

DOUGLAS BORS

# DATA ANALYSIS FOR THE SOCIAL SCIENCES

Integrating Theory and Practice



Los Angeles | London | New Delhi  
Singapore | Washington DC | Melbourne



Los Angeles | London | New Delhi  
Singapore | Washington DC | Melbourne

SAGE Publications Ltd  
1 Oliver's Yard  
55 City Road  
London EC1Y 1SP

SAGE Publications Inc.  
2455 Teller Road  
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd  
B 1/I 1 Mohan Cooperative Industrial Area  
Mathura Road  
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd  
3 Church Street  
#10-04 Samsung Hub  
Singapore 049483

---

Editor: Jai Seaman  
Assistant editor: Alysha Owen  
Production editor: Ian Antcliff  
Copyeditor: Richard Leigh  
Proofreader:  
Indexer:  
Marketing manager:  
Cover design:  
Typeset by: C&M Digital (P) Ltd, Chennai, India  
Printed in the UK

© Douglas Bors 2018

First published 2018

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act, 1988, this publication may be reproduced, stored or transmitted in any form, or by any means, only with the prior permission in writing of the publishers, or in the case of reprographic reproduction, in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

**Library of Congress Control Number: 00000000**

**British Library Cataloguing in Publication data**

A catalogue record for this book is available from the British Library

ISBN 978-1-4462-9847-3  
ISBN 978-1-4462-9848-0 (pbk)

At SAGE we take sustainability seriously. Most of our products are printed in the UK using FSC papers and boards. When we print overseas we ensure sustainable papers are used as measured by the PREPS grading system. We undertake an annual audit to monitor our sustainability.

# 1

## Chapter contents

<b>1.1</b> Purpose	4
<b>1.2</b> The general framework	4
<b>1.3</b> Recognizing randomness	8
<b>1.4</b> Lies, damn lies, and statistics	9
<b>1.5</b> Testing for randomness	10
<b>1.6</b> Research design and key concepts	14
<b>1.7</b> Paradoxes	19
<b>1.8</b> Chapter summary	20
<b>1.9</b> Recommended readings	20

# OVERVIEW

**KEY CONCEPTS:** randomness, experiment, quasi-experiment, observational designs, question of difference, question of association, categorical data, measurement data, null hypothesis, Simpson's paradox, Type I error, Type II error, variables, population, sample, random sample, independent variable, dependent variable.

## 1 1 PURPOSE

The first purpose of this chapter is to introduce you to a few concepts and themes that will be present, directly or indirectly, throughout this book. If there is one concept that is omnipresent, if not explicitly then at least implicitly, it is *randomness*. As will be seen, the concept underlies other phrases used either to refer to the presence of randomness or to its absence. To claim that two groups of people differ in some respect is to say that group membership is not completely random; for example, height is not random with respect to basketball players versus non-basketball players. To say that two groups of people do not differ in some regard is to say that group membership is random; for example, the maximum speed at which a car can travel is probably unrelated to the car's colour. To assert that two events are related is to say that they do not occur randomly with respect to each other; for example, tsunamis are associated with earthquakes. To state that two events are unrelated is to say that they occur randomly with respect to each other. Related to our use of randomness are four key questions: What is *expected*? What is *observed*? What is the *difference between the expected and the observed*? How much of a difference can be *expected due to chance alone*?

The second purpose of this chapter is to review some basic strategies and principles of empirical research. We differentiate the basic forms of research (experimental, quasi-experimental, and observational designs) and review the main characteristics of each.

## 1 2 THE GENERAL FRAMEWORK

Statistics, which are the numbers researchers use to describe their data and to test the trustworthiness or *replicability* of their findings, can feel convoluted and mysterious both for students and for researchers. In this section I offer you a three-part framework; if you use it, it will make the material

The great comedian George Burns said at his 100th birthday party, 'once you reach 100 you have it made. Very few people die after the age of 100.'

in this book, and the statistics you encounter in everyday life, more easily understood.

*Part 1.* As a researcher, you begin with a question (or questions) about the nature of the world, or at least that aspect of the world which interests you. Let us start with the

simplest type of research, where there is only one question. Regardless of your topic, your question will take one of two basic forms.

The first form is one of *difference*. For example, imagine yourself as a political science professor who wishes to know if your students prefer term papers or essay examinations as a means of evaluation. Or think of yourself as a clinical psychologist wishing to know if cognitive behavioural therapy (CBT) reduces your patients' anxiety symptoms more than does the most commonly prescribed anxiolytic (a medication to reduce anxiety). In both of these examples you suspect that one set of scores will be different from the other: more students will prefer one form of assessment over the other; the CBT group on average will show fewer symptoms than the anxiolytic drug patients.

The second form a research question may take is one of *relation* or *association*. For example, you may manage a coffee shop and are interested in customer behaviour. Is the choice of beverage (coffee versus tea) associated with gender (men versus women)? Or if you are an educational psychologist you may suspect that there is an association between the number of hours per week a student works off-campus and his or her grades at the end of term. In both of these examples you suspect that one set of scores will be related (or will predict) the other set of scores. Perhaps a greater proportion of women will prefer tea than will men. Perhaps the more hours a student works off-campus the lower his or her grade point average will tend to be. The type of question – difference versus association – orients you towards appropriate statistical procedures. Questions of differences are linked with one family of statistical tests, and questions of association are linked with another family of tests.

Questions of differences and questions of association are not as dissimilar as they may appear. They are usually two sides of a single coin, with one question implying the other. Furthermore, a research project in psychology and in the social sciences often entails more than one question, and it may involve both questions of differences and questions of associations. For example, you may be a ‘sportologist’ wishing to know why some baseball players hit more home runs than others. You suspect that the taller the player the more home runs he will hit (this is a question of a possible association between height and the number of home runs). You may also suspect that players who use aluminium bats will hit more home runs than will players who use the old-fashioned wooden bats (this is a question of a possible difference between types of bats).

