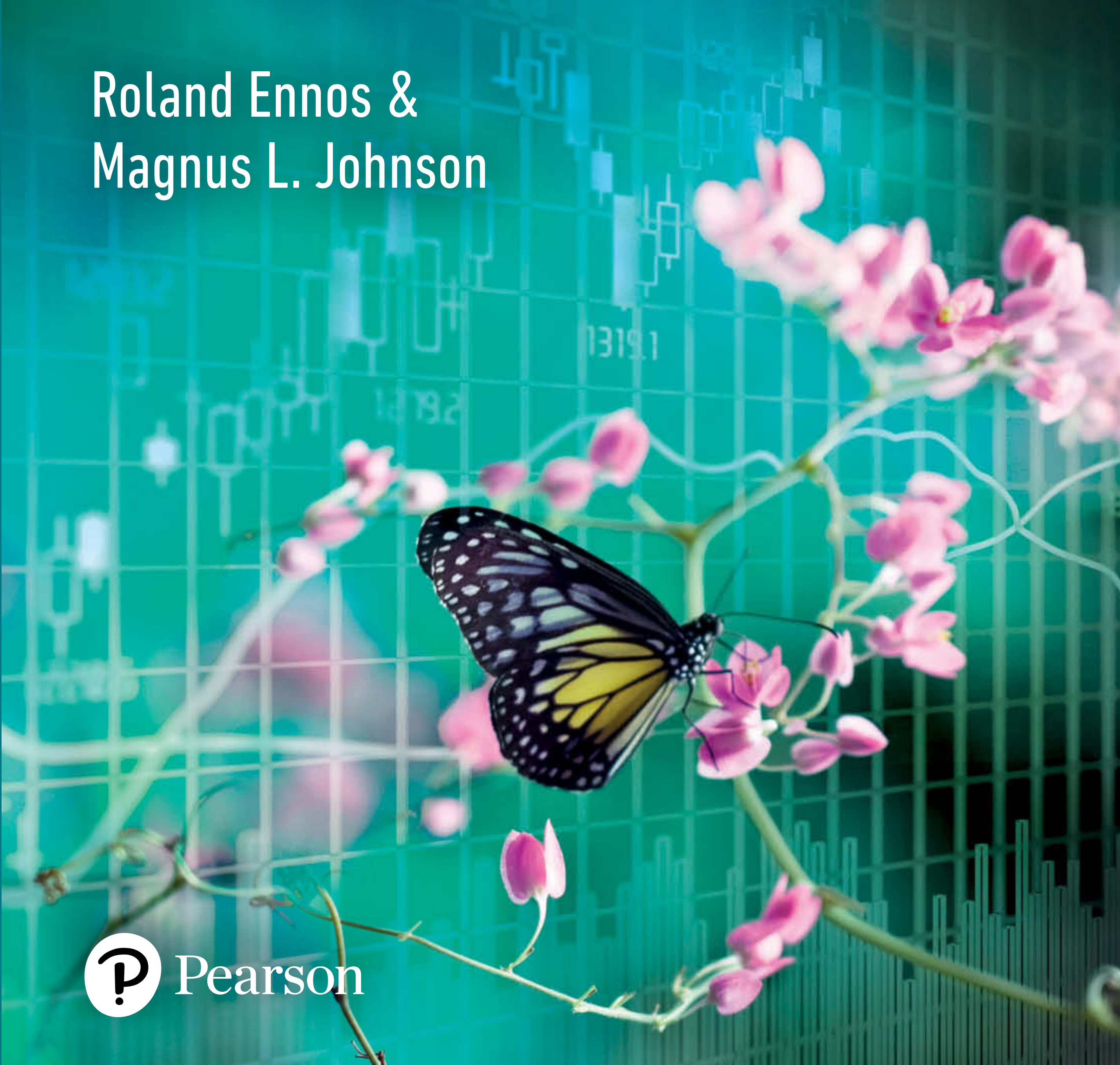
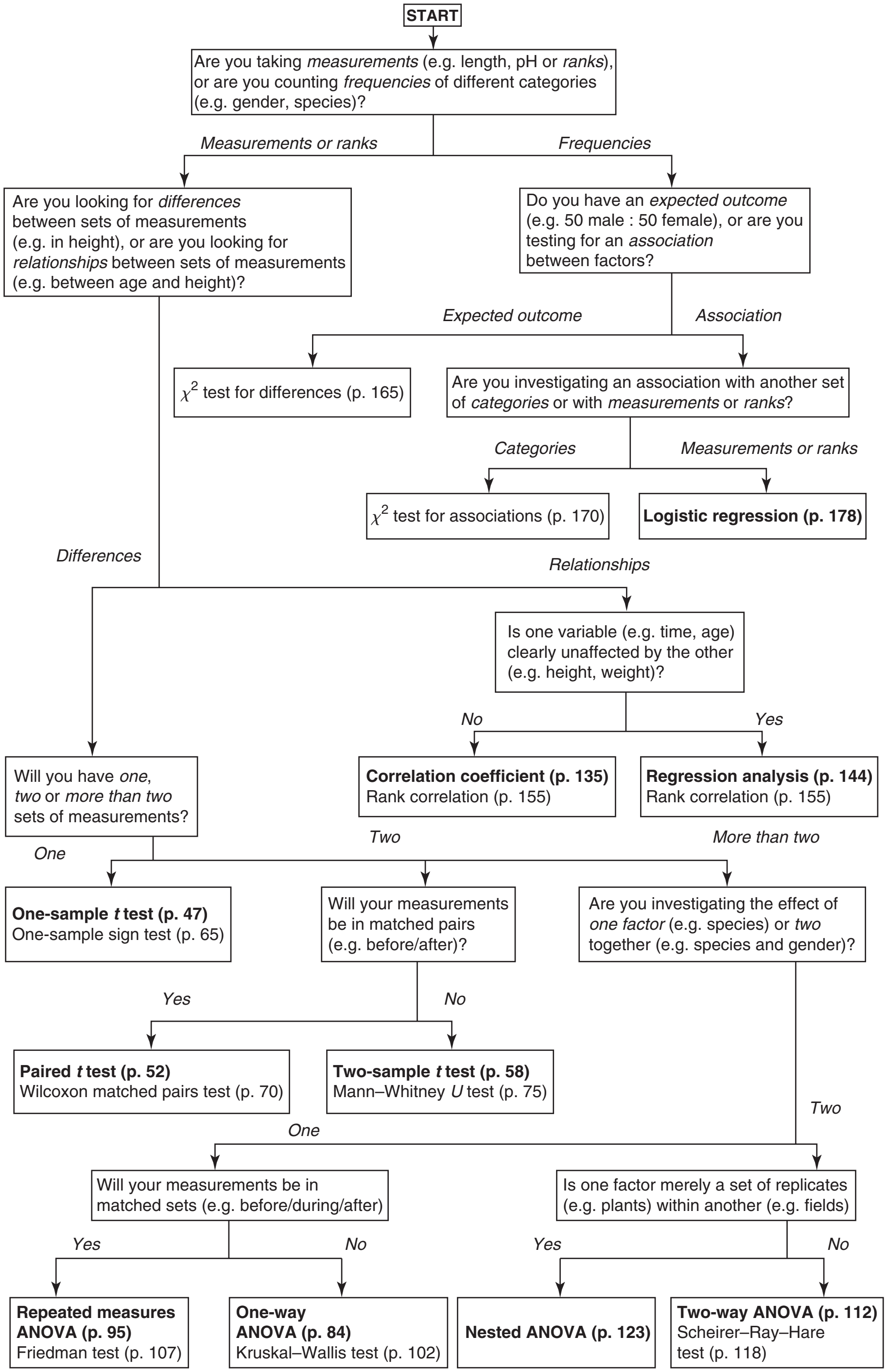
Statistical and Data Handling Skills in Biology

