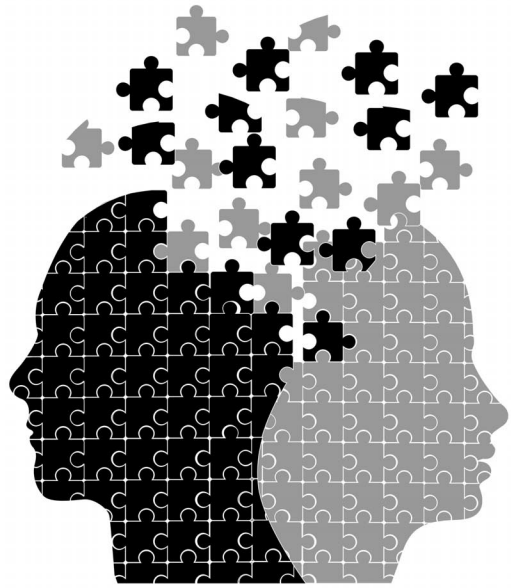
Psychological Research

Only study guide for PYC3704



P Kruger
HC Janeke

Department of Psychology
University of South Africa
Pretoria

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

PYC3704/1/2012–2018

98778870

3B2

PYC3-style

CONTENTS

APPENDICES

Introduction

Welcome to this module on psychological research. It is designed to introduce you to some of the techniques used by psychologists and other researchers in the social sciences to collect and investigate information and develop theories about human behaviour and mental processes.

Apart from occasional tutorial letters that you may receive, this study guide contains all the study material you will need for the module. There is no prescribed book.

How to approach this module

The key to completing this module successfully lies in understanding that the material presented here is different from much that you have encountered in psychology. You should, therefore, not use the same approach to studying this material as you would when, for example, studying social psychology, psychopathology or developmental psychology. A field such as developmental psychology is 'content-rich' and therefore requires that you become adept at techniques such as memorising and summarising large amounts of material, and understanding broad theoretical principles. By contrast, this module is 'conceptually-dense', that is, it is based on an intricate network of very precisely formulated arguments rather than on a large amount of content. Your approach should therefore not be one of memorising and summarising (although, of course, one always has to do some of this), but of making sure that you understand the arguments in detail. In a sense this is what we are trying to teach you here: how to think clearly and reason carefully.

What does this mean in practice? Simply this – you should work more slowly. It may be useful to read through a section quickly to get an overview of what it is about, but always go back and work through it more slowly, making sure that you have taken trouble to understand every step along the way. Each section in the study guide is in effect an unfolding argument, and you should try to see how it unfolds. Inevitably, there will be some parts of the argument that you may not quite understand. This does not mean that you have to get stuck – use your intuition to get an impression of what is meant and go on. But don't go on until you have made a good effort to understand the material. At the very least, you should be able to indicate what you don't understand.

How this study guide is organised

The study guide consists of six topics. Each of these has been broken down into a number of study units. Each study unit has been further divided into sections.

Dividing the material into topics, study units and sections are all ways of helping you to plan your studies sensibly. It is not possible to work through the entire study guide, or even one topic, in a single sitting, but you may be able to manage one study unit at a time. Many students find that even this is too much and they, therefore, do a few sections at a time so as to complete a study unit over a number of days. However much you are able to do at a time, don't make the mistake of leaving the work until just before the exam. You should rather work systematically through each of the six topics in the study guide so that you have covered everything in detail before the exam.

The topics are organised as follows:

- Topic 1 is a general introduction; meaning of various terms, how data is produced that can be analysed.
- Topic 2 introduces the notion of probability.
- Topic 3 explains how hypotheses are set up in such a way that we can use data to test them using the notion of probabilities.
- Topics 4–6 are applications of the techniques and procedures that were introduced in Topics 2–3, introducing you to some of the statistical procedures or tests that have been develop to test statistical hypotheses of various kinds. More specifically
  - Topic 4 presents a test to compare a sample mean with a specific population value;
  - Topic 5 presents tests for comparing group means from two samples; and
  - Topic 6 introduces correlation and shows how to test for relationships between two variables.

  A number of appendices are also provided. They contain a general example of data that we use for various examples throughout the course, as well as some background information. You will be referred to the appendices where they are relevant.

  Note that you are also expected to be familiar with the information in the appendices, and it will be assumed that you have studied it for assignments and exams!

Each topic ends with a set of exercises, multiple-choice questions and solutions.

These are very similar to the questions you will get in the exam, so, if you can do these, you are well on your way to success. Carefully read the feedback given for each of the questions. A useful technique in preparing for the exam is to try to formulate your own multiple-choice questions. Take each of the exercise questions and try to think up a similar question of your own – covering the same material, but asking a slightly different question about it.

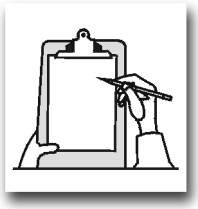
We suggest you do all the exercises and questions after completing a particular topic. Try to answer them first without looking up the answers. Do not mark the answers in the study guide, but on a separate sheet of paper, so you can test your knowledge later. In this way you will see which questions you can answer, and which ones you cannot answer. This will help you identify the difficult questions, and force you to rethink the reasons for the answers.

You should repeat this process when you study for the exam. Work through all the questions: the ones at the end of the topics and those that you encountered in your assignment questions. First try to answer the items without looking at the answer we provided (which is why you should not mark the correct answers in the study guides). Try and *understand* the reasons provided for the correct answers. It is hopeless to try and blindly memorise all of the questions.

If you really do not understand the information in the study guide, or the explanations given for the correct answers, make a note of the particular questions and concepts that you feel that you do not understand. You can then contact your lecturers to help you with this. It is much easier to help you if we know which particular concepts or questions give you difficulty.

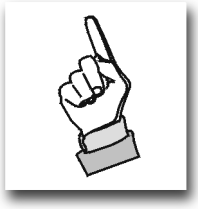
Icons

You will find that several different icons are used in the study guide. This is what they mean:



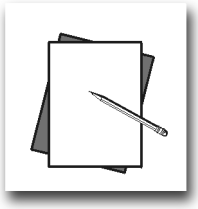
Icon 1: Summary of the important points of this topic

This is a list of key concepts for this topic or section of the work: it alerts you to the important issues or ideas that you should learn about in this section



Icon 2: Take note of this

This tells you to stop and consider some important or interesting additional information



Icon 3: Do an exercise

This indicates an exercise that you should do.

Icon 4: Beware of this

This alerts you to be aware of some potential pitfall or problem.

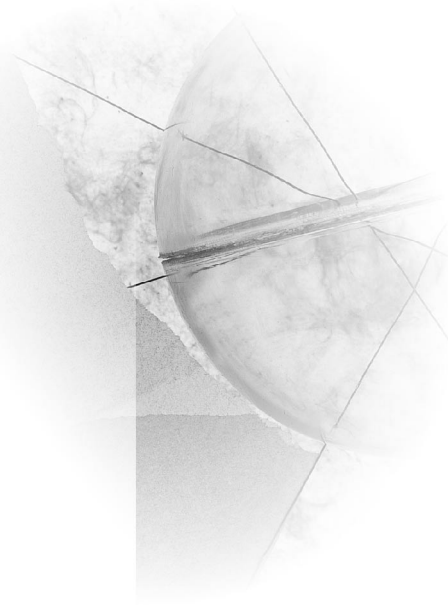
We hope that you will find the techniques used and reasoning processes involved in psychological research interesting and challenging, and that it will stimulate your interest in and enthusiasm for the more scientific aspect of psychology. It is only through research that a body of knowledge that can be applied in other more practice-oriented branches of psychology can be uncovered and evaluated.

We wish you a pleasant and rewarding year.

(Note: The names and contact details of your lecturers are given in the 101 tutorial letter for PYC3704)

TOPIC 1

Quantitative methods in research psychology



Quick overview

This topic serves as a general introduction to quantitative research in psychology. You are introduced to the way that constructs form explanatory concepts that are abstracted from observed behaviour, phenomena and events, and how they form the building blocks of theories. You learn how the constructs are turned into variables through measurement. We explain how data is collected by measuring a representative sample drawn from the population that is being studied, and how this necessarily leads to measurement error. You also learn how a research hypothesis is formed as a clear statement in terms of a relationship among the constructs (and the variables by which they are measured).

This topic is divided into the following study units:

- *Study unit 1.1:* Quantitative research in psychology
- *Study unit 1.2:* Constructs as the building blocks of theories
- *Study unit 1.3:* How constructs are made visible through measurement
- *Study unit 1.4:* Collecting information by sampling data
- *Study unit 1.5:* The research hypothesis

STUDY UNIT 1.1
Quantitative research in psychology

Psychology is a discipline that endeavours to collect information and develop theories about human behaviour and mental processes. The aim is to establish

facts that are related to psychological phenomena, that are valid and can be justified on scientific grounds. Everything that psychologists know about people about their behaviour, cognitive processes and emotional expressions comes from research of some kind. It is, therefore, necessary that you know something about how research is performed. In this course you are introduced to some of the techniques that have been developed and that are widely used in research, in social and other sciences.

All scientific knowledge begins with description of the phenomena being studied, based on careful observation. Knowledge based on observation of physical events is referred to as *empirical* knowledge (as distinct from knowledge based on contemplation, unexplained insights, mystical experiences or claims by authority figures). However, the act of simply observing phenomena and describing them or collecting facts about them is usually not sufficient. The next step in the scientific process is to go beyond the level of description by attempting to develop *explanations* for the things we observe: we want to know not only *what* the facts are, but also *why* they appear to be as they are. In other words, we want to develop *theories*, which explain why things are as they appear to be when we observe them. Of course, theoretical concerns are not the only reason why someone would do research. Sometimes a researcher is focused on more practical concerns: for example, whether a specific psychotherapy technique is effective. But even in these cases, a researcher would like to find out not only that it works better, but also *why*, and *what* the effectiveness (or lack of effectiveness) of the technique tells us about human behaviour, cognitive processes, experiences and emotions.

Powerful techniques have been developed over the years to formulate and test these theories. These techniques make it possible to disentangle the relationships among the many factors that may interact to produce complex phenomena. In this course we introduce some of these techniques, focusing on the use of *inferential statistics* in *quantitative research*.

Quantitative methods refer to situations where information (referred to as data) is available as *numbers*, which are the consequences of *measurements* of some kind. Statistics refer to the study of probabilities, which becomes relevant when we deal with data that are imperfect or incomplete, namely, data that contain *measurement error*. Techniques exist that can be used to reach decisions based on such imperfect data. An *inference* is a conclusion that follows from existing information, by generalising from the specific information to the general type of phenomenon, where the conclusion is not absolutely certain. So in summary inferential statistics are techniques for making generalisations based on imperfect numeric data, where the conclusions have a high probability of being true, but you can never be completely certain.

The techniques of inferential statistics that we show you in this study guide are widely used in various scientific disciplines, including the social sciences in general. It is necessary for you to know how they work, why they work and when they can be used. Without this knowledge much of the published literature in psychology and social science would be difficult or impossible to understand.

This course requires a basic familiarity with arithmetic and some very elementary algebra but it is *not* a course in mathematics! What is important to understand are the *principles* and the *reasoning* processes involved. Some use is made of mathematics, but the mathematical knowledge required is no more than what can be expected in high school arithmetic. For students who feel that their skills in interpreting formulas or equations and doing calculations are a bit rusty, we have included an appendix (Appendix E) to remind them of some of the rules of arithmetic. We also urge you to find a book on introductory arithmetic to sharpen your skills in using numbers and doing calculations.

The goal of this course is not only or even primarily to teach you specific statistical procedures, but is also (and more importantly) to develop your analytical skills: the ability to analyse the components of a research problem, develop a way to investigate it and draw appropriate conclusions by careful reasoning. One of the advantages of learning about how some forms of research actually work is that you learn to look at the research claims you read about – sometimes even in the popular media – with a more critical eye, and to take blanket statements like, 'it has been proven by science!' with a pinch of salt.

If you think of this course as an introduction to the scientific method and the development of your ability to analyse and solve problems in a careful and logical way, you may actually begin to enjoy it. We hope this will be your experience.

STUDY UNIT 1.2
Constructs as the building blocks of theories

1.2.1 Constructs

Psychologists try to develop explanations for human experiences and behaviour. To do this, they often have to make use of abstract concepts that serve as explanations for the behaviour they observe.

Look at the following list:

- *Anxiety*
- *Intelligence*
- *Introversion-extraversion*
- *Projection*
- *Achievement motivation*
- *Superego*
- *Job satisfaction*
- *Stress*
- *Self-esteem*

Imagine trying to see these things. What would they look like? How big are they? What colour are they? It is obvious that in general these words do not refer to physical objects, but have been abstracted out of our experience of human behaviour to serve as explanations for certain aspects of behaviour. Concepts

such as these are sometimes referred to as 'constructs'. They are in a sense 'made up' concepts that we use to explain things (like behavioural patterns) that we can observe, but cannot see in themselves (at least, not directly).

> Where do constructs come from? This is a bit like asking where words come from. Often, after studying a problem, a scientist may have a creative insight. Sometimes existing words are given a new meaning. Kurt Danziger writes about how the word 'emotion' was first used in something like our present understanding of the term by the eighteenth century philosopher David Hume (relating *feeling* to the idea of *inner motion*; the same origin as the concept of 'motives'), but only began to replace the older term 'passions' during the nineteenth century. Sometimes neologisms are invented by theorists. When R B Cattell was looking for a word for 'emotional flatness' (for Factor A of his Sixteen Personality Factor Test), he used an Italian word derived from art and called it 'sizia' (preparing a flat canvas for painting is referred to as 'sizing' it). Conversely, in natural science, when Newton was looking for a word to describe the 'pull' of the earth and planets, he used a word that was mostly used in a psychological context at that time: 'gravity' can also be used to describe a kind of emotional heaviness or seriousness.

1.2.2 Theories

Psychologists are interested to find out which constructs are important (in the sense of being required or useful to explain human behaviour) and how they work together in a pattern, or what their interrelationships are. One of the objectives of psychology is not only to describe human behaviour, but also to find explanations for it. Constructs and how they interact fill the role of explanatory mechanisms in psychology. We try to find out which constructs offer an appropriate explanation of the behaviour or events we perceive, and what the pattern of their interactions with other constructs may be. In this sense, it can be said that *constructs are the building blocks of theory*.

The way in which the word 'theory' is used in a scientific sense needs to be considered. You sometimes hear people say, when they refer to some idea (usually one that they do not agree with) that 'it is only a theory'. It is implied in this expression that a theory is only a guess. This represents a mistaken view of what the word 'theory' means to a scientist (in fact, this confuses the meaning of 'theory' with 'hypothesis', which is a kind of an intelligent guess, as is explained in more detail in Topic 3, later on in this study guide). In science, a theory is a framework for facts. It is some kind of description that tells you how the facts are connected, and why the facts are as they are (where the word 'facts' refers to things or events that were observed and described in a careful way). Constructs and their interrelations (how they affect each other, their patterns of interaction) are used in this way to develop theoretical explanations of why people behave in certain ways in certain contexts, or why mental phenomena appear to be as they are.

Some examples of constructs in psychology.

To explain why people sometimes act out of motives that they are unaware of, Freud suggested the concept of the unconscious. This is an example of a construct: something that (by definition) cannot be observed directly, but seems to explain those things that can be observed (in this example, what Freud observed was maladaptive and self-destructive behaviour). A person studying absenteeism and low morale at a company may conclude that what is at fault is the 'organisational climate' in the organisation, which then also serves as an explanatory construct.

Try thinking of a few more examples of these 'invisible forces' in the theories you have encountered in your psychology studies, which are used to explain some behaviour or expressions of people's moods or attitudes.

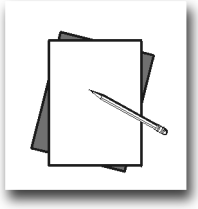
STUDY UNIT 1.3
How constructs are made visible through measurement

1.3.1 Measurement

Anything that can be differentiated in terms of *type* or *amount* can be measured. In quantitative psychological research we deal with constructs that are defined in such a way that they can be used to classify our observations into two or more categories, but often also so as to give an indication of the size or intensity of some phenomenon. Most of the constructs we deal with exist not only as a type (or *quality*) but also as a *quantity*. Quantification becomes relevant when we ask not only *what* the construct is that is in force in the situation, but also to *what extent* it appears – that is, *how much* of it exists, or to what *intensity* it exists.

In psychological research, quantification depends on our ability to *operationalise* the particular construct. This refers to that fact that we have to devise a systematic procedure or operation to make the construct visible, in such a way that we can measure it. Tying a construct to the way in which it is observed and measured in this way has the added advantage that different psychologists have a way of knowing that they are referring to the same phenomenon as long as they use the same measurement procedure.

'Measurement' refers to a process whereby numbers are allocated to something according to a rule. So what we need to do is to find specific standard procedures by which a specific construct can be observed, in such a way that a numeric value can be allocated.



Look at the following list of constructs. Try and imagine how you would go about measuring each of them. Try to think of suitable measurements before you look at our list of suggestions below.

- *Intelligence*
- *Anxiety*
- *Trustworthiness*
- *Psychotic behaviour*

Here are some possible ways in which each of these constructs can be measured:

Intelligence: a test of skills at solving certain problems; measured by counting the number of problems of a certain kind which a person can solve successfully;

Anxiety: a questionnaire asking self-report questions related to typical behaviour patterns that are known to be associated with high anxiety;

Trustworthiness: ratings by friends and acquaintances on a specific scale designed to measure this construct;

Psychotic behaviour: ratings by panel of clinical psychologists according to clinical criteria established relative to some diagnostic system, such as the Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV).

The link between observing a construct and measuring it is so close that when we talk about 'observation' in quantitative research, we often imply the process of measurement. The taking of a measurement is regarded as an act of observation. This is true even when we ask people to report on their feelings or attitudes by way of a questionnaire. For example, if we apply a questionnaire to measure people's attitude towards soccer, the construct would be 'attitude towards soccer' and the process of finding a score by means of the questionnaire would be regarded as the act of observation. The attitude scale (questionnaire) is the measurement device that makes this construct visible to the researcher, so the questionnaire and its application represent the operational definition or operationalisation of the attitude. It is the process by which the (invisible) construct is measured and, therefore, made visible, in the form of a number (quantity) on a particular scale.

Because of this link between measurement and construct, the procedure of operationalisation can also be said to provide an *operational definition* of a construct, as it can be seen as a practical demonstration of what the construct is. This idea has to be used carefully however. For example, letting someone do an Anxiety Scale test and obtain a score on it to measure the construct of 'anxiety' cannot be said to 'be' the anxiety. It is how the anxiety was made manifest, how it was made apparent (by subjects responding to a list of questions that refer to typical behaviour of people who are anxiety prone).

The subdiscipline that deals with psychological measurement is *Psychometrics*. This discipline is concerned with issues such as the *validity* of measurements (whether or not they actually measure what they claim to measure) and their *reliability* (whether the measurements are consistent). These things do not concern us in this course (except to remind us that psychometric measurement is difficult and often imprecise). We are more interested in what we can do with the

measurements once we have collected them. We usually assume that we are working with measurements where validity and reliability have been established to a satisfactory degree.

Something one should always be aware of when considering measurements is the *level* of the measurement. This has important implications for how the measurement can be used; for example, which arithmetic operations can be carried out and which ones are invalid. (See Appendix B for more on measurement levels.)

Note that we distinguish only between two kinds of measurement in this course. The first type is measurements where the *quantity* or *intensity* of some construct is considered (including both ratio and interval levels as defined in Appendix B), which we refer to as a quantity or a measurement of intensity. Secondly, we use measurements that indicate *category membership* (i.e. nominal scale measurements). It is however important to take the level of a measurement into account because it affects the type of calculations that can be made with a particular variable.

1.3.2 Variables

A construct that has been measured in some way produces a *variable*. A variable refers to a number that can take on any one of a range of possible values. They can be *discrete* (when only whole numbers like 1, 2, 3 are allowed) or *continuous* (what mathematicians refer to as 'real numbers'). In some cases variables also take on values smaller than zero to produce negative numbers. (See Appendix E if you feel your numeracy skills need brushing up.)

Variables can be contrasted with *constants*, which are numbers that can only take on a single size. *Example:* IQ tests are standardised in such a way that an average IQ is 100. As long as this practice remains, this number is constant. Another example is that if you measure a certain construct on a five-point scale that ranges from 1 to 5, the central point (and, therefore, the theoretical mean value of the scale, if the data is distributed in a fairly symmetrical way) is always at a constant value of 3.

So the (visible) variable reflects the intensity of the underlying (invisible) construct, in terms of how it was measured. We say that the variable is *manifest* (it is visible in the sense that we can observe it) and the construct is *latent* (it is invisible in the sense that we need some way to make it appear). So the latent construct is made manifest by the use of an appropriate measurement procedure.

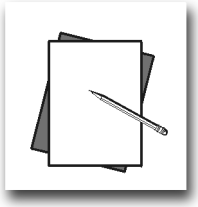
However, researchers are not only interested in the size or intensity of constructs (as reflected in variables) but also how they interact. This is because research is usually aimed at not only determining which constructs could be responsible for behaviour or mental processes, but also how these constructs interact to produce this specific set of events (the behaviours), which the researcher observes in the measurements. Note that since the constructs and their interactions become visible in the researchers' measurements, the interaction of constructs is often expressed in terms of the interactions of (two or more) variables.

When researchers refer to the interaction of specific variables (and you will come across such references often in technical articles, books and in this study guide), you have to keep in mind that when they refer to how variables interact, they are really interested in the interaction of the underlying constructs. The variables are just the way in which the constructs are made visible within the measurement procedure. We investigate the interactions among variables to infer something about the interactions among constructs.

One important distinction that we often need to make when we study the interaction among variables, is the distinction between the *dependent* and the *independent* variable. When a researcher focuses on the interaction of only two variables at a time, the dependent variable is usually the one that the researcher is interested in, the variable that is the focus of the research. The independent variable is something that the researcher manipulates, to see how this affects the dependent variable (in other words, the dependent variable is dependent on the independent variable). When there are only two variables involved, we often indicate the dependent variable with a 'y' and the independent variable with an 'x'. Keep in mind, however, that this is only a convention; using different symbols would be perfectly legitimate. The way that an independent variable affects a dependent variable can be expressed symbolically like this:

X $\longrightarrow$ **Y**

This notion can be extended to cases of interactions of more than two variables; for example, a set of independent variables can work together to produce a dependent variable, or there can be two sets of variables (dependent and independent) involved. Sophisticated techniques have been developed to analyse such complex interactions. However, in this course we focus mainly on the interactions between two variables at a time.



In each of the following statements of a research problem, identify which one is the *dependent* and which one the *independent* variable:

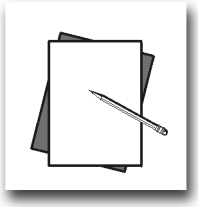
i Couples who are willing to openly discuss their feelings, have more stable marriages.
ii Men are more likely to be aggressive than women.
iii People who were abused as children are more likely to abuse their own children than people who were not abused as children.

Answers:

i Here, 'willingness to discussing feelings' is supposed to have an effect on 'stability of marriage'. Therefore 'stability of marriage' is the dependent variable and 'willingness to discussing feelings' is the independent variable.
ii In this statement, the claim is made that 'expression of aggression' is affected by 'gender', so *gender* is the independent variable on which the *expression of aggression* (the dependent variable) depends.
iii In this statement the claim is made that the predisposition to abuse one's

children (the dependent variable) is dependent on whether one was abused as a child (the independent variable).

Something else to keep in mind, even when we consider two variables only, is the possible effect of *hidden* variables. Hidden variables are effects on the dependent variable that we may be unaware of, or that we choose to ignore. Very often the events or behaviour that we observed are the consequence of many interacting factors, and we have to analyse the situation carefully to try and identify as many things as possible that may interfere with our ability to find a clear relationship between a dependent variable and some specific independent variable. One of these hidden effects that researchers in psychology often have to contend with is that people change their behaviour when they realise that someone is paying extra attention to them (usually referred to as the 'Hawthorne effect').



Imagine two groups of school children are compared on the construct 'mathematical ability' using an appropriate standardised test, one group coming from an up-market suburb in a big city, and the other group from a rural environment. Let us assume that, using the techniques of statistical hypothesis testing from this course, the researcher finds a difference exists in favour of the urban sample.

Can you think of any hidden variable – factors that may influence the children's mathematical ability – beyond the fact that they come from either an urban or a rural environment? Consider it for a moment, before reading our comment below.

Comment: What the researcher should consider is that a hidden variable may exist: the group of urban school children may have access to better quality education. Note that the difference that the statistical test shows could well be real, but the conclusion could be false, because the hidden variable 'quality of mathematics tuition' is not taken into account. It is not a false relationship, it is just that we cannot assume that the one variable affects the other one directly based purely on the fact that we found that relationship. Such hidden variables are sometimes also referred to as *nuisance variables*, because they interfere with the ability of the researcher to make sensible conclusions.

Keep in mind that techniques of statistical inference can show you that a difference (or other interaction effect) *exists*, but it cannot tell you *why* it exists. The reasoning is up to you. *The procedures do not replace the need for critical thinking.*

When we do research on the relationships among variables, a great deal of thinking and testing of alternative possibilities may be necessary. The statistical

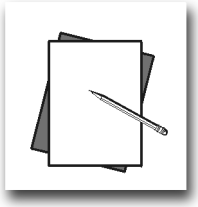
techniques can tell us that certain relationships exist, but which constructs are involved, and how they work together to form a pattern, and even which is the dependent and which the independent variable (in a given situation), depends on a critical consideration and testing of possibilities. The theories being investigated are often complex, and they are developed by abstracting suspected factors (the constructs that we postulate) from an even more complex situation. We explore the interactions of different sets of variables within different contexts for different groups or samples of persons, and often test our theories in a piecemeal fashion. Making sure that the variables you investigate are relevant and comprehensive depends partly on the theories being tested – which is why we prefer to begin with a theory of which variables are relevant and how they may be related, and not just throw in any variables we may think of. Making sure that your choice of variables makes sense is something that should be considered during the *design phase* of your research project (i.e. how the research programme is structured).

STUDY UNIT 1.4
Collecting information by sampling data

1.4.1 Data

When several measurements are collected from a number of people, the collected information is referred to as the *data* (while a single item of information is a *datum)*. Data are all the variables for all the cases in the research. These are often displayed in the form of a *spreadsheet* – with rows each representing a case (information from a single source, like a person or group of people or a specific object or event), and the variables (the set of measurements) arranged in columns.

Keep in mind that this arrangement is merely a convention: this is how most computer programs would arrange the data, but there is no intrinsic reason why the cases should not be put in columns and the variables in rows (as long as the computer or the human data analyst knows what the arrangement is).



Look at the table with data related to the AIDS scenario in Appendix A. This is typical of how data would be presented in a spreadsheet. You cannot really make much sense of it just by looking at it. Just looking at one or more pages of numbers tells you very little. So the problem is how to make *sense* of it, how to use it to form conclusions about the underlying constructs that were measured to produce these numbers, and the possible patterns that exist among them.

1.4.2 Descriptive statistics versus inferential statistics

A distinction exists between *inferential statistics* and *descriptive statistics*. The

second category refers to a set of quantities used to *summarise* aspects of numerical data. Examples that you may be familiar with are means, range, variance and standard deviation (see Appendix C for a quick introduction). These summary quantities are sometimes referred to as *parameters* (when they refer to the whole collection or population of data; see section 1.4.3 below).

Inferential statistics refers to the use of statistical techniques to make generalisations about the relationships among (two or more) variables. Here the patterns that may exist in the data are carefully investigated. The major part of this course (PYC3704) is designed to introduce you to the appropriate way to use these inferential procedures. Statistical techniques or tests that make a number of assumptions about the nature and distributions of the descriptive statistics that can be calculated (the parameters; see section 1.4.3 below) are referred to as *parametric statistics.* Here, inferential and descriptive statistics are usually closely linked, as the techniques for making inferences often make use of the summary values provided in descriptive statistics to explore the relationships among the variables (e.g. to compare the group means among different groups of people).

In fact, the statistical techniques we describe later on in this study guide (in Topics 4 to 6) mainly relate to testing statements or hypotheses about relationships related to descriptive statistics (the concept of a 'hypothesis' is explained in more detail below and in Topic 3). You should, however, be aware that there are statistical techniques that do *not* involve descriptive statistics as such (referred to as *non-parametric methods*). These techniques make use of other characteristics of the data than the summaries provided by the descriptive statistics. We consider one example of this in Topic 6: the chi-square test. Techniques such as these are often used when the measurement level is weak (ordinal or nominal data; see Appendix B), or when the sample size is too small to make meaningful use of more traditional methods, because the summary statistics cannot be trusted as reliable summaries of the data.

1.4.3 Populations and samples

The entire collection of cases that you are interested in when you make your measurements for a particular construct is referred to as the *population*. The population depends on which people or objects or events you are interested in studying. For example, if you are interested in exam stress among grade 12 learners in South Africa, your population would be *all* the grade 12 learners in South Africa who are required to write exams.

Note that if you define your cases in a certain way, some of the variables could become constants. For example, if your research refers to post-natal depression among women in the Western Cape, your population is all women in the Western Cape who have recently been pregnant, with both sex (female) and province of residence (Western Cape) taken as constants, even though they may serve as variables in some other research project.

Because populations can be very large, and we rarely have access to them, we would draw a *sample* of observations from the population and use that sample

to infer certain things about the population's characteristics. The most appropriate sample is usually a simple random sample, where each individual has the same chance of being included. If our samples are not random, they may lack external validity: it may not be possible to generalise beyond the group from which we drew the sample. For example, if all the grade 12 pupils we drew for our sample come from urban environments in Gauteng, we could not be sure the results could be accepted as valid for other pupils from other types of environment: we would be unsure whether the results could be generalised.

A short diversion on the issue of sampling

There are various types of samples that can be drawn – sometimes with the goal of deliberately bringing out certain characteristics, or to force the sample to reflect various categories of participants.

One of the most effective methods of sampling is ***random sampling***, which involves selecting a subset in such a way that each member of the population has an equal probability of being included in the sample. From a statistical point of view, a more satisfactory definition of random sampling is that it is a method of drawing a sample from a population in such a way that every possible sample of a particular size has the same probability of being selected.

At its simplest level, random sampling can be done by writing down the names of each of the individuals in a population on a small piece of paper, throwing all the pieces of paper into a large hat, shaking the hat and then, while blindfolded, drawing out only a few of the names from the hat. In practice, the method typically involves constructing tables of random numbers, and using these to select individuals in the population at random. However, the approach is not very efficient when dealing with large populations and in such cases alternative methods of sampling are typically employed such as ***systematic sampling*** (i.e. selecting individuals at fixed intervals), ***stratified sampling*** (i.e. dividing the population into homogeneous subgroups, and drawing random samples from the subgroups), or ***cluster sampling*** (i.e. sampling individuals from well-delineated areas – called 'clusters' – who have characteristics found in the rest of the population).

In practice, most research in social sciences cannot use samples that are truly random, simply because of the ethical requirement that participation should be voluntary, as people cannot be compelled to participate in research. Also, there may be situations where a researcher has no choice but to make use of the research participants he or she can find, for financial or other reasons. We refer to this as a *convenience sample*. For example, if the research population consists of people who suffer from a very rare neurological disease, the researcher must wait for those people who present themselves, in as far as they are willing and able to participate in the research.

Descriptive statistics that refer to the population and those that refer to the sample are treated somewhat differently. A numeric characteristic that refers to a

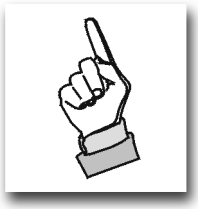
whole population is called a *parameter* and is often (but not always) signified by Greek letters, by convention. For example, the population mean is a parameter, and it is indicated by the Greek letter μ, pronounced 'muu'. The population standard deviation is indicated by the Greek letter σ, pronounced 'sigma'.

Population parameters like these are rarely known, since the only way to determine them would be to collect the relevant data from the entire population. However, there are occasions where we can know their values. For example, there is a standardised nine-point scale used by Psychometricians (referred to sometimes as a 'stanine' scale), where the scale is constructed in such a way that the scores are normally distributed with a mean of 5 and a standard deviation of 1.96 (the meaning of 'normally distributed' is explained in Topic 2). These can then be assumed to reflect the population mean and population standard deviation (μ and σ) respectively. In a similar way, IQ tests are usually constructed with a mean of 100 and a standard deviation of 15.

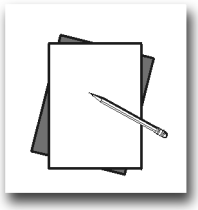
Numeric characteristics of a sample are usually referred to as *statistics*. These statistics are usually symbolised by Roman characters or by special characters, such as $\bar{\mathbf{x}}$ for the sample mean and **s** for the sample standard deviation. The character $\bar{x}$ can be referred to as 'x-bar'. (Note that because it can be tricky to produce this symbol on many word processors – and old typewriters – some books recommend the use of '**m**' for the mean. However, in this study guide we stick with convention, and use $\bar{\mathbf{x}}$ throughout).

When statistical testing is performed in inferential statistics, a researcher would measure a sample of people (or other entities) that was drawn from a particular population. If it is a parametric test, the relevant statistics would be calculated (from the sample data), and these statistics would be used to make inferences about the nature of the population. You must be very clear about this: we use the sample to represent the population, and do our calculations on the sample data, but ultimately we want to determine the situation in the population. To do this, we often have to estimate the (population) parameters by using the (sample) statistics. A researcher seldom knows the values of the population parameters, but the values of sample statistics can be calculated by means of clearly formulated mathematical procedures and these can be used as estimates of the parameters of the corresponding population. You learn more about how this is done in the topics that follow.

As we indicate below (in section 1.4.4), where we consider measurement error, there is always a certain degree of error in our measurements. A major source of this measurement error is due to the fact that we use samples to represent populations when we collect our data. While the sample represents the population, the representation is always imperfect. Random errors creep in. This is why we need to use techniques developed by statisticians based on the study of probabilities: probability theory gives us a way to deal with this kind of data , and it guides us to make decisions with a known probability of being valid, but not with absolute certainty. (*Note:* In Topic 2 we elaborate on this.)



One final thing to look out for is a problem with the terminology. A *statistic* is a sample measurement characteristic, as we have explained above, but you will also encounter the word in the sense of a *test statistic* in later topics (especially from Topic 4 onwards). A test statistic is the quantity you calculate (often by making use of sample statistics) to test a statistical hypothesis. You learn more about this in the topics that follow, but be aware of the possibility of confusion. When we refer to these test quantities, we always refer to the name in full – 'test statistic', and when we use the term 'statistic' on its own it refers to a descriptive statistic that describes an aspect of the sample data. Also the difference should usually be clear from the context.



Let us suppose you want to do research on *work stress of single mothers*. You want to know whether their stress levels have an influence on how strictly they discipline their children. Write down a description that defines the appropriate population in your study. Then look at our answer below. Also see if you can identify the dependent and independent variables in this study.

Answer. The population would be *all women who are not currently married, who have children and who have a job*. The fact that they have stress (or not) is not part of the definition of the population, since stress level is one of the variables you want to study. You should not only look at single working mothers who are stressed, but rather at single working mothers who have a range of levels of stress. *Stress level* is in fact the independent variable, with *strictness of discipline* as the dependent variable. This is because you want to know whether increased stress affects discipline, not so much the other way round.

So you can see from this that the population is the total number of that specific group that your research is about. It should be obvious that you could not observe all the people in the world who meet these criteria. Instead, you would draw a sample of persons who meet these criteria. You can then observe these sample members to make some inferences about conditions in the whole population (e.g. ask them to do some kind of test that measures 'work stress', which is the construct that you are interested in).

1.4.4 Measurement error

One of the consequences of using samples to represent populations is that this always leads to a certain degree of *measurement error*, no matter how rigorous our sampling procedure is. Another source of measurement error lies in the fact that our measurements are imprecise, that the measurement of a psychological construct is only more or less accurate. This measurement error is a kind of hidden variable, which we always presume to exist in social scientific research.

This is referred to as the *error component* or the *error term*. This is one of the major reasons for using statistical probability theory in our data analysis: we

assume that any variable we measure contains a 'true' element and an 'error' component. Furthermore, we assume that the mean of the error component is zero. We can do this because it is reasonable to assume that positive deviations and negative deviations from the perfect score (measurements that are too high or too low) will cancel each other out. We also need to make an additional assumption, namely, that these error terms are distributed around this mean of zero in a normal distribution (the notions of a 'distribution' in general and a 'normal distribution' in particular are explained in Topic 2).

If the observed measurement is indicated by $\mathbf{x_0}$, the 'true' measurement (the actual intensity of the construct that the measurement represents) by **x**, and **e** is used to indicate an error component, we can write the relationship symbolically as

$$x = x_0 + e$$

Note that the actual measurement we make is $\mathbf{x_0}$, and we use this to make a guess at what the 'true' measurement **x** would be, relying on the use of statistical methods to deal with the effect of the error component **e**. Since we assume the error component varies around a mean of zero according to a statistical distribution, we can also refer to the error term **e** as the *error variance* (and the term 'variance' is of course used in the sense of a description of the 'spread' of measurements, as discussed in Appendix C).

All of this is just an elaborate way of saying that when we measure a construct we always assume that our measurement is only more or less accurate, and that there is always an error hidden in these measurements, but we have developed ways (by the use of probability theory) to deal with it. (The use of probability theory and the relevance of the normal distribution is explained in Topic 2.)

STUDY UNIT 1.5
The research hypothesis

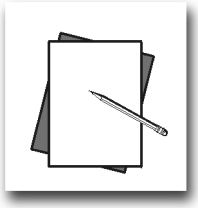
We now tie these concepts together by considering the cornerstone of the scientific method: the research hypothesis.

An hypothesis can be informally described as an educated guess. As we indicated above, research usually tries to establish relationships among constructs in order to develop a theory or to test an existing theory. Usually, the theory makes it possible for us to make some kind of prediction of how constructs should be interrelated. We formulate this relationship as an hypothesis, and we test the hypothesis (using statistical methods) to see if the prediction is true. If it is not true, there is something wrong with the theory, and we need to reconsider it.

We normally start with a research question. This could be an implication of a theory – something that seems to be implied by the theory or some kind of practical problem, which is stated in general terms.

Using our existing knowledge about plausible answers, we reformulate the

research question in terms of a conjecture or supposition, which has the goal of helping the researcher select what he or she has to observe in order to answer the research question. This is the research hypothesis (although there could be more than one), which expresses the problem in terms of very specific relationships among constructs that we expect to find (if our guess is true). It is important that this possible relationship should be clear and unambiguous. An hypothesis that is stated clearly and specifies exactly what is to be observed and what should be true if it is valid, is often called an *operational hypothesis*. However, this is just another name for a research hypothesis where the relationship between the measurements (representing the construct as variables) is written out in clear and explicit detail. You can think of the research hypothesis as a description of relationships that should hold among the constructs (two or more). The operational hypothesis is then the way the research hypothesis is expressed in the form of the relationships among the variables produced when the constructs are measured. But the operational hypothesis is usually taken as equivalent to the research hypothesis, so the distinction is rarely made in practice.



An example of a research hypothesis would be: 'motorists are more likely to express frustration at delays on the road on work days than on weekends'. Which are the constructs that may be at work here?

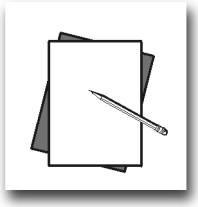
It looks as though there are two clear constructs: the amount of frustration that motorists show, and a dichotomous variable (see Appendix B if you are not sure what this means) dividing the time they are observed into 'work days' and 'weekends'.

If it is not possible to relate the research question to possible constructs, and to the variables that should be measured to represent the constructs, it may be that there is something wrong with the question. It is important that research hypotheses should be testable *in principle*. That is to say, it should at least be possible to say what kind of situation would exist if the hypothesis is true (or if it is false). Note that this does not imply that the test of the hypothesis should be obvious or even practical, only that it is *possible*. An hypothesis that can never be tested, no matter what you do, is not a scientific hypothesis.

An example of a question that is not really amenable to scientific research is the following. An Idealist philosopher may claim the whole world – the sum of all your experiences – is an illusion. This is not a claim that can ever be proven or refuted by finding empirical evidence, because no matter what evidence is provided, it can be said to be part of the illusion.

Be aware, however, that the fact that a question is not a scientific question does not imply that it is not important; it just means that no amount of empirical research will decide it. For example, the question of whether the death penalty is

morally acceptable is not a scientific question, since you cannot ever answer it by collecting data, and you cannot produce evidence that will prove it false. On the other hand, the claim that the death penalty serves as a deterrent to crime *is* a scientific question, since you can support or refute it by comparing societies that have rejected the death penalty with societies that enforce it. Whether only questions of a scientific nature have a place in psychology (at least in its role as a science) is a philosophical issue that we do not attempt to address here. Note, however, that the claim that humans act on moral grounds (or not) is clearly of importance to psychological theorising. This claim is, however, in itself an empirical question. Whether you can infer from what people *do* (the empirical question), what they *ought* to do (the moral dimension) is another matter.



An educational psychologist has the idea that students (at a residential university) who review their work every day after the relevant lectures will do better in exams than students who only look at their notes a few days before the exam, trying to study everything in one intense burst of concentration. This is based on the psychologist's own observations, but also seems to be in line with a theory on optimal memory strategies that she supports.

Try to consider this problem on your own. First, write out the research question as you understand it, then think about which constructs and variables are relevant. Then try to write out the research question that is implied as a research hypothesis. (We suggest you write out your thoughts on a separate piece of paper before you look at our attempt to answer it below.)

Answer: Here, the general statement that forms the research question could be: *How do those students who review their work regularly after the relevant lectures do in their exams relative to the students who do not do so, but prefer to study all their material in one intense burst of concentration? Does the first group do better than the second group in their exams, as the memory theory seems to suggest?*

This question now has to be reworked as a research (or operational) hypothesis. It seems obvious that the population could be all students at residential universities (in other words, those who receive specific lectures on specific days). A relevant sample will have to be drawn from these students. The two relevant constructs seem to be 'exam performance' and 'study method'. Regarded as measurements (i.e. variables), the exam mark (a percentage) could be used for the former, which would be a quantitative measurement (of ratio or at least interval level). 'Study method' could be represented as a nominal or categorical scale with only two categories, those that review their work after every lecture and those that study intensely before the exam. This information can be collected by questionnaire, personal interviews or observation (the first option is easiest, but may be less reliable). (Note that students who use both methods or who have

some other method of studying may exist. These students may either be dropped from the analysis, or put into an 'all the rest' category.)

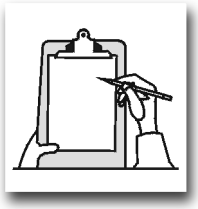
It may also be useful at this stage to consider the possible effect of hidden variables: do the same techniques work for all students or does the subject of study make a difference? In other words, do the same strategies work in the same way for – for example – a student of a text-based subject like history and a student who studies a formal, logic-based subject like mathematics? If it is likely to be relevant, the collection of this information should be specified at the beginning. It is impossible to find the data once the sample of research participants is no longer available.

Having considered the appropriate form of the data, we are now ready to rewrite the research question as a more formal research hypothesis. A possible formulation could be the following:

> *Students in the category of students who review their work regularly after the relevant lectures should generally score higher in their exams than students who fall in the category of those who prefer to study all their material in one intense burst of concentration before the exams.*

This statement clearly shows us which variables to consider (exam marks and study method) and how we expect them to be related (one group should generally get higher marks than the other). A bit of reflection will also suggest to us that a reasonable way to deal with this comparison would be to compare the two group means, as these are the appropriate parameters that show where the exam marks are centred for each group.

The next step in the research process is to turn the research hypothesis into a statistical hypothesis: a formal hypothesis that can be tested by statistical techniques. This subject is explained in Topic 3, but before we can proceed with that we need to learn more about probability theory. That is the subject matter of Topic 2.



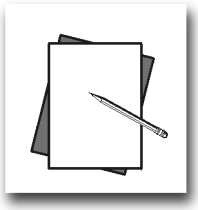
Summary of important points in this topic

After studying this topic, you should be familiar with the following terms and ideas, and how they are used in psychological research:

- *Constructs*: concepts that act as explanations for phenomena, events and behaviour and are abstracted from observations.
- *Theories:* a theory is a frame of reference for facts that attempts to account for why things are as they are; a claim about how constructs are related to produce phenomena, which has been validated by research.
- *Measurement:* allocating numbers according to a rule in order to quantify constructs. In order to measure something, a way must be found to *operationalise* it.
- *Variables:* constructs that were measured and are represented by numbers; the variable makes the *latent* construct *manifest* to the researcher. There can also be *hidden* variables of which the researcher may be unaware.
- *Measurement error:* a consequence of the fact that our measurements are

imperfect, and we use *samples* of data to represent *populations*. The variable we observe contains an unknown amount of measurement error.

- *Data*: collections of information, which could be quantities representing many measurements for many individuals or objects.
- *Descriptive statistics vs. inferential statistics:* using certain values to summarise data and making use of statistical techniques to make *inferences* or generalisations about them.
- *Populations and samples:* the whole group to which a set of measurements refers can be represented by using a smaller group to represent them.
- *A research hypothesis:* a statement about a possible relationship among constructs that may explain some set of observations that one intends to investigate.



Additional exercises

Exercise 1

Below is a list of questions from a biographical questionnaire. How many variables do you see and what construct does each one measure? What is the level of each measurement?

1. **What is your age in years?** ..... ☐

2. **What is the highest educational qualification you have completed?**

Primary school	☐	1
Some years in secondary school	☐	2
Completed secondary school	☐	3
Trade certificate	☐	4
Diploma or associate diploma	☐	5
Bachelor degree from an university or equivalent institution ...	☐	6
Postgraduate degree or diploma	☐	7

3. **What is your marital status?**

Never married	☐	1
Married	☐	2
Divorced	☐	3
Widowed	☐	4

4. **What is your employment status? (Mark one that applies best)**

Employed full time	☐	1

Employed part time		2
Self-employed		3
Unemployed		4
Retired		5
Student		6
Scholar		7
Full-time home duties/family responsibilities		8
Other		9

Discussion of Exercise 1

There are four variables, measuring the constructs (1) Age; (2) Highest level of education; (3) Marital status and (4) Employment status. 'Age' can be regarded as a quantitative measurement of *ratio* level, if we take zero as the point of birth. 'Highest level of education' can probably be regarded as an *ordinal level* measurement, since we cannot assume that the levels are an equal distance apart. It would, however, probably be a good idea to collapse the categories 'Trade certificate' and 'Diploma or associate diploma' into a single category, as it is not clear that one is higher than the other on a scale. The other two variables, 'Marital status' and 'Employment status' are nominal level measurements. For example there is no reason why 'Married', 'Divorced' and 'Never married' should be placed in any particular order.

Be careful not to regard the categories of a nominal scale variable like 'married' and 'never married' as if each one is a different variable.

Exercise 2

Imagine that you hear the claim that there are certain consequences that follow from the process of getting an education. In the process of getting an education, people are confronted with many new ideas and this affects the way they view society. For example, they are likely to become more aware of ideas such as personal freedom and human rights. The exposure to ideas such as these may influence their views on such issues as people's right to choose how they want to live and not be forced to do things against their will.

See if you can turn this scenario into a clear research question that can be investigated by quantitative methods. How will you measure these constructs?

Discussion of Exercise 2

What you need to do here is to see if you can find something in the scenario that could be clearly observed and measured. One obvious construct seems to be 'level of education'. Another may be 'views on individual freedom'. 'Confronting new ideas' does not seem to be a very promising choice for a construct.

So a possible research question could be, 'do people who have a higher level of education have stronger views on individual freedom?'