*Part 2.* As an empirical researcher you collect data. Regardless of your area of interest, the observations usually take one of two general forms.

The first form your observations can take is that of *frequency data* or *categorical data*. Remember, as a political science professor you wished to know if among your students term papers are more popular than essay examinations as a form of evaluation. You are keeping count of the number of students in the two categories: those who prefer a term paper versus those who prefer an essay examination. As a manager of a coffee shop you were keeping track of the frequencies in four categories: the number of women who prefer coffee, the number of women who prefer tea, the number of men who prefer coffee, and the number of men who prefer tea.

The second form your observations can take is that of *measurement data*. As a clinical psychologist you wished to know if two groups of patients (CBT versus anxiolytic) differ in terms of their average number of anxiety symptoms. You are recording the number of symptoms each patient exhibits. It is possible that no two patients will exhibit the same number of symptoms. As an educational psychologist interested in hours worked and academic performance, you are recording the actual number of hours per week each student works off-campus and his or her grade. It is possible that no two students in your study will have worked the same number of hours or have exactly the same grade.

I need to warn you: the two types of data are not as different as they may at first appear, nor do they encompass all possible types of data. And often one type of data can be transformed or treated as if it were the other type. Examples of this transformation will appear at the end of Chapter 3.

---

As we will see in Chapter 2, frequency/categorical data and measurement data can be further divided into four types of number scales: nominal, ordinal, interval, and ratio. Where nominal and ordinal number scales are described as being frequency/categorical data, interval and ratio scales are considered as

measurement data. As will become apparent in Part II of this book, for purposes of analysis ordinal data (such as percentile scores on an examination) often form an intermediate form of data or are transformed into a type of measurement data called z-scores, which are discussed in detail in Chapter 3.

---

We now have two basic research questions and two types of data. Earlier we said that each research question is linked with its own family of statistical tests. The same may be said with respect to the two types of data. Frequency data are linked with one family of statistical tests and measurement data are associated with another family of tests.

There are four families of statistical test:

Tests for a question of difference with frequency data

Tests for a question of relation with frequency data

Tests for a question of difference with measurement data

Tests for a question of relation with measurement data.

Keep in mind that this framework is not carved in stone, nor are the boundaries between the four categories impermeable. Rather, the framework is a guideline for following the flow of this book. It will help you to cut through what appear to be so many unrelated procedures and formulae and to see the general storyline and character types.

*Part 3.* We have seen that there are different families of statistical tests which reflect an intersection of the type of question the researcher asks and the type of data he or she has collected. Surprisingly, almost all statistical tests – at least those covered in this book – have the same underlying logic based on a few simple questions.

Question 1: What do you as a researcher *expect* to find?

You may have taken a course that introduces you to research methodology and know that what the researcher expects to find is often defined in two ways. One, as a researcher you have an

educated guess as to what you will find. For example, from past experience you expect that students prefer term papers over essay examinations. The *expectation* with which we usually begin a data analysis is a negation of what we actually expect. That is, we expect that our observations are only random; for example, students have no clear preference.

For the past hundred years there have been serious debates about the value of the null hypothesis test as developed by Neyman and Pearson (1933) and Fisher (1935). Several alternatives to testing the null hypotheses have been proposed, and all have their supporters as well as their critics. There will be more details concerning this debate in Chapter 4.

This is the famous *null hypothesis* which is considered in detail in Chapter 4. For now, the null hypothesis assumes that whatever your research idea (alternative hypothesis) may be, it is wrong, meaning that the data do not confirm it. This depiction of the null hypothesis is another over-generalization (and there are also important alternatives to null hypothesis testing), but this model helps us to get started.

---

The primary reason for the null hypothesis, and for assuming randomness, is that it allows us to know exactly what to expect. To claim that where a student sits during an examination (front versus back of the room) affects his or her mark is rather vague. How will it affect his or her mark and by how much? There could be an infinite number of predictions to

be made. It is easier to say that nothing is going on other than randomness. With this position there is only one prediction: the averages of the two groups (front and back of room) will be identical. If there is sufficient evidence to reject this position, then there is indirect evidence that seating location is related to examination marks.

---

Question 2: What do you *observe*?

You will always observe some difference between what you expect and what you observe. It is almost impossible to find exactly no difference or exactly no association. Remember, the expectation (null hypothesis) is randomness. For example, the frequency with which women prefer coffee will be no different than the frequency with which men prefer coffee, and the examination marks of those seated at the front of the room will not differ from those who are seated at the back of the room. But the frequencies will not be exactly the same, nor will the examination marks. Things are almost never as expected.

Question 3: Is this difference *trustworthy*?

The question statistical tests are designed to answer relates to the issue of the likelihood that the difference between the expected and the observed is real. 'Real' is a slippery concept. Here it means that if we repeated the observations we would observe a similar difference between the expected and the observed. If tomorrow we again recorded the choice of beverage of women and men, would the difference in relative frequencies be roughly the same? If next semester we compared the examination marks of those who are seated at the front and those seated at the back of the room, would the difference in the averages be about the same? Or were the differences all a tempest in a teapot (pun intended)? We call the reliability of findings the *replicability* of the findings.

Question 4: How much of a difference can we expect *due to chance alone*?