Fourth Edition

Roland Ennos &
Magnus L. Johnson



Pearson


START
Are you taking measurements (e.g. length, pH or ranks), or are you counting frequencies of different categories (e.g. gender, species)?
Measurements or ranks
Frequencies
Are you looking for differences between sets of measurements (e.g. in height), or are you looking for relationships between sets of measurements (e.g. between age and height)?
Do you have an expected outcome (e.g. 50 male : 50 female), or are you testing for an association between factors?
Expected outcome
Association
χ2 test for differences (p. 165)
Are you investigating an association with another set of categories or with measurements or ranks?
Categories
Measurements or ranks
χ2 test for associations (p. 170)
Logistic regression (p. 178)
Differences
Relationships
Is one variable (e.g. time, age) clearly unaffected by the other (e.g. height, weight)?
No
Yes
Correlation coefficient (p. 135)
Rank correlation (p. 155)
Regression analysis (p. 144)
Rank correlation (p. 155)
Will you have one, two or more than two sets of measurements?
One
Two
More than two
One-sample t test (p. 47)
One-sample sign test (p. 65)
Will your measurements be in matched pairs (e.g. before/after)?
Are you investigating the effect of one factor (e.g. species) or two together (e.g. species and gender)?
Yes
No
Paired t test (p. 52)
Wilcoxon matched pairs test (p. 70)
Two-sample t test (p. 58)
Mann–Whitney U test (p. 75)
Two
One
Will your measurements be in matched sets (e.g. before/during/after)?
Is one factor merely a set of replicates (e.g. plants) within another (e.g. fields)
Yes
No
Yes
No
Repeated measures ANOVA (p. 95)
Friedman test (p. 107)
One-way ANOVA (p. 84)
Kruskal–Wallis test (p. 102)
Nested ANOVA (p. 123)
Two-way ANOVA (p. 112)
Scheirer–Ray–Hare test (p. 118)

