse off (R 100m – R 60m) and if they had chosen to sell they would be R 80m (R 100m – R 20m) worse off. These figures are the opportunity losses for the strategies under the increasing demand state of nature.

The complete set of opportunity loss figures are given in Table 1.3.

Table 1.3

Strategy	State of future demand		
	Increasing	Steady	Decreasing
Invest	0	10	50
Franchise	40	0	20
Sell	80	30	0

From Table 1.3 the maximum opportunity loss from investing is R 80m, from franchising, R30m and from selling, R 50m. The minimum of these is the R30m from franchising, so according to the minimax regret decision rule this is the strategy they should adopt.

Exercise 9.4

In this case there are three possible states of nature - increasing, steady and decreasing future demand – so we assign each one probability of one-third. The investing strategy represents a one-third chance of R100m pay-off, a one-third chance of a R40m pay-off and a one-third chance of a -R30m pay-off. To get the EMV of the strategy we multiply the pay-offs by the probabilities assigned to them:

$$EMV(Invest) = \frac{1}{3} \times 100 + \frac{1}{3} \times 40 + \frac{1}{3} \times (-30) = 33,333 + 13,333 + (-10) = 36,666$$

Similarly, the EMVs for the other strategies are:

$$EMV(Franchise) = \frac{1}{3} \times 60 + \frac{1}{3} \times 50 + \frac{1}{3} \times 0 = 20 + 16,667 + 0 = 36,666$$

$$EMV(Sell) = \frac{1}{3} \times 20 + \frac{1}{3} \times 20 + \frac{1}{3} \times 20 = 20$$

According to the equal likelihood approach they should choose to either invest or franchise as both have a higher EMV than selling.

Exercise 9.5

Pay-off table:

	Win	Lose
Standard fee	13 500	-1 500
No win no fee	10 000	0

- (a) Maximum returns: 13 000 (standard fee), 10 000 (no win no fee) Choose standard fee.
- (b) Minimum returns: -1 500 (standard fee), 0 (no win no fee) Choose no win no fee.
- (c) Opportunity losses:

	Win	Lose
Standard fee	0	1 500
No win no fee	3 500	0

Maximum regrets: 1 500 (standard fee), 3 500 (no win no fee) Choose standard fee

(d) $P(\text{win}) = P(\text{lose}) = 0,5$

$$EMV(\text{standard fee}) = 0,5 \times 13\,500 + 0,5 \times (-1\,500) = 6\,000$$

$$EMV(\text{no win no fee}) = 0,5 \times 10\,000 + 0,5 \times 0 = 5\,000, \text{ Choose standard fee.}$$

Exercise 9.6

- (a) Ice Cream (b) Mix (c) Mix (d) Mix

Exercise 9.7

- (a) Don't vaccinate (b) Vaccinate (c) Vaccinate (d) Vaccinate

Exercise 9.8

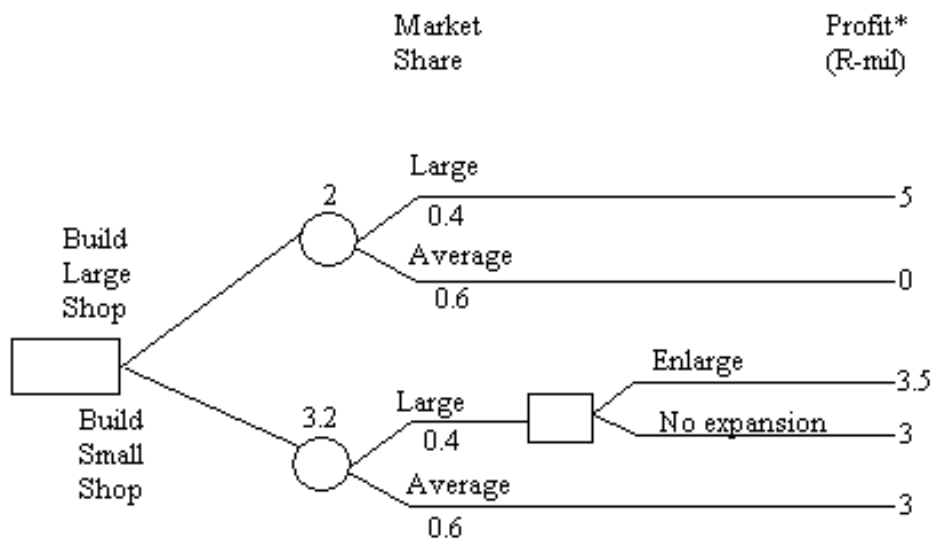
- (a) Don't discount (b) Discount (c) Don't discount (d) Don't discount