How great a difference between 'what is expected' and 'what is observed' is required for you to conclude that the difference between the two is trustworthy and not a random accident? We typically wish to be 95% confident of our conclusion. Determining the necessary size of the difference between the expected and the observed is the heart of the issue. This is the tricky part of statistics. If the required



Keep this framework in mind; it will help you learn the material. Good luck in your statistics course, and I hope you enjoy the remainder of this book.

size of a difference between the expected and the observed were a fixed amount, the problem would be simple, but it is not. The necessary difference, however, is contingent upon how

much of a difference is expected due to chance alone. And what is expected due to chance alone is contingent upon several factors, depending upon the nature of the research and the type of data.

To summarize, you have a research question. You collect data. You have an expectation, negative in nature. You have an observation. You find a difference between the two. And you need to determine if that difference is reliable.

### 1 3 RECOGNIZING RANDOMNESS

Can you tell which of the following ten strings of ten digits were produced randomly by blindly picking numbers out of a hat, that is, picked without any aim, purpose, plan, criteria, or design? (There are large and equal numbers of the digits 0, 1, 2, ..., 9 in the hat.)

- String 1: 9,7,4,3,7,2,2,8,0,2
- String 2: 3,3,6,8,9,9,8,0,2,7
- String 3: 6,6,6,0,3,9,8,9,9,5
- String 4: 1,3,5,4,5,0,2,9,1,1
- String 5: 1,1,9,6,7,4,1,2,8,4
- String 6: 3,1,5,4,6,8,1,6,0,7
- String 7: 0,9,9,3,0,5,0,4,6,0
- String 8: 7,3,6,9,8,2,7,3,5,2
- String 9: 0,7,7,3,0,6,4,1,2,8
- String 10: 6,8,6,9,5,8,9,6,1,0

See <https://study.sagepub.com/bors> for the answer and to judge your ability to identify when things are not random.

 Web Link 1.1 for the answer and to judge your ability to identify when things are not random.

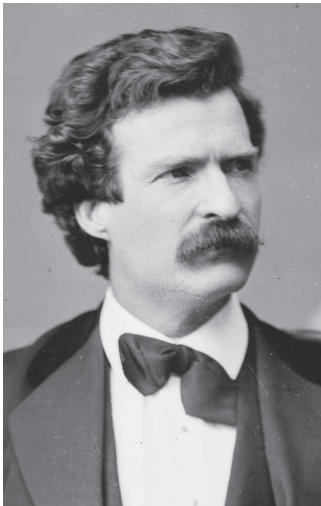
Are you able to differentiate those occasions when your favourite sports team is on a *real* winning streak and shows promise from those occasions when the team is having a lucky string of wins and will soon sink back into mediocrity? Clearly it is difficult to describe ahead of time what to expect if the events are not random. There are so many different criteria one might use. For example, too many of the same number or too many of the same number in a row. On the other hand, we will see that expectations are easier to describe when the events or observations are assumed to be random.

## 1 4 LIES, DAMNED LIES, AND STATISTICS

In today's world numerical information and statistics are everywhere. It often appears that we live in a world where everything is reduced to numbers, even when it seems absurd to do so: 'Our ice cream is twice as tasty as is our leading competitor's.' Truly there can be misuses and abuses of numbers and statistics. As Mark Twain (Figure 1.1) lamented, there are 'lies, damned lies, and statistics'. One way in which abuse occurs on a daily basis is through the technique of 'cherry picking'. Cherry picking is a process where politicians, advertisers, and even researchers report only some of the evidence and ignore or suppress the rest. This biased reporting and analysis of data can occur both intentionally or unintentionally. For the purposes of this book, we must ignore the scientific crime of intentional cherry picking.

If you are a baseball fan, imagine you are told that a particular batter (batter A) has a higher *overall* batting average than batter B, but batter B has a higher batting than batter A against both left-handed pitchers and right-handed pitchers. (Overall batting average is derived by combining data from the batters' performances against both right-handed and left-handed pitchers.) Is this a lie? Is it a damned lie? Or can it possibly be true?

The statistical description and testing of empirical data, however, can play a legitimate and central role in answering most research questions in the psychological, social, and biological sciences. This does not mean that the findings and conclusions from all research studies are 'true' and can be accepted uncritically. A sound appreciation of the fundamentals of the various statistical procedures used by researchers is crucial for understanding the appropriateness, the value, and the limitations of research findings and conclusions, as well as for avoiding the statistical lies and damned lies we encounter on a daily basis. This applies just as much, if not more, to those of us who are only consumers of statistics as it does to those who are engaged in research. As you will see in this book, there are a few basic notions that underpin nearly everything we use to describe and analyse our data.



**Figure 1.1** Photo of Mark Twain. His name means 'Take note! the water is two fathoms (12 feet or 3.66 meters) deep, on average']

Implicit in all difficult-to-comprehend stories is a simple – though not necessarily simplistic – plotline. Statistics is such a story. If we could ask Shakespeare about writing a play, he would tell us that there are only a few good plots, but that they can be told with many twists and turns and in many social, familial, and historical settings. Each variation requires the dramatist to adapt the plot to the particulars, even though the underlying theme and moral remain the same. English literature students often find it difficult to understand Shakespeare's plays. Without discerning the underlying plot, the dialogue and details make little sense. Furthermore, the difficulty often is exacerbated by the language: although Shakespeare's plays are written in English, it is not the vernacular of today.

Many students experience introductory statistics textbooks in an analogous manner. They often are so perplexed by the language

and details that they lose the plot: the purpose of a particular analysis. Browse through any textbook and you will find strange symbols, elaborate formulae, and common words with unexpected usages (e.g., *power*). Most students approach their statistics course as they approach a Shakespearean play, with trepidation, and they find much of the material initially incomprehensible. But when we carefully examine each play, act, and scene of statistics we find beneath the plethora of terms, tests, and formulae a small set of common themes and characters, as described earlier. Understanding these common themes is vital for understanding the particulars of any given statistical test. This first chapter presents the core plotline of the epic story of statistics as it is commonly employed in psychology, sociology, education, political science, and areas of biology. Let us begin; to quote *Henry V*, 'The game's afoot.'