This page intentionally left blank

Statistical and Data Handling Skills in Biology

Visit the *Statistical and Data Handling Skills in Biology,* Third Edition, Companion Website at
www.pearsoned.co.uk/ennos
to find valuable student learning material including:

- An Introduction to SPSS version 24 for Windows
- An Introduction to RStudio

Statistical and Data Handling Skills in Biology

Fourth Edition

Roland Ennos
University of Hull in the School of Environmental Sciences

Magnus L. Johnson
University of Hull in the School of Environmental Sciences

Pearson

Harlow, England • London • New York • Boston • San Francisco • Toronto • Sydney • Dubai • Singapore • Hong Kong
Tokyo • Seoul • Taipei • New Delhi • Cape Town • São Paulo • Mexico City • Madrid • Amsterdam • Munich • Paris • Milan

PEARSON EDUCATION LIMITED
KAO Two
KAO Park
Harlow CM17 9NA
United Kingdom
Tel: +44 (0)1279 623623
Web: www.pearson.com/uk

First published 2000 (print)
Second edition published 2007 (print)
Third edition published 2012 (print and electronic)
Fourth edition published 2018 (print and electronic)

ISBN: 978-1-292-08603-3 (print)
978-1-292-08606-4 (PDF)
978-1-292-13311-9 (ePub)

British Library Cataloguing-in-Publication Data
A catalogue record for the print edition is available from the British Library

Library of Congress Cataloging-in-Publication Data
Names: Ennos, A. R., author. | Johnson, Magnus, author.
Title: Statistical and data handling skills in biology / Roland Ennos, Magnus Johnson.
Description: Fourth edition. | New York : Pearson Education, 2017. | Includes index.
Identifiers: LCCN 2017048353 | ISBN 9781292086033 (Print) | ISBN 9781292086064 (PDF) | ISBN 9781292133119 (ePub)
Subjects: LCSH: Biometry.
Classification: LCC QH323.5 .E57 2017 | DDC 570.1/5195—dc23
LC record available at https://lccn.loc.gov/2017048353

10 9 8 7 6 5 4 3 2 1
22 21 20 19 18

Print edition typeset in 9/12.5pt Stone Serif ITC Pro by iEnergizer Aptara®, Ltd.
Printed and bound in Malaysia

NOTE THAT ANY PAGE CROSS REFERENCES REFER TO THE PRINT EDITION

Dedication

To our gardeners, Yvonne and Rowan

This page intentionally left blank

Brief contents

This page intentionally left blank

Contents

Supporting resources

Visit **www.pearsoned.co.uk/ennos** to find valuable online resources

Companion Website for students

- An Introduction to SPSS version 24 for Windows
- An Introduction to RStudio

For more information please contact your local Pearson Education sales representative or visit **www.pearsoned.co.uk/ennos**

List of figures and tables

Figures

Tables

This page intentionally left blank

Preface

It is six years since the third edition of *Statistical and Data Handling Skills in Biology* was published, and we are grateful to Pearson Education for allowing us the opportunity to update and expand the text for a fourth edition.

Errors in the previous edition have been corrected. However, the chief change has resulted from the recruitment of Magnus L. Johnson, a marine biologist and enthusiastic user of the free statistical package R, as a co-author. Magnus has overseen the replacement of instructions for the use of the ageing package, MINITAB, by introducing guidance on how to carry out tests with RStudio, a user-friendly version of R. The text therefore can act as an introduction to this highly flexible package, giving students the ability, with abundant help available on the web, to carry out a wide range of complex statistical analyses. As examples are worked through in both packages this book may be of particular interest to students and tutors trying to make the jump from SPSS to R.

The text has also been restructured to make it easier to use. The first chapter has been rewritten to introduce students more simply to the importance of statistics. The chapter on dealing with numbers, always marginal in a statistics book, has been replaced by a chapter describing how to write about and present statistics in papers, theses and reports. This should help students avoid common mistakes and present statistical information clearly, concisely and correctly.

Like the earlier editions, the text is based on courses we have given to students at the University of Manchester and the University of Hull. We are heavily indebted to those students who have taken these courses for their feedback. With their help, and with that of several of Pearson Education's reviewers, many errors have also been eliminated, and we have learnt much more about statistics, although we take full responsibility for those errors and omissions that remain.

Finally, we would like to thank Yvonne and Rowan for their unfailing support during the writing of this and previous editions of the text.

This page intentionally left blank

Publisher's acknowledgements

We are grateful to the following for permission to reproduce copyright material:

Screenshots

Screenshots on pages 26, 55, 62 from MS Excel, Microsoft product screenshots reprinted with permission from Microsoft Corporation; Screenshots on pages 24, 25, 28, 34, 39, 49, 54, 60, 61, 67, 72, 77, 88, 93, 97, 97, 104, 109, 114, 120, 125, 125, 139, 141, 146, 157, 167, 168, 172, 173, 180, 206, 209 from SPSS, IBM, Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation. SPSS Inc. was acquired by IBM in October, 2009. IBM, the IBM logo, ibm.com, and SPSS are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "IBM Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml; Screenshots on pages 25, 27, 28, 31, 35, 38, 41, 50, 55, 62, 69, 74, 79, 89, 92, 94, 99, 100, 105, 110, 116, 118, 121, 127, 140, 148, 154, 158, 169, 175, 181, 182, 195, 197, 199, 207, 207, 210 from RStudio Integrated Development Environment (IDE) for R, RStudio and Shiny are trademarks of RStudio, Inc.