To measure level of education would probably not be difficult: a person's highest educational qualification could serve as a measurement (but see the previous exercise: you may have to consider what level this measurement represents). For 'views on individual freedom' you may have to find or create some kind of questionnaire. Some kind of rating scale that measures the intensity with which people express their belief in the importance of individual freedom may be feasible.

Multiple-choice questions and solutions

Questions:

1. Inferential statistics is a branch of statistics concerned with ...
    1. inferring numerical properties of sample data.
    2. inferring properties of samples from assumptions.
    3. estimating properties of populations from data.

2. Research in psychology is primarily about ...
    1. gathering facts.
    2. testing theories of human behaviour.
    3. making assumptions.

3. The main purpose of psychological research is to ...
    1. test theories empirically.
    2. select random samples.
    3. apply inferential statistics.

4. A theory is a network of ...
    1. relations among facts that were proven to be true.
    2. explanations for observed phenomena in terms of constructs.
    3. hypotheses that were observed.

5. Constructs are ...
    1. theoretical in nature.
    2. empirical in nature.
    3. identical to variables.

6. An item in a psychological test may be considered to be a ...
    1. manifest variable.
    2. test construct.
    3. both of the above.

7. Constructs such as anxiety can be ...
    1. observed directly.
    2. empirical in nature.
    3. defined in terms of behaviour.

8. Operational definitions of psychological constructs should define constructs ...
    1. in terms of observable behaviour.
    2. in terms of other constructs.
    3. through measurement.

9. Operational definitions enable us to ...

    1. bridge the gap between theory and observations.
    2. observe constructs.
    3. do both of the above.

10. The nature of an hypothesis refers to a rule that tells us ...

    1. which variables cause, or are associated with, which variable.
    2. which value of an independent variable is associated with which value of a dependent variable.
    3. the nature of the variables as constructs.

11. Suppose an hypothesis states that X causes Y. We cannot predict Y exactly because ...

    1. of unknown population or sample values.
    2. of other variables that also influence Y.
    3. we might not know the values of X.

Consider the following hypothesis and then answer items 12 and 13:

Hypothesis: The viewing of violent video material is related to aggressive behaviour.

12. The independent variable in the hypothesis above is ...

    1. viewing of violent video material.
    2. aggressive behaviour.
    3. There is no independent variable.

13. The hypothesis above ...

    1. associates values of the independent variable with values of the dependent variable.
    2. implies which levels of 'aggressive behaviour' are caused by viewing violent video material.
    3. is a theory about how violence on videos and aggressive behaviour are connected.

14. The observation that 'a child hits another child' can be considered to be ...

    1. an observation of a manifest variable.
    2. a behavioural consequence of 'aggressive behaviour'
    3. both of the above.

15. An hypothesis is general or universal in nature in the sense that ...

    1. the postulated relation is linear.
    2. the relation is proposed for a population.
    3. the postulated relation will hold in a specific sample.

16. A sample is ...

    1. a segment of a population
    2. a representation of the population
    3. random

17. An inference is ...

    1. another term for hypothesis
    2. an inspired guess
    3. a generalisation based on existing information

18. 'Operationalisation' refers to ...

    1. the process of forming an hypothesis
    2. the process of finding a practical way of measuring a construct
    3. finding a practical operation or procedure to do the research

19. To say that a construct is 'latent' is another way of saying it is ...

    1. hidden from direct observation
    2. abstract
    3. a concept that forms part of a theory

20. Parameters are ...

    1. another word for descriptive statistics
    2. values that indicate certain important aspects of the data obtained from a sample
    3. values that summarise aspects of population data

21. In psychological research, a theory is best understood as ...

    1. a list of constructs
    2. an educated guess
    3. an explanation of how facts are connected

22. To operationalise a construct means to ...

    1. explain the construct to a lay person
    2. classify the construct
    3. set up criteria so that the construct can be measured

23. The main difference between constructs and variables, is that constructs are (a) ... and variables are (b) ...

    1. (a) qualitative
       (b) quantitative
    2. (a) latent
       (b) manifest
    3. (a) constant
       (b) variable

24. A psychologist has a theory that anxiety influences the exam marks that students obtain in statistics. In this example exam performance is ...

    1 a latent variable
    2 a dependent variable
    3 an independent variable

25. In psychological research the term 'hidden variables' refers to ...

    1 all the latent constructs in a theory
    2 all the unknown variables that can exert an influence on the independent variables in an experiment

3 other variables that may also affect the dependent variable, but that the researcher does not explicitly focus on in a research study

---

Answers to multiple-choice questions

1. Option 3 is correct. Note that 1 and 2 are incorrect because we calculate statistics for samples and then make inferences concerning population parameters.
2. Option 2 is correct. Options 1 and 3 are not altogether wrong, but 2 is more correct. We do make assumptions in research, but this is not the essential goal of research.
3. Option 1 is correct. Again, although we do select samples (option 2) and apply inferential statistics (option 3), these are not as general an aim of research as is option 1.
4. Option 2 is correct. Option 1 would also have been correct if the word 'proved' had been replaced with 'proposed'. Option 3 is obviously incorrect as hypotheses are theoretical in nature.
5. Option 1 is correct. It is correct because we know that constructs are abstract concepts proposed by the researcher for scientific use. Option 2 is incorrect as 'empirical' means the opposite of theoretical. Although we indicated that we would use the terms 'constructs' and 'variables' interchangeably, these two terms are not identical. Strictly speaking, the latter term refers to constructs that have been measured.
6. Option 1 is correct. Option 2 makes no sense: constructs are usually abstract concepts that cannot be observed directly.
7. Option 3 is the more correct answer. Options 1 and 2 are incorrect because constructs cannot be observed directly and the word 'empirical' refers to things or events that can be observed.
8. The correct answer is option 1. Option 2 is incorrect because ''operational'' refers to practical procedures by which constructs are made visible. Constructs are often explained in relation to other constructs, but this is not an 'operational' definition. Option 3 is incorrect because, although the defining of constructs in terms of observable behaviour leads to measurement, it is not identical to the measurements.
9. Option 3 is correct because both 1 and 2 are correct. Operational definitions link constructs to observable phenomena so that we can observe them.
10. Option 2 is correct. Option 1 is incorrect because, although hypotheses are about variables, they are about the relation between the variables, not about causes. The nature or rule of the relation is about how the independent variable relates to the dependent variable. An hypothesis is a statement of relationships among variables, not about the nature of variables, so option 3 is not correct.
11. Option 2 is correct. There are many known and unknown (or hidden) variables that also influence Y.
12. The correct option is 1. The independent variable is that variable which affects the dependent variable; or, conversely, the dependent variable depends on the independent variable.
13. Option 1 is correct: as the values of the independent variable change, the dependent variable should also change (if the hypothesis is valid). Option 2 is incorrect because this particular hypothesis does not state that one variable 'causes' another.

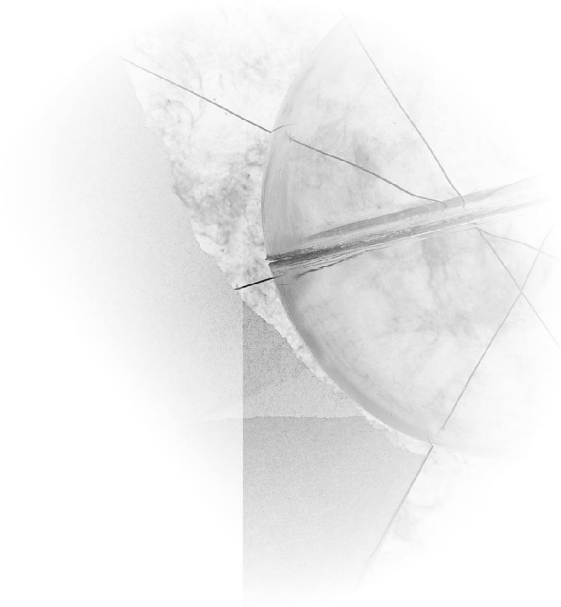
14. Option 3 is the correct answer. The observation that 'a child hits another child' can be considered to be a behavioural consequence or manifestation of the construct 'aggressive behaviour'. We infer that the behaviour is the consequence of the construct, and the construct is made visible (manifest) by the behaviour.
15. Option 2 is correct. Hypotheses that are true only for one or more samples are of no value in building theories that will allow us to understand and explain behaviour. Option 3 is incorrect as hypotheses are not stated for samples.
16. The correct option would be option 2. To say it is 'a segment of a population' as option 1 claims, seems to imply that a specific subgroup is deliberately chosen, and the whole idea is that the sample should represent the population as well as possible. Option 3 is not true of all samples. Random samples are desirable, but are not always possible or practical.
17. The correct option would be option 3. An inference is not the hypothesis, it is a conclusion based on information, where you state that something you have found has more general implications. It is not an hypothesis (as option 1 suggests), although you may use an inference to develop an hypothesis. For example, on the basis of the popularity of TV dramas that contain high levels of violence, you infer that people are entertained by watching violent behaviour. You can then turn this into an hypothesis for further study. 'An inspired guess', as stated in option 2 may be a good description of an hypothesis, but not for an inference, which is based on specific information.
18. Here the best answer would be 2. 'Operationalisation' is where you make the construct (which is usually an abstract concept, so it is difficult to observe it clearly) visible by finding some suitable way to measure it. You need it to be able to test an hypothesis, but it is not in itself 'the process of forming an hypothesis', as suggested in option 1. Option 3 'finding a practical operation or procedure to do the research' is a bit too vague, and would be more suitable as a definition of research design.
19. The correct answer is option 1. The fact that constructs are usually abstractions does play a role in making them latent, but it does not imply that 'abstract' and 'latent' mean the same thing, as suggested in option 2. Option 3 is true of constructs in general, but this is not related to the fact that they are not directly observable.
20. The correct answer is option 3. Options 1 and 2 are false because while the word 'parameters' does refer to descriptive statistics, it does not refer to all descriptive statistics. It is used only for those descriptive statistics that relate to the population, not to those that describe aspects of the sample.
21. Option 3 is correct (see section 1.2.2). Option 1 is incorrect because a theory is not just a list of constructs, but an account of how constructs are related to one another. Option 2 is incorrect because a theory is not just a guess, but a description of how observations fit together in an explanatory framework.
22. Option 3 is correct. The primary aim of operationalisation is to describe a construct clearly and unambiguously so that it can be measured and tested in a research study.
23. Option 2 is correct (see section 1.3.2). Option 1 is incorrect because variables and constants can both be qualitative or quantitative. Option 3 is incorrect because a construct is an abstract theoretical entity and not a constant (i.e. a numerical value that is constant).
24. Option 3 is correct. In the study the psychologist proposes that 'anxiety' has an effect on 'exam performance'. 'Anxiety' is, therefore, the independent

variable, and 'exam performance' is the variable that is actually being measured, the dependent variable. Option 1 is incorrect because exam performance is a manifest and not a latent variable.

25. Option 3 is correct (see section 1.3.2). Option 1 is incorrect because the term 'hidden variables' is used to denote other variables that could have been, but were not specifically controlled for in a research study, and does not to refer to all the possible constructs in a theory (there would be a very large number of such constructs). Option 2 is also incorrect, because although other variables (called covariates) could also influence the independent variables, the term 'hidden variable' specifically relates to variables that influence the measurement of the dependent variables.

TOPIC 2

Probability



Quick overview

In this topic you are introduced to the study of probability. Formal research is a process of testing hypotheses, and these hypotheses are evaluated by a form of reasoning that depends on probabilities. You learn how probabilities form a distribution of possible outcomes, and about the central role of the normal distribution. We show you how a measurement of any normally distributed variable can be transformed into its equivalent value relative to a standard normal distribution. The topic concludes with the notions of the sampling distribution and the central limit theorem. This theorem is of central importance in the development of statistical tests about the population means of sets of data.

This topic is divided into the following study units:

- *Study unit 2.1* Introduction to the study of probability
- *Study unit 2.2* Discrete probability distributions and the binomial distribution
- *Study unit 2.3* Continuous probabilities and the normal curve
- *Study unit 2.4* Sampling distributions and the central limit theorem

STUDY UNIT 2.1
Introduction to the study of probability

In this topic we present a somewhat intuitive introduction to probability theory. Since the theory actually involves a fair amount of mathematics, we do not explore its more technical aspects (did we hear a sigh of relief?). Nevertheless, it is important to develop at least a basic awareness of probabilities and

probability distributions, because these concepts play an important part in the interpretation of the results of psychological research studies.

Psychologists often want to formulate general statements about populations, and they want to interpret the effects of experimental conditions on criterion variables. They are not usually content to report merely that the arithmetic mean of an experimental group, subjected to a particular treatment, is higher or lower than that of a control group when tested on some variable X. They also want to make general statements such as 'The difference between the two groups is statistically significant, because the probability of obtaining these results by chance is very small.' Probability theory provides the logical basis for evaluating such statements, and for choosing between different interpretations of a statistical analysis. Moreover, in inferential statistics (the theme of this module) we make use of probability theory to infer the characteristics of a population on the basis of the characteristics of a sample.

Because uncertainty and probability are central concepts in psychological research, conducting and interpreting psychological research studies require at least a basic understanding of probability theory. We, therefore, explore the underlying theory briefly in this topic.

2.1.1 Defining probability

Probability can be studied in three ways. It can be approached in an a priori or **classical manner** in which the focus is purely on reasoning and mathematical deduction. It can also be studied in an **empirical** or **frequentist manner**, where probability is analysed in terms of the relative frequency of an event's occurrence by actual observations and conducting experiments. A third alternative is to think of probability in a purely **subjective manner,** as a degree of belief in something happening. Thus, a statement such as 'I think the South African cricket team has a 70% chance of beating Australia' expresses a subjective view, which is not really based on any calculation, but instead on gut feeling and past experience. This subjective approach is commonplace in everyday parlance and an interesting area to explore, but we do not delve deeper into it in this module. Our interest lies in the two technical, mathematical notions of probability.

2.1.1.1 The classical approach

The classical approach to probability theory has its origin in games of chance, and is used to help us estimate the likelihood of something happening based on reasoning alone. The approach works by analysing something happening in terms of all the possible outcomes associated with that something. The 'something happening' is called an **event**. In statistics an 'event' could be almost anything. The following are all examples of events:

- a coin landing heads up
- an ace selected from a deck of cards
- a red ball picked out of a box containing balls of various colours
- a score of 115 on a psychometric test of emotional intelligence
- a classification of 'female' for the next astronaut selected to go into space
- a Martian landing in a spaceship on the top of Table Mountain

In each of these examples, the event is a particular occurrence (e.g. the coin lands heads up) where various other events, called **outcomes** (i.e. results) are possible.

We obtain the probability of an event happening by dividing the occurrence of the event by the total number of possible outcomes. Alternatively stated, we determine the number of ways in which the event that we are interested in can occur, and divide this number by all the possible events (i.e. outcomes). The formula is as follows:

$$p(E) = \frac{\textit{Number of favourable events}}{\textit{Number of possible outcomes}}$$

where a **favourable event** is the specific event in which we are interested, and outcomes are the set of logically possible events relating to the particular problem. In the coin example there are two possible outcomes: the coin could fall either *heads* (one event) or *tails* (another event). These two events are the possible outcomes. In this formula the symbol **p** denotes probability, **E** represents the particular event of which we want to calculate the probability, and each of the different possible outcomes is assumed to be **equiprobable** (i.e. equally likely to occur).

An example will help to illustrate. Suppose we have a box that contains four balls, coloured red, blue, green and yellow. We denote the individual balls using the first letter of its colour (e.g. '**R**' is the red ball) as shown below:

R, B, G, Y

If we throw all the balls into a box, and randomly pick a ball, what is the chance that we shall pick a ***red*** ball? Here the favourable event is the red ball, and the number of possible outcomes is the four balls in the box. Applying the formula above we, therefore, have the following:

p (Red) = ¼ = 0.25

Let us try a slightly more difficult question. Suppose we pick two balls from the box. What are the chances that one is yellow and the other red? In other words, what are the chances of getting the following outcome?

Y, R

The first thing to do is to make a list of all the possible combinations of picking two balls, one at a time. We call this list of all the possible outcomes the **sample space**, and it is denoted by *S*. To determine the sample space, we start by combining R with B, then R with G, then R with Y and so on, and arrive at the following list of different pairs (we treat two combinations that contain the same colour balls, such as *R,Y* and *Y,R* as one pair; the order in which they are drawn does not matter in this case).

R , Y; R , B; R , G; Y , B; Y , G; B , G

The list of pairs above constitutes all the possible outcomes associated with this problem, and as we can see exactly six different pairs of coloured balls can be drawn from the box. The favourable outcome of red and yellow appears only

once in the list (*R,Y*). Hence, the probability of drawing a pair containing a yellow and a red ball is as follows:

$$\mathbf{p\ (R,Y) = \frac{1}{6} = 0.167}$$

The same approach can easily be applied to the other problems mentioned above (see p 28). In the coin tossing example mentioned above, we have the following sample space: $S = \{heads, tails\}$. We are interested in heads landing on top, therefore our probability is

$$\mathbf{p(heads) = ½}$$

Likewise, we can calculate the probability of drawing an ace by noting that there are four aces in a deck of 52 cards. The favourable events are, therefore, the four aces, and the sample space is the list of all 52 cards.

Thus, to determine the probability of drawing four aces from the deck we simply divide the four favourable events by the list of all the possible outcomes, yielding

$$p(ace) = \frac{No\ of\ aces}{No\ of\ cards\ in\ deck} = 4/52 = 1/13 = 0.0769$$

Observe that the numerical values for these probabilities derive from the nature of the particular games. When a coin is flipped there are *two* possible results (heads or tails) so that the probability of its landing heads up is 1/2 (1 out of 2). Similarly, there are four aces in a standard deck of 52 cards, so that the probability of drawing an ace is 4/52, which can then be simplified to 1/13, or 1 out of 13. The formula is based on the assumption that all the outcomes are equally probable (i.e. they have exactly the same chance of occurring), and that we know (i.e. can calculate) the number of different possible outcomes. If the outcomes are not all equal, the formula will not yield an accurate estimate of the probability of something occurring. The formula gives a theoretical definition of probability, and is often called the 'classical model of probability'.

2.1.1.2 The relative frequency approach

In general, the theoretical probability of an event occurring can be approximated by the **relative frequency**, or proportion of times that the event occurs. Hence

$$p(E) = \frac{number\ of\ observations\ of\ E}{number\ of\ times\ the\ experiment\ was\ performed} = \frac{f(E)}{n}$$

where f denotes frequency, n the number of times the experiment is performed and f(E) the frequency of the events.

$$For\ example:\ \frac{N\ of\ heads}{1\ million\ flips\ of\ coin} = \frac{1}{2} = \frac{500\ 000}{1000\ 000} = 0.5$$

We now consider some important points related to this 'relative frequency' definition of probability:

- The above formula represents the probability of an event occurring, given a particular **statistical experiment** (which is also called a **random experiment)**, and the probability is, therefore, estimated on the basis of the results of the statistical experiment.

Please note that in statistics, the term '**experiment'** is used rather broadly to refer to a process of observation and measurement. Thus, an experiment may refer to simple tasks such as counting the number of times a coin lands heads up, or identifying the proportion of married individuals in a group of adults. It may also relate to more complicated processes such as obtaining and evaluating data to predict trends in the economy, or determining the relative probability of different possible causes of a disease.

- The outcome of a statistical experiment is a **random variable**. It is called random because its value cannot be predicted and is known only after the experiment has been performed. For example, we do not know beforehand whether a coin will fall with head or tails on top – the experimental process can randomly generate either a head or tail as result.

2.1.2 Some basic terminology related to the theory of probabilities

All the possible outcomes of a statistical experiment are called the **sample space** (which we also call a **population**) of the experiment. For the coin-flipping experiment we have a sample space of two events:

$S = \{heads, tails\}$ because there are only two possible outcomes if a coin is flipped once.

If a coin is flipped twice, we have four possible outcomes, and the sample space therefore becomes

$S = \{(heads, heads); (heads, tails); (tails, heads); (tails, tails)\}$

If we throw a six-sided die and want to determine which of the sides is on top when the die falls, there are six possible outcomes, and the sample space is

$S = \{1; 2; 3; 4; 5; 6\}$

Let us now return to the concept of probability as a relative frequency. In terms of this approach, we determine the probability of an event by observing how frequently (i.e. the proportion of times) it occurs over the long run. Statisticians have formulated a general principle that can be used as a guideline when interpreting probability as a relative frequency.

The principle is called the **law of large numbers**, and it states the following:

> *If an experiment is done repeatedly, and if the outcomes are independent of one another, the observed proportion of favourable occurrences of an event will eventually approach its theoretical probability.*

What the law states is that a probability value should be seen as a theoretical limit on which the relative occurrence of an event (outcome) can be expected to converge over time **in the long run**. For example, in the above coin-flipping example, the probability of the coin coming up heads or tails on any flip is not influenced by the result of the previous flip. Each flip is independent of the other, and the theoretical probability of heads coming up remains the same, that is, p(heads) = 1/2 = 0.5. In terms of the law of large numbers, we can make the following prediction: If we flip the coin repeatedly, even though we do not know whether heads or tails will come up on any particular flip, the actual proportion

of heads will eventually get close to 0.5. Thus, as the experiment gets repeated over and over, the relative frequency or proportion of heads will approximate the theoretical probability of 0.5. You may try such an experiment yourself. Flip an ordinary coin 100 times and note the number of heads outcomes. It is unlikely that you will find that this number will be exactly 50. However, we can predict with reasonable confidence that the relative frequency of heads versus tails will be just about even if the coin is flipped a million times.

Some incorrect thinking about probability stems from a misunderstanding of the law of large numbers. For example, the **gambler's fallacy** is based on the assumption that if a certain event has not occurred in a number of trials, its probability of occurring in the next trail increases. Thus someone might notice that a coin has landed heads up seven times in a row, and incorrectly thinks that it is now time for it to land tails up. He or she might then start betting on tails coming up. Here the mistake lies in not realising that whether the event occurs or not, its probability is not altered because each flip of a coin is an independent event and the probability stays the same (i.e. 0.5).

The example below gives an interesting illustration of the role that the law of large numbers plays in statistical interpretation.

In ESP (extrasensory perception) research, psychologists make use of a Zanier deck of cards consisting of 100 cards, each showing one of five symbols: a square, a circle, a star, a plus, or a wavy line, as indicated below:

In a standard Zanier deck there are 100 cards, 20 for each symbol. Suppose Dr Venkman is conducting an ESP experiment. He shuffles the deck, holds up each card with the symbol pointing away from his subject (so that she cannot see what symbol is printed on the card), and asks her to guess what symbol is on the card. He repeats the same experiment 100 times. How many cards must she guess correctly before we can legitimately declare her clairvoyant? Of course, the answer is that she must consistently score higher than someone who does not have ESP. Therefore, the critical issue is the following:

> *How many cards out of 100 will someone who does not have ESP guess correctly?*

We know that the deck contains 20 cards with stars on them, so that the probability of the next card drawn from the deck being a star is

$$P(☆) = 20/100 = 0.20$$

The other four symbols have exactly the same probability of occurring (also 0.20), and because there are 100 cards in the deck, the subjects should guess correctly 20 times out of 100 cards just by chance. Hence, to be clairvoyant someone must guess more than 0.20 of the cards correctly. But how many more than 0.20?

In a famous study conducted in 1938 by Pratt and Woodruff, 32 participants completed altogether 60 000 trials with Zanier cards, and managed to make

12 489 correct guesses. This may not seem to be a big deal, because even if the subjects had no ESP ability, we would still expect them to get 0.2 x 60 000 = 12 000 cards correct just by chance. The participants in the experiment guessed correctly just slightly more than this, namely, 0.208 (12 489/60 000 = 0.208) of the times instead of the 0.20 that is to be expected.

Nevertheless, the odds against the observed result occurring by chance alone (from a statistical viewpoint) are greater than a million to one! The main reason why such a seemingly slight difference could be so statistically significant stems from the large sample size – the larger the sample size, the closer the observed frequency can be expected to be to the true probability. In an experiment involving 60 000 trials we expect a relative frequency very close to the true probability, and even a slight difference (the 0.008 in this experiment; i.e. 0.208 – 0.200 = 0.008) from this theoretical probability becomes very significant. (See Topic 3, section 3.3.3 on the issue of 'effect size'.)

2.1.3 A few useful facts about probabilities

The probability value tells us at a glance how frequent or infrequent the event is, and what the likelihood is of obtaining a favourable outcome associated with it. In the case of a game of chance such as playing roulette, or throwing a die, a calculation of probabilities will give us information about our chances of success.

- Probabilities can be expressed as percentages (e.g. a 10% probability), as fractions (e.g. a 1/10 probability), or as a decimals (e.g. a 0.10 probability). All these uses are quite commonplace, but in psychological research probabilities are typically written in decimal format. This is mainly because probability values in decimal format can easily be compared. We know that 0.008 is less than 0.009, whereas the difference between two fractions such as 13/27 and 14/31 is more difficult to evaluate.
- A probability value represents a **proportion** (i.e. the proportion of outcomes supporting the event). A proportion is a decimal number between 0 and 1 and indicates the fraction of the total. In the example above, 0.20 represents the proportion of times (out of a maximum of 1.0) that the star symbol can be selected by chance.
- In the rest of this module, we often refer to the probability of an event (or statistic) as its **p-value**. So you will find statements such as 'the p-value of x is 0.02', where x would be some or other statistic (e.g. the t-statistic) that you encounter later in this module. Stating that a p-value is 0.02 means that there is a 0.02 probability that the particular result (value of x) can occur by chance.
- When decimal notation is used to describe probabilities, they fall in a range between 0 and 1, with values closer to 1 indicating a greater likelihood (or chance of success) than values close to zero.
- Because probabilities fall in a range from 0.0 to 1.0 when expressed decimally, a probability can never be higher than 1 or lower than 0. The general rule is written symbolically as follows: $0 \leqslant p \leqslant 1$. Note that a probability can be 0, but to say that a probability is 0 is actually the same as

saying that the event is impossible and can never happen. Likewise, to say that the probability of an event is 1 is to assert that it is an absolute certainty. In actual practice, probabilities fall within these two extremes.

- You will typically encounter reference to probabilities in expressions such as ''p > 0.05''. This statement is interpreted as ''the probability value is higher than 0.05''.

  NB: Make sure that you know how to evaluate the operators '>' (bigger than), '<' (smaller than), '$\leqslant$' (smaller or equal to), '$\geqslant$ (bigger or equal to), because these are used when the probability values associated with a statistic are given. Refer to the mathematical revision in Appendix E if you are unsure about the appropriate interpretation and usage of these operators.
- The probability of an event *not* happening is (1 – p(E)). For example, in the coloured ball example higher up, the probability of drawing the pair R,Y was 1/6 so that there were 5/6 chances of not drawing this pair. Note, therefore, that p(R,Y) = 1/6 and p(Not R,Y) = 1 – 1/6 = 5/6.
- The sum of the probabilities of all simple events in S (sample space) equals 1. This characteristic follows from two facts, namely, that (a) the sample space lists all the possible outcomes associated with a given statistical experiment, and (b) when all the outcomes are summed together we have the maximum possible probability, which is 1. For example, in the case of the Zanier experiment mentioned previously there are five possible outcomes: S= {star, square, circle, plus-sign, wavy-line}. Each of these possible outcomes has an equal probability of 0.20 of occurring, and thus when summed together we have

  p(star) + p(square) + p(circle) + p(plus) + p(wavy-line)
  = 0.20 + 0.20 + 0.20 + 0.20 + 0.20 = 1.

2.1.4 Rules for combining probabilities

Let us first introduce some more terminology.

Two events are said to be **independent** if the occurrence of one has no effect on the probability of the other occurring. In the coin flipping example, the probability of the coin landing on its head each time is independent of the result of the previous flip (i.e. the coin has no memory).

Two events are said to be **mutually exclusive** if the occurrence of one precludes the occurrence of the other. For example, if a single coin is flipped, the events heads and tails are mutually exclusive. The coin can fall with either heads or tails up, but not both at the same time.

There are two important rules for combining probabilities, and these are both influenced by whether the events are dependent or independent.

2.1.4.1 The additive rule

The **additive rule** is **p(A or B) = p(A) + p(B**). This rule is used when two or more events are mutually exclusive. The additive rule is used to determine the sum of two or more probabilities, and is signalled by the use of the word 'or' (i.e. the probability of A **or** B).

An illustrative example: Jane and Magda take part in a competition in which one contestant wins a prize. There are 20 contestants altogether so that, everything else being equal, the probability of actually winning the competition is 1/20 = 0.05. What is the probability that either Jane or Magda will win? We obtain the answer by adding the two probabilities (using the formula above), so that we have: p(Jane wins or Magda wins) = 0.05 + 0.05 = 0.1. There is therefore a 0.1 probability that either Jane or Magda will win.

When dealing with events that are *not* mutually exclusive, a more general form of the additive rule is used, namely, **p(A or B) = p(A) + p(B) – p(A and B)**. The more general rule allows for the possibility that there may be an overlap between the probabilities – which is why p(A and B) must be subtracted as shown above.

For example, what is the probability of drawing either an ace or a heart from a well-shuffled, standard deck of 52 cards? First note that p(ace) = 4/52 = 1/13 and p(heart) = 13/52 = 1/4. However, we cannot just add these probabilities together because one of the 13 hearts is the ace of hearts. Therefore, adding 1/13 and 1/14 together will result in one of the cards being counted twice (first as one of the aces, and then as one of the hearts). To remedy the problem, we apply the general form of the additive rule given above, yielding the following:

$$\begin{aligned} \text{p(ace or heart)} &= \text{p(ace)} + \text{p(heart)} - \text{p(ace and heart)} \\ &= 1/13 + 1/4 - 1/52 \\ &= 0.0769 + 0.250 - 0.0192 \\ &= 0.3077 \end{aligned}$$

2.1.4.2 The multiplicative rule

The **multiplicative rule** states that **p(A and B) = p(A) x p(B**) where A and B are both independent events. This rule is used to determine the product of two or more probabilities and is indicated by the word 'and' (i.e. the probability of A **and** B).

Let us use the example given above, assume that there is more than one prize, and that the probability of winning any prize is 0.1. What is the probability that Jane and Magda will *both* win prizes? In this case we compute the product of the two probabilities, which gives us: p(Jane wins *and* Magda wins) = 0.1 × 0.1 = 0.01. This means that they have only a 0.01 probability of both winning a prize.

Example:

There is a saying that a monkey striking randomly at a keyboard will eventually produce the complete works of Shakespeare (sometimes referred to as the 'infinite monkey theorem'). What are the odds of a monkey producing a single word from Shakespeare? Let us take the word 'methinks', which has a good Shakespearian ring to it.

To make it simple, let us suppose our monkey has a simplified keyboard with only 30 letters on it: the alphabet (in capital letters only), the space bar, and a few punctuation marks. The odds of producing any letter at random is, therefore, 1/30. The monkey is hitting each key totally randomly, which

implies that the production of each letter is independent of any other. This means we can apply the multiplicative rule.

So for eight letters we get

$$p\ (\text{'METHINKS'}) = \frac{1}{30} \times \frac{1}{30} \times \frac{1}{30} \times \frac{1}{30} \times \frac{1}{30} \times \frac{1}{30} \times \frac{1}{30} \times \frac{1}{30}$$

$$= \frac{1}{30^8} = \frac{1}{656100000000} = \frac{1}{6.561 \times 10^{11}}$$

$$= 0.000\ 000\ 000\ 001\ 524\ 158$$

(which can be written as 1.524158×10^{-12}).

These odds are so small that if our monkey types at, on average, two letters a second, producing this one word by chance should happen once in about 10 000 years! It has been calculated that the chances of producing a complete play like 'Hamlet' in the time that has passed since the beginning of the universe to now would be extremely small.

In the formulation of the multiplicative rule given above we assume that the probabilities of the two events, A and B, are independent of one another. However, in some cases a particular probability is conditional on something else happening. For example, the probability of event A occurring may be conditional on the prior occurrence of event B. Conditional probabilities are written as **p(B|A)**, where **|** indicates that a condition applies. p(B|A) is read as 'the probability of B given A.' Likewise p(A|B) is read as 'the probability of A given B', or equivalently, as 'the probability of A happening on condition that B has occurred'.

The multiplicative rule that we use when we have conditional probabilities is

p(A and B) = p(A) x p(B|A) *(Formula 1)*

Suppose we let A denote 'Marie wins the race' and B|A stand for 'Marie gets a trophy given that she won the race'. We further assign a probability of 0.5 to A, and a probability of 0.6 to B|A. Therefore, the probability that Marie will win the race and get a trophy is

p(A and B) = (0.5) x (0.6) = 0.3.

Note that from the formula for conditional probability, using simple algebra, we can derive formula 2 below.

p(B|A) = p(A and B)/ p(A) *(Formula 2)*

Let us assume that we know that the chance of Marie winning the race and also a trophy is 0.3. We also know that the probability of winning the race is 0.6. What is the conditional probability of her winning a trophy provided she had won the race?

We use formula 2, insert the given probabilities and, therefore, have

p(B|A) = 0.3 /0.5 = 0.6

Conditional probabilities are used in an important rule called *Bays' rule*. See the box below, which gives a brief explanation of this rule.

Note: the information below is only intended for interest, you don't have to study, or know how to apply Bayes' rule in the examination.

Bayes' rule

Thomas Bayes was an eighteenth century mathematician who worked out a mathematical method for using prior probabilities to estimate the probability of future (posterior) events. The approach is based on the use of conditional probabilities and a simple (it gets more complicated) form of the formula is

$$p\,(A|B) = \frac{p(B|A) \times p(A)}{P(B|A) \times p(A) \times p(B|A') \times p(A')}$$

In this formula, we need to know the probability of B given A, and the probability of A. A' stands for the complement of A, which is given by p(1–A). Now, with this information we can actually determine the conditional probability of A given B, which makes the formula very useful. What Bayes' rule actually does is to give formal expression to something we all intuitively know to be true: that you can use probabilities calculated after examining the evidence from some observations made in the past to estimate probabilities of something happening in the future.

Bayes' rule and extensive further developments based on the idea, have led to an extremely important area in statistics called **Bayesian statistics.** Bayesian statistics are extensively used in statistics as well as in the cognitive, decision and computer fields (artificial intelligence), financial sciences, and even brain sciences. It is, however, a research and application area that is very mathematical in character, and it is, therefore, not really taught in most undergraduate psychology courses.

When are we required to consider events as being interdependent? The Lotto is an example of such a situation: if you draw one number, you cannot draw the same number again (unlike the previous example of a monkey hitting a keyboard at random, where striking a letter does not preclude the possibility of striking that same letter again). The fact that a number that was drawn in the Lotto cannot be drawn again affects the affect probability of the next number to appear in the sequence of numbers.

The implication of this is explained below.

Example:

What is the sample space of the Lotto draw? We do not go into this in detail, but deal with it conceptually.

First note that to win you need to have the correct six numbers. *These numbers may occur in any sequence.* This means that we are looking at a particular combination of numbers. The sample space consists of all

combinations of six numbers between the numbers 1 and 49. All we need to do is to calculate how many such combinations there are. Suppose it is 10 million, then any one particular combination has a 1 in 10 million chance of winning the lotto. But let's do the exact calculation of the sample space (as given in the following website: *http://www.math.mcmaster.ca/fred/Lotto/*):

There is a total of 13 983 816 different groups of six numbers that could be drawn from the set 11, 2, ..., 491. To see this we observe that there are 49 possibilities for the first number drawn, following which there are 48 possibilities for the second number, 47 for the third, 46 for the fourth, 45 for the fifth, and 44 for the sixth. If we multiply the numbers 49 x 48 x 47 x 46 x 45 x 44 we get 10 068 347 520. However, each possible group of six numbers (combination) can be drawn in different ways depending on which number in the group was drawn first, which was drawn second, and so on. There are 6 choices for the first, 5 for the second, 4 for the third, 3 for the fourth, 2 for the fifth, and 1 for the sixth. Multiply these numbers out to arrive at 6 x 5 x 4 x 3 x 2 x 1 = 720. We then need to divide 10 068 347 520 by 720 to arrive at the figure 13 983 816 as the number of different groups of six numbers (different picks). Since all numbers are assumed to be equally likely and since the probability of any number being drawn must be one, it follows that each pick of six numbers has a probability of 1 /(13 983 816) = 0.00000007151. This is roughly the same probability as obtaining 24 heads in succession when flipping a fair coin!

Suppose you play 10 distinct combinations. What are your chances of winning? *Answer*: 10 x 0.00000007151 = 0.0000007151. What would it cost you to play all the combinations? *Answer*: We saw above that there were 13 983 816 distinct combinations. Now, if you were to play them all at R2.50 a combination, you would pay 13 983 816 x 2.50 = R34 959 540 (about R35 million!).

STUDY UNIT 2.2
Discrete probability distributions and the binomial distribution

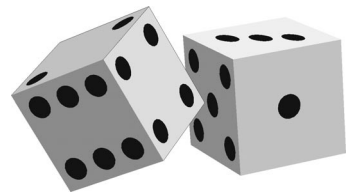
2.2.1 A probability model for discrete events: the even distribution

Consider the normal coin again. Provided the coin is unbiased, we know that the coin is as likely to fall on heads as on tails, and that only 'chance' determines which of the two sides will end up on top. We can use the following probability 'model' to predict the outcome of such a coin:

$$p(\text{heads}) = p(\text{tails}) = 1/2 = 0.5.$$

In statistics, a *model* can be a table with values, a computer program, a set of equations, or a formula. The important property of such a model is that it can take particular values as input and then generate an output. The model is, therefore, just a method or mechanism for calculating an answer. The model shown above is extremely simple, and simply states that the coin has an equal chance of falling on either heads or tails.

What would the probability model be for a normal die? We know that when we roll a die, the sample space (i.e. set of possible outcomes) is 1; 2; 3; 4; 5 and 6.



We assume that the outcomes are equally likely, and our model is, therefore,

$$p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6 \approx 0.167...$$

Since all the outcomes are equally probable, this is referred to as an even distribution and it can be represented graphically as shown in Figure 2.1. In this graph, the horizontal axis shows the possible outcomes, and the vertical axis shows the probability of each of these outcomes. This graph, therefore, represents the *probability distribution* of a six-sided die.

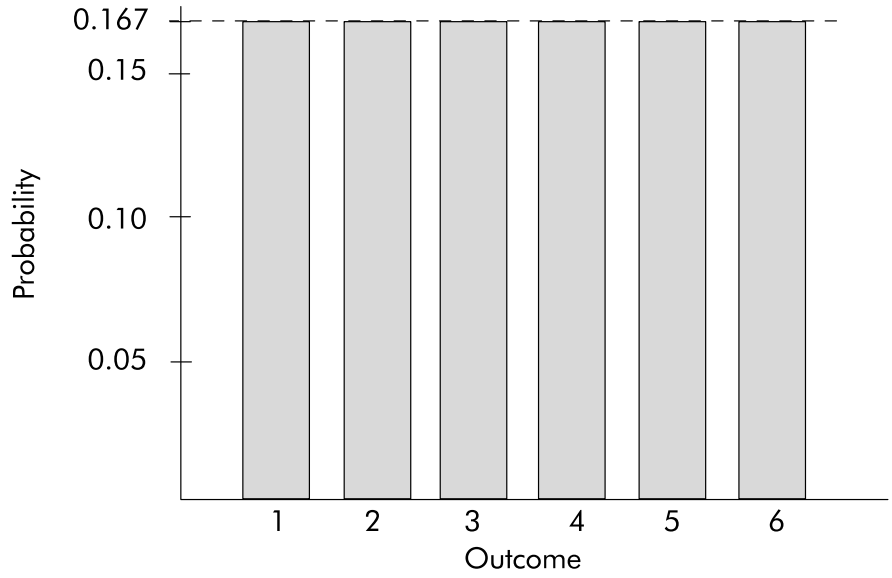


FIGURE 2.1: Even distribution of probabilities for a six-sided die

2.2.2 A probability distribution for multiple flips of a coin

Suppose we decide to flip a coin three times, noting the number of times 'heads' appears as the outcome, and repeat this experiment eight times. What prediction can we make about the number of heads that will come up? Note that this is a slightly different problem from those that we have considered so far, because here we are not just interested in the result of a single experiment, but in the repeated performance of the same statistical experiment for a specific number of times. Alternatively stated, up till this point we have only dealt with the issue of whether heads or tails comes up after each flip. Now we want to determine how many times heads comes up in series of three flips, and we are going to repeat this 'experimental trial' a number of times.

Try this yourself as an exercise. Flip a coin three times, repeat the experiment eight times, and note the number of times it lands with heads on top. Repeat this

process eight times (you are performing the experiment/research eight times) and enter your results in the following table:

TABLE 2.1: Counting the number of heads for eight experimental trials

Experiment	Number of heads
1	
2	
3	
4	
5	
6	
7	
8	

You will probably find that the repetition of the experiment yielded different results, you did not get exactly the same number of heads with each sequence of three flips, but a number (of heads) that ranged between 0 and 3. In the above exercise, the experiment was only performed eight times, but if we repeated the experiment a very large number of times (many millions of times) a distinct pattern of results will begin to emerge. What is this pattern and what predictions can we make, or model can we use to predict the number of heads for this type of experiment?

We can approach this problem by simply using logical and mathematical deduction instead of experimentation. Because the numbers 0, 1, 2, or 3 heads represent all the possible distinct outcomes, we can actually construct a table showing all the possible outcomes (eight) of tossing a fair coin three times as follows. This is represented in Table 2.2 below.

TABLE 2.2: All the possible outcomes for 3 tosses of an unbiased coin, repeated 8 times

Outcome No	First toss	Second toss	Third toss
1	Head	Head	Head
2	Head	Head	Tail
3	Head	Tail	Head
4	Head	Tail	Tail
5	Tail	Head	Head
6	Tail	Head	Tail
7	Tail	Tail	Head
8	Tail	Tail	Tail

From this table of all eight possible outcomes, we can summarise the information, and construct probabilities as shown in Table 2.3.

TABLE 2.3: Probability distribution for tossing a balanced coin 3 times, with the experiment repeated 8 times

No of heads	Outcome [H=heads; T=tails]	No of outcomes with that number of heads	Probability
0	*TTT*	1	1/8 = 0.125
1	*HTT; THT ; TTH*	3	3/8 = 0.375
2	*THH; HTH; HHT*	3	3/8 = 0.375
3	*HHH*	1	1/8 = 0.125

Table 2.3 is a ***probability distribution*** for this 3-toss coin experiment. Note that the table shows the probabilities associated with the four distinct logical possibilities that can occur (i.e. 0, 1, 2 or 3 heads). It shows that the highest or most likely probability is that we shall get 1 or 2 heads, and that it is less likely that we shall obtain 0 or 3 heads if we toss the coin 3 times.

The particular distribution that applies to the coin example is called a ***binomial distribution***. In a binomial distribution the random variable is discrete (a whole number or integer; see Appendix E). This distribution has been derived mathematically and, therefore, the formula for it is known. This makes it possible to work out the probabilities of specific outcomes without repeating complicated experiments a large number of times, as long as we know our variables are of a certain kind. We, therefore, take a closer look at this distribution.

2.2.3 The binomial distribution

The binomial distribution was introduced above as a probability model for the coin testing experiment. However, this distribution is not just applicable to a coin falling either heads or tails, but is a much more general distribution that applies to any statistical experiment with a random variable that generates two discrete outcomes. For example, the following can all be described using a binomial distribution:

- A diagnosis of patients as either diabetic or non-diabetic.
- Student answers classified as either correct or incorrect on a multiple-choice question.
- A participant in a parapsychology experiment guessing a card correctly or incorrectly.
- The gender of a newborn baby born in a hospital (i.e. either male or female).
- Psychology students passing or failing their statistics course.
- People watching a movie who either like it or don't like it.

More specifically, the binomial distribution applies in all cases where a random variable has the following properties:

- The random variable is for a sample that consists of a fixed number of experimental trials. Probabilities can be computed for trials of various lengths, but the length of the trials must be kept constant for the determination of each probability in the distribution.
- The random variable has only two mutually exclusive and collectively exhaustive events, typically labelled as 'success' and 'failure'. The terms 'success' and 'failure' apply to any outcome that has a binary character, such as 'yes' and 'no', 'hit' or 'miss', 'pass' or 'fail', '0' or '1', 'heads' or 'tails', 'correct' or 'incorrect'.
- The probability of an event being classified as a success, ***p***, and the probability of an event being classified as a failure, **1 – *p***, are both constant in all the experimental trials. This simply means that the probabilities cannot change during the trials. You cannot start a probability with one value (e.g. 0.5 for heads), and then later change this to 0.6.
- The event (success or failure) of any single experimental trial is independent of (i.e. not influenced by) the event of any other trial.

Using the binomial distribution avoids having to determine the probabilities using a list of all possible outcomes and applying the multiplication rule. In addition, this distribution does not require that the probability of success is 0.5, thereby allowing you to use the formula to determine a probability distribution for situations where the two outcomes do not have an even probability of occurring. Using the Binomial formula you could, for example, work out a probability distribution for the case where a coin is biased so that the chance of it falling heads up is 0.6, (and tails, therefore, 0.4).

Binomial distributions can be symmetrical or skewed. Whenever $p = 0.5$, the binomial distribution will be symmetrical regardless of how large or small the value of the sample size, n. However, when $p \neq 0.5$, the distribution will be skewed. If $p < 0.5$, the distribution will be positive or right-skewed; if $p > 0.5$, the distribution will be negative or left-skewed. The distribution will become more symmetrical as p gets closer to 0.5 and as the sample size, n, gets larger.

You typically determine binomial probabilities by using either a formula (see below), or a table of binomial probabilities, or statistical software (such as the Microsoft Excel function *BINOMDIST*).

The formula used to determine binomial probabilities looks quite intimidating, and is shown below.

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

Before seeing how to apply the formula, note that there is an operator in the formula that some of you may be unfamiliar with, namely '!'. The symbol '!' is called *factorial* and works as follows: n! = (n)(n–1)(n–2)...(1). Furthermore, 1! = 1, and 0! = 1. For example: $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$.

NB: *You will* **not** be asked to apply this formula in the examination, but you may have to do so in an assignment.

In the formula ***x*** denotes the probability of the number of 'successes' that we want to determine for the specific random variable, ***n*** stands for the number of trials, and ***p*** stands for the probability associated with any single, independent outcome of the variable.

For example, suppose we want to calculate the probability of only 1 head coming up in 10 tosses of a coin, then ***x*** = 1, ***n*** = 10, and ***p*** = 0.5 (assuming that the coin is not biased). If we insert these values into the equation we get:

$$\begin{aligned} \text{p(1 head)} &= 10!/1!(10–1)! \times .5^{1}(1–.5)^{10-1} \\ &= 10!/1(9!) \times .5^{1}(.5)^{9} \\ &= 10/1 \times .5 \times .001953 \text{ (rounded off to 6 decimal digits)} \\ &= 0.009765 \end{aligned}$$

(*Note*: $.5^{9}$ = is 0.5 to the power of 9; see Appendix E)

You should keep in mind that we don't have to calculate the full number when we are dividing 10! by 9!. We can use the information that 10! is exactly 10 times more than 9!. Factorial calculations can usually be simplified using cancellation as follows:

$$\frac{10!}{9!} = \frac{10 \times 9!}{9!} = 10$$

This follows because even though 9! = 9 × 8 × 7 × 6 × 5 × 4 × 3 × 2 × 1 is a huge number, when divided by itself it produces 1.