## 1 5 TESTING FOR RANDOMNESS

Here we have ten scores on a quiz in macroeconomics: 1, 7, 4, 5, 3, 4, 5, 2, 6, and 3. Inexplicably – as with most events in economics – we find the scores sorted into two groups. Group A is comprised of 1, 4, 2, 5, and 3. Group B is comprised of 7, 5, 3, 4, and 6. Do you think that those scores have been sorted 'randomly' into those two groups? This can be viewed as the rudimentary version of the 'to be or not to be' question of statistics.

The notion of randomness is one element of the plot that underlies much of this book. The idea of randomness – to reference Shakespeare one more time, 'like the sun; it shines everywhere' – is present explicitly or implicitly in all: collecting the data, analysing the data, and drawing conclusions. We all have a sense of what we mean when we say that something is random, but it might be difficult to clearly articulate that sense. In fact, humans have difficulty creating or recognizing randomness. According to the *Oxford English Dictionary* online, 'random' refers to something that was made or happened without method or conscious decision. For example, we speak of acts of random violence, horrendous acts that cannot be explained. But we usually discover that these acts were not random, but planned, albeit that the plan often appears crazy.

As mentioned above, sometimes we assume that we know the meaning of a word that is used by an author, when in fact the author's meaning is considerably different. For example, in a Shakespearean play the word 'abuse' usually means to deceive, as when King Lear says to his daughter Cordelia in Act IV 'do not abuse me'. In the domain of statistics, one important meaning of the word *random* is that all members of a set have the same chance or probability of appearing or of being selected as a member of a subset. The key words here are *set*, *chance*, *probability*, *appearing*, and *selected*. Let us begin by looking at some examples.

We may wish to randomly *select* ten students (subset) from a class of 100 (set). What does 'random' mean in terms of selection? In brief, it means that all 100 students in the class have the same *chance* or *probability* of being *selected* for our subset of ten. What would an equal chance look like? It means that no member or subset of members of the set is more likely than any other to be selected. In our example, tall students in the class of 100 are no more likely to be selected than short students; those who sit at the front of the class are no more likely to be selected than those who sit at the back; those who are eager to be selected are no more likely to be chosen than are

those who wish to avoid selection, etc. It is as if the selection daemon or mechanism is blind to all characteristics of the members of the set.

Let us jump ahead and anticipate material to be covered in greater detail in Chapter 3 and state that all members of the class of 100 have a one-in-a-hundred chance of being selected:  $1/100$ . Said another way, they all have a 1% chance of being selected; or as we will express it later, they have a *probability* of 0.01. If there were only 50 students in the class, then each student would have a  $1/50$  chance of being selected, or a probability of 0.02.

### Random with respect to ...

As mentioned earlier, randomness in the domain of statistics also means that observing one outcome or event does not change the likelihood of observing another outcome or event. That is, there is no systematic relationship between two possible observations. For example, with respect to the heights and the weights of our 100 students, we might ask if the heights of the students are random with respect to their weights. That is, do all of the weights have an equal chance of *appearing* or being associated with all heights? This would mean that regardless of a student's height, all weights would have the same chance of being associated with that student. In looking at the exam grades of the 100 students, we might ask if the students' grades are random with respect to where they sat during the final exam: front versus back of the room. As we will see, we can render this into the question of 'did one section of the room have a higher average grade than the other?' or 'are grades related to section of the room?' – a question of *difference* versus a question of *association*. We may have good reason to think or *hypothesize* that weight is not random with respect to height, and we may have good reason to think that grades are not random with respect to seating location. It is useful to know what to expect if randomness existed in both cases.

### Expected value versus observed value

Although we have focused on randomness, underlying all of these examples is an issue of *expectation*. For purposes of statistical testing there are two types of expectations. What would we expect to see if things were random? What would we expect to see if they were not? Recall that at the outset of the chapter we said that it was not easy to describe what a string of digits would look like if they were *not* random. We also asked if ten quiz scores were randomly sorted into two groups. As we will discover, it is far easier to state what is expected if things are random than it is to state what we would expect if they were not random.

In most forms of data analysis the expectation, or *expected value*, commonly takes the form of a magnitude or a frequency. The expected value is normally, but not always, what reasoning tells us we should observe, if things are random. If we flip a coin a number of times, then half of the time it should come up heads. Or if seating section is unrelated to exam score, then the average exam scores of the two sections should be the same. The *observed value* is simply the magnitude or the frequency that we actually find. The most common magnitude is a type of average. A frequency is

the number of times something is observed. This could be the number of heads observed after ten tosses of a coin. What is usually of interest to researchers is the discrepancy between the *expected value* and the *observed value* and the size of any discrepancy.

Rarely will the *observed value* correspond exactly to the *expected value*. If we knew that there were an equal number of men and women enrolled in our university and if introductory statistics were a required course, then introductory statistics students might be a random subset of the university's student population. Therefore, we might expect to find an numbers of men and women in our statistics class: 50% men and 50% women (*expected value*). In the terminology used above, we would expect our selection or subset of students (the class) to reflect the university student body (set). If we found that only 49% of the students (*observed value*) in the class were men, should we question our knowledge that there were equal numbers of men and women enrolled in our university? What if we observed only 40% men in the class, or 25% men, or even 1% men? There are good reasons why we might not observe exactly a 50%–50% split in every statistics class. Due to chance alone, we would expect to see some variation in the percentages of men and women. While in some semesters we might see more men than women, in other semesters there would be more women than men. While in some semesters we might see relatively small differences in the percentages of men and women, in other semesters we might see somewhat larger differences. In everyday life we usually refer to all of these purely chance disparities as accidental.