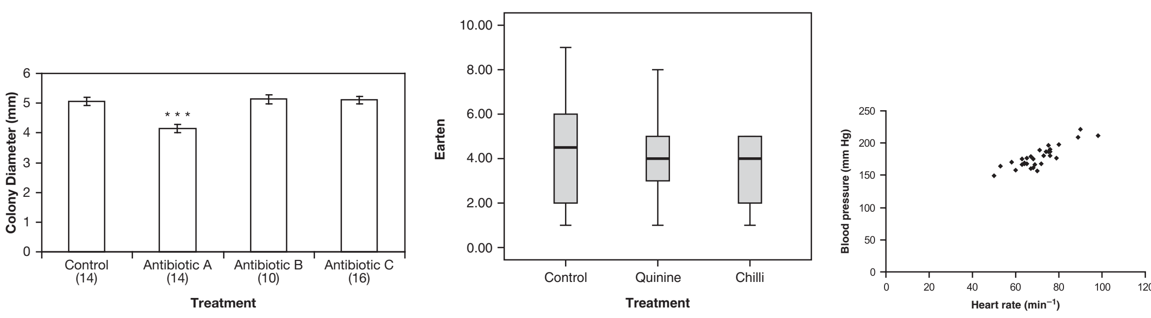
This page intentionally left blank

1 An introduction to statistics

1.1 Becoming a biologist

If you're reading this text, you're probably a biologist and have been told you have to learn about statistics. However, if like most biologists, you're not a great fan of mathematics, you may feel some resentment about this. Why should you have to learn about this dry subject rather than concentrate on what you're really interested in: the human body, animals, plants, microbes or ecosystems? How relevant is it, and why is it so dull? Unfortunately, however, you need to know about statistics for two reasons.

First, you'll have to be able to understand what lecturers are talking about when they describe the research other biologists have done. For instance, you may have already been shown the following sorts of graphs and tables (Figure 1.1). You'll need to know the meanings of all the bars, asterisks, lines and brackets as well as the meanings of the numbers in the figure legends. It's only then that you'll know what the researchers have actually done and can decide whether they have come to sensible conclusions.



(a) (b) (c)

	White	Asian	Afro Caribbean	Total
Have the condition	35 (38.6)	9 (10.2)	13 (8.2)	**57**
Healthy	162 (158.4)	43 (41.8)	29 (33.8)	**234**
Total	**197**	**52**	**42**	**291**

(d)

Figure 1.1 Typical figures that you might be shown in lectures or see in scientific papers: (a) a bar chart (Figure 5.3a), (b) a box and whisker plot (Figure 5.6), (c) a scatter plot (Figure 6.6) and (d) a contingency table (see below). You'll need to be able to interpret these and to know the meaning of the asterisks, bars, dots and brackets.

Second, as a biologist you'll have to analyse the results of laboratory practicals yourself. In your final year, you'll even have to carry out your own research project, for which you'll have to analyse your own results and design your own surveys and experiments. Statistics is therefore essential for biologists. But there are awkward questions that you are entitled to ask, and which this text has to answer, to give you valid reasons why the whole subject is so essential and why it seems so complicated.

1.2 Awkward questions

The first thing you are invariably told to do when carrying out a research project is to make repeated measurements: to include tens or even hundreds of people in surveys; or to have large numbers of replicates in experiments. This seems to be a great deal of wasted effort, so the first question that this text needs to answer is **why do biologists have to repeat everything?**

You are then told to subject your results to statistical analysis. You might reasonably feel that as you are studying biology, you should be able to leave the horrors of maths behind you. So, the second question that any biological statistics text needs to answer is **why do biologists have to bother with statistics?**

Many students also have a problem with the ideas behind statistics. You might well have already found that statisticians seem to think in a weird, inverted way that is at odds with normal scientific logic. So, this text also has to answer the question **why is statistical logic so strange?**

Finally, students often complain, not unreasonably, about the size of statistics books and the amount of information they contain. The reason for this is that there are large numbers of statistical tests, so this text also needs to answer the question **why are there so many different statistical tests?**

This opening chapter provides answers to these questions to help put the subject into perspective and encourage you to stick with it. This chapter introduces the information and the order in which it is set out throughout the text; it should help you work through the text, either in conjunction with a taught course or on your own. For those more experienced and confident about statistics, in particular those with an experiment to perform or results to analyse, you can go directly to the **decision chart for simple statistical tests** (Figure 1.2) introduced later in this chapter and also inside the front cover. This will help you choose the statistical test you require and direct you to the instructions on how to perform each test, which are given later in the text. Hence the material can also be used as a handbook to keep around the laboratory and consult when required.