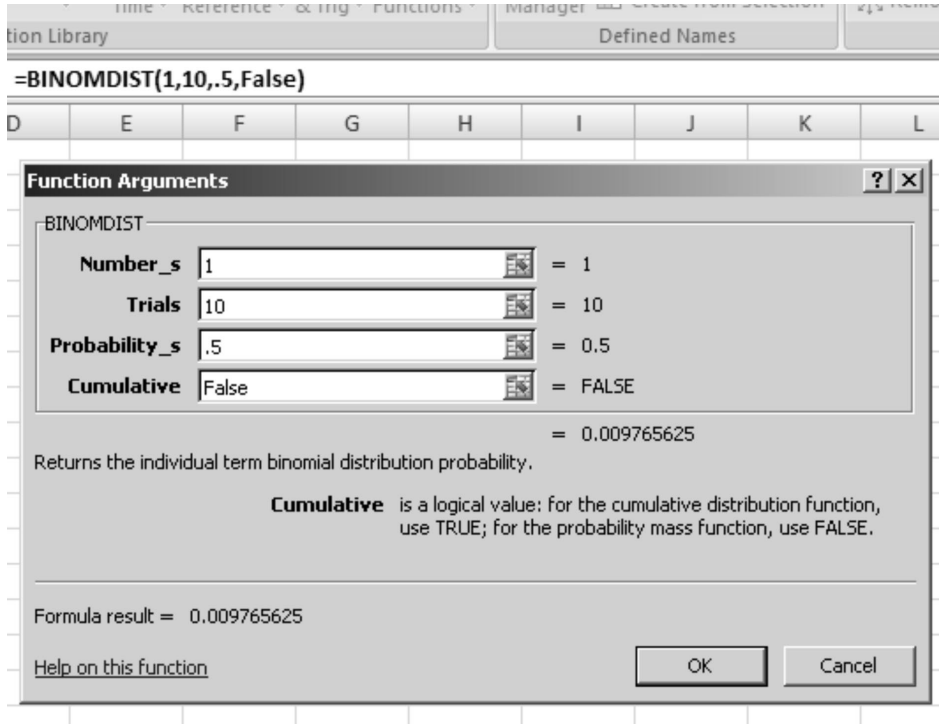
Example:

Because statistics often makes use of formulas, it is important that you should be able to insert values into a formula and work out the answer yourself. However, if you find it too difficult to apply this formula, and have *Microsoft Excel* on your computer, then you can use Excel to do the calculations for you. Open Microsoft Excel and then do the following selections: 'formulas' → 'more functions' → 'statistical functions' → 'BINOMDIST'.

In this function you have to fill in the following information:

'Number_s': The number of successes for which you want to determine a probability (insert 1);
'Trials': The number of trials (insert 10);
'Probability_s': The probability value for a single outcome (insert .5)
'Cumulative': (Insert 'False' if you insert 'True' here, the program will calculate cumulative probabilities; see section 2.3.1 below.)

Once you have inserted this information, Excel will immediately generate the probability value of 0.009765625. This is shown below:



As a further practical exercise, use either the formula or the Microsoft Excel function to determine the probabilities in the table below: The table represents the probability distribution associated with 10 flips of a coin. Thus, 11 experiments are performed (i.e. values of x from 0 to 10), and each experiment involves 10 trials. You must determine the probabilities for each of the 0 to 10 heads, and insert this into the column on the right.

TABLE 2.4 Probabilities of heads and tails coming up in 10 flips of a coin

Probability of 0 heads in 10 flips	
Probability of 1 heads in 10 flips	
Probability of 2 heads in 10 flips	
Probability of 3 heads in 10 flips	
Probability of 4 heads in 10 flips	
Probability of 5 heads in 10 flips	
Probability of 6 heads in 10 flips	
Probability of 7 heads in 10 flips	
Probability of 8 heads in 10 flips	
Probability of 9 heads in 10 flips	
Probability of 10 heads in 10 flips	

If you have worked out these probabilities correctly, you should have obtained

the values in Table 2.5 below (your answers might be slightly different if you rounded off the decimal digits differently):

TABLE 2.5 Probabilities of heads and tails coming up in 10 flips of a coin

Probability of 0 heads in 10 flips	0,0010
Probability of 1 heads in 10 flips	0,0098
Probability of 2 heads in 10 flips	0,0439
Probability of 3 heads in 10 flips	0,1172
Probability of 4 heads in 10 flips	0,2051
Probability of 5 heads in 10 flips	0,2461
Probability of 6 heads in 10 flips	0,2051
Probability of 7 heads in 10 flips	0,1172
Probability of 8 heads in 10 flips	0,0439
Probability of 9 heads in 10 flips	0,0098
Probability of 10 heads in 10 flips	0,0010

This distribution is represented visually in Figure 2.2 below.

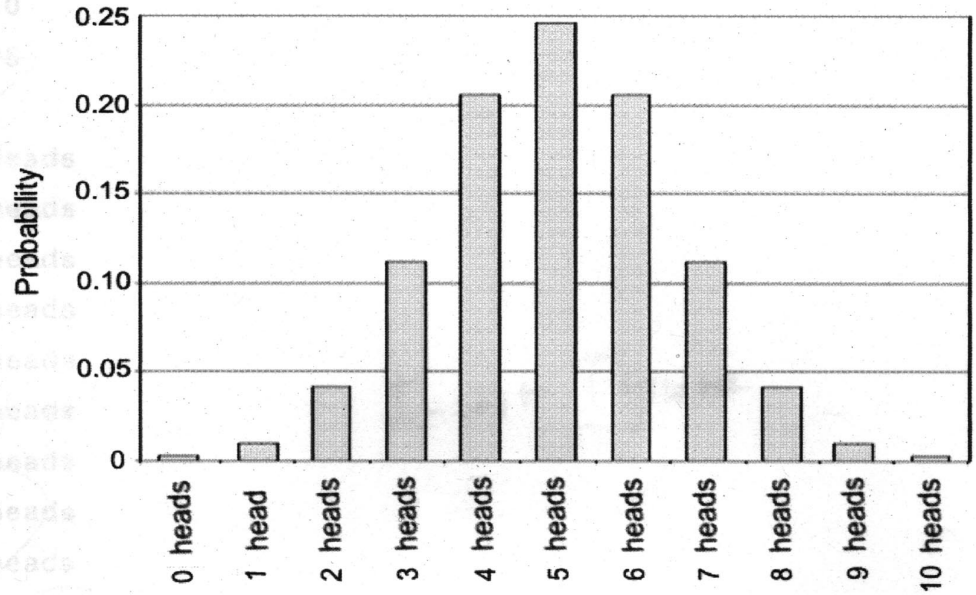


FIGURE 2.2: Probability distribution of number of heads for 10 flips of a coin

In Table 2.5, the probability of 0 heads in 10 flips is shown to be 0.0010. Thus there is a 0.001 probability of coming up with 0 heads in 10 flips (i.e., when all the outcomes are *tails*) should we repeat the experiment an infinite number of times. Let us for the moment imagine that by 'infinite' we mean one million repeated experiments, and that we actually performed that many experiments. The table predicts that we shall have obtained 0 heads in 10 flips of the coin in 0.0010 x 1 000 000 = 1 000 of the experiments (each experiment consists of 10 flips of the coin).

Question: What is the probability of coming up with four or less heads in 10 flips of a coin?

Answer: 0.2051 + 0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.377.

Note that the probability of 10 or less heads in 10 flips of the coin is 1 (since this range includes all possible outcomes). If the probability of 4 or less heads in 10 flips is 0.377, then the probability of 5 or more heads must be 1 – 0.377 = 0.623.

We can use the same table of probabilities whether the question is about heads or tails.

Question: What is the probability of getting 6 or more tails in 10 flips of the coin?

Answer: The same as the probability of getting 4 or less heads, namely 0.377.

STUDY UNIT 2.3
Continuous probabilities and the normal curve

In this study unit we explore the distribution of *continuous* variables. But in order to do that, we must first consider *ranges* of probabilities. These are referred to as *cumulative* probabilities.

2.3.1 Cumulative probabilities

Suppose a researcher studies the memory span of a small group of schoolchildren to determine whether this factor can help to account for individual differences in scholastic performance. She postulates that children with a longer memory span are able to retain more items in their memory at any moment, and are consequently able to learn more easily than children with shorter memory spans.

She conducts a test on a group of 36 children, which is done individually (i.e., one child at a time). The testing involves the presentation of a list of 15 words which she reads aloud, one at a time, to each child. After reading the words, she asks the child to recall as many of the words as he or she can, and makes a note of the items that each child was able to reproduce correctly. At the end of the study, she compiles the list of numbers shown below. Each number represents the total number of items correctly recalled by a particular child, yielding 36 totals, one for each child.

10; 10; 6; 7; 11; 4; 9; 5; 6; 7; 7; 8; 9; 8; 7; 4; 5; 3; 8; 2; 8; 9; 7; 10; 6; 11; 4; 9; 8; 12; 3; 5; 6; 7; 6; 5

Although this list contains 36 values, we note that, because some children remembered the same number of items, they obtained the same total scores. All in all, there are only 11 different totals.

Suppose we wish to determine the probability of the occurrence of each of these 11 outcomes. Using the formula that you learnt in study unit 2.1, we can do this for each of the 11 events by simply counting the frequency with which it occurred, and dividing by the total number of frequencies for all 11 outcomes. Proceeding in this way, we can compile a table, as shown below (Table 2.5), which presents the frequencies with which each of the 11 different totals occurred, relative to the total number of outcomes that are possible in principle, namely, 36.

TABLE 2.5: Relative frequencies of number of words remembered

x	2	3	4	5	6	7	8	9	10	11	12
f(x)	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

The values of these 11 totals range from 2 to 12, because no student managed to recall more than 12 correctly. Notice that the table sets out the probability of occurrence of each of these 11 totals, because it expresses the total's relative frequency as a fraction of the possible outcomes (i.e. 36). Furthermore, the table contains a **distribution** of the probabilities, because it tells us the probability of each of the possible totals in the range 2 to 12. If we regard the 36 children as a **population** (as if they were all the children in the world, and, therefore, all the possible participants in this experiment), such a distribution of probabilities is a listing of the probability of each particular outcome in the population.

For example, from the table one can see that the total with the highest probability is seven, because it occurs most frequently (6/36), whereas 2 and 12 are the least probable. There is only one chance out of 36 that a child selected at random from the group will only remember 2 items from the list, or as many as 12.

The information contained in the table is displayed graphically in the histogram in Figure 2.3.

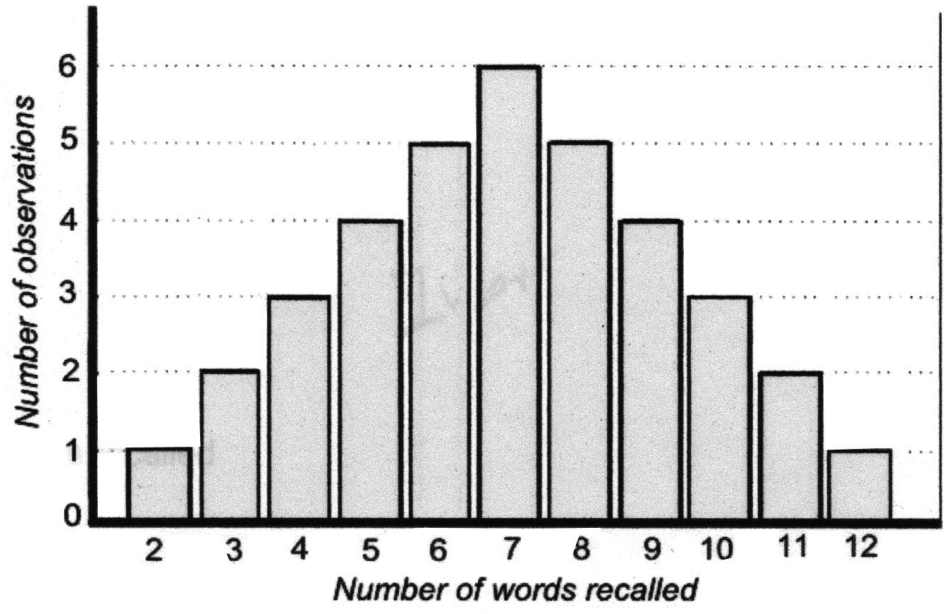


Figure 2.3: A histogram of the column frequencies of the table above

This histogram can be used to estimate the likelihood with which a child will recall a particular number of words. For instance, we can see immediately from the distribution of probabilities in the figure that the occurrence of a total of *seven* is more likely than the occurrence of a total of *four*. Notice also that the distribution has a symmetric shape, that is, it is higher in the centre and tapers off towards the sides.

We can use this distribution to calculate the probabilities of certain events, using a *relative frequency approach* (see section 2.1.1.2 above).

(*Note:* If you have difficulties with the numerical calculations, please consult Appendix E.)

We can, for example, calculate the probability of a child remembering nine words as:

$$\text{p(n words } = 9) = \frac{\text{Frequency at which 9 words were recalled}}{\text{Number of times the experiment was performed}} = \frac{4}{36} = 0.11$$

Using the same distribution, we can also calculate the probability of a child remembering nine words or more:

$$\text{p(n words } \geq 9) = \frac{\text{Frequency at which 9 or more words were recalled}}{\text{Number of times the experiment was performed}}$$

$$= \frac{4+3+2+1}{36} = \frac{10}{36} = 0.28$$

This value of 0.28 is known as a 'cumulative probability' and is interpreted as the probability that a child's total could be equal to or greater than 9.

Statisticians often equate 'cumulative probability' with the area under the curve (in a graph like the one above). We try to give you an intuitive feel for this, using the distribution above. Note that the graph consists of 36 little blocks as illustrated in Figure 2.4. (Each bar was divided into one or more blocks stacked on one another.)

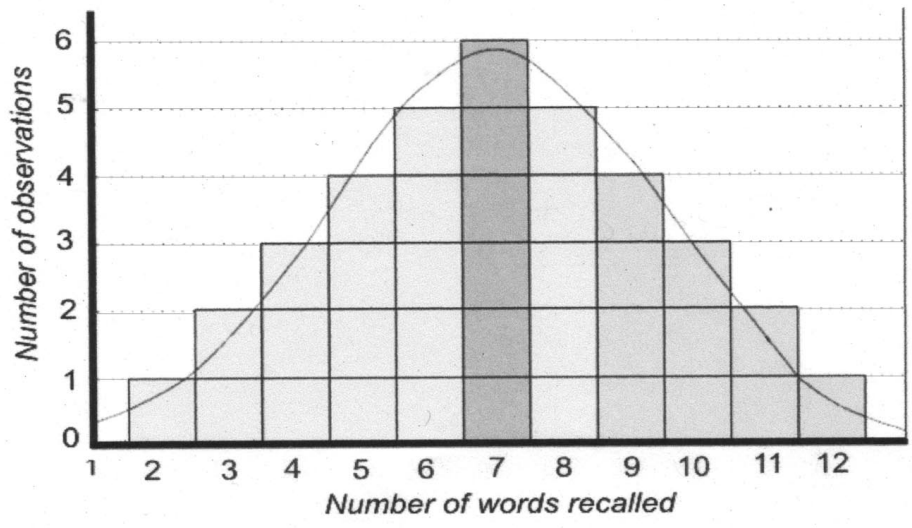


FIGURE 2.4: A probability distribution as an area under a curve

Let us set the surface of all 36 little blocks equal to 1 (because it represents 36/36 of the area). So the area of each little block is 1/36. The area associated with an outcome of *seven* is now six little blocks, or 6 x 1/36 = 6/36, which is exactly the probability of obtaining a total of seven. Clearly, the area under the curve when used in this way gives the probability of an event directly.

One more example: The area under the curve, to the right of the value 9 and including 9, is given by 10 little blocks, or simply 10/36, which is again the exact probability of a child obtaining a total of 9 or more.

The figure above shows that the probability distribution of the 11 totals in our example can be interpreted in terms of an area under a curve. As you can see from the figure above, we can draw a smooth curve that encloses all 11 totals. The figure illustrates that totals which occurred more frequently (totals 5 to 9) occupy a larger portion of the area under the curve than the totals which occurred less frequently (totals 2 to 4 and 10 to 12).

2.3.2 The normal curve

Up to now we have used examples of measurements that represent *discrete* measurements (e.g., the number of words remembered by children in an experiment). However, much of the data we have to deal with is measured on a *continuous* scale. If we want to consider measurements on a continuous scale, we have to take into account that there is an infinite number of points on such a scale, which in theory at least implies an infinite number of possible outcomes when we sample an event. Let us consider what this implies for the calculation of probabilities for a moment.

The distributions that we considered until now are based on **discrete variables**, that is, variables that take whole numbers (i.e. integers) as values. However, many naturally occurring variables such as age, weight and length are continuous. Continuous variables are **real numbers,** which can take on any value, within whatever limits its values may range between (see Appendix E).

For example, age is a continuous variable, because someone's age can be measured with an ever-increasing level of accuracy. Thus we can measure a particular person's age as 35 years, or 35 years and 6 months, or 35 years and 6 months and 5 days and 20 hours and 30 minutes and 0.15 seconds, and so on. Note, therefore, that a continuous variable allows for no gaps in the scale. The following example indicates how dealing with probability questions in the case of continuous variables can create special problems.

Suppose John is 20 years, 3 days, 2 minutes and 0.5 of a second old. Peter, on the other hand, is 80 years, 10 days, 5 minutes and 0.1 of a second old. There is an infinite number of other ages possible between John and Peter's ages. We can represent this as a continuous line, as follows:

John _____ **Peter**

Think of this solid line as consisting of an infinite number of dots – each representing a particular age.

So what is the sampling space of age? As a matter of fact, an infinite number of ages is possible, depending on how fine we want to measure it**.** Suppose we now want to determine the probability that a person could be exactly 30 years old. To do this we have to determine

$$p(Age = 30.0) = \frac{Number\ of\ people\ who\ are\ 30\ years\ old}{Number\ of\ ages\ a\ person\ could\ possibly\ be}$$

But as we saw above, thinking of age as a measurement on a continuous scale implies that the 'Number of ages a person could possibly be' is infinite. Dividing any number with an infinitely large number produces an answer of zero (you cannot literally divide with '∞' as it is not a number, but a concept).

It follows from this that the probability of obtaining any particular value from an infinite range of possible values is zero. Our usual probability formula does not work!

If we want to ask questions related to probabilities about objects that come from a sample space with infinite possible outcomes, we have to ask the questions in terms of a *range or interval* of possibilities (i.e., an interval on the continuous line).

For example, what is the probability that someone will be older than 20 but younger than 30? If we know how age is distributed in the population, we can use the notion of *cumulative probabilities* (introduced in section 2.3.1 above) to calculate the probability of obtaining an outcome that falls within a particular range. As you will see in our discussion of the normal curve that follows, this is equivalent to calculating areas under a continuous curve.

If you go back to look at the probability distribution in the graph for Figure 2.3, you will notice that the general form of this probability distribution is perfectly symmetrical: the right side of the distribution is a mirror image of the left side. Real data are rarely distributed in such a perfectly symmetrical form. Nevertheless, this type of symmetrical distribution is of considerable theoretical importance. A similar distribution of outcomes is often found when we make measurements on a continuous scale. In fact, research conducted by mathematicians and statisticians suggests that such a symmetrical distribution serves as a good abstract model for the distributions associated with a host of everyday variables.

Many of the scores that we use are also clustered around the average, and tail off to the ends of the distribution. Because it can be used to describe the distribution of many naturally or 'normally' occurring continuous variables, this type of symmetrical probability distribution is called a **normal distribution**. It is also commonly referred to as the **normal curve**, because the distribution can be plotted by a bell-shaped curve, as shown in Figure 2.5.

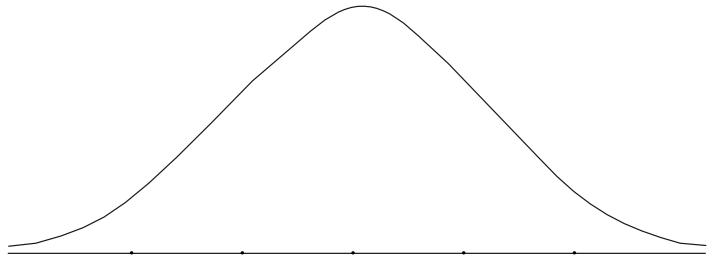


FIGURE 2.5: The normal curve

As the figure shows, the curve has a distinctly bell-shaped appearance, because it tapers off in a symmetrical way from the centre. The curve displayed in the figure above is somewhat idealised, because, for most data sets, a plot of the probability distribution will only approximate such a normal distribution, and the curve will not be nearly as smooth as that shown in the figure. However, for arbitrarily large populations, we would actually expect the data to fit a normal curve, and the greater the number of values from which the curve is drawn, the smoother, or more continuous, it will become. Incidentally, you will recall that this is what we expect as a result of the law of large numbers dealt with in study unit 2.1.

If the curve is based on a sufficiently large data set, it will eventually approach the theoretical probability distribution, that is, it will become a smooth, normal curve.

Many psychological and educational variables are distributed approximately normally, so that the normal curve can be used as a theoretical model for interpreting the distribution of these variables. The distributions relating to psychological variables such as measures of reading ability, introversion, job satisfaction and memory can all be plotted on a normal curve, and psychometric tests are often standardised in such a way that they conform to this distribution. Almost all the statistical tests discussed in this module assume normal distributions. Furthermore, many psychological measurements work very well even if the distribution is only approximately normally distributed. Some tests work well even with very wide deviations from normality. Also, apart from its theoretical significance, the normal distribution is useful because it is easy to work with in practice, and because many kinds of statistical tests can be derived for normal distributions.

An interesting demonstration of the normal distribution can be found on the Internet at

http://www.ms.uky.edu/~mai/java/stat/GaltonMachine.html

In the example on this website, balls are dropped from the top and then have to pass through a series of pins before falling to the bottom of the webpage. The pins serve as random events, pushing the balls into various different paths as they fall downwards. As the balls stack up at the bottom of the page, they eventually begin to form a distribution that gradually becomes increasingly normal in shape.

So far we have referred to the normal distribution as if it were a single distribution, but there is actually a whole family of distributions that have the same general shape. A number of different normal curves are shown in Figure 2.6.

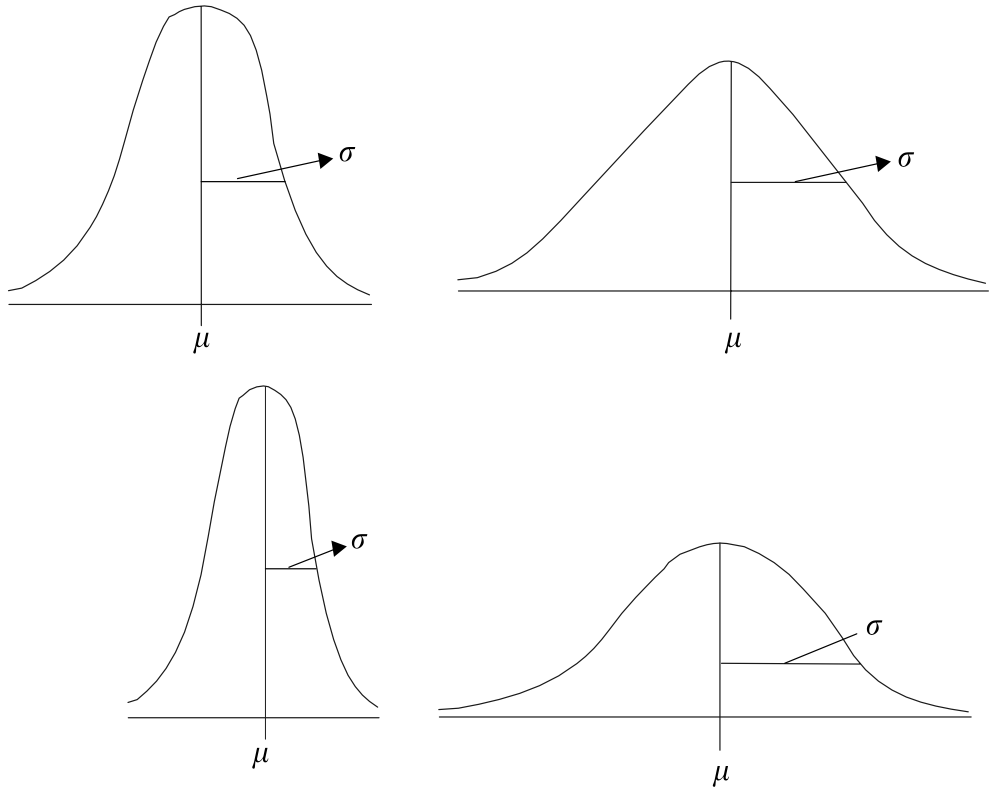


FIGURE 2.6: A few different normal curves

These distributions are all normal, because they are all symmetric with scores that are concentrated more in the middle than in the tails of the curve. Statisticians have derived a rather complicated-looking equation (or formula) which describes the normal curve, and have shown that it contains only two variables, the mean (μ) and the standard deviation (σ), with the rest of its terms being constants. The formula produces distributions that are all bell-shaped, but the actual shape of the curve – how high it is or how spread out it is – depends only on the mean and the standard deviation of the distribution concerned. They share a number of key properties, such as the following:

- They are bell-shaped. The most observations occur at the midpoint of the curve.
- They are symmetrical. The left side is a mirror image of the right side.
- They are continuous. Theoretically, the values which the variables can assume are infinite and are measured on a truly continuous scale so that the curve is smooth.
- Their curves are *asymptotic*, which means that the two tails never touch the horizontal axis, moving ever closer to infinity, because there is always some probability that more extreme values will occur.

2.3.3 The standard normal distribution

The different normal curves in Figure 2.6 differ solely because of differences in

their mean (μ) and standard deviation (σ). There is, however, one form of the normal distribution that is of special importance. This curve has a mean of $\mu = 0$ and a standard deviation of $\sigma = 1$ and is known as the **standard normal distribution**, and is by convention indicated with the letter '**z**' (so it is also referred to as the z-distribution). The measures on this distribution are referred to as **standard scores** or **z-scores**.

We have shown you that if we know how events are distributed, we can determine the probability that the event will occur. The standard normal distribution makes it possible for us to apply this knowledge to normal distributions in general. It provides us with a standardised scheme for interpreting a distribution of probabilities, as long as we know it is normal, or approximately normal.

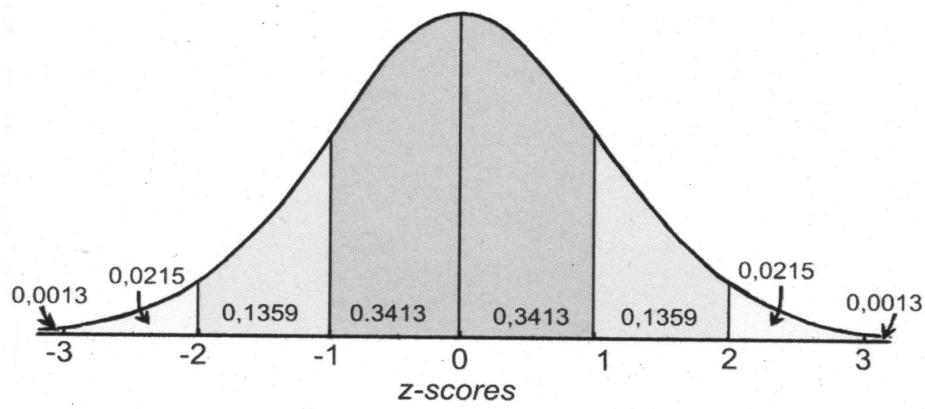


FIGURE 2.7: The standard normal distribution

The distribution of area under the standard normal curve is shown in Figure 2.7. The formula for the normal curve is known and portions of the area under the curve can be determined. In other words, areas that lie under the curve between different values of z-values can be calculated. However, here we are dealing with an interval under a *continuous* curve, and you would require knowledge of an advanced technique to do it (determining areas under a curve can usually be performed by using something called integral calculus, which is a technique for summing portions under a continuous function). In the case of the z-function, these calculations are usually published in tables, and you will find a table for the standard normal (z) distribution in Appendix D. You will find similar tables in most books on statistics.

Figure 2.7 shows the approximate proportions of scores distributed under the area covered by the curve.

- The total area under the curve gives the probability of the interval $-\infty$ and $+\infty$, and is equal to +1 (i.e., the probability of any value of z falling between minus and plus infinity is equal to 1).
- Because the distribution is symmetrical, 0.5 of the area lies to the left of the mean and the same proportion to the right of the mean.
- Approximately 0.341 of the area lies between the mean and 1 standard deviation in each direction.

- Roughly two-thirds, or 0.682 (0.341 x 2) of the area of the curve lies within one standard deviation of the mean.
- Approximately 0.477 (i.e. 0.3413 + 0.1359) of the area lies between the mean and 2 standard deviations in each direction.
- Approximately 0.954 (i.e. 0.477 x 2) of the area lies within 2 standard deviations from the mean.
- Approximately 0.998 (i.e. 0.954 + (0.0215 x 2)) of the area lies within three standard deviations from the mean.

As we saw in section 2.3.1, the area under the curve of a distribution of probabilities translates directly into a statement about probabilities, so you can use the tables to read off the probabilities that z-values will fall within certain ranges under the curve.

For example, the fact that 0.5 of the area lies to the right of the mean (where z=0) means the probability of a z-value falling between 0 and any positive value of z is $p(z \geqslant 0) = 0.5$. You should also keep in mind that on a continuous line, the difference between $p(z \geqslant 0)$ and $p(z > 0)$ is too small to matter: $p(z \geqslant 0) = p(z > 0) = 0.5$.

The area from z=0 to z=2 can be calculated by adding 0.3414 + 0.1359 = 0.4773, but an easier way would be to look at the tables in Appendix D. Look for 2.00 in the z column, and then for the value given for the 'Mean to z', which is 0.4772.

If you want to read the area between z=1 and z = 2 from the table, you are going to have to calculate it. Look up the value for 'Mean to z' for 2, which is 0.4772, then look up the value for 'Mean to z' for z = 1, which is 0.3413. Then subtract the latter from the former: 0.4772 – 0.3413 = 0.1359. If you look at the graph in Figure 2.7, you will see how this works: you are subtracting the area that you are *not* interested in (from z = 0 to z = 1) from the greater area (z = 0 to z = 2) to get the area between z = 1 and z = 2.

The probability of a z-value of between two and three standard deviations is $p(2 < z < 3) = 0.0215$. There is only a small probability of getting a z-value of greater than 3 in a purely random event: $p(z > 3) = 0.0013$: just a bit more than one in ten. The same is of course true for areas to the left: $p(-3 < z < -2) = 0.0215$ and $p(z < -3) = 0.0013$.

If you calculate or look up the area from the mean to z = 3, you get 0.4987. So the probability of a random z score of between z = –3 and z = 3 is 2 x 0.4987, which is more than 99%: $p(-3 < z < 3) = 0.9974$. You will notice, if you look in the tables, that the area beyond z = 4 is not even given. The area between z = –4 and z = 4 is so close to 100% that it can be regarded as 1.

Because of the symmetry of the curve around 0, the probabilities that you will find in the tables in Appendix D are given for positive values of z only. If you have to calculate the probability for a negative z-value, ignore the sign, but keep in mind that you are working on the left-hand side of the distribution when you interpret the results.

2.3.4 The z transformation

The standard normal curve presents a standardised distribution of probability values, which is very useful in hypothesis testing (as we show in Topic 3). Any variable (x) that comes from a normal distribution can be transformed to its representation on a standard normal distribution, provided that we know the mean and the standard deviation of the variable scores.

The formula for the transformation is

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

where **x** represents the variable, μ is the population mean, and σ the standard deviation of the population from which x was obtained.

A z-score is the original measurement transformed into a point on a standard normal distribution. Therefore, all the characteristics of the standard normal distribution apply. For example, the size of the z-score always reflects the number of standard deviations that a particular score lies above or below the mean.

While its major use is in calculating probabilities, as we show in Topic 3, transforming a score from a normal distribution to its associated z-score has an additional benefit. Transforming a set of measurements, each with a different mean and a different standard deviation, into a z-score can be used to compare an individual across different distributions. After transformation, all the scores will fall on a common standard normal distribution with a mean of 0 and a standard deviation of 1, which makes it possible to compare them directly.

Example: We find that a specific person has an IQ of 120 on an IQ test (standardised to a mean of 100 and a standard deviation of 15 on a normal distribution). The same person also gets 8 on a 9-point test for his aptitude for mathematics (where the scale has a mean of 5 and a standard deviation of 1.5 on a normal distribution). It is difficult to compare these scores, but transforming them into equivalent z-scores makes it easier.

For IQ:

$$z_1 = \frac{120 - 100}{15} = -\frac{20}{15} = \frac{4}{3} = 1.333$$

For mathematical aptitude:

$$z_2 = \frac{8 - 5}{1.5} = \frac{3}{1.5} = \frac{4}{3} = 2.000$$

From this we can infer that his mathematical aptitude is higher than we would expect when we judge him by his IQ.

In practice, we are rarely able to calculate the mean of a population of scores and the standard deviation of the population σ, because we seldom have population scores available. In such cases we can draw a representative sample from the population and use the sample statistics $\bar{\mathbf{x}}$ and **s** to calculate z, as follows:

$$z = \frac{x - \bar{x}}{s}$$

Example: A student gets 56% for an exam in cognitive psychology and wants to know where he or she stands in relation to the rest of the class. Suppose the mean mark for all the students is 52%, and the standard deviation is 4. If we assume further that the marks are normally distributed, the z-score corresponding to 56% can be calculated using the formula given above:

$$z = \frac{56 - 52}{4} = \frac{4}{4} = 1$$

The student gained a mark that is one standard deviation above the average.

If we look up a z-score of 1 in the z-table (in Appendix D), we find that it corresponds to a portion of 0.84 when rounded off to two decimal places (we need to look at the 'larger portion' as can be seen from the grey area in Figure 2.8). This means that the student did better on this test than approximately 0.84 (or 84%) of the group.

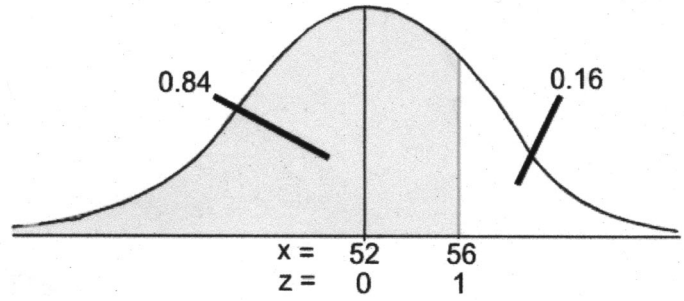


FIGURE 2.8

Now consider the case where the same student achieves 56% for research methodology and the mean for this course is 65% with a standard deviation of 6. Then the corresponding z-score is

$$z = \frac{56 - 65}{6} = \frac{-9}{6} = -1.5$$

This time, the mark is 1.5 standard deviations below the average mark, as indicated by the negative sign. Since the normal curve is symmetrical, we can look at the table values for a positive value of z = 1.5, as long as we remember which part of the graph we are interested in. If we look up the z-value associated with 1.5 in the table, we see that for a z-score of 1.5 the smaller portion of the distribution is 0.0668, and the larger area is 0.9332. So the student did worse than 93.3%, or better than 6.68% of the class. Note that we can tell from the negative sign that the student performed *below* average and, therefore, that his mark places him in the *bottom* 0.0668 proportion of the class. (It always helps to draw a picture in these cases, such as the one in Figure 2.9.)

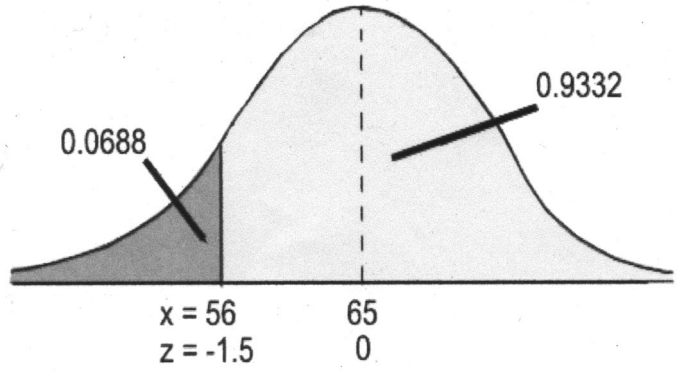


FIGURE 2.9

STUDY UNIT 2.4
Sampling distributions and the central limit theorem

2.4.1 Sampling and sampling distributions

As we explained in Topic 1 (section 1.4.3), we distinguish between the **population** that is being studied, and the **sample** of data that we use to represent this population in our study. The purpose of sampling is to use a relatively small number of cases to draw conclusions about a much larger group. The group you wish to study is the population and the group you actually involve in your research is the sample; in other words, the sample *represents* the population. Once you have obtained certain results based on the sample, you can go on to generalise (or apply) your results to the population.

The main reason for using samples is to save the researcher time and money. Sampling is a useful short-cut, leading to results that are almost as accurate as those in which the complete population is studied, but for a fraction of the cost. Most studies are subject to a law of diminishing returns in that once a certain number of cases/individuals has been studied, each successive case will not add much to our understanding of the emerging patterns in the results. It is, in any case, often impossible to study the entire population since it may not be practical to try and find all the members of a specific group. For example, if you were to do a study on working mothers, how could you possibly observe *all* the people in the world (or even in a specific country) who belong to this population?

When selecting a random sample, a researcher hopes that it will be representative, but he or she has no guarantee that it will be. Any two separate samples will probably be different even if they are selected from the same population on a random basis. The samples will consist of different individuals and different scores, even though – because of appropriate sampling techniques – we can expect them to be broadly similar and a reasonable approximation to the population.

If we were to draw all the possible random samples of 30 people from a large population, such as the population of Unisa students, the number of samples that could be drawn would be impossibly large. Many of these samples would

deviate from the actual population in some way purely by chance, and would, therefore, contain a margin of error. Thus, the notions of sampling and error are intimately connected (see section 1.4.4).

Because the samples are different, the statistics calculated on the basis of samples will also vary from sample to sample, in spite of the fact that the samples were taken from the same population. This last point is very important: **because a sample is randomly chosen from a larger population, any statistic such as the value of the sample mean, will vary from sample to sample.**

Even with small populations it is possible to obtain many thousands of different samples, so that it may seem hopeless to try to establish some simple rules for the relationship between samples and populations. Fortunately, despite the huge number of different possible samples that can be selected from a single population, the set of possible samples themselves forms a simple pattern that makes it possible to predict the characteristics of a sample with some accuracy. The ability to predict the characteristics of a sample is based on the concept of the *distribution of sample statistics*.

The **sampling distribution of a statistic** is the set of all possible values of the statistic when all possible samples of a fixed size are taken from the population. It is important that you understand this simple definition fully. The sampling distribution refers to the variation of a statistic, for example, the sample mean ($\bar{x}$), from sample to sample. Note that here we are not concerned with the variation of individual elements in the sample, or individual elements in the population, but with the variation of a *summary value* (such as the mean) for a sample.

The sampling distribution refers to variation over a hypothetical set of all possible samples. This may be a rather difficult concept to grasp. It is easy to visualise the variation of individual elements in a sample because the values are there for you to see. It is also easy to think of the variation of individual elements in a population because you can picture the set of individual units. But it is much more difficult to imagine the set of all possible samples because (1) we typically deal with one or two samples so that the idea of a sampling distribution is not really intuitive, and (2) the set of all possible samples is typically extremely large (conceptually infinitely many samples).

Because the concept of a sampling distribution is rather abstract, it is useful to consider a simple example. Suppose that an entire population consists of only five medical residents working in the emergency section (ER) of Johannesburg General Hospital, and that we want to determine the average age of the residents. We could proceed by selecting a sample of two residents, determine their individual ages, calculate their mean age and use this to estimate the corresponding population parameter.

Of course, nobody would actually draw a sample from such a small population, but we have deliberately chosen this example for its simplicity. Also, normally, the population values would not be known in advance. (For why would you then have to take a sample?) But suppose that we know the five residents have the following ages:

33; 28; 45; 43; 47.

The population mean for 'age' and the population standard deviation for the variable can now be calculated:

$$\mu = \frac{1}{n}\sum x \frac{33+28+45+43+47}{5} = \frac{196}{5} = 39.20 \text{ years; and}$$

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{n}} = \sqrt{\frac{272.8}{5}} = \sqrt{54.56} = 7.39 \text{ years}$$

Keep in mind that these are *population parameters*; both of these values refer to all the individual units in the population. (*Note:* The way that these parameter values were calculated is shown in Appendix C.)

Suppose, however, that instead of merely drawing a single sample, we now continue to draw *all* the possible samples of size 2 from the population. All in all there are 10 such samples. They are shown in the Table 2.6 below, together with the mean for each sample.

Take note of the following facts about the data in Table 2.6:

- It is a theoretical table of *all possible samples* of size 2. Normally, a researcher would choose only a single one of these samples (by selecting two of the five individuals at random) and he or she would then make decisions based on this single sample.
- The table presents a sampling distribution in the rightmost column. The sampling distribution of $\bar{x}$ refers to the variation of the sample mean over all the possible samples from the population.

TABLE 2.6: All 10 samples of size 2 that can be drawn from the population

Sample units		**Sample mean ($\bar{x}$)**
33	28	30.5
33	45	39.0
33	43	38.0
33	47	40.0
28	45	36.5
28	43	35.5
28	47	37.5
45	43	44.0
45	47	46.0
43	47	45.0
Mean of all these means		**39.2**

Now consider this sampling distribution of $\bar{x}$ in more detail. Some values of $\bar{x}$ are above the population mean, and some values of $\bar{x}$ are below the population

mean (which we know to be $\mu = 39.2$ from our calculation above). As you can see in the table, the mean of $\bar{x}$ over all possible samples is $\Sigma x/n = 392/10 = 39.2$, which is the same as the population mean (i.e., 39.2). The mean of the sample means, therefore, provides an accurate estimate of the population mean, μ. Note that $\bar{x}$ varies from sample to sample, but since we have drawn all possible samples here (each with probability of 0.1 of occurring, as represented in Table 2.6), we have a probability distribution of $\bar{x}$ over all possible samples, and can, therefore, compute the expected value of $\bar{x}$.

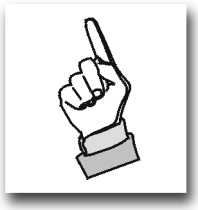
The crucial point to note here is that the statistic computed from an individual sample is an element from the distribution of all possible samples, and this latter distribution is called the **sampling distribution** of the statistic.

Normally, of course, we'll not know what our true population parameter is, and we would have calculated the mean from only a single sample – but we can still apply the basic principle: that our sample mean will be a reliable estimate of our population mean. We can also estimate the size of the error we would make if we used the sample mean as an estimate of the population mean. This is referred to as the **standard error**, and it is specified in the **central limit theorem**, which we discuss next.

2.4.2 The central limit theorem

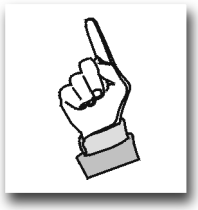
In the above example, we have shown that a distribution of sample means can be obtained for a very simple, specific situation. However, in most cases it will not be possible to list all the samples and to compute all the sample means. Fortunately, the general characteristics of the distribution of sample statistics such as the mean are specified by a mathematical theorem called the **central limit theorem**.

It is as follows:



If a simple random sample of size n is selected from a population with mean μ and standard deviation σ, the sampling distribution of means obtained from all possible samples is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$.

The central limit theorem gives a precise description of the distribution that you will obtain if you selected every possible sample, calculated every sample mean, and constructed the distribution of the sample mean. The importance of the theorem lies in the fact that we can use it to describe a sampling distribution without actually having to sample a population of raw scores 'infinitely', and because of this we can calculate the extent to which any sample mean approximates the mean of the population from which it was drawn.



Some interesting facts about this theorem should be noted:

- This theorem gives the sample distribution of the sample means for *any* population, irrespective of the shape, mean or standard deviation of the original population.

- The distribution of sample means will become more normal as sample size (n) increases, so that with larger and larger samples the shape of the distribution of sample means will become increasingly normal in form. In fact the distribution of sample means approximates a normal distribution very rapidly: by the time the sample size reaches n = 30, the distribution is very close to perfectly normal.

Just as the normal distribution is defined by its mean and standard deviation, so the distribution of sample means is described by the same two quantities. The central value of the sampling distribution equals the population mean (i.e. the mean of the distribution of all possible means is the same as the mean of the population from which the samples were drawn, or $\mu_{\bar{x}} = \mu$) while the standard deviation of the sample means is estimated by a value we call the **standard error** of the mean. Like a standard deviation, the standard error of the mean tells us by what average amount the sample means deviate from the mean of the sampling distribution. It is an estimate of the size of the error we shall make if we use the mean of the distribution of sample means as an estimate of the true population mean, that is, if we use $\mu_{\bar{x}}$ to estimate μ.

The standard error is denoted by $\sigma_{\bar{x}}$. The σ indicates that we are describing a population, and the subscript $\bar{x}$ informs us that we are dealing with a population of sample means. The standard error is given by dividing the population standard deviation by the square root of the sample size:

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

Figure 2.10 is an illustration of the relationship between a population and a number of samples drawn from that population, and the population mean and standard deviation in relation to the mean of all the samples and the standard error.

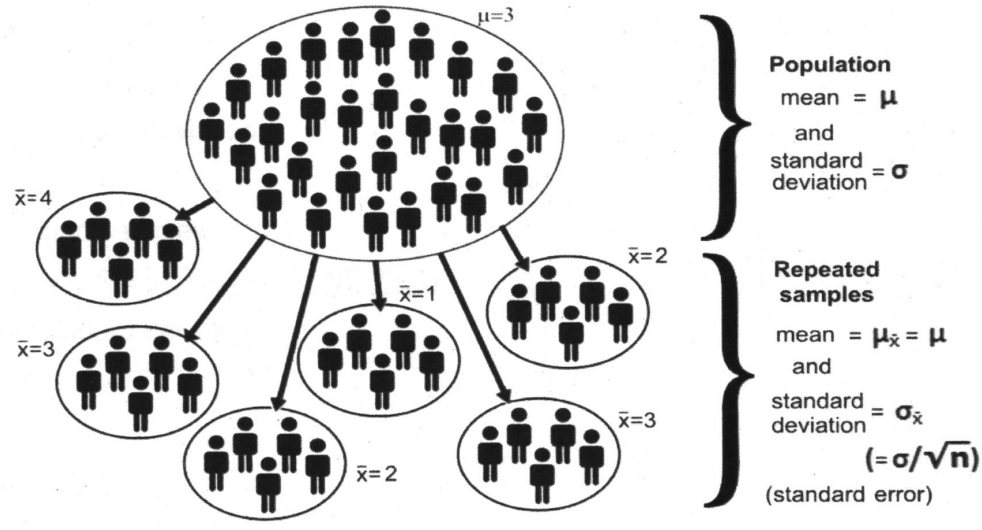


FIGURE 2.10: Samples and sample means in relation to a population

The standard error is an extremely valuable measure because we can use it to estimate how well a sample mean approximates its population mean in general,

that is, how much error you can expect on average between the sample mean ($\bar{x}$) that you calculated from your sample and the population mean (μ) that you are trying to estimate. In other words, it is an indication of the size of the error that you make by using a sample of a particular size (n) to determine the population mean.

This amount of error will decrease as the size of the sample increases. For example, suppose we have $\sigma = 10$ and then compute $\sigma_{\bar{x}}$ for three different sample sizes.

- For n = 4 we have

  $$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{4} = 5$$

- For n = 25 we have

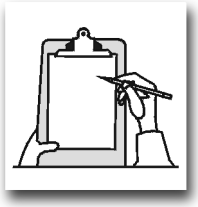
  $$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{25} = 2$$

- And for n = 100, we have

  $$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 10/\sqrt{100} = 1$$

With each increase in sample size there is a corresponding decrease in the standard error, which in turn implies that $\bar{x}$ more closely approximates μ for larger and larger samples. So if the standard error is small we know that the sampling statistic is a relatively accurate reflection of the underlying population parameter. Conversely, the larger the standard error the greater the sampling error, and likewise the greater the difference between the sampling statistic (e.g. the mean, $\bar{x}$) and the corresponding population parameter (i.e. μ).

