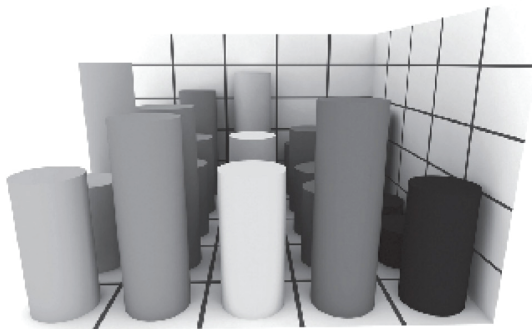


BASIC STATISTICS



only study guide for **STA1510**
XTA1510

T P Mohlala

DEPARTMENT OF STATISTICS

**UNIVERSITY OF SOUTH AFRICA
PRETORIA**

© 2017 University of South Africa

All rights reserved

Printed and published by the
University of South Africa
Muckleneuk, Pretoria

STA1510/1/2018–2020

70662258

Layout done by the Department of Statistics

TABLE OF CONTENTS

INTRODUCTION	1
STUDY UNIT 1	
1.1 Sample versus population	7
1.2 Types of variables	8
1.3 Levels of measurement	9
STUDY UNIT 2	
2.1 Tables and charts for categorical data	11
2.2 Tables and charts for numerical data	11
2.3 Summary	16
STUDY UNIT 3	
3.1 Central tendency	17
3.2 Measures of dispersion	19
3.3 Summary	25
STUDY UNIT 4	
4.1 Introduction	27
4.2 Assigning probability to an event	29
4.3 Calculation of probability	31
4.4 Self-assessment exercise	47
4.5 Solutions to the self-assessment exercise	50
4.6 Summary	52
STUDY UNIT 5	
5.1 Introduction	53
5.2 Probability distribution for discrete random variables	56
5.3 The binomial distribution	61
5.4 The Poisson distribution	65
5.5 Self-assessment exercise	67
5.6 Solutions to the self-assessment exercise	69
5.7 Summary	71

STUDY UNIT 6

6.1	Introduction	73
6.2	Normal and standardised normal distributions	74
6.3	Self-assessment exercise	85
6.4	Solutions to the self-assessment exercise	88
6.5	Summary	92

STUDY UNIT 7

7.1	Introduction	93
7.2	Sampling distribution of the mean	95
7.3	Sampling distribution of the proportion	102
7.4	Self-assessment exercise for section 7.2	106
7.5	Self-assessment exercise for section 7.3	108
7.6	Solutions to the self-assessment exercise for section 7.2	109
7.7	Solutions to the self-assessment exercise for section 7.3	115
7.8	Summary	117

STUDY UNIT 8

8.1	Introduction	119
8.2	Confidence interval estimate for the mean when the population standard deviation is known	121
8.3	Confidence interval estimate for the mean when the population standard deviation is unknown	123
8.4	Confidence interval estimate for the proportion	126
8.5	Self-assessment exercise	128
8.6	Solutions to the self-assessment exercise	129
8.7	Summary	132

STUDY UNIT 9

9.1	Introduction	133
9.2	Fundamental concepts of hypothesis testing	134
9.3	Hypothesis testing for the mean	137
9.4	Hypothesis testing for the proportion	143
9.5	Self-assessment exercise	145
9.6	Solutions to the self-assessment exercise	146
9.7	Summary	149

STUDY UNIT 10

10.1	Introduction	151
10.2	Basic concepts of Chi-square testing	151
10.3	Testing for independence of two variables	152
10.4	Summary	156

STUDY UNIT 11

11.1	Introduction	157
11.2	The simple linear regression line	158
11.3	Introduction to correlation analysis	160
11.4	Summary	162

INTRODUCTION

“Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write”.
---H.G Wells

STATISTICS!

The word *statistics* leads to different reactions ranging from disgust to admiration. Do you believe that statistics is difficult and irrelevant? I sincerely hope not, but if you do, this is just the module that will enable you to confront this feeling and gradually change it to a feeling of astonishment about the need and the strength of statistical theory and the clever scientific applications of statistics in different disciplines. My aim is to convince you that statistics is indeed a timely topic forming an essential part of your intellectual development!

Most probably this module will be the only contact you have with statistics during your studies for your degree or diploma and my aim is to make this introduction to statistics as interesting as possible.

We are all heading somewhere and, being a student, you must be heading for an exciting new career or for an improvement of your academic skills. Never underestimate the power of your mind regarding your final destination with the degree you are enrolled for. Feeling negative about statistics because you are not interested in numbers, or because you do not understand the need for a module in statistics, will influence the way you study as well as the way you progress.

Have an “open mind” for what you are about to learn; let it catch your imagination. I am asking for a real effort, seeing that it will be an investment in your future personal and professional life. Set your mind on success and think in terms of your journey’s end, which is that degree or diploma you want to complete!

Let's GO!

Tell yourself: I want to learn, like and use statistics!

I am sure that there are many questions in your mind and seeing that you are a distance learner, I will anticipate some of those questions:

Question 1

How will I benefit by gaining this knowledge?

Statistical skills take on different forms and without realising it, you are using many of them every day of your life. Soon, everybody will be forced to have a certain level of numeracy at school level and statistics has been introduced at school level as well. Learning concepts by heart has little meaning in statistics, because it is a subject about perception, insight and the ability to apply knowledge. Allow logic to direct you through the rules and results.

How will this benefit you? Statistics will enrich you with knowledge relevant in different walks of life, because it is living knowledge, applicable wherever it fits in. It is definitely not only about collecting information, called data. We will go beyond data and you will become an explorer, turning information into wisdom. After the completion of this module, you should understand more about life, the role of decision making and the importance of scientific knowledge in governance and control.

Question 2

What is the nature of the statistics in this module?

The authors of the prescribed book explain in the preface that their aim was to present statistics in an interesting and useful way. In this they succeeded, and also in their use of modern technological advances. This module is a service module for students from different disciplines and with varying background knowledge. It is therefore different from the more mathematical presentations offered to students from the College of Science, or students from the College of Economic and Management Sciences who are majoring in statistics. This is a stand-alone module, as it may not be a prerequisite for a module at any level in statistics.

Question 3

How should I go about this module?

Keep in mind that different students have different study methods, so it is not really possible to give you an indication of the time you will spend preparing for this module. The time you spend studying is not necessarily correlated to intelligence. The extremely important fact I do want to stress is that Statistics needs continuous, steady attention! You need time for reflection on the knowledge you have attained. Please make a study timetable for all the modules for which you are enrolled, taking the assignment due dates and your personal circumstances into consideration.

Question 4

Can I continue with statistics once I have completed this module?

As said, this is a service module. You cannot present this module for exemption from any other major statistics module at Unisa. Furthermore, this module cannot form part of a major in statistics. We trust that this module will open your mind and create an interest in statistics, but if you want to major in statistics, you will have to start again at level one. The reason for this rather depressing statement lies in the depth of knowledge and the method of presentation in this module, which is too different from the more mathematical presentation typical of modules forming part of a major in statistics.

Question 5

What is a wrap-around (or textbook guide)?

This is a textbook guide you are reading at the moment. It is a way of talking to you in a manner similar to a lecturer at a residential university talking to his/her students. I know that words on a piece of paper can never substitute personal contact, but I will try to come as close to that as possible. In this guide I include summaries on certain sections, discuss difficult sections, indicate the sections that need to form part of your examination preparation and then I also give you what is called activities. They are like worked out examples, but I give you the opportunity to try them yourself before you look at my solutions. The process to follow is as follows:

- Study the particular section in the textbook (given in a block at the beginning of each study unit).
- Read the corresponding section in this textbook guide.
- Attempt to answer the questions in the activity relevant to that section. Do not look at my solutions at the end of each study unit before you have tried really hard to do them yourself.

You may have more questions and if they are serious and you have concerns, contact your lecturer. You may find that you are able to answer your own questions as the year rolls on!

Outline of this module

The basic statistics topics you need to study are all included in the prescribed textbook. In fact, there are additional topics in this textbook which do not form part of this module. A clear indication of the chapters designated for examination purposes is given below, but read the notes on the different study units for finer details.

You have to know the following chapters for examination purposes:

- Chapter 1: Introduction
- Chapter 2: Organizing and visualizing data
- Chapter 3: Numerical Descriptive Measures
- Chapter 4: Basic Probability
- Chapter 5: Discrete Probability Distributions
- Chapter 6: The Normal Distribution
- Chapter 7: Sampling and Sampling Distributions
- Chapter 8: Confidence Interval Estimation
- Chapter 9: Fundamentals of Hypothesis Testing: One-Sample Tests
- Chapter 11: Chi-Square Tests
- Chapter 12: Simple Linear Regression

You cannot do this module without the prescribed book. Also make sure that you buy the correct edition. You should nurse all your prescribed books, even after the

completion of the different modules. They will become precious references in your current or future career. Note that at the time of the development of this module, the policy is that your assignments and examination paper will all contain only multiple-choice questions owing to the large number of students enrolled for this module. The majority of the questions in the activities given in this wrap-around will also be multiple choices. However, some basic principles can be explained in more detail in a standard question.

The content of this module has been divided into 11 study units.

Study material

Your study material consists of

- a prescribed book: *details found in Tut 101*
- this study guide

Notes

- The textbook you have to buy yourself and as soon as possible!
- Tutorial Letter 101, containing general information as well as the assignment questions, forms part of the study material you received during the registration period.

Please make sure that you receive both the study guide and Tutorial Letter 101 during registration. Once you have bought the prescribed book, you will be ready to start with your new and exciting learning. If you have access to the internet, log on to myUnisa and join the discussion forums for the different modules you are enrolled for. Being a distance learner can lead to isolation, so get connected or meet regularly with a peer group which is also registered for this module. Look around you and see if you can find *statistical information* in your community, involve your *parents, friends*, etc.

I wish you all the best in your studies.

TP Mohlala

STUDY UNIT 1

Key questions for this unit

What is statistics?

How do you collect and summarise data?

What types of variables exist?

1.1 SAMPLE VERSUS POPULATION

The first chapter is very general and offers you an excellent background to the information that follows. Most probably you will come across concepts which are new to you; others may seem familiar, but you are not certain about their significance or meaning. Is this the stage where you find statistics boring and a lot of dead facts? Hey, come on! Nothing really good in life comes easy! Before you can use a language, you have to learn the boring vocabulary and only when you can manoeuvre the words into sentences, the language starts to make sense and you can enjoy it in all its beauty! Be patient, there are basic facts and concepts you have to learn! Let us quickly run through some “starters”.

Make sure you understand the meaning of a population and a sample. Remember that the measures for a

- sample are called statistic/s
- population are called parameters

The information in the table below will become more and more clear as we continue with the chapter. If you do not know what is meant by measures of location or spread, use this table for further reference.

Table 1.1 Sample versus population

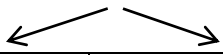
	Sample Subset of the observations	Population All possible observations
Measure <i>Measures of location</i> Average Middle element Most frequent element <i>Measures of spread</i> Range Standard deviation (SD)	Statistic Sample mean Sample median Sample mode Sample range Sample SD	Parameter Population mean Population median Population mode Population range Population SD

This brings us to variables and the difference between qualitative and quantitative variables. The characterisation of data is the starting point of any statistical analysis, so know your data! I would like to train you to evaluate published statistical analyses in order to decide whether results are really trustworthy. This process starts with the data type and the corresponding correct form of analysis.

1.2 TYPES OF VARIABLES

This brings us to variables and the difference between qualitative and quantitative variables. Please understand that there is a difference between types of variables and scales of measurement. Variables are classified as either qualitative (think in terms of quality of life) or quantitative (if you quantify something, you could count it). The information can be given in table form. Once you know your variable is quantitative, it helps to ask yourself whether you have actually counted (then discrete) or measured (then continuous) when you gathered the values.

Table 1.2 Qualitative and quantitative variables

Qualitative variable (several categories)	Quantitative variable 	
	<i>Discrete</i> Only specific values you counted	<i>Continuous</i> Any value within interval you measured
<i>Data</i>	<i>Data</i>	<i>Data</i>
Data only as frequencies (count elements in categories) Frequencies can also be expressed as percentages	Generated by counts of elements	Generated by measurements of some aspect of the elements

1.3 LEVELS OF MEASUREMENT

Once a variable has been measured, you must know how to analyse the data (the measurements put together), but in order to do this, you have to look at the variable under a magnifying glass. Four levels, called scales, of measurement are given. Data are actually either nominal or ordinal. There is little to say about nominal data, but ordinal data can be defined as interval or ratio. Make sure that you understand the difference between the data types.

Table 1.3 Scales of measurement

Nominal	Ordinal		
	\sphericalangle Ordinal	\downarrow Interval	\searrow Ratio
Categories or labels. If numbers are used, they have no numerical meaning.	Preferences are ordered. Numbers are ranked, but ranks do not represent specific measurements.	Numerical labels indicate order and distance. Unit of measurement exists but no absolute zero.	Absolute zero present and multiples have meaning.

Activity 1.1

Question 1

Which of the following statements about the variable type is *incorrect*?

1. Whether or not you own a Panasonic television set is a qualitative variable.
2. Your status as either a full-time or part-time student is a quantitative variable.
3. The number of people you know who attended the graduation last year is a quantitative, discrete variable.
4. The price of your most recent haircut is a quantitative, discrete variable.
5. Cyril's travel time from his home to the examination centre is a quantitative, continuous variable.

Question 2

Which of the following quantitative variables is not continuous (i.e. it is discrete)?

1. Your weight
2. The circumference of your head (in centimetres)
3. The time it takes Jerome to walk from his home to the taxi pickup point

4. The length of your forearm from elbow to wrist
5. The number of coins in your pocket

Feedback on the activity

Question 1

Option 2

1. *Correct.* The options are “yes” and “no”. The variable does not track measurements of elements, but will generate counts of the number of people in each category.
2. *Incorrect.* The options are that you are a full-time or a part-time student and it is a qualitative variable for the same reason as given in 1.
3. *Correct.* The answer is some countable number of people, making the answer a number and this number can take on only whole numbers as values, therefore it is a discrete variable.
4. *Correct.* The answer is a price in rands and cents. These are numbers, therefore quantitative, and furthermore these numbers can extend only to the second decimal place, so it is discrete.
5. *Correct.* The answer is a number, so quantitative, but time can be accurately measured to any level of accuracy. So the quantitative variable is also continuous.

Question 2

Option 5

1. *Continuous.* The answer lies in an interval.
2. *Continuous.* The measurement lies somewhere in an interval of possibilities.
3. *Continuous.* The time it takes Jerome to walk from his home to the taxi pickup point is a measurement.
4. *Continuous.* The length of your forearm from elbow to wrist is a measurement.
5. *Not continuous.* The number of coins in your pocket is a discrete value; they are easy to count.

STUDY UNIT 2

Key questions for this unit

How do you draw up a table for statistical data?

What different graphs exist?

2.1 TABLES AND CHARTS FOR CATEGORICAL DATA

The emphasis in this study unit is on appropriate methods for analysing quantitative data, but it concludes with an introduction to the comparison of qualitative variables through cross-tabulation.

It is frustrating to try and make sense of raw data (simply collected information). Even non-statisticians have the need to do something with such information. The most elementary manipulation would be to arrange data from small to large, or in alphabetical order, or for categorical data you can draw up a summary table and use a bar chart, pie chart or pareto chart to display the data.

2.2 TABLES AND CHARTS FOR NUMERICAL DATA

In the previous section we decided that a frequency distribution is much better than simply writing down the measurements as they are being recorded. However, once data have been placed in the different classes, some of the information is lost. You no longer have the data measurements, only the counts per class. Different charts are used to display the data and one method is the stem-and-leaf display, which is a very sweet visual presentation. No information about the data is lost! The data are also grouped: the stems correspond to the class intervals of the frequency distribution, but the leaves give more detail – they record the actual data values in that class interval. Note that the stem-and-leaf computer printout of MINITAB gives an additional first column indicating a count of the number of leaves per stem. If you read that there are 37 values in this category or lower, but you can see with your own

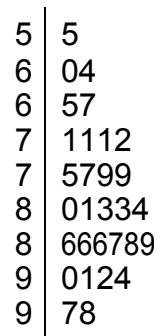
eyes that there are only eight leaves, namely 5, 5, 5, 5, 7, 8, 8, 9 and 9, in that line, do not be upset. Why is this not a lie? In the text it was explained that it is possible to break up each stem into two or more lines. Look again at the printout and you will see that the stems have been doubled: there are two 1s, two 2s, etc. Because there were too many numbers per stem, they split each stem into two parts. For example, for the stem 1 all the numbers from 10 to 14 are written with the first line and all numbers from 15 to 19 are written with the second line. An outlier is a data value very distant from most of the others.

Of course, if the interest is in the *form* of the data (symmetric, skewed), repeating stems may not be used in the stem-and-leaf display.

Activity 2.1

Question 1

The following stem-and-leaf plot gives the ages of people living in Block A of a retirement village:



Which statement is *incorrect*?

1. There are 39 data points in the stem-and-leaf display.
2. The youngest resident is 55 years old.
3. Eight people are between 70 and 80 years old.
4. There are no outliers in this data set.
5. Half of the people are younger than 83.

Question 2

The following stem-and-leaf plot displays a set of values where the stem is formed by the units and the leaf represents the decimal digits.

3		0167
4		333
5		258
6		99
7		4

Which one of the following statements is *incorrect*?

1. The number with the highest frequency is 4.3.
2. There are 13 values being represented in this stem-and-leaf display.
3. The original data are 3.0, 3.1, 3.6, 3.7, 4.3, 5.2, 5.5, 5.8, 6.9 and 7.4.
4. The smallest and largest values are 3.0 and 7.4, respectively.
5. If arranged in increasing order, the middle value will be 4.3.

Three methods to help portray a frequency distribution are the histogram, the percentage polygon and the cumulative percentage polygon.

Activity 2.2

Question 1

The following comments refer to histograms. Identify the *incorrect* statement.

1. Histograms graphically display class intervals as well as class frequencies.
2. Histograms are appropriate for qualitative data.
3. Whereas stem-and-leaf displays are ideal for small data sets, large data sets are better presented in histograms.
4. Histograms are good tools for judging the shape of a data set, provided the sample size is relatively large.
5. Only estimates of the centre, variability and outliers of a data set can be determined from a histogram.

Question 2

The following comments refer to different graphical representations. Which statement is *incorrect*?

1. Adjacent rectangles in a histogram share a common side, while those in a bar chart have a gap between them.
2. A pie chart is a circular display divided into sections based on the number of observations within the segments.
3. With a stem-and-leaf plot the intervals for the stems are restricted in length, but this is not true for a histogram.
4. Histograms as well as bar charts represent frequencies according to the relative lengths of a set of rectangles.
5. Box plots give a direct look at the centre, variability, outliers and shape of a data set.

Feedback on the activities

Activity 2.1

Question 1

Option 1

1. *Incorrect.* There are 30 data points. (We do not count the stems.) You can simply count the digits to the right of the vertical bar.
2. *Correct.* The first stem of 5 is for people in their fifties and the leaf is a five, making the age 55.
3. *Correct.* With stem 7 there are eight values, namely 71, 71, 71, 72, 75, 77, 79 and 79.
4. *Correct.* 55 (the smallest number) is not markedly lower than the second-lowest value of 60. Furthermore 98 (the highest value) is very close to 97 and also not an outlier.
5. *Correct.* Half of the people are indeed younger than 83. There are 30 people, so half of them means 15 people. If you count the ordered ages starting at 55, the 15th person is 81 years old.

Question 2

Option 3

3	0167
4	333
5	258
6	99
7	4

1. *Correct.* The number with the highest frequency is 4.3 because there are three data points with a value of 4.3. (Highest frequency means occurring the most.)
2. *Correct.* Counting the leaves, you will get 13 values. If you have a problem, look at the listed numbers in 3 below and count them.
3. *Incorrect.* The list given below does not repeat the values that occur more than once. The original data values are 3.0, 3.1, 3.6, 3.7, 4.3, 5.2, 5.5, 5.8, 6.9 and 7.4. The repeat values 4.3, 4.3 and 6.9 should also appear in the list and then there are 13 data points.
4. *Correct.*
5. *Correct.* In a stem-and-leaf plot the values are arranged in ascending order. If you have 13 items arranged in order, then item number 7 is in the middle position with 6 items on either side of it. Starting with 3.0, you will find the second value of 4.3 in position 7.

Activity 2.2

Question 1

Option 2

Statement 2 is the only statement that is incorrect, as histograms are appropriate for quantitative data.

Question 2

Option 5

Box plots give a direct look at the centre, variability and outliers, but not shape.

2.4 SUMMARY

Once you have familiarised yourself with study units 1 and 2, you should be able to

- use graphical displays to describe sample data and to gain insight into the nature of the sampled population
- categorise qualitative data and evaluate different graphical displays
- interpret and compare bar charts, pie charts, histograms and stem-and-leaf displays
- communicate the information contained in different statistical summary measures

STUDY UNIT 3

Key questions for this unit

What measures of central tendency are used?

What measures of dispersion are used?

3.1 CENTRAL TENDENCY

There are three measures of central tendency:

- arithmetic mean (an average)
- median (the middle value in an ordered arrangement of the data)
- mode (the value(s) with the highest occurrence in the list of values)

In the discussion on the arithmetic mean you are introduced to mathematical notations such as μ , $\sum x_i$, \bar{x} . These symbols are like little “pictures” and you should read them in that way. If you were in a lecture hall, the lecturer would not say “mew” for μ , but he/she would say “population mean”. Let me give you the words for the symbols, as I trust that this will help you in your learning.

Symbol	Pronunciation	Read as
μ	mew (like a cat)	population mean
Σ	sigma	the sum of
x_i	ex eye	all x -values
$\sum x_i$	sigma ex eye	sum of all the x -values
\bar{x}	ex bar	sample mean
N		total of the population values
n		total of the sample values

You can now imagine the lecturer saying the following sentences:

The population mean is equal to the sum of all the data values in the population

divided by the number of data values in the population for $\mu = \frac{\sum x_i}{N}$.

The sample mean is equal to the sum of all the data values in the sample divided by

the number of data values in the sample for $\bar{x} = \frac{\sum x_i}{n}$.

Note

Be careful not to write $\mu = \frac{\sum x_i}{n}$.

What is wrong?

μ refers to the *population* and you cannot divide by the number of values n in the sample if you are referring to the population parameter.

Writing $\bar{x} = \frac{\sum x_i}{N}$ would be just as disastrous!

Note

It would help if you take note of the following:

- Greek letters are used for the population parameters, for example μ .
- Standard alphabet letters are used for the sample statistics, for example \bar{x} .

The relationship between the mean, median and mode is determined by the shape of the distribution, which can be symmetric, positively skewed or negatively skewed.

Activity 3.1

Question 1

Read the following statements and identify the *incorrect* statement:

1. The mean is one of the most frequently used measures of central tendency.
2. When the mean is greater than the mode, we say it is negatively skewed.
3. When the mean is greater than the median, we say it is positively skewed.

4. When a distribution is bimodal, it will be impossible for the mean, median and both modes to be equal.
5. The measure most affected by extreme values is the mean.

Question 2

Certain measures have been calculated for the following small sample data set:

15 17 23 11 9 20 45 9 13

The *incorrect* calculation is the following:

1. The mean for this data is double the value of the mode.
2. The value of 45 is an outlier.
3. The mode for this data set is 9.
4. The median is 15.
5. After removing the outlier, the mean is 13.

3.2 MEASURES OF DISPERSION

Describing data using only the mean, median and mode can lead to disaster. Remember that statistical analysis should be based on all the measurements, but because this is not practical in most of the cases, we make summaries. These summaries have to represent all the data in the best possible way, which implies that more than a description of the centre values of the data set is needed.

The measures of spread discussed in this section are the

- range (the difference between the highest and lowest values)
- quartiles (a division of the data in four equal-sized groups)
- variance/standard deviation (an indication of the average deviation from the mean)

The *range* is a concept that is easy to understand and very few students have problems with this.

The reference to *quartiles* is much more complex, as the division of the data into equal groups lead to very important features of the data set. The interquartile range plays a very important role in the analysis of statistical data. The best application of

quartiles is found in a box plot. Make sure that you understand and are able to interpret a given box plot.

The *variance* and *standard deviation* are but one mathematical calculation apart, but in general, people prefer to refer to standard deviation and not to variance. You need to know how to calculate these measures and understand their meaning as you come across them in different walks of life.

As we move through the prescribed book, you will find that the number of symbols you have to recognise increases. Variance and standard deviation have their own symbols and again there are clear distinctions between the sample and the population measures. Do you want another summary and sentences?

Symbol	Pronunciation	Read as
σ	sigma (same as for Σ)	population standard deviation
σ^2	sigma squared	population variance
s		sample standard deviation
s^2		sample variance
$\frac{\sum (x_i - \mu)^2}{N}$	population variance Note division by N .	sum of the squares of the differences between the values and the population mean, divided by the total number of population values
$\frac{\sum (x_i - \bar{x})^2}{n-1}$	sample variance Note division by $(n - 1)$.	sum of the squares of the differences between the values and the sample mean, divided by the total number of sample values minus one
$\sqrt{\frac{\sum (x_i - \mu)^2}{N}}$	population standard deviation	square root of the population variance
$\sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$	sample standard deviation	square root of the sample variance

Note

A reminder that

- Greek letters are used for the population parameters, for example σ and σ^2 .
- standard alphabet letters are used for the sample statistics, for example s and s^2 .

Make sure that you will be able to draw and interpret a box plot.

For the interpretation and uses of the standard deviation, Chebyshev's theorem and the Empirical Rule are described.

Activity 3.2

Question 1

Given below are the summary statistics for data described as fastest ever driven.