Although we *expect* to find a 50%–50% split in the number of men and women in the class, we also recognize that it is highly unlikely that we will *observe* an equal number of men and women. Thus, a 49%–51% split is probably not too large a discrepancy for us to continue assuming that there are an equal number of men and women at the university, whereas a 1%–99% split might cause us to question our assumption. What about intermediate discrepancies such as 40%–60%? We recognize that there will be class-to-class fluctuations in the percentages, but how much fluctuation is reasonable? Where do we draw the line when it comes to maintaining or rejecting our assumption of equal percentages of men and women? As we will find, there is no single, absolute threshold upon which we can make our decision. Furthermore, any conclusion that we draw may be wrong, no matter at what level we set the threshold. Consequently, it is being able to describe reasonable fluctuation and how to determine the threshold that is key.

For purposes of making decisions, we usually begin by retaining our expectation – or the assumption it is based upon – unless the chance of obtaining the *observed value* is less than 5%. Thus, if we could determine the threshold where the difference between the *expected value* (e.g., 50% men and 50% women) and the *observed value* is 5% then we could draw a tentative conclusion about whether or not to retain our assumption.

## Randomization testing

In this section we introduce the notion of *randomization testing*. This most basic form of statistical testing will be discussed in more detail beginning in Chapter 5. To describe the randomization test we return to our earlier example of scores on a quiz in macroeconomics: 1, 7, 4, 5, 3, 4, 5, 2, 6, and 3. Remember, we inexplicably found the scores sorted into two groups. Let us assume that the two groups represent left-handed students and right-handed students. The right-handed group's scores

were 1, 4, 2, 5, and 3. The left-handed group's scores were 7, 5, 3, 4, and 6. Do we think that those scores have been sorted 'randomly' into those two groups? That is, do we think that, with respect to quiz scores, there is no difference between left-handed students and right-handed students?

How might we determine where the threshold is for our 95% confidence? If we assume (expect) that the scores are randomly sorted into the two groups and that handedness is not a factor, then the scores as currently grouped are just one of a large number of other possible random sorts. We might then compare the present grouping with other random sorts. The question is, how? If the current grouping is random, then we *expect* the averages of the scores of the two groups to be equivalent. This is our null hypothesis. But the observed means are not equivalent. The average of the right-handed students is 3 and the average of the left-handed students is 5. The difference between the two groups is 2, not 0. The question now is, does this difference have a less than a 5% chance of being due to chance alone?

How can we determine what are the chances of randomly sorting the quiz scores and observing a difference of 2? The procedure is simple, although tedious. We write those ten quiz scores on slips of paper and put them in a hat. Then we pull them out and place the first score drawn in group A, the second score drawn in group B, the third in group A, the fourth in group B, and so on, until we have a new sort of the ten scores. We then record the difference in the averages of the new sort. Then we put back the scores that had been drawn into the hat and start over again. We obtain another sort, another two averages, and we record the difference. We do this a very large number of times. Ideally, we repeat the process for every possible sort.

---

After one random sort, group A could contain the scores of 1, 2, 3, 3, and 4; group B would then contain the scores 4, 5, 5, 6, and 7. The difference in the averages would be  $2.6 - 5.4 = -2.8$ . After another random sort, group A could contain the scores 1, 2, 4, 6, and 7; group B would contain the scores 3, 3, 4, 5, and 5. The difference in the averages would be  $4 - 4 = 0.0$ . There are far too many possible sorts for us to enumerate them here.

---

Once we complete recording all those differences in the averages we have the basis for determining the differences we might expect to see due to chance alone. First, we order the differences from the smallest to the largest. Next, we count the number of differences in the ordering which are greater than 2. If we divide this number by the total number of sortings, we have the proportion of differences that are greater than 2. If we went through this process we would find that about 4% of the differences are larger than 2.

Before we reach any conclusion about handedness and quiz scores, there is one more twist in the process. Although in the case of the original quiz scores the left-handed students did better than the right-handed student, it could actually be the other way around. The right-handed students may have accidentally done better than the left-handed students by an average score of 2. Thus, the chances of a difference of 2, regardless of the direction, are about 8%. Earlier we said that the chances would need to be less than 5% before we would no longer assume randomness. Stated more formally, the chances need to be less than 5% before we reject the null hypothesis.

Because the chances of our observed difference between what is expected and what is observed are greater than 5%, we conclude that there is insufficient evidence to support the idea that handedness influences quiz scores. Of course, we could be wrong. When we incorrectly fail to reject the assumption of the null hypothesis, we call this a *Type II* error. When the chances of the difference between the expected and the observed are less than 5% we reject the null hypothesis. We may be correct. On the other hand, we may be wrong. Remember, all the outcomes that were ordered from smallest to largest were randomly created. Thus, 5% of the time we will incorrectly think that something is affecting the sort, in our example the handedness of the students. This mistake is called a *Type I* error.

The problem with the randomization testing of the null hypothesis is more than tediousness. Even with the use of computers, it becomes unviable when the research design is more complex and the number of scores increases. Try producing all the possible sorts for the ten scores in our example. In large part, this book is about the short-cuts taken to make the analyses possible. These short-cuts require making certain assumptions. The testing of the assumptions and the 'correcting' for any violations become an important part of statistical analyses. We will cross those bridges when we come to them.

Finally, this was only an example using the question of *difference* and *measurement* data. There are other variations of the randomization test for other combinations of research questions and data types. The example of handedness and quiz scores illustrates the basic framework, however.

