Statistics for the Humanities

John Canning

John Canning,

First Edition 2014.

Brighton, UK

A website accompanying this book is available at www.statisticsforhumanities.net

@statistics4hums

Comments/ corrections/ questions? Please email j.canning@brighton.ac.uk

This book makes use of images available under Creative Commons licenses or are in the public domain. Except where otherwise noted the images have been published online under creative commons licences and are used in this book in good faith. Further details about Creative Commons are available at http://www.creativecommons.org

Cover Illustration: Altar of Domitius Ahenobarbus, Louvre Photo: Marie-Lan Nguyen (2007) Public domain
http://commons.wikimedia.org/wiki/File%3AAltar_Domitius_Ahenobarbus_Louvre_n2.jpg

Figure 0.1 cc BY StuSeeger (2008) http://www.flickr.com/photos/56274840@N02/galleries/72157633089348310

Figure 0.2 cc BY Amri HMS (2007) http://www.flickr.com/photos/kulimpapat147/3541489234/

Figure 0.3 cc BY Paul Downey (2007) http://www.flickr.com/photos/psd/8419933758/

Figure 2.1 cc BY SA Steven Isaacson (2006) http://www.flickr.com/photos/spi/190603061/

Figure 2.2 Photographer unknown (1963). Repository: American Jewish Historical Society. Parent Collection: American Jewish Congress records, undated, 1916-2006 (I-77) http://digital.cjh.
org/R/?func=dbin-jump-full&object_id=1432347 No known copyright restrictions.

Figure 2.3 cc BY Alex Wellerstein. http://www.flickr.com/photos/restricteddata/6322465279/

in/photostream/

Figure 3.1 cc BY John Canning (2013)

Figure 4.4 cc BY Andrew Loomis (1938) Collections Canada.

http://collectionscanada.gc.ca/pam_archives/index.php?fuseaction=genitem.displayItem&lang=eng&rec_nbr=2842836

Figure 5.1 cc BY WattAgNet.com (2009) www.WATTAgNet.com. Image found at

http://www.245246flickr.com/photos/wattagnet/6721769653/in/photostream/

Figure 6.1 cc BY Charles D P Miller (2007) http://www.flickr.com/photos/cdpm/3644503386/

Figure 9.1 © R Foundation (2013). Used under the terms outlined at http://www.r-project.org/screenshots/screenshots.html

Figure 11.4 User: Wujaszek. This image is in the public domain because its copyright has expired in the European Union and those countries with a copyright term of life of the author plus 70 years.
http://en.wikipedia.org/wiki/File:William_Sealy_Gosset.jpg

Figure 13.2 This image is in the public domain because its copyright has expired in the European Union and those countries with a copyright term of life of the author plus 70 years.
http://en.wikipedia.org/wiki/File:Karl_Pearson_2.jpg

Figure 15.1 cc BY SA Markus Hagenlocher (2006) http://en.wikipedia.org/wiki/File:Futterr% C3%BCbe.jpg

Figure 15.2 cc BY SA Brian Sims (2011) http://en.wikipedia.org/wiki/File:Occupy_Oakland_ 99_Percent_signs.jpg

Figure 16.1 No copyright. This image is in the public domain due to its age. Author died in 1858, material is public domain. http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.
jpg

Figure 16.2 cc BY John Canning (2013)

Figure 18.1 cc BY Smabs Sputzer http://www.flickr.com/photos/10413717@N08/2397168892/

Figure 18.2 User: Bletchley This image is in the public domain because its copyright has expired in the European Union and those countries with a copyright term of life of the author plus 70 years. http://en.wikipedia.org/wiki/File:R._A._Fischer.jpg

Figure 20.17 cc BY SA User:BenFrantzDale (2005) http://en.wikipedia.org/wiki/File:Chartjunk-example.

svg

Figure 20.18 cc BY User:sacks08 (2010) http://www.flickr.com/photos/29071316@N06/4789882070/

Figure 20.19 cc BY SA Doug Kerr (2007) http://www.flickr.com/photos/dougtone/2720153713/

Figure 20.20 Coin image: cc BY Ivan Walsh (2009) http://www.flickr.com/photos/ivanwalsh/3782836923/

Graph: cc BY John Canning

Appendix cc BY Henry Lawford (2008) https://www.flickr.com/photos/herry/2988080770

Appendix cc BY Michiel2005 (2011) https://www.flickr.com/photos/govert1970/6432136985

Contents

Preface...v

Chapter 0: The mathematics behind statistics....................................1

Chapter 1: Introduction..7

Chapter 2: How many and how big?...11

Chapter 3: Summarising data..15

Chapter 4: Measuring spread..23

Chapter 5: Sampling..33

Chapter 6: Measuring change...37

Chapter 7: Conclusions though evidence..43

Chapter 8: Key concepts in statistics..47

Chapter 9: Using statistics intelligently...53

Chapter 10: Comparing Groups: the chi-square test........................55

Chapter 11: Comparing two groups: the Student's t-test..................59

Chapter 12: Analysis of variance 67

Chapter 13: Understanding relationships..75

Chapter 14: Predicting new observations from known data.............79

Chapter 15: Ranking data..87

Chapter 16: Everything happens somewhere: Spatial data..............97

Chapter 17: Having confidence in data..103

Chapter 18: Association causation and effect................................107

Chapter 19: Collecting your own data...113

Chapter 20: Presenting data..117

Chapter 21: The next steps: Developing understanding in statistics 125

Appendix: Dealing with non-decimal units......................................127