Suppose the speed is given in kilometres per hour, then:

	Males 87 students			Females 102 students		
Median		110			89	
Quartiles	95		120	80		95
Extremes	55		150	30		130

Determine

1. the fastest speed driven by anyone in the group
2. the slowest speed driven by a male
3. the cut-off speed indicating that 25% of the men drove at that speed or faster
4. the proportion of females who had driven 89 km/h or faster
5. the number of females who had driven 89 km/h or faster
6. the differences (if any) between male and female drivers (also interpret your answers)
7. the range for males and females
8. the interquartile range for male and female drivers
9. the distribution of male and female data by drawing two box plots

Question 2

Which of the following statements is *correct*?

1. The range is found by finding the difference between the high and low values and dividing the answer by 2.
2. The interquartile range is found by finding the difference between the first and third quartiles and dividing this value by 2.
3. The mean is a measure of the deviation in a data set.
4. The standard deviation is expressed in terms of the original units of measurement but the variance is not.
5. The median is a measure of dispersion.

Feedback on the activities

Activity 3.1

Question 1

Option 2

When the mean is greater than the mode, we say that it is positively skewed.

Question 2

Option 5

1. *Correct.* The mean of the given data is 18 and the mode is 9 (there are two 9s) and we all know that 18 is double 9.
2. *Correct.* The value nearest to 45 is 23 and in general the data consist mostly of much smaller numbers, so 45 can be considered an outlier.
3. *Correct.* We have already discussed the mode and saw that it is 9.
4. *Correct.* To determine the median, the data must be ordered (from small to large or vice versa): 9, 9, 11, 13, 15, 17, 20, 23, 45
If there are nine values, the middle one is in position five (four values on each side). In this position we have the 15.
5. *Incorrect.* Remove 45 and the total is 117, which must be divided by 8. The answer should have been 14.625 ($117 \div 8$). If you are interested, the incorrect answer given was calculated by dividing 117 by 9 instead of 8.

Activity 3.2

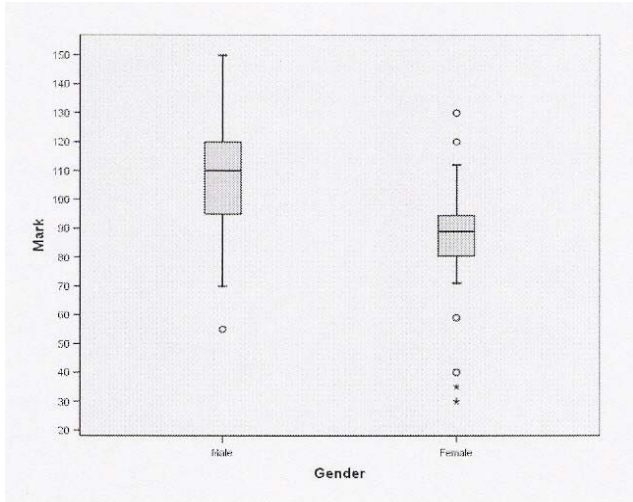
Question 1

	Males 87 students			Females 102 students		
Median		110			89	
Quartiles	95		120	80		95
Extremes	55		150	30		130

- The fastest speed driven by anyone in the group is 150 km/h. Top speeds for males and females are indicated in the table as extremes. (Maximum for females is 130 km/h.)
- The slowest speed driven by a male is 55 km/h.
- The cut-off speed indicating that 25% of the men drove at that speed or faster implies the value of the upper quartile for males, which is 120 km/h.
- To find the proportion of females who had driven 89 km/h or faster, you have to notice that 89 is the value of the female median. The median divides the data into two equal parts, so the proportion of the data above 89 is 50% expressed as a percentage, or 0.5 expressed as a fraction.
- The *number* of females who had driven 89 km/h or faster is (as said) half of the number of females.
If there are 102 female students, half of them will be 51.
- Use the table to interpret the differences between male and female drivers. Some of the obvious differences are the following:
 - The median speed for males is higher than that for females.
 - The highest speed was recorded by a male.
 - The lowest speed was recorded by a female (which is what can be expected if the mean values differ the way they do).
 - The upper quartile of females is the same as the lower quartile of the males (speed 95).
- Range
Males: $(150 - 55) = 95$
Females: $(130 - 30) = 100$

8. Interquartile range for males and females are $(120 - 95) = 25$ and $(95 - 80) = 15$, respectively.

9.



Question 2

Which of the following statements is *correct*?

Option 4

1. *Incorrect.* The range is found by finding the difference between the high and low values – not divided by 2.
2. *Incorrect.* The interquartile range is found by finding the difference between the first and third quartiles – not divided by 2.
3. *Incorrect.* The mean is a measure of *central tendency*.
4. *Correct.* The standard deviation is expressed in terms of the original units of measurement, but the variance is not.
5. *Incorrect.* The median is a measure of *central tendency*.

3.3 SUMMARY

Once you have familiarised yourself with this chapter, you should be able to

- compare the considerations when using the mean, median and mode of quantitative data
- understand the influence of outliers on the mean, median and mode and their correlation with the shape of the distribution
- evaluate the meaning of dispersion as conveyed by the range, the quartiles, MAD and variance/standard deviation
- say whether values in a given data set are outliers or not with special reference to a box-and-whisker plot
- use the standard deviation and mean to determine the coefficient of variation for both sample and population

STUDY UNIT 4

Key questions for this unit

Define probability. Distinguish between the three types of probability.

What is meant by the concepts “event”, “joint event”, “and complement of an event” and “sample space”?

Under what conditions is $P(A/B) = P(A)$?

What does it mean if we say that two events are mutually exclusive?

Why can mutually exclusive events not also be independent?

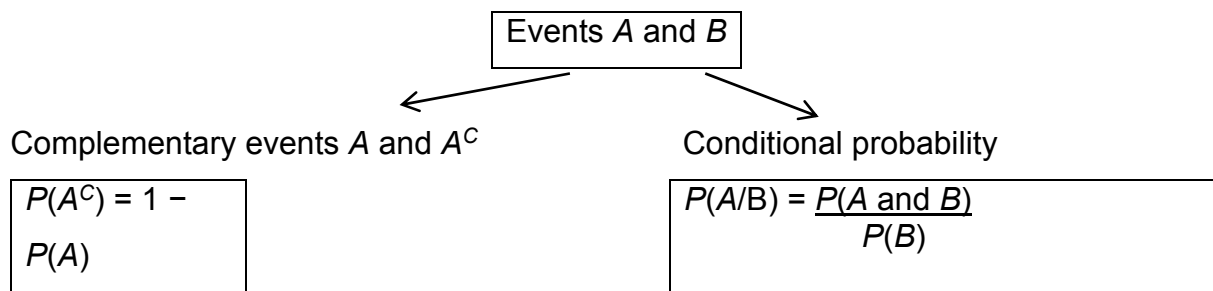
4.1 INTRODUCTION

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 make simple probability conclusions in our daily lives. Sometimes these determinations are based on facts, while others are subjective. If the probability of an event is high, one would expect that it would occur rather than not occur. If the probability of rain is 95%, it is more likely that it would rain than not rain.

The principles of probability help bridge the words of descriptive statistics and inferential statistics. Reading this unit will help you learn different types of probabilities, how to calculate probability and how to revise probabilities in light of new information. Probability principles are the foundation for the probability distribution, the concept of mathematical expectation and the binomial and Poisson distributions, topics that are discussed in study unit 5.

Activity 4.1: Overview Study skill

Draw a mind map of the different sections/headings you will deal with in this study unit. Then page through the unit with the purpose of completing the map.

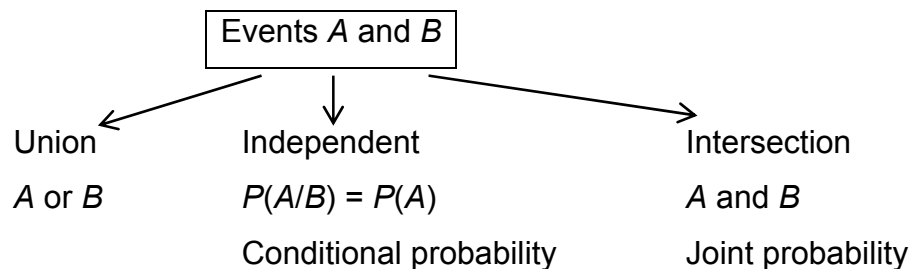


Events A and B are mutually exclusive

↳ $P(A \text{ and } B) = 0$

Events A and B are independent

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



Probability rules

<p>Multiplication rule</p> $P(A \text{ and } B) = P(A/B) \times P(B)$ <p>If A and B are INDEPENDENT, then</p> $P(A \text{ and } B) = P(A) \times P(B)$	<p>Addition rule</p> $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ <p>If A and B are MUTUALLY EXCLUSIVE, then $P(A \text{ or } B) = P(A) + P(B)$</p>
---	---

Activity 4.2: Concepts Conceptual skill Communication skill

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Probability		
Event		
Joint event		
Exhaustive event		
Venn diagram		
Complement of event		
Sample space		
Simple probability		
Joint probability		
Marginal probability		
General addition rule		
Conditional probability		
Decision trees		
Independence		
Multiplication rule		

4.2 ASSIGNING PROBABILITY TO AN EVENT

This section describes procedures for assigning probabilities to events and outlines the basic requirements that must be satisfied by probabilities assigned to simple events. Probabilities can be assigned to simple events (or, for that matter, to any events) using the classical approach, the relative frequency approach or the subjective approach.

Whatever method is used to assign probabilities to the simple events that form a sample, two basic requirements must be satisfied:

1. Each simple event probability must have a value from 0 to 1.
2. The sum of the probabilities assigned to the simple events in a sample space must be equal to 1.

The probability of any event A is then obtained by summing the probabilities assigned to the simple events contained in A .

How do I know whether I should combine two events A and B using “and” or “or”?

Solution:

The key here is to fully understand the meaning of the combined statement.

$P(A \text{ and } B)$ = probability that A and B will both occur, while $P(A \text{ or } B)$ = probability that A or B or both will occur. Sometimes it will be necessary to reword the statement of a given event so that it conforms to one of the two expressions given above. For example, suppose your friend Rajab is about to write two exams and you define the events as follows:

A : Rajab will pass the statistics examination.

B : Rajab will pass the accounting examination.

The event “Rajab will pass at least one of the two exams” can be reworded as “Rajab will either pass the statistics exam or he will pass the accounting exam, or he will pass both exams”. This new event can therefore be denoted by $(A \text{ or } B)$.

On the other hand, the event “Rajab will not fail either exam” is the same as “Rajab will pass both his statistics exam and his accounting exam”. This event can therefore be denoted by $(A \text{ and } B)$.

Example 4.1

An investor has asked his stockbroker to rate three stocks (A , B and C) and list them in the order in which she would recommend them. Consider the following events:

L : Stock A doesn't receive the lowest rating.

M : Stock B doesn't receive the lowest rating.

N : Stock C receives the highest rating.

- (i) Define the random experiment and list the simple events in the sample space.
- (ii) List the simple events in each of the events L , M and N .
- (iii) List the simple events belonging to each of the following events: L or N , L and M , and M .
- (iv) Is there a pair of mutually exclusive events among L , M and N ?
- (v) Is there a pair of exhaustive events among L , M and N ?

Solution:

- (i) The random experiment consists of observing the order in which the stockbroker recommends the three stocks. The sample space consists of the set of all possible orderings:
 $S = \{ABC, ACB, BAC, BCA, CAB, CBA\}$
- (ii) $L = \{ABC, ACB, BAC, CAB\}$; $M = \{ABC, BAC, BCA, CBA\}$; $N = \{CAB, CBA\}$
- (iii) The event $(L \text{ or } N)$ consists of all simple events in L or N or both;
 $(L \text{ or } N) = \{ABC, ACB, BAC, CAB, CBA\}$. The event $(L \text{ and } M)$ consists of all simple events in both L and M ; $(L \text{ and } M) = \{ABC, BAC\}$. The complement of M consists of all simple events that do not belong to M ; $M' = \{ACB, CAB\}$.
- (iv) No, there is not a pair of mutually exclusive events among L , M and N , since each pair of events has at least one simple event in common.
 $(L \text{ and } M) = \{ABC, BAC\}$
 $(L \text{ and } N) = \{CAB\}$
 $(M \text{ and } N) = \{CBA\}$
- (v) Yes, L and M are an exhaustive pairs of events, since every simple event in the sample space is contained either in L or M or both. That is, $(L \text{ or } M) = S$.

4.3 CALCULATION OF PROBABILITY

Probability can be regarded as a fraction. In a multiple-choice test, a typical question has five possible answers. If an examination candidate makes a random guess on one such question, what is the probability that the response is wrong?

Solution:

The probability is $\frac{4}{5}$, because out of the five answers there are four ways to answer incorrectly. Each question can be represented as follows:

Type of answer	Number	Probability
Correct	1	$\frac{1}{5}$
Incorrect	4	$\frac{4}{5}$
Total	5	$\frac{5}{5}$ or 1

Before we calculate the probabilities, let us first discuss the meaning of the words.

At least two: This means that two is the minimum value and if we say at least two children, it means two or three or four or ... children.

$$P(X \geq 2) = P(x = 2) + P(x = 3) + P(x = 4) + \dots$$

At most two: This means that two is the maximum value. At most two children means no child or one child or two children.

$$P(X \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

No more than two: This means that two is the maximum number, that is, two or one or zero children.

$$P(X \leq 2) = P(x = 0) + P(x = 1) + P(x = 2)$$

Less than two: This means that two is not included and we are only interested in the values smaller than two that is, zero or one.

$$P(X < 2) = P(x = 0) + P(x = 1)$$

More than two: This means that two is not included and we are only interested in the values greater than two that is, three, four, five, etc.

$$P(X > 2) = P(x = 3) + P(x = 4) + P(x = 5) + \dots$$

Example 4.2

Consider the following table in which wild azaleas were classified by colour and by the presence or 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, which one of the following probabilities is *incorrect*?

1. $P(\text{a fragrance}) = \frac{130}{200}$

2. $P(\text{colour is orange}) = \frac{68}{200}$

$$3. \quad P(\text{is orange and has a fragrance}) = \frac{58}{200}$$

$$4. \quad P(\text{is orange known that it has a fragrance}) = \frac{58}{130}$$

$$5. \quad P(\text{has a fragrance given that it is orange}) = \frac{58}{130}$$

Solution:

Option 1: *Correct*

Option 2: *Correct*

Option 3: *Correct*

Option 4: *Correct*

Option 5: *Incorrect* – $P(\text{has a fragrance given that it is orange}) = \frac{58}{68}$

General addition rule

When two events A and B occur simultaneously, the general addition rule is applied for finding $P(A \text{ or } B)$ = probability that event A will occur or event B will occur or both will occur.

Formula: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Example 4.3

Consider the table of wild azaleas (example 4.2). If event A denotes that the flower is

orange and event B that it has a fragrance, then: $P(A \text{ or } B) = \frac{68}{200} + \frac{130}{200} - \frac{58}{200} =$

$$\frac{140}{200}$$

Note that the word OR in probability theory denotes ADDITION.

If A and B cannot occur simultaneously, then $P(A \text{ and } B) = 0$.

Events A and B are mutually exclusive if they cannot occur simultaneously.
--

Example 4.4

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

A: 41% B: 9% AB: 4% O: 0.46%

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

$$P(A \text{ or } B \text{ or } AB) = P(A) + P(B) + P(AB) = 0.41 + 0.09 + 0.04 = 0.54$$

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

For many exclusive events $A_1, A_2, A_3, \dots, A_n$, the addition rule may be written as:

$$P(A_1 \text{ or } A_2 \text{ or } A_3 \text{ or } \dots \text{ or } A_n) = P(A_1) + P(A_2) + P(A_3) + \dots + P(A_n)$$

Multiplication rule

The multiplication rule finds the probability that events *A* and *B* will both occur. Two events are independent if one may occur irrespective of the other. For example, event *A*, the patient has tennis elbow, and event *B*, the patient has appendicitis, are intuitively independent.

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

For many independent events $A_1, A_2, A_3, \dots, A_n$, the multiplication rule can be written as:

$$P(A_1 \text{ and } A_2 \text{ and } A_3 \dots \text{ and } A_n) = P(A_1) \times P(A_2) \times P(A_3) \times \dots \times P(A_n)$$

Note that in probability the word AND denotes MULTIPLICATION.

Example 4.5

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.

The probability is $0.6 \times 0.6 \times 0.6 \times 0.6 \times 0.6 = 0.078$.

Calculation of objective probabilities (sections 4.1 and 4.2 of the textbook)

Objective probabilities can be classified into three categories. These categories are

- marginal probability
- joint probability
- conditional probability

The definition and calculation of each type is described next.

Marginal probability

A *marginal* probability is the probability of only a single event A occurring. It is written as $P(A)$. A single event is an event that describes the outcomes of only one random variable. A frequency distribution describes the occurrence of only one characteristic of interest at a time and it is used to estimate marginal probabilities.

Example 4.6

In table 4.1 the random variable industry type is described by the frequency distribution in the second column.

Table 4.1

Industry type	Number of JSE firms
Mining	35
Finance	72
Service	10
Retail	33

Let B = event (finance). Then $P(B) = \frac{72}{150} = 0.48$.

Joint probability

A joint probability is the probability of both event A and event B occurring simultaneously on a given trial of a random experiment. A joint event describes the behaviour of two or more random variables (i.e. the characteristics of interest) simultaneously. It is written as:

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

Example 4.7

Table 4.2

Industry	Company size (in R million turnover)			
	Small (0 to less than 10)	Medium (10 to less than 50)	Large (50 and above)	Total
Mining	0	0	35	35
Finance	9	21	42	72
Service	6	3	1	10
Retail	14	13	6	33
Total	29	37	84	150

Let A = event (small company) and B = event (finance company).

In the sample, there are 9 out of 150 JSE listed companies which are both small and finance companies.

Then $P(A \text{ and } B) = \frac{9}{150} = 0.06$.

Conditional probability

Conditional probability is the probability of one event A occurring given information about the occurrence of another event B . A conditional event describes the behaviour of one random variable in light of known additional information about a second random variable. Conditional probability is defined as:

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

The essential feature of the conditional probability is that the sample space is reduced to the outcomes describing event B (the given prior event) only, and not all possible outcomes as for marginal and joint probabilities.

Example 4.8

- Let A = event (large company)
- Let B = event (retail company)

Then the probability of selecting a company from the JSE sample which is large *given* that the company is known to be a retail company is calculated as follows:

Retail	Small	Medium	Large	Total
	14	13	6	33

There are 6 large companies out of 33 retail companies (refer to table 4.2).

$$\text{Therefore } P(A/B) = \frac{6}{33} = 0.1818.$$

Using the formula:

$$P(A \text{ and } B) = \frac{6}{150} \text{ (a joint probability)}$$

$$P(B) = \frac{33}{150} \text{ (a marginal probability)}$$

$$\text{Then } P(A/B) = \frac{\frac{6}{150}}{\frac{33}{150}} = \frac{6}{33} = 0.1818 \text{ (a conditional probability)}$$

A conditional probability, denoted $P(A/B)$ is the probability that an event A will occur *given* that we know that an event B has already occurred.

The key to recognising a conditional probability is to look for the phrase “given that” or its equivalent. For example, the statement of a conditional probability might read “The probability that A will occur when B occurs” or “The probability that A will occur if B occurs”. In each of these cases, you can reword the statement using “given that” instead of “when” or “if”. Therefore, both of these statements refer to conditional probability.

Probability rules

This section outlines three rules of probability that allow you to calculate the probabilities of three special events [A , (A or B) and (A and B)] from known probabilities of various related events. The three rules are as follows:

1. *Complement rule:* $P(A') = 1 - P(A)$
2. *Addition rule:* $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
3. *Multiplication rule:* $P(A \text{ and } B) = P(A)P(B/A)$ or $P(A \text{ and } B) = P(B)P(A/B)$

We note that the addition rule and the multiplication rule can be expressed more simply under certain conditions.

If A and B are *mutually exclusive* events, then $P(A \text{ and } B) = 0$, so the addition rule becomes:

$$P(A \text{ or } B) = P(A) + P(B)$$

If A and B are independent events, then the multiplication rule becomes:

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

Activity 4.3 Application skills

Question 1

A soft drink company holds a contest in which a prize may be revealed on the inside of the bottle cap. The probability that each bottle cap will reveal a prize is 0.1 and winning is independent from one bottle to the next. What is the probability that a customer will win a prize when opening his third bottle?

1. $(0.1)(0.1)(0.9) = 0.009$
2. $(0.9)(0.9)(0.1) = 0.081$
3. $(0.9)(0.9) = 0.81$
4. $1 - (0.1)(0.1)(0.9) = 0.991$
5. $(0.9)(0.9)(0.9) = 0.729$

Question 2

Suppose two people each have to select a number from 00 to 99 (therefore 100 possible choices). The probability that they will both pick the number 13 is

1. $\frac{2}{100}$

2. $\frac{1}{100}$

3. $\frac{1}{200}$

4. $\frac{1}{10\,000}$

5. $\frac{2}{10\,000}$

Question 3

Use the same information as in question 2. The probability that both people will pick the same number is equal to

1. $\frac{2}{100}$

2. $\frac{1}{100}$

3. $\frac{1}{200}$

4. $\frac{1}{10\,000}$

5. $\frac{2}{10\,000}$

Question 4

For three mutually exclusive events the probabilities are as follows:

$P(A) = 0.2$, $P(B) = 0.7$ and $P(A \text{ or } B \text{ or } C) = 1.0$. The value of $P(A \text{ or } C)$ is equal to

1. 0.3

2. 0.5

3. 0.9

4. 0.6

5. 0.1

Feedback on the activity

Question 1

Option 2

The first two bottles must definitely not reveal a prize on the inside of the bottle top.

The probability of not winning with one bottle is $1 - 0.1 = 0.9$. The probability not to win with two bottles is $(0.9)(0.9) = 0.81$. The probability of winning with the third bottle is $(0.9)(0.9)(0.1) = 0.081$.

Question 2

Option 4

For the first person to select the number 13 the probability is $\frac{1}{100}$. The choice of the

second person is independent of the first selection and the probability to select 13 for the second person is also $\frac{1}{100}$. To find the probability of two independent events, use

the rule $P(A \text{ and } B) = P(A) \times P(B) = \frac{1}{100} \times \frac{1}{100} = \frac{1}{10\,000}$.

Question 3

Option 2

In the previous question we found the probability that both will select one specific

number (13) is $\frac{1}{10\,000}$. For this question, any doubling of numbers is considered for

the probability. There can be two 1s, two 2s, ..., two 99s and if you count these possibilities, there are 100 such double combinations.

$P(\text{select the same number})$ is $100 \times \frac{1}{10\,000} = \frac{1}{100}$.

Question 4

Option 1

Given that $P(A \text{ or } B \text{ or } C) = 1.0$ and the fact that these events are mutually exclusive.

Therefore $P(A) + P(B) + P(C) = 1.0$. Filling in the given probabilities we get

$0.2 + 0.7 + P(C) = 1.0$. If we solve this for $P(C)$, then $P(C) = 0.1$. Therefore

$P(A \text{ or } C) = P(A) + P(C) = 0.2 + 0.1 = 0.3$.

Tree diagrams

The following contingency table shows sampled data for four provinces in South Africa and three types of music people listen to. We are going to use this table to illustrate different types of probability.

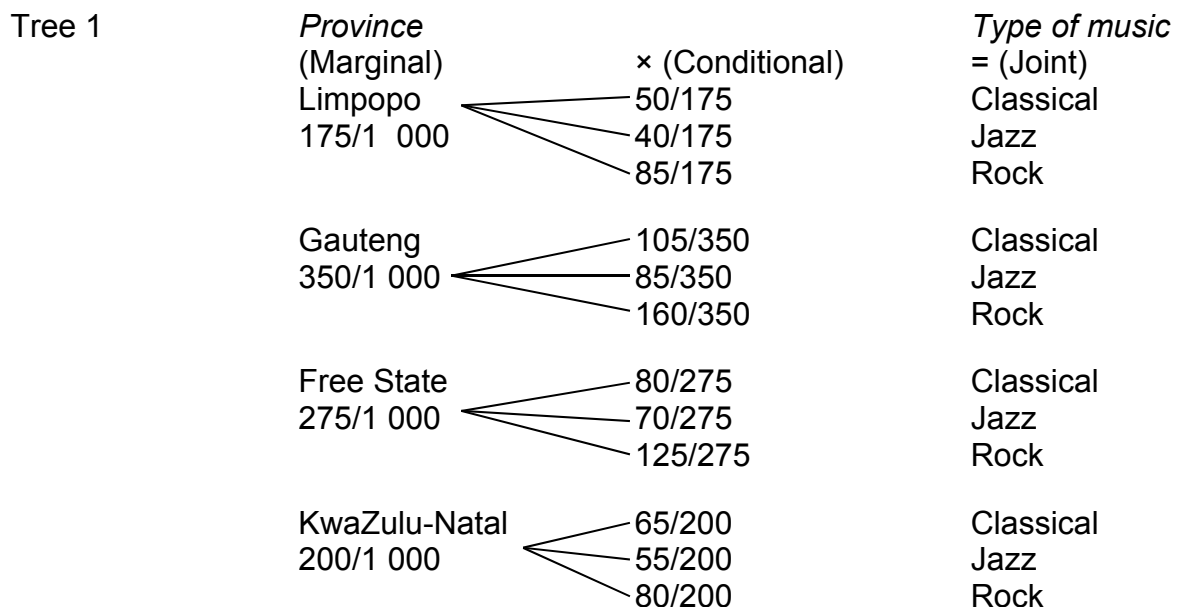
	Classical	Jazz	Rock
Limpopo	50	40	85
Gauteng	105	85	160
Free State	80	70	125
KwaZulu-Natal	65	55	80

To use the table, we have to total the number of people in each province and the number of people who listen to each type of music.