## 1 6

## RESEARCH DESIGN AND KEY CONCEPTS

In this part of this chapter we review the basic empirical research designs and key concepts. Here we cover only those designs and details which are needed to provide the necessary background in research methodology for understanding the statistical procedures presented in this book. For more detailed coverage of empirical research methodologies see Nestor and Schutt (2015) and Privitera (2016).

There is an important distinction to be made between generalizing across people and generalizing across situations (Aronson et al., 2007). In the first case you are making a prediction about how people in the same population (your university students) will respond. In the second case you are making a prediction about how others elsewhere in the same circumstance may respond. This second type of generalization is made clear in our discussion of the experiment.

All researchers collect data from sources or subjects. Because most of our research in psychology, the social and biological sciences, is directed at human or animal activity, we shall refer to *subjects*.

Central to all research designs are the key elements of variables, population, sample selection, and sample assignment.

All empirical research asks about either a possible difference or a possible association between two or more variables. *Variables* are anything that can take more than one value or form. They can be *differences in amount*, such as quiz scores or the number of anxiety symptoms reported by patients. These are called quantitative or *measurement* variables. Variables can reflect *differences in form*, such as choice of beverage or the type of treatment used to treat severe anxiety. These are called qualitative or *categorical* variables.

The meaning of the term *population* in statistics is different from the everyday or geographical usage of the term. In statistics 'population' refers to the entire set of events to which the researcher wishes to generalize his or her findings. Imagine that you are a clinical psychologist working in a university counselling centre and you wish to know whether cognitive behavioural therapy or the new anxiolytic works better for the students who visit the counselling centre. What is the population to which you wish to generalize? It is not every young adult on the planet. It is not even all university students, or even all students at your university. The population of interest to you is all the students who walk into your counselling centre and report symptoms of anxiety. For a detailed discussion of the issue of generalization see Cook and Campbell's (1979) groundbreaking work on the topic.

Researchers do not test or observe all members of the population to which they wish to generalize. Furthermore, it is usually impossible to observe all members of a population. Instead, researchers observe a subset of the population. This subset drawn from the population is called a *sample*. Samples are used to make inferences about populations. In order to make such inferences the sample must be representative of the population. The best way to ensure that is to randomly sample from the population. A *random sample* means that all members of the population have an equal chance of being drawn for the sample. Ideally, the sample is a mini version of the population. It is almost impossible to adhere to this ideal. Most random samples are really *quasi-random* samples. A quasi-random sample usually uses some systematic device to select subjects, such as every third person who walks through the counselling centre door who reports anxiety symptoms. There are other sampling methods that are beyond the scope of this chapter, such as the matching strategy (Rubin, 1973). Much of the academic research carried out at universities, particularly in psychology, uses the *convenience sampling* technique. This method samples from the population members who are conveniently available to participate in the research (e.g., university students). To the extent that undergraduate university students represent the young adult population, there is little bias in the sample. This is likely the case when the researcher is studying colour perception. If the researcher is investigating political attitudes, university students may not be an unbiased sample of young adults. For most purposes in this book we will assume that the selected samples are either random or at least representative of the population to which the researcher wishes to generalize.

In some research designs, once a sample has been selected the members must be assigned to different treatment conditions. It is important that members of the sample are randomly assigned to conditions. Random assignment requires that all members of the sample have an equal chance of being assigned to a given condition. Just as the overall sample should be a mini version of the population, so each condition should initially be a mini version of the overall sample. As in the case of sample selection, it is almost impossible to adhere to this ideal of equal chances with assignment. Most condition assignments are in practice *quasi-random* in nature. A quasi-random assignment uses a simple technique for the assignment of subjects. For example, the odd-numbered subjects in your anxiety treatment study will be assigned to the cognitive behavioural therapy group, and the even-numbered subjects will be in the anxiolytic group.

Figure 1.2 summarizes three basic research designs, the names of the variables associated with them, and the type of possible conclusions associated with each design.



Design type	Variable types	Possible conclusions
True experiment	Independent Dependent Nuisance Confounding	Causal/influence
Quasi-experiment	Pseudo-independent Dependent Nuisance Confounding	Marker/association
Observational	Predictor Criterion Nuisance Third-variable	Association

Figure 1.2

independent variable or are due to chance alone (Privitera, 2016). Stated differently, it offers the researcher an opportunity to determine if one variable is the ‘cause’ of another.

## Experiment

Although experiments can be run inside or outside of a laboratory, we shall focus on experiments conducted inside a laboratory. Many argue that the laboratory experiment is the ideal form of research. Others argue that it has serious limitations. The reason why it is deemed by many to be ideal is that the experiment is designed to allow the researcher to evaluate whether the changes in the dependent variable are related to the manipulation of the

Random selection of subjects is the best way of ensuring the *external validity* of the research. That is, random selection allows the researcher to generalize to the population of interest. The random assignment of subjects to conditions is a way to ensure the

*internal validity* of the experiment. An internally valid experiment is one where only the independent variable could be responsible for changes in the dependent variable, all else being a constant across the treatment conditions.

There are two forms of an experiment: the *between-subjects design* and the *within-subjects design*. In the between-subjects design subjects in the selected sample are (quasi-)randomly assigned to different treatment conditions. For example, the odd-numbered subjects in your anxiety treatment study are assigned to the cognitive behavioural therapy (CBT) group, and the even-numbered subjects assigned to the anxiolytic group. The two types of variables of importance in an experiment are the *independent variable* (IV) and the *dependent variable* (DV). The IV is called independent because it is theoretically independent or random with respect to all other variables in the population of interest. The type of treatment your patients receive is your IV: CBT or the anxiolytic. Because the IV levels are a difference in kind or type of treatment, your IV is qualitative or categorical. The DV is the variable whose values are dependent (not random) on the values or levels of the IV. Your DV is the anxiety level of your patients after treatment. The logic is as follows:

- 1 if the two groups of patients are equivalent prior to treatment,
- 2 and if the only difference in the two groups is the difference in treatment,
- 3 then any observed difference in the anxiety levels in the two groups post treatment can be attributed to the differences in the treatment.