Exercise 9.9

- a) Increase (b) Decrease (c) Same (d) Same

Exercise 9.10

a)



**Profit defined as first year's profit less cost of the shop*

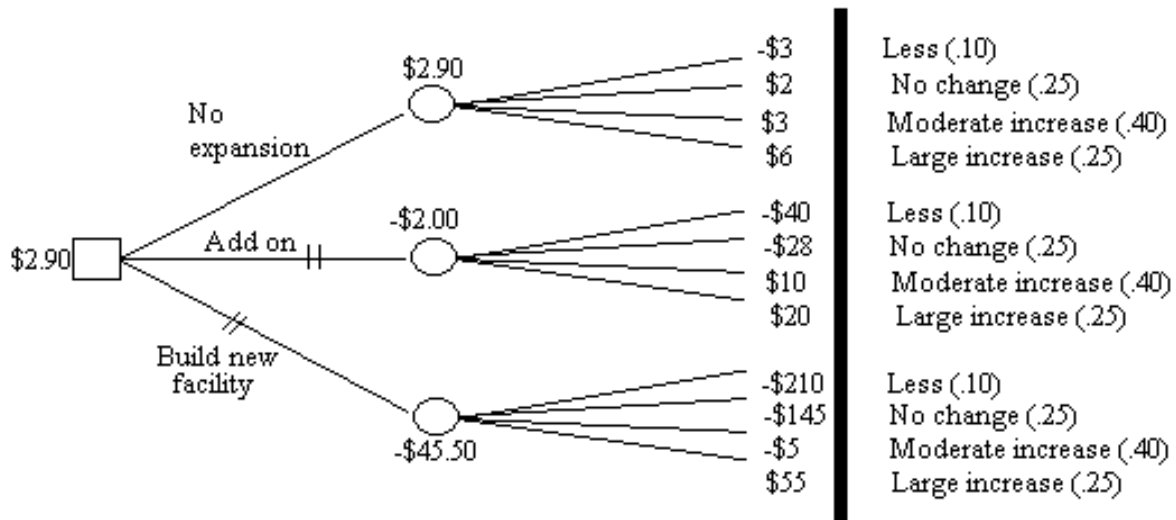
b)

Build the small shop and wait for the result of the advertising campaign. If results are positive, enlarge the shop.

Note: The advertising cost is not relevant to the decision as it will be incurred regardless of the size of the shop.

Exercise 9.11:

a)



b)