1.3 Why biologists have to repeat everything

At first sight, it seems strange that biologists have to repeat everything when they are conducting surveys or analysing experiments. After all, physicists don't need to do it when they are comparing the masses of sub-atomic particles. Chemists don't need to when they're comparing the pHs of different acids. And engineers don't need to when they are comparing the strengths of different-shaped girders.

They can just generalise from single observations; if a single neutron is heavier than a single proton, then that will be the case for all of them.

However, if you decided to compare the heights of fair- and dark-haired women, it is obvious that measuring just one fair-haired and one dark-haired woman would be insufficient. If the fair-haired woman was taller, you couldn't generalise from this single observation to tell whether fair–haired women are *on average* taller than dark-haired ones. The same would be true if you compared a single man and a single woman, or one rat that had been given growth hormone and another that had not. Why is this? The answer is, of course, that in contrast to sub-atomic particles, which are all the same, people (in common with other organisms, organs and cells) are all *different* from each other. In other words, they show **variability**, so no one person or cell or experimentally treated organism is typical. It is to get over this problem that biologists have to do so much work and have to use statistics.

To overcome variability, the first thing you have to do is to make **replicated observations** of a **sample** of all the observations you could possibly make. You can do this in two ways.

1. You can carry out a **survey**, sampling at random from the existing **population** of people or creatures or cells. You might measure 20 fair-haired and 20 dark-haired women, for instance.
2. You can create your own samples by performing an **experiment**. Your experimental subjects are then essentially samples of the infinite **population** of subjects that you *could* have created if you had infinite time and resources. You might, for instance, perform an experiment in which 20 **experimentally treated** rats were injected with growth hormone and 20 other **controls** were kept in exactly the same way, except that they received no growth hormone.

1.4 Why biologists have to bother with statistics

At first glance, it is hard to know exactly what you should do with all the observations that you make, given that all creatures are different. This is where statistics comes in; it helps you to deal with the variability. First, it helps you to examine exactly how your observations vary; in other words, to investigate the **distribution** of your samples. Second, it helps you to calculate reasonable **estimates** of the situation in the whole population; for instance, working out how tall the women are *on average*. These estimates, known as **descriptive statistics**, are introduced in Chapter 2.

Descriptive statistics summarise what you know about your samples. However, few people are satisfied with simply finding out these sorts of *facts*; they usually want to answer *questions*. You might want to know whether one group of the women was on average taller than the other, or whether the rats that had been given the growth hormone were heavier than those which hadn't. You can answer questions such as these by carrying out **hypothesis testing**. If you compared two groups of organisms, you would undoubtedly find that they were at least slightly different (for instance, the fair-haired women might be taller than the dark-haired women), but there could be two reasons for this. It could be because there really is a difference in height between fair- and dark-haired women. However, it is also possible that you obtained this difference *by chance* by virtue of the particular people you chose. To discount this possibility, you have to carry out a **statistical test** (in this case, a two-sample *t* test) to work out the probability that the apparent effects *could* have occurred by chance. If this probability was small enough, you could make the judgement that you could discount it and decide that the effect was **significant**. In this case, you would then have decided that fair-haired women are *significantly taller* than dark-haired women.

1.5 Why statistical logic is so strange

All of this has the consequence that the logic of hypothesis testing is rather counterintuitive. When you are investigating a subject in science, you typically make a hypothesis that something interesting is happening—for instance, in our case that fair-haired women are taller than dark-haired women—and then set out to test it. In statistical hypothesis testing, you do the opposite. You construct a **null hypothesis** that *nothing interesting* is happening, in this case that fair- and dark-haired women have the same mean height, and then test whether this null hypothesis is likely to be true. Statistical tests have four main stages.

null hypothesis
A preliminary assumption in a statistical test that the data shows no differences or associations. A statistical test then works out the probability of obtaining data similar to your own by chance.

Step 1: Formulating a null hypothesis

The null hypothesis you must set up is the opposite of your scientific hypothesis: that there are no differences or relationships. (In the case of the fair- and dark-haired women, the null hypothesis is that they are the same height.)

Step 2: Calculating a test statistic

The next step is to calculate a **test statistic** which measures the size of any effect (usually a difference between groups or a relationship between measurements) relative to the amount of variability there is in your samples. Usually (but not always), the larger the effect, the larger the test statistic.

Step 3: Calculating the significance probability

Knowing the test statistic and the size of your samples, you can then calculate the probability of getting the effect you have measured, just by chance, *if the null hypothesis were true*. This is known as the **significance probability**. Generally, the larger the test statistic and sample size, the smaller the significance probability.

significance probability
The chances that a certain set of results could be obtained if the null hypothesis were true.