Because of the central limit theorem, we now know something useful about the distribution of any mean – no matter how original data were distributed, means tend to be distributed normally with a known mean and known error when we try to estimate it (for a sample of a reasonable size). This is information that will come in very handy when we try to make inferences about populations based on sample data. The importance of this theorem and the use of the standard error in statistics will become clearer in Topics 3 and 4, where we see how it is used in actual statistical analyses.

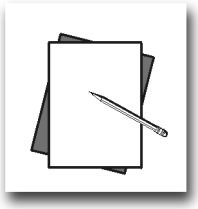


Summary of major points in this topic

After you have worked through all the study units included in this topic, you should know

- how to calculate and interpret probability values
- the law of large numbers
- the additive and multiplicative probability rules for independent events
- how to determine and interpret a conditional probability
- how to interpret a probability distribution
- how the binomial distribution represents probabilities for a discrete variable
- the standard normal distribution curve as a probability distribution for continuous variables

- how to determine and interpret a z-score
- the difference between samples and populations and the symbols used to describe them
- the relevance of the central limit theorem to psychological research
- how to determine and interpret the standard error of a sample mean



Multiple-choice questions and solutions

Questions:

1. A researcher randomly selects a child from a group of 300 boys and 400 girls to participate in a research experiment. What is the probability that the child selected will be male?

   1. 0.5
   2. 0.43
   3. 0.57

2. With reference to the question above, what is the probability that the child selected will *not* be a male?

   1. 0.5
   2. 0.43
   3. 0.57

3. A street magician shows you a deck of 52 playing cards and asks you to randomly pick a card from the pack, without showing him what card you take. He then correctly informs that you picked an ace. What is the probability that he could have guessed correctly simply by chance?

   1. 1/52
   2. 1/13
   3. 1 /27

4. Select the statement below that provides the most accurate formulation of the law of large numbers:

   1. Any probability value will eventually converge on an average value.
   2. Probability estimates are prone to considerable error because one can never predict probabilities accurately.
   3. If a statistical experiment event is performed a large number of times, a specific outcome will tend to converge on its theoretical probability.

5. A student writes in her research report that p(Hypothesis 1: true) $\leqslant$ –0,3. Upon reading this, her supervisor becomes angry. Why?

   1. The student has given no reason for her conclusion.
   2. A probability cannot be negative.
   3. One should always state a confidence level when formulating an hypothesis, and the student did not do this.

6. The standard normal distribution has a mean of ..... and a standard deviation of .....

   1. 0; 1

2. 1; 0
3. 1; 1

7. A ..... is completely described by the mean and the standard deviation.

    1. normal distribution
    2. population
    3. sample

8. A researcher at Unisa administers an attitude scale to a group of research participants. Their average score is 2, and the standard deviation is 2. Suppose a participant obtains a score of 2, what is her z-score?

    1. 2
    2. 0
    3. 1

9. The first exam in a statistics course yielded a normal distribution of scores with a mean of 35 and a standard deviation of 10. If you were to select the score of one student at random, what is the probability that the score would be 45 or above?

    1. 0.34
    2. 0.66
    3. 0.16

10. Which of the following statements about population parameters is the most accurate?

    1. They are essential for making statements about probability.
    2. They are usually unknown.
    3. They are essential, but cannot be estimated from sample information.

11. The larger the sample, the more likely it is that the sample will accurately reflect the population mean.

    1. True
    2. False

12. A researcher is interested in the IQ of students at his college. The researcher believes that the IQ of college students, measured by means of a standardised test, has a mean of 110 and a standard deviation of 15. The researcher takes a random sample of college students and finds that the mean IQ is 120. Which of the following situations provides the strongest evidence that the mean IQ of his students is greater than 110? (Note: n = size of sample.)

    1. n = 10
    2. n = 50
    3. n = 100

13. When the sample size (n) decreases, the dispersion of the sample means

    1. becomes less.
    2. becomes greater.
    3. remains the same.

Questions 14 and 15 are based on the following research scenario:

A national survey of college students indicates that students drink an average of 4.1 alcoholic beverages per week. A researcher randomly selects 30 college students and asks each one how many alcoholic beverages he or she consumes per week. The researcher finds that the students surveyed drank 182 alcoholic beverages during the week.

14. What is the population mean?

    1. 30
    2. 4.1
    3. 6.1

15. What is the sample mean?

    1. 30
    2. 4.1
    3. 6.1

16. A university researcher is interested in the incomes of her psychology graduates. A national survey shows that university graduates (from all departments) earn R127 500 on average, per year, with a standard deviation of R30 000. The researcher believes that her college's alumni make more than R127 000 a year (i.e. the researcher believes that the population of psychology students is different from the population of all university students). The researcher says: 'I talked to two graduates [a sample of 2] and they both make over R200 000 a year! Obviously, our students make more than R127 500 per year after university.' Do you agree with this researcher? If not, indicate why not. Cite statistical evidence in support of your answer (but no calculations are necessary).

Solutions

1. To calculate the probability that the child will be a boy, we use our formula

$$\text{p(child was a boy)} = \frac{\text{Number of favourable outcomes}}{\text{Number of possible outcomes}} = \frac{300}{700} \approx 0.43$$

   Therefore, there is a 0.43 probability that the child will be a boy, and option 2 is correct.
2. Since we have determined that p(boy) = 0.43, we know that p(not boy) will be 1 – 0.43 = 0.57. There are only two possibilities (boy or not boy) so one is the other subtracted from 1. Therefore, option 3 is correct.
3. There are four aces in an ordinary deck of 52 cards. The probability of picking an ace by chance is

$$\text{P(ace)} = \frac{4}{52} = \frac{1}{13}$$

   This shows that the correct option is 2.
4. The correct option is 3. Option 1 is not correct, because it omits the important condition of the law of large numbers, namely, that the event has to be performed a large number of times before the theoretical probability will be reached.

5. Probability values fall in the range 0 to 1, and are always positive. Option 2 is correct because any supervisor would be annoyed if a student used a negative probability value. A negative p-value has no interpretation.
6. The definition of the standard normal distribution (see section 2.3.3) is that it has a mean of 0 and a standard deviation of 1. Option 1 is, therefore, correct.
7. Any normal curve can be generated provided that we know its mean and standard deviation. Option 1 is, therefore, correct. Populations and samples are not necessarily normally distributed, so that further information may be needed to describe them. Therefore, options 2 and 3 are incorrect.
8. Since the score that you obtained is exactly the same as the mean, we know that your z-score is 0, because it does not deviate from the mean at all. Option 2 is, therefore, correct.
9. The z-score for a score of 45 is

$$z = \frac{45 - 35}{10} = \frac{10}{10} = 1$$

The score, therefore, lies one standard deviation above the mean, so that 0.84 of the scores lie below it. The probability of a score of 45 or higher is, therefore, 1 – 0.84 = 0.16. The correct option is, therefore, 3.

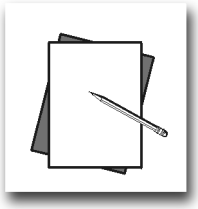
10. Population parameters are usually unknown and have to be inferred from sample data. Therefore, option 2 is correct. Since population parameters are unknown, they cannot be *essential* to make statements about probability. Option 1 is, therefore, incorrect. Option 3 is also incorrect because it incorrectly states that population parameters cannot be estimated from sampling information, but the whole process of statistical inference is actually concerned with inferring information about a population from sample data.
11. The statement is true, so option 1 is correct. Larger samples will contain less individuals with only extreme values and are more likely to have a normal distribution. Larger samples will, therefore, reflect the population mean more accurately than small samples.
12. From the previous question we know that larger samples are more likely to describe the population mean accurately. Since option 3 is the largest sample, we can expect it to provide the best indication of the true population mean.
13. Alternative 2 is correct. The spread or dispersion of the sampling distribution of means is given by the standard deviation of the sampling distribution of means (i.e. the standard error). The formula for determining this standard deviation is σ/n (see section 2.4.2), and the value becomes larger as n decreases.
14. Our best guess about the population mean is given by the national survey, because it presumably involved a very large sample of students. We can, therefore, assume that the population mean is 4.1 and that option 2 is correct. Option 1 is the number of students (and not the mean) and is, therefore, incorrect. Option 3 is also incorrect because it is the mean of the small sample (n=30) investigated by the researcher (see the calculation in the next question below), and this mean is, therefore, less reliable as an indicator of the population mean than the data obtained from the national survey.

15. The sample mean is obtained by

$$\bar{x} = \frac{\sum x}{n} = \frac{182}{30} = 6.1$$

(rounded off to one decimal place). Option 3 is, therefore, correct.

16. The researcher is generalising on the basis of the two students who were questioned about their salaries. The researcher's sample (only two students) is probably too small to make valid inferences about the population of graduate students.



The questions above were all very easy. Here are a few slightly more difficult questions to test your understanding of the study units in this topic.

17. Why is a probability distribution always of a theoretical nature?

Answer:

A relative frequency can only be used as a probability distribution if we assume that probability theory applies. Probability theory is a theoretical approach that makes use of theoretical constructs such as probability distributions to describe empirical phenomena.

18. Suppose we randomly draw a sample of five scores from a population and calculate the mean for this sample. The same procedure is repeated 10 times.

    1. Why are the mean values for the samples not the same?
    2 Given the 10 samples plus the fact that the mean of the population distribution is unknown, what would the best estimate of the population mean be?

Answer:

1. Each sample provides a different estimate of the population because of random sampling error.
2. The best estimate of the population mean can be obtained by calculating the mean of the 10 means. NB: The 10 means can now be considered in the same way as any set of 10 scores, and we might be interested in the mean, standard deviation, frequency distribution, et cetera, of this set of scores.

19. Does random sampling ensure a sample that is representative of the population?

Answer:

No it does not, because of sampling error that will play a role even if random sampling is used. However, since sampling is random, it does make it possible for one to derive a probability distribution for a particular sample statistic, such as the sample mean. Given an hypothesised value for the population mean, we can judge the likelihood of our sample mean under the derived probability distribution of means.

20. Suppose that the population distribution of a dependent variable is assumed to be normal, with a mean of 50 and a standard deviation of 10. Suppose, further, that samples of the same size are drawn randomly from the population with replacement.

1. What would the mean of the sample means approach as the number of samples of the same size that are drawn approaches infinity?
2 Can the frequency distribution of all the samples be fully specified?

   There could be an infinitely large number of samples, each with its own mean and standard deviation.

Answer:

1. It will approach the numerical value of the population mean.
2 Yes, provided that the population mean is specified (hypothesised) and the distribution of sample means can be assumed to be normal. The latter assumption may be made if it can be assumed that the population has a normal distribution, or if a very large sample is selected from the population.

21. How would you derive a probability distribution for the mean?

Answer:

Assume that the theoretical distribution of sample means of the same size, selected randomly from the population, is normal with a mean that is equal to the population mean, and a standard deviation equal to $\sigma/\sqrt{n}$ (if σ is known). The distribution can now be transformed to the standard normal distribution if σ is known. (Note: If σ is not known, we shall have to use another distribution, as you will see in Topic 3 (i.e., the t-distribution).)

22. Suppose a researcher transforms each score in a non-normal population to a z-score. Will these scores be normally distributed? Are the z-scores in the z-tables normally distributed?

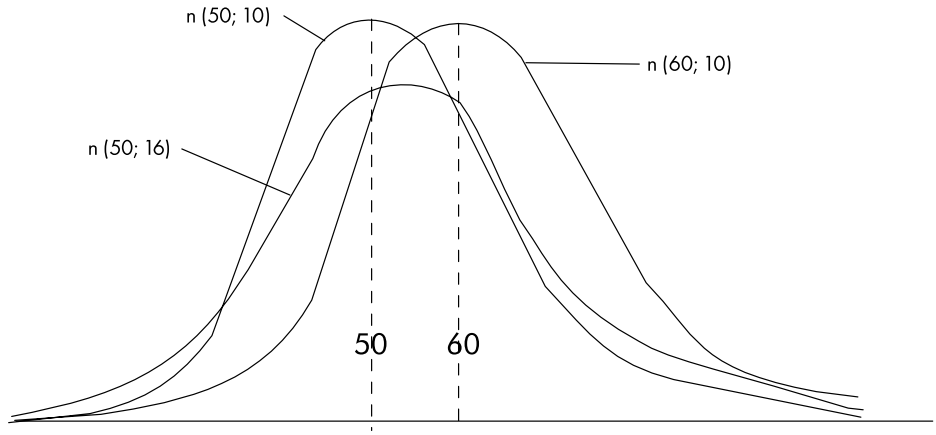
Answer:

No. The z-transformation does not change the shape of the original distribution. The z-scores in the z-tables are normally distributed because the z-table specifies, by definition, the standard normal distribution.

23. Make a freehand drawing of each of the following normal distributions on the same scale:

1. a normal distribution with mean = 50, variance = 100
2. a normal distribution with mean = 50, variance = 256
3. a normal distribution with mean = 60, variance = 100

Answer:



Note: Remember, the standard deviations in the graph are the square roots of the variances

24. Why is the central limit theorem important?

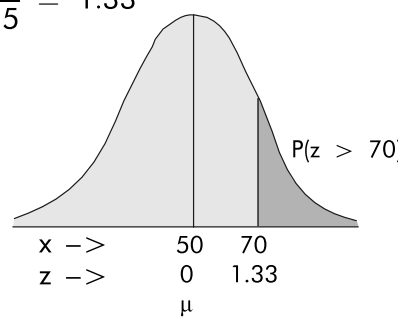
Answer:

It often happens that we doubt the assumption that the population distribution is normal. However, the central limit theorem states that, for a large sample size, the sampling distribution of a mean is close to normal, irrespective of the shape of the population distribution of the original data. This enables us to make inferences about means and develop test statistics for means.

25. Suppose a population distribution is normal with $\mu = 50$ and $\sigma = 15$.
    1. What is the raw z-score for a raw score of 70 in the population?
    2. Suppose the size of the population is 10 000. How many scores greater than 70 are there in the population?
    3. Suppose a single case is selected from this population using a random selection process. What is the probability that the score will be greater than 70? What is the difference between 'case' and 'score'?
    4. What is the probability that a single case selected randomly will have a score between 45 and 55?
    5. Suppose 25 cases are randomly selected from the population and the mean for this sample is 45. Is it possible to obtain a sample mean of 45 when one assumes that the population mean is equal to 50? Is it possible for the researcher to derive a theoretical distribution of the mean without selecting even a single sample?

Answer:

1. $z = \frac{70 - 50}{15} = \frac{20}{15} = 1.33$



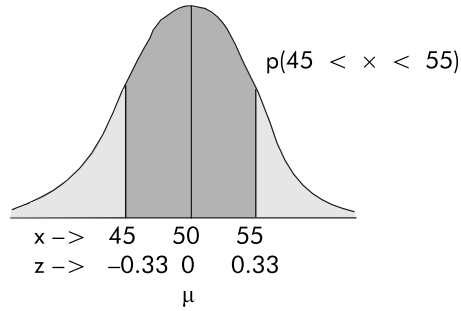
2. p(z > 70) = 0.092 (from z-table in Appendix D, using the smaller portion as indicated in the graph above.)

   So we can estimate that 0.092 x 10 000 = 920 scores in the population should be greater than 70.

3. As we have already calculated for 2, p(x > 70) =0.92

   A 'case' is the particular entity being observed, whereas a 'score' is a numerical value reflecting some characteristic of the case being considered.

4.



For $x = 45$ we find $z = \frac{45 - 50}{15} = -0.33$

And for $x = 55$ we find $z \frac{55 - 50}{15} = 0.33$

These two values are therefore an even distance away from the mean (of z=0) on the standard normal distribution. We can, therefore, look up $p(x > 45) = p(z > -.33)$ which is 0.6293, and subtract from it $p(x > 55) = p(z > 33)$ which is 0.3707 (the larger portion and the smaller portion respectively for $z = 0.33$ on the standard normal distribution tables in Appendix D). This gives $p(45 < x < 55) = 0.6293 - 0.3707 \approx 0.26$.

5. Yes, any result is possibly due to sampling error. However, some results (sample means) in this example will be less probable than others. It is possible to derive a theoretical distribution for the mean provided one knows the population standard deviation and hypothesises a value for the mean of the population.

TOPIC 3

General principles of statistical hypothesis testing

Quick overview

In this topic we explain how a research hypothesis can be transformed into a statistical hypothesis. This statistical hypothesis is a formal expression of the research hypothesis, which enables us to test it. You learn how the process of testing such an hypothesis works, and how the notion of probability is used as a basis for making a decision or inference based on the outcome of the observations that are derived from a sample. You learn about the risk of making errors, and how to evaluate the size of an effect that you have observed.

The present topic is organised into the following study units:

- *Study unit 3.1:* Translating a research hypothesis into a statistical hypothesis
- *Study unit 3.2:* Using data from a sample to calculate the probability of a particular result
- *Study unit 3.3:* Making a decision regarding the null and alternative hypothesis

STUDY UNIT 3.1
Translating a research hypothesis into a statistical hypothesis

3.1.1 Introduction

In Topic 1, we emphasised that the basic purpose of research is to test a theory

by establishing whether certain relationships exist among constructs. A theory was defined as a network of relations between constructs that has been validated by research, and a research hypothesis was described as a statement concerning a possible relationship between two or more such constructs that we want to test as part of the validation for the theory. Part of the research process involves refining the research hypothesis until it suggests or implies the following:

- how the constructs involved will be measured
- what the research population is
- the nature of the relationship being investigated

Once the research has reached this level of refinement, the research hypothesis is often referred to as an operational hypothesis. However, when it is clear from the context what hypothesis we are dealing with, we tend to use the terms 'research hypothesis' and 'operational hypothesis' interchangeably (see Topic 1).

The next stage is to translate the research (or operational) hypothesis into a *statistical hypothesis* to test, on the basis of sample observations, whether the relationship proposed in the research hypothesis indeed exists.

The present topic explains how such statistical hypotheses are derived from the research hypothesis and are to be tested by means of sample data.

We explain the testing of hypotheses below using a simple example, beginning with the formulation of a research hypothesis, deriving the appropriate statistical hypotheses from it, and ending with a conclusion that can be related back to the original research hypothesis.

3.1.2 The translation of the research hypothesis into a statistical hypothesis

Consider a simple research hypothesis such as the following:

> *Unisa students tend to have higher intelligence scores than the general population.*

It seems as if there are two constructs that we need to consider here: the construct 'IQ score' and the construct 'group membership' which we shall use to indicate 'Unisa students' versus 'general population'. The value we are investigating (the dependent variable) is IQ score, and we want to know whether it is affected by group membership. We can also think of it as determining whether you can predict something about IQ when you know into which group a person falls (e.g. If you know someone is a Unisa student, is this information of any use in guessing what his or her IQ might be?).

On the face of it, 'group membership' is a nominal scale measurement, since you can only be in one of two categories (which we can indicate with numbers), such a 'Unisa student' is coded as 1 and 'general population' is coded as 2. These numbers are arbitrary (it will not matter if we code them the other way round) and it is also a dichotomy (only two possible values exist; see Appendix B for details on measurement scales).

The other variable (*IQ test score*) can be regarded as a continuous measurement, and we shall assume that it is a measurement of at least an interval scale level (see Appendix B). This has the advantage that we can calculate means, and we can reason that an appropriate way to test the relationship would be to compare the two group means. (Remember that a mean is an indication of the *central tendency* of a measurement, that is, where in the distribution of data a specific measurement is focused or concentrated; see Appendix C.) This seems to be the appropriate aspect of the data to compare in this particular case. We can, therefore, consider calculating the two means, one for each group (Unisa students and general population), to see if the Unisa group mean is bigger.

However, as it so happens, we know something about the IQ of the general population that makes the research process much simpler: the IQ test that we are using was standardised in such a way that the population mean IQ for the general population is always 100, and the standard deviation is 15. So for practical purposes we can take these two values as constants: for the general population, $\mu = 100$ and $\sigma = 15$. This means that we only need to determine the values of the mean and the standard deviation for a sample of Unisa students, and compare these to the constant values for the population. Note that testing the IQs of a sample from the general population would be a perfectly legitimate way to do the research, but it is simply not necessary (which can save time and costs in this case).

We are now ready to state the research hypothesis in the form of a statistical hypothesis, which is to say we can express it symbolically as follows:

H_1: $\mu > 100$

Here H_1 indicates a (alternative) hypothesis. In this particular hypothesis, the symbol represents the mean or average intelligence score of all Unisa students (i.e. the population of Unisa students). The value 100 is the mean intelligence score of the general population, which is a constant value that is known to us. The symbol '>' stands for 'greater than' or 'larger than' (as we show in Appendix E).

This statement is, therefore, just a symbolic way of saying: The mean intelligence score of Unisa students is greater than 100.

What would be our prediction if our expectation is wrong that is, if the mean IQ does *not* differ from that of the general population at all? In this case, the mean IQ value for Unisa students would not differ from 100, so we would write:

H_0: $\mu = 100$

This hypothesis is a symbolic expression of the possibility that Unisa students are not different from the general population as far as the mean of their intelligence scores is concerned; in other words, there is no difference between the population mean (μ) of Unisa students and the population mean of the IQ score of the general population, which we know should be 100.

This hypothesis is referred to as the 'null hypothesis' because it is the hypothesis

that implies no effect: it claims that there is no difference to be found between the IQs of Unisa students and those of the general population. By convention, the null hypothesis is usually indicated with the symbol H_0. The other hypothesis, the one we are comparing with H_0, is indicated with H_1, and is referred to as the 'alternative hypothesis' (if there is more than one alternative hypothesis being tested, which is sometimes possible, we would indicate them with consecutive numbers, such as H_2, H_3, etc.).

Note that the null hypothesis is a precise statement: it gives a very specific prediction of what the mean IQ score of Unisa students should be. On the other hand, the alternative hypothesis specifies a range of possibilities: according to the way it is expressed, the mean IQ of Unisa students could be anything from just larger than 100 to infinity (indicated as ∞). The fact that a mean IQ of infinity is somewhat implausible (even for Unisa students!) is not relevant: the point is just that *some* value greater than 100 is indicated. The fact that one statement is precise (the null hypothesis) and the other not (the alternative hypothesis) has very specific consequences for how we proceed to test the hypotheses (as we explain in section 3.2.2 below).

Take note that a research hypothesis always translates into two mutually exclusive hypotheses (i.e. both cannot be true at the same time): a null and an alternative hypothesis. Also remember at this stage that, in Topic 1, we referred to quantities such as as parameters (population parameters). These particular statistical hypotheses are, thus, statements about the value of a particular population parameter. In our example, the parameter of interest happens to be the population mean, as this is the appropriate summary value to test when comparing the central value of two groups (or one group and a constant) when measures are on an interval scale, and we want to infer something about where the measurements for a particular group are concentrated.

As another example, let as consider formulating the research hypothesis somewhat differently:

> *Unisa students tend to have intelligence scores that differ from that of the general population.*

This research hypothesis differs from the previous one in that this time we do not specify whether we expect Unisa students to have a higher or lower intelligence than the general population, but only that their intelligence *differs* in some way from that of the general population. This hypothesis translates into the following statistical hypotheses:

> **H_0: $\mu = 100$**
> **H_1: $\mu \neq 100$**

Note that the symbol '$\neq$' means 'not equal to'. (It is also conventional to state H_0 first and then H_1, although this is not important.)

The table below (Table 3.1) summarises the kinds of research hypotheses discussed above and the statistical hypotheses into which they translate, using the Unisa IQ problem as a general example. Note that the null hypothesis formulation is always the same (always stating that the Unisa population mean

IQ *does not* differ from the general population), but that the alternative hypothesis varies according to how the research hypothesis is formulated. Note that each statistical hypothesis should always have both a null and an alternative statistical hypothesis.

TABLE 3.1: Directional and non-directional hypotheses

Research hypothesis	**Statistical hypotheses**
Unisa students are more intelligent than the general population.	**H_0: $\mu = 100$** **H_1: $\mu > 100$** Here, H_1 specifies that only sample means greater than 100 will be considered as possible evidence for rejecting H_0. Only sample results different from 100 in a particular direction will thus be considered. For this reason, H_1 here indicates that a *directional* or *one-tailed* test of H_0 is required.
Unisa students are less intelligent than the general population.	**H_0: $\mu = 100$** **H_1: $\mu < 100$** Again, this implies that only one tail of the distribution will be considered, but now the *directional test* is in the opposite direction because, this time, means of less than 100 are predicted.
In terms of their level of intelligence, Unisa students are different from the general population.	**H_0: $\mu = 100$** **H_1: $\mu \neq 100$** Where both values of the mean, either greater than or smaller than 100 are to be considered, a *non-directional* or *two-tailed test* is required.
In terms of their level of intelligence, Unisa students are no different from the general population.	**H_0: $\mu = 100$** **H_1: $\mu \neq 100$** Here, the values being tested for are actually the same as in the case of a difference being predicted. The difference is that we are actually aiming to test for the validity of the null hypothesis – i.e. H_0 rather than H_1. We thus see that the research hypothesis sometimes translates directly into the null hypothesis. But because any difference in either direction is of relevance, the alternative hypothesis is also a *non-directional* or *two-tailed test.*

Note that, in the table above, the null hypothesis always contains the 'equal to' symbol '='. The null hypothesis is the hypothesis that *no effect exists*, and in cases where we are testing a mean, this implies that two group means (or a group mean and a specific constant value) do *not* differ.

The alternative hypothesis can contain any of the symbols '$>$', '$<$' or '$\neq$' respectively, the symbols for 'larger than', 'smaller than' or 'not equal to'. When a comparison is between a value that is greater (more) than another, we use the symbol '$>$' and when a comparison is between a value that is smaller (less than) than another, we use '$<$'. The statistical test that must be performed in either of these cases is a *directional* or *one-tailed* statistical test (we use these expressions interchangeably). When we do *not* specify what the direction of the difference should be, and both a larger and a smaller difference between means are considered as relevant, the symbol '$\neq$' must be used. The statistical test to be performed will now be a *non-directional* or *two-tailed* test.

Once the statistical hypotheses have been formulated, the object of the statistical procedure is to try to choose between H_0 and H_1 on the basis of observations. You should keep in mind that these statements are mutually exclusive (if one is true the other cannot be true), but one or the other should be valid. The problem is to find a way to decide which one.

STUDY UNIT 3.2
Using data from a sample to calculate the probability of a particular result

3.2.1 Obtaining sample data and sample results

Once we have stated the null and alternative statistical hypotheses, we must plan how to study the populations involved. Let's again consider our original statistical hypotheses:

H_0: $\mu = 100$
H_1: $\mu > 100$

Remember, as before, that μ is the mean intelligence score of all Unisa students, and 100 is the assumed mean intelligence score of the general population.

In this case, we have to find out what the mean IQ score for Unisa students actually is, so that we can compare it to the given value of 100. As indicated before (in Topic 1, Section 1.4.3) we do not have to test each and every one of the Unisa students to get a mean value for the IQ of the entire Unisa student population. That would be expensive, time-consuming and usually practically impossible to do. Instead, we identify an appropriate *sample* of Unisa students, and use the group mean for the sample to represent the population parameter. In other words, we obtain a random sample of Unisa students, apply an intelligence test to each student in the sample, and then calculate a value for the sample mean. We use this sample mean (indicated by $\bar{x}$) as a substitute for the population mean (μ). It is important to take note of this: we set up our hypothesis in terms of *population parameters*, but we test it through the use of *sample statistics* (see Topic 1, section 1.4.3 for a discussion of this).

Suppose we select a random sample of 64 Unisa students. We then apply an intelligence test to each student and obtain his or her intelligence score. Let us assume that when we calculate the mean of the IQ scores of this random sample we find that the result is $\bar{x} = 104$.

Let us first consider this result at face value. Does it look like this result could imply that the alternative hypothesis is in fact true? The answer appears to be in the affirmative, since the alternative hypothesis states that $\mu > 100$, and the sample value of 104 is definitely larger than 100.

In fact, we are not yet entitled to conclude that the alternative hypothesis is true. This is because of the problem of *sampling error*. This error exists partly because we are using a *sample* to make conclusions about a *population*, in addition to which we are using a test that is only accurate to a certain degree. It is because of this random error that we require the use of statistical tests to see whether the result is in fact adequate for us to make a decision about the hypothesis. (See section 1.4.4 on the problem of the error term in measurement.)

We shall take this problem up again below (in section 3.2.2). However, before we consider this, let us first look at some other possible outcomes of our calculation of the sampling mean.

What would we do if we calculated the sample mean and found it to be $\bar{x} = 96$? Would we be able to decide that the null hypothesis can be rejected and the alternative hypothesis accepted? The answer is, clearly, that we could not, because the alternative hypothesis suggests we should expect to find a mean for the IQ measurement of the Unisa students that is *more* than 100, and 96 is definitely not more than 100 (we can write this symbolically as $96 \not> 100$).

The probability that a value of 96 is greater than 100 is in fact zero. The same arguments would hold if, based on our calculations from the sample data, we found a sample mean ($\bar{x}$) of exactly 100: the probability that 100 is greater than 100 is also zero. So before we even begin the process of calculating probabilities of results, we need to look at the sample result and use our common sense. If we are doing a *directional test* and our sample outcome is in the *wrong direction*, further analysis is not necessary. We know the alternative hypothesis cannot be true.

On the other hand, if we were testing a non-directional alternative hypothesis like $H_1: \mu \neq 100$, a deviation from 100 in *either* direction in our sample mean (either more than 100 or under 100) would imply that we may possibly have found a result that favours the alternative hypothesis, and further testing is required to establish this.

3.2.2 Calculating the probability of the sample result under the null hypothesis

Let us return to testing the hypothesis related to the research hypothesis stating that Unisa students have IQs that are higher than that of the general population, leading to the statistical hypothesis:

H_0: μ = 100
H_1: μ > 100

As before, we consider the situation where we found a sample mean for the Unisa students' IQ scores of $\bar{x}=104$ from a random sample of $n=64$ students.

As explained before, we cannot be sure that this value of 104 is *sufficiently*

higher than 100 to be sure that the alternative hypothesis is true. This follows because we are aware that our measurement contains a degree of random error due to the way in which the sample statistic was obtained (see Topic 1, section 1.4.4 for a discussion of sampling and measurement error). Obviously we can see by inspection that 104 > 100, but is this difference of merely four points on an IQ scale (104 – 100 = 4) big enough to compensate for possible random measurement errors?

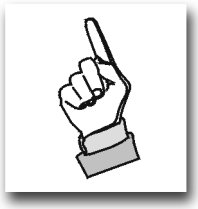
To test this, we need to find a way to calculate the probability of obtaining a result like $\bar{x} = 104$ from the sample if we assume that the value of μ (in the population from which the sample was drawn) is actually equal to 100. In other words, we want to know, if the actual value for the mean of Unisa students' IQ scores in the population is $\mu = 100$ (as claimed in the null hypothesis), what the chances are that we would find a sample mean of $\bar{x} = 104$, as a consequence of random errors created by our sampling and measurement procedures.

This probability of obtaining the value of 104 purely by chance, due to random measurement errors, when the null hypothesis is actually true, is referred to as the *p-value.*

This p-value is an extremely important concept in inferential statistics. Note that the null hypothesis states that the true population mean is 100, and that any sample deviation from this – such as a mean of 104 – is due to chance or random sampling error. The p-value is a probability value that indicates what the chances are that our sample value would be 104 or greater when it should 'really' reflect the population mean of 100, but misses the mark due to measurement error. Our task is to make up our minds whether we suspect that 104 is probably due to chance and is, therefore, really an approximation for 100, or whether we believe that 104 is sufficiently higher than 100 to support our alternative hypothesis, in spite of possible measurement error. If we accept the difference as significant, we would conclude that we can reject H_0 in favour of H_1.

The reason why we have to derive this p-value relative to an assumption that H_0 is true is that the probability that H_1 is true cannot really be calculated directly. It would be very convenient if we could calculate two p-values, one as if H_0 were true and another as if H_1 were true. Then one could simply choose the hypothesis that leads to the bigger p-value on the basis that this is the statement that is most probably true. Unfortunately, H_1 does not state an exact value of the population mean. A statement like $\boldsymbol{H_1: \mu > 100}$ refers to *any* of a possible *range* of values for μ, as long as it is larger than 100 (and since we assume that the variable IQ is measured on a continuous scale, the range of possible values of μ is in fact infinite). So we cannot determine what the distribution(s) of this range of means would be: there is an infinite number of distributions of the means implied in this statement.

So what we do instead is to calculate how far from the expected mean ($\mu = 100$) our observed mean ($\bar{x}=104$) is, and determine from this the probability that this difference is not 'real' but just a consequence of chance (random error). In other words, we determine the probability of getting this sample result, on a sample of this size (n=64), if H_0 were true.



We use the expression 'under the null hypothesis' by which we mean , 'assuming that the hypothesis H_0 is true'. Similarly, the phrase 'under H_1' would mean, 'assuming that H_1 is true'.

It is in determining the p-value that the issue of the statistical distribution of means becomes important. We need to know something about the general probability distribution of means to be able to make probability judgements about means. It so happens that our exploration of the distribution of sample means (which was discussed in study unit 2.4) taught us something very useful: because of the *central limit theorem* (section 2.4.2), we know all the important characteristics of the sampling distribution of the mean. We know what its shape will be, and we know its mean and its standard deviation.

Keep in mind that in order to apply this theorem we have to know both the mean and the standard deviation for the IQ scores of the general population. However, we know that the population mean is 100 and, fortunately, we also know that the population standard deviation is 15. This is because we happen to know that IQ tests are standardised in such a way that the population mean IQ for the general population is $\mu = 100$, and the population standard deviation is $\sigma = 15$ (as we mentioned before, in section 3.1.2).

Applying the central limit theorem we can use this information to obtain the sampling distribution of the mean, and determine that its mean would be

$$\mu_{\bar{x}} = \mu = 100$$

and its standard deviation (i.e. the standard error) would be

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{64}} = \frac{15}{8} = 1.875$$

Our task is now to calculate the probability that a result of 104 or greater could have occurred under this distribution. Figure 3.1 below illustrates this: it shows how the mean value would be distributed if we assume H_0 is true.

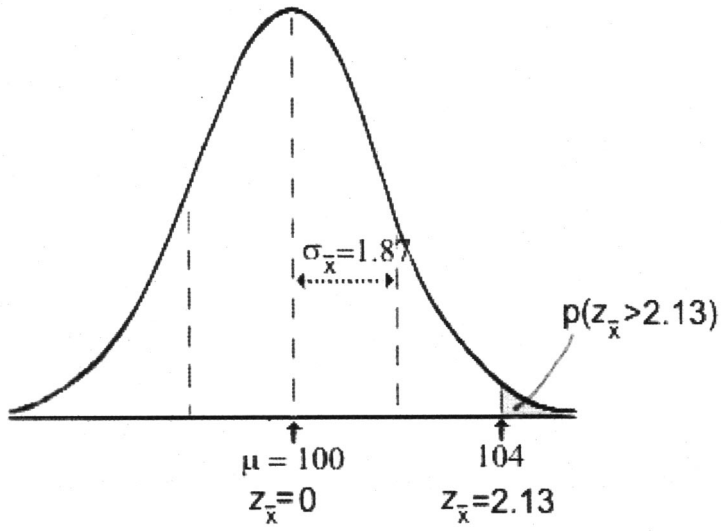


FIGURE 3.1: Distribution of the mean of the IQ scores

Note that this graph does not represent the distribution of IQ scores in the general population (with μ=100 and σ=15). It is the distribution of the ***mean*** *of these IQs as represented in the general population (with $\mu_{\bar{x}}$=100 and $\sigma_{\bar{x}}$=1.875). We want to compare our sample mean ($\bar{x}$=104) to this distribution, not to the distribution of IQs in general. (If the graph was drawn to scale, it would have been even more narrow.)*

The probability that $\mu > 104$ is given by the shaded area to the right of 104, and we give it a special name. We call this probability the *one-tailed or directional p-value* of the sample result, since we are considering one side of the distribution only, that is, doing *one-sided testing*.

We can determine this p-value by the same procedure that we used before (in Topic 2): that is, by first converting the value of the data point x to an equivalent point on a z-distribution (i.e. a standard normal distribution with a mean of 0 and a standard deviation of 1). Since the probability distribution of z is known, we can then read the probability of x exceeding a specific point directly from the z-tables (such as the tables in Appendix D).

In section 2.3.4 we learnt that the equation or formula for converting a x-value (any normally distributed variable) to a z-value (the standard normal distribution), relative to the mean and standard deviation of the distribution of x (μ and σ respectively), is

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

Since the probability distribution of z is known, we can then read the probability of x exceeding a specific point directly from the z-tables (such as the tables in Appendix D).

What we would like to do now is use this formula not for a raw data element (x) but for a mean value ($\bar{x}$). Since this implies that we are now working not with a distribution of raw data but rather with a distribution of means, we need to adapt the formula for z a bit. By substituting the symbols in the equation above with the symbols that refer to the distribution of a mean ($\bar{x}$), we derive the following formula:

The equation for this z-value (which we shall refer to as $z_{\bar{x}}$) now becomes

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

Here, $\bar{x}$ is the mean we are testing for, $\mu_{\bar{x}}$ is the mean of the distribution of these means, and $\sigma_{\bar{x}}$ is the standard deviation of the distribution of the means (also referred to as the standard error). These are the distribution parameters that we have already worked out above, using the central limit theorem. We can substitute the values that we found previously, as follows:

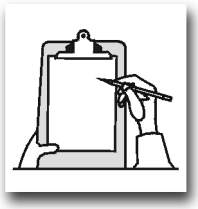
$$z_{\bar{x}} = \frac{104 - 100}{1.875} = \frac{4}{1.875} = 2.133$$

This is the z-value that is equivalent to a sample mean of $\bar{x} = 104$, as illustrated in Figure 3.1 above.

It follows from our calculation that $p(\bar{x} > 104) = p(z_{\bar{x}} > 2.133)$; that is to say, the probability that $\bar{x}$ is greater than 104 is the same as the probability that $z_{\bar{x}}$ is greater than 2.133. This conversion makes it possible to find out what this probability is, because we can consult the tables of the z-distribution in Appendix D. To find it, you have to look for the area of the *smaller portion* under the distribution (you can see from the shaded area in the graph in Figure 3.1 that this is the area we need to consider).

You should find that $p(z_{\bar{x}} > 2.133) = 0.0166$. (Check Appendix D to see whether you agree.)

This is the probability that we shall find a sample mean of $\bar{x} > 104$ if we assume that the null hypothesis is actually true (i.e. $\mu = 100$). We discover that the probability that the null hypothesis is true (the p-value) is actually quite small. So small in fact that we would probably be justified in concluding that the alternative hypothesis is acceptable, and that we seem to have found support for our assertion that the IQs of Unisa students are generally higher than those of the general population (but see the discussion of significance level in section 3.3.1).

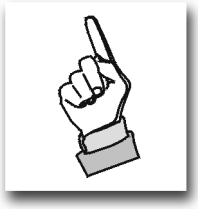


Here is a summary of the important points regarding the p-value:

- The p-value gives the probability of obtaining the sample result under H_0.
- If the p-value is very small, the probability is very small that the sample result would occur under H_0, and one should consider rejecting H_0 in favour of H_1.
- The smaller the p-value, the more likely that the null hypothesis is false and should be rejected in favour of the alternative hypothesis.

The calculation of the p-value entails calculating the area under the curve for $z > 2.13$ (i.e., the small grey area in figure 3.1). As was explained in section 2.3.3, we use statistical tables such as the table in Appendix D to determine these probabilities as they require the use of advanced mathematical procedures. Such tables are available in many statistics handbooks for a variety of test statistics. (Note that the table in Appendix D supplies one-tailed values for a z-distribution, so if you want to do non-directional testing, you have to multiply by 2.)

Alternatively, you could use a computer program, which usually supplies a two-tailed p-value. In such a case, if you required a one-tailed p-value, you would have to calculate it by dividing the two-tailed p-value by 2. Due to wide availability of microcomputers and appropriate statistical software, it is rarely necessary to use tables these days.



The relationship between one-tailed and two-tailed p-values can be summarised as follows:

- One-tailed p-value = (two-tailed p-value) / 2
- Two-tailed p-value = (one-tailed p-value) x 2

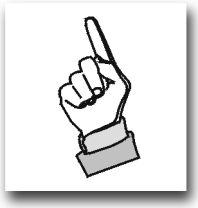
The important point to remember is that the p-value indicates more or less how likely the particular result we have observed in our data is *if the null hypothesis were true*; or, as we say, 'under the null hypothesis'.

3.2.3 The test statistic

In this discussion, we have in fact developed our first statistical test. It is not a very useful test in general, because it will only work if we know beforehand what the population standard deviation (σ) is, which is rarely the case.

When we transform a value such as 104 in this way to an equivalent z-score so that we can use the z-tables to determine the p-value, this z-statistic is referred to as a 'test statistic'. We use special symbols to denote such test statistics. In the present case, we use the symbol $z_{\bar{x}}$, which indicates *the z-test for a single sample mean*.

In the topics that follow throughout the rest of this study guide we deal with several of the more popular test statistics, their underlying assumptions, when they are used, et cetera. One of the tasks a researcher faces is to decide on the appropriate test statistic to use. We can refer to a test statistic as a variable that has a known theoretical probability distribution. In other words, the probabilities of various values for the test statistic can be calculated (as explained in Topic 2), although this usually requires using appropriate computer programs. Examples of test statistics that you will encounter in this course are the $z_{\bar{x}}$, the $t_{\bar{x}}$ and the χ^2, but many others exist for different types of relationships among variables.



Criteria of relevance in choosing the appropriate test statistic usually include one or more of the following:

- The nature of the relationships between the variables that you are trying to investigate
- The statistical distribution of the population
- The size of the sample
- Certain parametric requirements assumed in the test

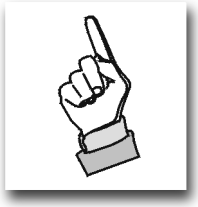
See the diagram in Appendix F for all the test statistics that fall within the scope of the present module and for the symbols we use to represent these test statistics.

STUDY UNIT 3.3
Making a decision regarding the null and alternative hypotheses

3.3.1 Using a decision rule for evaluating statistical hypotheses

We hinted above that small p-values would lead one to reject the null hypothesis, because it shows that the probability of H_0 being true is not very high. But how small must the p-value be? The practice in empirical research is to decide what size p-values would be considered small enough to justify rejecting the null hypothesis *before* the research is actually conducted. We do this by specifying a

'cut-off' p-value so that, if the calculated p-value of our sample result is smaller than this 'cut-off' p-value, the null hypothesis is rejected. This 'cut-off' p-value is called the **significance level** of the statistical test procedure. We will use the symbol 'α' to denote this significance level. The symbol 'α' is pronounced 'alpha' and is the Greek letter equivalent to the normal 'a' in our (Roman) alphabet. By convention, this value is often set at either 0.05 or 0.01. The α-value specifies the maximum risk that we are willing to take of making an error if we reject the null hypothesis (see section 3.3.3 below for more details on this).



The decision rule for H_0 is simply as follows:

> *If the p-value of the sample result is smaller than α (i.e. if the p-value $< \alpha$), the null hypothesis is rejected. If the p-value is not smaller than α (i.e if the p-value $\geqslant \alpha$)*, the null hypothesis is not rejected.

Let us suppose we had set $\alpha = 0.05$ and we are testing H_0: $\mu = 100$ against H_1: $\mu > 100$. Suppose the sample result is 104 and after we calculate the test statistic ($z_{\bar{x}}$), we find that it has a one-tailed p-value of 0.0166 associated with it. This p-value is smaller than 0.05 (in other words: $0.0166 < 0.0500$). Therefore, the null hypothesis (H_0) is rejected and the alternative hypothesis (H_1) is accepted. We refer to this result as *significant* or *statistically significant*.

The expression 'not rejecting the null hypothesis' makes it sound as if we are accepting that the null hypothesis is true, but there is a difference between 'not rejecting the null hypothesis' and accepting it. The reason why we prefer to avoid saying that we 'accept the null hypothesis' is as follows. H_0 is a very precise statement: it says very something specific, like $\mu = 100$. We shall, however, rarely find that our sample mean ($\bar{x}$) is *exactly* equal to 100. If we obtained (for the example above) a sample mean of $\bar{x}=102$ and our statistical test leads to the conclusion that this is not a significant result (i.e. we find $p > \alpha$), it means that this value does not differ from 100 *to a sufficient degree* for us to conclude that H_0 is false. But we have *not* found that 102 is exactly the same as 100 (as the null hypothesis literally suggests). We would say that our result does not enable us to conclude that H_0 is false, or even that the result favours the null hypothesis, but that we cannot accept it is *literally* true. Even if our sample mean is $\bar{x}=100$ exactly, there is a remote possibility that this is a chance event (due to measurement or sampling error). What we do know in such a case is that there is no indication that H_1 can be true and no reason to do a test to confirm this.

Note, however, that the problem does not occur when we look at H_1: if you reject H_0, you can safely say that you accept H_1; and not being able to reject H_0 implies that H_1 is rejected.

Finally, after the researcher has made his or her decisions regarding the statistical hypotheses, he/she must follow this up by drawing a conclusion regarding the research hypothesis. Often, but not always, the research hypothesis translates into the alternative hypothesis so that, when the null hypothesis is rejected in favour of the alternative hypothesis, this usually implies that the research hypothesis has been confirmed. If the null hypothesis could not

be rejected we say that the research hypothesis could not be confirmed (but, remember, we do not accept it!). We do not say that the research hypothesis has been shown to be false. The whole process, thus, starts with a research hypothesis and ends with a conclusion as to whether the research hypothesis has been confirmed or not on the basis of probabilities.

3.3.2 Type I and II errors and the power of a statistical test

The null hypothesis specifies what we can expect to find in the population if no effect exists (i.e., if there is no relationship among the variables). We draw an unbiased sample from the population to measure the effect, but we need to keep in mind that the effect we observe when we take the measurement can be caused either by a 'real' underlying relationship among variables or be the consequence of measurement error (which is partly due to sampling error). We calculate a test statistic that is an indication of how far the observed effect – as reflected in the sample data – deviates from what the null hypothesis tells us to expect (if it were true).

The test statistic is a value with a known probability distribution: we can use it to determine what the probability is of finding an effect of a particular size, which we refer to as the p-value. It is because of our knowledge of the *probability distribution* of the test statistic that we can determine the p-value (this notion of a probability distribution was explained in Topic 2). We compare this p-value with a level of significance (α) that we chose before we did the sampling and made the observation. This is chosen by the researcher, based on the risk of being wrong when rejecting the null hypothesis that he or she is willing to take. If the p-value associated with the test statistic is smaller than this α-value, the null hypothesis is rejected and the alternative hypothesis accepted. If not, the null hypothesis is not rejected.

The p-value represents the probability that the null hypothesis is true: that the effect we see in our observation is due to chance effects like measurement error. If this probability is small, we conclude that H_0 is not true, and we reject it. This p-value is also a direct indication of the probability that the null hypothesis is being *mistakenly* rejected. In other words, it shows the probability that the researcher is rejecting a null hypothesis *that is actually true.* This result may be improbable, given our observations in the data from the sample, but improbable things actually happen all the time.

This kind of mistake – rejecting the null hypothesis when it is in fact true – is referred to as a *Type I error.* Note that we can never know for sure whether a Type I error was actually made when we reject the null hypothesis. This is because we shall never know what the 'true' value of the population parameter (e.g. the mean) is, since it is not practical to study entire populations. The whole point of drawing a sample and doing a statistical test is to make some kind of guess or inference regarding the population parameters that we are interested in. Based on this, however, we can estimate the probability that we shall make an error if we reject the null hypothesis, namely, a Type I error.