The expected monetary value for no expansion is

$$(-\$3)(.10) + (\$2)(.25) + (\$3)(.40) + (\$6)(.25) = \$2,90$$

The expected monetary value for adding on is

$$(-\$40)(.10) + (\$28)(.25) + (\$10)(.40) + (\$20)(.25) = -\$2,00$$

The expected monetary value for building a new facility is

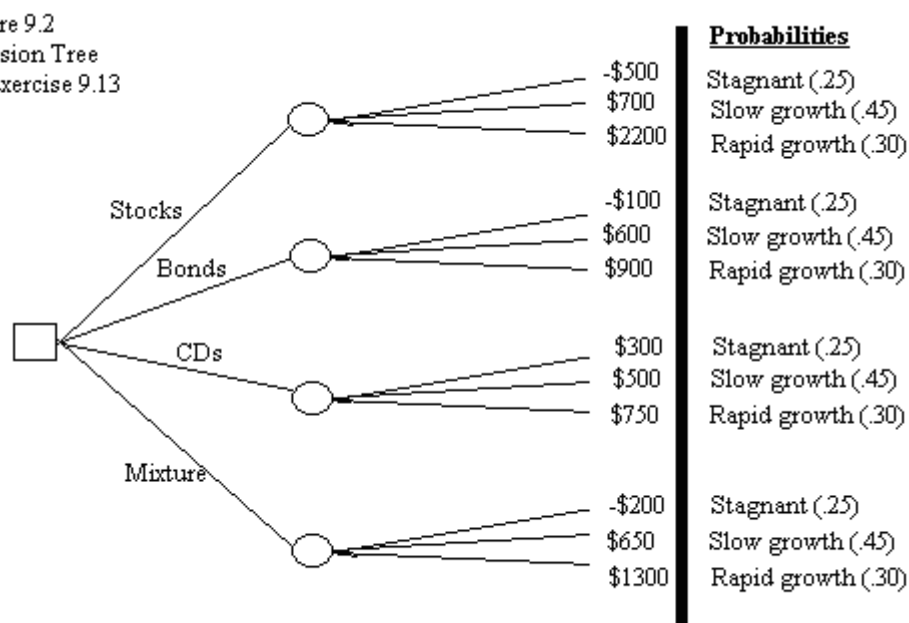
$$(-\$210)(.10) + (\$145)(.25) + (\$5)(.40) + (\$55)(.25) = \$45,50$$

The decision maker who uses the EMV criterion will select the co-expansion decision alternative because it results in the highest long-run average payoff, \$2, 90. It is possible that the decision maker will only have one chance to make this decision at this company. In such a case, the decision maker will not average \$2,90 for selecting no expansion but rather will get a payoff of -\$3,00, \$2,00, \$3,00 or \$6,00, depending on which state of demand follows the decision.

Exercise 9.12

a)

Figure 9.2
Decision Tree
for Exercise 9.13



Expected monetary value computation for the state of the economy

We compute the expected monetary value for the R10 000 investment problem displayed with the associated probabilities. We use the following calculations to find the expected monetary value for the decision alternative stocks.

$$\text{Expected value for stagnant economy} = (.25)(-\$500) = -\$125$$

$$\text{Expected value for slow-growth economy} = (.45)(\$700) = \$315$$

$$\text{Expected value for rapid-growth economy} = (.30)(\$2200) = \$660$$

The expected monetary value of investing in stocks is:

$$-\$125 + \$315 + \$660 = \$850$$

The calculations for determining the expected monetary value for the decision alternative bonds follow.

$$\text{Expected value for stagnant economy} = (.25)(-\$100) = -\$25$$

$$\text{Expected value for slow-growth economy} = (.45)(\$600) = \$270$$

$$\text{Expected value for rapid-growth economy} = (.30)(\$900) = \$270$$

The expected monetary value of investing in bonds is:

$$-\$25 + \$270 + \$270 = \$515$$

The expected monetary value for the decision alternative CDs is found by the following calculations.

$$\text{Expected value for stagnant economy} = (.25)(\$300) = \$75$$

$$\text{Expected value for slow-growth economy} = (.45)(\$500) = \$225$$

$$\text{Expected value for rapid-growth economy} = (.30)(\$750) = \$225$$

The expected monetary value of investment in CDs is:

$$\$75 + \$225 + \$225 = \$525$$

The following calculations are used to find the expected monetary value for the decision alternative mixture.

$$\text{Expected value for stagnant economy} = (.25)(-\$200) = -\$50,00$$

$$\text{Expected value for slow-growth economy} = (.45)(\$6500) = \$292,50$$

$$\text{Expected value for rapid-growth economy} = (.30)(\$1300) = \$390,00$$

The expected monetary value of investment is mixture is:

$$-\$50,00 + \$292,50 + \$390,00 = \$632,50$$

A decision maker using expected monetary value as a strategy will choose the maximum of the expected monetary values computer for each decision alternative.