	Classical	Jazz	Rock	Total
Limpopo	50	40	85	175
Gauteng	105	85	160	350
Free State	80	70	125	275
KwaZulu-Natal	65	55	80	200
Total	300	250	450	1 000

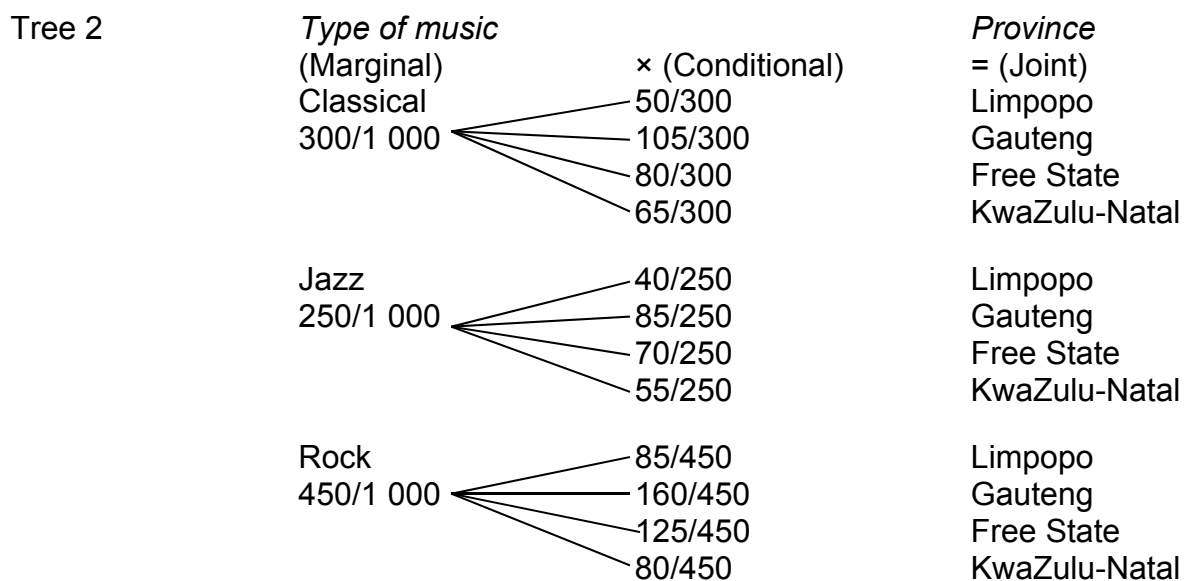
Tree diagrams

The information given in the contingency table above can just as well be given in a probability tree as follows.



Notice that the provincial probabilities are all marginal probabilities and the probabilities for each type of music in each of the four provinces are all conditional probabilities. So the probability that someone listens to classical music, for example, changes depending on the province in which he/she lives.

The probability tree we developed above shows the provinces first and the types of music preferred as contingent upon the province in which someone lives. But suppose we want to show the tree the other way around, that is, suppose we want to show the types of music first and the provincial identifications as contingent upon the types of music preferred.



Notice now that the music types are marginal probabilities and the probabilities for each province are all conditional probabilities. So the probability that someone lives in Limpopo, for example, changes depending on which type of music he/she listens to.

The joint probabilities can be calculated directly from the contingency table or using the tree diagrams, for example:

$$\begin{aligned}
 P(\text{Classical and Gauteng}) &= P(\text{Classical}) \times P(\text{Gauteng/Classical}) \\
 &= \frac{300}{1\,000} \times \frac{105}{300} \\
 &= \frac{105}{1\,000}
 \end{aligned}$$

Question 1

Which statement is *incorrect*?

1. A marginal probability is the probability that an event will occur regardless of any other events.
2. A joint probability is the probability that two or more events will all occur.
3. If $P(A) = 0.8$ and $P(B) = 0.5$ and $P(A \text{ and } B) = 0.24$, we can conclude that events A and B are mutually exclusive.
4. Given the same information as in option 3, the events A and B cannot be independent.
5. Two events cannot be mutually exclusive as well as independent.

Question 2

A study was conducted at a small college among first-year students living on campus. A number of variables were measured. The table below provides information regarding the number of roommates and end-of-term health status for the first-year students at this college. Health status for individuals is measured as poor, average and exceptional.

Health status	Number of roommates		
	None	One	Two
Poor	15	36	65
Average	35	94	40
Exceptional	50	50	25

Which one of the following statements is *incorrect*?

1. The probability that a randomly selected first-year student with no roommates will have poor end-of-term health status is 0.15.
2. The probability that a randomly selected first-year student with 1 roommate will have poor end-of-term health status is 0.20.
3. The events $H = \{\text{student has poor health status}\}$ and $N = \{\text{student has no roommates}\}$ are mutually exclusive.

4. The events $H = \{\text{student has poor health status}\}$ and $N = \{\text{student has no roommates}\}$ are dependent.
5. If you find a person with average health, you can be 52% sure that he/she comes from a room with only one roommate.

Question 3

In the Barana Republic there are two producers of fridges, referred to as Cool and Dry. Assume that there are no fridge imports. The market shares for these two producers are 70% for Cool and 30% for Dry. One executive at Dry proposes a longer warranty period to be offered at a slight extra cost in order to increase market share. A market research company appointed by Dry conducts a census and finds the following: 50% of the owners of a fridge made by Cool like the proposal, 30% are indifferent to it, while the remaining owners oppose it. Among the owners of a fridge made by Dry, 70% like the proposal, 20% are indifferent to it and the remaining owners oppose it.

- a. A fridge owner is selected at random. What is the probability that the person will own a fridge made by Dry?
- b. A fridge owner is selected at random. What is the probability that the owner will be opposed to the proposal of a new warranty at an extra cost?

Hint: Make a tree diagram.

Feedback on the activity

Question 1

Option 3

1. *Correct.* A marginal probability is the probability that an event will occur regardless of any other events.
2. *Correct.* A joint probability is the probability that two or more events will all occur.
3. *Incorrect.* For mutually exclusive events $P(A \text{ and } B) = 0$.
4. *Correct.* Given the same information as in option 3, the events A and B cannot be independent. If events A and B were independent, then

$P(A \text{ and } B) = P(A) \times P(B) = 0.7 \times 0.6 = 0.42$. However, the problem states that $P(A \text{ and } B) = 0.35$, not 0.42.

5. *Correct.* Two events cannot be mutually exclusive as well as independent.

Question 2

Option 3

1. To be able to answer this question, it is advisable that you add the rows and columns of the given table.

Health status	Number of roommates			Total
	None	One	Two	
Poor	15	36	65	116
Average	35	94	40	169
Exceptional	50	50	25	125
Total	100	180	130	410

Correct. Of the 100 students with no roommates, 15 had poor health and

$$\frac{15}{100} = 0.15.$$

2. *Correct.* Of the 180 students with 1 roommate, 36 had poor health and

$$\frac{36}{180} = 0.20.$$

3. *Incorrect.* If these two events were to be mutually exclusive, the cell where “poor health” and “no roommates” cross should have had a zero (and there is a 15).

4. *Correct.* Use the multiplication rule to prove that the events are not independent. Test whether $P(H \text{ and } N) = P(H) \times P(N)$.

$$P(H \text{ and } N) = \frac{15}{410} = 0.037$$

$$P(H) \times P(N) = \frac{116}{410} \times \frac{100}{410} = 0.069 \neq 0.037$$

This implies that the two events are dependent.

5. *Correct.* This is a conditional probability, which you can simply read from the table as $\frac{94}{180} = 0.5222$.

You can also calculate it with the formula $P(A/B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{\frac{94}{410}}{\frac{180}{410}} = 0.5222$.

Question 3

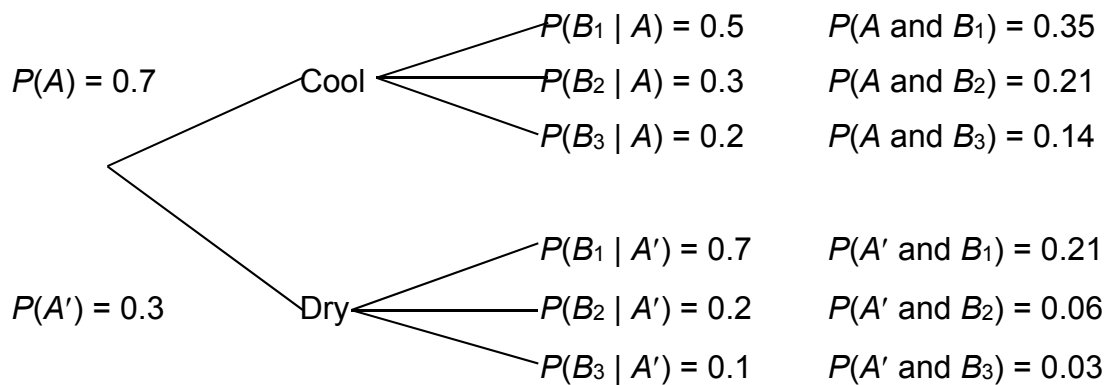
The following tree diagram gives the question information in a concise manner:

Let A = Cool fridge and A' = Dry fridge.

B_1 = like proposal

B_2 = indifferent to proposal

B_3 = oppose proposal



- a. The probability that the person will own a fridge made by Dry is equal to 0.30.
- b. The probability that the owner will be opposed to the proposal of a new warranty at an extra cost is $0.03 + 0.14 = 0.17$.

4.4 SELF-ASSESSMENT EXERCISE

Question 1

Which statement is *correct*?

1. Probability takes on a value from 0 to 1.
2. Probability refers to a number which expresses the chance that an event will occur.
3. Probability is zero if the event A of interest is impossible.
4. The sample space refers to all possible outcomes of an experiment.
5. All the above statements are *correct*.

Question 2

Assume that X and Y are two independent events with $P(X) = 0.5$ and $P(Y) = 0.25$.

Which of the following statements is *incorrect*?

1. $P(X') = 0.75$
2. $P(X \text{ and } Y) = 0.125$
3. $P(X \text{ or } Y) = 0.625$
4. X and Y are not mutually exclusive.
5. $P(X|Y) = 0.5$

Question 3

Refer to the following contingency table:

Event	C_1	C_2	C_3	C_4	Total
D_1	75	125	65	35	300
D_2	90	105	60	45	300
D_3	135	120	75	70	400
Total	300	325	200	150	1 000

Which one of the following statements is *incorrect*?

1. $P(C_1 \text{ and } D_1) = 0.075$
2. $P(D_1) = 0.3$
3. $P(C_1 \text{ or } D_1) = 0.6$
4. $P(D_3/C_4) = 0.4667$
5. $P(C_4/D_3) = 0.175$

Question 4

A survey asked people how often they exceed speed limits. The data were then categorised into the following contingency table showing the relationship between age group and response.

		Exceed limit if possible		
		Always	Not always	Total
Age	Under 30	100	100	200
	Over 30	40	160	200
	Total	140	260	400

Which one of the following statements is *incorrect*?

1. Among the people over 30, the probability of always exceeding the speed limit is 0.20.
2. Among the people under 30, the probability of always exceeding the speed limit is 0.5.
3. The probability that a randomly chosen person is over 30 and will not always exceed the speed limit is 0.4.
4. 10% of the people in the survey always exceed the speed limit.
5. Among the people who always exceed the speed limit 71.43% are under 30.

Question 5

Numbers 1, 2, 3, 4, 5, 6, 7, 8, 9 are written on separate cards. The cards are shuffled and the top one turned over. Let A = an even number and B = a number greater than

6. Which one of the following statements is *incorrect*?

1. The sample space is $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$
2. $P(A) = 4/9$
3. $P(B) = 1/3$
4. $P(A \text{ and } B) = 1/9$
5. $P(A \text{ or } B) = 7/9$

Question 6

If A and B are independent events with $P(A) = 0.25$ and $P(B) = 0.60$, then $P(A/B)$ is equal to

1. 0.25
2. 0.60
3. 0.35
4. 0.85
5. 0.15

Question 7

Given that $P(A) = 0.7$, $P(B) = 0.6$ and $P(A \text{ and } B) = 0.35$, which one of the following statements is *incorrect*?

1. $P(B') = 0.4$
2. A and B are not mutually exclusive.
3. A and B are dependent.
4. $P(B/A) = 0.6$
5. $P(A \text{ or } B) = 0.95$

Question 8

The Burger Queen Company has 124 locations along the west coast. The general manager is concerned about the profitability of the locations compared to major menu items sold. The information below shows the number of each menu item selected by profitability of store.

	Baby Burger M_1	Mother Burger M_2	Father Burger M_3	Nachos M_4	Tacos M_5	Total
High profit R_1	250	424	669	342	284	1 969
Medium profit R_2	312	369	428	271	200	1 580
Low profit R_3	289	242	216	221	238	1 206
Total	851	1 035	1 313	834	722	4 755

If a menu order is selected at random, which statement is *incorrect*?

1. $P(M_5) = 0.1518$
2. $P(R_3) = 0.0501$
3. $P(R_2 \text{ and } M_3) = 0.0900$
4. $P(M_2/R_2) = 0.2335$
5. $P(R_1/M_4) = 0.4101$

4.5 SOLUTIONS TO THE SELF-ASSESSMENT EXERCISE

Question 1

Option 5

Question 2

$$P(X') = 1 - P(X) = 1 - 0.5 = 0.5$$

$$P(X \text{ and } Y) = P(X) \times P(Y) = 0.5 \times 0.25 = 0.125$$

$$P(X \text{ or } Y) = P(X) + P(Y) - P(X \text{ and } Y) = 0.5 + 0.25 - 0.125 = 0.625$$

$P(X \text{ and } Y) = 0.125 \neq 0$, therefore X and Y are not mutually exclusive.

$$P(X|Y) = P(X \text{ and } Y)/P(Y) = \frac{0.125}{0.25} = 0.5$$

Option 1

Question 3

$$P(C_1 \text{ and } D_1) = \frac{75}{1000} = 0.075$$

$$P(D_1) = \frac{300}{1000} = 0.3$$

$$P(C_1 \text{ or } D_1) = 0.3 + 0.3 - 0.075 = 0.525$$

$$P(D_3|C_4) = \frac{70}{150} = 0.4667$$

$$P(C_4|D_3) = \frac{70}{400} = 0.175$$

Option 3

Option 5

Question 4

$$P(\text{always}|\text{over 30}) = \frac{40}{200} = 0.20$$

$$P(\text{always}|\text{under 30}) = \frac{100}{200} = 0.5$$

$$P(\text{over 30 and not always}) = \frac{160}{400} = 0.4$$

$$\text{Percentage people always exceeding} = \frac{140}{400} = 0.35 = 35\%$$

$$P(\text{under 30}|\text{always}) = \frac{100}{140} = 0.7143 = 71.43\%$$

Option 4

Question 5

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = \frac{4}{9} + \frac{1}{3} - \frac{1}{9} = \frac{6}{9} = \frac{2}{3}$$

Option 5

Question 6

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A) = 0.25$$

Option 1

Question 7

$$P(B') = 1 - P(B) = 0.4$$

$P(A \text{ and } B) \neq 0$, A and B are not mutually exclusive events.

$P(A)P(B) \neq P(A \text{ and } B)$, therefore A and B are dependent events.

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{0.35}{0.7} = 0.5$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.7 + 0.6 - 0.35 = 0.95$$

Option 4

Question 8

$$P(M_5) = \frac{722}{4\,755} = 0.1518$$

$$P(R_3) = \frac{1\,206}{4\,755} = 0.2536$$

$$P(R_2 \text{ and } M_3) = \frac{428}{4\,755} = 0.0900$$

$$P(M_2|R_2) = \frac{369}{1\,580} = 0.2335$$

$$P(R_1|M_4) = \frac{342}{834} = 0.4101$$

Option 2

4.6 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to

- link random circumstances and probability to everyday life
- define a sample space, an event and complementary events
- grasp the idea of a Venn diagram displaying the sample space and the events within
- differentiate between the union and intersection of events
- differentiate between marginal, joint and conditional probability
- clarify the difference between mutually exclusive and independent events
- appreciate the different fundamentals of counting in probability
- describe and apply the basic rules for probability
- use contingency tables and tree diagrams to solve more complex questions on probability

STUDY UNIT 5

Key questions for this unit

Define a discrete probability distribution.

How would you construct a probability distribution for a discrete random variable?

Distinguish between discrete and continuous random variables.

How would you calculate the expected value and the variance of a discrete random variable?

How would you calculate the expected value and the variance of a binomial distribution?

How would you calculate the expected value and the variance of a Poisson distribution?

5.1 INTRODUCTION

In study unit 4 you learnt much about probability in general. In this study unit we discuss discrete random variables and their probability distributions. Probability distributions are classified as either discrete or continuous, depending on the random variable. A random variable is a variable that can take on different values according to the outcome of an experiment. It is described as random because we don't know ahead of time exactly what value it will have following the experiment. For example, when we toss a coin, we don't know for sure whether it will land heads or tails. Likewise, when we measure the diameter of a roller bearing, we don't know in advance what the exact measurement will be. Random variables are either discrete or continuous. In this unit the emphasis is on discrete random variables and their probability distributions. In the next unit we will cover random variables of the continuous type.

- A random variable is discrete if it can assume only a countable number of possible values (0, 1, 2, 3, ...).
- A continuous random variable assumes an uncountable number of possible values; it can take on any value in one or more intervals of values. Levine et al provide the following definition of a probability function:

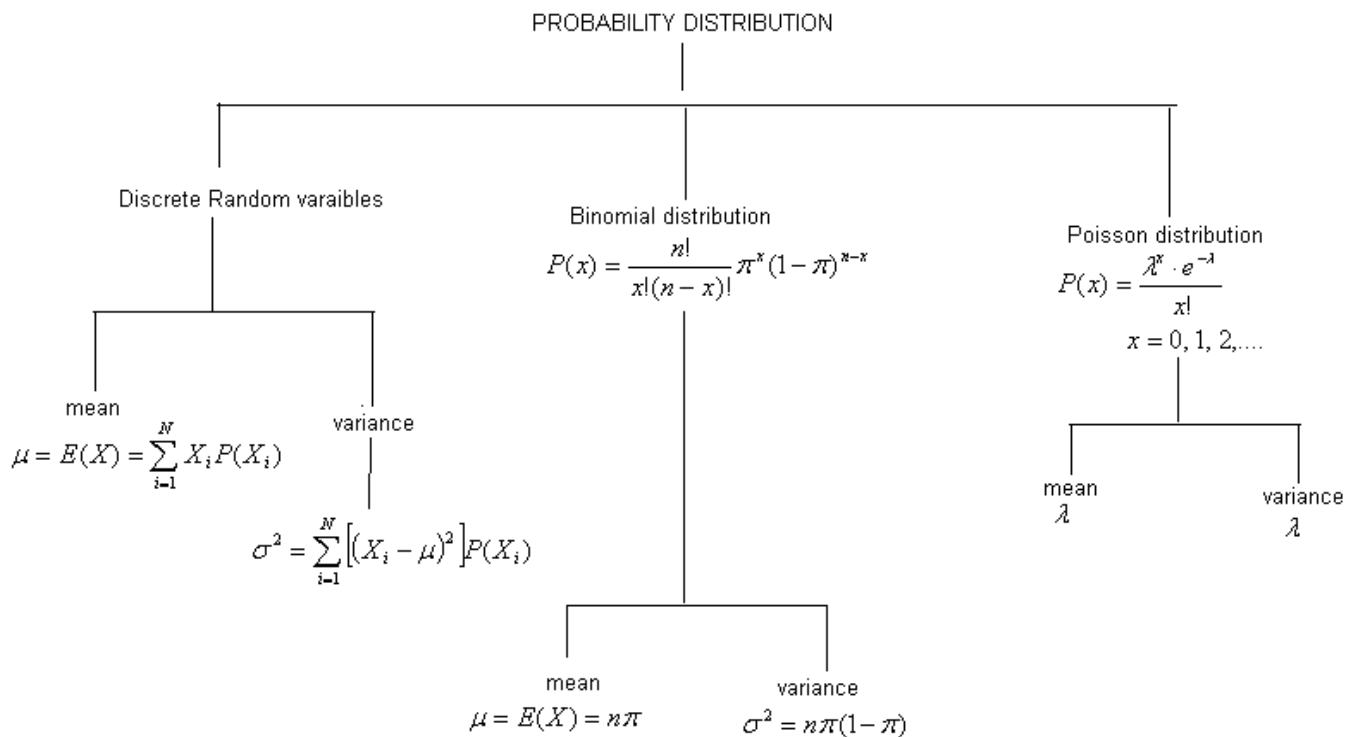
A probability function, denoted by $p(x)$, specifies the probability that a random variable is equal to a specific value. More formally, $p(x)$ is the probability that the random variable X takes on the value x , or $p(x) = P(X = x)$.

The two key properties of a probability function are

- $0 \leq p(x) \leq 1$ for any value of x
- $\sum p(x) = 1$, that is, the sum of the probabilities for all possible outcomes x for a random variable X equals one.

Activity 5.1: Overview Study skill

Draw a mind map of the different sections/headings you will deal with in this study unit. Then page through the unit with the purpose of completing the map.



PROBABILITY DISTRIBUTION

Discrete random variables

mean

variance

Binomial distribution

mean

variance

Poisson distribution

mean

variance

Activity 5.2: Concepts	Conceptual skill	Communication skill
------------------------	------------------	---------------------

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Probability distributions		
Discrete random variables		
Continuous random variables		
Binomial distributions		
Poisson distributions		
The mean of the binomial distribution		
The variance of the binomial distribution		
The mean of the Poisson distribution		
The variance of the Poisson distribution		
The standard deviation		

5.2 PROBABILITY DISTRIBUTION FOR DISCRETE RANDOM VARIABLES

Levine et al define the probability distribution for a discrete random variable as a mutually exclusive listing of all possible numerical outcomes along with the probability of occurrence of each outcome (see section 5.1). That is, if X is a discrete random variable associated with a particular chance experiment, a list of all possible values that X can assume together with their associated probabilities is called a discrete probability distribution. The total probability of all outcomes is 1.

5.2.1 Expected value of a discrete random variable

The mean μ of a discrete probability distribution for a discrete random variable is called its expected value and this is referred to as $E(x)$ or μ . It is calculated as the sum of the product of the random variable X and its corresponding probability, $P(X)$, as follows:

$$\mu = E(X) = \sum_{i=1}^N X_i P(X_i)$$

where

X_i = the i th outcome of the discrete random variable X

$P(X_i)$ = the probability of occurrence of the i th outcome of X

Example 5.1

Based on her experience, a professor knows that the probability distribution for X = number of students who come to her office on Wednesdays is given below.

x	0	1	2	3	4
$P(X=x)$	0.10	0.20	0.50	0.15	0.05

What is the expected number of students who visit her on Wednesdays?

1. 0.50
2. 0.70
3. 1.85
4. 0.90
5. 0.30

Solution:

The expected number (the mean) is calculated as the sum of the product of the random variable X and its corresponding probability, $P(X)$, as follows:

$$\begin{aligned}\mu = E(X) &= \sum_{i=1}^N X_i P(X_i) \\ &= (0 \times 0.1) + (1 \times 0.20) + (2 \times 0.5) + (3 \times 0.15) + (4 \times 0.05) \\ &= 1.85\end{aligned}$$

Option 3

5.2.2 Variance and standard deviation of a discrete random variable

The variance of a probability distribution is calculated by multiplying each possible squared difference $(X_i - \mu)^2$ by its corresponding probability, $P(X_i)$, and then summing the resulting products as follows:

$$\sigma^2 = \sum_{i=1}^N [(X_i - \mu)^2] P(X_i)$$

where

X_i = the i th outcome of the discrete random variable X

$P(X_i)$ = the probability of occurrence of the i th outcome of X

Please note that we have to calculate the mean first before we think of calculating the variance of a discrete random variable.

The standard deviation is the positive square root of the variance of a discrete random variable:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^N (X_i - \mu)^2 P(X_i)}$$

Example 5.2

Let the probability distribution for $X =$ number of jobs held during the past year for students at a college be as follows:

x	1	2	3	4	5
$P(X=x)$	0.25	0.33	0.17	0.15	0.10

The standard deviation of the number of jobs held is

1. 8.000
2. 1.3682
3. 2.5200
4. 1.2844
5. 1.6496

Solution:

We first calculate the mean.

$$\begin{aligned}\mu = E(X) &= \sum_{i=1}^N X_i P(X_i) \\ &= (1 \times 0.25) + (2 \times 0.33) + (3 \times 0.17) + (4 \times 0.15) + (5 \times 0.10) \\ &= 2.52\end{aligned}$$

Then we use the mean to calculate the variance:

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^N [(X_i - \mu)^2] P(X_i) \\ &= (1 - 2.52)^2 \times 0.25 + (2 - 2.52)^2 \times 0.33 + (3 - 2.52)^2 \times 0.17 + (4 - 2.52)^2 \times 0.15 + (5 - 2.52)^2 \times 0.10 \\ &= 1.6496\end{aligned}$$

The standard deviation is $\sigma = \sqrt{\sigma^2} = \sqrt{1.6496} = 1.2844$.

Option 4

If you have not mastered the calculation of the mean, the variance and the standard deviation of a discrete random variable, you can now work through section 5.1 of Levine et al again, otherwise try the following activity before looking at its solutions.

Activity 5.3	Application skills
--------------	--------------------

Question 1

The telephone calls coming into a switchboard and their respective probabilities for a 3-minute interval are as follows:

x	0	1	2	3	4	5
$P(X=x)$	0.60	0.20	0.10	0.04	0.03	0.03

How many calls might be expected over a 3-minute interval?

1. 0.04
2. 3
3. 0.2
4. 0.79
5. 3.75

Question 2

The probability distribution of a discrete random variable is shown below.

x	0	1	2	3
$P(X=x)$	0.25	0.40	0.20	0.15

Find the *incorrect* statement.

1. This is an example of a discrete probability distribution.
2. The expected value of x is 1.25.
3. The variance of x is 2.55.
4. If $x = 0$, the answer of 0 after multiplication by $P(x)$ means that the probability associated with the value $x = 0$ has no influence on the answers of the mean and the variance.
5. The standard deviation of x is 0.9937.

Question 3

Use the data set given in question 2 and find the *incorrect* statement.