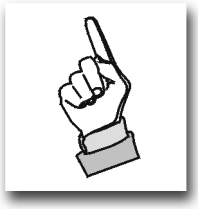
The convention in most research projects is to fix the value of the level of

significance (α) in advance of the actual research, thereby limiting the probability of a Type I error to this value. The value of α represents the *maximum risk* that we are willing to take of making a Type I error by rejecting H_0 in error. So, by setting α in advance, and only rejecting H_0 if the p-value is smaller than α, we are protecting ourselves against the probability of making a Type I error of larger than α. To reiterate: the p-value gives the probability of an error of Type I exactly, whereas α is deliberately chosen by us as the maximum probability of making a Type I error that we are willing to risk. When we find that p-value $< \alpha$, we say that we have found a (statistically) significant result, and because of this we can reject H_0 in favour of H_1.

What if the p-value is *not* smaller than the α-level and we decide not to reject H_0? Now we run the risk of *not* rejecting H_0 when – in fact – H_0 is false and H_1 is true. This is referred to as a *Type II error*. The decision not to reject H_0 is based on the test statistic, from which we determine a p-value that is *not* smaller than our chosen level of significance (i.e., p-value $> \alpha$). However, this outcome may also be the result of measurement error. The effect may be close to what H_0 predicts purely by chance.

This error of failing to reject a null hypothesis that is really false is indicated by the Greek symbol β (pronounced 'beta') which indicates the probability associated with this risk. So while α indicates the risk that a researcher is prepared to take of making a Type I error (rejecting H_0 when the researcher should not do so), β is used to indicate the opposite risk – the risk he (or she) is taking of making a Type II error (*not* rejecting H_0 when in fact he should).

We cannot set the β-value in advance of the research as we do with the α-value. In fact, because H_1 does not specify a specific mean for the population distribution, but a range of possible values, we cannot derive the sampling distribution of the mean under H_1 and we do not calculate the size of β directly. There are ways of determining what it is, but these depend on decisions regarding how sensitive we require our test to be (see the discussion on *effect size* below). Generally, though, the smaller α, the larger β. If we wish to avoid Type I errors, we set α to a small value such as 0.01 or even 0.001, but if we want to avoid Type II errors, we could set α to a larger value.



The relationships between these two types of errors and the hypotheses can be summarised in Table 3.2 below:

TABLE 3.2: The relationships between hypotheses and types of error

	H_0 is not rejected	**H_0 is rejected**
H_0 should not be rejected	*Correct*	*Type I error*
H_0 should be rejected	*Type II error*	*Correct* ('Power' of test)

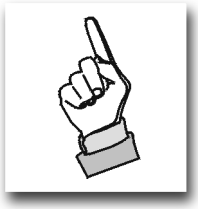
The ability of a statistical test to detect a significant relationship between

variables when such a relationship does in fact exist, is referred to as its *power*. This is the *inverse* of a Type II error: it is the probability of rejecting H_0 when, in fact, it is false and H_1 is true. To put it succinctly, it is the probability of *correctly* rejecting a false null hypothesis (as indicated in the lower right-hand cell of Table 3.2 above).

The *power* of a test is calculated by subtracting the probability of a Type II error from one (i.e., *power = 1 – β*). It can be thought of as a measure of the ''accuracy'' of the text.

The power of a test is related to how sensitive the test should be (see section 3.3.4 on effect size below) as well as the sample size (n) that you are going to use.

In practice, we usually control only the α-level when we use a particular statistical test. But, given a fixed α-level, there are ways of increasing the power of a test even if we do not actually calculate the value of 1 – β.



There are a few things we could do, namely:

- Increase the sample size.
- Decrease error such as sampling error (by means of a careful choice of appropriate sampling techniques), measurement error (by using more reliable tests), or error due to external variables (e.g., by controlling external variables by eliminating or controlling them in the research design). In effect, this results in a smaller standard error of the sampling distribution of the mean.
- A third way to influence the power of our procedures is by our choice of statistical test. In general, parametrical statistical tests (tests based on population parameters or sampling estimates of them: see Topic 1, section 1.4.2) tend to be more powerful than equivalent non-parametric techniques, which is why we prefer to use them when we can.

3.3.3 Effect size

A major determinant of the sensitivity or power of a statistical test is sample size (which is why we can increase sample size to enhance power). When the sample is large, even smaller effects will have statistical significance. The reason is that the larger the sample, the less error variance can be expected (variance purely due to randomness). This is due to a principle called the *law of large numbers*, which states that on average the result obtained from a large number of trials should be close to the expected value, and will tend to become closer as more trials are performed (this law is described in section 2.1.2). This implies that when sample sizes are large, even sample effects that seem insignificant can produce small p-values, leading to the rejection of H_0.

Let us illustrate this with an example.

Suppose, as before, we are testing the IQ of Unisa students in comparison with the general population, with the following hypothesis

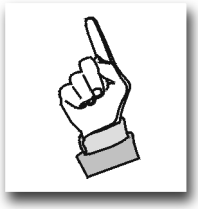
H_0: $\mu = 100$
H_1: $\mu > 100$

Let us assume we find a sample mean of $\bar{x} = 102$. We can calculate the z-statistic ($z_{\bar{x}}$) for each of three different sample sizes (using the same procedure as explained above in section 3.2.2), to see how this affects the p-value.

TABLE 3.3: The relationship between sample size and p-value (for z)

n	$z_{\bar{x}}$	p-value
10	0.422	0.337
100	1.333	0.092
1000	4.219	0.000

So from Table 3.3 we can see that for a difference in IQ of only 2 points from the population mean (100 – 102), we find a highly significant result if we base this on a sample of n=1000.



(Note: p-value = 0.000 implies by convention that there are at least three zeroes after the decimal point; so we can be sure $p < 0.0005$. When we calculated this p-value with a computer, the result was 0.000012.)

This sample mean of 102 does not appear to be an important result from a psychological perspective. We hardly expect that two persons who differ by only two points in their IQs would perform noticeably differently in tasks of a cognitive nature.

The implication is that we have to be careful how we interpret significant results. A p-value of smaller than our chosen level of significance (α) simply implies that, relative to this sample, it is improbable that the effect we see in our observations is purely due to chance. It does not imply that the effect is big or important. This is something that we have to decide by looking at what the data *means*.

One way that statisticians have suggested to deal with this problem is by the notion of *effect size*. Different procedures exist to determine the effect size of a result. In the case of a comparison between means, one way of calculating this is by the use of *Cohen's d*. We do this by expressing the mean difference that we observed relative to the standard deviation:

$$\textit{Effect size} = d = \frac{\textit{mean difference}}{\textit{standard deviation}}$$

A result of $d > 1$ would imply a difference of greater than one standard deviation between the means, which is quite large.

Let us go back to the problem where we were comparing a sample mean of $\bar{x} = 104$ with an expected value of $\mu = 100$ and with a given population standard

deviation of $\sigma = 15$ (see section 3.2.2 above). We would calculate the effect size as follows:

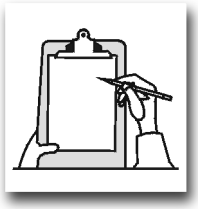
$$d = \frac{\textit{mean difference}}{\textit{standard deviation}} = \frac{\bar{x} - \mu}{\sigma} = \frac{104 - 100}{15} = \frac{4}{15} = 0.267$$

As a rule of thumb we can interpret the effect size as follows:

Around 0.2:	'small
Around 0.5:	'medium'
Around 0.8:	'large'

So, in such a case, based on the sample of n=64, we would conclude that even though Unisa students exceed the general population in IQ to a greater extent than we would expect by chance, the difference is not very large.

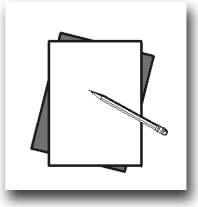
Effect size, power and sample size are interrelated; you can determine one if you have information regarding the other two. For example, if you set your desired effect size and know the power of the test, you can use this to determine what an optimal sample size would be to use the test effectively.



Summary of major points in this topic

Once you understand the logic of statistical hypothesis testing dealt with in the present topic, you should understand and be able to do the following:

- Translate the research hypothesis into statistical hypotheses.
- Define the parameter(s) used in statistical hypotheses. In the first study unit of this topic, the parameter of interest is the population mean.
- Define what is meant by 'level of significance'.
- Identify the appropriate sample statistic. In the present topic, this is the sample mean.
- Define what is meant by the p-value and find the p-value associated with a particular value of the sample mean.
- Distinguish between a sample statistic and a test statistic and convert a sample mean to a z-test statistic.
- Know what a Type I error is, and why can we never know if a Type I error has been committed.
- You should know what a Type II error is.
- You should know what is meant by the power of a statistical test.
- You should know what the effect size of a particular outcome refers to and how it is calculated.
- Once a decision is made concerning the statistical hypothesis, you should be able to draw a conclusion concerning the research hypothesis.



Examples, exercises and solutions

We now revisit the issues dealt with above by working through some examples.

Each of the examples below gives a research scenario that we study within the

framework of what we have learnt so far in Topics 1, 2, and 3. We can summarise the central issues as a number of important steps or stages:

A. Formulate the research or operational hypothesis

Try to formulate this hypothesis so that the following aspects are implied:

- the constructs between which a relation is being postulated (see Topic 1)
- the nature or rule of the relation (see Topic 1)
- the research population (see Topic 1)
- how constructs are being measured
- the research design

B. Translate the research hypothesis into statistical hypotheses and test these on the basis of data from sample(s)

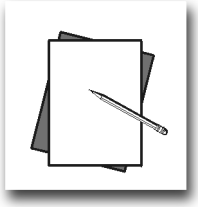
The following steps need to be taken:

Step 1: State the statistical hypotheses and set the value of a.

Step 2: Select a random sample(s) and calculate appropriate statistics. Look at the data from a common-sense, non-statistical point of view. What does it seem to tell you?

Step 3: Select an appropriate test statistic. Calculate the appropriate p-value (e.g. decide if this should be a directional or non-directional p-value) and compare it with the a value. If the p-value is smaller than a, reject H_0 and accept H_1. If not, do not reject H_0.

C. *Draw a conclusion regarding the research/operational hypothesis*



Example 1

Suppose we believe that viewing violence causes one to become more violent. Suppose a test to measure a person's violent behaviour is standardised on a very large student sample considered to be representative of the general student population. The test requires a student to induce an electric shock in a monkey. Students are led to believe that they can adjust the severity of the shock by adjusting the voltage between 0 and 600 on a dial, but do not know that only a very mild constant shock is possible. The severity of shock selected by each student is recorded. The mean for this standardisation group is found to be 300 and the standard deviation 50.

Question 1: Suppose that we plan to subject a single, random sample of students to the viewing of violent material and then subject each student to the test described above to measure his or her violent behaviour. State a research or operational hypothesis.

Answer: A population of students (call this population A) that could potentially be subjected to the viewing of violent material will show more violent behaviour, as measured by the severity of the shock they are prepared to give to the monkey, compared with students who are not subjected to such violent material (call this population B).

Comments: Note that these populations (A and B) as such do not actually exist! We cannot in practice subject all students to the viewing of violent behaviour. But we state the hypothesis in this way to emphasise that the research hypothesis is about some relation between the viewing of violent behaviour and the subsequent tendency towards violent behaviour in a population of students and not in some sample. Of course, we *will* study a random sample of students, but this is to find out about the relation between the two variables in the population.

Also note that this hypothesis implies

- the constructs between which a relation is being postulated, namely, 'viewing of violence' and 'violent behaviour'
- the nature or rule of the relation (i.e. the more the viewing of violence, the more the likelihood of violent behaviour)
- the research population, namely, students
- how constructs are being measured ('Severity of shock' measures constitute the dependent variable.)

Question 2: State the null and alternative statistical hypotheses and set the value of α to 0.05.

Answer: H_0: $\mu = 300$
H_1: $\mu > 300$

Comments: Here μ is the mean 'severity of shock administered' score of a population of students that could potentially be subjected to the viewing of violent material, and 300 is the mean of the population that is not subjected to the viewing of violent material. The value of α is set to 0.05.

Question 3: Suppose we assume that the standard deviation of both these populations is 50 and that the mean of population B is known to be 300. We select a single sample of 100 students randomly and have them view a 5-minute segment of film taken from a movie that shows a man being beaten by a group of gangsters. Thereafter, these students are subjected to the 'test' of violent behaviour. The mean score is calculated and found to be 312. The standard deviation is also calculated and found to be 30. Does the value of 312 appear large? Is there a logical chance that H_1 might be true? Is there anything else about this sample result that should be of concern?

Answer: The sample mean of 312 is larger than 300. This is what we would expect under H_1. So there is a chance that the sample result of 312 might be more likely under H_1 than H_0. The difference is only 12 (312 – 300 = 12), and we are in some doubt whether this is an important difference. We assumed that the population standard deviation is 50, but we find that the sample standard deviation is 30. Is the assumption that the population variance is 50 incorrect?

Comments: It is sometimes difficult to judge the 'practical importance' of a sample result and methods have been developed to assist the researcher in this regard. However, these methods fall outside the scope of the present module. As far as the population standard deviation is concerned, we can perform a statistical test on the hypothesis '$\sigma = 50$', but this test also falls outside the scope of this module.

Question 4: Perform a statistical test procedure on the sample result. Give reasons for your choice of the test statistic.

Answer: The appropriate test statistic is the $z_{\bar{x}}$ test of a single sample mean, because of the following considerations:

- The null hypothesis is about a population mean (μ); the appropriate sample result is, therefore, a sample mean ($\bar{x}$).
- The population distributions are assumed to be normal in shape with a known standard deviation ($\sigma = 50$).

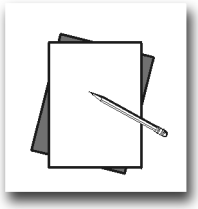
Comments: The z-test statistic is calculated as follows:

$$z_{\bar{x}} = \frac{312 - 300}{\frac{50}{\sqrt{100}}} = \frac{12}{5} = 2.4$$

From the z-tables in Appendix D, the p-value (for the area above $z = 2.4$) is 0,0082. This is smaller than 0,05; therefore, reject H_0 and accept H_1.

Question 5: What is the conclusion regarding the research hypothesis?

Answer: From a statistical point of view, the research hypothesis is confirmed. In other words, the viewing of violent material results in increased violent behaviour among students.



Example 2

Study the AIDS evaluation scenario in Appendix A. Suppose that, after studying various research applications of the 'Attitude to AIDS' questionnaire, the trainers are satisfied that the general population of employees in South Africa has a mean of 20.0 and a standard deviation of 3.5. The trainers are interested in whether the large company from which their sample of 40 was selected has a more negative attitude to AIDS than the general population of workers in South Africa.

Question 1: State the research hypothesis.

Answer: The company's workers have a more negative attitude to AIDS (as measured by the AIDS attitude questionnaire) than the general population of workers in South Africa.

Comments: The research population is clearly 'employees in South Africa'. Note, however, that the two populations associated with the two levels of the independent variable so that the statistical hypotheses may be tested are as follows:

- employees of the specific company (call this population A) that requested the research, and
- employees in all companies in South Africa (population B).

These two populations represent a way of thinking by the statistician in order to test the statistical hypotheses, and must not be confused with the researcher's research population.

Question 2: State the null and alternative statistical hypotheses and set the value of $\alpha = 0.01$.

Answer: H_0: $\mu = 20$
H_1: $\mu < 20$

Here μ is the mean 'Attitude to AIDS' score of all employees in the company, and the mean 'Attitude to AIDS' score for all employees in South Africa is assumed to be 20. The significance level (α) is set to 0.01.

Comments: Note here that we accept that the 'Attitude to AIDS' scores constitute interval-scale measurements. Parameters and their corresponding statistics, such as the mean and standard deviation, are thus appropriate summary values of the distributions involved.

Question 3: Suppose we assume that the standard deviation of the AIDS questionnaire is 3.5 for both statistical population distributions involved and that these distributions are normal in shape. We consider the 40 subjects in Appendix A of this module as a random sample from the company. Go to Appendix A and let your eye roam over the 40 'Attitude to AIDS' scores. Can you compute the mean of these 40 scores? We calculated the mean and found it to be 18.55. The standard deviation was also calculated and found to be 3.47 (see Appendix C for the appropriate formulas). The results are summarised in the following table:

Sample size	Mean	Standard deviation	Minimum score	Maximum score
40	18.55	3.47	11	25

Does the value 18.55 appear to be notably smaller than 20? Is there a logical chance that H_1 might be true? Does the sample standard deviation of 3.47 look likely given the assumption that the population standard deviation is 3.5?

Answer: The number 18.55 is smaller than 20, as we would expect under H_1. There is thus a chance that H_1 might explain the results better than does H_0. The difference does not look noteworthy, as a difference of 20 – 1.55 = 1.45 on a scale that ranges from 11 to 25 does not look like an important difference. The sample

standard deviation of 3.47 does look close to the assumed population standard deviation of 3.5.

Question 4: Perform a statistical test procedure on the sample result. Use a significance level of $\alpha = 0.01$. Give reasons for your choice of the test statistic.

Answer: The appropriate test statistic is the $z_{\bar{x}}$ of a single sample mean, because of the following considerations:

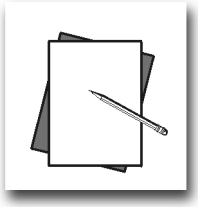
- The null hypothesis is about a population 'mean'; the appropriate sample result is, therefore, a sample 'mean'
- The population distributions are assumed to be normal in shape with a known standard deviation, namely, 3.5.

$$z_{\bar{x}} = \frac{18.55 - 20}{\frac{3.5}{\sqrt{40}}} = \frac{-1.45}{0.55} = -2.64$$

From the z-tables in Appendix D, the p-value for the area under the curve, left of z = –2.64 is 0.0041. This is smaller than 0.01: therefore reject H_0 and accept H_1.

Question 5: What is the conclusion regarding the research hypothesis?

Answer: As the alternative hypothesis is accepted, and this hypothesis corresponds closely to the research hypothesis, we may consider the research hypothesis to be confirmed or verified. It does appear that the employees of this company are more negative to AIDS than the general employee in South Africa.



Multiple-choice questions and solutions

Exercise 1: Study the passage below and then answer questions 1 to 8.

A psychologist hypothesises that sleep deprivation affects cognitive performance negatively. He selects a sample of 16 students randomly and deprives them of sleep for 10 hours over and above the normal 14 hours during which they are awake in a day. He then measures each student's performance on a computer game, which requires cognitive skill to perform. It is also known that the general population of students has a mean score of 1 200 for this particular computer game, with a standard deviation of 200. Suppose it is found that the mean score for the sample is 1 050 and that the level of significance is set at 0.05.

Question 1.1: What is the null hypothesis?

1. The mean score of all students in the computer game is 1 200.
2. The mean score of all students deprived of sleep in the computer game is 1 200.
3. The mean score of 16 students in the computer game is 1 200.

Question 1.2: What is the alternative hypothesis?

1. The mean score of all students in the computer game is less than 1 200.
2. The mean score of all students deprived of sleep in the computer game is less than 1 200.
3. The mean score of all students in the computer game is greater than 1 200.

Question 1.3: Several assumptions about the population of scores have to be made so that the sampling distribution of the mean may be derived. Which one of the following assumptions is **not** made?

1. The shape or form of the distribution is normal.
2. The sampling process is random.
3. The population of scores is known.

Question 1.4: If α is set at 0.05, which of following statements is true?

1. The probability of a Type I error will not exceed 0.05.
2. The p-value of the study cannot exceed 0.05.
3. Both of the above.

Question 1.5: The mean value of the sampling distribution of the mean is

1. 1 200.
2. 1 050.
3. Unknown under H_0.

Question 1.6: What is the value of the standard error?

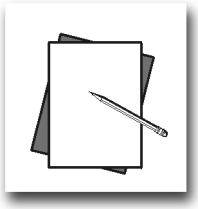
1. 50
2. 200
3. 12.5

Question 1.7: The value of the $z_{\bar{x}}$ -statistic is

1. –0.59.
2. 3.0.
3. –3.0.

Question 1.8: What can you conclude about the hypothesis (stated in 1.1 and 1.2 above), using the value of $z_{\bar{x}}$-calculated in the previous question and testing at the significance level of $\alpha = 0.05$?

1. The null hypothesis cannot be rejected. Sleep deprivation does not affect cognitive performance in a negative way.
2. The null hypothesis should be rejected. Sleep deprivation does not have an effect on cogniive performance.
3. The null hypothesis can be rejected. The results show that sleep deprivation has a negative effect on cognitive performance.



Exercise 2: Study the passage below and then answer questions 1 to 8.

A researcher has developed a measurement of 'attitude to capital punishment' among adults such that the higher a score on the test, the more a person expresses himself in favour of capital punishment or displays a positive attitude to capital punishment. The researcher standardised the test on a large sample of men, and, after several different studies, concluded that the general level of attitude to capital punishment of the general population of men can be assumed to be normal with a mean of 30 and a standard deviation of 11. Forty adult men were then selected randomly and shown a video in which hanging for murder was shown in detail. The test was then administered to the sample and the following results were found: $\bar{x} = 29$ and $s^2 = 100$.

Question 2.1: The research hypothesis states that

1. the sample becomes more negative in its attitude.
2. exposure to capital punishment influences attitudes negatively.
3. exposure to capital punishment influences attitudes.

Question 2.2: The alternative statistical hypothesis is

1. $H_1: \mu > 30$.
2. $H_1: \mu < 30$.
3. $H_1: \mu \neq 30$.

Question 2.3: The mean of the sampling distribution of the mean is

1. 29.
2. 30.
3. unknown.

Question 2.4: The standard deviation of the sampling distribution of the mean is

1. 100.
2. 10.
3. 1.74.

Question 2.5: The value of $z_{\bar{x}}$ is

1. –0.57.
2. –0.45.
3. –0.1.

Question 2.6: Suppose the value of $z_{\bar{x}}$ is found to be 1.5. What is the p-value?

1. 0.0668
2. –0.0668
3. 0.1336

Question 2.7: Suppose the appropriate p-value was found to be 0.07 and the level of significance is 0.10. What should the decisions be regarding the statistical hypotheses?

1. Reject the null hypothesis and accept the alternative hypothesis.

2. Do not reject the null hypothesis.
3. Accept the null hypothesis, but reject the alternative hypothesis.

Question 2.8: Suppose the appropriate p-value was found to be 0.07 and the null hypothesis was rejected because of this p-value. Has a Type I error been committed?

1. No.
2. Yes.
3. It is not possible to say.

Solutions to exercises

Solutions to Exercise 1: Multiple-choice questions 1.1 to 1.8

Question 1.1: The correct answer is option 2. Note that option 1 refers to the population of students not deprived of sleep and merely states what we know of this population. Option 3 is clearly incorrect as hypotheses are never about samples.

Question 1.2: The correct answer is option 2 because the researcher expects sleep deprivation to affect cognitive performance negatively. Option 1 is incorrect because we already know that the population of students in general has a mean of 1 200 and we do not test this. We are testing, however, whether a population of sleep-deprived students would score lower than 1 200. Option 3 is also incorrect as it does not refer to a population of sleep-deprived students.

Question 1.3: The correct answer is option 3. This assumption may be true as far as all those students not deprived of sleep are concerned, but is not true as far as the population of students deprived of sleep is concerned. Option 1 is incorrect since we do actually assume that populations are normally distributed. Option 2 is also incorrect, because we do assume that samples are randomly selected, because then it becomes possible to derive the sampling distributions of sample statistics.

Question 1.4: The correct answer is option 1. The p-value associated with the actual result of the study might be smaller than 0.05, thus indicating a probability of a Type I error smaller than 0.05 but not larger. The actual p-value of the result can of course be any size between 0 and 1. Option 2 is, therefore, incorrect and, therefore, also option 3.

Question 1.5: The correct answer is option 1, because we derive the mean value of the sampling distribution of the mean under the null hypothesis, and because we know that $\mu_{\bar{x}} = \mu$, which is 1 200 under the null hypothesis.

Question 1.6: The correct answer is option 1. The standard error is the standard

deviation of the sampling distribution of the mean. From section 2.4.2 we know the formula is

$$\sigma_{\bar{\chi}} = \frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{16}} = \frac{200}{4} = 50$$

Question 1.7: The correct answer is option 3. This is because

$$z_{\bar{\chi}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{1050 - 1200}{50} = -3.0$$

Question 1.8: The correct answer is option 3. Look up the p-value for $z_{\bar{x}} = -3.0$ in die z-tables in Appendix D. You should find that $p(z < -3.0) = 0.0013$, which is the appropriate p-value for this test. Keep in mind that since the distribution is symmetrical, the 'smaller portion' at the far left (for a negative z-score) will be equivalent to the area at the far right (for a positive score). This p-value is smaller than the level of significance of $\alpha = 0.05$, and it implies that the null hypothesis can be rejected in favour of the alternative hypothesis. In other words, sleep deprivation has a significant negative effect on cognitive performance. Option 1 is false and in option 2 the wrong conclusion is made after the null hypothesis is rejected.

Solutions to Exercise 2: Multiple-choice questions 2.1 to 2.8

Question 2.1: Option 3 provides the clearest expression of the research hypothesis. Option 1 is obviously incorrect because a hypothesis is never about samples. Option 2 is incorrect because there is no explicit indication in the problem scenario that the researcher expects that attitudes will become more negative (even though a person reading the scenario may expect that this is what would happen). We have no option but to state a research hypothesis with an alternative hypothesis that is nondirectional.

Question 2.2: Option 3 is correct because, as shown above, the alternative hypothesis has to be non-directional.

Question 2.3: Option 2 is correct because we know that $\mu_{\bar{x}} = \mu$ and that $\mu = 30$ under the null hypothesis. Option 1 is incorrect because it refers to a sample result and not a population mean.

Question 2.4: The correct answer is option 3. We know that

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{11}{\sqrt{40}} = \frac{11}{6.32} = 1.74$$

Question 2.5: The correct answer is option 1 because

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{29 - 30}{1.74} = -0.57$$

Question 2.6: The correct answer is option 3. The one-tailed p-value is 0.0668

(from Appendix D), but because we have a non-directional alternative hypothesis, the two-tailed p-value is 2 x 0.0668 = 0.1336. Note that option 2 is obviously incorrect since a probability cannot be negative!

Question 2.7: The correct answer is option 1. Note that option 3 is obviously incorrect as the null hypothesis is never accepted. The decision regarding the null hypothesis is to reject it or not to reject it.

Question 2.8: The correct answer is option 3. Because we can never know what the true situation is in the population, we do not know, and will never know, if the decision to reject the null hypothesis has been an error. We can, however, calculate the probability that a Type I error has been made. This is given by the p-value.

TOPIC 4

Statistical hypothesis testing: testing means for a single sample

Quick overview

In this topic we focus on statistical tests that are aimed at testing an hypothesis about a sample mean. This is a special case where a mean based on a single sample is to be compared to a specific value – a population mean that is treated as a given. In the process we introduce the t-test.

Our discussion of single-sample group design has been organised into the following study units:

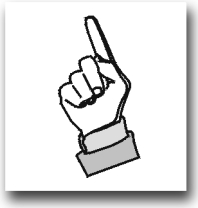
- *Study unit 4.1:* Comparing a single mean to a constant value
- *Study unit 4.2:* Testing a single mean when the population standard deviation is unknown

STUDY UNIT 4.1
Comparing a single mean to a constant value

4.1.1 The general rules for testing an hypothesis

The first three topics of this module were designed to teach the basic principles of testing formal hypotheses through the techniques of inferential statistics. The remaining topics (4–6) are an application of these principles. Each of the remaining topics discusses the appropriate test for specific kinds of statistical hypothesis, depending on what it is we are hoping to establish through our test –

that is to say, which kind of possible relationships among the variables we are investigating. In this course (PYC3704) we only consider tests of relationships of up to two variables at a time; but keep in mind that more advanced methods for testing more complex relationships do exist.



In all statistical testing you will, however, find that the basic underlying procedure remains the same:

- Reformulate a research hypothesis into an appropriate statistical hypothesis (contrasting a null hypothesis with an alternative hypothesis). This shows us what relationship among variables we need to test for.
- Select an appropriate level of significance (α) to test against (i.e. the maximum probability of making an error if we reject the null hypothesis that we are willing to accept).
- Choose a test statistic that would be appropriate, given the type of relationship among variables that was specified in the statistical hypotheses plus certain characteristics of the distribution and population parameters that may be required.
- Draw a sample from the relevant population of data.
- Calculate the required sample statistics that may be required to do the test.
- Calculate the test statistic. This shows the magnitude of the observed relationship relative to the expected relationship if the null hypothesis were true (i.e. 'under the null hypothesis').
- Determine the p-value, which tells you what the probability of this observed relationship (indicated by the test statistic) would be under the null hypothesis.
- If the p-value is small, that is to say, less than the chosen level of significance, reject the null hypothesis – or else, do not reject it.
- Relate the results to the research hypothesis to come to a conclusion.

4.1.2 Statistical tests for means: a single-sample test to use when σ is known

In Topic 3 (section 3.2.2), in the process of explaining the logic of statistical testing in general, we introduced you to the $z_{\bar{x}}$ test for single-sample comparisons. This is used when you have only one sample of data of a variable from which a mean could be derived, and you want to compare this mean with a specific constant value.

We said before (in Topic 1) that inferential statistics refers to the comparison of two or more variables. In a case like this, only one variable seems to be involved. If we are interested in comparing means, there should be at least two of them (we do not consider the possibility of more than two groups being compared in this course, although specialised tests exist for this).

When we compare means, the dependent variable – the variable that we are investigating – would be a measurement of some kind of at least an interval level of measurement (see Appendix B in this regard). It seems as though another variable should exist, an independent variable to divide the population into (at least) two groups. If there is only one sample, where is this second

variable, that is, the nominal scale variable that divides the population into categories or groups? As a matter of fact, there is a second variable implied, because the sample group is compared with the population, which implies two groups being compared. It is just that the population mean being tested is known beforehand, so it is not necessary to draw a second sample to find an estimate for it.

Another way to think about this is in terms of the research design. Here you have a single sample, from which a sample mean is calculated, and the type of design can be referred to as a *single-sample groups design* (because there is one group being sampled).

Let us consider another example of the process of testing a mean derived from a single sample ($\bar{x}$) against a population mean (μ) which is known, in those situations where the population standard deviation (σ) is known beforehand.

During a literature search a researcher comes across a questionnaire which, according to the author, can be used to measure the construct 'general optimism'. With the questionnaire is a set of norm values that can be used to convert the score derived from a combination of the questionnaire items to a so-called 'stanine' scale. This is a 9-point scale standardised in such a way that by using a table of norms we should obtain a population mean of 5 and a population standard deviation set to 1.96. We are going to test the claim that the norms do in fact provide a scale with a mean of 5.

First, we establish the research question:

Are the results of the 'general optimism' scale for members of the general population centred around the scale mean of 5, based on the published norm table?

This can now be converted into a statistical hypothesis, by writing it in symbolic form:

H_0: $\mu = 5$
H_1: $\mu \neq 5$

Note that we want to know if the score deviates from 5 but we have no preference for the *direction* in which it differs, which is why we need to do two-tailed or non-directional testing. We are also interested less in the alternative hypothesis than in checking the validity of the null hypothesis. This is a perfectly acceptable goal, and it has no real effect on the procedure for conducting the statistical test; only on how we deal with the results.

Let us assume we drew a random sample of persons of size $n = 50$ and asked them to complete the 'general optimism' scale. After collecting the results, we used the norm tables to convert each of the results to a stanine value. We find that the sample mean of these results is $\bar{x} = 4.50$.

Because we know what the population standard deviation should be ($\sigma = 1.96$) we can proceed to do the $z_{\bar{x}}$ test. To do this, we set the level of significance to $\alpha = 0.05$. As explained in Topic 3, this represents the *maximum* probability of making

an error if we rejected H_0 that we would be willing to consider before proceeding to reject the null hypothesis. Note that setting this level of significance is something that we must always do *before* proceeding to do the test.

Something else to be on guard about is that certain conditions need to be met before we may use this test: we assume that the distribution of scores in the population from which the sample is drawn has a standard deviation that is close to σ, and we assume that this population is normally distributed. This assumption can be relaxed if the sample size is greater than about 30 (i.e. $n > 30$).

As we saw in Topic 3 (section 3.2.2), the formula for the $z_{\bar{x}}$ test statistic is

$$z_{\bar{x}} = \frac{(\bar{x} - \mu)}{\sigma_{\bar{x}}} = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

Remember that we are testing a mean, and $\sigma_{\bar{x}}$ is the standard deviation of the distribution of means, or standard error. Replacing the symbols with the relevant variables produces

$$z_{\bar{x}} = \frac{(4.5 - 5)}{\frac{1.96}{\sqrt{50}}} = \frac{-0.5}{\frac{1.96}{7.071}} = \frac{-0.5}{0.277} = -1.805$$

To find out whether this result is significant, we need to consult the z-tables (in Appendix D). Since the tables only give results for positive values of the z-value, we can ignore the sign. (In mathematical terms, we use the *absolute value* of –1.805, symbolised by $|-1.805| = 1.805$. (See Appendix E.)

We are interested in the regions right at the ends of the distribution (referred to as the 'smaller area' in the table). So, according to the table, if $z_{\bar{x}}$=1.805 we shall find a p-value of 0.035. Note, however, that the tables imply that we are doing one-tailed or directional testing, so to get the non-directional value we need to multiply by 2, which produces a two-tailed p-value of 2 x 0.035 = 0.070 (see Topic 3, section 3.2.2).

We, therefore, find that the p-value is *not* less than the level of significance of $\alpha = 0.05$, which implies that we *cannot* reject the null hypothesis in favour of the alternative hypothesis. Relating this to the research hypothesis, we conclude that the scale does seem to be evenly distributed around the theoretical scale mean of 5, as the author of the questionnaire claimed.

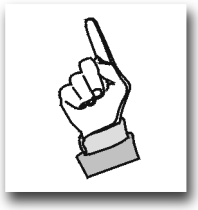
STUDY UNIT 4.2
Testing a single mean when the population standard deviation is unknown

The use of a test based on the z-distribution is unfortunately limited to cases where the population standard deviation (σ) is known. It is actually quite rare for this standard deviation to be known. When this population parameter is not

available to a researcher, the test statistic cannot be calculated, and consequently the $z_{\bar{x}}$ test is not often used in practice.

4.2.1 Introducing the t-test

So what should we do if σ is not known? We have to find some way to estimate it. The obvious thing would be to use the sample variance (s^2) as an estimate for the population variance (σ^2). In practice, it was found that this leads to a result that is more likely to underestimate σ^2 than to overestimate it. The resulting z-value would be a bit larger than the one we would get if one used the true population variance (making a Type I error more likely). To compensate for this, a somewhat altered distribution was developed from the z-distribution, which became known as the *t-distribution*.



The important point is that – as in the case of the z-distribution – the t-distribution is a statistical distribution with a probability distribution that can be determined, which means that we can use it to predict the chances of obtaining specific outcomes when testing for comparisons of means when the population standard deviation σ is unknown.

More precisely, the t-distribution refers to a range of possible distributions determined by the *degrees of freedom* implied in the research design. Degrees of freedom (*df*) relate to the number of independent pieces of information that remains unknown once we have estimated one or more parameters of the population. As the sample size increases, the t-distribution moves closer to the z-distribution.

In t-tests for one sample, the degrees of freedom are equal to n-1. We mention this because if you decide to do t-tests by hand and look up the relevant p-value in a table, you will need the df to be able to find the relevant p-value for a chosen level of significance (α). Computer programs would calculate this value directly.

The t-distribution is used in a variety of test statistics, mostly related to testing aspects of means, and it is a very common test that is available in most statistic packages (including the widely available *Microsoft Excel* program).

We explain the use of this test in the special case where a mean is to be compared to a specific constant value, but without knowing what the population standard deviation is, by way of example.

Suppose we are interested in the effect of anxiety on the performance of PYC3704 students in the exam. We want to investigate the hypothesis that students with high anxiety do less well in exams than the population of PYC3704 students in general. We happen to know from experience over many years that the general population of PYC3704 students obtains a mean exam result of 60%. We are confident enough to accept that this 60% mean is an accurate reflection of the underlying population mean, and have reason to believe that the results are normally distributed around this mean, but we do not know the population standard deviation. We assume that the data from which the sample

was drawn is normally distributed, which is a requirement of this test (this assumption can be relaxed for a large sample size n).

We now have to draw a sample of students who come from a population of students who display 'high anxiety'. To do this, we first have to specify what we mean by 'high anxiety'. One way of doing this would be to test all students in PYC3704 and obtain anxiety scores for them. We then use some operational definition of 'high anxiety' to identify a group of students that can be regarded as individuals having 'high anxiety'. For example, we may conclude that students with a score of seven or higher on our nine-point anxiety scale suffer from 'high anxiety'. These students then form the population from which our sample will be drawn.

We can now formulate the following statistical hypothesis:

H_0: $\mu = 60$ in the exam
H_1: $\mu < 60$ in the exam

Here μ represents the mean of the population distribution of examination scores of PYC3704 students with high anxiety. This value of μ is to be compared with the mean examination scores of all PYC3704 students, which is represented by a constant of 60. So you can see that there are really two variables being compared: 'examination score', and 'anxiety group membership' (with two levels: students with high anxiety vs. students in general).

Note that the null hypothesis states that the high-anxiety group will not achieve a mean exam score that differs significantly from that of PYC3704 students in general (i.e., it will not differ from 60%), while the alternative hypothesis is a *directional* hypothesis, asserting that high-anxiety students will achieve an expected score that is *significantly less* than the mean percentage for the general population (which we treat as a constant population value that is given).

Another thing that we want to point out is that this population constant of 60 represents the expected population *mean*, but the fact that this is a mean that represents a score out of a possible total of 100 (i.e. it is a percentage) is not really relevant here. Think of it as a population mean out of a measurement scale that just happens to vary from a possible minimum of 0 to a possible maximum of 100. We therefore omit the percentage sign for the rest of this discussion.

We also decide at this stage to test at a significance level of $\alpha = 0.05$.

Our problem at this point is that the population standard deviation σ (and, therefore, the population variance σ^2) is unknown. As suggested above, we can, however, use s as an estimate for σ. This leads to a one-sample t-test, as we explained above. The formula for this t-statistic is

$$t_{\bar{x}} = \frac{(\bar{x} - \mu)}{s_{\bar{x}}}$$

Look at the formula for the one sample z-test (in section 3.2.2): this is the same formula, but with $s_{\bar{x}}$ replacing $\sigma_{\bar{x}}$. As before (in section 2.4.2), this is the standard

deviation of the distribution of the means (or *standard error*), which we can calculate using the central limit theorem:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Substituting this in the formula for $t_{\bar{x}}$, we also express $t_{\bar{x}}$ as follows:

$$t_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Suppose we now draw a random sample of size n=25, and we find that the mean of this sample is an exam score of 55% (i.e. $\bar{x} = 55$) with a sample standard deviation of 15 (i.e., s = 15). From this we calculate the standard error as

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

Clearly, the 'observed value' of 55 is smaller than the 'expected value' of 60. This is a difference in the correct direction, which fits with our stated alternative hypothesis. The question is, however, whether this difference between the scores of 55 and 60 *is statistically significant*, or likely to be just due to chance. Only if it is significant should we have the confidence to reject the null hypothesis and accept that high-anxiety students do in fact perform less well than PYC3704 students in general.

We can use the t-statistic to determine this, using the values given above.

$$t_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{55 - 60}{\frac{15}{\sqrt{25}}} = \frac{-5}{\frac{15}{5}} = \frac{-5}{3} = -1.667$$

A result of $t_{\bar{x}} = -1.667$ with a sample size of n = 25 provides a p-value of p = 0.05496. (We calculated the p-value using a computer program.) Remember also that we should do a *directional* t-test (testing H_1: $\mu <$ our population mean). Computer programs often only provide the p-value for non-directional testing (i.e., for the two-tailed t-test). In this case, the non-directional p-value would be p = 0.10992, which should be divided by two to get the one-tailed value of p = 0.05496 (we could round this off as p = 0.055).

Since this value of p is *not* less than the significance level of $\alpha = 0.05$, which we have chosen before, we *cannot reject the null hypothesis in favour of the alternative hypothesis*, and our conclusion is that the sample result does not support the alternative hypothesis. Given that we are not willing to make an error of more than 5% ($\alpha = 0.05$) of rejecting the null hypothesis in error (the Type I error), the sample score of 55 – although obviously smaller than the given population mean of 60 – is not *sufficiently* smaller to warrant the conclusion that the difference is not merely due to sampling or measurement error. This leads us to conclude that PYC3704 students with high anxiety scores do not perform significantly worse in their exams than PYC3704 students in general.

A good way to think about the procedure for deciding whether or not to reject the

null hypothesis is to state it in terms of what is known as the *decision rule* (see also Topic 3, section 3.3.1). For example:

> If the p-value is equal to or less than the chosen significance level α, reject H_0 in favour of H_1. Otherwise, do not reject the null hypothesis in favour of the alternative hypothesis.

In our example, the application of the decision rule would read as follows:

> *Because $p = 0.055$ is not equal to or less than $\alpha = 0.05000$, do not reject the null hypothesis in favour of the alternative hypothesis.*

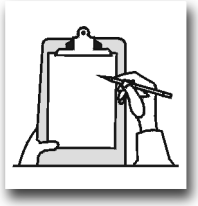
The decision rule above is used in most of the solutions to exercises in this study guide. While it can be applied without problems to two-tailed tests, you should be careful in the case of a one-tailed test. Make sure your interpretation of the results actually makes sense, relative to the alternative hypothesis. If you set up a one-tailed alternative hypothesis like the one above (H_1: $\mu < 60\%$ in the exam), and you find that the computed sample mean is $\bar{x} = 65$, then this alternative hypothesis cannot possibly be true (65 is never smaller than 60, so there is no sense in asking if it is so much smaller that it is significant!). In such a case, you know beforehand that you cannot reject the null hypothesis.

Of course, the same situation holds in the opposite direction: if you set up an alternative hypothesis like H_1: $> 60\%$ in the exam, and you find a sample mean of 60 or below, you cannot conclude that this sample mean probably comes from a population that favours the alternative hypothesis.

Another point to keep in mind when you are doing one-tailed testing, as was mentioned before, is to make sure that the computed p-value is the correct one-tailed value. If the two-tailed value has been given by the computer program, you will need to divide it by two.

If you look through some statistics handbooks, you will find that other methods also exist for making the decision of whether to reject H_0 or not. For example, if you use tables, you could proceed by determining a *critical value*. This implies finding the value which – for example – the t-statistic must *exceed* if you want to be able to conclude that the p-value is less than a particular level of significance (α). Note that the bigger the t-value the greater the likelihood of rejecting H_0 (as is the case with z-statistics), because it refers to how far the observed value of the sample statistic differs from the population parameter that was provided and refers to the areas on the edges of the distribution (as explained in Topics 2 and 3).

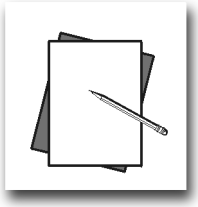
Another method is to use *confidence intervals.* In this case, you evaluate a sample result relative to what is known as a 'rejection region', which *surrounds* the population parameter. We mention this because this technique is often recommended by some statisticians who fear that the artificially exact p-values provided by computers give a false sense of certainty. We do not discuss this method here. We trust you understand that probabilities are, by definition, not exact values, and the *exact* p-value does not tell you very much.



Summary of major points in this topic

After studying this topic, you should be able to

- calculate and use the z-statistic for a single-groups design ($z_{\bar{x}}$) to test an hypothesis when the population standard deviation (σ) is known.
- calculate and use the t-statistic for a single-groups design ($t_{\bar{x}}$) to test an hypothesis when the population standard deviation (σ) is unknown.
- determine whether or not to reject the null hypothesis using a decision rule.



Exercises and solutions

Here are some exercises you can do to test your knowledge of this topic. We suggest that you do the exercises before looking at the answers. Note that some of the exercises refer to the AIDS evaluation scenario given in Appendix A.

> Now for some good news: we do not expect you to memorise the equations for calculating test statistics such as $t_{\bar{x}}$ or $z_{\bar{x}}$, nor will you have to perform such calculations in the exam. What is important is the principles underlying the use of these statistics: how and why would you want to use them. We do, however, feel that working through these exercises is important for a proper understanding of these principles, so please do them. You will also get some questions that require calculations in some of your assignments.

Multiple-choice questions and solutions

1. A researcher hypothesises that a population mean for the variable 'aggression' is significantly different from some known specific value. She draws a sample from an identified population of subjects and computes the sample mean so that she can use the $t_{\bar{x}}$ statistic to test her hypothesis. What type of comparison is implied in this procedure?

    1. Comparison of sample means.
    2. Single sample with σ known.
    3. Single sample with σ unknown.

2. For what reason would we calculate the $t_{\bar{x}}$ test statistic? It is used to determine whether a ...

    1. population parameter such as a mean differes significantly from a given value
    2. relationship exists between a population parameter and a sample parameter.
    3. sample mean is distributed randomly.

3. Consider the following from the case study presented in Appendix A: The mean pre-knowledge score for workshop participants was 6.05, and the standard deviation 1.79. Use this to estimate the standard error (which is the standard deviation of the sampling distribution of the mean).

1. 0.400
2. 0.283
3. Not enough information.

4. Referring to the AIDS evaluation scenario (Appendix A), suppose that the mean population pre-knowledge score concerning HIV/AIDS is 5.8. Which is the correct t-statistic for the data in question 3?

   1. 0.1396
   2. –0.625
   3. 0.625

5. Consider the following from the case study (Appendix A): The mean post-knowledge score for workshop participants was 10.65 and the standard deviation 3.199. What is the standard deviation for the sampling distribution of the mean?

   1. 0.506
   2. 0.715
   3. 0.159

6. Referring to the case study regarding AIDS (Appendix A), suppose that the mean population post-knowledge score concerning HIV/AIDS is 9.4. Which is the correct t-statistic for the data in question 5?

   1. 1.748
   2. –1.748
   3. 0.391

7. Assume that we are testing the following hypothesis: Individuals who attend workshops on HIV/AIDS have a higher level of knowledge about the disease than the population at large.

   Suppose we know the knowledge level score for the general population (from previously compiled norms), and we find a t-statistic of 1.71, which leads to a p-value of 0.0518. Which interpretation is correct?

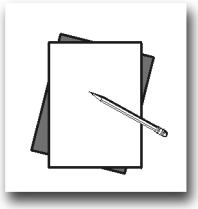
   Assume $\alpha = 0.05$

   1. Cannot reject the null hypothesis on a 5% level of significance. Individuals who attend workshops on HIV/AIDS do not have different levels of knowledge about the disease when compared to the general population.
   2. Reject the null hypothesis in favour of the alternative hypothesis on a 5% level of significance. Individuals who attend workshops on HIV/AIDS have a higher level of knowledge about the disease than the general population.
   3. Not enough information is given to decide whether or not the null hypothesis should be rejected for any given level of significance.

8. When the results of a t-test were entered into a computer program to calculate the p-value, a one tailed result of $p = 0.042$ was returned for a test of an alternative hypothesis that states H_1: $\mu_1 \neq \mu_2$.

   What can we conclude if we choose a significance level of $\alpha = 0.05$?

   1. Reject the null hypothesis in favour of the alternative hypothesis.
   2. Do not reject the null hypothesis.
   3. Reject the alternative hypothesis in favour of the null hypothesis.



Solutions to multiple-choice questions

1. Correct answer: Option 3. Only one sample was drawn from a single population of subjects, to be compared with a mean that was known (so we do not need a second sample) and with σ unknown. If we knew the population standard deviation (σ), we could use a $z_{\bar{x}}$ test.
2. Correct answer: Option 1. Option 2 is wrong because the sample parameter is used to estimate the population parameter, but this is not what is compared in the test. Option 3 is wrong because, although a random sample is a precondition for doing the test, this is not what is being tested.
3. Correct answer: Option 1. The standard deviation of the sampling distribution of the mean is estimated by using the following equation:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where n is the size of your sample and s is the standard deviation of the sample (the formula is explained in Topic 4.2.1). In this case, the sample size is 20 because 20 participants attended the workshop.