Critical values..131

Preface

> *Well-rounded graduates, equipped with core quantitative skills, are vital if the UK is to retain its status as a world leader in research and higher education, rebuild its economy, and provide citizens with the means to understand, analyse and criticise data. Quantitative methods facilitate `blue skies' research and effective, evidence-based policy. Yet, the UK currently displays weak quantitative ability in particular, but not exclusively in the humanities and social sciences. [1]*

Just 15% of students in England study mathematics beyond GCSE level [2] However, many of this non-mathematics studying majority find that they need mathematical skills for the advanced study of other subjects, including humanities and social science subjects at school or university or in their job. As a recent report into the teaching of mathematics noted, this is not a new problem, but there has been a significant increase in mathematical requirement for jobs. [3] Without mathematical, and in particular statistical skills whole areas of the social sciences and humanities are inaccessible to research students and future academics. With a few exceptions statistics rarely forms part of the humanities curriculum. Where these courses do exist students often dislike them and regard them as hurdle to be cleared rather than a skill to be developed and nurtured for the long term. Concern has also been expressed about the increasing tendency of students to take the GCSE mathematics exams a year early. [4]The problem is not that students take the subject early then move onto more advanced studies - rather they clear the maths hurdle in order to be able to focus their attention on other subjects.

The humanities student or academic who wishes (or needs) to study statistics is left in something of a quandary. The internet, the first port of call to most people seeking to fill a knowledge gap, is bewildering to the beginner. Wikipedia, whatever its vices, can be a helpful starting point for finding information on a variety of subjects, but its statistical articles are complex and not accessible for those new to the subject. Good material is available online, but it is not that easy to find. There are numerous introductory books to statistics, but the reader will often find the examples given biased towards the sciences and social sciences. Consequently the applicability of the examples given can be difficult to discern for the humanities student. A second issue has been the rise of software for statistical analysis over the past 20 years. Statistical analysis software is wonderful; vast data sets can be processed in matter of seconds. The problem for the student beginner is that these software packages are taught alongside statistics and the student is attempting to learn two major new skills simultaneously.

Like 85% of 16-year-olds, I `dropped' mathematics after GCSE. However, I soon found that my studies of geography at A-level and at university required some knowledge of statistics. Most importantly this was not simply going back to my early mathematical studies, but learning entirely new skills altogether. My career path bought me into the then Subject Centre for Languages, Linguistics and Area Studies, based at the University of Southampton. In this environment I found that my knowledge of statistics was invaluable. My colleague Angela Gallagher-Brett and I designed workshops to teach social science research methods to humanities academics, so that they could conduct research into teaching and learning. [5] The need for an introductory book for those working in the humanities became increasingly apparent to me.

Although I felt the need for the book, I wasn't sure if it would ever be written, not least by me. However, when the British Academy recognised the need to support Quantitative Skills in the humanities and social sciences, the real possibility of a book arose.

First and foremost, I would like to thank the British Academy for their grant which made it possible for me to write the book. Anandini Yoganathan, Senior Policy Advisor has been supportive from the initial conception of the book right through to its publication. I am also grateful to Vivienne Hurley, Director of Programmes, for her input. I would also like to thank colleagues at Southampton for their support, comments and feedback including Lisa Bernasek, Kate Borthwick, Erika Corradini, Alison Dickens, Angela Gallagher-Brett, Laurence Georgin, Liz Hudswell and Mike Kelly at the LLAS Centre. Graeme Earl, David Wheatley and Rich Harris provided helpful comments. Peter Mitchell from the University of Sheffield provided some very detailed comments and corrections.

Hannah Burd provided extensive comments on an early drafts of the book.

At home my wife Michelle and young sons, Samuel and Elijah have been always supportive throughout.

1. British Academy (2012) Society counts: Quantitative Skills in the Social Sciences and Humanities. London: British Academy.

2. ibid p.3 For the benefit of non-UK readers GCSE's are the exam taken by students in England, Wales and Northern Ireland at the age of about 16.

3. Carol Vorderman, Christopher Budd, Richard Dunne, Mahzia Hart, Roger Porkess (2011)A world-class mathematics education for all our young people Report commissioned by the Conservative Party.

4. Department for Education (2011) Early entry to GCSE examinations.(London: DfE)

5. Canning, J. and A. Gallagher-Brett (2010) Building a bridge to pedagogic research: teaching social science research methods to humanities practitioners. Journal of Applied Research in Higher Education, 2, 3-9

Copyright and creative commons licenses and permissions.

This book makes use of images available under Creative Commons licenses or are in the public domain. Except where otherwise noted the images have been published online under creative commons licences and are used in this book in good faith. Further details about Creative Commons are available at http://www.creativecommons.org

Cover Illustration: Altar of Domitius Ahenobarbus, Louvre Photo: Marie-Lan Nguyen (2007) Public domain http://commons.wikimedia.org/wiki/File%3AAltar_Domitius_Ahenobarbus_Louvre_n2.jpg

Figure 0.1 CC BY StuSeeger (2008) http://www.flickr.com/photos/56274840@N02/galleries/72157633089348310

Figure 0.2 CC BY Amri HMS (2007) http://www.flickr.com/photos/kulimpapat147/3541489234/

Figure 0.3 CC BY Paul Downey (2007) http://www.flickr.com/photos/psd/8419933758/

Figure 2.1 CC BY SA Steven Isaacson (2006) http://www.flickr.com/photos/spi/190603061/

Figure 2.2 Photographer unknown (1963). Repository: American Jewish Historical Society. Parent Collection: American Jewish Congress records, undated, 1916-2006 (I-77) http://digital.cjh.
org/R/?func=dbin-jump-full&object_id=1432347 No known copyright restrictions.