1. $P(x > 1) = 0.35$
2. $P(x \leq 2) = 0.65$
3. $p(1 < x \leq 2) = 0.20$
4. $P(0 < x < 1) = 0.00$
5. $P(1 \leq x < 3) = 0.60$

Feedback on the activity

Question 1

Remember that the expected number is also the mean of a discrete random variable, calculated as:

$$\begin{aligned}\mu = E(X) &= \sum_{i=1}^N X_i P(X_i) \\ &= (0 \times 0.60) + (1 \times 0.20) + (2 \times 0.10) + (3 \times 0.04) + (4 \times 0.03) + (5 \times 0.03) = 0.79\end{aligned}$$

Option 4

Question 2

1. *Correct.* The variable takes on discrete values, therefore the statement is correct. Remember, in section 5.2 of this unit we defined the probability distribution for a discrete random variable as a mutually exclusive listing of all possible numerical outcomes along with the probability of occurrence of each outcome, which is exactly the case in this option.
2. *Correct.*

$$\begin{aligned}\mu = E(X) &= \sum_{i=1}^N X_i P(X_i) \\ &= (0 \times 0.25) + (1 \times 0.40) + (2 \times 0.20) + (3 \times 0.15) \\ &= 1.25\end{aligned}$$

3. *Incorrect.* This figure was incorrectly calculated. It should be:

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^N [(X_i - \mu)^2] P(X_i) \\ &= (0 - 1.25)^2 \times 0.25 + (1 - 1.25)^2 \times 0.40 + (2 - 1.25)^2 \times 0.20 + (3 - 1.25)^2 \times \\ &0.15 \\ &= 0.9875\end{aligned}$$

4. *Correct.* You can see it if you study the calculation of the mean and the variance.

5. *Correct.* $\sigma = \sqrt{\sigma^2} = \sqrt{0.9875} = 0.9937$

Question 3

1. *Correct.* We add from 2 (greater than 1) up to 3 as follows:

$$P(x > 1) = P(x = 2) + P(x = 3) = 0.20 + 0.15 = 0.35$$

2. *Incorrect.* Here we take values from 0 to 2. You could also consider this question as at most two as discussed in study unit 4.

$$P(x \leq 2) = P(x = 0) + P(x = 1) + P(x = 2) = 0.25 + 0.40 + 0.20 = 0.85$$

3. *Correct.* In this case 1 is not included, but 2 is.

$$P(1 < x \leq 2) = P(x = 2) = 0.20$$

4. *Correct.* $P(0 < x < 1) = 0.00$ because between 0 and 1 there is no discrete value for x .

5. *Correct.* Here 1 is included but 3 is not.

$$P(1 \leq x < 3) = P(x = 1) + P(x = 2) = 0.40 + 0.20 = 0.60$$

Now that you understand discrete random variables, we can discuss their probability distributions. This is a very small but important section in statistics. There are quite a few discrete probability distributions, though Levine et al emphasise only two, namely the binomial distribution and the Poisson distribution.

5.3 THE BINOMIAL DISTRIBUTION

The binomial distribution describes the probability distribution resulting from the outcome of a binomial experiment. A binomial experiment usually involves several repetitions (trials) of the basic experiment. The binomial probability distribution gives us the probability that a success will occur x times in n trials, for $x = 0, 1, 2, \dots, n$.

Characteristics

- The experiment must consist of n identical trials.
- Each trial has one of two possible mutually exclusive outcomes: success or failure (success refers to the occurrence of the event of interest).

- The probability (π) that the trial will result in a success remains the same from trial to trial.
- The trials are independent of one another (the outcome of a trial does not affect the outcome of any other trial).
- The probability distribution of the number of successes x of the random variable X in n trials of a binomial experiment is:

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

where n = number of trials or sample size

π = probability of success on each trial

x = number of events of interest in the sample

The mathematical sign ! is called the factorial sign of a positive integer n . It is interpreted as the product of all positive integers less than or equal to n . For example, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$, $4! = 4 \times 3 \times 2 \times 1 = 24$ and $0! = 1$.

5.3.1 The mean of the binomial distribution

The mean μ of the binomial distribution is equal to the sample size n multiplied by the probability of an event of interest π .

$$\mu = E(x) = n\pi$$

5.3.2 The variance and the standard deviation of the binomial distribution

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)} = \sqrt{n\pi(1-\pi)}$$

Example 5.3

A textile firm has found from experience that only 20% of the 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 three qualified persons?

$$n = 5, \pi = 0.20, P(x \geq 3) = ?$$

Please do not forget that at least three means add, starting at three.

$$\begin{aligned}
P(x \geq 3) &= P(x = 3) + P(x = 4) + P(x = 5) \\
&= \frac{5!}{3!(5-3)!} 0.20^3(1-0.20)^{5-3} + \frac{5!}{4!(5-4)!} 0.20^4(1-0.20)^{5-4} + \frac{5!}{5!(5-5)!} 0.20^5(1-0.20)^{5-5} \\
&= 0.0512 + 0.0064 + 0.0003 \\
&= 0.0579
\end{aligned}$$

You can now attempt the following typical exam questions. Please try to answer them before looking at the solutions.

Activity 5.4 Application skills

Question 1

A salesman selling new cars knows that he sells a car to one customer out of ten who enters the showroom. The probability that he will sell a car to exactly two of the next three customers is

1. 0.027
2. 0.973
3. 0.000
4. 0.090
5. 0.901

Question 2

Use the information given in question 1. Let X be the number of cars the salesman sells to the next three customers. Which one of the following statements is *incorrect*?

1. X has a binomial distribution.
2. The expected number of cars sold if $n = 3$ is 0.3.
3. The variance of this distribution is 0.27.
4. $P(X \leq 1) = 0.9720$
5. $P(X > 2) = 0.0280$

Question 3

Suppose that 62% of new cars sold in a country are made by one big car manufacturer. A random sample of 7 purchases of new cars is selected. The probability that 4 of these purchases are cars made by this car manufacturer is

1. 0.5800
2. 0.5714
3. 0.2838
4. 0.4200
5. 0.7162

Feedback on the activity

Question 1

$$n = 3, \pi = \frac{1}{10} = 0.1, P(x = 2) = ?$$

$$\begin{aligned} P(x = 2) &= \frac{3!}{2!(3-2)!} 0.1^2(1 - 0.1)^{3-2} \\ &= 0.027 \end{aligned}$$

Question 2

1. *Correct.*
2. *Correct.* $E(x) = n\pi = 3 \times 0.1 = 0.3$
3. *Correct.* $\sigma^2 = n\pi(1 - \pi) = 3 \times 0.1(1 - 0.1) = 0.27$
4. *Correct.* $P(x \leq 1) = P(x = 0) + P(x = 1)$
$$\begin{aligned} &= \frac{3!}{0!(3-0)!} 0.1^0(1 - 0.1)^{3-0} + \frac{3!}{1!(3-1)!} 0.1^1(1 - 0.1)^{3-1} \\ &= 0.7290 + 0.2430 \\ &= 0.9720 \end{aligned}$$
5. *Incorrect.* $P(x > 2) = P(x = 3) = \frac{3!}{3!(3-3)!} 0.1^3(1 - 0.1)^{3-3} = 0.001$

Question 3

$$n = 7, \pi = 0.62, P(x = 4) = ?$$

$$\begin{aligned} P(x = 4) &= \frac{7!}{4!(7-4)!} 0.62^4(1 - 0.62)^{7-4} \\ &= 0.2838 \end{aligned}$$

5.4 THE POISSON DISTRIBUTION

The Poisson distribution is a discrete distribution for an event for which the probability of occurrence over the given span of time, space or distance is extremely small.

There is no specific upper limit to the count (n is unknown), although a finite count is expected. The Poisson distribution tends to describe phenomena such as:

- customers' arrival at a service point during a given period of time, for example the number of motorists approaching a toll booth, the number of hungry people entering a restaurant or the number of calls received by a company call centre. In this context it is also useful in the management science technique called queuing (waiting-line) theory
- defects in manufactured materials, such as the number of flaws in wire or pipe products over a given number of feet, or the number of knots in wooden panels for a given area
- the number of work-related deaths, accidents, divorces, suicides or homicides over a given period of time

Although it is closely related to the binomial distribution, the Poisson distribution has a number of characteristics that makes it unique. These include the following:

- The number of successes that occur in a specified interval is independent of the number of occurrences in any other interval.
- The probability that 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.
- If X is a Poisson random variable, the probability distribution of the number of successes of x is:

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

$$x = 0, 1, 2, \dots$$

where

λ = average number of successes occurring in the given time or measurement

$e = 2.71828$ (the base of natural logarithms)

Example 5.4

The average number of a certain radio sold per day by a firm is approximately Poisson, with a mean of 1.5. The probability that the firm will sell at least two radios over a three-day period is equal to

1. 0.5578
2. 0.1255
3. 0.9389
4. 0.0447
5. 0.4422

Solution:

Recall that this distribution has no upper bound. Therefore we have to express it at least in another equivalent way such as:

$$\begin{aligned} P(x \geq 2) &= 1 - P(x \leq 1) \\ &= 1 - \{P(x = 0) + P(x = 1)\} \\ &= 1 - \left\{ \frac{4.5^0 \cdot e^{-4.5}}{0!} + \frac{4.5^1 \cdot e^{-4.5}}{1!} \right\} \\ &= 1 - \{0.0111 + 0.0500\} \\ &= 0.9389 \end{aligned}$$

Option 3

Example 5.5

A bank receives on average 6 bad cheques per day. The probability that it will receive exactly 4 bad cheques on a given day is

1. 0.0892
2. 0.1393
3. 0.2851
4. 0.1339
5. 0.6667

Solution

Given that $\lambda = 6$, $P(x = 4) = ?$

$$P(x = 4) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} = \frac{6^4 \cdot e^{-6}}{4!} = 0.1339$$

Option 4

5.5 SELF-ASSESSMENT EXERCISE

This section contains questions based on the entire study unit. Please attempt them before referring to the solutions.

Question 1

Bank robbers brandish firearms to threaten their victims in 80% of the incidents. An announcement that six bank robberies are taking place is being broadcast. The probability that a firearm is being used in at least one of the robberies is

1. 0.0015
2. 0.7379
3. 0.0001
4. 0.9999
5. 0.0016

Question 2

In an urban region, health officials anticipate that the number of births this year will be the same as last year, when 438 children were born – an average of $\frac{438}{356}$ or 1.2 births per day. Daily births have been distributed according to a Poisson distribution. The distribution can be represented as:

x	0	1	2	3	4	5	6	7
$P(X = x)$	0.3012	0.3614	0.2169	0.0867	0.0260	0.0062	0.0012	0.0002

What is the probability that at least two births will occur on a given day?

1. 0.3374
2. 0.8795
3. 0.3795
4. 0.7831
5. 0.6626

Question 3

The following probability distribution for an infinite population with the discrete random variable x is given:

x	0	1	2	3
$P(x)$	0.2	0.1	0.3	0.4

Which statement is *incorrect*?

1. The mean of x is 1.9.
2. The probability that x is at most 1 is equal to 0.3.
3. The variance of x is 1.29.
4. The standard deviation of x is 1.14.
5. The probability that x is at least 0 is equal to 0.2.

Question 4

A drug is known to be 80% effective in curing a certain disease. If four people with the disease are to be given the drug, the probability that more than two are cured is

....

1. 0.8464
2. 0.1536
3. 0.5000
4. 0.1808
5. 0.8192

Question 5

Referring to question 4, the expected value of people cured is ...

1. 0.80
2. 0.20
3. 3.20
4. 0.64
5. 1.00

Question 6

Given a Poisson random variable X where the average number of successes occurring in a specified interval is 1.8, $P(X = 0)$ is equal to

1. 0.1653
2. 0.2975
3. 1.0000
4. 0.0000
5. 0.4762

5.6 SOLUTIONS TO THE SELF-ASSESSMENT EXERCISE

Question 1

Solution

$$\begin{aligned}P(x \geq 1) &= 1 - P(x \leq 50) \\&= 1 - \{P(x = 0)\} \\&= 1 - \frac{6!}{0!(6-0)!} 0.80^0 (1 - 0.80)^{6-0} \\&= 1 - 000064 \\&= 0.9999\end{aligned}$$

Option 4

Question 2

$$P(x \geq 2) = 1 - P(x \leq 1) = 1 - \{P(x = 0) + P(x = 1)\} = 1 - \{0.3012 + 0.3614\} = 0.3374$$

Option 1

Question 3

1. *Correct.*

$$\begin{aligned}\mu = E(x) &= \sum_{i=1}^N X_i P(X_i) \\&= (0 \times 0.20) + (1 \times 0.10) + (2 \times 0.30) + (3 \times 0.4) = 1.9\end{aligned}$$

2. *Correct.* $P(x \leq 1) = P(x = 0) + P(x = 1) = 0.2 + 0.1 = 0.3$

3. *Correct.*

$$\begin{aligned}\sigma^2 &= \sum_{i=1}^N [(X_i - \mu)^2] P(X_i) \\ &= (0 - 1.9)^2 \times 0.2 + (1 - 1.9)^2 \times 0.1 + (2 - 1.9)^2 \times 0.3 + (3 - 1.9)^2 \times 0.4 \\ &= 1.29\end{aligned}$$

4. *Correct.* $\sigma = \sqrt{1.29} = 1.14$

5. *Incorrect.* $P(x \geq 0) = P(x = 0) + P(x = 1) + P(x = 2) + P(x = 3) = 1.0$

Question 4

$$\begin{aligned}P(x > 2) &= P(x = 3) + P(x = 4) \\ &= \frac{4!}{3!(4-3)!} 0.80^3 (1-0.8)^{4-3} + \frac{4!}{4!(4-4)!} 0.8^4 (1-0.8)^{4-4} \\ &= 0.4096 + 0.4096 \\ &= 0.8192\end{aligned}$$

Option 5

Question 5

$$E(X) = n\pi = 4 \times 0.80 = 3.2$$

Option 3

Question 6

Given that $\lambda = 1.8$, $P(x = 0) = ?$

$$P(x = 0) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} = \frac{1.8^0 \cdot e^{-1.8}}{0!} = 0.1653$$

Option 1

5.7 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to

- recognise and define a discrete probability distribution
- construct a probability distribution for a discrete random variable
- understand the concept of a Bernoulli process and its application in consecutive trials, as associated with the binomial distribution
- differentiate between the binomial and Poisson distributions
- determine the probability that a binomial variable will assume a given value and a Poisson variable a value within a given range

STUDY UNIT 6

THE NORMAL DISTRIBUTION

Key questions for this unit

How would you calculate probabilities from the normal distribution?

Can you distinguish between discrete and continuous probability distributions?

Can you use the normal table to calculate probabilities?

Can you determine the Z-variable given the area under the normal curve?

6.1 INTRODUCTION

The normal distribution is the most important distribution in statistics and the key reasons for this include:

- Numerous continuous variables common in business have distributions that closely resemble the normal distribution.
- The normal distribution can be used to approximate various discrete probability distributions.
- The normal distribution provides the basis for classical statistical inference because of its relationship to the *Central Limit Theorem* (see sections 6.1 and 6.2 of Levine et al).

You must make sure that you know the characteristics of the normal distribution and how to use the normal table (E.2) to determine probabilities.

Activity 6.1: Overview Study skills

Draw a mind map of the different sections/headings you will deal with in this study unit. Then page through the unit with the purpose of completing the map.

CONTINUOUS PROBABILITY DISTRIBUTION

Normal Density Function

Where

mean

standard deviation

where

parameter of the distribution

Activity 6.2: Concepts	Conceptual skill	Communication skill
------------------------	------------------	---------------------

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Continuous probability distributions		
Discrete probability distributions		
The standard normal probabilities		
The mean of the normal distribution		
The standard deviation of the normal distribution		
The area under the normal curve		

6.2 NORMAL AND STANDARDISED NORMAL DISTRIBUTIONS

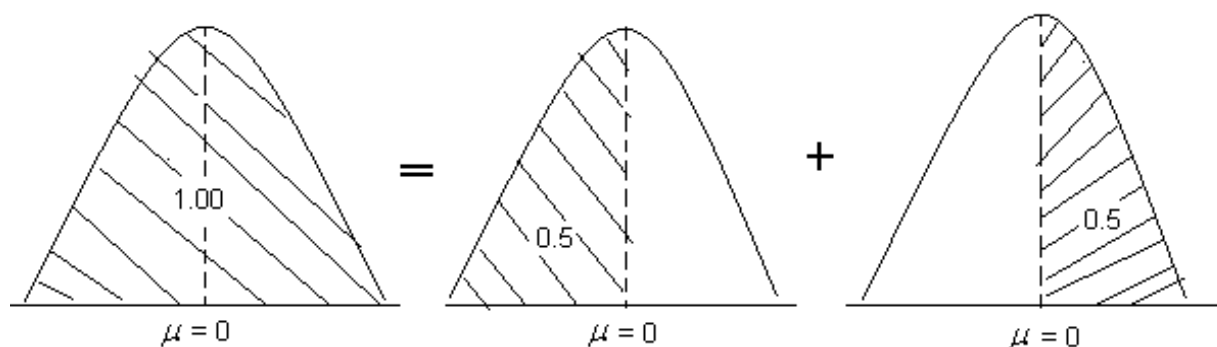
The most important fact to understand is that although a normal distribution is also a probability distribution function, it has its own characteristics, with the most obvious one being the fact that the variable it describes is continuous. It changes the concept of probability to an area instead of the height of a bar. Many students have problems to understand what we mean by the area beneath a curve, especially because we say that this area is determined through the mathematics of calculus. So, if you are

one of those, do not worry or imagine it as so difficult. You have been hearing about area since primary school and it was always the product of length and breadth ($l \times b$). Now area is still the product of two values, even though it is not such a perfect “box”. The nice part is that it is not your problem how the area of such a funny “box” is determined – you simply read off the answer from a table (see E.2).

Read through sections 6.1 and 6.2 of the textbook at least once!

In addition, there are a few things you have to know about the normal distribution.

- The form of the distribution is described as bell-shaped, meaning that it is symmetric (if it were possible to cut out the line forming the bell, you would be able to fold it double with the two halves fitting on top of each other).
- The normal curve is indicated within a system of axes – two perpendicular lines (like those you may have used in Grade 9 to present a straight-line graph).
- The values of the continuous variable X (in the notation of Levine et al) are indicated on the horizontal axis.
- There is a difference between the *probability distribution* and the probabilities.
 - The *distribution* is only the line forming the bell and is called the density function, indicated as $f(X)$ (a function of the variable X).
 - The *total probability* is represented by the area between that density function (bell-shaped line) and the x -axis.
- The total area equals one, but can be divided into sections, determined by the values given to the variable X on the horizontal axis as shown in the following graphs:



- The centre X -value, where you would fold the curve to indicate the symmetry, represents the mean μ for that particular distribution (see graphs above).

- It is not only the placement of the mean μ that determines the distribution of a particular normal distribution, the standard deviation σ (how the values are spread around the mean) also determines the form of the distribution.
- If the values of μ and σ are used to standardise each individual X -value using the formula $\frac{X - \mu}{\sigma}$ (see equation 6.2), then the original normal distribution will be transformed into a so-called *standard normal distribution* whose mean = 0 and standard deviation = 1.
- The values in the normal table give the areas for the probabilities of the standard normal distribution, that is, the one whose mean = 0 and standard deviation = 1.
- The previous statement implies that all general normal distributions must first be standardised using the formula $\frac{X - \mu}{\sigma}$ before the normal table may be used.
- In this study guide we also insert another version of the normal table (taken from Utts and Heckard: *Mind on statistics*, 3rd edition) for your convenience. Here two separate tables are given – one for negative Z -values and one for positive Z -values. Some students find it easier to calculate areas under the normal curve, using these two tables.

Standard Normal Probabilities (for $z < 0$)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0266	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1036	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1768	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3697	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4860	.4640	.4601	.4761	.4721	.4681	.4641

In the Extreme (for $z < 0$)

z	-3.09	-3.72	-4.26	-4.75	-5.20	-5.61	-6.00
Probability	.001	.0001	.00001	.000001	.0000001	.00000001	.000000001

S-PLUS was used to determine information for the "in the extreme" portion of the table.

Standard Normal Probabilities (for $z > 0$)



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8119	.8186	.8212	.8238	.8264	.8239	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
1.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9986	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

In the Extreme (for $z > 0$)

z	3.09	3.72	4.26	4.75	5.20	5.61	6.00
Probability	.999	.9999	.99999	.999999	.9999999	.99999999	.999999999

S-PLUS was used to determine information for the “in the extreme” portion of the table.

The normal table can be used either to determine an area under the normal curve for a given Z -value or *backwards* whenever the value of the Z -variable has to be determined for a given area.

You should now go through sections 6.1 and 6.2 again. There is nothing wrong with you if you have to go through a chapter a number of times. It may even be that you need to break up a chapter into small sections and repeat them over and over until you understand what we are trying to teach you.

Remember that statistics is about understanding and then building a mental structure based on the underlying theory. I think you are now ready to do a few activities. Always study the feedback carefully, as I do a lot of explaining there (for those of you who could not manage the activity yourself).

Activity 6.3 Study skill

Question 1

Assume X is normally distributed with mean $\mu = 15$ and standard deviation $\sigma = 3$. Use the approximate areas under the normal curve to evaluate the following statements.

The *incorrect* statement is

1. $P(X \geq 15) = P(X \leq 15) = 0.5$
2. $P(12 \leq X \leq 18) = 0.955$
3. $P(X \leq 9) = 0.0228$
4. $P(X = 20) = 0$
5. $P(X \geq 12) = 0.8413$

Question 2

A retailer finds that the demand for a very popular board game averages 100 per week with a standard deviation of 20. If the seller wants to have adequate stock 95% of the time, how many of the games must she keep on hand?

Question 3

Identify the *incorrect* statement.

1. The average waiting time at the checkout counter for a large grocery chain is 2.45 minutes with a standard deviation of 24 seconds (0.40 minutes). If we assume that the distribution of waiting times is normal, the probability that a customer will have to wait more than 3 minutes for checkout is 0.9162.
2. Considering the information in option 1, the proportion of the customers who are served between 1 minute and 2.5 minutes is 0.5518.
3. Suppose the monthly demand for car tyres at a tyre dealer is normally distributed with a mean of 250 tyres and a standard deviation of 50 tyres. The number of tyres the store must have in stock at the beginning of each month in order to meet the demand 95% of the time is 332.25.
4. A circus performer who gets shot from a cannon is supposed to land in a safety net. The distance he travels is normally distributed with a mean of 55 metres and a standard deviation of 4.7 metres. His landing net is 16 metres long and the mid-point of the net is positioned 55 metres from the cannon. The probability that the performer will miss the net on a given night is 0.0892.
5. The probability that the circus performer in option 4 will hit the net is equal to 0.9108.

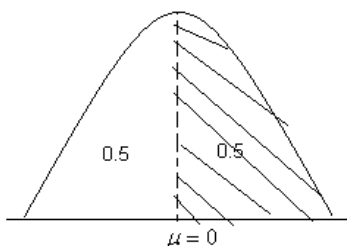
Feedback on the activity

Question 1

1. *Correct.* We begin by standardising as in equation 6.2 of Levine et al.

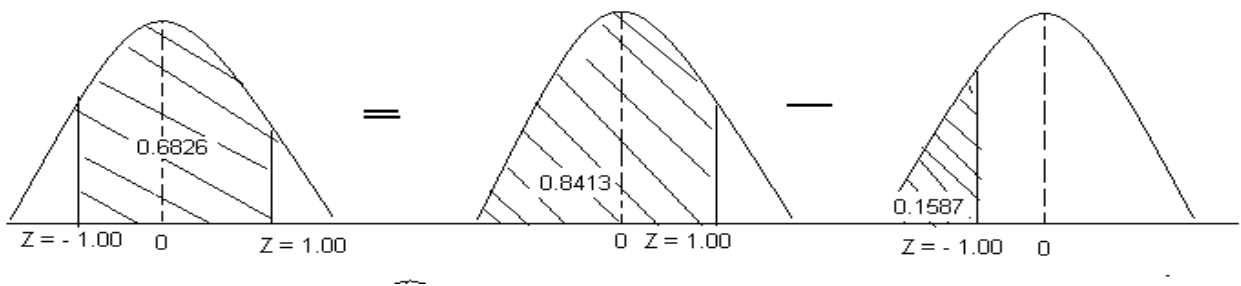
$$P(X \geq 15) = P(X \leq 15) = P\left(\frac{X - \mu}{\sigma} < \frac{15 - 15}{3}\right) = P(Z \leq \frac{15 - 15}{3}) = P(Z \leq 0) = 0.5$$

The mean lies at the centre of the distribution and therefore divides the total area of 1 into half (each half represents 0.5 of the total area) as shown in the figure below.

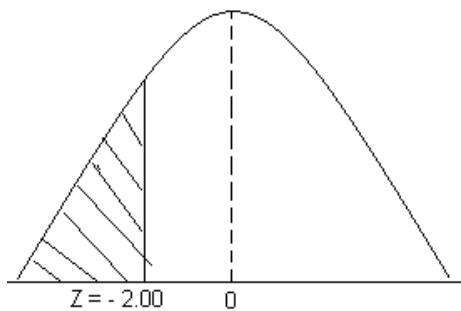


2. *Incorrect.* Standardise the random variables before reading off the respective probabilities from the graphs. You can use the tables in the study guide or table E.10 in Levine et al.

$$\begin{aligned}
 P(12 \leq X \leq 18) &= P\left(\frac{12-15}{3} \leq \frac{X-\mu}{\sigma} \leq \frac{18-15}{3}\right) \\
 &= P(-1.00 \leq Z \leq 1.00) \\
 &= 0.8413 - 0.1587 \\
 &= 0.6826
 \end{aligned}$$



3. *Correct.* $P(X \leq 9) = P\left(\frac{X-\mu}{\sigma} \leq \frac{9-15}{3}\right) = P(Z \leq -2.00) = 0.0228$

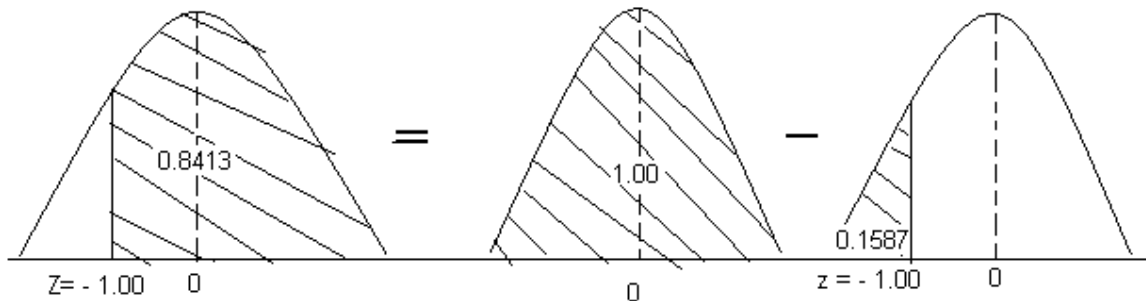


4. *Correct.* $P(X = 20) = 0$

If the variable is continuous, we assume that its probability for any fixed value is always zero! Remember, the continuous variable lies somewhere within a small interval, but we cannot give a fixed value to it.