For example, if the CBT treatment group manifests fewer symptoms than the anxiolytic treatment group, you might conclude that the 'cause' of (reason for) the difference is that CBT is a superior treatment.

At this point we need to introduce the key notion of *operational definition* (Bridgman, 1927). Thus far we have used general descriptions for our variables. Someone might ask, 'What form of CBT and how many sessions?' They are asking you to operationally define your treatment. An operational definition describes a variable either in quantitative terms, or how it was measured, or in terms of its composition. For example, you may operationally define CBT at 1 hour per week for 12 weeks of standard CBT (Figure 1.3).

You may operationally define the anxiolytic treatment as 1 mg of Oxazepam per day for 12 weeks (Figure 1.4).

Finally, you may operationally define your DV (anxiety level) as the number of symptoms subjects check off on a standard symptom checklist.

Of course there are innumerable variables that will influence the number of symptoms a patient checks off on the list: how long the patient has been anxious, how severely he or she experiences some of those symptoms, how many hours he or she slept last night, just to name a few. We assume, however, that because the subjects were randomly assigned to the two treatment conditions, the two groups are roughly equivalent on these other variables. These additional variables which can influence an individual's score are often referred to as *nuisance* or *extraneous variables*. A *confounding variable* is a special type of nuisance variable. It is an extraneous variable that is associated with both the IV and the DV. As a consequence, it is impossible to determine if the changes in the DV are due to changes in the IV or the changes in the confounding third variable. Interpretation of the findings is rendered hopeless. Such variables are undesirable. Despite the researcher's best intentions of randomly assigning subjects to conditions, confounding variables will occur from time to time. For example, due to chance alone your CBT condition may be comprised primarily of men, and your anxiolytic group may be comprised primarily of women. Thus, any difference between the two treatment groups in terms of the DV (reported number of symptoms) will also be found between men and women. Interpreting your results will be a hopeless task. For example, is CBT a more effective treatment than the anxiolytic, or is it that men (who were the majority in the CBT group) are less anxious in general?

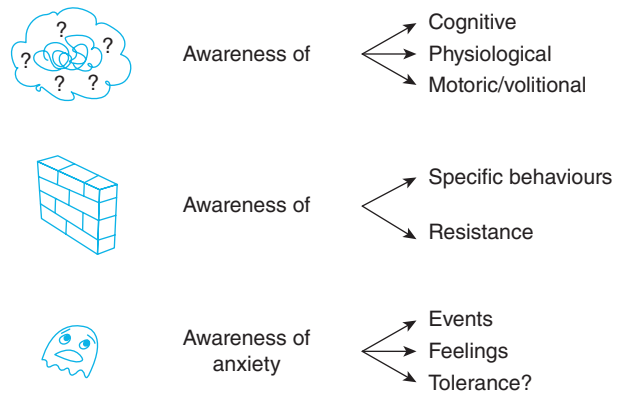


Figure 1.3

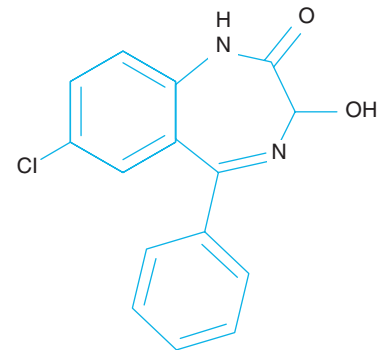


Figure 1.4

There are two important reasons for using operational definitions. First, they allow the reader to know what exactly the researcher studied. Second, they allow other researchers to try to replicate the original findings, which is a crucial part of establishing and confirming knowledge.

In the within-subjects design all subjects are tested in all conditions. Having the same subject in all conditions provides a statistical problem, but it is easily solved. Other than that, the logic of the within-subjects design is no different than that of the between-subjects design.

---

Experiments are usually considered a research strategy designed to answer questions of differences. For example, your anxiety treatment experiment can be viewed as asking a question about the difference

in the effectiveness of the two treatments, CBT and anxiolytic. The question guiding your experiment can be turned into one of association: is outcome (reported number of symptoms) related to the type of treatment?

---

While many researchers maintain that the experiment is the ideal for conducting empirical research, there are others who have serious reservations. The reservations centre on a specific aspect of what was described above as external validity: *ecological validity*. Critics of the laboratory experiment, particularly when it is the only strategy some researchers are willing to use, argue that what is found in the laboratory is not necessary what will be found in the natural or social world (Schmuckler, 2001). An obvious example is often described. Drugs are usually tested on young male rats that have one disorder, but usually they are prescribed for elderly female humans who have several disorders and are receiving multiple treatments. Can we be sure that the effect of the drug (and side effects) observed in the lab will also be the effect when they are prescribed to humans?

### Quasi-experiment

The *quasi-experiment* is often confused with the true experiment described above. A quasi-experiment can have the appearance of an experiment and can be run under controlled laboratory conditions. The key difference is that in a quasi-experiment the IV is not an IV, far from it. Examples of quasi-experiments include research exploring the difference between men and women on a spatial abilities test; native English speakers and native French speakers on a short-term memory test; and tall people and short people on a personality test. The key is that no one can walk into your lab and be assigned to a gender, or to a first language, or to a height, or any other characteristic of someone. These are called *subject variables*.