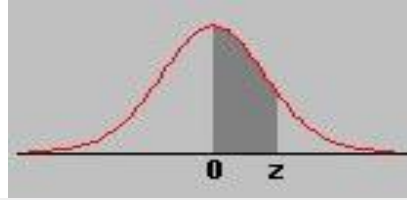
$$\text{Maximum of } \{\$850, \$515, \$525, \$632, 5\} = \$850$$

REFERENCES

- Andre, F (2004). *Business Mathematics and Statistics* (6th edition). Thomson: UK.
- Arsham, H (2006). *Statistical Thinking for Managerial Decisions*.
Retrieved from home.ubalt.edu/ntsbarsh/Business-stat/opre504.htm - 747k. June 2006.
- Bancroft, G and O'Sullivan G (1981). *Mathematics and statistics for Accounting and Business Studies*. McGraw-Hill Book Company (UK) Ltd.
- Black, K (2000) *Business Statistics: Contemporary Decision Making* (3rd edition) South-Western College Publishing – Thomson Learning.
- Buglear, J (2005). *Qualitative methods for business*. Elsevier Butterworth Heinemann.
- Croucher, JS (1998). *Introductory Mathematics and Statistics for Business* (3rd edition). McGraw-Hill Book Company, Sydney.
- Daniel, W. and Terrell, J (1995). *Business Statistics: for management and economics* (7th Edition). Houghton Mifflin Company, Boston.
- Groebner, DF, Shannon, PW, Fry, PC, Smith, KD (2013). *Business Statistics* (9th edition). Pearson.
- Stine, R, Foster, D (2013). *Statistics for Business: Decision Making and Analysis* (2nd edition). Pearson
- Business Donnelly, RA (2013). *Statistics*. Goldey-Beacom College. Pearson.
- Keller, G (2001). *Applied Statistic with Microsoft Excel*. Duxbury Thomson Learning.
- Lind, D Marchal, WG and Wathen, SA (2005). *Statistical Techniques in Business and Economics* (12th edition). New York: McGraw-Hill.
- Steyn, AGW et al (1994). *Modern Statistics in practice*. JL Van Schaik Academic, Pretoria.
- Wegner, T (2012). *Applied Business Statistics Methods and Excel-based Applications*. Juta and Co Ltd. Cape Town.
- Kvanli, AH, Guynes, CS, Pavur, RJ (2000). *Introduction to Business Statistics* (5th edition). South-Western College Publishing

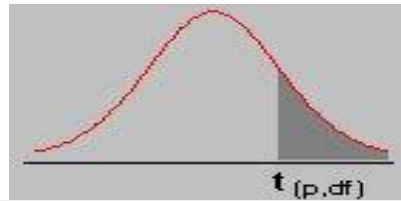
TABLES

TABLE 1: THE STANDARD NORMAL DISTRIBUTION



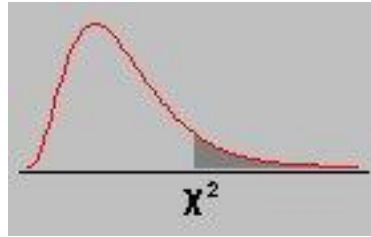
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

TABLE 2: THE T-DISTRIBUTION



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
inf	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905

TABLE 3: THE CHI-SQUARE DISTRIBUTION



df/p	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.64464	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100	13.70069	17.27501	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032	14.84540	18.54935	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.33976	15.98391	19.81193	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.33850	19.36886	23.54183	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	10.86494	13.67529	17.33790	21.60489	25.98942	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.56200	18.33765	22.71781	27.20357	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	13.23960	16.34438	20.33723	24.93478	29.61509	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.13730	22.33688	27.14134	32.00690	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	18.11390	21.74940	26.33634	31.52841	36.74122	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297	46.97924	50.89218	53.67196