5. *Correct.*

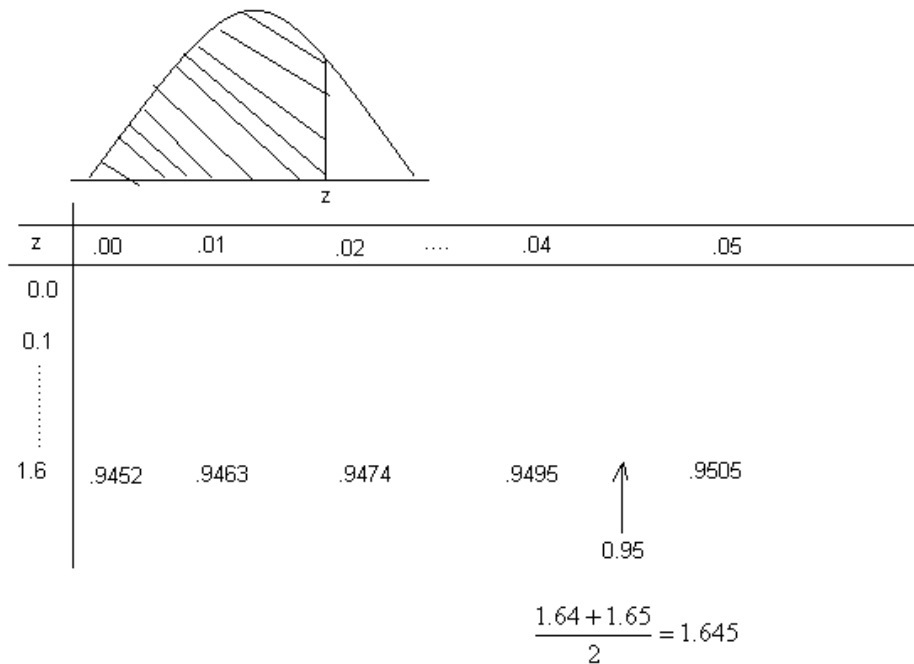
$$P(X \geq 12) = P\left(\frac{X - \mu}{\sigma} \geq \frac{12 - 15}{3}\right) = P(Z \geq -1.00) = 1.00 - 0.1587 = 0.8413$$



Question 2

Option 1

This question is about working backwards. First standardise the random variable, then look for the Z-value corresponding to the area under 0.95. Please note that this should be read off from the body of the table as shown below.



$$P(X < w) = 0.95$$

$$P\left(\frac{X - \mu}{\sigma} < \frac{w - 100}{20}\right) = 0.95$$

$$P\left(Z < \frac{w - 100}{20}\right) = 0.95$$

$$\frac{w - 100}{20} = 1.645$$

$$w = 100 + (20 \times 1.645) = 132.9$$

Question 3

1. *Incorrect.* The correct answer is 0.0838.

Note that you always have to use the same units – in this case, use only *minutes*.

$$P(X > 3) = P\left(\frac{X - \mu}{\sigma} > \frac{2.45}{0.40}\right) = P(Z > 1.38) = 1.00 - 0.9162 = 0.0838$$

From the normal table, the value corresponding to 1.38 (rounded to two decimals) is 0.9162. This is the given answer, but not the correct one.

Remember how the normal table is tabulated? The areas are tabulated *cumulatively* from the mean up to the listed value, but the question specifies the area greater than 1.38 (to the right of 1.38). For the correct answer you therefore have to subtract 0.9162 from 1.00.

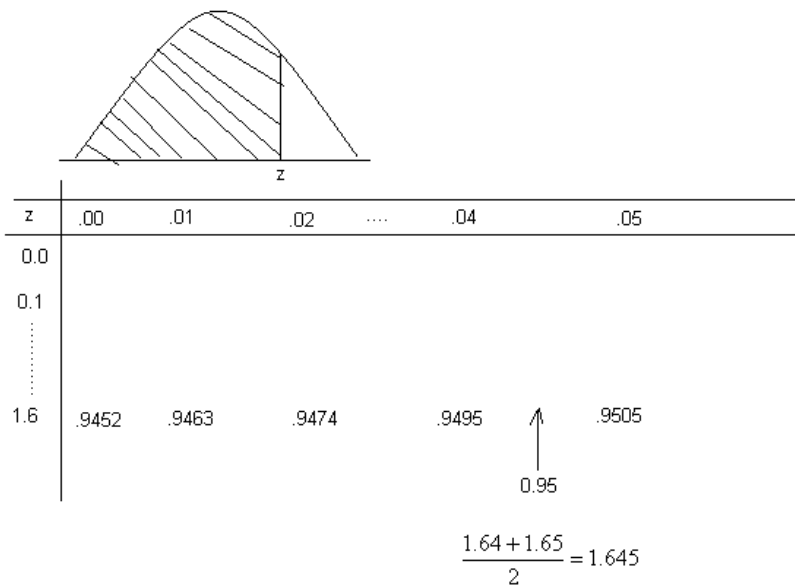
2. *Correct.* Using the normal table:

$$\begin{aligned} P(1 < X < 2.5) &= P\left(\frac{1 - 2.45}{0.40} < \frac{X - \mu}{\sigma} < \frac{2.5 - 2.45}{0.40}\right) = P(-3.63 < Z < 0.13) \\ &= 0.00014 + 0.5517 = 0.5518 \end{aligned}$$

3. *Correct.* This question is about working backwards.

$$P(X < w) = P\left(\frac{X - \mu}{\sigma} < \frac{w - 250}{50}\right) = P\left(Z < \frac{w - 250}{50}\right) = 0.95$$

To meet the demand 95% of the time implies that we are looking for a z-value such that 0.95 of the area lies to the left of it. We use the normal table (see E.2) to look for the value of 0.95 inside the normal table, because this is an area. The z-value which corresponds to an area of 0.95 is 1.645. This value is the z-value, but we have to use it to find the w-value.



$$\frac{w - 250}{50} = 1.645$$

$$w = 250 + (50 \times 1.645) = 332.25$$

4. *Correct.* According to the information, the 16 m net is placed in such a way that it begins at $(55 - 8 = 47)$ metres and stretches up to $(55 + 8 = 63)$ metres from the cannon. The performer will miss the net by falling short or falling past the net. In terms of the normally distributed variable, this comment means that $P(X \leq 47)$ or $P(X \geq 63)$.

$$\text{Standardise: } P\left(Z \leq \frac{47 - 55}{4.7}\right) \text{ or } P\left(Z \geq \frac{63 - 55}{4.7}\right)$$

$$P(Z \leq -1.70) \quad \text{or} \quad P(Z \geq 1.70)$$

The table value for $P(Z \leq -1.70)$ is 0.0446, which means that $P(Z \geq 1.70) = 1.00 - 0.9554 = 0.0446$. Combining the total probabilities is twice 0.0446, which is 0.0892.

5. *Correct.* I hope that you realise that it is not necessary to repeat the calculation. The person can only hit the net or miss the net. This means that the sum of the probabilities must equal one. The probability for a hit of the net is therefore $1 - 0.0892 = 0.9108$.

6.3 SELF-ASSESSMENT EXERCISE

Question 1

Which one of the following is not a characteristic of a normal distribution?

1. The normal variable can take on only discrete values.
2. It is a symmetrical distribution.
3. The mean, median and mode are all equal.
4. It is a bell-shaped distribution.
5. The area under the curve is equal to 1.

Question 2

Given that Z is a standard normal random variable, a negative value of z indicates that the ...

1. value Z is to the left of the mean.
2. value Z is to the right of the median.
3. standard deviation of Z is negative.
4. area between zero and Z is negative.
5. area to the right of Z is equal to 1.

Question 3

If Z is a normal variable with $\mu = 0$ and $\sigma = 1$, the area to the left of $Z = 1.6$ is

1. 0.4452
2. 0.9452
3. 0.0548
4. 0.5548
5. 0.5000

Question 4

Use the normal table to find the Z -value Z_1 if the area to the right of Z_1 is 0.8413. The value of Z_1 is

1. 1.36
2. -1.36
3. 0.00
4. -1.00
5. 1.00

Question 5

Let Z be a Z -score that is unknown but identifiable by position and area. If the area to the left of Z is 0.9306, then the value of Z must be

1. 1.48
2. 0.9603
3. -1.48
4. 0.4306
5. -0.0694

Question 6

Which of the following statements is incorrect?

1. $P(Z \geq 1.63) = 0.0516$
2. $P(Z \geq 0.5) = 0.3085$
3. $P(Z < -1.63) = -0.0516$
4. $P(Z > 1.28)$
5. $P(-1 \leq Z \leq 1) = 0.6826$

Question 7

If the mean is 20 minutes and the standard deviation is 5 minutes, then the area between 22 and 25 minutes for a normal curve is

1. 0.1554
2. 0.3413
3. 0.4967
4. 0.1859
5. 0.0185

Question 8

A bakery finds that the average weight of its most popular package of biscuits is 200.5 g with a standard deviation of 10.5 g. What proportion of packages will weigh less than 180 g?

1. 0.4744
2. 0.0256
3. 0.5226
4. 0.4713
5. 0.9744

Question 9

The average labour time to sew a pair of jeans is 4.2 hours with a standard deviation of 0.5 hours. If the distribution is normal, then the probability of a worker finishing a pair of jeans in more than 3.5 hours is

1. 0.0808
2. 0.4192
3. 0.5808
4. 0.9192
5. 0.9808

Question 10

A retailer finds that the demand for a popular board game averages 50 per week with a standard deviation of 20. If the seller wants to have adequate stock 99% of the time, how many games must she keep on hand?

1. 81.0
2. 89.2
3. 50.0
4. 70.0
5. 96.6

6.4 SOLUTIONS TO THE SELF-ASSESSMENT EXERCISE

Question 1

The normal variable can take on only continuous values.

Option 1

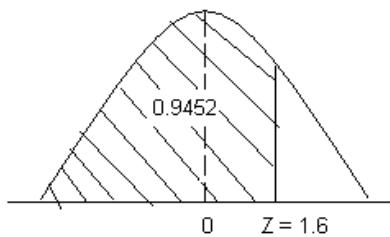
Question 2

The value Z is to the left of the mean.

Option 1

Question 3

$$P(Z < 1.6) = 0.9452$$

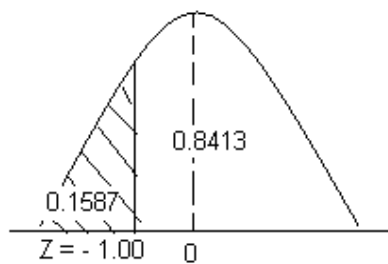


Option 2

Question 4

$$P(Z > Z_1) = 0.8413$$

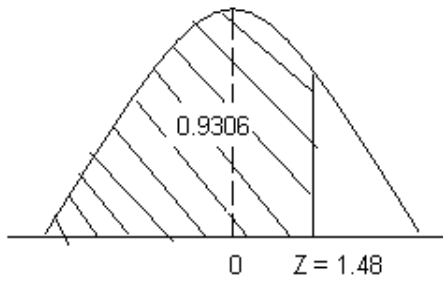
$$P(Z < Z_1) = 0.1587 \Rightarrow Z_1 = -1.00$$



Option 4

Question 5

$$P(Z < z) = 0.9306 \Rightarrow z = 1.48$$

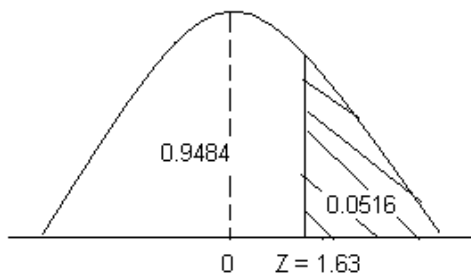


Option 1

Question 6

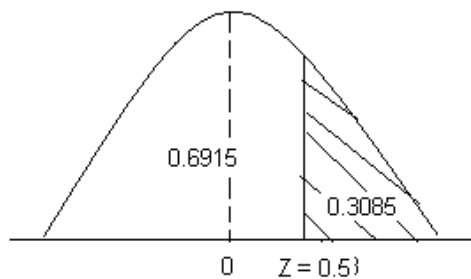
Option 1: *Correct.*

$$P(Z \geq 1.63) = 0.0516$$



Option 2: *Correct.*

$$P(Z \geq 0.5) = 0.3085$$

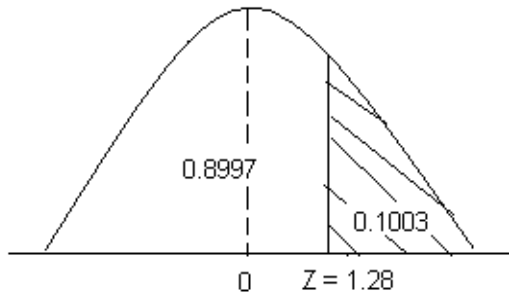


Option 3. *Incorrect!* Remember, the area under the graph cannot be negative.

$$P(Z < -1.63) = 0.0516$$

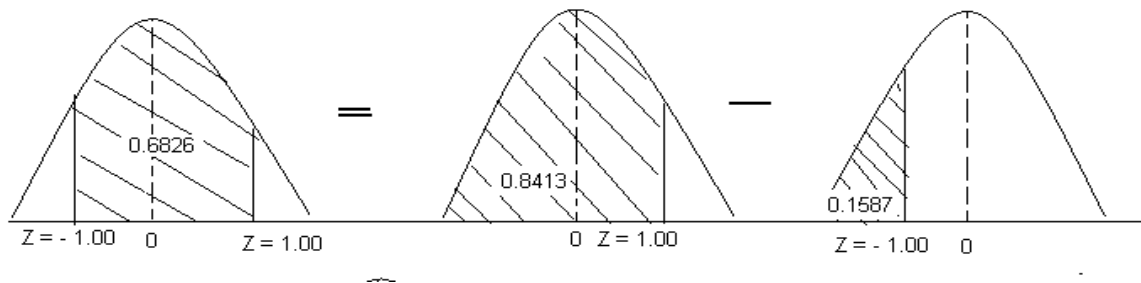
Option 4. *Correct.*

$$P(Z > 1.28) = 0.1003$$



Option 5. *Correct.*

$$P(-1 < Z < 1) = 0.6826$$



Question 7

$$\begin{aligned} P(22 \leq X \leq 25) &= P\left(\frac{22-20}{5} \leq Z \leq \frac{25-20}{5}\right) \\ &= P(0.4 \leq Z \leq 1) = 0.8413 - 0.6554 = 0.1859 \end{aligned}$$

Option 4

Question 8

$$\begin{aligned} P(X < 180) &= P\left(Z < \frac{180-200.5}{10.5}\right) \\ &= P(Z < -1.95) = 0.0256 \end{aligned}$$

Option 2

Question 9

$$\begin{aligned} P(X > 3.5) &= P\left(Z > \frac{3.5-4.2}{0.5}\right) \\ &= P(Z > -1.4) = P(Z < 1.4) = 0.9192 \end{aligned}$$

Option 4

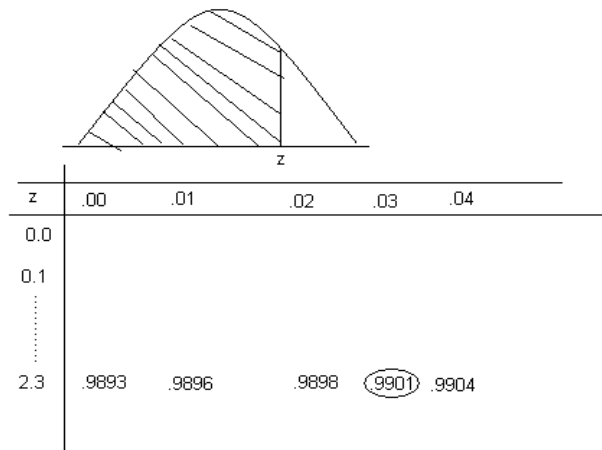
Question 10

This question is about working backwards.

$$P\left(Z \leq \frac{a-50}{20}\right) = 0.99$$

$$\frac{a-50}{20} = 2.33$$

$$a = 50 + (20 \times 2.33) = 96.6$$



Option 5

6.5 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to

- understand the idea that probability is given in terms of an area if the variable is continuous
- identify the normal distribution as a continuous distribution and use it appropriately
- recognise the characteristics of the normal distribution based on its symmetry
- appreciate the link between general normal distributions and the standardised normal distribution
- use the normal table to find specific areas within given limits
- exploit the backwards use of the normal table, that is, determine the z-value given an area.

STUDY UNIT 7

SAMPLING DISTRIBUTION

Key questions for this unit

What is meant by the concepts “estimate”, “inference”, “sample mean”, “population mean”, “statistic”, “parameter”, “sample proportion” and “population proportion”?

Define a sampling distribution.

Distinguish between a sampling distribution of the mean and a sampling distribution of proportion.

What is the purpose when making a statistical inference?

What is the benefit of the Central Limit Theorem?

7.1 INTRODUCTION

A sampling distribution is classified as either a sampling distribution of the mean or a sampling distribution of proportion, depending on the possible samples selected or the proportion of items in a population having a certain characteristic of interest.

In study unit 3 we defined sample mean (\bar{X}), population mean (μ), statistic and parameter. In this and the next study unit we will use these concepts, but in a different manner, including sample proportion and population proportion.

- Inference means that we are making an assumption or a deduction about the population based on the sample data where data are gathered by drawing a sample from the population.
- A sample must be representative of the population.
- A sampling distribution is the distribution of the results if you actually selected all possible samples.

Activity 7.1: Overview Study skill

Draw a mind map of the different sections/headings you will deal with in this study unit. Then page through the unit with the purpose of completing the map.

Sampling distribution

Sampling distribution of the mean

Sampling distribution of proportion

In many situations the population is so large that you cannot gather information on every item. Instead, statistical sampling procedures focus on collecting a small representative group of the larger population. Analysing the results obtained from the sample is less time-consuming, less costly and more practical than an analysis of the entire population.

Activity 7.2: Concepts Conceptual skill Communication skill

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Sample mean		
Population mean		
Inference		
Statistic		
Parameter		
Sample proportion		
Population proportion		
Sampling distribution		
Unbiased		
Standard error of the mean		
Standard error of proportion		

7.2 SAMPLING DISTRIBUTION OF THE MEAN

Definition: The sampling distribution of the mean is the distribution of the means of all possible samples of a given size.

The sample mean (\bar{X}) is unbiased because the mean of all the possible sample means is equal to the population mean (μ). Alternatively, the sample mean (\bar{X}) is unbiased because the expected value of the sample mean (\bar{X}) is equal to the population parameter:

$$E(\bar{X}) = \mu$$

STANDARD ERROR OF THE MEAN

The standard error of the mean ($\sigma_{\bar{X}}$) is equal to the standard deviation (σ) in the population of all possible sample means.

Steps

1. Calculate the population standard deviation (σ), if it is not calculated.
2. Find the sample size n .

The standard deviation of the mean can be used to do this.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Activity 7.3

Question 1

Suppose a random sample of $n = 25$ observations is selected from a population that is normally distributed with the mean equal to 106 and the standard deviation equal to 12. Determine the mean and the standard deviation of the sampling distribution of the sample mean \bar{X} .

Question 2

A random sample of n observations is selected from a population with a standard deviation $\sigma = 2$. Calculate the standard error of the mean for these values of n :

- a. $n = 5$
- b. $n = 49$

Question 3

Population A consists of all values of the invoices of a certain company. The mean of population A is R350 and the standard deviation is R100. Population B consists of all samples of 16 values drawn from population A. The mean of population B is

1. R100
2. R250
3. R350
4. R450
5. R25

Feedback on the activity

Question 1

Steps

1. Population standard deviation $\sigma = 12$.
2. The sample size $n = 25$.

\therefore The standard error of the mean is equal to $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{25}} = \frac{12}{5} = 2.4$.

The population mean $\mu = 106$.

Question 2

- a. When $n = 5$:

Standard deviation $\sigma = 2$

Standard error of the mean $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{5}} = \frac{2}{2.2361} = 0.8944$

- b. When $n = 49$:

Standard deviation $\sigma = 2$

Standard error of the mean $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{49}} = \frac{2}{7} = 0.2857$

Question 3

The mean of population A is equal to the mean of population B.

Option 3

What distribution will the sample mean \bar{X} follow?

In the previous study unit, we saw that a random variable X is normally distributed with the mean μ and the standard deviation σ . If we are now sampling from a population that is normally distributed with the mean μ and the standard deviation σ , then, regardless of the sample size n , the sampling distribution of the mean is normally distributed with the mean $\mu_{\bar{X}} = \mu$ and the standard error of the mean

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}.$$

How do you calculate the probability for the sampling distribution of the mean?

Steps

1. Determine the population mean μ and the sample mean \bar{X} .
2. Determine the sample size n .
3. Determine the number of the sample mean (\bar{X}) for which we want to find the probability.
4. Find the value of Z , called the “test statistic”:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \text{or} \quad Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

where $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean ($\sigma_{\bar{X}}$)

Activity 7.4

Question 1

A normal distribution with the population mean $\mu = 100$ and the standard deviation $\sigma = 12$ is given. You select a sample of $n = 36$.

1. What is the probability that the sample mean \bar{X} is less than 95?
2. What is the probability that the sample mean \bar{X} is between 95 and 97.2?
3. What is the probability that the sample mean \bar{X} is above 102.2?
4. There is a 65% chance that the sample mean \bar{X} is above which value?

Question 2

Given an infinite population with a mean of 75 and a standard deviation of 12, the probability that the mean of a sample of 36 observations, taken at random from this population, will exceed 78 is

1. 0.4332
2. 0.0668
3. 0.0987
4. 0.9013
5. 0.9332

Feedback on the activity

Question 1

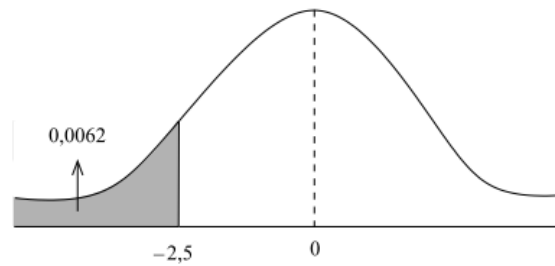
1. $P(\bar{X} < 95) = ?$

Steps

1. Use the transformation formula called the “test statistic” $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

2. Substitute the values into this formula: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{95 - 100}{\frac{12}{\sqrt{36}}} = \frac{-5}{\frac{12}{6}} = \frac{-5}{2} = -2.5$

3. Determine the equivalent number of the sample mean for which we want to find the probability $P(\bar{X} < 95) = P(Z < -2.5)$. Now determine the area that is less than -2.5 .



4. Find the value using the cumulative standard normal distribution table E.2 (from the appendix).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-6.0										
-5.5										
.....										
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	
.....										

$$P(Z < -2.5) = 0.0062$$

2. $P(95 < \bar{X} < 97.2) = ?$

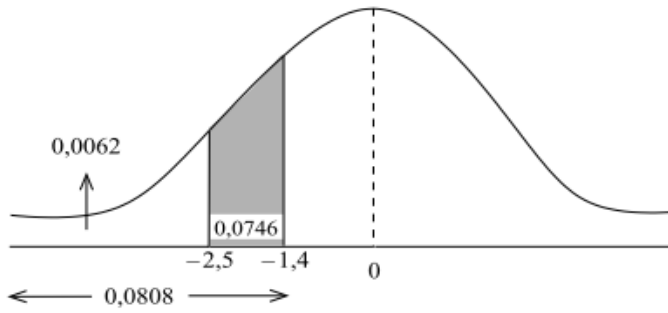
Steps

$$1. Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \quad \mu = 100 \quad n = 36 \quad \sigma = 12$$

$$2. \text{ If } \bar{X} = 95, \text{ then } Z = \frac{95 - 100}{\frac{12}{\sqrt{36}}} = \frac{-5}{\frac{12}{6}} = \frac{-5}{2} = -2.5$$

$$\text{ If } \bar{X} = 97.2, \text{ then } Z = \frac{97.2 - 100}{\frac{12}{\sqrt{36}}} = \frac{-2.8}{\frac{12}{6}} = \frac{-2.8}{2} = -1.4$$

3. $P(95 < \bar{X} < 97.2) = P(-2.5 < Z < -1.4)$ and we now determine the area that is between -2.5 and -1.4 .