Therefore:

$$s_{\bar{x}} = \frac{1.79}{\sqrt{20}} = \frac{1.79}{4.472} = 0.4$$

4. Correct answer: Option 3. The t-statistic is calculated using the following formula:

$$t_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}} = \frac{6.05 - 5.8}{0.4} = \frac{0.25}{0.4} = 0.625$$

that is, the mean pre-knowledge score for the sample minus the mean pre-knowledge score for the sampling distribution of the mean, divided by the standard deviation of the sampling distribution of the mean that you calculated in question 3.
5. Correct answer: Option 2. The standard deviation of the sample distribution of the mean can be estimated by using the following equation:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where n is the size of your sample and s is the standard deviation of the sample. In this case the sample size is 20 because 20 participants attended the workshop. Therefore:

$$s_{\bar{x}} = \frac{3.199}{\sqrt{20}} = \frac{3.199}{4.472} = 0.715$$

6. Correct answer: Option 1. The t-statistic is calculated using the following formula:

$$t_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}} = \frac{10.65 - 9.4}{0.715} = \frac{1.25}{0.715} = 1.748$$

That is, the mean post-knowledge score for the sample minus the

mean post-knowledge score for the sampling distribution of the mean divided by the standard error that you calculated in question 5.

If you subtracted the mean post-knowledge score for the sample from the mean post-knowledge score of the population, you would have arrived at answer 2.

7. Correct answer: Option 1 is correct. The application of the decision rule should read:

   Because 0.0518 is not equal to or less than the chosen significance level of 0.0500, do not reject the null hypothesis in favour of the alternative hypothesis.

8. Correct answer: Option 2. We cannot reject the null hypothesis. Remember that when you receive a one-sided p-value for a two-sided test, the p-value should be multiplied by two. Now 2 x 0.042 = 0.084, which is larger than our chosen level of significance of $\alpha = 0.05$.

TOPIC 5

Statistical hypothesis testing: comparing two samples

Quick overview

This topic will consider the situation where two sample means are to be compared. The difference between independent samples and dependent samples and how this affects the testing of hypotheses will be explained.

The present topic is organised into the following study units:

- *Study unit 5.1:* Testing for differences between the means of two independent groups
- *Study unit 5.2:* Testing for differences between the means of two dependent groups
- *Study unit 5.3:* Using differences scores to compare two independent groups

The topic ends with a number of multiple-choice questions with solutions.

STUDY UNIT 5.1
Testing for differences between the means of two independent groups

In this topic, we concern ourselves with the comparison of two population means from two groups, which we can indicate with μ_1 and μ_2. Before we come to the actual tests, we first need to consider the difference between independent and dependent groups.

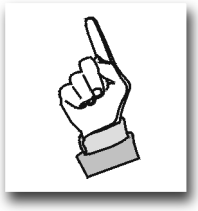
5.1.1 *Independent versus dependent groups*

Samples are considered as comprising *independent groups* if the composition of the one sample in no way affects, in any systematic way, the composition of the other sample. The two samples come from two groups that have no obvious relationship. For example, where one sample is measurements of a construct like 'self-esteem' among men, and the other among women, but both groups were sampled purely randomly.

On the other hand, the concept of *dependent groups* refers to situations where the samples are related, and it implies that each subject in one group can be systematically paired off with a subject from the other group. For this reason, a dependent groups research design is often referred to as a *matched-pairs* design.

Sometimes dependent samples are produced when the researcher deliberately matches subjects into pairs, based on the value of some hidden or 'nuisance' variable. For example, a researcher wishes to establish whether male students do better in research methodology than female students. He would, however, like to ensure that the two groups are equal regarding intelligence, and decides to 'match' the samples in terms of intelligence. For every male student with a specific IQ score in one group, the researcher deliberately selects a female student with the same IQ for the second group. Another example of such a design would be a repeated measures design, where the same research participant is observed under more than one treatment or experimental condition. For example, to test the effectiveness of a psychotherapy technique, people can be tested *before* the treatment begins, and again *afterwards*. The two sets of measurement (indicated by two variables) can be regarded as two samples of data, which is to be compared to see whether some kind of change has taken place. Dependent samples are also sometimes referred to as *correlated* samples (see Topic 6 for more on the meaning of this term).

An implication of this is that dependent samples do not have to be the same size (but they can be), while dependents samples have to be of the same size, as each individual measurement from one sample should be matched to an individual in the other sample.



Make sure that you do not confuse the notion of dependent versus independent *samples* with the distinction between dependent and independent *variables* (Topic 1, section 1.3.2). While the latter refers to the relationships among variables – how one may affect the other – in the case of samples it is a relationship among the *groups* from which the data were collected (i.e., where the variables were measured) that is of concern.

5.1.2 *The t-test (t_c) for differences between the means of two independent groups*

We first consider the situation where we want to compare the means from two independent groups, which comes from two samples that were drawn independently. The goal of the test that follows is actually to determine

whether the two samples come from a single population: in other words, whether the grouping variable (independent variable) affects the measurement (dependent variable) in any way.

By now you will have examined the data from the AIDS evaluation scenario (see Appendix A of this study guide) from a number of different angles. However, in this study unit for the first time we really consider the central question posed in the scenario, namely

Was the AIDS training successful or not?

You will recall that there were two samples of employees in the scenario – those who attended an AIDS training workshop and those who did not.

Read the AIDS scenario in Appendix A again and note the following aspects:

- Two random samples of 20 employees each were selected from the population 'all employees in the company'.
- Group 1 serves as the treatment group and group 2 as the control group.
- Both groups were tested before and after the workshop on some of the tests.
- On some of the variables, scores were obtained for both samples only *before* the workshop.

The question, 'Was the AIDS training successful or not?' can be asked with regard to each of two variables, namely

- knowledge about AIDS – did the training increase knowledge about AIDS?
- attitude to AIDS – did the training make employees more positive in their attitude to AIDS?

We concentrate for the time being on the 'knowledge of AIDS' variable. The research question is

Did the training increase employees' knowledge about AIDS?

We begin by stating the following hypothesis:

Training increases employees' knowledge of AIDS in comparison to those employees who do not receive the training

Note that we can infer the following information from this hypothesis:

The independent variable has two levels (or values), namely 'those who receive training' (call this population 1) and 'those who do not receive training' (call this population 2). These two levels/categories/values imply two populations that need to be compared with regard to their scores on the dependent variable.

The easiest way to test this hypothesis is to begin by assuming that the two groups had the same 'knowledge of AIDS' before training. We now need only show that the group that received training scored significantly higher on the 'post-training knowledge of AIDS' variable than the group that received no training. We are now ready to state statistical hypotheses and choose the level of significance (the α-level):

H_0: $\mu_1 = \mu_2$ (which can also be written as $\mu_1 - \mu_2 = 0$)
H_1: $\mu_1 > \mu_2$ (which can also be written as $\mu_1 - \mu_2 > 0$)

Here μ_1 is the mean of the post-training 'knowledge of AIDS' distribution of scores (call it population distribution 1) of employees who receive training (i.e., the treatment or treatment group) and μ_2 the mean of the post-training 'knowledge of AIDS' distribution of scores (call it population distribution 2) of employees who do not receive training (the control group). In other words: μ_1 is the mean of population distribution 1 and μ_2 is the mean of population distribution 2. We decide to test this against the significance level $\alpha = 0.01$.

To develop a test for difference between two means ($\mu_1 - \mu_2$) we need to know something about the statistical distribution. Statisticians have determined that the distribution of the difference between two normally distributed variables also produces a normally distributed variable.

Furthermore, they have found that as long as the two standard deviations (of the two groups being compared) do not differ significantly, we can estimate the standard deviation of the pooled means ($\sigma_{\overline{\chi_1} - \overline{\chi_2}}$) as follows:

$$\sigma_{\overline{\chi_1} - \overline{\chi_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

However, because we do not know the population standard deviations (σ_1 and σ_2), we need to estimate them with the sample standard deviations (s_1 and s_2). The standard deviation of the pooled means then becomes

$$s_{\overline{\chi_1} - \overline{\chi_2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

As before (section 4.2.1), this substitution leads, not to a z-distribution, but to a t-distribution. We can, therefore, specify the t-statistic for testing the difference between two means (which we shall indicate with t_c) with the following formula:

$$t_c = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(Note: using the subscript 'c' in the t_c test is just a convention, to distinguish the test for independent samples from other uses of the t-test).

Note that the term in the numerator (the top part of the fraction in the equation) should really be ($\bar{x}_1$-$\bar{x}_2$) – ($\mu_1 - \mu_2$) because we are comparing how far the difference we see in our sample data ($\bar{x}_1 - \bar{x}_2$) deviates from the difference predicted in the null hypothesis ($\mu_1 - \mu_2$), but because this would be equal to zero (H_0 can be written as $\mu_1 - \mu_2 = 0$), we do not include this term in the formula.

Let us now return to our example, where we want to test the research hypothesis '*training increases employees' knowledge of AIDS (group 1) in comparison to those employees who do not receive the training (group 2)*, which leads to our statistical hypothesis

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$

In order to use the t-test (t_c) statistic, we need to make two assumptions regarding the data: that the two populations being compared are *normally distributed* with the *same variance* (or standard deviation). (Remember that the square root of the variance is equal to the standard deviation.) We can also assume that the samples are independent – since the samples were selected randomly, we can safely consider them to be independent of each other. All of this makes the t_c-test an appropriate test.

The next step is to actually calculate the means and standard deviations of the two samples with regard to post-training 'knowledge of AIDS' scores. Table 5.1 below gives some statistics regarding the two samples.

TABLE 5.1: Descriptive statistics for the variable, post-training 'knowledge of AIDS'

Group	Sample size (n)	Mean	Standard deviation	Minimum	Maximum
1	20	10.65	3.20	4.0	15.0
2	20	6.15	3.18	4.0	15.0

The first task of the researcher is to look at the data. The sample result of interest here is the difference between the two sample means, namely $\bar{x}_1 - \bar{x}_2$, which is equal to $10.65 - 6.15 = 4.5$.

To us, this looks like an impressive or notable difference! We also note that group 1 has a mean greater than that of group 2 – as predicted by H_1 above. We now proceed to test whether this result of 4.5 is statistically significant.

We begin by calculating t_c, substituting with $\bar{x}_1$ and $\bar{x}_2$ as the means of the two groups, s_1 and s_2 as their standard deviations, and n_1 and n_2 as the sample sizes.

$$t_C = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$= \frac{(10.65 - 6.15)}{\sqrt{\frac{(3.20)^2}{20} + \frac{(3.18)^2}{20}}}$$

$$= \frac{(10.65 - 6.15)}{\sqrt{\frac{10.24}{20} + \frac{10.11}{20}}}$$

$$= \frac{4.50}{\sqrt{\frac{20.35}{20}}} = \frac{4.50}{\sqrt{1.0175}}$$

$$= \frac{4.50}{1.0087} = 4.4612$$

Note that the two sample sizes need not be the same for this test, although if they differ a great deal an adjustment to the formula for the t_c-test statistic is advised. (We do not discuss this requirement in any detail.)

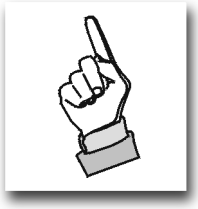
We used a computer program to compute the directional p-value associated with this value of $t_c = 5{,}54$, and found p-value = 0.00007. This p-value is clearly smaller than the level of significance, which we set at α=0.01. Note that a computer will often only provide the p-value for a two-tailed test. In such a case, you have to divide the given value with 2 to set the one-tailed p-value (see Topic 3.2.2).

The null hypothesis (H_0) is, therefore, rejected and the alternative hypothesis (H_1) accepted. The research hypothesis is, therefore, confirmed. We have managed to show that a three-day workshop, where training is based on experiential learning, increases employees' knowledge of AIDS.

Note that the mean difference we saw in our sample was in the right direction as specified in our alternative hypothesis. If it had differed in the wrong direction (i.e., $\bar{x}_1 < \bar{x}_2$), there would have been no need to continue with the test. (See, for example, Topic 3, section 3.2.1.)

Note also that if the computer program provides the probability in the form 'p-value = 0.000,' it means that the result is valid to at least three (3) decimal places (i.e. one can assume that p-value < 0.0005).

Tables for the t-distribution exist and are available in most statistics handbooks. Due to the general availability of computers these days, we do not use this method to find p-values. If you do want to use tables, you will have to calculate the degrees of freedom, which for this test would be df= (n_1+n_2-2).



Note: Even the most elementary statistics program makes provision for performing t-tests. Such programs usually require that we indicate which variable should be used to identify the two groups and which is the dependent variable. In addition, we have to choose between a t_c test for independent samples or a t_d test for dependent or correlated groups (the t_d test is explained in study unit 5.2 below).

Based on the p-value, the difference between the means seems quite impressive. Remember, however, that the t-value is sensitive to sample size (for a larger sample a smaller effect would be significant; see Topic 3, section 3.3.3). To evaluate our effect, let's do an effect size calculation:

In this case calculating Cohen's d for the effect size is

$$d = \frac{\textit{estimated mean difference}}{\textit{estimated standard deviation}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

Here s_p represents the *pooled* standard deviation from the sample (calculated from all data, irrespective of group membership). We calculated this pooled standard deviation to be s_p= 3.888.

The effect size can, therefore, be calculated as

$$d = \frac{10.65 - 6.15}{3.888} = \frac{4.50}{3.888} = 1.157$$

This is an effect of more than one standard deviation, so this difference is quite large (see Table 3.3 in section 3.3.3 for the interpretation of Cohen's d).

Although in this text we are concentrating on the 'AIDS study' in Appendix A, this general scenario is not at all unusual – we very often want to know if people who have been exposed to an intervention (e.g. a workshop, a psychotherapeutic technique, medication or a particular teaching method) differ in respect of some or other outcome variable (e.g. knowledge, insight, blood pressure, stress, examination performance or attitude to capital punishment) from those who have not been exposed to the intervention. If those who have undergone the intervention do indeed show a significant difference, then we have shown that the intervention has had an effect.

Other examples of such comparisons between the mean scores on a variable that was measured for two groups are the following:

- If the group that attended the AIDS workshop has higher post-workshop knowledge scores than the group that did not.
- If people who live in townships experience, on average, more perceived stress than those who live in the suburbs.
- If Mamelodi Sundowns has, on average, scored more goals per game in the last two years than Kaizer Chiefs.

We have formulated the research questions above as a difference between two means that were calculated from two groups of people for some measurement. Another way of thinking about the same problem is to consider it as a relationship between two variables. One variable is a nominal level measurement or categorical measurement that is used to indicate memebership of a group (this would be the independent variable), and the other is a measurement of intensity or quantity on at least an interval level (the dependent variable). It can be contrasted with the situation where two variables are compared when both are measurements of quantity or intensity (which will be discussed in Topic 6).

STUDY UNIT 5.2
Testing for differences between the means of two dependent groups

Often the two samples are not 'independent'. This happens when each subject in one sample is matched with regard to some characteristic (usually a nuisance or external variable that we wish to control) to a particular subject in the other sample. The samples are dependent if each measurement of a variable for a particular case can be paired with the measurement of a matching case in the other sample. The implication is that the two samples will always have to be of the same size (that is, $n_1 = n_2$). This design is, therefore, often referred to as a *matched-pairs design*. This implicit matching usually causes the scores to be

correlated (see Topic 6 for the meaning of this term). A typical example would be if the same research participants are measured twice, once before and again after an intervention. From the point of view of research design, we would refer to this type of comparison as a *two-sample repeated measures design*.

As far as the AIDS evaluation study in Appendix A is concerned, let us concentrate on the treatment group only, for the time being. Note that we obtain two sets of scores on the variable, 'knowledge of AIDS', namely, pre-training and post-training scores. This comes about because the same 20 subjects were tested twice – before and after training. In other words, each subject is matched with himself or herself, which is why the samples are regarded as dependent (see the discussion of dependent versus independent samples at the beginning of this topic in section 5.1.1). It is to compensate for this existing relationship between the samples that we require an adjustment in the t-statistic.

To develop this adjusted t-test, we use the two matched samples to create a new variable called '$\bar{d}$'. We do this by computing a 'difference score' between $\bar{x}_1$ and $\bar{x}_2$ so that $\bar{d}$ reflects the mean of the differences between the measurements before and after the workshop. This is reflected in Table 5.2 below, where these calculations are shown for the treatment group (taken from Appendix A) only, and subtracting each score before the workshop from the matched score thereafter (e.g. $d = x_2 - x_1$).

TABLE 5.2: Calculation of difference scores for the pre- and post-scores of 'knowledge' in the treatment group

Treatment group no.	Pre-training AIDS knowledge (x_1)	Post-training AIDS knowledge (x_2)	Difference ($d=x_2-x_1$)
1	8	9	1
2	7	12	5
3	3	7	4
4	4	11	7
5	6	6	0
6	2	4	2
7	8	10	2
8	5	12	7
9	9	15	6
10	4	10	6
11	6	6	0
12	7	15	8
13	5	12	7
14	6	9	3
15	7	14	7
16	6	11	5
17	8	13	5
18	6	9	3
19	7	13	6
20	7	15	8
$\bar{x}$	6.05	10.65	4.60
s	1.79	3.20	2.58

We work with the last column in Table 5.2 – the set of differences between the post-training and pre-training AIDS knowledge scores – as our raw data. We want to test whether these differences have generally *increased*, which is what we would expect if the workshop was effective.

We can formulate our statistical hypothesis as follows:

H_0: $\bar{D} = 0$ (i.e. the mean difference score shows no increase)
H_1: $\bar{D} > 0$ (i.e. the mean difference score shows an increase)

Here we use the symbol $\bar{D}$ to refer to the population mean of the 'difference' scores. We are, in fact, testing whether the population mean for the differences has increased significantly.

We refer to the mean *sample* difference score by means of the symbol $\bar{d}$. Note that this mean score was calculated from the difference scores in the last column of the table. The standard deviation of the sample of difference scores is indicated by the symbol $s_{\bar{d}}$.

What we want to do now is to find a t-test to test for the significance of these differences scores (we indicate this test statistic with the symbol $t_{\bar{d}}$). Note that if the calculated mean difference had been < 0, we would not have continued with the test, because it would have implied that the knowledge of AIDS had *decreased* in the treatment group, so there could be no probability of it having increased significantly. We could also have worked the mean difference out the other way round: using $d=x_1-x_2$ for the differences in each matched pair in Table 5.2 would imply testing whether the mean of the differences score in knowledge about AIDS for the treatment group was *lower before* the workshop than *after*. This would have implied testing the alternative hypothesis H_1: $\bar{D} < 0$. This would be a perfectly legitimate way to do the test, as long as you interpret the result correctly, by keeping in mind what the hypotheses *mean*.

It so happens that we are familiar with this particular test statistic already! Since we are in fact comparing a single mean (the difference score) with a specific constant (zero), this is just an application of the t-test for one sample when the population standard deviation (σ) is unknown (i.e. the $t_{\bar{x}}$ test statistic from Topic 4). All we need to do is substitute the sample mean ($\bar{x}$) with the mean of the differences ($\bar{d}$). So the test statistic is

$$t_{\bar{d}} = \frac{\bar{d} - \bar{D}}{\frac{s_{\bar{d}}}{\sqrt{n}}}$$

If you compare this with the formula in Topic 3 (section 3.2.2), you will see that all that was done here was to replace $\bar{x}$ with $\bar{d}$, μ with $\bar{D}$, and s with $s_{\bar{d}}$. We can also replace $\bar{D}$ with zero, since the particular null hypothesis that we want to test states that $\bar{D} = 0$.

So our formula becomes

$$t_{\bar{d}} = \frac{\bar{d}}{\frac{s_{\bar{d}}}{\sqrt{n}}}$$

This t-test statistic (t_d) can now be used for the comparison of means from two matched or dependent samples.

When we substitute the values from Table 5.2 above using the mean and standard deviations of the difference score, we get

$$t_{\bar{d}} = \frac{4.6}{\frac{2.58}{\sqrt{20}}} = \frac{4.6}{\frac{2.58}{4.472}} = \frac{4.6}{0.577} = 7.972$$

This t-value is so large that the p-value is bound to be very small (as in the case of z-statistics, t-values of above 3 are seldom not significant). We computed it and found it to be 0.00000008. Most computer programs will report such a p-value to four figures or so after the decimal point as 0.0000. This p-value is clearly smaller than 0.01 (the α value we chose), so that H_0 must be rejected and H_1 accepted.

We could also check the effect size to see whether the difference is as large as it seems. For differences scores we would calculate *Cohen's d* as follows:

$$\textit{Cohen's } d = \frac{\bar{d} - \bar{D}}{s_{\bar{d}}} = \frac{4.6 - 0}{2.58} = 1.783$$

This confirms that our effect is very large.

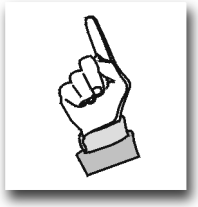
We have shown, therefore, that there was a very significant increase in the 'knowledge of AIDS' score for those who participated in the training. Note that we cannot conclude that the training *caused* this increase in knowledge. Before we can make such a claim, we would have to establish what happened to the 'knowledge of AIDS' scores of those employees who did not participate! If their 'knowledge of AIDS' showed a similar increase, then the training itself cannot get the credit, and some other hidden variable or variables are likely to be involved.

The strategy of creating 'difference' scores for dependent sets of measures is a highly useful and an important one in the social sciences. We often test and retest subjects on some variable, sometimes even more than two times (e.g., the treatment group can be tested again after six months or so to see if they retained their knowledge).

STUDY UNIT 5.3
Using differences scores to compare two independent groups

In the discussion of the t_d-test statistic above, we chose to look at the treatment group only. The possibility should, however, be considered that the improvement we saw between scores before and after the intervention (workshop) was caused by factors outside the AIDS workshop: even the experience of doing the test can affect the outcome (since the second time a test participant may remember his or her responses from the first time). This is why we use a control group: we want to take the possibility of hidden variables into account as far as possible. The problem is that we now seem to have four mean scores for four groups to

contend with: two groups of people (treatment and control) who are to be compared for two dependent measurements (knowledge of AIDS before and after the workshop). If the workshop was effective, we would expect little change in the control group means, but that the change (improvement) in the means of the treatment group should be significantly greater.



Methods to compare such multiple mean sets exist; for instance, analysis of variance techniques (which we do not deal with in this module). Another method is, however, suggested if we think in terms of the difference scores for each of the two (treatment and control) groups.

- We know that – if all went well – we expect to find *no change* in the control group when we compare the mean after the workshop with the one before.
- We also know that a *positive change* should occur in the scores from before and after the workshop in the treatment group.
- We also would like to see the change in the treatment group to be *greater than* the change (or lack of change) in the control group.

We can state this problem in terms of the difference scores for each of these groups: we expect the mean of difference of the treatment group to be greater than the mean of the difference in the control group. In other words, there should be a larger increase in the knowledge scores for those attending the training (the treatments group) than for those who did not attend the training (the control group).

If we refer to mean difference of the treatment group in the population as $\bar{D}_t$ and mean difference of the control group as $\bar{D}_c$ then the statistical hypothesis we want to test is

H_0: $\bar{D}_t = \bar{D}_c$ (or $\bar{D}_t - \bar{D}_c = 0$): There is no change between treatment and control group reflected in the means of the difference scores from before and after the workshop.

H_1: $\bar{D}_t > \bar{D}_c$: There is a change, which favours the treatment group.

To test this hypothesis, we need to calculate the mean of the difference scores in the control group in the same way as we calculated the mean difference in the treatment group before (Table 5.2). This is presented in Table 5.3 below.

TABLE 5.3: Calculation of difference scores for the pre- and post-scores of 'knowledge' in the control group

Control group no.	Pre-training AIDS knowledge (x_1)	Post-training AIDS knowledge (x_2)	Difference ($d=x_2-x_1$)
21	9	8	-1
22	6	8	2
23	4	2	-2
24	5	6	1
25	5	8	3
26	3	2	-1
27	9	9	0
28	4	2	-2
29	10	11	1
30	3	3	0
31	7	7	0
32	6	3	-3
33	6	9	3
34	7	3	-4
35	6	12	6
36	7	3	-4
37	9	8	-1
38	4	4	0
39	9	9	0
40	6	6	0
$\bar{x}$	6.25	6.15	-0.10
s	2.124	1.725	2.4257

Assume we decide to test at $\alpha = 0.01$.

From Table 5.3 we can see that the control group (who did *not* do the workshop) did slightly worse on the second test than on the first test. Such a small difference is, however, probably just the effect of random (measurement) error. We compare the sample statistics of the two groups (from Tables 5.2 and 5.3) in Table 5.4 below:

TABLE 5.4: Descriptive statistics for the variable d: change in knowledge between post-training and pre-training scores

Group	**n**	**Mean of differences**	**S.d. of differences**
Treatment	20	4.60	2.58
Control	20	–0.10	2.43

Notice that the change in mean difference score is in the right direction: the mean difference score of the treatment group is greater than that of the control group. If this mean difference score had *decreased*, there would be no possibility of detecting an increase, and there would really be no point in continuing to do a test (see section 3.2.1).

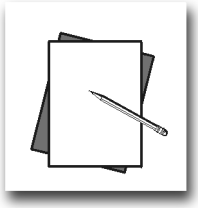
We now have to decide which statistical test we should use. Note that the data

above refer to two *independent* groups, given that research participants were randomly allocated to the treatment or control groups. The dependent grouping factor – the fact that each individual has been tested twice, once before and once after the workshop – has been removed. It is 'hidden' in the fact that we are not looking at the sets of measurement but at the *difference between them*.

The appropriate test, therefore, is the $\mathbf{t_c}$ test for independent samples. Note that the two group standard deviations in Table 5.3 also seem fairly similar, which means that this condition for using the test is met. Since we use means of differences and standard deviations of difference scores, we can write the formula (from section 5.1.2) as

$$t_c = \frac{(\bar{d}_1 - \bar{d}_2)}{\sqrt{\frac{s_{d1}^2}{n_1} + \frac{s_{d2}^2}{n_2}}}$$

Here we substitute $\bar{x}_1$=4.60 and s_{d_1}=2.58 for the treatment group (1) and $\bar{x}_2$=–0.10 and s_{d_2} =2.43 for the control group.



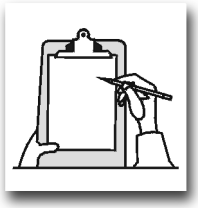
Exercise

We suggest you try to calculate this t-test statistic by yourself. When we calculated this test statistic (using a computer program), we found that t_c= 5.93142 with a p-value of 0.0000 (i.e. at least four zeros after the decimal point). This is for two-sided testing, but is so small it needs not be adjusted for the one-tailed test (we would need to divide it by two otherwise; see Topic 3 section 3.2.2).

The p-value is clearly smaller than the level of significance (p-value $< \alpha = 0.01$). The null hypothesis (H_0) can, therefore, be rejected and the alternative hypothesis (H_1) accepted. The research hypothesis is, therefore, confirmed. We have managed to show that the three-day workshop did indeed increase the employees' knowledge of AIDS relative to that of employees who did not attend this workshop.

What if we wanted to compare three or more groups? One can compare three groups using analysis of variance (a procedure not covered in this module), or one can use t-tests to compare two groups at a time until one has compared all three groups with one another. It would probably be wise to use a smaller level of significance since the probability of a Type I error increases as you do more statistical tests on the same data.

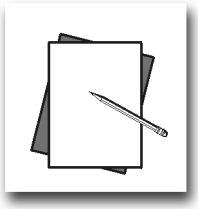
There are many statistical procedures (e.g. various forms of analysis of variance, regression analysis and factor analysis) that allow for more complex situations than those covered in this module. However, if you understand the basic principles of inferential statistics and the commonly-used statistics discussed in this module, moving on to more complex statistics will be much easier (should you wish to do so at a later stage).



Summary of the important ideas in this topic

This topic dealt with the comparison of two population groups on the basis of a random sample from each of these populations; in other words, in this topic we looked at statistical tests performed on the data from two samples. After having studied this topic, you should be able to

- identify situations where it is appropriate to use a t-test.
- list the three factors that determine the size of a t-value – the size of the difference between the means, the standard deviations of the two samples and the sample sizes.
- indicate when to use a t-test for independent samples and when to use a t-test for dependent samples.
- know how differences scores can be used in comparisons of two means.
- list and explain two assumptions underlying the t-test – normality of distribution and equality of variance (or standard deviations).
- state a research hypothesis for a research scenario requiring the comparison of two statistical populations, formulate the appropriate statistical hypotheses, and test sample results for statistical significance.
- know how to calculate the effect size when two sample means are compared.



Multiple-choice questions and answers

Questions

Study the AIDS evaluation scenario in Appendix A with regard to the variable, 'attitude to AIDS'. Assume that the two groups (treatment and control) were equal with regard to their attitude scores before training commenced. State a research hypothesis implying that the training was successful in improving attitudes, then state the statistical hypotheses. Now answer questions 1 to 18 below.

1. The research hypothesis implies that
    1. the two samples are dependent.
    2. the dependent variable affects the independent variable.
    3. the treatment group will score higher than the control group on the dependent variable.

2. The research population is
    1. all employees in South Africa.
    2. all employees in the company.
    3. undefinable.

3. The samples are
    1. randomly selected.
    2. randomly assigned to experimental and control groups.
    3. both of the above.

4. Measurements on the 'attitude to AIDS' test can be viewed as a/an
    1. nominal scale.

2. interval scale.
3. ordinal scale.

5. If the alternative hypothesis is $\mu_1 > \mu_2$, then μ_1 is

   1. the sample mean of the post-attitude scores of group 1.
   2. the population mean of post-attitude scores from which group 1 was selected.
   3. the unknown parameter under the null hypothesis.

6. The appropriate test statistic is

   1. t_c
   2. z_c
   3. $t_{\bar{d}}$

7. The assumptions underlying the t-test are

   1. equal distributions and normal variances.
   2. unknown standard deviations.
   3. normal population distributions with equal variances.

8. Consider the following statistics regarding the post-training 'attitude' scores:

Group	Mean	Std Dev
1	21.65	2.99
2	20.40	3.05

   What are the values in the table called?

   1. Population parameters
   2. Sample statistics
   3. Test statistics

9. Based on the data in question 8, what is the value of the t-test statistic?

   1. +2.3
   2. –2.3
   3. +1.3

10. Suppose the two-tailed p-value for the t-test of the differences between the two means in question 8 is 0.19. If a is set to 0.10, what is the decision regarding H_0?

    1. Do not reject H_0 because $0.19 > 0.10$.
    2. Accept H_1 because $0.19 > 0{,}1$.
    3. Reject H_0 because $0.095 < 0.10$.

11. The p-value gives

    1. the probability that the difference between 21.65 and 20.40 could be significant.
    2. the probability that the difference between 21.65 and 20.40 could be due to chance.

3. the probability that the difference between 21.65 and 20.40 could be true.

12. Suppose you are convinced that you cannot make the assumption that the two groups are equal with regard to pre-training 'attitude to AIDS'. How would you establish if the training was successful in improving 'attitudes to AIDS'?

    1. Perform statistical tests on pre-training attitude scores.
    2. Perform statistical tests on the difference between pre- and post-training attitude scores.
    3. Perform a t-test for dependent groups.

13. Suppose one feels uncomfortable with the assumption that population distributions are normal. It is helpful if

    1. sample sizes are large.
    2. sample standard deviations are known.
    3. (a) and (b).

14. Suppose we want to determine if the control group in the AIDS study (Appendix A) improved as far as its attitude scores are concerned. One should

    1. compare pre- and post-training scores on the 'attitude' variable for the control group.
    2. compare the control group with the experimental group on attitude scores.
    3. perform a t-test on the post-training scores.

15. If three groups are being compared with regard to their mean scores, one

    1. should consider an analysis of variance strategy.
    2. should join two of the groups and proceed with a t-test.
    3. could do either of the above.

16. When the differences between two sample means look large, one

    1. will find the result to be statistically significant.
    2. may still find that the result is not statistically significant.
    3. should reject the null hypothesis in favour of the alternative hypothesis.

17. Suppose one finds an impressive difference between two sample means, but the result is found to be statistically insignificant. One may

    1. repeat the study with larger samples.
    2. decrease the α level.
    3. recalculate the p-value, but for a larger α level.

18. Suppose the t-statistic is found to be –1.2. This indicates

    1. a non-significant result.
    2. a significant result if, on the basis of H_1, we expect a negative t-value.
    3. a significant result if the appropriate p-value is smaller than α.

Solutions

1. Option (3) is correct. Option (1) is incorrect because the treatment and control groups are independent. Option (2) is incorrect as the independent variable affects the dependent variable and not the other way around.
2. Option (2) is correct. The issue is whether training is effective or not and not whether the company is different from other employers in South Africa.
3. Option (3) is correct. Note that in the case of true experimental designs, random assignment to groups is always possible, whether random sampling from the general population took place or not. Therefore, random assignment to groups does not imply that random selection took place in the first place. The researcher sometimes uses a non-random sample and then divides the sample randomly into an experimental and a control group. This latter process is known as random assignment to groups.
4. Option (2) is correct. See Appendix B.
5. Option (2) is correct. Option (1) is incorrect because hypotheses are never about sample statistics but about population parameters. Option (3) is too vague to be correct.
6. Option (1) is correct as we are comparing two independent samples with regard to their means.
7. Option (3) is correct.
8. Option (2) is correct as these values refer to calculations based on a sample of data (see section 1.4.3).
9. Option (3) is correct. The t_C value is calculated as follows:

$$\begin{aligned} t_c &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{21.65 - 20.40}{\sqrt{\frac{2.99 \times 2.99}{20} + \frac{3.05 \times 3.05}{20}}} \\ &= 1.3 \end{aligned}$$

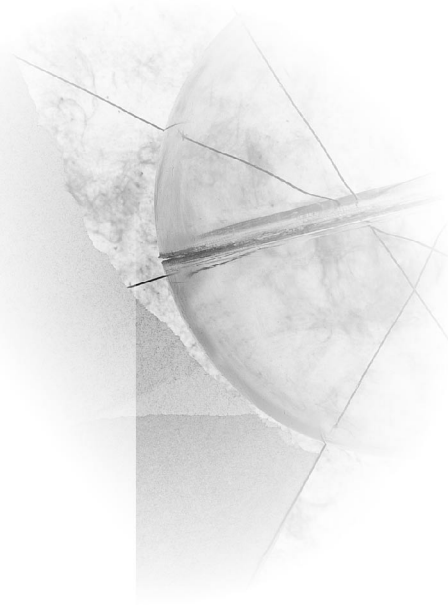
10. The correct option is (3). Because the alternative hypothesis is directional, implying that group 1 will score higher than group 2, which is indeed the case, we apply a directional test. This requires that we divide 0.19 by 2, which gives 0.095. As 0.095 is less than 0.10 (the significance level), the null hypothesis should be rejected and the alternative hypothesis accepted.
11. Option (2) is correct. The p-value always reflects the probability that the result is due to chance.
12. Option (2) is correct. We have to compare the two groups with regard to their mean 'difference' score. The idea is to show that group 1 became more positive in attitude relative to group 2.
13. Option (1) is correct. The larger the sample sizes, the more the sampling distribution will tend to be normal, irrespective of the shape of the population distribution (see Topic 2 on the central limit theorem). Option (2) is incorrect. Sample standard deviations are always known, as we can always calculate them.
14. The correct option is (1).
15. Option (1) is correct.
16. Option (2) is correct. Given that a particular result is in the right direction (favouring H_1), it is necessary to perform a statistical test to see whether the

result is significant. A large effect can fail to be significant if the sample is too small. The smaller the sample, the greater the probability of finding a result that appears to be significant *purely by chance*. This relates to issues like the power of the test and effect size (see sections 3.3.2 and 3.3.3).

17. Option (1) is the only one of the three options that makes sense. If a large effect does not yield a significant result, it may be that the statistical test lacks power (see section 3.3.2) and increasing the sample size will increase the power of the test (i.e. its ability to detect significant differences).
18. The correct option is (3). You have to know what the p-value is that is to be compared with the level of significance, α.

TOPIC 6

Testing hypotheses about a relationship between two variables



Quick overview

In this topic we consider the comparison of two variables from the same sample, and how hypotheses that relate to the relationships between them can be tested. The correlation between two variables which are quantitative measurements will be discussed. We also discuss the situation where both variables are categorical (nominal level) measurements, and the use of the chi-square test for contingency tables.

This topic is divided into the following study units:

- *Study unit 6.1:* Correlation: measuring the association between variables
- *Study unit 6.2:* A test of association between two nominal variables: the χ^2 test for contingency tables

Introduction

Researchers are often interested in establishing whether a relationship exists between two variables, or whether a specific kind of relationship exists. They are also likely to be interested in knowing whether the relationships found in the sample of data can be generalised to the whole population from which the sample was drawn.

In this topic, we consider the problem of testing an hypothesis of association between two variables from a single sample. We consider two instances, namely,

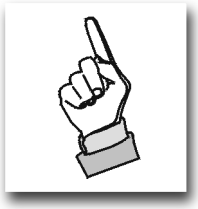
- the notion of the relationship between two continuous variables and how the size of the relationship can be expressed in terms of a *correlation* between them (the index of association is the *Pearson product-moment correlation coefficient*). This coefficient can also be used as a test statistic.
- a test of the association between two categorical (nominal scale) variables, where both variables are considered to be nominal (or categorical) and the appropriate test statistic is the Pearson χ^2 test of association for contingency tables.

STUDY UNIT 6.1
Correlation: measuring the association between variables

6.1.1 Visualising a correlation

Remember that variables are observable attributes of events or conditions that we have measured in some way (i.e. which have been given a numeric value according to a scale). When we have two such variables, we may want to ask the following questions:

- Is there a relationship between the variables: in other words, is there some kind of interdependency?
- What is the shape and size of the relationship?



Correlation is a measurement of the extent to which a measurement on one variable is related to a measurement on another variable for the same sample of individual cases.

This can be visualised by way of a graphical representation called a *scatter plot*. A scatter plot is a graph that represents the measurements of two variables on two perpendicular axes, usually called the x-axis (horizontal axis or abscissa) and the y-axis (vertical axis or ordinate).

To create a scatter plot is fairly simple. As an example, we plot the scores for the questionnaire regarding a person's attitude to persons with HIV/AIDS before and after the workshop referred to in the research scenario presented in Appendix A (referred to as 'Pre-attitude' and 'Post-attitude' respectively in the rest of this discussion). Since these variables represent measurements of the same construct for the same sample of subjects, one would expect some kind of relationship to exist, in spite of the intervention (i.e. the workshop).

The actual scores for each of the two variables under consideration are given under the headings 'Pre-attitude' and 'Post-attitude'.

For your convenience, the values for *Pre-attitude* (x) and *Post-attitude* (y) in Appendix A are repeated here:

TABLE 6.1: Pre-Attitude and Post-Attitude – Attitudes before and after AIDS workshop

Pre-Attitude (x):	14 20 24 21 21 22 23 24 20 18 12 14 17 19 19 15 17 17 19 14 15 19 25 20 22 21 24 23 19 19 11 15 16 21 18 16 17 18 19 14
Post-Attitude (y):	21 23 23 26 24 21 25 25 25 23 16 19 19 21 24 19 18 23 22 16 18 22 21 22 22 18 23 23 25 24 16 18 17 17 25 18 18 23 22 16

First you will need a graph paper, such as the one provided in Figure 6.1 below.

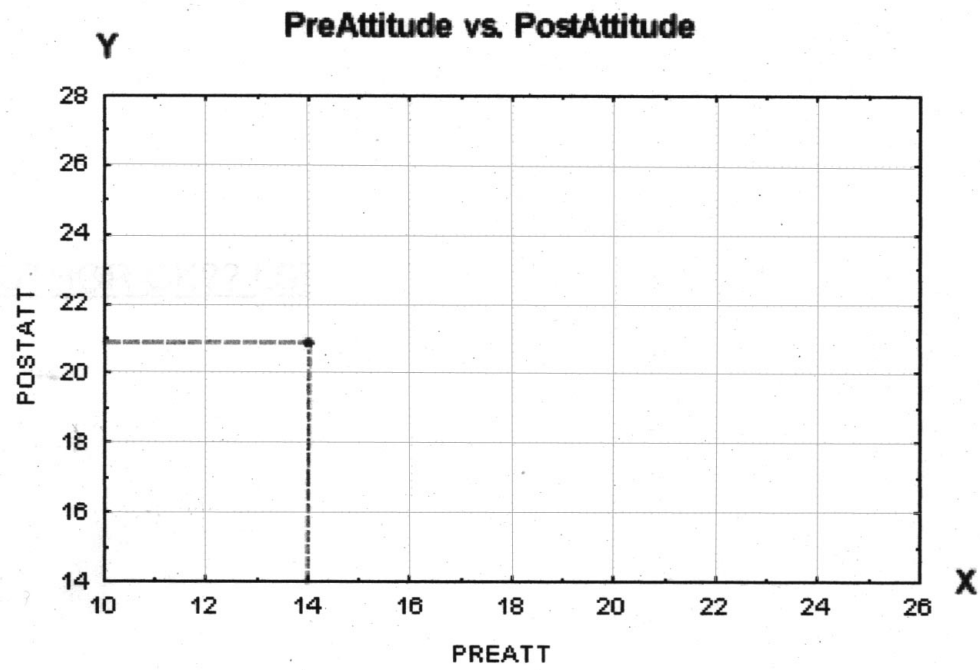


FIGURE 6.1: Position of the first research participant on the scatter plot

The two axes of the graph, **x** and **y**, represent the two variables, *Pre-attitude* and *Post-attitude*. For this example, we put *Post-attitude* on the y-axis (vertical) and *Pre-attitude* on the x-axis. It is usual to put the variable that is being affected (the *dependent* variable) on the y-axis and the variable that is causing the effect (the *independent* variable) on the x-axis. It probably makes more sense to say the attitude *after* the workshop is a consequence of the attitude *before* than the other way round, but, for the computation of a correlation coefficient, it does not really matter much which of the variables is taken as dependent and which is taken as independent variable.

To create the scatter plot, proceed as follows. Find the first *Pre-attitude* value (i.e. the score of the first subject on the Attitude Questionnaire as tested before the workshop) in the table. This value (the first x-value, which we can refer to as x_1) is 14, so lightly draw a line vertically upwards from the point that represents '14' on the x-axis (as indicated by the broken line in Figure 6.1). Take the matching y_1 value from the *Post-attitude* column (which is 21) and lightly draw a light line

horizontally towards the right from the point that represents '21' on the y-axis. Mark the spot where the two lines cross with a dot.

Repeat this procedure for each pair of x- and y-values. The dots now represent the relationship between students' scores for the two variables, *Pre-attitude* and *Post-attitude*. You can check your scatter plot by comparing it with Figure 6.2.

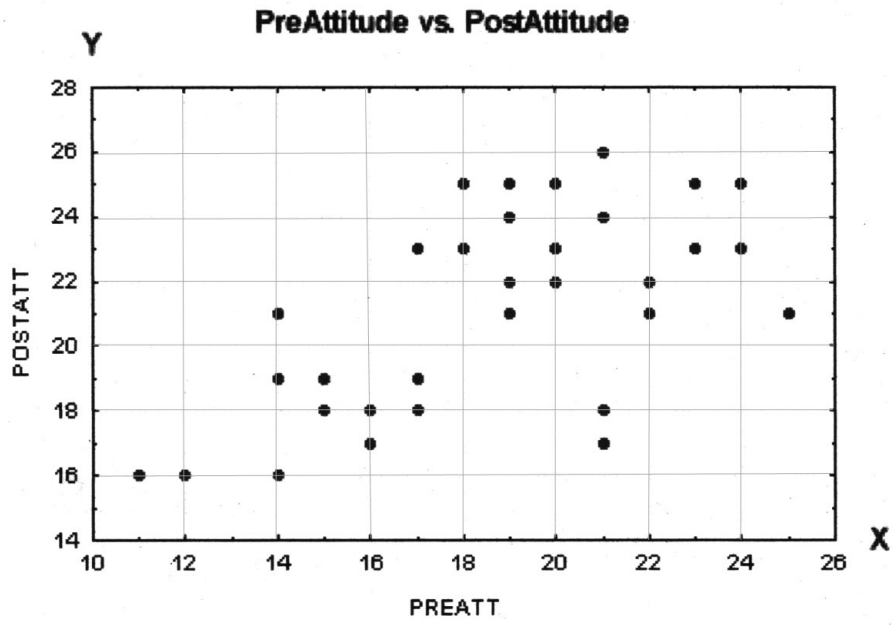
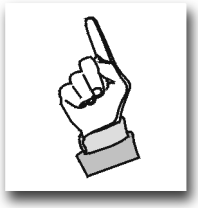


FIGURE 6.2: Scatter plot of Pre-attitude vs. Post-attitude

If you look carefully at your scatter plot (or at Fig. 6.2), you will notice that, in general, persons scoring high on the Pre-attitude questionnaire also scored high on the Post-attitude questionnaire. Similarly, persons scoring low on the Pre-attitude questionnaire scored low on the Post-attitude questionnaire, and moderate scores on one questionnaire were usually matched with moderate scores on the other. Because of this, the dots seem to be roughly distributed around a line going from below at the left to the top at the right. This tendency to form a line is referred to as the linearity of the relation, and it is this linearity (or lack of it) – the extent to which the relationship approximates a straight line – that the correlation coefficients we are interested in are designed to measure. It is possible to calculate the best approximate straight line going through the set of points, which is known as the *regression line*.



Study the different scatter plots in Figure 6.3. You will notice certain trends. The closer the dots in the plot are to a straight line, the closer the correlation coefficient is to 1 (it can be either a positive number (+1) or a negative one (–1)). The more arbitrary or spread out the dots, the closer the correlation coefficient is to 0. If the plot seems to form a line from lower left to upper right, the correlation is positive. On the other hand, if the line runs from upper left to lower right, the correlation is negative.