Step 4: Deciding whether to reject the null hypothesis

The final stage is to decide whether to reject the null hypothesis or not. By convention it has been decided that you can reject a null hypothesis if the significance probability is less than or equal to 1 in 20 (a probability of 5%, or 0.05). If the significance probability is greater than 5%, however, you have no evidence to reject the null hypothesis (*but this does not mean you have evidence to support it!*).

The 5% cut-off is actually a compromise to reduce the chances of biologists making mistakes about what is really going on. For instance, there is a 1-in-20 chance of finding an apparently significant effect, even if there wasn't a real effect. If the cut-off point had been lowered to, say, 1 in 100, or 1%, the chances of making this sort of mistake (known to statisticians as a **type 1 error**) would be

type 1 error
The detection of an apparently significant difference or association, when in reality there is no difference or association between the populations.

type 2 error
The failure to detect a significant difference or association, when in reality there is a difference or association between the populations.

reduced. On the other hand, the chances of failing to detect a real effect (known as a **type 2 error**) would be increased by lowering the cut-off point.

As a consequence of this probabilistic nature, performing a statistical test does not actually allow you to *prove* anything conclusively. If your test tells you there is a significant effect, there is still a small chance that there might not really have been one. Similarly, if your test is not significant, there is still a chance that there might really have been an effect.

1.6 Why there are so many statistical tests

Even if we accept that statistical tests are necessary in biology and can cope with the unusual logic, it is perhaps not unreasonable to expect that we should be able to analyse all our results using just a single statistical test. However, statistics texts such as this one contain large numbers of different tests. Why are there so many? There are three main reasons for this. First, there are several very different ways of quantifying things and hence different types of data that you can collect. Second, data can vary in different ways. Third, there are very different questions you might want to ask about the data you have collected.

1.6.1 Types of data

measurement
A character state which can meaningfully be represented by a number.

normal distribution
The usual symmetrical and bell-shaped distribution pattern for measurements that are influenced by large numbers of factors.

parametric test
A statistical test which assumes that data is normally distributed.

non-parametric test
A statistical test which does not assume that data is normally distributed, but instead uses the ranks of the observations.

(a) Measurements The most common way of quantifying things about organisms is to take **measurements** (of things such as height, mass or pH), to give what is also known as **interval data**. This sort of data can vary **continuously,** like weight (e.g. 21.23 or 34.651 kg), or **discretely**, like the numbers of hairs on a fruit fly (e.g. 12 or 18). As we shall see in Chapter 2, many of these measurements vary according to the **normal distribution**. There is a set of tests, the so-called **parametric tests**, that assume that this is the case. On the other hand, many measurements do not vary in this way. This sort of data either has to be **transformed** until it does vary according to the normal distribution (Chapter 3) or, if that is not possible, must be analysed using a separate set of tests, the **non-parametric tests**, which make no assumption of normality.

(b) Ranks On many occasions, you may only be able to put your measurements into an order, without the actual values having any real meaning. This **ranked** or **ordinal data** includes things like the pecking order of hens (e.g. 1st, 12th), the seriousness of an infection (e.g. none, light, medium, heavy) or the results of questionnaire data (e.g. 1 = poor to 5 = excellent). This sort of data *must* be analysed using **non-parametric tests**.

frequency
The number of times a particular character state turns up.

(c) Categorical data Some features of organisms are impossible to quantify in any way. You might only be able to classify them into different **categories**. For instance, birds belong to different species and have different colours; people could be diseased or well; and cells could be mutant or non-mutant. The only way of quantifying this sort of data is to count the **frequency** with which each category occurs. This sort of data is usually analysed using χ^2 (chi-squared) tests or logistic regression (Chapter 7).

1.6.2 Types of questions

Statistical tests are designed to answer two main types of questions: Are there differences between sets of measurements? Or, are there relationships between them?

(a) Testing for differences between sets of measurements There are many occasions when you might want to test to see whether there are **differences** between two or more sets of measurements. For instance, we have already looked at the case of comparing the height of fair- and dark-haired women. An even more common situation is when you carry out experiments; you commonly want to know if experimentally treated organisms or cells are different from controls. Or you might want to compare two sets of measurements taken on a single group of organisms; for instance, you might want to know if the medical condition of patients was different before and after treatment. Tests to answer these questions are described in Chapter 4. Alternatively, you might want to see if several different types of organisms (for instance, five different bacterial strains), or ones that had been subjected to several types of treatments (for instance, wheat subjected to different levels of nitrate and phosphate), were different from each other. Tests to answer these questions are described in Chapter 5.

(b) Testing for relationships between measurements Another thing you might want to do is to take two or more measurements on a single group of organisms or cells and investigate how the measurements are **related**. For instance, you could investigate how people's heart rates vary with their blood pressure; how weight varies with age; or how the concentrations of different cations in neurons vary with each other. This sort of knowledge can help you work out how organisms operate or enable you to predict things about them. Chapter 6 describes how statistical tests can be used to quantify relationships between measurements and work out if the apparent relationships are real.