Since our statistics tables are all going to the left:

$$P(-2.5 < Z < -1.4) = P(Z < -1.4) - P(Z < -2.5) = 0.0808 - 0.0062 = 0.0746$$

3. $P(\text{above } 102.2) = P(\bar{X} > 102.2) = ?$ $\mu = 100$ $n = 36$ $\sigma = 12$

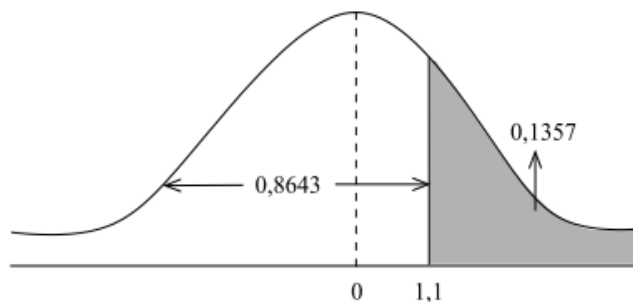
Steps

1. Use the transformation formula called the “test statistic” $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

2. Substitute the values into this formula: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{102.2 - 100}{\frac{12}{\sqrt{36}}} = \frac{2.2}{\frac{12}{6}} = \frac{2.2}{2} =$

1.1

3. Determine the equivalent number of the sample mean for which we want to find the probability $P(\bar{X} > 102.2) = P(Z > 1.1)$ by determining the area that is greater than 1.1.



4. Find the value using the cumulative standard normal distribution table E.2 (from the appendix).

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0										
0.1										
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
.....										

$$P(Z > 1.1) = 1 - P(Z < 1.1) = 1 - 0.8643 = 0.1357$$

4. $P(\bar{X} > a) = 0.65$

$$P\left(Z > \frac{a - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0.65 \quad \text{Now substitute } \mu = 100, n = 36, \sigma = 12.$$

$$P\left(Z > \frac{a - 100}{\frac{12}{\sqrt{36}}}\right) = 0.65 \quad \text{Find the Z-value corresponding to 0.65 from the}$$

cumulative standard normal table by looking inside the table.

$$P\left(Z > \frac{a - 100}{\frac{12}{\sqrt{36}}}\right) = 0.65$$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0										
0.1										
.....										
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
.....										

$$\frac{a - 100}{\frac{12}{\sqrt{36}}} = -0.38$$

$$\frac{a - 100}{2} = -0.38 \quad a = 2 \times -0.38 + 100 = 99.24$$

Question 2

$$P(\bar{X} > 78) = ?$$

Steps

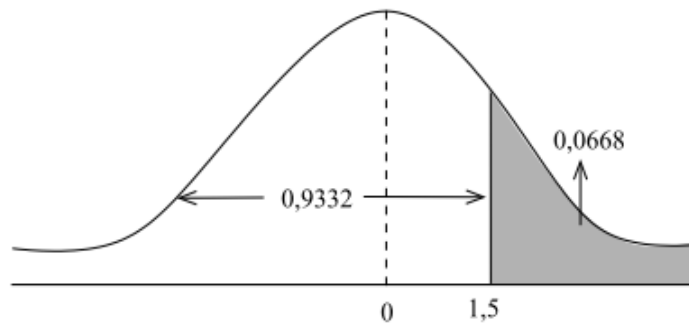
1. Use the transformation formula called the "test statistic" $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

$$\mu = 75 \quad \sigma = 12 \quad n = 36 \quad \bar{X} = 78$$

2. Substitute the values into this formula: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{78 - 75}{\frac{12}{\sqrt{36}}} = \frac{3}{\frac{12}{6}} = \frac{3}{2} = 1.5$

3. Determine the equivalent number of the sample mean for which we want to find the probability.

$P(\bar{X} > 78) = P(Z > 1.5)$, so determine the area that is greater than 1.5.



4. Find the value using the cumulative standard normal distribution table E.2.

$$P(Z > 1.5) = 1 - 0.9332 = 0.0668$$

The Central Limit Theorem is important when using statistical inference to draw conclusions about a population without knowledge of the specific shape of the population distribution. This theorem states that the sum of a large number of independent observations form the same distribution, under certain general conditions an approximate normal distribution.

7.3 SAMPLING DISTRIBUTION OF PROPORTION

The sample proportion is represented by p and the population proportion is represented by π . Therefore p is a statistic and π is a parameter.

The statistic p is used to estimate the population proportion parameter π . This sample proportion is an unbiased estimator of the population proportion, because the expected value of the sample statistic is equal to the relevant population parameter, that is, $E(p) = \pi$.

The sample proportion formula:

$$p = \frac{X}{n} = \frac{\text{number of items having the characteristics of interest}}{\text{sample size}}$$

THE STANDARD ERROR OF PROPORTION

The standard error of proportion σ_p is given by $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$.

In many instances, you can use the normal distribution to estimate the sampling distribution of proportion. When the parameter π is unknown, then the standard error

of proportion is given by $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$.

When do you assume that the sampling distribution of proportion is approximately normally distributed?

It is normally distributed when $n \times \pi$ and $n \times (1 - \pi)$ are each at least 5.

How do you calculate the probability for the sampling distribution of proportion?

Steps

1. Determine the population proportion π and the sample proportion p .
2. Determine the sample size n .
3. Determine the number of the sample proportion (p) for which we want to determine the probability.
4. Find the value of Z called the “test statistic”:

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \quad \text{or} \quad Z = \frac{p - \pi}{\sigma_p}$$

where

the standard error of proportion (σ_p) = $\sqrt{\frac{\pi(1-\pi)}{n}}$ when π is known

the standard error of proportion (σ_p) = $\sqrt{\frac{p(1-p)}{n}}$ when π is unknown

Activity 7.5

Question 1

In a random sample of 64 people, 48 are classified as “successful”.

- Determine the sample proportion p of “successful”.
- If the population proportion is 0.80, determine the standard error of proportion.

Question 2

Suppose that we randomly select a sample of $n = 100$ units from a population and that we calculate the sample proportion p of these units that fall into a category of interest.

If the true population proportion π equals 0.9:

- Find the mean and the standard deviation of the sampling distribution of p .
- Calculate the following probabilities about the sample proportion p .
 - $P(p \geq 0.96)$
 - $P(0.855 \leq p \leq 0.945)$
 - $P(p \geq 0.915)$

Feedback on the activity

Question 1

- The sample proportion $p = \frac{X}{n} =$
$$\frac{\text{number of items having the characteristics of interest}}{\text{sample size}}$$

$$= \frac{48}{64} = 0.75$$

- Population proportion $\pi = 0.80$

$$\therefore \text{Standard error of proportion } \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.80(1-0.80)}{64}} = 0.05$$

Question 2

- The population of all possible sample proportions has mean $\pi = 0.9$.

$$\text{The standard deviation } \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.9(1-0.9)}{100}} = \sqrt{0.0009} = 0.03$$

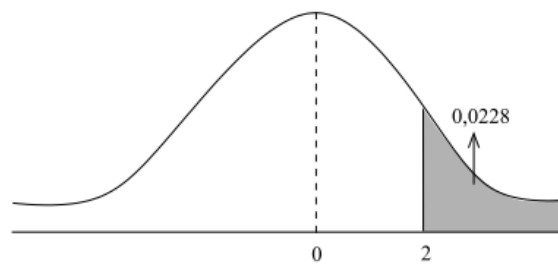
b. (i) $P(p \geq 0.96)$

Steps

1. The population proportion mean $\pi = 0.9$ and the sample proportion $p = 0.96$.
2. The sample size $n = 100$.
3. The value of Z called the “test statistic”:

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad \text{or} \quad Z = \frac{p - \pi}{\sigma_p} = \frac{0.96 - 0.9}{0.03} = 2$$

4. $P(p \geq 0.96) = P(Z \geq 2) = 0.0228$



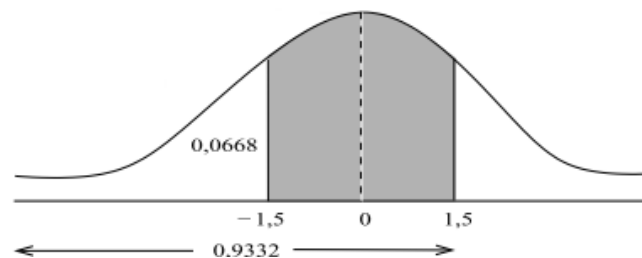
(ii) $P(0.855 \leq p \leq 0.945) = ?$

Steps

1. The population proportion mean $\pi = 0.9$.
2. The sample proportion $p = 0.855$ and $p = 0.945$.
3. The sample size $n = 100$.
4. The value of Z called the “test statistic”:

$$\text{if } p = 0.855, \text{ then } Z = \frac{p - \pi}{\sigma_p} = \frac{0.855 - 0.9}{0.03} = -1.5$$

$$\text{if } p = 0.945, \text{ then } Z = \frac{p - \pi}{\sigma_p} = \frac{0.945 - 0.9}{0.03} = 1.5$$



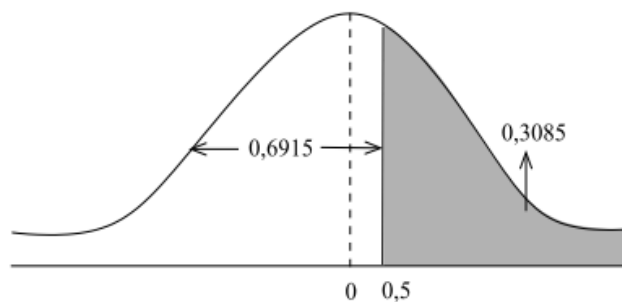
5. $P(0.855 \leq p \leq 0.945) = P(-1.5 \leq Z \leq 1.5) = P(Z \leq 1.5) - P(Z \leq -1.5)$
 $= 0.9332 - 0.0668 = 0.8664$

(iii) $P(p \geq 0.915) = ?$

Steps

1. The population proportion mean $\pi = 0.9$ and the sample proportion $p = 0.915$.
2. The sample size $n = 100$.
3. The value of Z called the “test statistic”:

$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} \quad \text{or} \quad Z = \frac{p - \pi}{\sigma_p} = \frac{0.915 - 0.9}{0.03} = 0.5$$



4. $P(p \geq 0.915) = P(Z \geq 0.5) = 1 - 0.6915 = 0.3085$

7.4 SELF-ASSESSMENT EXERCISE FOR SECTION 7.2

Question 1

Time spent using e-mail per session is normally distributed, with a population mean of 8 minutes and a population standard deviation of 2 minutes. Select a random sample of 16 sessions.

- a. What is the probability that the sample mean is between 7.8 and 8.2 minutes?
- b. If you select a random sample of 100 sessions, what is the probability that the sample mean is between 7.8 and 8.2 minutes?

Question 2

Consider an infinite population with a mean of 160 and a standard deviation of 25. A random sample of size 64 is taken from this population. The standard error of the mean equals

1. 0.391
2. 6.4
3. 2.50
4. 9.766
5. 3.125

Question 3

The standard error of the mean is the ...

1. standard deviation of the sampling distribution.
2. squared value of the population variance.
3. same value as the population standard deviation.
4. same for distributions of all sample sizes.
5. mean of the sampling distribution.

Question 4

A manufacturing company packages peanuts for Piedmont Airlines. The individual packages weigh 1.4 g with a standard deviation of 0.6 g. For a flight of 152 passengers receiving the peanuts, the probability that the average weight of the packages is less than 1.3 g is

1. 0.0202
2. 0.2040
3. 0.9798
4. 0.4798
5. 2.0500

Question 5

The fill amount of bottles of a soft drink is normally distributed, with a mean of 2.0 litres and a standard deviation of 0.06 litre. You select a random sample of 36 bottles.

- a. What is the probability that the sample mean will be between 1.99 and 2.0 litres?
- b. What is the probability that the sample mean will be below 1.98 litres?
- c. What is the probability that the sample mean will be greater than 2.01 litres?
- d. The probability is 99% that the sample mean amount of soft drink will be at least how much?
- e. The probability is 99% that the sample mean amount of soft drink will be between which two values symmetrically distributed around the mean?

7.5 SELF-ASSESSMENT EXERCISE FOR SECTION 7.3

Question 1

In each of the following cases, find the mean, the variance and the standard deviation of the sampling distribution of the sample proportion p .

- a. $\pi = 0.5, n = 250$
- b. $\pi = 0.98, n = 1\ 000$

Question 2

A political pollster is conducting an analysis of sample results in order to make predictions on election night. Assuming a two-candidate election, if a specific candidate receives at least 55% of the votes in the sample, then that candidate will be forecast as the winner of the election. If you select a random sample of 100 voters, what is the probability that a candidate will be forecast as the winner when the true percentage of her vote is:

- a. 50.1%?
- b. 49%?

Question 3

According to Gallup's poll on personal finances, 46% of US workers feel that they will have enough money to live comfortably when they retire. You select a random sample of 200 US workers.

- a. What is the probability that the sample will have between 45% and 55% workers who say they have enough money to live comfortably now and expect to do so in future?
- b. The probability is 90% that the sample percentage will be contained within what symmetrical limits of the population percentage?

7.6 SOLUTIONS TO THE SELF-ASSESSMENT EXERCISE FOR SECTION 7.2

Question 1

a. $P(7.8 < \bar{X} < 8.2) = ?$

The given information is $\mu = 8$, $\sigma = 2$, $n = 16$.

Steps

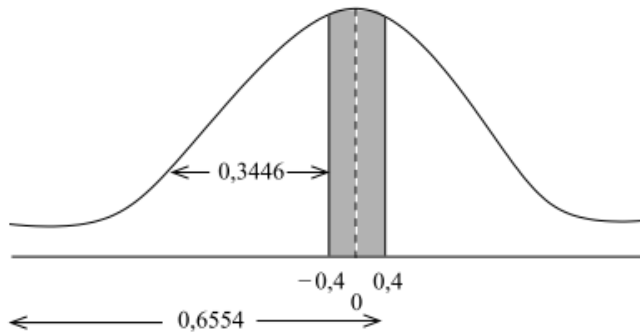
1. The transformation formula is the test statistic $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

2. Substitute the values into the Z formula.

$$\text{If } \bar{X} = 7.8, \text{ then } Z = \frac{7.8 - 8}{\frac{2}{\sqrt{16}}} = \frac{-0.2}{\frac{2}{4}} = \frac{-0.2}{0.5} = -0.4$$

$$\text{If } \bar{X} = 8.2, \text{ then } Z = \frac{8.2 - 8}{\frac{2}{\sqrt{16}}} = \frac{0.2}{\frac{2}{4}} = \frac{0.2}{0.5} = 0.4$$

3. $P(7.8 < \bar{X} < 8.2) = P(-0.4 < Z < 0.4)$, now determine the area between -0.4 and 0.4 .



4. The value using the cumulative standard normal distribution table is:

$$P(-0.4 < Z < 0.4) = P(Z < 0.4) - P(Z < -0.4) = 0.6554 - 0.3446 = 0.3108$$

- b. $P(7.8 < \bar{X} < 8.2) = ?$

The given information is $\mu = 8$, $\sigma = 2$, $n = 100$.

Steps

1. Use the transformation formula called the “test statistic” $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

2. Substitute the values into the Z formula $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

$$\text{If } \bar{X} = 7.8, \text{ then } Z = \frac{7.8 - 8}{\frac{2}{\sqrt{100}}} = \frac{-0.2}{\frac{2}{10}} = \frac{-0.2}{0.2} = -1$$

$$\text{If } \bar{X} = 8.2, \text{ then } Z = \frac{8.2 - 8}{\frac{2}{\sqrt{100}}} = \frac{0.2}{\frac{2}{10}} = \frac{0.2}{0.2} = 1$$

3. $P(7.8 < \bar{X} < 8.2) = P(-1 < Z < 1)$, now determine the area between -1 and 1 .

4. The value using the cumulative standard normal distribution table is:

$$\begin{aligned} P(-1 < Z < 1) &= P(Z < 1) - P(Z < -1) \\ &= 0.8413 - 0.1587 = 0.6826 \end{aligned}$$

Question 2

Steps

1. Population standard deviation $\sigma = 25$.
2. The sample size $n = 64$.

$$\therefore \text{The standard error of mean } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{25}{\sqrt{64}} = \frac{25}{8} = 3.125$$

Question 3

Option 1

Question 4

$$P(\bar{X} < 1.3) = ?$$

Steps

1. The test statistic is $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

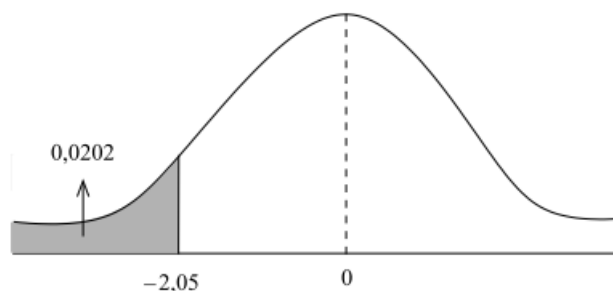
The given information is $\mu = 1.4$, $\sigma = 0.6$, $n = 152$, $\bar{X} = 1.3$.

2. Substitute the values into the Z formula:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1.3 - 1.4}{\frac{0.6}{\sqrt{152}}} = \frac{-0.1}{\frac{0.6}{12.3288}} = \frac{-0.1}{0.0487} = -2.05$$

3. Determine the equivalent number of the sample mean for which we want to find the probability.

$$P(\bar{X} < 1.3) = P(Z < -2.05), \text{ now determine the area that is less than } -2.05.$$



4. Find the value using the cumulative standard normal distribution table E.2 (from the appendix).

$$P(Z < -2.05) = 0.0202$$

Question 5

Given information: population mean $\mu = 2.0$, sample mean $\bar{X} = 1.99$, sample size $n = 36$.

a. $P(1.99 < \bar{X} < 2.0) = ?$

Steps

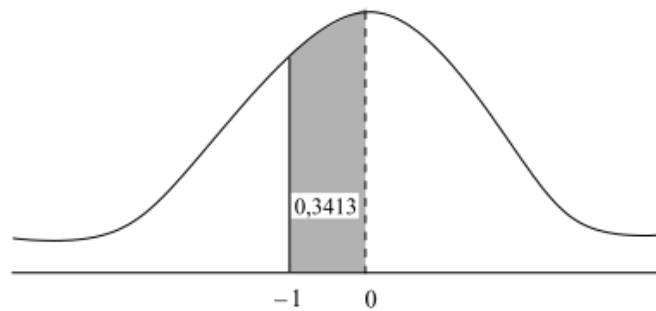
1. The test statistic is $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

2. Substitute the values into the Z formula:

$$\text{when } \bar{X} = 1.99, \text{ then } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1.99 - 2.0}{\frac{0.06}{\sqrt{36}}} = \frac{-0.01}{\frac{0.06}{6}} = \frac{-0.01}{0.01} = -1$$

$$\text{when } \bar{X} = 2.0, \text{ then } Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.0 - 2.0}{\frac{0.06}{\sqrt{36}}} = \frac{0}{\frac{0.06}{6}} = \frac{0}{0.01} = 0$$

3. $P(1.99 < \bar{X} < 2.0) = P(-1 < Z < 0) = P(Z < 0) - P(Z < -1)$, now determine the area between 0 and -1.



4. The value using the cumulative standard normal distribution table E.2:

$$P(Z < 0) = 0.5$$

$$P(Z < -1) = 0.1587$$

$$\therefore P(-1 < Z < 0) = P(Z < 0) - P(Z < -1) = 0.5 - 0.1587 = 0.3413$$

b. $P(\bar{X} < 1.98) = ?$

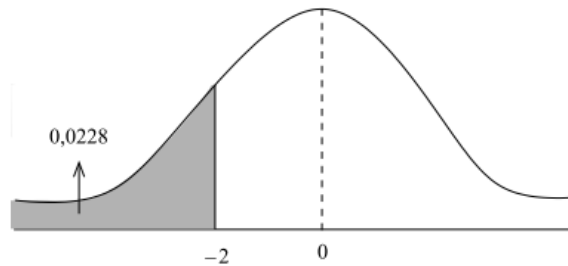
Steps

1. The test statistic is $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

2. Substitute the values into the Z formula:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1.98 - 2.0}{\frac{0.06}{\sqrt{36}}} = \frac{-0.02}{\frac{0.06}{6}} = \frac{-0.02}{0.01} = -2$$

3. $P(\bar{X} < 1.98) = P(Z < -2)$, now determine the area that is less than -2 .



4. The value using the cumulative standard normal distribution table is $P(Z < -2) = 0.0228$.

c. $P(\bar{X} > 2.01) = ?$

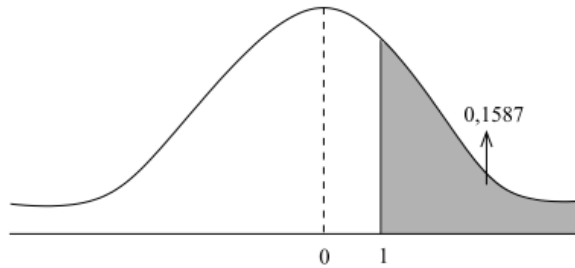
Steps

1. The test statistic is $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$.

2. Substitute the values into the Z formula:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.01 - 2.0}{\frac{0.06}{\sqrt{36}}} = \frac{0.01}{\frac{0.06}{6}} = \frac{0.01}{0.01} = 1$$

3. $P(\bar{X} > 2.01) = P(Z > 1)$, now determine the area which is greater than 1.



4. The value using the cumulative standard normal distribution table is:

$$P(Z > 1) = P(Z < -1) = 0.1587$$

d. $P(\bar{X} > a) = 0.99$

$$P\left(Z > \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 0.99$$

$\bar{X} = \mu + Z \times \frac{\sigma}{\sqrt{n}}$ the Z-value corresponding to the area 0.99 is -2.33 (using the cumulative standardised normal table)

$$= 2.0 - 2.33 \times \frac{0.06}{\sqrt{36}} = 1.9767$$

e. The area between A and B equals 0.99. The Z-value corresponding to the area 0.99 is 2.58. The A and B values associated with a known probability is given by:

$$\begin{aligned} A &= \mu - Z \times \frac{\sigma}{\sqrt{n}} \\ &= 2.0 - 2.58 \times \frac{0.06}{\sqrt{36}} = 1.9742 \end{aligned}$$

$$\begin{aligned} B &= \mu + Z \times \frac{\sigma}{\sqrt{n}} \\ &= 2.0 + 2.58 \times \frac{0.06}{\sqrt{36}} \\ &= 2.0258 \end{aligned}$$

7.7 SOLUTIONS TO THE SELF-ASSESSMENT EXERCISE FOR SECTION 7.3

Question 1

a. $\pi = 0.5$ $n = 250$

The mean $\pi = 0.5$.

The standard deviation for proportion:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.5(1-0.5)}{250}} = \sqrt{0.001} = 0.0316$$

The variance $\sigma_p^2 = (0.0316)^2 = 0.001$.

b. $\pi = 0.98$ $n = 1\,000$

The mean $\pi = 0.98$.

The standard deviation for proportion:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.98(1-0.98)}{1000}} = \sqrt{0.000019} = 0.0044$$

The variance $\sigma_p^2 = (0.0044)^2 = 0.000019$.

Question 2

Given information: sample proportion $p = 55\% = 0.55$, sample size $n = 200$

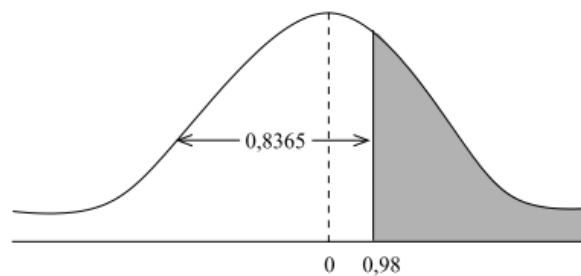
a. Population proportion $\pi = 50.1\% = 0.501$ $P(p > 0.55) = ?$

Steps

1. The test statistic $Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.55 - 0.501}{\sqrt{\frac{0.501(1-0.501)}{100}}} = \frac{0.049}{\sqrt{0.0025}} = \frac{0.049}{0.05} =$

0.98

2. $P(p > 0.55) = P(Z > 0.98)$, now determine the area greater than 0.98.



2. The value using the cumulative standard normal distribution table is:

$$P(Z > 0.98) = 1 - P(Z < 0.98) = x = 0.1635$$

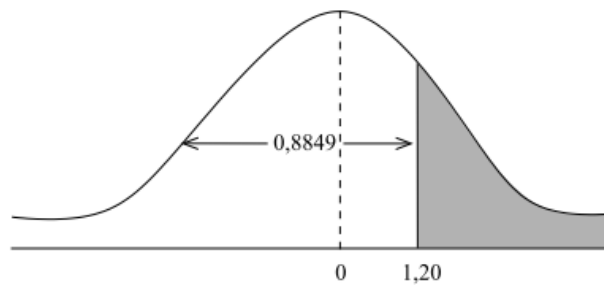
b. Population proportion $\pi = 0.49$.

Given information: $p = 0.55$; $n = 100$

Steps

1. The test statistic $Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{0.55 - 0.49}{\sqrt{\frac{0.49(1 - 0.49)}{100}}} = \frac{0.06}{0.05} = 1.20$

2. $P(p > 0.55) = P(Z > 1.20)$, now determine the area greater than 1.20.



3. The value using the cumulative standard normal distribution table is:

$$P(Z > 1.20) = 1 - P(Z < 1.20) = 1 - 0.8849 = 0.1151$$

Question 3

a. $P(0.45 < p < 0.55)$

Given information: population proportion $\pi = 0.46$, sample proportions $p = 0.45$ and $p = 0.55$, $n = 200$

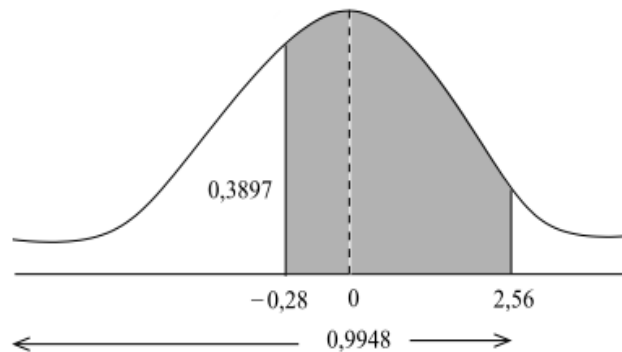
Steps

1. The test statistic $Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$.

$$\text{when } p = 0.45, \text{ then } Z = \frac{0.45 - 0.46}{\sqrt{\frac{0.46(1 - 0.46)}{200}}} = \frac{-0.01}{0.0352} = -0.2841$$

$$\text{when } p = 0.55, \text{ then } Z = \frac{0.55 - 0.46}{\sqrt{\frac{0.46(1 - 0.46)}{200}}} = \frac{0.09}{0.0352} = 2.5568$$

2. $P(0.45 < p < 0.55) = P(-0.2841 < Z < 2.5568)$, now determine the area between -0.2841 and 2.5568 .