Correlation coefficients that measure the linear relationship between two variables, such as the Pearson product-moment correlation coefficient, can

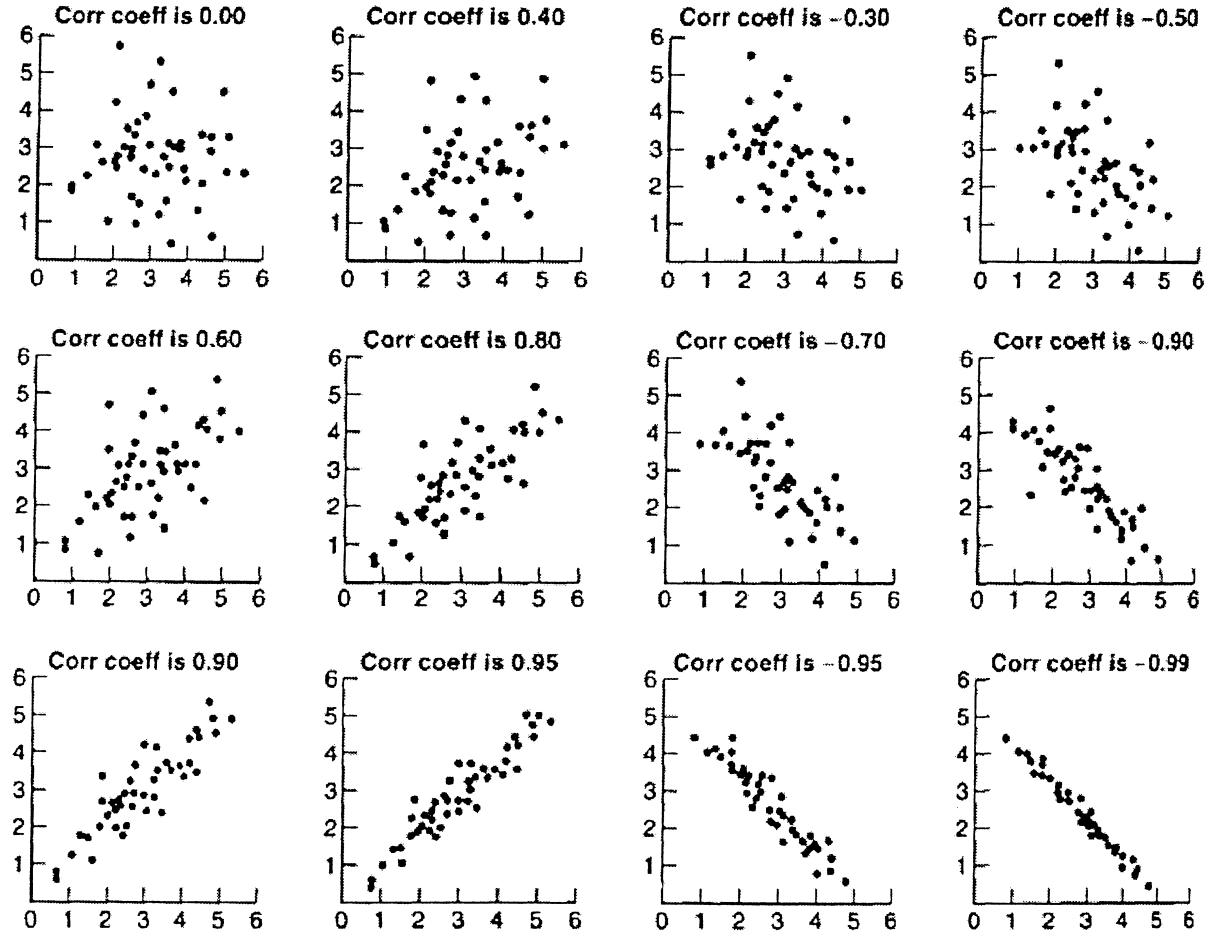
have a continuous value that ranges from –1 to 1 (a positive value is usually written without the sign, so '1' is presumed to mean '+1'). We use 'r' as the symbol that represents a correlation coefficient (as in the case of the Pearson product-moment correlation coefficient), and the following applies:

r = 1 implies a perfect *positive linear relationship* (the dots in a scatter plot will run from lower left to upper right in a perfectly straight line)

r = 0 implies *no linear relationship* at all (the dots may be scattered all over the place)

r = -1 implies a perfect *negative linear relationship* (the dots will run from upper left to lower right in a straight line)

When positive relationships occur, this implies that as one variable gets larger, so does the other. When negative relationships occur, this implies that as one variable gets larger, the other gets smaller. An example of a negative correlation might be *age* and *health status*: the older you get, the less healthy you are likely to be.



Source: From *Statistics* (2nd ed., pp. 119, 121) by D. Freedman, R. Pisani, R. Purves, and A. Adhikari, 1991, New York: W.W. Norton

FIGURE 6.3: Scatter plots representing different values of the correlation coefficient

Here are some frequently asked questions for you to consider:

What does it imply if there is a straight line, but it is parallel to either the x-axis (horizontal) or the y-axis (vertical)?

In such a case, there is no correlation, since, as the one variable varies, the other one remains the same (i.e., one of the variables is actually a constant).

Is the correlation coefficient relevant if a good nonlinear relationship exists, that is, if the dots lie in a line that forms a U-shaped or S-shaped curve, or some other pattern, but do not simply form a straight line?

The Pearson correlation coefficient only measures the extent to which the data vary around a straight line. A perfect, nonlinear relationship can produce a reasonable Pearson's r, or even a very small one, so r tells one very little about nonlinear relationships. (Indexes that measure such non-linear relationships do exist.)

6.1.2 Calculating the Pearson product-moment correlation coefficient

To illustrate how correlation coefficients are computed, we discuss the most popular one, namely, Pearson's product-moment correlation coefficient. As we mentioned before, the coefficient is usually written as r (in cases where the two variables to be correlated are expressed as x and y, the coefficient is sometimes expressed as r_{xy}), and it provides a quantity that indicates the extent to which the relationship between two variables is linear: that is, the extent to which the shape of the relationship can be approximated by a straight line.

The general form of the Pearson's r can be expressed as follows:

$$r = \frac{cov(x;y)}{\sqrt{var(x)var(y)}}$$

In this equation, *cov(x;y)* refers to the *covariance* of x and y, and it represents the extent to which two variables – x and y – vary *together*. Var(x) and var(y) refer to the *variances* of x and y respectively, where the variance is the square of the standard deviation (see Appendix C). Var(x) is, therefore, a measurement of the extent to which a variable varies on its own. From this you can see that the Pearson's r expresses a relationship of the variability of *two variables taken together*, with the variability of *each of the two variables on its own*.

It is because two variables taken together can never vary more than the same two variables taken separately that the correlation coefficient can never be larger than 1 (or smaller than –1). Statistics like variances and means are referred to as *moments* of a distribution, and, since these are multiplied (i.e. a product is computed), Pearson's r is also known as the *product-moment correlation coefficient*.

To compute the Pearson's r, you use raw scores in the following formula:

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

This may seem terribly complicated, but if you use a table of the data listing all the variables you need, you can do it in a systematic way. Insert the two variables that you want to correlate (**x** and **y**), calculate their respective squares ($\mathbf{x^2}$ and $\mathbf{y^2}$), their product (**x** multiplied with **y**, written as **xy**) and add these to find their sums (Σ).

(*Note:* an example of calculations using a table such as this is given in Appendix E. Be careful not to confuse **(Σx)(Σy)** with **(Σxy)**; that is, the product of the sums with the sum of the products. Summing first, then multiplying the results lead to a different result than multiplying first and then adding the results.)

You can then put this information in the formula above to find the value of Pearson's r.

TABLE 6.2: Calculations for Pearson's r

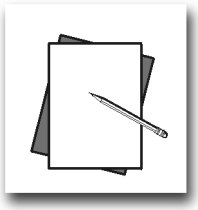
	Pre-Attitude		Post-Attitude		Product
Case	**x**	**x^2**	**y**	**y^2**	**xy**
1	14	196	21	441	294
2	20	400	23	529	460
3	24	576	23	529	552
4	21	441	26	676	546
5	21	441	24	576	504
6	22	484	21	441	462
7	23	529	25	625	575
8	24	576	25	625	600
9	20	400	25	625	500
10	18	324	23	529	414
11	12	144	16	256	192
12	14	196	19	361	266
13	17	289	19	361	323
14	19	361	21	441	399
15	19	361	24	576	456
16	15	225	19	361	285
17	17	289	18	324	306
18	17	289	23	529	391
19	19	361	22	484	418
20	14	196	16	256	224
21	15	225	18	324	270
22	19	361	22	484	418
23	25	625	21	441	525
24	20	400	22	484	440
25	22	484	22	484	484
26	21	441	18	324	378
27	24	576	23	529	552
28	23	529	23	529	529
29	19	361	25	625	475
30	19	361	24	576	456
31	11	121	16	256	176
32	15	225	18	324	270
33	16	256	17	289	272
34	21	441	17	289	357
35	18	324	25	625	450
36	16	256	18	324	288
37	17	289	18	324	306
38	18	324	23	529	414
39	19	361	22	484	418
40	14	196	16	256	224
Sum (Σ)	742	14234	841	18045	15869

To demonstrate this, let us calculate the Pearson's r for *the Pre-attitude* and *Post-attitude* data that we plotted before on the scatter plot (in Figure 6.2).

Set up the table as shown in Table 6.2, then substitute the sums from the bottom row of the table for the formula, as follows:

$$\begin{aligned} r &= \frac{40(15869) - (742)(841)}{\sqrt{[40(14234) - (742)^2][40(18045) - (841)^2]}} \\ &= \frac{634760 - 624022}{\sqrt{(569360 - 550564)(721800 - 707281)}} \\ &= \frac{10738}{\sqrt{(18796)(14519)}} = \frac{10738}{\sqrt{272899124}} \\ &= \frac{10738}{16519.6587} \\ &= 0.65001 \end{aligned}$$

So the correlation between. *Pre-attitude* and *Post-attitude* is r = 0.65. Note that the correlation is positive, as we might have expected from the scatter plot: subjects with a higher Pre-attitude score are also likely to get a high Post-attitude score.



Exercise

It would be interesting to determine whether attitude is related to knowledge: can one state that the more the employees know about HIV/AIDS, the more positive their attitude to infected persons will be? To test this, a first step would be to determine the correlation between *Pre-Attitude* and *Pre-Knowledge*, to see whether a relationship exists before the effect of the workshop becomes important. We encourage you to do this as an exercise using the information in Appendix A (the AIDS workshop scenario) in the equation given above. But do the scatter plot first to see if you can predict the answer from the graph. (The correlation coefficient that you should find is given right at the end of this topic.)

6.1.3 Using Pearson's r to test an hypothesis

The Pearson correlation coefficient is really a *descriptive statistic*: it describes the relationship between two variables. In this section we show you how it can also be used as a test statistic to test hypotheses about relationships among variables.

Up to now we have worked with sample data only, using actual data to plot a relationship and to compute the strength of the linear relationship using Pearson's r. In practice, a researcher would want to generalise this information to the population from which the particular sample was drawn, that is, to establish whether Pearson's r as computed from sample data is substantial enough so that one may conclude that a significant relationship actually exists in the population

between the particular variables being studied. What we want to know is, at what point is the relationship that we have calculated – from sample data – substantial enough that we may conclude that an actual relationship exists in the population, and that it is not just a consequence of measurement and sampling error?

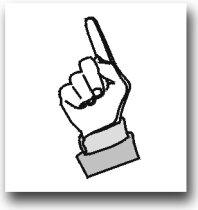
In order to do this, we have to distinguish the correlation coefficient that we computed from sample data from the 'ideal' correlation coefficient referring to the whole population (given that we are not able to collect actual data for the whole population). Where we used 'r' for the sample coefficient, we shall use the Greek letter 'ρ' (pronounced 'rho') for the population correlation coefficient.

We now need to set up a formal statistical hypothesis to test whether the sample result can be generalised to the population. We already know that, if no linear relationship exists between two variables, r will be equal to 0. The further r deviates from zero (in either a positive or negative direction), the stronger is the indication that a linear relationship exists.

So the null hypothesis – the hypothesis of no effect – will state that no relationship exists:

H_0: $\rho=0$

Note that as usual the hypothesis is formulated in terms of the *population* correlation coefficient, for we want to draw a conclusion about the whole population, and not just about the representative sample for which we have data.



What we usually want to know is whether this null hypothesis can be rejected, that is, whether we may conclude that an actual (significant) relationship does in fact exist. Three possible alternative hypotheses can be formulated:

(i) H_1: $\rho \neq 0$ This implies that a relationship that differs significantly from zero does in fact exist, but we are making no 'educated guesses' as to whether it is a positive or negative relationship: we just want to know whether there is in fact a relationship.

(ii) H_1: $\rho > 0$ This implies that we want to establish whether a significant relationship of greater than zero exists, that is, a significant positive relationship.

(iii) H_1: $\rho < 0$ This implies that we want to establish whether a significant relationship of less than zero exists, that is, a significant negative relationship.

Note that the first of these is a *two-tailed* or *non-directional* test (deviations from zero can be either positive or negative), while the other two are *one-tailed* or *directional* tests. It is important that you decide which of the hypotheses you want to test for before you do the testing, using your knowledge of the research problem that interests you (i.e. do not do the test first, then decide which hypothesis best fits the results: that would be cheating!).

To actually test these hypotheses, we need a test statistic so that we can derive the p-value to indicate whether the null hypothesis may be rejected in favour of the alternative hypothesis (one of the three possibilities above). Statisticians have found that it is possible to derive a probability distribution for the population correlation coefficient ρ directly. In other words, the value of r can be used as its own test statistic (if the sample is not too small).

Computers can use the calculated r-value and the sample size (n) to provide a p-value directly. Therefore, we do not describe the test in any detail: just be aware that r is both a descriptive statistic and (if you want to generalise it to a population) a test statistic.

After determining the p-value, one can decide whether to reject the null hypothesis in favour of the alternative hypothesis under consideration, by seeing whether or not the computed p-value is less than the level of significance (α) that you have chosen. Remember that the p-value produced by a computer is usually appropriate for a two-tailed test, so, if you are interested in a directional or one-tailed result, the two-tailed p-value should be divided by two before comparing it to the level of significance.

As an example, let us consider again the question of whether there is a relationship between Pre-attitude and Post-attitude in our HIV/ AIDS sample. We have seen before that a positive correlation exists, but now we want to test whether this result exceeds the correlation that we can expect purely by chance.

First, we need to formulate the problem as a research hypothesis. Since the Pre- and Post-attitude measures are measurements of the same construct at different times (before and after a workshop), it seems sensible to suppose that they should be positively correlated. So it seems sensible to formulate our alternative hypothesis as a *directional* hypothesis: that there is a positive relationship between attitude measurements before and after the workshop, as represented by the variables *Pre-attitude* and *Post-attitude* respectively. Given that Pearson's r is an appropriate measurement of the relationship between the variables, this becomes

H_0: $\rho = 0$
H_1: $\rho > 0$

where ρ is the population correlation between the measurements, *Pre-attitude* and *Post-attitude*.

We also have to decide on the level of significance (α) that we want to use. Let us decide that we want to perform our test on a 1% level of significance, that is, we set $\alpha = 0.01$. The next step would be to calculate r, using the *Pre-attitude* and *Post-attitude* sample data. As we have done this before (in section 6.1.2), we know that the result will be $r = 0.65$.

With an input of $r = 0.65$ and a sample size of $n=40$, the following p-value was found by our computer program:

p-value = 0.000006

Since we are doing a directional test, this p-value has to be divided by two. This

produces a result of p = 0.000003, which is considerably less than our preferred significance level value of $\alpha = 0.01$. So we conclude that we can reject the null hypothesis in favour of the alternative hypothesis. In other words, we can conclude that there is a significant positive relationship between employees' attitude measurements before and after the workshop.

6.1.4 Interpreting the correlation coefficient

6.1.4.1 Effect size for correlations

One should, however, be careful as to how one interprets a significant result. To clarify this, consider the relationship between the calculated significance (the **p-value)** and the sample size (**n**).

If you randomly put three dots on a blank square of paper, they may, purely by chance, fall into something approximating a straight line. If you make a hundred marks on the same piece of paper, also in a totally random way, the chance of them falling in a straight line is, however, a lot less. This tells you something about the relationship between **r** (a measure of whether the dots on a scatter plot fall in a straight line) and the number of dots (the sample size **n**): the smaller **n**, the more likely it is that the plot will represent a straight line *purely by chance*. Therefore, for a smaller sample **n**, the test must be much more conservative. You must, therefore, put up a bigger hurdle to be crossed before you conclude that the result is not the consequence of chance. You, therefore, require a larger value of **r** before you can conclude that the result is not a chance event due to sampling or measurement error, but an actual representation of the state of affairs in the population.

The consequence of this is that, for a large sample, a relatively modest correlation can turn out to be significant. For example, for a sample of n = 40 (as in the HIV/AIDS research project in Appendix A), the value of **r** must be at least r = 0.26 for $\alpha = 0.05$ (a 5% level). If we increase the sample size to 100, a smaller result of r = 0.16 would be significant at the same level of $\alpha = 0.05$.

This shows that, for a large value of **n**, a very modest **r** can be significant. The implication of this is that significance does *not* indicate that a relationship is large. It merely tells you that some relationship exists (perhaps a modest one), and that it is large enough not to be regarded as purely due to the effect of chance, given the size of the sample.

The squared correlation (r^2) measures the proportion of variance in one variable that can be determined from its relationship with the other, or how much variance they have in common. It can be used as an indication of the size of the effect.

The table below gives an indication of how to interpret this effect.

TABLE 6.3: Evaluating r^2

$r^2 = 0.01$	Small effect
$r^2 = 0.09$	Medium effect
$r^2 = 0.25$	Large effect

In the example above, we obtained a correlation of r = 0.65 between *Pre-attitude* and *Post-attitude*. If we square this we get $r^2 = 0.42$. This implies that the two variables share about 42% of their variance, which can be interpreted as a very large effect.

6.1.4.2 The problem of causation

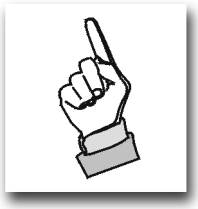
Correlation implies a relationship between variables, but can we say that one variable *causes* the other? Not necessarily. While causation implies correlation (if smoking causes lung cancer, you would expect a measurement of the amount of smoking to be positively correlated with a measurement of the incidence of lung cancer), the converse is not true. If you do find that two variables, **x** and **y**, are correlated, it may mean that one causes the other, but you need more evidence to be certain. Something that needs to be considered is the possibility of one or more *hidden* variables: that the correlation between **x** and **y** is caused by a third variable, which influences both of them (see also Topic 1, section 1.3.2, regarding this).

Consider the likelihood that a high, positive correlation exists between children's *foot size* and their *spelling ability*. This is not because foot size influences spelling ability (or because spelling ability influences foot size), but because older children, who can spell better, are also likely to be bigger and have larger feet. So a third variable, which influences the other two, is the child's *age*.

To determine causation, further research is usually necessary to eliminate the influence of possible alternative variables (a process of establishing the internal validity of the research model), and taking into account such factors as temporal precedence (how the variables are related to each other in time), and so on.

6.1.4.3 When to use Pearson's r

The Pearson correlation coefficient is used particularly when the two variables under consideration are both of a ratio or interval level of measurement (see Appendix B). In practice, it has been found that the coefficient works quite well even in cases where the measurement level is ordinal, that is, where the variables indicate sequence or rank order only, and not actual size or intensity. You may, however, consider using a different coefficient if your sample is on the small side and not normally distributed, or if you are interested in relationships that are not necessarily linear (on a straight line).



Both variables are not always measured on an ordinal, interval or ratio scale. Some ways of dealing with data where one of the variables or both are nominal scale measurements are given in the following scenarios:

- Both variables are categorical in nature (nominal scale measurements). *In such cases we often perform a χ^2-test. This is explained in Study Unit 6.2.*
- One variable is a dichotomous variable (i.e., a categorical variable consisting of two categories only, like 'gender') and the other variable is an interval or ratio scale. *Perform a t-test. The experienced researcher may even calculate an 'r' value between the two variables and be able to make sense of a significant r-value.*
- One variable is a categorical variable consisting of three or more categories and the other variable is an interval or ratio scale. *Perform a one-way Analysis of Variance F-test. This procedure does not form part of this syllabus.*

STUDY UNIT 6.2
A test of association between two nominal variables: the χ^2 test for contingency tables

6.2.1 Setting up an hypothesis about the relationship between two categorical variables

Up to this point we have considered the case where both variables in a possible relationship are measured on a quantitative measurement scale, usually up to the interval or ratio level. What happens if both variables are nominal or categorical? Let us look at an example.

Referring back to our original AIDS workshop research scenario (introduced in Appendix A), suppose that a researcher wants to establish whether employees' membership of the different job categories in the company (managerial, clerical or technical, as reflected in the *Status* variable) has had any influence on their responses to the question ''Do you believe that HIV testing should be made compulsory before an individual is appointed to any position in the company?'' Note that both of these variables are purely nominal: employees are allotted the numbers managerial = 1, clerical = 2 and technical = 3, but these numerals are used merely as labels to show they belong to different groups; it does not imply any relationship (in the sense of a measurement) between the groups. In the case of the answers to the question on HIV testing, the values Yes = 1, No = 2, were allotted to the responses, but it could just as well have been the other way round.

The research hypothesis that we want to test can be stated as follows:

> *Employees' opinions as to whether HIV testing should be compulsory for new appointments, depend on their job classification in the company.*

Note that the nature of the relation is not specified. We cannot, therefore, derive

a directional, alternative statistical hypothesis. Our alternative hypothesis needs to be *non-directional*.

An appropriate test in this case would be the *Pearson chi-square test* of association (or χ^2_p), which is a general test for determining whether two variables are *distributed* in the same way. We use this technique to compare what is expected if the null hypothesis (of no relationship between the variables) were true, with what is observed in the sample data.

So the technique works on the same basic principle as the other statistical tests we have been discussing: we are comparing observed data with what can be expected if the null hypothesis is true. In this particular instance, we use the χ^2_p test statistic to determine what the chances are that an observed number of objects or responses falling in each of a number of cross-tabulated categories are the result purely of random sampling error.

The first step in our test for the relationship between job category (the variable *Status*) and responses to the question about compulsory HIV testing (the variable *HIVemploy*), is to set up our formal statistical hypothesis:

H_0: There is no relationship between job status and an employee's opinion about compulsory HIV testing (or, job status and opinion about compulsory HIV testing are independent variables).

H_1: A relationship between job status and opinion about compulsory HIV testing exists (or, these two variables are not independent; in other words, they do affect each other).

Let us decide beforehand that we shall be testing at a 5% level of significance (i.e. $\alpha = 0.05$).

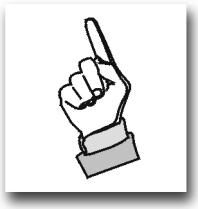
6.2.2 The contingency table

After setting up the hypotheses to be tested for, the next step is to create a contingency table, which is a table indicating the number of individual objects falling in each cell of cross-tabulated data. In other words, it is a two-dimensional table in which each observation is classified in terms of two categories simultaneously.

For our example, the table would appear as follows:

Table 6.4: A contingency table for cross-tabulations of *HIV-Employ* versus *Status*

		Status			
		1 (Managerial)	2 (Clerical)	3 (Technical)	*Row total*
HIV-Employ	1 (Yes) 2 (No)	5 (O_{11}) 5 (O_{21})	9 (O_{12}) 7 (O_{22})	5 (O_{13}) 9 (O_{23})	**19 ($O_{1.}$)** **21 ($O_{2.}$)**
	Column total	**10 ($O_{.1}$)**	**16 ($O_{.2}$)**	**14 ($O_{.3}$)**	**40 ($O_{..}$)**



This table represents the *observed* data, and is in effect a count (referred to as *frequencies* or *frequency counts*) of the number of individual employees that fall into each possible category. The frequencies 5, 9, 5, 5, 7, and 9 in the contingency table are called ***cell*** *frequencies*. For example, there are 5 employees that are managerial staff (Status = 1) **and** who also responded 'yes' to the question about compulsory HIV testing (HIV-Employ=1).

These observed frequency values can be referred to by the general symbol O_{ij} where 'i' indicates the row number and 'j' indicates the column number. These symbols are given in brackets in the table above so that we have:

$\mathbf{O_{11}}$ = 5 ... (row 1, column 1)
$\mathbf{O_{12}}$ = 9 ... (row 1, column 2)
$\mathbf{O_{13}}$ = 5 ... (row 1, column 3)
$\mathbf{O_{21}}$ = 5 ... (row 2, column 1)
$\mathbf{O_{22}}$ = 7 ... (row 2, column 2)
$\mathbf{O_{23}}$ = 9 ... (row 2, column 3)

The frequencies in each row are summed to produce a row total, called the ***row frequencies***, and are indicated as follows:

$\mathbf{O_{1.}}$ = 19 ... (row 1)
$\mathbf{O_{2.}}$ = 21 ... (row 2)

The ***column frequencies*** are the totals for each column in the table, and are indicated as:

$\mathbf{O_{.1}}$ = 10 ... (column 1)
$\mathbf{O_{.2}}$ = 16 ... (column 2)
$\mathbf{O_{.3}}$ = 14 ... (column 3)

The total number of persons in the sample is 40 and this overall frequency can be indicated as follows:

$\mathbf{O_{..}}$ = 40 ... (total)

Note that if you add the row frequencies together you should get the same number as you get from adding the column frequencies, and this is the same as for adding the value of each individual cell (the O_{ij} values), since this represents the original sample size.

It is important to note that the relation between the variables, in this case *Status* and *HIV-Employ* is described by the ***cell*** and not by the row or column frequencies. These cell frequencies represent the way the information is distributed relative to the two variables *Status* and *HIV-Employ*. These cell frequencies are often referred to as the ***observed*** or ***empirical*** cell frequencies.

The question now is: How would these cell frequencies be distributed under the null hypothesis, that is, if H_0 is actually true? Asked differently: What are the ***expected*** frequencies if the two categorical variables are truly independent?

We can indicate these ***expected cell frequencies*** by E_{ij} and they are computed as follows:

$\mathbf{E_{11} = (O_{1.} \times O_{.1})/O_{..}} = (19 \times 10)/40 = 4.75$ ... (row 1, column 1)
$\mathbf{E_{12} = (O_{1.} \times O_{.2})/O_{..}} = (19 \times 16)/40 = 7.60$ ... (row 1, column 2)
$\mathbf{E_{13} = (O_{1.} \times O_{.3})/O_{..}} = (19 \times 14)/40 = 6.65$ ... (row 1, column 3)
$\mathbf{E_{21} = (O_{2.} \times O_{.1})/O_{..}} = (21 \times 10)/40 = 5.25$ ... (row 2, column 1)
$\mathbf{E_{22} = (O_{2.} \times O_{.2})/O_{..}} = (21 \times 16)/40 = 8.40$ ... (row 2, column 2)
$\mathbf{E_{23} = (O_{2.} \times O_{.3})/O_{..}} = (21 \times 14)/40 = 7.35$ ... (row 2, column 3)

If you study this example, it should be clear that to find the expected frequency for a particular cell, the row total *for that row* is multiplied by the column total *for that column* and this result is then divided by the *overall total*. These expected frequencies show what the results would have been like if the distribution of frequencies through the cells were homogeneous, in proportion to the respective row and column totals.

If the ***observed*** frequencies correspond precisely with the ***expected*** frequencies, we know that the null hypothesis cannot be rejected. But the ***observed*** frequencies will seldom be precisely equal to the ***expected*** frequencies – *even if H_0 is not rejected* – because of sampling error. In the Table 6.5 below, the ***expected*** cell frequencies are given in brackets after the ***observed*** cell frequencies.

TABLE 6.5: A contingency table for cross-tabulations of *HIV-Employ* versus *Status* with expected frequencies added

		Status			
		1 (Managerial)	2 (Clerical)	3 (Technical)	*Row total*
HIV-Employ	1 (Yes)	5 (4.75)	9 (7.60)	5 (6.65)	**19**
	2 (No)	5 (5.25)	7 (8.40)	9 (7.35)	**21**
	Column total	10	16	14	40

It is the differences between these expected and observed frequencies that interest us, that is, we want to know how far the actual (observed) results are removed from the expected situation, if there is no interaction effect.

6.2.3 Computing the chi-square test statistic

The Pearson chi-square test statistic, which we indicate with χ^2_p, is a calculation of the difference between the observed and expected frequencies.

The formula is

$$\chi^2_p = \Sigma_{ij} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

and what this means is that the expected value for each cell in the contingency

table is subtracted from the observed value for that cell, squared, and divided by the expected value for that cell. Then all of these terms are added together to yield χ_p^2.

The easiest way to do this is to begin by creating a table, such as the one below.

TABLE 6.6: Calculations for computing the chi-square test statistic (χ_p^2)

Row	Column	O	E	(O-E)	$(O-E)^2$	$(O-E)^2/E$
1	1	5	4.75	0.25	0.0625	0.0132
1	2	9	7.60	1.40	1.9600	0.2579
1	3	5	6.65	–1.65	2.7225	0.4094
2	1	5	5.25	–0.25	0.0625	0.0119
2	2	7	8.40	–1.40	1.9600	0.2333
2	3	9	7.35	1.65	2.7225	0.3704
						$\chi_p^2 = 1.2961$

In this table **O** represents the observed frequencies, **E** the expected frequencies, **(O-E)** these expected frequencies subtracted from observed frequencies (for that row), $\mathbf{(O-E)^2}$ that value squared, and $\mathbf{(O-E)^2/E}$ is this last value divided by the expected frequencies. All of the values in the last column are then added together (Σ) to produce the χ^2 test statistic (χ_p^2).

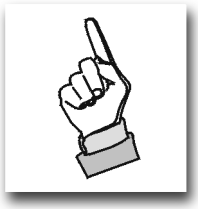
Note that if the difference between the ***expected*** frequency and ***observed*** frequency of each cell is zero, the χ_p^2-value will be zero. The larger the differences between expected and observed frequencies, the larger the χ_p^2-value will be. A large χ_p^2 statistic will, therefore, clearly lead to the rejection of H_0. But how large must this χ_p^2 value be?

The formal probability distribution of chi-square (χ^2) is known, and the values of χ_p^2 as yielded by the formula above are distributed approximately according to this distribution, with **(r-1)(k-1)** degrees of freedom, where **r** represents the number of rows and **k** is the number of columns. For our example:

$$\mathbf{(r-1)(k-1) = (2-1)(3-1) = 1 \times 2 = 2}$$

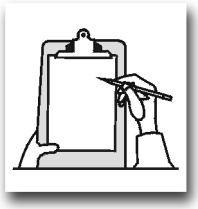
The value of χ_p^2 with the degrees of freedom can be used to calculate the p-value, which is to be compared to our level of significance (remember, we have chosen $\alpha = 0.05$ before). The p-values are calculated by computer, or by looking it up in appropriate tables.

With $\chi_p^2 = 1.2961$ and the degrees of freedom at 2, we shall find a p-value of $p = 0.5231$. Since this value is *not* less than the criterion of 0.05, the null hypothesis cannot be rejected. We, therefore, conclude that we have no reason to suppose that employees from different job categories (*Status*) responded differently to the question about whether or not HIV testing should be compulsory for new employees (*HIV-Employ*). The two variables do not seem to be mutually interdependent.



Some points worth noting about the χ^2-test include the following:

1. The value of the test statistic χ^2_p can never be less than zero.
2. The test can only be non-directional (you can hypothesise that a difference exists, but cannot say anything about the direction of the difference). Because of that, a computer would produce the correct value for p. Therefore, this computed value of p should *not* be divided by 2 for the χ^2_p-test (unlike the test for the Pearson's r that we discussed before).
3. The approximation of the test statistic to the χ^2 distribution is conditional on the assumption that the values of the expected cell frequencies are not too small. As a rule of thumb, the expected frequency for each cell should be 5 or more. (Expected frequencies can sometimes be increased by combining categories.) This rule need not be enforced too rigidly; in some cases, expected frequencies of as small as 1.5 may be acceptable, if there are not too many such small expected frequencies and if the number of cells is fairly large. However, no expected frequencies may be as small as zero (you cannot divide by zero, so you cannot use the equation for χ^2_p if $E_{ij} = 0$ for any row i or column j).
4. While this chi-square test is usually recommended when both values are categorical or on a nominal level of measurement, this is not an absolute requirement for the test. In fact, the test is applicable for any data where the information can be organised into categories and can be represented by a contingency table.
5. The χ^2-test is an example of a *non-parametric test*. No use was made of such descriptive parameters as means or standard deviations (given that it is used for nominal level measurements these values would not be available). (See Topic 1 (section 1.4.2) on parametric vs. non-parametric testing.)

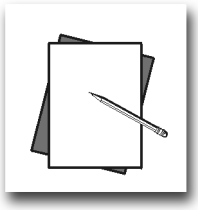


Summary of ideas in this topic

Having completed this topic, you should

- be able to express the linear relationship between two variables as a Pearson product-moment correlation coefficient **r** (you do not need to memorise the actual equation).
- be able to recognise a *Pearson's* **r** as a value between –1 and +1.
- be able to interpret what a particular value of **r** implies about the type of relationship that exists.
- know when a linear relationship is positive or negative, or non-existent.
- know how to create and interpret a scatter plot.
- know how to set up hypotheses for testing the significance of a particular value of **r**.
- know how to interpret the result of a particular test for the significance of **r**.
- know that a significant correlation need not imply that one variable causes the other.
- be able to interpret the relationship that exists between the Pearson's r and the sample size **n**.
- know how to calculate the effect size for a correlation coefficient ($\mathbf{r^2}$).
- know that the correlation coefficient **r** cannot be computed if either or both variables are categorical, that is, were measured on a nominal scale of measurement.

- know that if variables fall on a nominal scale, Pearson's chi-square (χ^2) test can be used to test for a relationship.
- know that the chi-square test compares the observed sample data with the way that the data would be distributed if there were no interaction between variables, that is, if the null hypothesis were true.
- know when to use the χ^2 test, that is, when data can be classified or cross-tabulated into cells and none of the expected cell frequencies are too small, and are definitely larger than zero.



Multiple-choice questions and solutions

Questions:

1. Two of the following are correlation coefficients. Which one is not valid?

   1. –0,02
   2. 1.0
   3. 1.1

2. Which best describes the purpose of calculating a correlation coefficient? To determine

   1. the strength of the linear relationship between two variables.
   2. whether a relationship exists between two variables.
   3. whether the relationship between two variables is positive or negative.

3. A correlation coefficient can vary

   1. from 1 to 10.
   2. from –1 to 1.
   3. from 0 to 1.

4. If you get a Pearson correlation of 0 between two variables, you may conclude that

   1. there is no correlation at all.
   2. there is no linear correlation.
   3. as it is impossible to get a correlation of 0, a calculation error was made.

5. What do you call a variable that can take on only one of two possible values?

   1. dichotomous
   2. biserial
   3. binary

6. Another name for the Pearson's r is

   1. Spearman's rho.
   2. product-moment correlation.
   3. Pearson's chi square.

7. The Pearson's r is likely to be used to establish a relationship between two variables when

   1. both variables are measurements on a nominal or categorical scale.

2. continuous variables are correlated with nominal variables.
3. both variables are continuous measurements.

8. The correlation between the two variables, Predictability and Locus-of-Control, is given as r = –0.2861. This means that

   1. as the value of Predictability becomes smaller, so does the value of Locus-of-Control.
   2. as the value of Predictability becomes smaller, the value for Locus-of-Control becomes larger.
   3. as the value of Locus-of-Control becomes larger, so does the value for Predictability.

9. A correlation coefficient of r = 0.3531 was found between Duration and Control. If the original data are to be plotted on a scatter plot, it will be found that

   1. the dots in the scatter plot cannot be approximated to a straight line.
   2. the dots will approximate to a straight line from bottom left to top right.
   3. the dots will approximate to a straight line from top left to bottom right.

10. If duration were to be correlated with itself, the result would be

    1. r = 1.
    2. r = 0.
    3. r cannot be calculated, for a variable cannot be correlated with itself.

11. The Pearson r represents the relationship between

    1. the observed and the expected frequencies of the data.
    2. the covariance of two variables and the product of the variance of each.
    3. the variance of one variable and the variance of the other.

12. Suppose the alpha-value was set to $\alpha = 0.05$. A chi-square test statistic of $\chi^2 = 3.5$ is calculated, and it is found that the matching p-value is 0.04. Which of the following statements is true?

    1. The null hypothesis can be rejected since 0.05 is less than 3.5.
    2. The null hypothesis cannot be rejected since 0.05 is not less than 0.04.
    3. The null hypothesis should be rejected since 0.04 is less than 0.05.

13. When you want to establish whether a relationship exists between two nominal-level or categorical variables, the best test to use is

    1. the Pearson r.
    2. the t-test.
    3. the chi-square test.

14. What is the expected frequency (given the null hypothesis) for the top left cell in the following contingency table used to calculate the chi-square statistic?

	Group A	Group B	Column total
Type 1	2	8	10
Type 2	8	2	10
Row total	10	10	20

1. 5
2. 2
3. 8

15. The chi-square test statistic χ_p^2 is used to compare
    1. the observed frequency distribution and the expected frequency distribution.
    2. the covariance of two variables and the variance of each.
    3. the variance of a variable and its frequency distribution.

16. Which of the following values of Pearson's r represents the strongest relationship between two variables?
    1. r = 0.058
    2. r = 0.44
    3. r = –0.70

Solutions:

1. The answer is (3). A correlation coefficient can never be larger than 1.
2. Option (1) provides a good description of what a correlation coefficient is used for.
3. The correct answer is (2).
4. The correct answer is (2). Two variables exhibiting a correlation coefficient of close to zero may actually show a good non-linear correlation.
5. The correct answer is (1).
6. The correct answer is (2).
7. The correct answer is (3). Correlation cannot be used for nominal scale measurements, except in the case of dichotomies (i.e. only two categories).
8. The correct answer is (2). A negative correlation implies that, as one variable gets larger, the other diminishes.
9. The correct answer is (2). This is a positive correlation, so, as one variable gets larger, so does the other.
10. The correct answer is (1). A variable correlated with itself will form a perfect correlation, since the variables being compared are perfect matches.
11. The correct answer is (2). Option (1) refers to the chi-square test.
12. The correct answer is (3). The general rule in hypothesis testing is to see if the calculated p-value is equal to or less than the predefined level of significance (α), in which case the null hypothesis can be rejected.
13. The correct answer is (3). The Pearson correlation coefficient cannot be used for nominal scale measurements (except when both measurements are dichotomies, in which case more appropriate tests than Pearson's r exist), nor can the t-test be used in such a case.
14. The correct answer is option 1. Multiply the column total for group A with the row total for type 1, and divide by the overall total:

$$E = \frac{(10 \times 10)}{20} = \frac{100}{20} = 5$$

15. The correct answer is (1). The second option (2) refers to the Pearson's r and option (3) makes no sense at all.
16. The correct answer is (3). The fact that the correlation is negative indicates that, as one variable gets larger, the other gets smaller, but the absolute

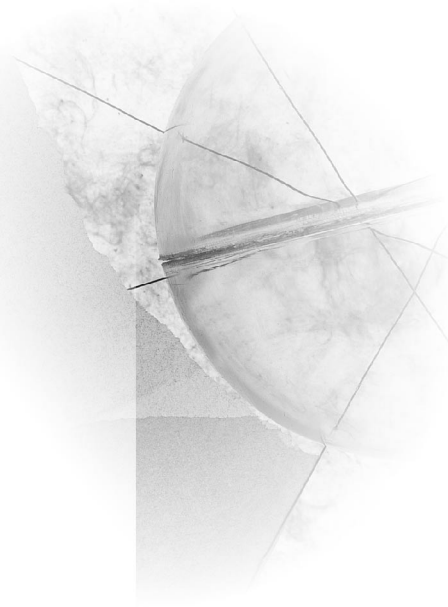
value of the correlation coefficient (ignoring the sign) shows how strong the correlation is.

Solution to the exercise on computing a correlation coefficient (section 6.1.2)

The actual Pearson's correlation between Pre-knowledge and Pre-attitude comes to r = –0.309: so there does not seem to be a correlation between Knowledge and Attitude; in fact, the calculated relationship is negative. If you had set up an hypothesis, it would probably have been that Knowledge and Attitude are positively related, so this result clearly implies that the null hypothesis cannot be rejected in its favour. Even if you test for any relationship at all (a non-directional test), the relationship is not statistically significant (p = 0.052, which is not smaller than α, if this was chosen as 0.05), so we conclude that no actual relationship exists. Note that we draw this conclusion even though the relationship is nearly significant. Once you have selected your critical value for α, you should keep to it.

APPENDIX A

AIDS evaluation scenario



Throughout this study guide we refer to this example as the 'AIDS evaluation scenario' and we show how statistical tests are performed on the data. The scenario is typical of the type of situation that psychology students and graduates are likely to encounter, and our purpose in this study guide is to impart the necessary quantitative analysis skills so that you will be able to draw sensible conclusions and make useful recommendations when faced with such a scenario. We suggest that you read through the scenario carefully before doing anything else, as many of the explanations in the rest of the study guide refer to this scenario. However, if at any point you find that you can't quite remember what the scenario is about, you can always return to this page and read it through again.

AIDS evaluation scenario

Recently, a large South African company noticed that its employees were attending far more funerals than in the past and that more employees were taking sick leave. It appeared that this situation was due to the HIV/AIDS pandemic, as almost all employees had reported losing friends or family to the disease, and some were known to be infected themselves. There were also cases where HIV-infected employees had been badly treated by colleagues, who told them that it was their fault that they had been infected and that they should go and die at home. It appeared that, in most cases, it was male employees who made such remarks, but some female employees had also reportedly done so too.

Clearly, there was a need for a training programme that would educate all employees, making them aware of the facts about HIV/AIDS, and assisting them

in dealing with their attitudes about the disease and their behaviour towards people with HIV/AIDS.

The company asked the Unisa Centre for Applied Psychology (UCAP) to hold a three-day workshop on HIV/AIDS. The workshop was based on an experiential learning process. By actively interacting with others and by doing 'fun' things and reflecting on them, participants were encouraged to explore their attitudes to and knowledge (or lack of knowledge) of HIV/AIDS issues.

In addition, the Centre was contracted to conduct research to determine whether or not the training was successful. A list of all male and female employees in the company was made available. From this list, the trainers selected a random sample of 20 male and 20 female employees. A random procedure was then used to divide the 20 males into a group A (10 employees) and a group B (10 employees). In the same way, a group A and a group B were obtained for the female employees. It was then decided randomly which one of the male and female groups would serve as the treatment group (consisting of 10 male and 10 female employees), and which as the control group (also consisting of 10 male and 10 female employees). The employees in the treatment group were then invited to participate in the three-day workshop. Fortunately, they were all able to do so. The control group did not participate in the workshop. The purpose of this group was to control for nuisance variables; in effect, to make sure that any changes in the attitudes of the employees that were due to influences in the general environment or merely due to the passage of time, were accounted for.

The trainers applied a questionnaire before (pre) and after (post) the workshop to all 40 employees; in other words, the questionnaire was applied to both the treatment and the control group. Towards the end of the workshop, many of the participants in the treatment group gave very positive feedback on the training, making comments such as, 'It has opened my eyes', 'It was not just a good experience, but a WOW experience', 'We gained in knowledge and understanding, but, most importantly, it challenged me very deeply – and personally'. These qualitative data were analysed and included in a report that was submitted to management and presented as a lecture to all employees.

However, the quantitative data still needed to be statistically analysed. This is what we shall do in this module.

Data sources

The UCAP evaluators collected the following kinds of data –

Biographical variables: Age, gender, educational level and job category (status) of each employee.

Questionnaire variables: Pre- and post-scores on 'knowledge' and 'attitudes' for each employee. Employees had to answer a series of factual questions about HIV/ AIDS (such as, 'Can one get HIV/AIDS from other people sneezing?'), and each person's knowledge score was the number of questions he/she answered

correctly. Employees also had to say if they agreed or disagreed with statements such as, 'I think HIV-positive people should not be allowed to work', and each person's attitude score was the number of times he/she indicated a constructive attitude to people with HIV/AIDS. So a person with a knowledge score of 10 and an attitude score of 12 knows more about HIV/AIDS and has a more constructive attitude than somebody with a knowledge score of eight and an attitude score of five.

As an extra check regarding their attitudes to HIV/AIDS infected individuals, employees were asked the following question at the beginning of the project (i.e. before the workshop): 'Do you believe that HIV-testing should be made compulsory, before an individual is appointed to any position in the company?'

They were required to answer either 'Yes' or 'No'.

Biographical and questionnaire data were available for each of the 20 employees who attended the workshop (the treatment group; this is sometimes also referred to as the 'experimental' group) as well as for the 20 employees who did not attend the workshop, but were part of the control group. The evaluators expected that the employees who attended the workshop would get higher knowledge and attitude scores than those who did not, and that employees who attended the workshop would get higher knowledge and attitude scores after the workshop (post-test scores) than before the workshop (pre-test scores). It was also expected that men would have less constructive attitudes (would get lower attitude scores) than women.

The data

Table 1 shows all the quantitative data available for evaluating the impact of the workshop. Each of the columns in the table represents a particular variable (there are 10 variables), while each row represents a person or subject (there are 40 subjects).

What the variables mean

- *Subject number* is simply a number from 1 to 40 given to each employee who was included in the study – subject numbers are useful to help one keep the data organised and to ensure that data are not lost.
- *Group* indicates if the employee was in the group of employees who participated in the workshop or in the group that did not participate in the workshop; here 1 stands for 'belongs to the group that participated in the workshop' (also known as the *treatment group* because they are the participants who underwent the 'treatment' of attending the workshop) while 2 stands for 'belongs to the group that did not participate in the workshop' (also known as the control group). As you can see, employees 1 to 20 participated in the workshop (they were in the treatment group), while employees 21 to 40 did not (they were in the control group). Age is each employee's age, rounded to the nearest year.
- *Gender* indicates whether each person is female (1) or male (2).

- *Status* refers to an employee's job category in the organisation, namely, Managerial (1), Clerical (2) or Technical Services (3).
- *HIV-employ* reflects their answers to the question on whether HIV testing should be compulsory, that is, either 'Yes' (1) or 'No' (2).
- *Pre-attitude* is each person's attitude score before the training course.
- *Post-attitude* is his or her attitude score after the course.
- *Pre-knowledge* is each person's knowledge score before the course.
- *Post-knowledge* is his or her knowledge score after the course.

You will notice that employees in the control group (employees 21 to 40) also have pre- and post-scores. This is because, like those in the treatment group, they also completed the questionnaire twice – once before and once after the workshop – even though they did not attend the workshop.

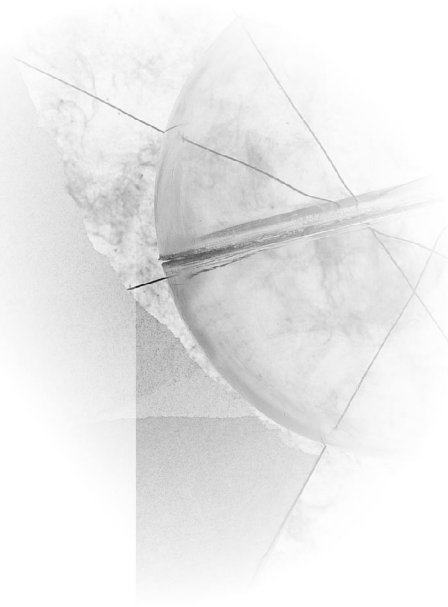
TABLE A.1: AIDS evaluation data

Case	Group	Age	Gender	Status	HIV-Employ	Pre-Attitude	Post-Attitude	Pre-Knowledge	Post-Knowledge
1	1	4	1	2	1	14	21	8	9
2	1	30	1	1	1	20	23	7	12
3	1	40	1	3	1	24	23	3	7
4	1	35	1	2	1	21	26	4	11
5	1	34	1	3	2	21	24	6	6
6	1	29	1	1	2	22	21	2	4
7	1	27	1	2	2	23	25	8	10
8	1	28	1	3	1	24	25	5	12
9	1	30	1	2	1	20	25	9	15
10	1	43	2	2	1	18	23	4	10
11	1	35	2	2	2	12	16	6	6
12	1	29	2	1	1	14	19	7	15
13	1	33	2	1	2	17	19	5	12
14	1	35	2	1	1	19	21	6	9
15	1	48	2	3	2	19	24	7	14
16	1	38	2	1	2	15	19	6	11
17	1	44	2	2	2	17	18	8	13
18	1	39	2	3	2	17	23	6	9
19	1	37	1	2	1	19	22	7	13
20	1	40	2	3	2	14	16	7	15
21	2	43	1	3	1	15	18	9	8
22	2	30	2	3	2	19	22	6	8
23	2	45	1	2	2	25	21	4	2
24	2	51	1	2	1	20	22	5	6
25	2	45	1	3	2	22	22	5	8
26	2	29	1	1	1	21	18	3	2

Case	Group	Age	Gender	Status	HIV-Employ	Pre-Attitude	Post-Attitude	Pre-Knowledge	Post-Knowledge
27	2	37	1	2	1	24	23	9	9
28	2	40	2	3	1	23	23	4	2
29	2	38	2	1	2	19	25	10	11
30	2	33	2	2	2	19	24	3	3
31	2	46	2	3	2	11	16	7	7
32	2	34	2	2	1	15	18	6	3
33	2	43	1	3	2	16	17	6	9
34	2	49	2	3	1	21	17	7	3
35	2	46	2	2	2	18	25	6	12
36	2	33	2	3	2	16	18	7	3
37	2	51	2	2	2	17	18	9	8
38	2	38	1	1	1	18	23	4	4
39	2	45	1	2	1	19	22	9	9
40	2	40	1	1	2	14	16	6	6

APPENDIX B

Measurement levels



As explained in Topic 1, we measure things (like psychological constructs) by allocating numbers according to some rule. The *level* of a measurement matters, because it indicates which operations we can perform on the data, and also helps us to interpret it. We usually distinguish between four levels of measurement: the nominal, ordinal, interval and ratio levels of measurement.