Figure 2.3 CC BY Alex Wellerstein. http://www.flickr.com/photos/restricteddata/6322465279/

in/photostream/

Figure 3.1 CC BY John Canning (2013)

Figure 4.4 CC BY Andrew Loomis (1938) Collections Canada.

http://collectionscanada.gc.ca/pam_archives/index.php?fuseaction=genitem.displayItem&lang=eng&rec_nbr=2842836

Figure 5.1 CC BY WattAgNet.com (2009) www.WATTAgNet.com. Image found at

http://www.245246flickr.com/photos/wattagnet/6721769653/in/photostream/

Figure 6.1 CC BY Charles D P Miller (2007) http://www.flickr.com/photos/cdpm/3644503386/

Figure 9.1 © R Foundation (2013). Used under the terms outlined at http://www.r-project.org/screenshots/screenshots.html

Figure 11.4 User: Wujaszek. This image is in the public domain because its copyright has expired
in the European Union and those countries with a copyright term of life of the author plus 70 years.
http://en.wikipedia.org/wiki/File:William_Sealy_Gosset.jpg

Figure 13.2 This image is in the public domain because its copyright has expired in the European Union and those countries with a copyright term of life of the author plus 70 years.
http://en.wikipedia.org/wiki/File:Karl_Pearson_2.jpg

Figure 15.1 CC BY SA Markus Hagenlocher (2006) http://en.wikipedia.org/wiki/File:Futterr% C3%BCbe.jpg

Figure 15.2 CC BY SA Brian Sims (2011) http://en.wikipedia.org/wiki/File:Occupy_Oakland_ 99_Percent_signs.jpg

Figure 16.1 No copyright. This image is in the public domain due to its age. Author died in 1858, material is public domain. http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.
jpg

Figure 16.2 CC BY John Canning (2013)

Figure 18.1 CC BY Smabs Sputzer http://www.flickr.com/photos/10413717@N08/2397168892/

Figure 18.2 User: Bletchley This image is in the public domain because its copyright has expired in the European Union and those countries with a copyright term of life of the author plus 70 years. http://en.wikipedia.org/wiki/File:R._A._Fischer.jpg

Figure 20.17 cc BY SA User:BenFrantzDale (2005) http://en.wikipedia.org/wiki/File:Chartjunk-example.svg

Figure 20.18 cc BY User:sacks08 (2010) http://www.flickr.com/photos/29071316@N06/4789882070/

Figure 20.19 cc BY SA Doug Kerr (2007) http://www.flickr.com/photos/dougtone/2720153713/

Figure 20.20 Coin image: cc BY Ivan Walsh (2009) http://www.flickr.com/photos/ivanwalsh/3782836923/

Graph: cc BY John Canning

Appendix cc BY Henry Lawford (2008) https://www.flickr.com/photos/herry/2988080770

Appendix cc BY Michiel2005 (2011) https://www.flickr.com/photos/govert1970/6432136985

Chapter 0: The mathematics behind statistics

"Quantification, even of a comparatively simple kind, arouses fear among many students, especially those who come to history with a background in arts subjects. Even those who approach history from a social-science background —students who also study sociology or politics —are easily disheartened when confronted with historical numbers." [1]

'Fear' is one of the words most associated with the study of statistics. Students opt to study arts and social science subjects believing that they have left their studies of mathematics firmly in the past then along comes a compulsory statistics course at university. If you are someone who feels fear and trepidation when statistics is mentioned it is comforting to know that the mathematics you have already studied will take you some distance in the study of statistics. The only tool you need to do the exercises in this book is a basic pocket calculator to add, subtract, multiply and divide. No special hardware or software is required. A second comfort for many students is the realisation that statistics involves the making of personal, subject judgements. Statistics is as much art as science. If you haven't studied mathematics for a while this chapter may help you revise things you are likely to have studied before. Don't worry if you can't take it all in at the first try. You might like to revisit this chapter from time to time.

Figure 1: A basic calculator with the functions +,−,× and ÷ is sufficient for all the exercises in this book. Suitable smartphone apps are widely available.

Numbers and types of data

Statistics is about numbers of course. The Oxford English Dictionary defines 'statistics' as "The systematic collection and arrangement of numerical facts or data of any kind." Numbers can be used in all sorts of different ways. This section is going to go through a few terms.

Interval data

A man who is 179 cm tall is taller than a man who is 178cm tall and shorter than a man who is 180cm tall. 19 degrees centigrade is warmer than 18 degrees centigrade and cooler than 20 degrees centigrade. The person who gets 19 out of 20 on their exam does better than the person who gets 18 out of 20 and less well than a person who gets 20 out of 20. These are examples of interval data. Unlike in the case of ordinal data (see below), interval data is measured along a scale where the differences between the points are meaningful. The difference between 170cm and 180cm is the same as the difference between 160cm and 170cm.