3. The value using the cumulative standard normal distribution table is:

$$\begin{aligned} P(-0.2841 < Z < 2.5568) &= P(Z < 2.5568) - P(Z < -0.2841) \\ &= 0.9948 - 0.3897 = 0.6051 \end{aligned}$$

- b. The area between A and B represents 0.90. The Z -value corresponding to the area 0.90 is 1.645. The A and B values associated with a known probability is given by:

$$A = \pi - Z \times \sqrt{\frac{\pi(1-\pi)}{n}} = 0.46 - 1.645 \times \sqrt{\frac{0.46(1-0.46)}{200}} = 0.4021$$

$$B = \pi + Z \times \sqrt{\frac{\pi(1-\pi)}{n}} = 0.46 + 1.645 \times \sqrt{\frac{0.46(1-0.46)}{200}} = 0.5179$$

$$\therefore P(0.4021 < p < 0.5179)$$

7.8 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to

- understand the concept of the sampling distribution
- calculate probabilities related to the sample mean and the sample proportion
- understand the importance of the Central Limit Theorem

STUDY UNIT 8

CONFIDENCE INTERVAL ESTIMATION

Key questions for this unit

What is meant by the concepts “point estimate”, “standard deviation known”, “standard deviation unknown”, “level of confidence”, “level of significance”, “critical value”, “degrees of freedom” and “student’s t distribution”?

Define a confidence interval estimate.

Distinguish between a confidence interval estimate for the mean and a confidence interval estimate for proportion.

Distinguish between a confidence interval estimate for the mean when σ is known and a confidence interval estimate for the mean when σ is unknown.

What is the purpose of constructing a confidence interval estimate?

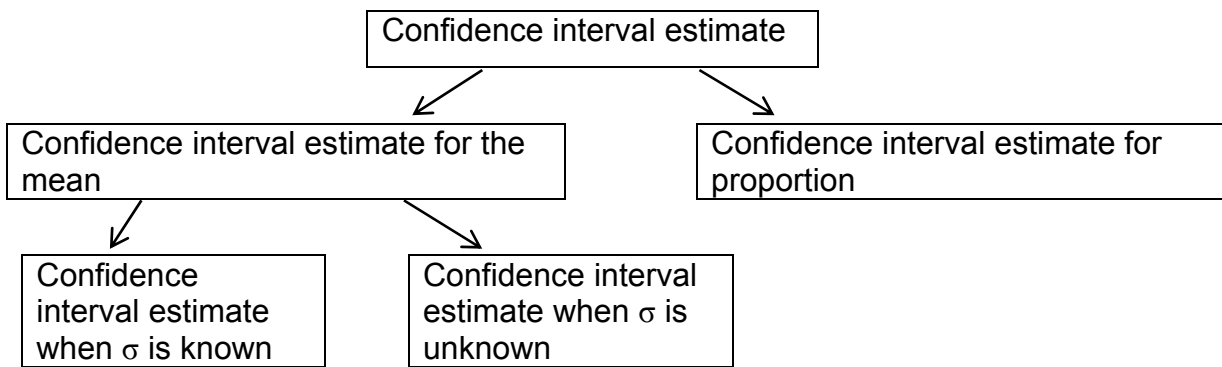
8.1 INTRODUCTION

In this study unit we use inferential statistics, the process of using sample results to estimate unknown population parameters such as population mean or population proportion. We estimate population parameters using either point estimates or interval estimates.

A point estimate is the value of a single sample statistic. For example, the sample mean \bar{X} is the point estimate of the population mean μ and the sample proportion p is the point estimate of the population proportion π . A confidence interval estimate is a range of numbers, called an interval, constructed around the point estimate. It is called a confidence interval, because we associate a degree of confidence that the real value of the population mean lies within this interval. Of course, the interval may or may not contain the true value of the population mean or proportion. Note that even a statistician cannot be 100% sure either way. So, what we do is to indicate a level of confidence that the true population mean or population proportion will lie within the confidence interval.

Activity 8.1 Overview Study skill

Draw a mind map of the different sections/headings you will deal with in this study unit. Then page through the unit with the purpose of completing the map.



There are two options to consider for a confidence interval estimate of the mean, depending on whether the population standard deviation σ is known or the population standard deviation σ is unknown.

Activity 8.2: Concepts Conceptual skill Communication skill

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Confidence level		
Confidence interval		
Point estimate		
Population standard deviation σ known		
Population standard deviation σ unknown		
Critical value		
Degrees of freedom		

8.2 CONFIDENCE INTERVAL ESTIMATE FOR THE MEAN WHEN THE POPULATION STANDARD DEVIATION IS KNOWN

When the population standard deviation σ is known, the normal distribution is used in the construction of the interval.

Steps

1. Determine the sample mean \bar{X} .
2. Determine the population standard deviation σ .
3. Determine the sample size n .
4. Determine the Z -value called the “critical value” corresponding to the level of confidence $(1 - \alpha)\%$.
5. The confidence interval estimate for the population mean is:

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$$

$$\left(\bar{X} - Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}, \quad \bar{X} + Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} \right)$$

Lower limit of the interval Upper limit of the interval

6. Substitute the values into the above formula.

Activity 8.3

Question 1

The owner of a large shopping centre is besieged with complaints about the shortage of parking space. He feels that the 1 000 spaces are adequate. In an effort to address the problem, he obtains a sample of the average number of cars in the parking lot during prime hours. The sample of 40 has a mean of 952. Assume a population standard deviation of 396. The 95% confidence interval estimate for prime-hour parking is

1. 790.46 to 1 112.54
2. 849.31 to 1 054.69
3. 829.28 to 1 074.72
4. 932.60 to 971.40
5. 952.00 to 1 052.00

Question 2

If $\bar{X} = 120$, $\sigma = 24$ and $n = 36$, construct a 99% confidence interval estimate of the population mean μ .

Feedback on the activity

Question 1

Steps

1. Sample mean $\bar{X} = 952$.
2. Population standard deviation $\sigma = 396$.
3. Sample size $n = 40$.
4. Select $Z_{\frac{\alpha}{2}} = 1.96$ at a 95% confidence interval.
5. The confidence interval estimate formula is $\bar{X} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$.
6. Substitution of the values into Z formula: $952 \pm 1.96 \times \frac{396}{\sqrt{40}}$
 952 ± 122.7217
 $(952 - 122.7217, 952 + 122.7217)$
 $(829.2783, 1\ 074.7217)$

Question 2

Steps

1. Sample mean $\bar{X} = 120$.
2. Population standard deviation $\sigma = 24$.
3. Sample size $n = 36$.
4. Use the critical value $Z_{\frac{\alpha}{2}} = 2.58$ for a 99% confidence interval estimate.
5. Substitute the values into the Z formula: $120 \pm 2.58 \times \frac{24}{\sqrt{36}} = 120 \pm 10.32$
 $(120 - 10.32, 120 + 10.32)$
 $(109.68, 130.32)$

8.3 CONFIDENCE INTERVAL ESTIMATE FOR THE MEAN WHEN THE POPULATION STANDARD DEVIATION IS UNKNOWN

When the population standard deviation is unknown, we need to construct a confidence interval estimate of μ using the sample standard deviation S as an estimate of the population standard deviation σ . Returning to the idea of a confidence interval estimate of the mean (σ known), the normal distribution was used in the construction of the interval. If σ is unknown, the student's t distribution is used with $n - 1$ degrees of freedom.

- The degrees of freedom $df = n - 1$.
- The critical value of t for the appropriate degrees of freedom from the table for the t distribution. The confidence interval for the mean (σ unknown):

$$\bar{X} \pm t_{(n-1, \frac{\alpha}{2})} \times \frac{S}{\sqrt{n}}$$

$$\left(\bar{X} - t_{(n-1, \frac{\alpha}{2})} \times \frac{S}{\sqrt{n}}, \quad \bar{X} + t_{(n-1, \frac{\alpha}{2})} \times \frac{S}{\sqrt{n}} \right)$$

Lower limit of the interval Upper limit of the interval

Steps

1. Determine the sample mean \bar{X} .
2. Determine the sample standard deviation S .
3. Determine the sample size n .
4. Determine the degrees of freedom $df = n - 1$.
5. Find the critical value using the student's t table with $t_{(n-1, \frac{\alpha}{2})}$.
6. Substitute the values into the confidence interval estimate for the mean (σ unknown).

Activity 8.4

Question 1

For a selected month, the average kilowatt-hours used by 49 residential customers is 1 160 kWh and the standard deviation S is 1 085 kWh. Assume that the t -value for a 95% confidence interval is 1.6772. Determine the confidence interval estimate for the true mean.

Question 2

A stationery store wants to estimate the mean retail value of greeting cards that it has in stock. A random sample of 100 greeting cards indicates a mean value of R2,65 and a standard deviation of R0,44. Assuming a normal distribution, construct a 95% confidence interval estimate of the mean value of all greeting cards that the store has in stock.

Feedback on the activity

Question 1

Steps

1. The sample mean $\bar{X} = 1\ 160$.
2. The sample standard deviation $S = 1\ 085$.
3. The sample size $n = 49$.
4. The degrees of freedom $df = n - 1 = 49 - 1 = 48$.
5. The critical value equals 1.6772.
6. Substitute the values into the confidence interval estimate for the mean (σ unknown) formula:

$$1\ 160 \pm 1.6772 \times \frac{1\ 085}{\sqrt{49}}$$

$$1\ 160 \pm 259.966$$

$$(1\ 160 - 259.966, \quad 1\ 160 + 259.966)$$

$$(900.034, \quad 1\ 419.966)$$

Lower limit of the interval Upper limit of the interval

Question 2

Steps

1. The sample mean $\bar{X} = 2.65$.
2. The sample standard deviation $S = 0.44$.
3. The sample size $n = 100$.
4. The degrees of freedom $df = n - 1 = 100 - 1 = 99$.
5. The critical value at $t_{(99, 0.025)}$ equals 1.9842.
6. Substitute the values into the confidence interval estimate for the mean (σ unknown) formula:

$$2.65 \pm 1.9842 \times \frac{0.44}{\sqrt{100}}$$

$$2.65 \pm 0.0873$$

$$(2.65 - 0.0873, \quad 2.65 + 0.0873)$$

$$(2.5627, \quad 2.7373)$$

Lower limit of the interval Upper limit of the interval

HOW DO YOU CALCULATE THE PROBABILITY FOR THE SAMPLING DISTRIBUTION OF THE MEAN WHEN σ IS UNKNOWN?

Let us recall that in study unit 7.3 the probability of a sampling distribution of the

mean was calculated using the value $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. If the population standard deviation

is unknown, the following statistic is used: $t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$

This expression has the same form as the Z statistic except that S is used to estimate the unknown σ .

8.4 CONFIDENCE INTERVAL ESTIMATE FOR PROPORTION

This section is concerned with estimating the proportion of items in a population having a certain characteristic of interest. The unknown population proportion is π and the sample proportion is $p = \frac{X}{n}$ where X is the number of items in the sample having the characteristic of interest and n is the sample size. The confidence interval estimate for the proportion is given by:

$$p \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{p(1-p)}{n}}$$

where

p is the sample proportion

$Z_{\frac{\alpha}{2}}$ is the critical value found from the standardised normal distribution

Activity 8.5

Question 1

Companies are spending more time on screening applicants than in the past. A study of 102 recruiters by Execunet found that 77 conducted internet research on candidates. Construct a 95% confidence interval estimate of the population proportion of recruiters who conduct internet research on candidates.

Question 2

Closed caption movies allow the hearing impaired to enjoy the dialogue as well as the acting. A local organisation for the hearing impaired members of the community takes a random sample of 100 movies offered by a cable television company in order to estimate the proportion of closed caption movies offered. Fourteen movies were closed captioned. The cable television company says at least 5% of the movies shown are captioned. Use $Z_{\frac{\alpha}{2}} = 1.65$ to find a 90% confidence interval estimate for true proportion and comment on the cable television company's claim.

Feedback on the activity

Question 1

Steps

1. The sample proportion $p = \frac{X}{n} = \frac{77}{102} = 0.7549$.
2. The sample size $n = 102$.
3. The critical value $Z_{\frac{\alpha}{2}} = 1.96$ at a 95% confidence level.
4. Substitute the values into the confidence interval estimate for proportion

$$\text{formula: } 0.7549 \pm 1.96 \times \sqrt{\frac{0.7549(1-0.7549)}{102}}$$

$$0.7549 \pm 0.0835$$

$$(0.7549 - 0.0835, 0.7549 + 0.0835)$$

$$(0.6714, 0.8384)$$

Question 2

Steps

1. The sample proportion $p = \frac{X}{n} = \frac{14}{100} = 0.14$.
2. The sample size $n = 100$.
3. The critical value $Z_{\frac{\alpha}{2}} = 1.65$ at a 90% confidence level.
4. Substitute the values into the confidence interval estimate for the proportion

$$\text{formula: } 0.14 \pm 1.65 \times \sqrt{\frac{0.14(1-0.14)}{100}}$$

$$0.14 \pm 0.0573$$

$$(0.14 - 0.0573, 0.14 + 0.0573)$$

$$(0.0827, 0.1973)$$

The interval for the population proportion is 0.0827 to 0.1973, or approximately 0.08 to 0.20. The organisation for hearing impaired people can therefore be 90% confident that the proportion of closed caption movies offered is somewhere between 0.08 (8%)

and 0.20 (20%). The cable television company is correct in saying that at least 5% of the movies it shows are closed captioned.

8.5 SELF-ASSESSMENT EXERCISE

Question 1

Your statistics instructor wants you to determine a confidence interval estimate for the mean test score. Past experience indicated that test scores are normally distributed with a sample mean of 160 and a population standard deviation of 45. A 95% confidence interval estimate if your group has 36 students is

1. 145.3 to 174.7
2. 157.55 to 162.45
3. 152.5 to 167.5
4. 158.75 to 161.25
5. 160 to 174.7

Question 2

If $\bar{X} = 70$, $S = 24$ and $n = 36$, and assuming that the population is normally distributed, construct a 95% confidence interval estimate of the population mean μ .

Question 3

The data represent the overall miles per gallon (MPG) of 2008 SUVs priced under \$30 000.

23 20 21 22 18 18 17 17 19 19 19
17 21 18 18 18 17 17 16 20 16 22

Construct a 95% confidence interval estimate for the population mean miles per gallon of 2008 SUVs priced under \$30 000 assuming a normal distribution.

Question 4

The owner of a restaurant which serves continental food wants to study characteristics of his customers. He decides to focus on two variables: the amount of money spent by customers and whether customers order dessert. The results from a sample of 60 customers are as follows:

The amount spent = R38,54 and $S = R7,26$; 18 customers ordered dessert.

- a. Construct a 95% confidence interval estimate of the population mean amount spent per customer in the restaurant.
- b. Construct a 90% confidence interval estimate of the population proportion of customers who order dessert.

8.6 SOLUTIONS TO THE SELF-ASSESSMENT EXERCISE

Question 1

Steps

1. Sample mean $\bar{X} = 160$.
2. Population standard deviation $\sigma = 45$.
3. Sample size $n = 36$.
4. Use the critical value $Z_{\frac{\alpha}{2}} = 1.96$ for a 95% confidence interval estimate.

5. Substitute the values into the Z formula: $160 \pm 1.96 \times \frac{45}{\sqrt{36}}$

$$160 \pm 14.7$$

$$(160 - 14.7, 160 + 14.7)$$

$$(145.3, 174.7)$$

Option 1

Question 2

Steps

1. The sample mean $\bar{X} = 70$.
2. The sample standard deviation $S = 24$.
3. The sample size $n = 36$.
4. The degrees of freedom $df = n - 1 = 36 - 1 = 35$.
5. The critical value at $t_{(35, 0.025)}$ equals 2.0301.
6. Substitute the values into the confidence interval estimate for the mean (σ

unknown) formula: $70 \pm 2.0301 \times \frac{24}{\sqrt{36}}$

$$70 \pm 8.1204$$

$$(70 - 8.1204, \quad 70 + 8.1204)$$

$$(61.8796, \quad 78.1204)$$

Lower limit of the interval Upper limit of the interval

Question 3

Steps

1. The sample mean $\bar{X} = \frac{23 + 20 + 21 + 22 + \dots + 20 + 16 + 22}{22} = \frac{413}{22} = 18.7727$

2. The sample standard deviation $S = \frac{\sum(X_i - \bar{X})^2}{n-1}$

$$S^2 = \frac{(23 - 18.7727)^2 + (20 - 18.7727)^2 + \dots + (22 - 18.7727)^2}{22 - 1} = \frac{85.8636}{21} = 4.0887$$

$$\therefore S = \sqrt{4.0887} = 2.0221$$

3. The sample size $n = 22$.
4. The degrees of freedom $df = n - 1 = 22 - 1 = 21$.
5. The critical value at $t_{(21, 0.025)}$ equals 2.0796.
6. Substitute the values into the confidence interval estimate for the mean (σ

unknown) formula: $18.7727 \pm 2.0796 \times \frac{2.0221}{\sqrt{22}}$

$$18.7727 \pm 0.8965$$

$$(18.7727 - 0.8965, \quad 18.7727 + 0.8965)$$

$$(17.8762, \quad 19.6692)$$

Lower limit of the interval Upper limit of the interval

Question 4

a. Steps

1. The sample mean $\bar{X} = 38.54$.
2. The sample standard deviation $S = 7.26$.
3. The sample size $n = 60$.
4. The degrees of freedom $d f = n - 1 = 60 - 1 = 59$.
5. The critical value at $t_{(59, 0.025)}$ equals 2.0010.
6. Substitute the values into the confidence interval estimate for the mean (σ

unknown) formula: $38.54 \pm 2.001 \times \frac{7.26}{\sqrt{60}}$

$$38.54 \pm 1.8755$$

$$(38.54 - 1.8755, 38.54 + 1.8755)$$

$$(36.6645, 40.4155)$$

b. Steps

1. The sample proportion $p = \frac{18}{60} = 0.3$.
2. The sample size $n = 60$.
3. The critical value $Z_{\frac{\alpha}{2}} = 1.645$ at a 90% confidence interval.
4. Substitute the values into the confidence interval estimate for proportion

formula: $0.3 \pm 1.645 \times \sqrt{\frac{0.3(1-0.3)}{60}}$

$$0.3 \pm 0.0973$$

$$(0.3 - 0.0973, 0.3 + 0.0973)$$

$$(0.2027, 0.3973)$$

8.7 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to construct and interpret confidence interval estimates for the mean and the proportion.

STUDY UNIT 9

HYPOTHESIS TESTING

Key questions for this unit

What is meant by the concepts “acceptance region”, “rejection region”, “type I error”, “type II error”, “critical value”, “power of the test”, “two-tailed test”, “one-tailed test”?

Define hypothesis testing.

Distinguish between hypothesis testing for the mean and hypothesis testing for proportion.

What is the significance level of a test?

How do you make a decision in hypothesis testing?

9.1 INTRODUCTION

Hypothesis testing is the statistical assessment of a statement or idea regarding a population. This means that we state a claim or assertion about a particular parameter of a population. For instance, a statement could be as follows: “The mean weight of cereal boxes is 368 g.” Given the results of the weight of the cereal boxes, hypothesis testing procedures can be employed to test the validity of this statement at a given significance level for a sample weight of cereal boxes. You will examine the results of the sample to see whether it supports the stated claim. This type of problem introduces you to inferential statistics.

In the previous study unit you saw that a confidence interval can be used when we have to predict the value of a population parameter. The inclusion of the parameter was never certain, but we quantified the likelihood of the parameter lying within that particular interval through the expression of a confidence level (95%, 99% or ...). The same form of quantified uncertainty is used in hypothesis testing.

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Acceptance region		
Rejection region		
Critical value		
Two-tailed test		
One-tailed test		
Significance level of the test		
The null hypothesis		
The alternative hypothesis		

9.2 FUNDAMENTAL CONCEPTS OF HYPOTHESIS TESTING

In this section, we will present the basic concepts of hypothesis testing as follows:

1. The null hypothesis, denoted by H_0 , represents the current belief in a situation.
2. The alternative hypothesis, denoted by H_1 , is the opposite of the null hypothesis and represents a research claim or specific inference you would like to prove.
3. The level of significance (α) is the probability of rejection when the null hypothesis is true. It represents the risk level that you are willing to have of rejecting the null hypothesis when it is true. You select levels of 0.01 (1%), 0.05 (5%) or 0.10 (10%).
4. The confidence coefficient is the probability that you will not reject the null hypothesis H_0 when it is true and should it not be rejected the confidence coefficient is $(1 - \alpha) \times 100\%$.
5. The acceptance region (non-rejection) is any portion of the distribution resulting in the decision to fail to reject the null hypothesis if the observed statistic falls in this region.

6. The rejection region is any portion of the distribution resulting in the decision to reject the null hypothesis if the observed statistic falls in this region.
7. The critical value is the value that divides the acceptance region and the rejection region.
8. The power of the test is the probability that you will reject H_0 when it is false and it should be rejected.
9. A two-tailed test is a statistical test when the researcher is interested in testing both sides of the distribution.
10. A one-tailed test is a statistical test when the researcher is interested in testing one side of the distribution.

HOW TO SPECIFY THE NULL AND ALTERNATIVE HYPOTHESES

The null hypothesis specifies the parameter that is equal to some particular value and it is for the alternative hypothesis to answer the question. In order for you to specify the alternative, you must determine what the question asks.

Examples:

1. If the question asks whether the waiting time to place an order has changed in the past month from its previous population mean of 4.5 minutes and the population mean is μ , then

$$H_1: \mu \neq 4.5$$

and

$$H_0: \mu = 4.5, \text{ which means the population mean is equal to } 4.5.$$

Alternatively, the question might be: "There is sufficient evidence to conclude that the waiting time to place an order is not equal to (or is different from) the previous population mean $\mu = 4.5$." Therefore you have to perform a two-tailed test.

2. If the question asks whether there is sufficient evidence to conclude that the population mean is greater than 4.5, then

$$H_1: \mu > 4.5$$

and

$$H_0: \mu = 4.5$$

Therefore you have to perform a one-tailed test.

3. If the question asks whether there is sufficient evidence to conclude that the population mean is less than 4.5, then

$$H_1: \mu < 4.5$$

and

$$H_0: \mu = 4.5$$

Therefore you have to perform a one-tailed test.

Hypothesis testing is classified as either hypothesis testing for the mean or hypothesis testing for proportion, depending on the possible samples selected or the proportion of items in a population having a certain characteristic of interest.

THE STEPS METHOD OF HYPOTHESIS TESTING

Step 1

State the null hypothesis H_0 : population parameter (μ) = hypothesised value

Step 2

State the alternative hypothesis H_1 : summarises what the case will be if the null hypothesis is not true and can assume one of three possible forms:

- a. H_1 : population parameter (μ) \neq hypothesised value
- b. H_1 : population parameter (μ) $<$ hypothesised value
- c. H_1 : population parameter (μ) $>$ hypothesised value

Step 3

Choose the level of significance (α) just to provide a probability basis for deciding whether an observed difference between a sample statistic and a hypothesised value is a chance difference or a statistically significant difference.

Step 4

Determine the appropriate test statistic and calculate the value.

Step 5

Determine the critical values that divide the rejection and non-rejection regions.

Step 6

State the decision rule.

The decision rule is a statement that indicates the action to be taken, that is, to fail to reject H_0 or to reject H_0 .

- Reject H_0 when the value of the test statistic is greater than the critical value at a specific significance level, otherwise do not reject H_0 .
- Reject H_0 when the p -value is less than the significance level.

The p -value is the lowest level of significance at which the null hypothesis can be rejected. It is determined based on the test statistic value.

Step 7

State the conclusion. This conclusion should be based on the context of the problem and the level of significance should be included.

9.3 HYPOTHESIS TESTING FOR THE MEAN

We are again going to differentiate between two distinct cases: the population standard deviation is known or it is unknown. You will see that the first distinction to make is to see whether the population or the sample standard deviation is known and then you decide whether the population is normally distributed or not.

9.3.1 Population standard deviation σ known

When the population standard deviation σ is known, the normal distribution is used in the hypothesis testing.

Steps

1. State (or identify) the null hypothesis H_0 and the alternative H_1 .
2. Choose the level of significance (α).
3. Determine the sample mean \bar{X} .
4. Determine the population mean μ .
5. Determine the population standard deviation σ .
6. Determine the sample size n .
7. Calculate the test statistic:
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$
8. Make the statistical decision.

Activity 9.2

Question 1

The quality control manager at a light bulb factory needs to determine whether the mean life of a large shipment of light bulbs is equal to the specified value of 575 hours. State the null and alternative hypotheses.

Question 2

The p -value for a hypothesis test has been reported as 0.03. If the test result is interpreted using the $\alpha = 0.05$ level of significance as a criterion, will H_0 be rejected? Explain.

Question 3

For a sample of 12 items from a normally distributed population for which the standard deviation is $\sigma = 17.0$, the sample mean is 230. At the 5% level of significance, test $H_0: \mu = 220$ versus $H_1: \mu > 220$.

- a. Calculate the test statistic.
- b. Determine the p -value for the test.

Feedback on the activity

Question 1

$H_0: \mu = 575$

$H_1: \mu \neq 575$

Question 2

Given information: $\alpha = 0.05$ p -value = 0.03

The decision rule:

Reject H_0 if the p -value is less than the level of significance α .

Since $0.03 < 0.05$, we reject H_0 at a 5% level of significance.