1. *Nominal scale*
   On a nominal scale, numbers show category membership, but are otherwise arbitrary. They do not represent a size or intensity of something, but are only used as labels to distinguish among qualities or characteristics. They can also be referred to as categorical variables, or qualitative variables. This is because differences in the numbers represent differences in quality, character or type, but not in amount.

   For example, we could code a variable like 'region' into 1 = North; 2 = West; 3 = South; and 4 = East. But these four categories can be coded in a different sequence if we choose, without any information being lost. Note in the special case where there are only two options, for example, when we code 'Gender' as 1 = male and 2 = female, we refer to it as a *dichotomy*.

   The important point about nominal scale measurements is that you cannot do arithmetic with them. Adding them and obtaining an average makes no sense (e.g. adding telephone numbers to obtain an 'average telephone number').
2. *Ordinal scale*
   This scale is used only to place objects in an order or to order objects along a continuum. Numbers on this scale show the order in which entities are arranged or ranked but not their relative sizes. An example would be to arrange a class of school children in order of tallness: tall =1, medium height =2 and short = 3; or students can be classified as 1= passed; 2=

obtained a supplementary exam; and 3= failed the exam. The order does tell you something about the relationship among the numbers, but it is not likely that the distance between the numbers is a direct reflection of the distance between the entities, for example, the students' marks. There are some arithmetical manipulations that can be performed on such data, but adding and subtraction would not really work.

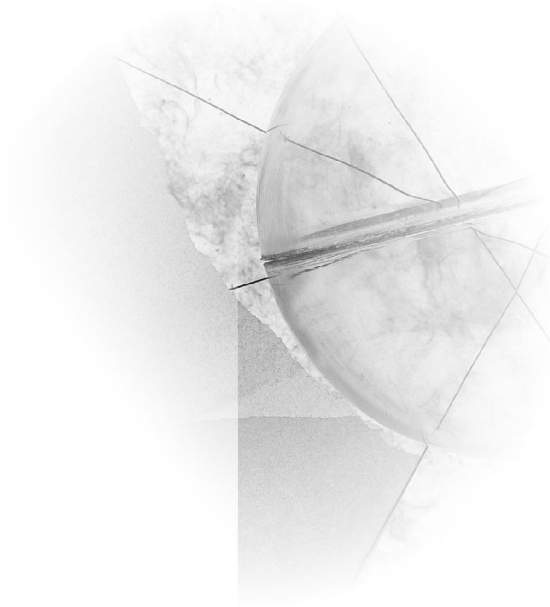
3. *Interval scale*
   On this scale, numbers represent actual sizes or quantities of something, that is, the distances between two numbers expresses a difference in intensity or quantity. However, on an interval scale the zero point is arbitrary. As an example, temperature in centigrade indicates zero as the freezing point of water (but other liquids have different freezing points, and the temperature can fall far below this point).
4. *Ratio scale*
   This level of measurement has all the characteristics of an interval scale, but zero is an absolute null point. For example: age in years, months and days, counted from birth, with zero taken as the moment of birth. All arithmetical manipulations can be performed on both interval and ratio scales.

Note that for the purposes of this course, we only distinguish between nominal scale measurements and measurements on an interval/ratio level. We refer to an interval/ratio level measurement as a quantity or a measurement, and to a nominal level measurement as a category or categorical variable.

Although statistical procedures exist for variables that were measured on an ordinal scale, in this course we are not concerned with any of these.

APPENDIX C

Descriptive statistics



Descriptive statistics are values used to describe the data. They are summary values that give some indication of an important characteristic of the data. The specific descriptive statistics that we deal with in this module are given below. We refer to these values as *parameters* when we are talking about data from a whole population, or as *statistics* when we are talking about a sample. (The difference between a population and a sample is described in Topic 1 of this study guide.)

◆ The arithmetic mean

The mean is one of the measurements used to show the central tendency of the data. It is used to determine where the data is centred, or the score that is most representative of the data. It is calculated by adding all the values of a measurement divided by the number of scores.

The symbol $\bar{x}$ is conventionally used to indicate a sample mean. For a sample of size n, where the measurements are indicated by x, the formula is the following:

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

Note that in some literature, the symbol used to indicate the sample mean is ***M***.

In the case of a population of data, the formula is the same, but the symbol used is μ. The only difference is that the mean is calculated using *all* the members of the population.

◆ The standard deviation

The way that a set of measurements is spread out or clustered together is its variability. One way to measure this would be to take the range; that is, the smallest value subtracted from the largest value for a variable. This measurement is, however, very sensitive to outliers: values that are very much larger or smaller than the rest. A more useful measurement of variability is the *standard deviation*. The symbol for a sample standard deviation is **s**, and for a population standard deviation, it is σ.

Calculating the standard deviation

There are two possible formulae to consider.

1. If you want to calculate the **sample standard deviation** (s) to use as an estimate of the population standard deviation, you would use the following formula:

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

   We divide by n-1 when we calculate the standard deviation using *sample data*, because it gives us a better estimate of the population standard deviation than dividing by n would give.
2. If we want to calculate the standard deviation of the specific data that we have for example, where we have the measurements for the *whole population* (so we do not need to estimate it) we can calculate the **population standard deviation** (σ) directly with the following formula:

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{n}}$$

   Note that the formula is the same as before, except that here we divide by **n** rather than **n-1**. We can use $\bar{x}$ (the sample mean) as an estimate of μ (the population mean) for a reasonably sized sample ($n > 30$) or for a sample that is more or less normally distributed.

Example:

The example that follows relates to Topic 2, section 2.4.1, where it is required that the standard deviation for a whole population is to be calculated.

Imagine the following measurements represent the ages of an entire population of five medical residents working in the emergency section (ER) of Johannesburg General Hospital in a particular year:

33; 28; 45; 43; 47.

To calculate the standard deviation, we first have to calculate the mean. There are five x-scores in the population data set, so the mean of this is:

$$\mu = \frac{\Sigma x}{n} = \frac{(33 + 28 + 45 + 43 + 47)}{5} = \frac{196}{5} = 39.2$$

(If we were estimating population statistics from a sample, we would use the same formula for the sample mean $\bar{\mathbf{x}}$, which would be used as an estimate of $\boldsymbol{\mu}$ for normally distributed data.)

We need to calculate the population standard deviation (σ) directly, since all of the population measurements are given. The formula we would use is, therefore, the second one given above; i.e.:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

We can now calculate the population standard deviation of the data as follows:

$$\begin{aligned}
\sigma &= \sqrt{\frac{\sum(x-\mu)^2}{n}} \\
&= \sqrt{\frac{(33-39.2)^2+(28-39.2)^2+(45-39.2)^2+(43-39.2)^2+(47-39.2)^2}{5}} \\
&= \sqrt{\frac{(-6.2)^2+(-11.2)^2+(5.8)^2+(3.8)^2+(7.8)^2}{5}} \\
&= \sqrt{\frac{38.44+125.44+33.64+14.44+60.84}{5}} \\
&= \sqrt{\frac{272.80}{5}} = \sqrt{54.56} \approx 7.39 \text{ (rounded off)}
\end{aligned}$$

Note that if we were calculating the *sample* standard deviation, the only difference would be that **n** in the formula above would be replaced by **n – 1**.

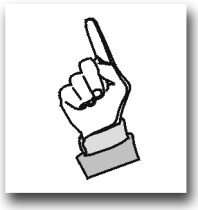
◆ The variance

The variance is just the square of the standard deviation. Conversely, the standard deviation is the square root of the variance. Variance gives an indication of how much the data varies around the mean; the 'width' of the distribution (in both directions). The advantage of using standard deviation is that it is expressed in the same units (the same measurement scale) as the original data, while the variance represents a measurement in squares (x^2).

- For a sample, the variance is $\mathbf{s^2}$
- For a population, the variance is $\boldsymbol{\sigma^2}$

◆ The correlation coefficient

Correlation is a measurement of the extent to which two variables vary together. One such measurement, the *Pearson product-moment correlation coefficient* is described in Topic 6 of this Guide. This coefficient is symbolised with **r** for a sample and ρ for a population.



You should take careful note of the following important distinctions between *samples* and *populations*. Summary values for populations are called 'parameters' and are usually denoted by Greek letters, while summary values for samples are called 'statistics' and are denoted by Roman letters. The following table contains some familiar summary values that you will encounter in this study guide and the symbols that are used for them.

Summary value	**Symbol**	
	Parameter	**Statistic**
Arithmetic mean	μ	$\bar{x}$
Standard deviation	σ	s
Variance	σ^2	s^2
Standard error of mean	$\sigma_{\bar{x}}$ $(=\sigma/\sqrt{n})$	$s_{\bar{x}}$ $(= s/\sqrt{n})$
Mean of sampling distribution of the mean	$\mu_{\bar{x}}$	
Correlation between two measurements	ρ	r

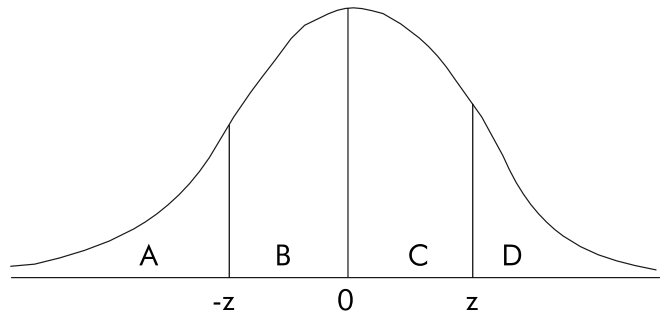
A researcher seldom knows the values of the population parameters, but the values of sample statistics can be calculated by means of clearly formulated mathematical procedures, and these can be used as estimates of the parameters of the corresponding population.

APPENDIX D

z-scores and areas under the normal curve

The z-distribution refers to a variable that is normally distributed with a population mean of $\mu = 0$ and a standard deviation of $\sigma = 1$ (generally referred to as the *standardised normal distribution*). This is explained in Topic 2 (section 2.3.3).

The table that follows gives probabilities in terms of the area under the curve of the standard normal distribution for a particular value of z.



In the figure above, the horizontal represents the scale of z-values. A, B, C and D represent various portions or areas under the curve. The various columns in the tables that follow give us the following information:

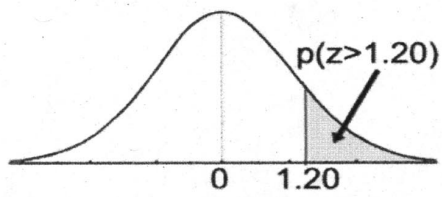
- The first column gives z-values 0 and larger.
- The column marked 'larger portion' refers to the portions A + B + C and gives the probability of a score equal to or less than z (for $z > 0$).
- The column 'smaller portion' refers to the portion D and gives us the probability of a score equal or greater than z (for $z > 0$).

- The 'mean to z' column refers to portion C and gives us the probability of a score between 0 and z (remember that the mean z-score is equal to 0).
- Because the z-distribution is symmetrical, values of $z < 0$ is just the inverse of the values for $z > 0$. Areas to the left of z are exactly the same as to the right of +z. To find portions or areas associated with negative values, note the following:
  - Portion A = Portion D
  - Portion B = Portion C

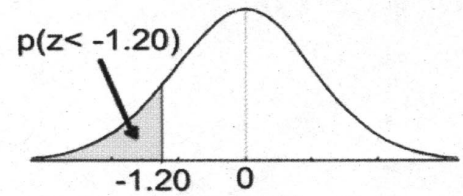
Also note the following:

- The total area is 1; that is, A+B+C+D= 1 [the whole area under the curve is defined as 1 (a p-value can never be larger than 1 and p = 1 is a probability of 100%)]
- Half of the area is A+B = C+D = 0.5 [What this 0.5 implies, is that half of the area (0.5 of 1) is to the right of 0 (the mean), and the remaining half of the area is to the left of 0 (also 0.5 of 1), and the probability of z falling in any particular half of the distribution is ½ = 0.5]

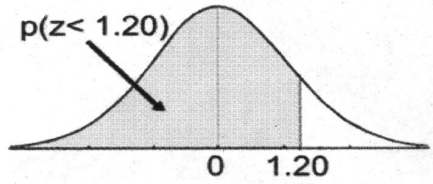
Examples



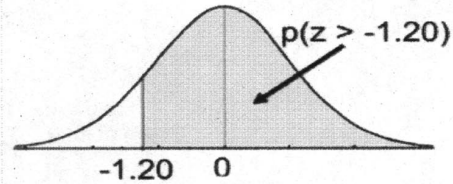
The probability of a z-score larger than 1.20 is 0.1151



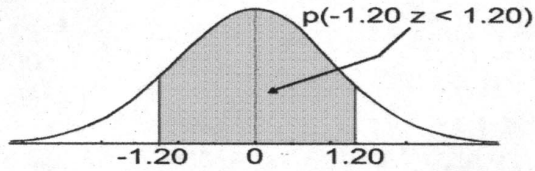
The probability of a z-score smaller than -1.20 is 0.1151



The probability of a z-score smaller than 1.20 is 0.8849 (or 1 - 0,1151 = 0.8849)



The probability of a z-score greater than -1,20 is 0.8849 (or 1 - 0,1151 = 0.8849)



The probability of a z-score between -1.20 and +1.20 is given by
$p(-1.20 < z < 1.20) = p(z < 1.20) - p(z < -1.20)$
$= 0.8849 - 0.1151 = 0.7698.$

Appendix D *(continued)*

The standard normal distribution (z)

z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
.00	0.0000	0.5000	0.5000	.45	0.1736	0.6736	0.3264
.01	0.0040	0.5040	0.4960	.46	0.1772	0.6772	0.3228
.02	0.0080	0.5080	0.4920	.47	0.1808	0.6808	0.3192
.03	0.0120	0.5120	0.4880	.48	0.1844	0.6844	0.3156
.04	0.0160	0.5160	0.4840	.49	0.1879	0.6879	0.3121
.05	0.0199	0.5199	0.4801	.50	0.1915	0.6915	0.3085
.06	0.0239	0.5239	0.4761	.51	0.1950	0.6950	0.3050
.07	0.0279	0.5279	0.4721	.52	0.1985	0.6985	0.3015
.08	0.0319	0.5319	0.4681	.53	0.2019	0.7019	0.2981
.09	0.0359	0.5359	0.4641	.54	0.2054	0.7054	0.2946
.10	0.0398	0.5398	0.4602	.55	0.2088	0.7088	0.2912
.11	0.0438	0.5438	0.4562	.56	0.2123	0.7123	0.2877
.12	0.0478	0.5478	0.4522	.57	0.2157	0.7157	0.2843
.13	0.0517	0.5517	0.4483	.58	0.2190	0.7190	0.2810
.14	0.0557	0.5557	0.4443	.59	0.2224	0.7224	0.2776
.15	0.0596	0.5596	0.4404	.60	0.2257	0.7257	0.2743
.16	0.0636	0.5636	0.4364	.61	0.2291	0.7291	0.2709
.17	0.0675	0.5675	0.4325	.62	0.2324	0.7324	0.2676
.18	0.0714	0.5714	0.4286	.63	0.2357	0.7357	0.2643
.19	0.0753	0.5753	0.4247	.64	0.2389	0.7389	0.2611
.20	0.0793	0.5793	0.4207	.65	0.2422	0.7422	0.2578
.21	0.0832	0.5832	0.4168	.66	0.2454	0.7454	0.2546
.22	0.0871	0.5871	0.4129	.67	0.2486	0.7486	0.2514
.23	0.0910	0.5910	0.4090	.68	0.2517	0.7517	0.2483
.24	0.0948	0.5948	0.4052	.69	0.2549	0.7549	0.2451
.25	0.0987	0.5987	0.4013	.70	0.2580	0.7580	0.2420
.26	0.1026	0.6026	0.3974	.71	0.2611	0.7611	0.2389
.27	0.1064	0.6064	0.3936	.72	0.2642	0.7642	0.2358
.28	0.1103	0.6103	0.3897	.73	0.2673	0.7673	0.2327
.29	0.1141	0.6141	0.3859	.74	0.2704	0.7704	0.2296
.30	0.1179	0.6179	0.3821	.75	0.2734	0.7734	0.2266
.31	0.1217	0.6217	0.3783	.76	0.2764	0.7764	0.2236
.32	0.1255	0.6255	0.3745	.77	0.2794	0.7794	0.2206
.33	0.1293	0.6293	0.3707	.78	0.2823	0.7823	0.2177
.34	0.1331	0.6331	0.3669	.79	0.2852	0.7852	0.2148
.35	0.1368	0.6368	0.3632	.80	0.2881	0.7881	0.2119
.36	0.1406	0.6406	0.3594	.81	0.2910	0.7910	0.2090
.37	0.1443	0.6443	0.3557	.82	0.2939	0.7939	0.2061
.38	0.1480	0.6480	0.3520	.83	0.2967	0.7967	0.2033
.39	0.1517	0.6517	0.3483	.84	0.2995	0.7995	0.2005
.40	0.1554	0.6554	0.3446	.85	0.3023	0.8023	0.1977
.41	0.1591	0.6591	0.3409	.86	0.3051	0.8051	0.1949
.42	0.1628	0.6628	0.3372	.87	0.3078	0.8078	0.1922
.43	0.1664	0.6664	0.3336	.88	0.3106	0.8106	0.1894
.44	0.1700	0.6700	0.3300	.89	0.3133	0.8133	0.1867

(Continues)

Appendix D *(continued)*

The standard normal distribution (z)

z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
.90	0.3159	0.8159	0.1841	1.35	0.4115	0.9115	0.0885
.91	0.3186	0.8186	0.1814	1.36	0.4131	0.9131	0.0869
.92	0.3212	0.8212	0.1788	1.37	0.4147	0.9147	0.0853
.93	0.3238	0.8238	0.1762	1.38	0.4162	0.9162	0.0838
.94	0.3264	0.8264	0.1736	1.39	0.4177	0.9177	0.0823
.95	0.3289	0.8289	0.1711	1.40	0.4192	0.9192	0.0808
.96	0.3315	0.8315	0.1685	1.41	0.4207	0.9207	0.0793
.97	0.3340	0.8340	0.1660	1.42	0.4222	0.9222	0.0778
.98	0.3365	0.8365	0.1635	1.43	0.4236	0.9236	0.0764
.99	0.3389	0.8389	0.1611	1.44	0.4251	0.9251	0.0749
1.00	0.3413	0.8413	0.1587	1.45	0.4265	0.9265	0.0735
1.01	0.3438	0.8438	0.1562	1.46	0.4279	0.9279	0.0721
1.02	0.3461	0.8461	0.1539	1.47	0.4292	0.9292	0.0708
1.03	0.3485	0.8485	0.1515	1.48	0.4306	0.9306	0.0694
1.04	0.3508	0.8508	0.1492	1.49	0.4319	0.9319	0.0681
1.05	0.3531	0.8531	0.1469	1.50	0.4332	0.9332	0.0668
1.06	0.3554	0.8554	0.1446	1.51	0.4345	0.9345	0.0655
1.07	0.3577	0.8577	0.1423	1.52	0.4357	0.9357	0.0643
1.08	0.3599	0.8599	0.1401	1.53	0.4370	0.9370	0.0630
1.09	0.3621	0.8621	0.1379	1.54	0.4382	0.9382	0.0618
1.10	0.3643	0.8643	0.1357	1.55	0.4394	0.9394	0.0606
1.11	0.3665	0.8665	0.1335	1.56	0.4406	0.9406	0.0594
1.12	0.3686	0.8686	0.1314	1.57	0.4418	0.9418	0.0582
1.13	0.3708	0.8708	0.1292	1.58	0.4429	0.9429	0.0571
1.14	0.3729	0.8729	0.1271	1.59	0.4441	0.9441	0.0559
1.15	0.3749	0.8749	0.1251	1.60	0.4452	0.9452	0.0548
1.16	0.3770	0.8770	0.1230	1.61	0.4463	0.9463	0.0537
1.17	0.3790	0.8790	0.1210	1.62	0.4474	0.9474	0.0526
1.18	0.3810	0.8810	0.1190	1.63	0.4484	0.9484	0.0516
1.19	0.3830	0.8830	0.1170	1.64	0.4495	0.9495	0.0505
1.20	0.3849	0.8849	0.1151	1.65	0.4505	0.9505	0.0495
1.21	0.3869	0.8869	0.1131	1.66	0.4515	0.9515	0.0485
1.22	0.3888	0.8888	0.1112	1.67	0.4525	0.9525	0.0475
1.23	0.3907	0.8907	0.1093	1.68	0.4535	0.9535	0.0465
1.24	0.3925	0.8925	0.1075	1.69	0.4545	0.9545	0.0455
1.25	0.3944	0.8944	0.1056	1.70	0.4554	0.9554	0.0446
1.26	0.3962	0.8962	0.1038	1.71	0.4564	0.9564	0.0436
1.27	0.3980	0.8980	0.1020	1.72	0.4573	0.9573	0.0427
1.28	0.3997	0.8997	0.1003	1.73	0.4582	0.9582	0.0418
1.29	0.4015	0.9015	0.0985	1.74	0.4591	0.9591	0.0409
1.30	0.4032	0.9032	0.0968	1.75	0.4599	0.9599	0.0401
1.31	0.4049	0.9049	0.0951	1.76	0.4608	0.9608	0.0392
1.32	0.4066	0.9066	0.0934	1.77	0.4616	0.9616	0.0384
1.33	0.4082	0.9082	0.0918	1.78	0.4625	0.9625	0.0375
1.34	0.4099	0.9099	0.0901	1.79	0.4633	0.9633	0.0367

(Continues)

Appendix D *(continued)*

The standard normal distribution (z)

z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
1.80	0.4641	0.9641	0.0359	2.25	0.4878	0.9878	0.0122
1.81	0.4649	0.9649	0.0351	2.26	0.4881	0.9881	0.0119
1.82	0.4656	0.9656	0.0344	2.27	0.4884	0.9884	0.0116
1.83	0.4664	0.9664	0.0336	2.28	0.4887	0.9887	0.0113
1.84	0.4671	0.9671	0.0329	2.29	0.4890	0.9890	0.0110
1.85	0.4678	0.9678	0.0322	2.30	0.4893	0.9893	0.0107
1.86	0.4686	0.9686	0.0314	2.31	0.4896	0.9896	0.0104
1.87	0.4693	0.9693	0.0307	2.32	0.4898	0.9898	0.0102
1.88	0.4699	0.9699	0.0301	2.33	0.4901	0.9901	0.0099
1.89	0.4706	0.9706	0.0294	2.34	0.4904	0.9904	0.0096
1.90	0.4713	0.9713	0.0287	2.35	0.4906	0.9906	0.0094
1.91	0.4719	0.9719	0.0281	2.36	0.4909	0.9909	0.0091
1.92	0.4726	0.9726	0.0274	2.37	0.4911	0.9911	0.0089
1.93	0.4732	0.9732	0.0268	2.38	0.4913	0.9913	0.0087
1.94	0.4738	0.9738	0.0262	2.39	0.4916	0.9916	0.0084
1.95	0.4744	0.9744	0.0256	2.40	0.4918	0.9918	0.0082
1.96	0.4750	0.9750	0.0250	2.41	0.4920	0.9920	0.0080
1.97	0.4756	0.9756	0.0244	2.42	0.4922	0.9922	0.0078
1.98	0.4761	0.9761	0.0239	2.43	0.4925	0.9925	0.0075
1.99	0.4767	0.9767	0.0233	2.44	0.4927	0.9927	0.0073
2.00	0.4772	0.9772	0.0228	2.45	0.4929	0.9929	0.0071
2.01	0.4778	0.9778	0.0222	2.46	0.4931	0.9931	0.0069
2.02	0.4783	0.9783	0.0217	2.47	0.4932	0.9932	0.0068
2.03	0.4788	0.9788	0.0212	2.48	0.4934	0.9934	0.0066
2.04	0.4793	0.9793	0.0207	2.49	0.4936	0.9936	0.0064
2.05	0.4798	0.9798	0.0202	2.50	0.4938	0.9938	0.0062
2.06	0.4803	0.9803	0.0197	2.51	0.4940	0.9940	0.0060
2.07	0.4808	0.9808	0.0192	2.52	0.4941	0.9941	0.0059
2.08	0.4812	0.9812	0.0188	2.53	0.4943	0.9943	0.0057
2.09	0.4817	0.9817	0.0183	2.54	0.4945	0.9945	0.0055
2.10	0.4821	0.9821	0.0179	2.55	0.4946	0.9946	0.0054
2.11	0.4826	0.9826	0.0174	2.56	0.4948	0.9948	0.0052
2.12	0.4830	0.9830	0.0170	2.57	0.4949	0.9949	0.0051
2.13	0.4834	0.9834	0.0166	2.58	0.4951	0.9951	0.0049
2.14	0.4838	0.9838	0.0162	2.59	0.4952	0.9952	0.0048
2.15	0.4842	0.9842	0.0158	2.60	0.4953	0.9953	0.0047
2.16	0.4846	0.9846	0.0154	2.61	0.4955	0.9955	0.0045
2.17	0.4850	0.9850	0.0150	2.62	0.4956	0.9956	0.0044
2.18	0.4854	0.9854	0.0146	2.63	0.4957	0.9957	0.0043
2.19	0.4857	0.9857	0.0143	2.64	0.4959	0.9959	0.0041
2.20	0.4861	0.9861	0.0139	2.65	0.4960	0.9960	0.0040
2.21	0.4864	0.9864	0.0136	2.66	0.4961	0.9961	0.0039
2.22	0.4868	0.9868	0.0132	2.67	0.4962	0.9962	0.0038
2.23	0.4871	0.9871	0.0129	2.68	0.4963	0.9963	0.0037
2.24	0.4875	0.9875	0.0125	2.69	0.4964	0.9964	0.0036

(Continues)

Appendix D *(continued)*

The standard normal distribution (z)

z	Mean to z	Larger Portion	Smaller Portion	z	Mean to z	Larger Portion	Smaller Portion
2.70	0.4965	0.9965	0.0035	2.90	0.4981	0.9981	0.0019
2.71	0.4966	0.9966	0.0034	2.91	0.4982	0.9982	0.0018
2.72	0.4967	0.9967	0.0033	2.92	0.4982	0.9982	0.0018
2.73	0.4968	0.9968	0.0032	2.93	0.4983	0.9983	0.0017
2.74	0.4969	0.9969	0.0031	2.94	0.4984	0.9984	0.0016
2.75	0.4970	0.9970	0.0030	2.95	0.4984	0.9984	0.0016
2.76	0.4971	0.9971	0.0029	2.96	0.4985	0.9985	0.0015
2.77	0.4972	0.9972	0.0028	2.97	0.4985	0.9985	0.0015
2.78	0.4973	0.9973	0.0027	2.98	0.4986	0.9986	0.0014
2.79	0.4974	0.9974	0.0026	2.99	0.4986	0.9986	0.0014
2.80	0.4974	0.9974	0.0026	3.00	0.4987	0.9987	0.0013
2.81	0.4975	0.9975	0.0025	...	...	...	...
2.82	0.4976	0.9976	0.0024	3.25	0.4994	0.9994	0.0006
2.83	0.4977	0.9977	0.0023	...	...	...	...
2.84	0.4977	0.9977	0.0023	3.50	0.4998	0.9998	0.0002
2.85	0.4978	0.9978	0.0022	...	...	...	...
2.86	0.4979	0.9979	0.0021	3.75	0.4999	0.9999	0.0001
2.87	0.4979	0.9979	0.0021	...	...	...	...
2.88	0.4980	0.9980	0.0020	4.00	0.5000	1.0000	0.0000
2.89	0.4981	0.9981	0.0019				

Using the z-tables

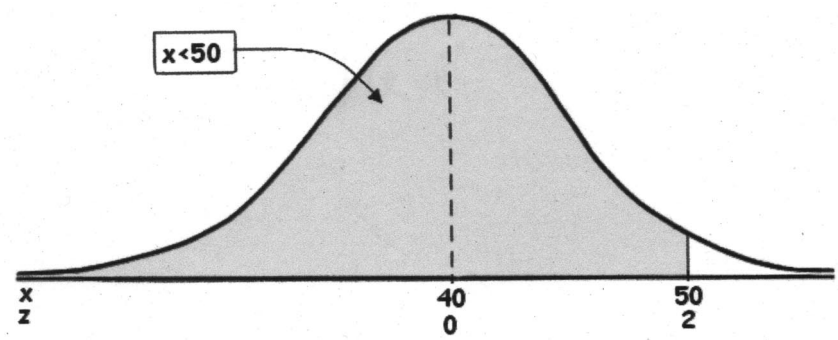
The portion of the table to use ('smaller portion' or 'larger portion') depends on the question.

The following example demonstrates this.

> Determine the probability that x will be less than 50 for a variable drawn from a normal distribution with $\bar{x} = 40$ and $s = 5$.

To answer a question like this, begin by drawing the graph. If the question is to calculate $p(x<50)$, which is just another way of writing 'the probability that x is smaller than 50', for a normal distribution with a mean of $\bar{x} = 40$ and standard deviation of $s = 5$, you already know that you are interested in the *left-hand side* of the graph – everything up to the point where $x = 50$.

But $x=50$ lies to the *right* of the mean (given as $\bar{x} = 40$), so you also know that it is the larger part of the graph that interests you, as it implies the whole area from minus infinity ($-\infty$) up to the point where $x = 50$ – which is above the mean of 40 – that is relevant.



To find the correct information on the table, you first need to transform the x-score into the equivalent point on the z-distribution (using the formula from Topic 2):

$$z = \frac{x - \bar{x}}{s} = \frac{50 - 40}{5} = \frac{10}{5} = 2$$

This implies that an x score of 50 translates to a z-score of 2. It follows that the probability that x is smaller than 50 is *the same as* the probability that z is smaller than 2. This can be written symbolically as $p(x < 50) = p(z < 2)$. Looking at the larger portion under the curve, we find that, according to the z-tables, this area is equal to 0.9772.

So, we may conclude that $p(x<50) = p(z<2) = 0.9772$: the probability that x is smaller than 50 is 0.9772 (or 97.72%).

What if you had to calculate the probability that x is larger than 50?

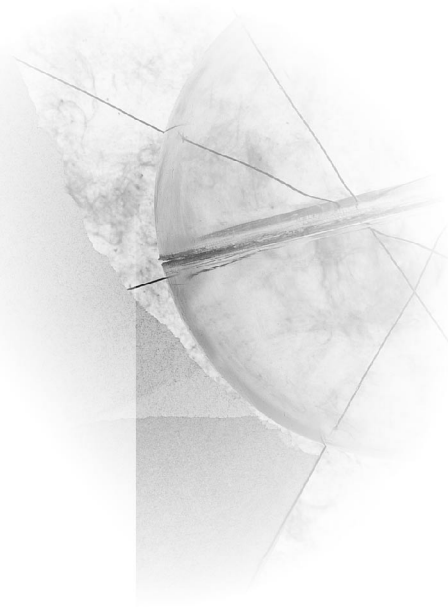
If you had to calculate $p(x>50)$, you would have followed exactly the same reasoning to decide that you need to use the *smaller* portion under the curve, since 50 lies to the *right* of the mean of $\bar{x} = 40$ (which would translate into the

statement that 2 lies to the right of the mean of 0 on the z-distribution). You would now be interested in the portion of the graph under the curve from $z = 2$ on towards the far right-hand side ($+\infty$). Also remember that the left hand side of the normal curve, from minus infinity to $z = 0$, is a mirror-image of the right-hand side. So what is valid to the right of $z = 2$ is also true of the area to the left of $z = 2$.

Note that there is a third possibility given in the table: 'mean to z'. This is the area from the middle of the graph (where $z = 0$) up to a particular z-value. You know that the area to either side of $z = 0$ is equal to 0.5, so this is equal to 0.5 added to get the larger portion, or to the given value subtracted from 0.5 to get the smaller portion. This information might be useful when you want to calculate the area under the curve between two values (as in, e.g., $p(-0.3 < z < 1.2)$: the probability that z lies between –0.3 and 1.2).

APPENDIX E

Review of Arithmetic



What follows is a short refresher of the basic arithmetic operations and some elementary rules of algebra. Although this is not primarily a course in mathematics, you do require some basic numeracy skills and a familiarity with the rules of arithmetic and some very basic algebra. It is included in case you have last seen this in high school, but have never used it since. If you feel you need more exercise, we urge you to find a simple introductory book to the rules of arithmetic.

Numbers

Integers — Whole numbers 0, 1, 2, 3, 4 ... etc., along with their negative values –1, –2, –3 ... etc. These can be written without a decimal component; e.g. 55, 347 and –15 are integers but $^{3}/_{8}$ and 6.34 are not integers. In other words, integers are *discrete* (only whole numbers like 1, 2, 3, and –7, –100, etc. are allowed, with nothing between them).

Real numbers — Real numbers express the whole set of possible numbers (including integers and negative numbers) along a continuum. Real numbers are *continuous*, including all possible fractions between any two numbers, usually expressed in decimal representation. This has the strange implication that between any two real numbers lie an infinite number of real numbers (you can always get a smaller fraction). Examples of real numbers are 7, 66.356, 0.02, –5.5 and $\frac{3}{4}$. Note that a real number can give an infinite decimal representation, such as 1/3=0.333333333 ... etc. Such numbers are usually rounded off. Note that we use decimal points in this module rather than

decimal commas (to conform to the APA rules for publishing), but either points or commas are acceptable, as long as you are consistent.

Fractions — Such numbers are proportions, of the form **a/b**, where **a** and **b** are integers. Any fraction can be written in a decimal form, as the '/' symbol just implies division, but – as mentioned above – not all fractions produce a finite result. A percentage is just a fraction out of 100 ('percent' means 'out of a hundred'). So: 2/5 = 40/100=0.40=40%. Any number divided by itself gives one (**a/a**=1 for any **a**). Note also that you *cannot* divide by zero; i.e. in **a/b** the **b** can *never* be 0 (so 0/0=1 would be false).

Numerator — The number on top of a fraction like the **a** in **a/b.**

Denominator — The number at the bottom of a fraction like the **b** in **a/b**. So in $\frac{7}{12}$ the number 7 is the numerator and 12 is the denominator.

Some standard symbols

+, –, x and ÷ — Add, subtract, multiply and divide, called the *operators* in arithmetic. Note that computer programmers often use the symbol '*' as a multiplication sign (so as not to confuse it with the letter 'X'), while mathematicians often use just a dot (for example a•b = a x b), or they may use no symbol at all: ab = a x b. For division the symbols '÷' and '**/**' are used interchangeably.

= — The equals sign. In **a**=**b** what is on the left must equal what is on the right; or else there is an error.

≠ — Not equal to. In **a** ≠ **b** the two numbers **a** and **b** *must* be different.

≈ — Approximately equal. This can be used when rounding off. For example, 66.975546 ≈ 66.98.

< — Smaller than. In **a** < **b**, **a** must be smaller than **b**. But note that between real numbers the difference can be infinitely small!

⩽ — Smaller than or equals to. In **a** ⩽ **b**, **a** must be equal to **b** or smaller than **b**.

> — Greater than. In **a > b**, **a** must be greater than **b**.

⩾ — Greater than or equal to. In **a** ⩾ **b, a** must be equal to **b** or greater than **b**.

a < x < b — The value of **a** is less than **x** and **x** is less than **b**. This also implies that **a** < **b.**

± — The meaning of a ± b would be '**a plus or minus b**.' It could be used to specify the range from a – b up to a + b.

| a | — '|**a**|' means take the *absolute value* of **a**; i.e. ignore the sign of **a**. So |–24| = |24| = 24. Also |12 – 47|= |–35|=35.

1/a — The reciprocal of **a**. So 1/5 is the reciprocal of 5.

a^2 Square the number **a**. So $23^2 = 23 \times 23 = 529$.

a^n This is **a** raised to the power **n**; i.e. multiplication is repeated **n** times. So $7^4 = 7 \times 7 \times 7 \times 7 = 2401$. Raising something to a power is also called *exponentiation*.

$\sqrt{b}$ The square root of **b**. The number that must be multiplied by itself to produce **b**. Example: $\sqrt{144} = 12$. Note that $\sqrt{b} = b^{\frac{1}{2}}$ and in real numbers you can never get a square root of a negative number: $\sqrt{-c}$ is invalid.

a! This is called a *factorial*, and is defined as $a! = (a-1) \times (a-2) \ldots (1)$. Also $1! = 1$, and $0! = 1$. For example: $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$.

∞ Infinity. A number that is as large as possible. Since there is always a larger number possible, ∞ is not a number but a concept. Negative infinity (the smallest number possible) is indicated by $-\infty$.

Σ Add together. The use of this symbol is discussed further below.

Addition and subtraction

- To subtract a large number from a small one, subtract the smaller from the larger one and make the result negative. For example $6 - 15 = -9$.
- The order of the operations is not important. For example: $-20 + 15 = 15 - 20 = 5$.
- When adding numbers where some have positive signs (e.g. +3) and others have negative signs (e.g. –5), one way to go about it is as follows:

  1. Add all the numbers with positive signs and remember that the answer also has a positive sign. Example: $+3 + 5 + 2 = +10$. Note that numbers with no sign, such as 5 or 6, can actually also be written as +5 and +6.
  2. Add all the numbers with negative signs and remember that the answer has a negative sign. *Example:* $-2 - 3 - 7$ adds up to 12 with a minus in front:

     $-2 - 3 - 7 = -(2 + 3 + 7) = -12$.

  3. Ignore the signs of the two totals and then subtract the smaller from the larger total. Then use the sign of the larger for the answer.

     Example: We want to add $-3 - 4 + 6 - 7 + 4 - 8 + 3$

     - Add the positive numbers: $6 + 4 + 3 = 13$.
     - Add the negative numbers: $3 + 4 + 7 + 8 = 22$.
     - Subtract the smaller number from the larger number: $22 - 13 = 9$.
     - Because the negative numbers came to a larger total ($22 > 13$), put a negative sign before the result, which becomes –9.

     So: $-3 - 4 + 6 - 7 + 4 - 8 + 3 = -9$.

Multiplication and division

- If there is no operator before a set of parentheses, multiplication is implied.

Two symbols next to each other also implies multiplication. So if **a** and **b** are two variables, **ab** = **a x b**.

Example: 2(4)(3)=2 x 4 x 3 = 24. Also (3+7)(–6)=10 x (–6) = –60.

- Two minuses make a plus. For example: –3 x –4 = 12. This is why you cannot take the square root of a negative number. No number divided by itself will produce –12.

  When there are many minuses, we know that the final answer will be a positive number if there is an even number of minuses, such as two, four or six minuses. If there is an uneven number of minuses, such as one, three, five or seven minuses, then the answer will also be a minus number.

  Examples:

  - –2 x –2 = +4 (even number of minuses)
  - –3 x –3 x –2 = +9 x –2 = –18 (uneven number of minuses)
  - –3 x –3 x –2 x –2 = 9 x 4 = 36 (even number of minuses)
  - –3 x (–2) = –3 x –2 = 6 (even number of minuses)
  - –3 x –2(–3) = –3 x 6 = –18 (uneven number of minuses)
  - –12 divided by 3 is the same as +12 divided by –3, namely –4
  - –12 divided by –3 can be written as (–12)/(–3) and the answer is +4

- You can never divide by zero: 6/0 is impossible. It approaches infinity, but 'infinity' (written as ∞) is not really a number; it just means infinitely large. Or, $-\infty$ is infinitely small.

The precedence or order of operations

- There are conventions for the order in which the operations of arithmetic must be performed. The precedence (order in which the operations are performed) of numerical operations is important when we have to find the answer to a complex expression or equation.

  Operations must be performed in the following order:

  1. Do what is in brackets or parenthesis first. If there are brackets within brackets, begin with the innermost set of brackets.
  2. Calculate squares, take square roots and do exponentiation.
  3. Multiply and divide.
  4. Add and subtract.

  Examples:

  3 + 2 x 4 = 3 + 8 = 11; but (3 + 2) x 4 = 5 x 4 = 20.

  Note that in adding and subtracting the order is not important:
  7 – 2 = –2 + 7 = 5

  So also for multiplication and division:

  7 x 10 ÷ 2 = (7 x 10)/2 = 70/2 = 7 x (10/2) = 7 x 5 = 35.
  4 x (6 – (2 x 2)) = 4 x (6 – 4) = 4 x 2 = 8 (start with the innermost set of parenthesis).

Doing calculations with fractions

- In a fraction like **a/b**, **a** is called the numerator and **b** is the denominator. To convert a fraction to a decimal, just divide the numerator by the denominator: 2/5 = 0.4.
- To multiply a fraction with a whole number, multiply the numerator by the number: $6 \times {}^{3}/_{4} = (6 \times 3) / 4 = 18/4$.
- *Multiplying fractions.* To multiply a series of fractions, just multiply all the numerators together, and multiply all the denominators together:

$$\frac{3}{4} \times \frac{2}{5} \times \frac{1}{6} = \frac{3 \times 2 \times 1}{4 \times 5 \times 6} = \frac{6}{120} = \frac{1}{20}$$

  Note how in the last step the fraction was simplified, because 6 can be divided into 120 without remainder.
- *Adding or subtracting fractions.* Fractions with a denominator that is the same can be added together by adding the numerators and dividing by the same denominator:

$$\frac{1}{4} + \frac{3}{4} + \frac{5}{4} = \frac{1+3+5}{4} = \frac{9}{4} = 2.25$$

  To add fractions with different denominators is a bit more complicated. You first have to convert the denominators to the same number. You can use the fact that a number divided by itself equals one, and that a number multiplied by one is equal to itself. The example will make this clear:

$$\frac{2}{3}+\frac{5}{12} = \left(\frac{2}{3} \times \frac{12}{12}\right) + \left(\frac{5}{12} \times \frac{3}{3}\right) = \frac{24}{36}+\frac{15}{36} = \frac{24+15}{36} = \frac{39}{36}$$

  The same principles would apply if you were subtracting fractions instead of adding them.

Basic Algebra

The two basic rules to remember when you try to solve an equation with an unknown term like 'x' are the following:

- If you add or subtract something from one side of an equation (or formula) you should add or subtract the same value from the other side.

  Example: x – 3 = 2

  To remove the '3', add it to both sides: x – 3 + 3 = 2 + 3, which leaves you with x = 5.

  For x + 7 = 4 we would subtract 7 from both sides: x + 7 – 7 = 4 – 7, so x = –3.
- If you multiply or divide a value on one side of the equation, you must also multiply/divide it on the other side (which translates into: if you multiply a term on one side, divide it on the other; if you divide it on one side, multiply it on the other).

Examples:

$3x = 24$: To remove the '3', divide by 3 on both sides: $3x/3 = 24/3$, which means $x = 8$.

$\frac{1}{2} y = 23$: To remove the ½, multiply by 2 on both sides: $2 \times \frac{1}{2} y = 2 \times 23$, so $y = 46$.

Rounding

If you do a series of calculations, maintain all decimal places in intermediate calculations and only round off the final answer to limit rounding-off errors. Final answers can be rounded off to two or three places. To leave a number like 89.4582485432 in this form gives an illusion of accuracy that is misleading.

Let us say we are rounding off to two places after the decimal. The rules are as follows:

- If the remaining decimal fraction to the right of the second decimal position is greater than 5, round the number up. If it is less than 5 then round it down.

  For example: 245.95**6**01 $\approx$ 245.96 and 45.56372 $\approx$ 45.56

- If the remaining fraction to the right of the second decimal is exactly 5, look at the decimal in the third position. If it is greater than 5, round up, else round down.

  Examples: 4.2356 $\approx$ 4.24 but 4.2353 $\approx$ 4.23. Also, 3.4555 $\approx$ 3.45.

  Note: In statistics, probabilities are sometimes expressed as **p = 0.0000**. This does not mean $p = 0$; it is a convention for indicating that $p < 0.00005$: i.e., there are at least four zeroes after the decimal point.

Using the summation sign (Σ)

'Σ' indicates adding numbers to produce a total; i.e. 'add the numbers in a column or in a row'.

Because we sometimes organise data in columns within a table (or a spreadsheet), this symbol is often used when we want to indicate that the numbers in some column should be added. Consider, for example, the table of figures below:

Case	**x**	**y**	**x^2**	**y^2**	**x – y**	**xy**
1	1	2	1	4	–1	2
2	2	3	4	9	–1	6
3	3	4	9	16	–1	12
Σ	6	9	14	29	–3	20

Note that:

- **x^2** and **y^2** are the squares of the numbers in columns **x** and **y** respectively (for each case).

- The column **x–y** is obtained by subtracting the number in column **y** from the value in column **x**.
- The column **xy** is obtained by multiplying the value in column **x** by the value in column **y**.
- $\Sigma\mathbf{x} = 1 + 2 + 3 = 6$.
- $\Sigma\mathbf{y} = 2 + 3 + 4 = 9$
- $\Sigma\mathbf{x^2} = 1+4+9=14$
- $\Sigma\mathbf{y^2} = 4 + 9 + 16 = 29$
- $\Sigma\mathbf{(x - y)} = -1 - 1 - 1 = -3$.
- $\Sigma\mathbf{xy} = \Sigma\ (\mathbf{xy}) = 2 + 6 + 12 = 20$
- $\Sigma\mathbf{xy}$ is the same as $\Sigma\mathbf{(xy)}$; in other words, first multiply each value in column **x** by the matching value in the **y** column, which then gives us column **xy**. Then add the numbers in column **xy** to get the total.
- Consider the expression $\Sigma\mathbf{x}\Sigma\mathbf{y}$. This means $(\Sigma\mathbf{x})$ x $(\Sigma\mathbf{y})$: First get $\Sigma\mathbf{x}$ and $\Sigma\mathbf{y}$ separately by adding all values for each case in each column, and then multiply them:

  $\Sigma\mathbf{x}\Sigma\ \mathbf{y} = (1 + 2 + 3) \times (2 + 3 + 4) = 6 \times 9 = 54$.

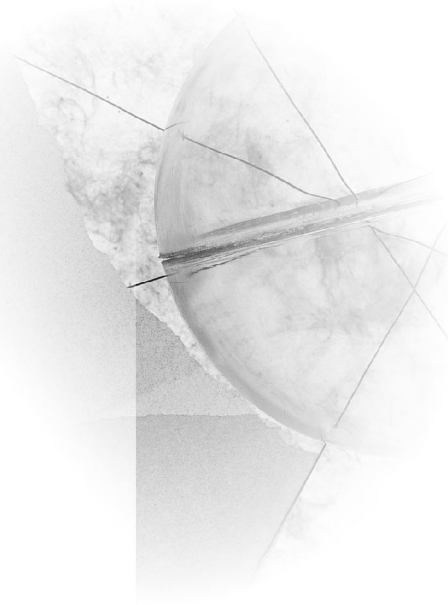
It is important to realise that $\Sigma\mathbf{xy}$ and $\Sigma\mathbf{x}\Sigma\mathbf{y}$ will *not* produce the same result!

Some more examples:

- $\Sigma\mathbf{(x–y)}\ \Sigma\mathbf{x^2} = (-1-1-1) \times (1 + 4 + 9) = -3 \times 14 = -42$
- $3\Sigma\mathbf{y^2} = 3\mathbf{(}\Sigma y^2) = 3 \times (4 + 9 + 16) = 3 \times 29 = 87$

APPENDIX F

Decision tree for test statistics



The diagram below shows the tests covered in this module, and when it is appropriate to use them. There are many other tests for other conditions. Keep in mind, however, there may be conditions attached to their use, such as distributions from which the samples come, or sample size, etc.