Ordinal (or ranking) data

First, second, third, fourth etc. are ordinal numbers. The person who wins a race comes first and the next person to finish comes second. The person who gets the highest exam result comes first and the person with the next highest second, the next highest third and so on. If I tell you that Jane got the highest score in the class, Peter got the next highest and Bill got the next highest you will know that Jane came first, Peter second and Bill third. However there is no information here about whether Jane got 1 more mark than Peter, 10 more marks than Peter or half a mark more than Peter. Ordinal data is not always in the form of numbers. A King-size bed is larger than a Queen-size bed which is larger than a double bed. The gap between 10th place and 20th place is not necessarily the same as the gap between 20th place and 30th place.

Nominal data

Nominal data is basically a name for a category of data. If I ask you whether you prefer cricket or football, cricket and football are categories. Cricket and football are not ordinal. Cricket does not come after football or before football. They are also not interval: cricket is not more than football or less than football. Similarly there is not a halfway point between cricket and football. Numbers are sometimes used as

Figure 2: The numbers these footballers wear on their shirts are nominal- they are simply to identify the players.

nominal data. Consider sport where the player wear numbers on their shirts

Continuous data and discrete data

Data can also be described in terms of being continuous or discrete. Continuous data can take any numerical value. For example a person can be 180cm tall or 180.1 cm tall, 180.2cm tall or 180.125987cm tall. In contrast discrete data can only take certain defined values. A village can have a population of 500 people, a population of 501 people or a population of 1000 people; it cannot have a population of 500.5 people, 501.092345 people or 999.23 people. Nominal data can also be described as discrete data.

Calculating percentages

A percentage (%) is a number as a fraction of 100. For example, if we have a spelling test of 100 words and get 100 right we say that we got 100%. Similarly if we got 50 out of 50 we would also say we got 100%. 6 out of 6 would be 100% as well. An easy way to calculate a percentage is to divide the number of spellings we got right by the number of spellings there were altogether, then multiply by 100.

Basic arithmetic

For example:

What percentage is 25 out of 50?

25÷50=0.5

0.5×100=50

So 25 out of 50 is 50%

What percentage is 5 out of 20?

5÷20=0.25

0.25×100=25

So 5 out of 20 is 25% Instead of writing on two lines we can write the formula on one line so

$$\frac{5}{20}\times 100 = 25$$

Positive and negative numbers

If your bank account is overdrawn then your balance will be a negative number, that is a number less than zero. Similarly if it is cold outside and the temperature is below zero degrees Centigrade then the temperature will be a negative number. Negative numbers are common in statistics and are often added, subtracted, multiplied or divided. The results of arithmetic with negative numbers are not always intuitive. Although a calculator will do the job for you it is useful to know (or estimate) whether the final answer will be a positive number or a negative number in order to check that you have not made a mistake.

Addition

Two negative numbers added together always results in a negative number (the brackets are not essential, but are used here for clarity):

$(-2) + (-3) = (-5)$

$(-3) + (-5) = (-8)$

If I am overdrawn at the bank by £ 100 and you are overdrawn by £50 then together we are overdrawn by £150 or you could say between us we have -£150. A positive number added to a negative number can be positive or negative. If the positive number is bigger than the negative number the result will positive. $5 + (-4) = 1$

If the negative number is bigger than the positive number then the result will be negative.

$4 + (-5) = (-1)$

Subtraction

A positive number subtracted from a positive number can be either a positive number or a negative number. $3 - 2 = 1$ $2 - 3 = (-1)$ Subtracting a positive number from a negative number results in a negative number. $(-3) - 1 = (-4)$

Multiplication

A positive number multiplied by a negative number always results in a negative number: [2]

$3 \times (-3) = (-9)$

A negative number multiplied by a negative number is always positive. Many statistical tests involve squaring negative numbers (that is multiplying a minus number by itself). $(-3) \times (-3) = 9$

Division

The rules for division are the same as for multiplication: A positive number divided by a negative number always results in a negative number

$4 \div (-2) = (-2)$

A negative number divided by a negative number is always positive.

$(-8) \div (-4) = (-2)$

Algebra

Algebra as used in this book is simply the use of letters to represent unknown numbers. For example, if we read that

$2 + a = 5$

what number is a? 2 plus something equals 5? As 2 + 3 = 5, then a = 3. We can also use these letters to show how to do an equation: We can write the formula for calculating a percentage as follows

$$\frac{a}{b} \times 100 = c$$

Using our spelling test example: *a* is the number of spellings we got right. *b* is the number of spellings there were altogether and c is the percentage of spellings we got right. In order to solve our formula we need to replace the letters with the numbers that we have.

Squares and Square roots

It is common in statistics to need to calculate squares and square roots.

Squares

To square a number, we simply multiply it by itself. For example 2 squared is the same as 2 × 2. This is usually written as 2^2. So $2^2 = 4$. 3 squared is the same as 3 × 3 or 3^2 so $3^2 = 9$.

Square root

A square root (written √) is the opposite of squaring a number so 9√ = 3 and 4√ = 2. As 4 and 9 are square numbers it is easy to find the square root. You will usually need to use a computer or calculator to work out the square root of a number.