Question 3

a. Steps

1. The null hypothesis $H_0: \mu = 220$ and the alternative $H_1: \mu > 220$.
2. The level of significance $\alpha = 0.05$.
3. The sample mean $\bar{X} = 230$.
4. The population mean $\mu = 220$.
5. The population standard deviation $\sigma = 17.0$.
6. The sample size $n = 12$.
7. The test statistic:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{230 - 220}{\frac{17.0}{\sqrt{12}}} = \frac{10}{4.9075} = 2.0377, \text{ which is approximately } 2.04.$$

- b. p -value: $P(Z > 2.04) = 1 - 0.9793 = 0.0207$ (using statistics table E.2 from the appendices)

9.3.2 Population standard deviation σ unknown

When the population standard deviation σ is not known, the hypothesis testing procedure is the same as when it is known. Make sure you remember which table to use. You use the t -distribution table if you are given the value of S and it is wrong to use the normal distribution in such a question.

Other reminders:

- The test statistic follows a t distribution having $n - 1$ degrees of freedom and the degrees of freedom are not the sample size.
- The table values are given as positive values. If you work in the left tail of the area under the curve, you have to put a minus sign before the table value.
- Working two-tailed, you have to divide the significance level by 2 and use the answer for the table. You then use that table value twice – once in the right tail with a positive sign and once in the left tail making it negative.

Steps

1. State (or identify) the null hypothesis H_0 and the alternative H_1 .
2. Choose the level of significance (α).
3. Determine the sample mean \bar{X} .
4. Determine the population mean μ .
5. Determine the sample standard deviation S .
6. Determine the sample size n ,
7. Calculate the test statistic: $t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$
8. Make the statistical decision.

Activity 9.3

Question 1

My daughter and I have argued the average length of our preacher's sermons on Sunday mornings. Despite my arguments, she thinks that the sermons are more than twenty minutes and this is not acceptable to her. For one year she randomly selected 12 Sundays and found an average time of 26.42 minutes with a standard deviation of 6.69 minutes. Assuming that the population is normally distributed and using a 0.05 level of significance, we decided to do a scientific analysis, using a hypothesis test. Calculate the test statistic and make a statistical decision.

Question 2

A random sample of 10 observations was drawn from a normally distributed population. The data values were 6, 4, 4, 7, 5, 5, 4, 5, 6 and 4. A person tested the hypothesis $H_0: \mu \geq 6$ versus $H_1: \mu < 6$ and scribbled down his calculations. When his friend came along and quickly wanted to copy his work, he did not read properly and wrote down that the

1. sample mean is equal to 4
2. sample variance is equal to 1
3. rejection region is $t < t_{(0.05, 10)} = -1.833$.
4. test statistic is $t = 3.0$.
5. conclusion is to reject H_0 , because the test statistic $t = -3.0 < -1.833$

Which of the above statements is *correct*?

Question 3

The credit manager of a large department store claims that the mean balance for the store's charge account customers is R410. An independent auditor selects a random sample of 18 accounts and finds a mean balance of $\bar{X} = R511,33$ and a standard deviation of $S = R183,75$. If the manager's claim is not supported by these data, the auditor intends to examine all charge account balances. If the population of account balances is assumed to be approximately normally distributed, what action should the auditor take at 5% level of significance?

Feedback on the activity

Question 1

Steps

1. The null hypothesis $H_0: \mu \leq 20$ versus the alternative $H_1: \mu > 20$.
2. The level of significance $\alpha = 0.05$.
3. The sample mean $\bar{X} = 26.42$.
4. The population mean $\mu = 20$.
5. The sample standard deviation $S = 6.69$.
6. The sample size $n = 12$.

7. The test statistic:
$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{26.42 - 20}{\frac{6.69}{\sqrt{12}}} = \frac{6.42}{1.9312} = 3.3244$$

8. Conclusion

Reject H_0 if the test statistic is greater than the critical value.

Critical value is $t_{n-1; \alpha} = t_{11; 0.05} = 1.7959$

Since the test statistic 3.3244 is greater than 1.7959, H_0 can be rejected. We conclude that there is enough evidence that the alternative H_1 is true and that my daughter is correct in thinking that the average length of sermons is more than 20 minutes.

Question 2

1. The sample mean
$$\bar{X} = \frac{\sum X_i}{n} = \frac{6+4+4+7+5+5+4+5+6+4}{10} = \frac{50}{10} = 5$$
2. The variance
$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{(6-5)^2 + (4-5)^2 + (4-5)^2 + \dots + (4-5)^2}{10-1} = 1.1111$$
3. The test statistic:
$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

the standard deviation $S = \sqrt{1.1111} = 1.0541$

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{5 - 6}{\frac{1.0541}{\sqrt{10}}} = \frac{-1}{0.3333} = -3$$

4. The rejection region is $t < t_{(0.05, n-1)} = t_{(0.05, 9)} = -1.833$
5. *Correct.*

Question 3

Steps

1. The null hypothesis $H_0: \mu = 410$ versus the alternative $H_1: \mu \neq 410$.
2. For this test, let the level of significance $\alpha = 0.05$.
3. The sample mean $\bar{X} = 511.33$.
4. The population mean $\mu = 410$.
5. The sample standard deviation $S = 183.75$.
6. The sample size $n = 18$.

7. The test statistic:
$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{511.33 - 410}{\frac{183.75}{\sqrt{18}}} = \frac{101.33}{43.3103} = 2.3396$$

8. Conclusion

Reject H_0 if the test statistic is greater than the critical value of 2.1098 or less than -2.1098.

Critical value is ± 2.1098 (from the table).

Since the test statistic 2.3396 is greater than 2.1098, H_0 can be rejected. We conclude that there is enough evidence that the alternative H_1 is true and that the auditor should proceed to examine all charge account balances.

9.4 HYPOTHESIS TESTING FOR PROPORTION

Many of the principles applied in this section are not new, as they have the same “pattern” as the technique of hypothesis testing for the population mean.

Steps

1. State the null hypothesis H_0 and the alternative H_1 .
2. Choose the level of significance (α).
3. Determine the sample proportion $p = \frac{X}{n}$.
4. Determine the population proportion π .
5. Determine the sample size n .

6. Calculate the test statistic for the proportion:
$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}}$$

7. Make the statistical decision.

Activity 9.4

Question 1

A random sample of 200 observations shows that there are 36 successes. We want to test at the 1% significance level whether the true proportion of successes in the population is less than 24% and make certain calculations.

Which one of the following statements is *incorrect*?

1. The value of p is $\frac{36}{200}$.
2. The appropriate hypotheses are $H_0: \pi = 0.24$ versus $H_1: \pi < 0.24$.
3. The critical value of Z (from the table) is $Z < -Z_{0.01} = -2.33$.
4. The standard error associated with this test is 0.0302.
5. The test statistic is 1.99.

Question 2

If, in a random sample of 400 items, 164 are defective, what is the sample proportion of the defective items?

Question 3

Refer to question 2. Suppose you are testing the null hypothesis $H_0: \pi = 0.40$ against $H_1: \pi \neq 0.40$ and you choose the level of significance $\alpha = 0.05$. What is your statistical decision?

Feedback on the activity

Question 1

1. *Correct.*
2. *Correct.*
3. *Correct.*

4. *Correct.* The standard error is $\sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0.24(1-0.24)}{200}} = 0.0302$

5. *Incorrect.* The test statistic $Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.18 - 0.24}{0.0302} = \frac{-0.06}{0.0302} = -1.9868$

Question 2

The sample proportion $p = \frac{X}{n} = \frac{164}{400} = 0.41$.

Question 3

This is a two-tailed test.

The decision rule:

- Reject H_0 when the value of the test statistic is greater than or less than the critical value at a specific significance level, otherwise do not reject H_0 .
- Reject H_0 when the p -value is less than the significance level.

$$\text{The test statistic } Z = \frac{p - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.41 - 0.40}{\sqrt{\frac{0.40(1-0.40)}{400}}} = \frac{0.01}{0.0245} = 0.4082$$

The critical value equals ± 1.96 (from the normal table).

Since 0.4082 is less than 1.96, we do not reject H_0 at a 5% level of significance.

9.5 SELF-ASSESSMENT EXERCISE

Question 1

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 was found to be 2.001 cm and the standard deviation to be 0.003 cm. Is there evidence to suggest that the machine is set too high? Use $\alpha = 0.05$.

Question 2

The light bulbs in an industrial warehouse have been found to have a mean lifetime of 1 030.0 hours, with a standard deviation of 90.0 hours. The warehouse manager has been approached by a representative of Extendabulb, a company that makes a device intended to increase bulb life. The manager is concerned that the average lifetime of Extendabulb-equipped bulbs might not be any longer than the 1 030 hours historically experienced. In a subsequent test, the manager tests 40 bulbs equipped with the device and finds their mean life to be 1 061.6 hours. Does Extendabulb really work? Use $\alpha = 0.05$.

Question 3

For a simple random sample of 15 items from a population that is approximately normally distributed, $\bar{X} = 82.0$ and $S = 20.5$. At the 0.01 level of significance, test $H_0: \mu = 90$ versus $H_1: \mu \neq 90$.

Question 4

The new director of a local YMCA has been told by his predecessors that the average member has belonged to the organisation for 8.7 years. Examining a random sample of 15 membership files, he finds the mean length of membership to be 7.2 years, with a standard deviation of 2.5 years. Assuming the population is approximately normally distributed and using the 0.05 level, does this result suggest that the actual mean length of membership may be some value other than 8.7 years?

Question 5

The career services director of Hobart University has said that 70% of the school's seniors enter the job market in a position directly related to their undergraduate field of study. In a sample consisting of 200 of the graduates from last year's class, 66% have entered jobs related to their field of study. Make the related decision. Use $\alpha = 0.05$.

9.6 SOLUTIONS TO THE SELF-ASSESSMENT EXERCISE

Question 1

Steps

1. The null hypothesis $H_0: \mu = 2.0$ and the alternative $H_1: \mu > 2.0$ (this is a one-tailed test).
2. The level of significance $\alpha = 0.05$.
3. The sample mean $\bar{X} = 2.001$.
4. The population mean $\mu = 2.0$.
5. The population standard deviation $\sigma = 0.003$.
6. The sample size $n = 50$.

7. The test statistic:
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.001 - 2.0}{\frac{0.003}{\sqrt{50}}} = \frac{0.001}{0.00042} = 2.357$$

The critical value is equal to 1.645.

8. Decision

Since $2.357 > 1.645$, H_0 is rejected at a 5% level of significance. The sample results suggest that the machine is set too high.

Question 2

Steps

1. The null hypothesis $H_0: \mu \leq 1\,030.0$ and the alternative $H_1: \mu > 1\,030.0$ (this is a one-tailed test).
2. The level of significance $\alpha = 0.05$.
3. The sample mean $\bar{X} = 1\,061.6$.
4. The population mean $\mu = 1\,030.0$.
5. The population standard deviation $\sigma = 90$.
6. The sample size $n = 40$.

7. The test statistic:
$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1061.6 - 1030.0}{\frac{90}{\sqrt{40}}} = \frac{31.6}{14.23025} = 2.2206$$

8. The critical value is equal to 1.645.

9. Decision

Since $2.206 > 1.645$, H_0 is rejected at a 5% level of significance.

The results suggest that Extendabulb does increase the mean lifetime of the bulbs. This firm may wish to incorporate Extendabulb into its warehouse lighting system.

Question 3

$$\text{The test statistic: } t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{82.0 - 90}{\frac{20.5}{\sqrt{15}}} = \frac{-8}{5.2931} = -1.5114$$

The critical value is equal to ± 2.9768 .

Decision

Since $-2.9768 < -1.5114 < 2.9768$, H_0 is not rejected at a 1% level of significance.

Question 4

Steps

1. The null hypothesis $H_0: \mu = 8.7$ versus the alternative $H_1: \mu \neq 8.7$.
2. The level of significance $\alpha = 0.05$.
3. The sample mean $\bar{X} = 7.2$.
4. The population mean $\mu = 8.7$.
5. The sample standard deviation $S = 2.5$.
6. The sample size $n = 15$ and the degrees of freedom $df = n - 1 = 15 - 1 = 14$.

$$7. \text{ The test statistic: } t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{7.2 - 8.7}{\frac{2.5}{\sqrt{15}}} = \frac{-1.5}{0.6455} = -2.3238$$

8. The critical values are $t = -2.145$ and $t = 2.145$.

9. Conclusion

Since the calculated test statistic falls in the rejection region, we reject H_0 .

At the 0.05 level, the results suggest that the actual mean length of membership may be some value other than 8.7 years.

Question 5

Steps

1. The value of p is 0.66.
2. The appropriate hypotheses are $H_0: \pi = 0.70$ versus $H_1: \pi \neq 0.70$.
3. The critical values of Z are -1.96 and 1.96 .
4. The test statistic:
$$Z = \frac{p - \pi}{\sqrt{\frac{\pi(1 - \pi)}{n}}} = \frac{0.66 - 0.70}{\sqrt{\frac{0.70(1 - 0.70)}{200}}} = \frac{-0.04}{0.0324} = -1.2346$$

The test statistic value falls between the two critical values. The null hypothesis is not rejected. We conclude that the proportion of graduates who enter the job market in careers related to their field of study could indeed be equal to the claimed value of 0.70. This analysis would suggest that the director's assertion not be challenged.

9.7 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to

- understand the fundamental concepts of hypothesis testing
- distinguish between hypothesis testing for the mean and hypothesis testing for proportion
- make a decision in hypothesis testing

STUDY UNIT 10

CHI-SQUARE DISTRIBUTION

Key questions for this unit

What is meant by the concepts “Chi-square”, “test of independence”, “observed and expected frequencies”, “critical area”, “contingency table”, “degrees of freedom”?

What are the steps to test whether two nominal variables are related?

Under what conditions should you use the χ^2 test of independence?

10.1 INTRODUCTION

This study unit focuses on hypothesis testing on two categorical variables having two or more categories. Chi-square analysis is used to test whether there is a relationship between these variables.

In this study unit we are going to consider only section 11.3 in the prescribed textbook. You may read the rest of the chapter if you are interested, but you will not be examined on that information.

10.2 BASIC CONCEPTS OF CHI-SQUARE TESTING

You have to understand the basic concepts of Chi-square testing and be able to test for the independence of two variables. Going through the general characteristics of the Chi-square distribution, you should understand that the χ^2 -distribution is

- continuous
- the sampling distribution of $((n - 1)s^2 / \sigma^2)$
- always positive because $(n - 1)$, s^2 and σ^2 are all always positive
- a family of distributions, where every family member is determined by its number of degrees of freedom (df)
- skewed, but as the df increases, the form of the χ^2 -distribution becomes more and more like the bell shape of the normal distribution
- tabulated and mostly used for right-tailed areas

- a reflection of the extent to which a table of observed frequencies differs from one constructed under the assumption that the particular null hypothesis is true

10.3 TESTING FOR INDEPENDENCE OF TWO VARIABLES

This is a special technique used to test whether or not two categorical variables could be independent of each other.

The test procedure involves the following steps:

- Start with a contingency table of observed frequencies reflecting the intersection of the various categories of the two variables.
- Formulate the null and alternative hypotheses.
- Construct tables of observed and expected frequencies.
- Calculate the value of the χ^2 test statistic.
- Identify the critical value of the χ^2 statistic at the significance level specified for the particular question.
- Draw a conclusion regarding the statements in the hypotheses after comparing the critical value and the value of the test statistic.

Summary of the Chi-square test for independence:

Test for independence of two variables

H_0 : Variables independent of each other

H_1 : Variables not independent of each other

$$df = (r - 1)(c - 1)$$

Expected frequency per cell, namely f_e

$$f_e = \frac{\text{row total} \times \text{column total}}{n}$$

Test statistic:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Activity 10.1: Concepts	Conceptual skill	Communication skill
-------------------------	------------------	---------------------

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Chi-square		
Test for independence		
Critical value		
Degrees of freedom		
Contingency table		
Observed frequency		
Expected frequency		

Activity 10.2

Question 1

The quality manager of a tyre manufacturing plant in Port Elizabeth wants to test whether the nature of defects found in manufactured tyres depends on the shift during which the defective tyres were produced. Formulate the hypothesis of this test.

Question 2

A large carpet store wishes to determine whether the brand of carpet purchased is related to the purchaser's family income. As a sampling frame, they mailed a survey to people who have a store credit card. Five hundred customers returned the survey and the results follow:

Family income	Brand of carpet		
	Brand A	Brand B	Brand C
High income	65	32	32
Middle income	80	68	104
Low income	25	35	59

The statements below refer to a test conducted on the data above to determine whether the brand of carpet purchased is related to the purchaser's family income at the 5% level of significance. Select the *incorrect* statement.

1. H_0 : Family income and brand of carpet are independent.
 H_1 : Family income and brand of carpet are dependent.
2. Rejection region: reject H_0 if calculated $\chi^2 > \chi_{0.05,4}^2 = 9.488$.
3. The estimated frequencies are as follows:

Family income	Brand of carpet		
	Brand A	Brand B	Brand C
High income	43.86	34.83	40.31
Middle income	85.68	68.04	98.28
Low income	24.46	32.13	46.41

4. The calculated χ^2 value is 27.372.
5. We can conclude that the brand of carpet purchased is related to the purchaser's family income.

Feedback on the activity

Question 1

H_0 : The nature of defects found in manufactured tyres and the shift during which they were produced are independent.

H_1 : The nature of defects found in manufactured tyres and the shift during which they were produced are related (dependence).

Question 2

Option 3

1. *Correct.*
2. *Correct.*
3. *Incorrect.*

The error lies in the table of expected frequencies.

The calculation of each cell is: $f_e = \frac{\text{row total} \times \text{column total}}{n}$

For example:

Family income	Brand of carpet		
	Brand A	Brand B	Brand C
High income	65	32	32
Middle income	80	68	104
Low income	25	35	59

Frequency for row 1, column 1: $f_e = \frac{129 \times 170}{500} = 43.86$

Two values were calculated incorrectly, indicated below in bold.

The table should be as follows:

Family income	Brand of carpet		
	Brand A	Brand B	Brand C
High income	43.86	34.83	50.31
Middle income	85.68	68.04	98.28
Low income	40.46	32.13	46.41

4. *Correct.* Because the calculation of the χ^2 -value is very tedious and you may not know where you made a calculation error, I will show you the manual calculation of this value:

f_0	f_e	$\frac{(f_0 - f_e)^2}{f_e}$
65	43.85	10.1892
32	34.83	0.2299
32	50.31	6.6638
80	85.68	0.3765
68	68.04	0
104	98.28	0.3327
25	40.46	5.9074
35	32.13	0.2564
59	46.41	3.4154

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

$$= 27.372$$

5. *Correct.* Because $27.372 > \chi_{0.01,4}^2 = 13.277$, we can reject H_0 and there is a significant relationship between the brand of the carpet and family income.

10.4 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to

- understand the nature and procedures involved in Chi-square testing in general
- calculate the expected frequencies for the Chi-square test of independence discussed in this study unit
- determine the value of the χ^2 test statistic
- apply the Chi-square distribution to test whether two nominal variables are independent or not

STUDY UNIT 11

REGRESSION AND CORRELATION ANALYSIS

Key questions for this unit

What is meant by the concepts “regression and correlation analysis”, “scatter plot”, “least-squares method”, “regression coefficients”, “dependent and independent variables”, “slope”, “y-intercept”, “interpolation”, “extrapolation”, “coefficients of correlation and determination”?

How would you calculate and interpret the regression coefficients?

How would you estimate the Y-variable using the X-variable?

How do you calculate and interpret coefficients of correlation and determination?

11.1 INTRODUCTION

Simple linear regression enables us to develop a model for the prediction of numerical variables based on the value of other variables. The variable we wish to predict is called dependent (y) and the one used for the prediction is called independent (x). In simpler terms, simple linear regression describes and evaluates the relationship between two variables.

In this study unit we are going to consider only sections 12.2 and 12.3 in the prescribed textbook. You may read the rest of the chapter if you are interested, but you will not be examined on that information.

Activity 11.1: Concepts	Conceptual skill	Communication skill
-------------------------	------------------	---------------------

Test your own knowledge (write in pencil) and then correct your understanding afterwards (erase and write the correct description). Often a young language may not have words for all the terms in a discipline. Can you think of some examples?

English term	Description	Term in your home language
Regression analysis		
Correlation		
Scatter plot		
Least-squares method		
Regression coefficients		
Dependent variable		
Independent variable		
Slope		
<i>Y</i> -intercept		
Interpolation		
Extrapolation		
Coefficient of correlation		
Coefficient of determination		

11.2 THE SIMPLE LINEAR REGRESSION LINE

If you understand the basics of the straight-line equation from school, you should have no problem to understand the concept of a linear regression line.

Suppose you have a data set of paired observations, that is, two observations have been recorded for each object under study. If the objects under study were people, suppose one set of observations indicate their lengths and the other set their masses. Then the paired observation per person would be the combined pair (length; mass). Should you take a piece of graph paper, draw two perpendicular axes for length and mass, respectively, and record the combined pair per person in this two-dimensional plane, you now have what statisticians call a scatter plot of the values.

Looking at a scatter plot, the question is what you can deduce. Remember that you are a scientist and you want proper backing for what you say! This is when you use a simple linear regression model, which is scientifically acceptable. With this regression line you may even make predictions for the one variable based on the values of the other. This means that if you consider the length of the person as the independent variable, you can use the regression line equation to predict the mass of a person once you know what his/her length is. Never lose perspective – you cannot say that the answer in such a case is absolutely correct. As a statistician, you are only estimating the value that can be expected at some level of certainty (never 100%).

The idea of a simple linear regression model, how to determine the equation, as well as the principle of the least-squares criterion are explained in detail in the prescribed textbook. Make sure that you understand the following:

- We are only considering linear relations, meaning that we only fit straight lines through the data.
- There is a difference between an observed value of a variable y_i and the estimated value of the same value \hat{y}_i . The difference between these two is called the error.
- The regression line always passes through the point (\bar{X}, \bar{Y}) . Stated differently, the means of the two variables given as a pair are always a pair of coordinates on the regression line.
- In the general form of the regression line $\hat{Y} = b_0 + b_1X$, the b_0 and b_1 are only symbols and will be substituted by numbers in the calculated equation for a specific data set. In school you most probably used the form $Y = mX + c$ for the straight-line equation and learnt that m indicates the slope and c the Y -intercept. Compare what you learnt in school with this “new” form and you will see that the Y -intercept is given by the number without an x , namely b_0 , and that the value with the x , namely b_1 (also called the coefficient of X), represents the slope of the straight line.
- To calculate the least-squares regression line manually takes a lot of time and is quite tedious. Still, it is a necessary exercise for you at this stage. There will be enough time for you later in your life to use software and simply interpret printouts.

Make sure that you understand the meaning of the required assumptions for the linear regression model.

Activity 11.2

Question 1

Consider the following data values for variables x and y :

X	5	4	3	6	9	8	10
Y	7	8	10	5	2	3	1

The regression coefficients were calculated as $b_0 = 13.223$ and $b_1 = 1.257$.

Select the *correct* statement.

1. The relationship between x and y appears to be linear and positive.
2. The least-squares regression line is $y = 13.223 - 1.257x$.
3. The least-squares regression line is $y = 1.257 - 13.223x$.
4. If $x = 2$, the estimated value of y from the relevant regression line is -25.189 .
5. An x -value of 11 resulted in an estimated value of -1.861 .

11.3 INTRODUCTION TO CORRELATION ANALYSIS

Correlation analysis takes data analysis a little further. In regression analysis, the relationship between two interval or ratio scale variables is expressed in terms of a least-squares regression line, whereas correlation analysis can measure the strength and the nature of the relationship between the variables. We will discuss both the coefficient of correlation and the coefficient of determination.

The coefficient of correlation

You should know that the correlation coefficient r is such that

- $-1 \leq r \leq 1$
- if $r > 0$ (positive), the two variables will both either increase or decrease
- if $r < 0$ (negative), the one variable will increase when the other variable decreases

- the strength of the relationship depends on the actual value of r (if r is close to +1 or close to -1 , the relationship is strong)
- the strength of the relationship is weaker the closer the value of r (positive or negative) is to zero

As in the case of the regression line, a lot of calculations are needed if you want to determine the value of r manually.

The coefficient of determination

This is simply the value of the square of the correlation coefficient r , namely r^2 . The value of r^2 indicates the proportion of the variation in y , as explained by the regression line $Y = b_0 + b_1X$. That is all you have to know about this coefficient. (See the first paragraph under the heading “The coefficient of determination”.)

Activity 11.3

Question 1

The statements in this question are based on the following data:

X	Y
2.6	5.6
2.6	5.1
3.2	5.4
3.0	5.0
2.4	4.0
3.7	5.0
3.7	5.2
$\sum X = 21.2$	$\sum Y = 35.3$

The correlation coefficient r was calculated as 0.327. Identify the *incorrect* statement.

1. There is a positive relationship between x and y .
2. $\bar{y} = 5.043$
3. The coefficient of determination is 0.5719.
4. The regression coefficient b_1 is also positive.
5. Only 10.7% of the variation in y is explained by the variation in x .

Feedback on the activities

Activity 11.2

Question 1

Option 2

1. *Incorrect.* The relationship between x and y appears to be linear and negative (as the x -values are increasing, the y -values are decreasing).
2. *Correct.* The least-squares regression line is $y = 13.223 - 1.257x$.
3. *Incorrect.*
4. *Incorrect.* The estimated value of y is -25.189 only when $x = 5$ is substituted into the equation given in option 3. The correct answer is 10.709 .
5. *Incorrect.* An x -value of 12 resulted in an estimated value of -1.861 .

Activity 11.3

Question 1

1. *Correct.* $r > 0$
2. *Correct.* $\bar{y} = 5.043$
3. *Incorrect.* $r^2 = (0.327)^2 = 0.107$
4. *Correct.*
5. *Correct.* The value of r^2 is $(0.327)^2 = 0.107$. Then only 10.7% of the variation in y is explained by the variation in x .

11.4 SUMMARY

Once you have familiarised yourself with this study unit, you should be able to

- give detailed descriptions of the individual terms in the simple regression line
- interpret the value of the coefficient of correlation
- interpret the coefficient of determination