(c) Testing for differences and relationships between categorical data There are three different things you might want to find out about categorical data. You might want to determine whether there are different frequencies of organisms in different categories from what you would expect; whether rats turn more frequently to the right in a maze than to the left, for instance. Alternatively, you might want to find out whether categorical traits, for instance people's eye and hair colour, are associated: are people with dark hair also more likely to have brown eyes? Finally, you might be interested in working out how quantitative measurements might affect categorical traits—for instance, are tall people more likely to have brown eyes? Tests to answer all these sorts of questions are described in Chapter 7.

1.7 Using the decision chart

The logic of the previous section has been developed and expanded to produce a **decision chart** (Figure 1.2 and on the inside front cover). Although not fully comprehensive, the chart includes virtually all the tests that you will likely encounter as an undergraduate. If you are already a research biologist, it may also

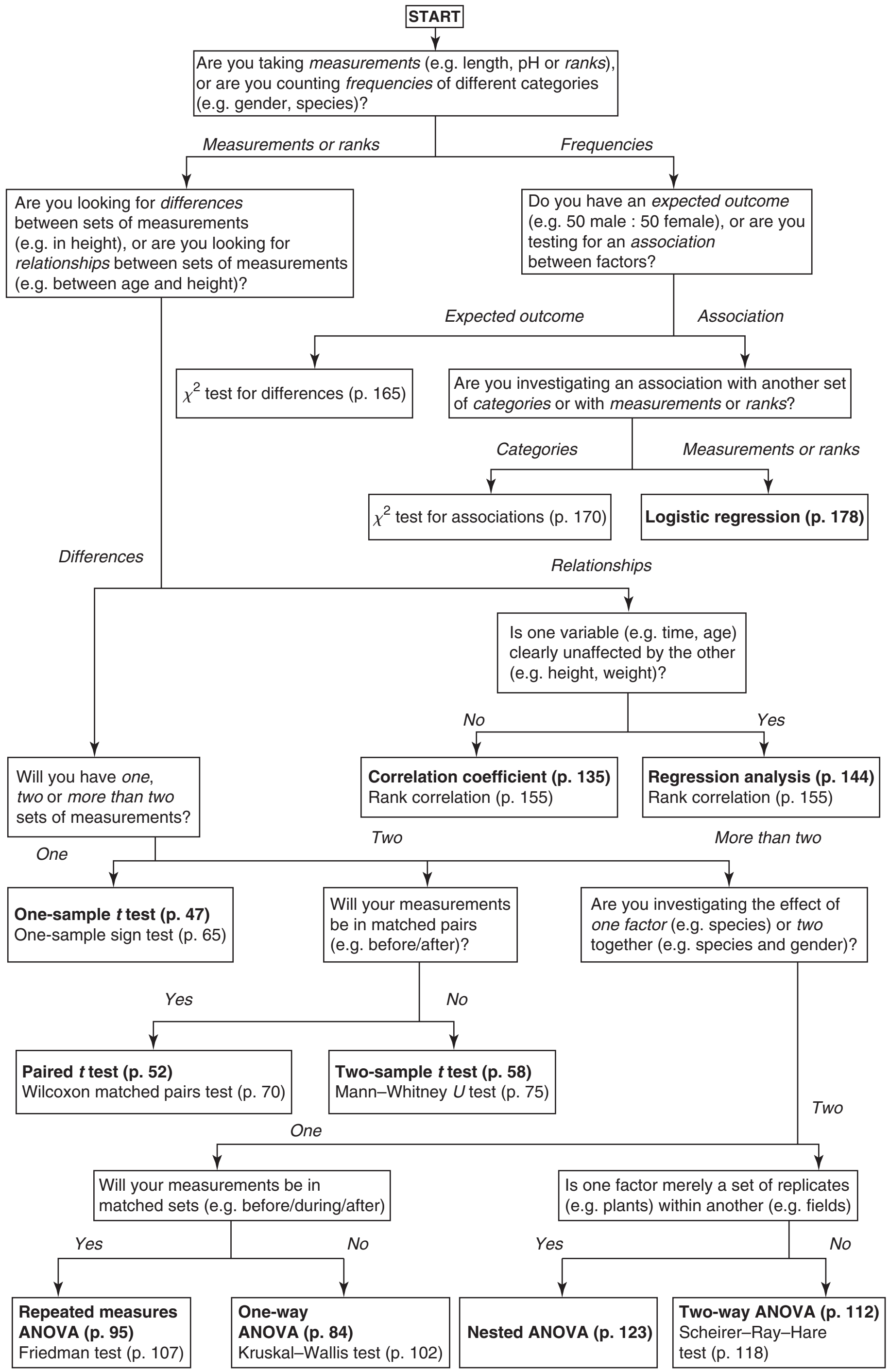


Figure 1.2 Decision chart for statistical tests. Start at the top and follow the questions down until you reach the appropriate box. The tests in normal type are non-parametric equivalents for irregularly distributed or ranked data.