This is not very intuitive. Fuller explanation can be found online, e.g.
https://www.youtube.com/watch?v=yrwI1OH9fw
Several possible ways of explaining this can be found at
http://mathforum.org/dr.math/faq/faq.negxneg.html

Rounding

Situations will arise when we get answers with a lot of decimal places. For example, according my calculator

2√ = 1.414213562373095

We may wish to describe this number more simply so that we can handle it better. For example we could round to the nearest whole number which would be 1, as 1.4 is nearer 1 than 2. Rounding to one decimal place would be 1.4 as 1.41 is nearer 1.4 than 1.5. Rounding to two decimal places would be 4.14 as 1.414 is nearer 4.14 than 1.45. To round the number 7.8954 to the nearest whole number would be 8 as 7.8 is closer to 8 than 7. Similarly if we were to round it to one decimal place then 7.89 would become 7.9 as 7.89 is nearer 7.9 than 7.8. Whilst rounding numbers up and down is done frequently in this book, it is not without its disadvantages. When number are rounded up or down and used in a number of calculations then 'rounding errors accumulate' leading to answers which are less accurate than would be the case if numbers were left alone. Suppose we did a survey of 200 people on their opinion of a particular issue, and the answers came back as follows:

Agree 49%.

Disagree 25.5%

Don't know 25.5%

49% + 25.5% + 25.5% = 100%

But if we round up to the nearest whole numbers we get

49% + 26% + 26% = 101%

Greek letters

Statistics makes a lot of use of Greek letters. These will be explained as needed. Σ is the most common Greek letter you will encounter. It is used to signify 'sum of'.

Brackets

In equations with different signs (+, -, ×,÷) multiplication and division always take precedent over addition and subtraction. In other words we DO NOT read equations left to right.

$3 \times 4 + 2 = 14$

Gives the same answer as

$2 + 3 \times 4 = 14$

Note that the 3×4 is always done first, even if it does not appear first in the equation. However, if brackets are used then the sum in brackets must be calculated first

$(2+3) \times 4 = 20$

Because 2+3=5. We then multiply this by 4 and get 20. To give another example:

$9 - 4 \times 2 = 1$

But

$(9-4) \times 2 = 10$

One way of remembering the order is BODMAS:

B= Brackets
O= Orders, powers or square roots.
D= Divisions
M= Multiplication
A= Addition
S= Subtraction

Thanks to Hannah Burd for altering me to this.

Greater than and less than

The signs greater than (>) and less than (<) are used to shown inequalities.

3 is greater than 2

$3>2$

3 is greater than a

$3>a$

3 is less than a

$3<a$

a is greater than b

$a>b$

a is less than b $a<b$

p is less than 0.05

$p<0.05$

p is greater than 0.01

$p>0.01$

≤ is less than or equal to and ≥ greater than or equal to:
3 is greater than or equal to a

$$3\geq a$$

3 is less than or equal to a

$$3\leq a$$

All the above calculations can be put together in one equation.

Putting it all together

Example 1:

$\sqrt{5+4}$

Is the same as

$\sqrt{9}$

Because 5+4=9.

Example 2:

$\sqrt{(4+5)\times 3}$

Is the same as

$\sqrt{27}$

Because
$(4+5)\times 3=27$

Example 3: As before everything in brackets is calculated first. If there are no brackets then multiplication and division ALWAYS take precedence over addition and subtraction

$\sqrt{4+5\times 3}$

Is the same as

$\sqrt{19}$

Because $4+5\times 3=19$

Example 4:

$\sqrt{\frac{6}{6}}$

Is the same as

$\sqrt{1}$

Because

$6\div 6=1$

Example 5:

$(3+3)^2$ Is the same as

6^2
Because
$3+3=6$

Example 6:

The same example as 5 but without the brackets. Squaring a number is a form of multiplication so it takes precedence.

$3+3^3$ Is the same as

$3+9$
Because
$3^2=9$

Coefficients and variables

A variable is a value which is able to vary. In the expression $3x$ the variable is x. If $x = 1$ then $3x = 3$. If $x = 2$ then $3x = 6$, if $x = 3$ then $3x = 9$ and so on. In this same expression 3 is a constant. Whatever value x has the 3 never changes. In this case the 3 comes together with x to make $3x$ or $3 \times x$. When a constant is used in an expression to be multiplied by a variable (in this case x) it becomes a special kind of constant called a coefficient. This book contains tests such as the Pearson Product Moment Correlation Coefficient and the Spearman's Rank Correlation Coefficient. These tests are described as coefficients because they use formulas which include coefficients.

Exercises

1. Calculate the following test scores as a percentage
    1. 6 out of 6
    2. 3 out of 9
    3. 85 out of 100
    4. 40 out of 50
2. Calculate:
    1. $(-10)+3$
    2. $(-3)\times(-3)$
    3. $4\times(-2)$
    4. $(-3)\div(-3)$
3. Find the value of x
    1. $6+x=10$
    2. $7\times x=7$
    3. $x-1=2$
    4. $x\div 3=3$

    5. $6x = 24$

4. Calculate:

    1. 3^2
    2. 4^3
    3. $(-3)^2$
    4. $\sqrt{9}$
    5. $\sqrt{1}$
    6. $\sqrt{4}$
    7. $\sqrt{20}$

5. Calculate:

    1. $4 \times 4 + 2$
    2. $2 + 2 \times 4$
    3. $(2 + 2) \times 4$
    4. $3 \times 3 \times 3 - 2$