Furthermore, any difference that you find with respect to some DV comes with an infinite number of confounding variables. The IV is best called here a *marker variable*. It indicates that there are differences between the groups, but it does not imply causality. The simple way to know if you are faced with a true experiment or a quasi-experiment is to ask the following question: Could the subjects be randomly assigned to the conditions? If the answer is no, then the research is quasi-experimental in nature.

### Observational or descriptive studies

Unlike in an experiment, subjects in a *descriptive or observational study* are not assigned, randomly or otherwise, to a limited set of conditions. These studies use samples to describe populations,

particularly the associations among two or more variables. In its simplest form, one variable (called a *predictor*) is used to try to predict changes in another variable (called a *criterion* or outcome variable). For example, a sociologist may be interested in knowing if the average income level (predictor) of a city is predictive of its crime rate (criterion). Of course, cities cannot randomly be assigned to average income levels. The goal of such research is to describe the nature and the strength of the possible associations.

At a more complex level, a number of predictor variables are used to improve the researcher's ability to predict events or an *outcome*. Imagine yourself as the dean of a medical school, and you wish to know how you can predict which applicants will do well in your programme. In addition to their average marks in medical school (criterion) you have a number of pieces of information (predictors) on students accepted into the programme in previous years: average university marks, marks on a standardized entrance examination, number of volunteer placements, among others. To some degree, all of the predictors will be associated with one another. Despite the problem of confounding, you wish to know which of the variables are the important predictors of performance in your medical school.

There are other forms of empirical research as well as variations and combinations of the three we have briefly described. Elaborations on these designs will be described as we move through the book and the relevant forms of data analysis are introduced.

## 1.7 PARADOXES

As we said, it is always possible to conclude that things are random when they are not, and it is always possible to conclude that things are not random when they are. In addition to these problems there are strange, almost magical occurrences that can happen when we begin analysing data. In this section

we will focus on only one of those occurrences, sometimes called the amalgamation paradox or Simpson's paradox (Simpson, 1951). The most general form this paradox takes is where a difference or association appears in different groups but disappears when the groups are combined. A real-life example found in Julious and Mullee (1994) will help make the point.

Julious and Mullee compared the rates of success with two common treatments for kidney stones. Overall, treatment A was effective in 273 of 350 (78%) of the cases. Overall, treatment B was effective in 289 of 350 (83%) of the cases. Thus, treatment B appears to be the slightly more effective of the two treatments. The magic happens when we take into account that there are large and small kidney stones, and physicians tend to prescribe treatment B to those patients with the small stones and treatment A to those patients with large stones. Figure 1.5 illustrates the relative effectiveness of the two treatments with respect to the two stone types separately.

The bottom row of the table repeats what we already know, that overall treatment B appears to be more effective. But when we take into account small versus large stones and different rates of treatment, we find that treatment A is more effective than treatment B for the treatment of small stones (93% versus 87%), and treatment A is also more effective than treatment B for the treatment of


	Treatment A	Treatment B
Small stones	(81/87) 93%	(234/270) 87%
Large stones	(192/263) 73%	(55/80) 69%
Overall	(273/350) 78%	(289/350) 83%

Figure 1.5

large stones (73% versus 69%). From Chapter 4 onwards we will be discussing this phenomenon as the third-variable problem.

### CHALLENGE QUESTION

The manager of the local baseball team had completed choosing eight of his starting nine players. He decided to choose the last player on the basis of the highest batting average. Player A was delighted because he had the highest batting average of the remaining players. The manager told him, however, that the opposing team had a right-handed pitcher and that although he (player A) had the highest batting average overall, player B had a higher batting average against right-handed pitchers. Player A was devastated but he understood. It rained and they never played the game. The next day the manager again decided to choose the ninth player on the basis of the highest batting average. Player A was delighted because he had the highest batting average of the remaining players. The manager told him, however, that the opposing team now had a left-handed pitcher and that although he (player A) had the highest batting average overall, player B had a higher batting average against left-handed pitchers. Player A was devastated and could not understand how he had the highest overall batting average but player B had a higher average against both right-handed and left-handed pitchers. (There are only left-handed and right-handed pitchers.) How is this possible? Can you make up a set of batting averages illustrating your answer?

 Web Link 1.2 for an answer to the challenge question.

## 1 8 CHAPTER SUMMARY

In this chapter we have presented a general framework to organize the remainder of the book. Statistical tests can be described as falling into one of four categories defined by type of research question and type of data.

- 1 Tests for a question of difference with frequency data
- 2 Tests for a question of relation with frequency data
- 3 Tests for a question of difference with measurement data
- 4 Tests for a question of relation with measurement data.

We also summarized basic research terminology (e.g., types of variables) and the three basic designs to be analysed in this book: experiments, quasi-experiments, and observational studies. Finally, we looked at an example of Simpson's paradox, an issue with which we will need to be concerned before rushing to any conclusion, regardless of the form of analysis.

## 1 9 RECOMMENDED READINGS

Mlodinow, L. (2008). *The drunkard's walk: How randomness rules our lives*. New York: Random House. This very popular book discusses how randomness plays more of a role in the everyday events of our lives than we are usually willing to admit or even suspect. The author also points out our propensity to over-interpret and attribute causes to these random events.

Nestor, P. G. and Schutt, R. K. (2015). *Research methods in psychology: Investigating human behavior*. Los Angeles: Sage.

This book provides a comprehensive overview of the methods used by psychologists to collect and analyse data concerning human behaviour. Of note is Chapter 5 on survey research, a topic we will not be able to cover in any detail.

Privitera, G. J. (2016). *Research methods for the behavioral sciences* (2nd ed.). Los Angeles: Sage.

This book offers a very comprehensive treatment of the research process from beginning to end. Of particular relevance are Chapters 4–6. The entire book should be required reading for any student in psychology or the social sciences. It is worthwhile to keep on your shelf as a handy reference book.