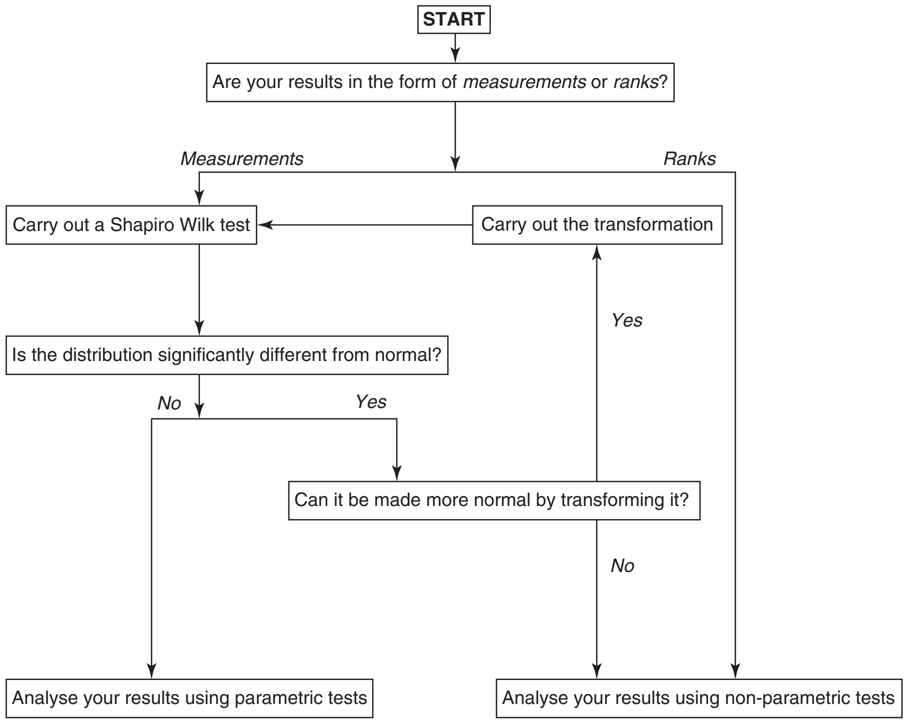


Figure 1.3 Flow chart showing how to deal with measurements and rank data. Start at the top, answer the questions and transform data where appropriate before deciding whether you can use parametric tests or have to make do with non-parametric ones.

include all the tests you are ever likely to use over your working life! If you follow down from the start at the top and answer each of the questions in turn, this should lead you to the statistical test you need to perform.

There is only one complication. The final box may have two alternative tests: a parametric test, shown in bold type, and an equivalent non-parametric test, shown in normal type. You are always advised to use the parametric test if it is valid, because parametric tests are more powerful in detecting significant effects. Use the non-parametric test if you are dealing with ranked data, irregularly distributed data that cannot be transformed to the normal distribution, or have measurements which can only have a few, discrete, values. Before deciding which tests to carry out, therefore, you need to investigate the distribution of your data (Figure 1.3 and on the inside back cover) to see whether it is valid to carry out parametric tests, or if it is possible to transform your data so that you can.

1.8 Using this text

1.8.1 Carrying out tests

Once you have made your decision, the chart will direct you to a page in the main section of this text (Chapters 4–7), which describes the main statistical tests. You should go to the page indicated, where details of the test will be described. Each test description will do five things.

1. It will tell you the sorts of questions the test will enable you to answer and give examples to show the range of situations for which it is suitable. This will help you make sure you have chosen the right test.
2. It will tell you when it is valid to use the test.
3. It will describe the rationale and mathematical basis for the test; basically, it will tell you how it works.
4. It will show you how to perform the test using a calculator and/or the computer-based statistical packages SPSS and RStudio.
5. It will tell you how to present the results of the statistical tests.

1.8.2 Designing experiments

As a research biologist, you will not only have to choose statistical tests and perform the analysis yourself; you will also have to design your own experiments. Chapter 8 will show how you can use the information about statistics set out in the main part of the text to design better experiments.

1.8.3 Complex statistical analysis

This text describes most of the statistical tests you will need to analyse straightforward experiments and surveys: ones that look at one or, at most, two factors. I would strongly recommend that you stay as far as possible within these limits when you design your experiments. There may be some occasions, however, particularly within some branches of biology, where you simply have to carry out and analyse rather more complex experiments or investigate huge sets of data. Chapter 9 describes some of the complex statistical techniques that can help you investigate several factors at once.

1.8.4 Presenting and discussing statistics

Many students are unsure about how much information they need to give about statistics in the write-ups of their practical classes, dissertations and, in due course, theses and scientific papers. The final chapter describes how to present information about what statistical tests you did and why, how to present the results of your tests and the level at which you should discuss these results. This will enable you to produce professional write-ups.