6. Calculate:

    1 . $\sqrt{\frac{6}{3}}$

    2. $\sqrt{\frac{12-3}{3}}$

    3. $(3 + 4)\times 2$

    4. 2×10^2

[1]Mark Freeman (2004) Teaching Quantification in History (Glasgow: HEA Subject Centre for History, Classics and Archaeology)

CHAPTER 1: INTRODUCTION

1 Using statistics in the humanities

Were the people who emigrated from England to North America in the nineteenth century poor people fleeing poverty or prosperous skilled labourers seeking opportunities in a new country? Has the number of the people speaking Welsh in Wales increased in the past twenty years? When Jane Austin died in 1817 she left assets worth around £800. Was £800 a lot of money in 1817? [1]

Following the 1948 US presidential election why did the Chicago Tribune feel confident enough to run the headline "Dewey defeats Truman" before all the votes were counted only for it to be become evident that Truman had actually won? [2]

The average life expectancy in the early 1800s was about 40 years of age. Does this mean that there were no old people? The 2011 UK census reveals that 59% of people in the UK identify themselves as 'Christian',[3] but only 15% attend a Christian place of worship at least once a month. [4] What does this tell us about the relationship between identity, belief and practice?

2 Types of Quantitative Data

The answers to all of these questions rely on some sort of understanding of numbers and interpreting them. This book is a beginner's guide to statistics which uses examples from the humanities subjects. The case studies used are from archaeology, history, languages, linguistics, religious studies and area studies. I do not assume any prior knowledge of statistics other than that you know how to add, subtract, multiply and divide with the aid of a pocket calculator.

Some sources such as surveys and censuses are created with the express purpose that they will generate numerical data which will then be analysed. Whether it is a company researching the market for a new project, the government surveying the whole population in order to plan services in the longer term or educationalists undertaking questionnaires of primary school teachers about their opinions on learning a second language, these sources are intrinsically numerical. In a democratic society the principle of elections is that those with most votes are chosen. Again, the data is numerical by its very nature. Other sources are not designed to be quantitative, but quantitative data can be derived from them with ease. From Parish Registers it is possible to derive numerical data about length of life, age at marriage, family size, infant mortality and maternal death in childhood. Changes in these factors can be monitored over decades and centuries and different geographical areas can be compared. Financial data in the form of prices, taxation and public and private spending can provide insight into historical and contemporary conditions. These can be monitored over time or countries might be compared. We can get numerical information on the occupation, age, nationality and health of emigrants to the USA from immigration data and from ships' passenger lists. Again, trends can be identified and monitored over a period of time. Quantitative data can also be derived from non-quantitative sources such as newspapers. How many incidents of arson were there in Sussex in the 1830s? Although not every incident would necessarily be reported in the newspaper it is possible to count up the number of stories or column inches given to certain topics. It is possible to derive quantitative data from any source. The Oxford English Dictionary annually announces new words which have been invented or have come into more common use in the previous year. [5] Recent additions have included 'credit crunch', 'staycation' and 'jeggings'. OED staff monitor written language use to identify new trends. Researchers in linguistics might count incidents of an individual using certain phrases or metaphors or how many times they pause or say 'um', 'err' or 'ah'. These can be compared between individuals, sexes, languages or native and non-native speakers.

3 Where do statistics come from?

Statistics prove …", or do they? Some people think of statistics as a way of proving or disproving a particular argument or relationship. We might begin an argument by saying “Statistics prove …” or “Statistics show …”. In fact statistics are not morally, ethically and epistemologically neutral. There are inherent biases in the statistics we and others collect and why we collect them. These biases reflect the values both of those who collect data or statistics (including governments) and how we interpret them. We can only use statistics that are available to us whether we collected them ourselves or acquired them from other surveys or from other documents. We cannot make arguments on the basis of statistics we do not have. The UK census has taken place every ten years since 1801 (with the exception of 1941), but the questions have changed to reflect changes in society and changing views of what the Government needs to know about people. From 1951 to 1991 the census asked people if they had an inside toilet. In 1951 a lack of access to inside sanitation was seen as an important indicator of social deprivation. However, by 1991 there were very few houses without an inside toilet and the question was dropped. For the first time in 2011 the population was asked how well they could speak English. Our own reasons for being interested in a particular topic also impact on how we use statistics.

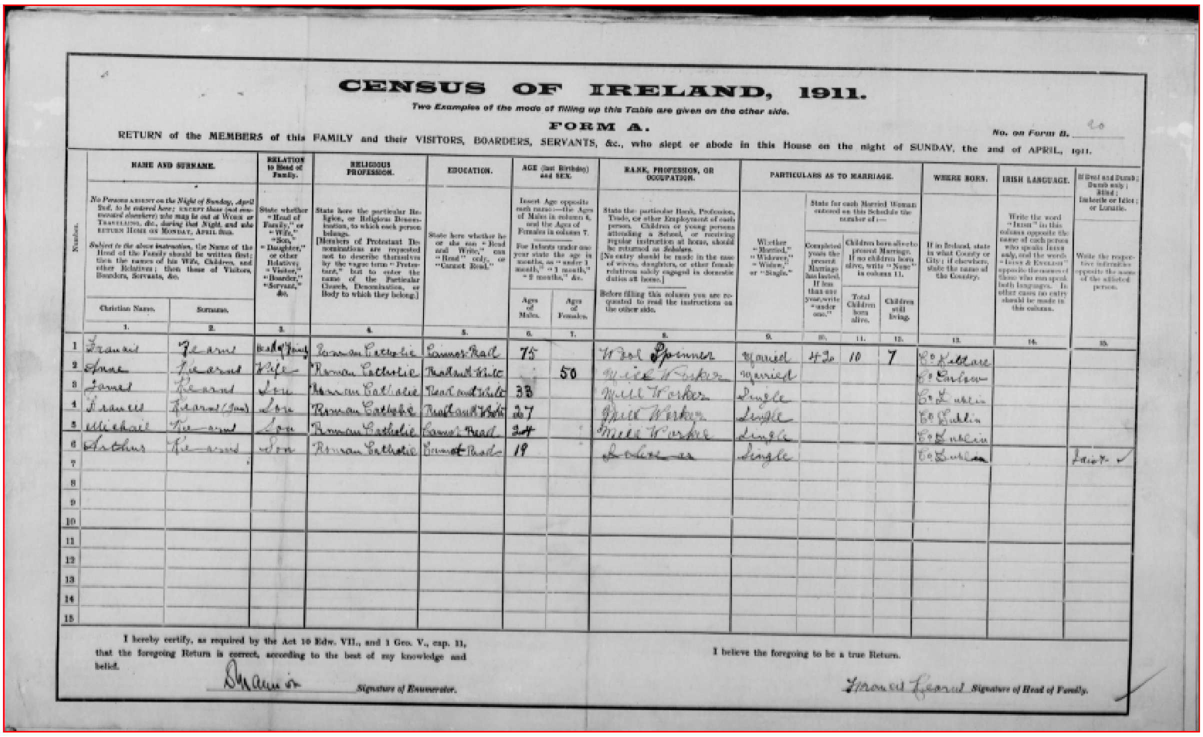
As researchers we all have our own values which are reflected in the things we are interested in and how we might use data relating to these topics. When we use the data produced by other people we are often examining them for a different purpose. The government of 1801 did not start collecting census data to make it easier for future generations to research their family history though many people use the census for this purpose. Taxation records were created for the purposes of collecting tax, not for producing maps of relative wealth in different parts of a city, though researchers have used these records to do just that. Today, schools count up how many pupils have free meals so that they know how many meals they need to cook and get reimbursement for. However, the proportion of children on free school meals is frequently used to measure social deprivation in any given school. A school will be considered ‘deprived’ if a high proportion of its pupils are eligible for free school meals. Schools do not ask parents to provide details of their income for the purpose of measuring deprivation. Statistics are also often used to support legal, moral or ethical arguments. Opinion polls are used by advocacy groups to demonstrate that the public is supportive, not supportive or doesn’t care about alcohol regulation, abortion time limits, gay marriage or euthanasia. Such polls can be useful for governments unsure whether to proceed with a particular piece of legislation or policy change, but the amount of support for an opinion does not prove that one side is wrong and one side is right and researchers should use this sort of data with caution. In chapter 19 we will be exploring survey design and some of the implications this can have on the conclusions we come to about attitudes and beliefs.

4 References

Table 1: 1911 Census return for the Kearns family of Dublin. Note that the 19-year old son Arthur is described as an ‘idiot’. Other people were listed as imbeciles or lunatics. These categories were considered scientific at the time.

Surname	Forename	Age	Sex	Relation to head	Religion	Specified Illnesses
Kearns	Francis	75	Male	Head of Family	Roman Catholic	0
Kearns	Anne	50	Female	Wife	Roman Catholic	0
Kearns	James	33	Male	Son	Roman Catholic	0
Kearns	Francis	27	Male	Son	Roman Catholic	0
Kearns	Michael	24	Male	Son	Roman Catholic	0
Kearns	Arthur	19	Male	Son	Roman Catholic	Idiot

Figure 1 Full original census return for the Kearns family. Unlike the UK Census, the 1911 Census of Ireland is available free online.



CENSUS OF IRELAND, 1911.

Two Examples of the mode of filling up this Table are given on the other side.

FORM A.

No. on Form B. 20

RETURN of the MEMBERS of this FAMILY and their VISITORS, BOARDERS, SERVANTS, &c., who slept or abode in this House on the night of SUNDAY, the 2nd of APRIL, 1911.

Number.	NAME AND SURNAME. No Persons ABSENT on the Night of Sunday, April 2nd, to be entered here; EXCEPT those (not enumerated elsewhere) who may be out at WORK or TRAVELLING, &c., during that Night, and who RETURN HOME ON MONDAY, APRIL 3RD. Subject to the above instruction, the Name of the Head of the Family should be written first; then the names of his Wife, Children, and other Relatives; then those of Visitors, Boarders, Servants, &c.		RELATION to Head of Family. State whether "Head of Family," or "Wife," "Son," "Daughter," or other Relative; "Visitor," "Boarder," "Servant," &c.	RELIGIOUS PROFESSION. State here the particular Religion, or Religious Denomination, to which each person belongs. [Members of Protestant Denominations are requested not to describe themselves by the vague term "Protestant," but to enter the name of the Particular Church, Denomination, or Body to which they belong.]	EDUCATION. State here whether he or she can "Read and Write," can "Read" only, or "Cannot Read."	AGE (last Birthday) and SEX. Insert Age opposite each name:—the Ages of Males in column 6, and the Ages of Females in column 7. For Infants under one year state the age in months, as "under 1 month," "1 month," "2 months," &c.		RANK, PROFESSION, OR OCCUPATION. State the particular Rank, Profession, Trade, or other Employment of each person. Children or young persons attending a School, or receiving regular instruction at home, should be returned as Scholars. [No entry should be made in the case of wives, daughters, or other female relatives solely engaged in domestic duties at home.] Before filling this column you are requested to read the instructions on the other side.	PARTICULARS AS TO MARRIAGE. Whether "Married," "Widower," "Widow," or "Single."	State for each Married Woman entered on this Schedule the number of:—			WHERE BORN. If in Ireland, state in what County or City; if elsewhere, state the name of the Country.	IRISH LANGUAGE. Write the word "IRISH" in this column opposite the name of each person who speaks IRISH only, and the words "IRISH & ENGLISH" opposite the names of those who can speak both languages. In other cases no entry should be made in this column.	If Deaf and Dumb; Dumb only; Blind; Imbecile or Idiot; or Lunatic. Write the respective infirmities opposite the name of the afflicted person.
	Christian Name.	Surname.				Ages of Males.	Ages of Females.			Completed years the present Marriage has lasted. If less than one year, write "under one."	Children born alive to present Marriage. If no children born alive, write "None" in column 11. Total Children born alive.	Children still living.			
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1	Francis	Kearns	Head of Family	Roman Catholic	Cannot Read	75		Wool Spinner	Married	42	10	7	Co Kildare		
2	Anne	Kearns	Wife	Roman Catholic	Read and Write		50	Mill Worker	Married				Co Carlow		
3	James	Kearns	Son	Roman Catholic	Read and Write	33		Mill Worker	Single				Co Dublin		
4	Francis	Kearns (Jun)	Son	Roman Catholic	Read and Write	27		Mill Worker	Single				Co Dublin		
5	Michael	Kearns	Son	Roman Catholic	Cannot Read	24		Mill Worker	Single				Co Dublin		
6	Arthur	Kearns	Son	Roman Catholic	Cannot Read	19		[illegible]	Single				Co Dublin		Idiot ✓
7															
8															
9															
10															
11															
12															
13															
14															
15															

I hereby certify, as required by the Act 10 Edw. VII., and 1 Geo. V., cap. 11, that the foregoing Return is correct, according to the best of my knowledge and belief.

D Mannion Signature of Enumerator.

I believe the foregoing to be a true Return.

Francis Kearns Signature of Head of Family.

1].National Archives website. Famous wills: Jane Austin http://www.nationalarchives.gov.uk/museum/item.asp?itemid=33 http://www.ons.gov.uk/ons/rel/census/2011-census/key-statistics-for-local-authorities-in-england-and-wales/rpt-religion.html

[2]Chicago Tribune November 3, 1948

[3]Office for National Statistics (2012) Religion in England and Wales

[4] J. Ashworth et al (2007) Churchgoing in the UK: A research report from Tearfund on church attendance in the UK(London: Tearfund)

[

CHAPTER 2: HOW MANY AND HOW BIG?

1 How many and how big?

Counting is one of the first skills we learn as children and is an important milestone in a child's development. Counting is simply the question "How many are there? How many people were killed in the Holocaust? How many people were transported from Africa to the Americas as slaves? How many people speak Basque? How many mosques are there in Birmingham? How many copies of Harry Potter novels have been sold? How many Norse burial sites are there in the Orkney Islands? How many people lived in Liverpool at the time of the 1851 census? An accident which kills fifty people is more likely to get more time on the news than a similar accident which only claims one life. The news that more people are speaking Basque than before or that more people are unemployed then were six months ago may lead to calls for changes in policy. The statistic that six million Jews were killed in the Holocaust invokes a sense of moral outrage.

2 Why count?

Some linguists have suggested that "humans possess an innate number sense. Counting is essentially the first stage of working with numbers. Whether consciously or not we use counting to make analytical and moral judgements, often by using non-statistical language as we talk about numbers. Think of words like up, down, common, uncommon, important, major and minor. Although these words do not relate directly to any specific scale they invoke judgements about numerical scale. We think of success and historical events in terms of numbers, even if these numbers are unknown. Would Martin Luther King have come to prominence if he was one of only a small number of people supporting the Montgomery Bus Boycott? Would the Paris riots of 1968 still be talked about if only a handful of people participated over the course of a couple of hours? Even if we do not use statistics in our work we use words which employ ideas of size and scale (see Table 1).

Counting helps us to identify trends which are taking place or took place in the past;

For Example:

1. We can see if numbers of Welsh speakers are going up or down or how they went up or down in the past.

2. We can see how the population of Liverpool is going up or down or how it went up or down in the past.

3. Counting can challenge conventional wisdom about social phenomenon such as numbers of nineteenth century brides who were pregnant at the time of their marriage.

4. We can use numbers to make a qualitative judgement about the importance, scale, severity or impact of an event. For example, two earthquakes can have the same magnitude, but if one takes place in a city and another in an uninhabited area, the former is likely to impact on more people than the latter.

Table 1 Everyday words with an idea of numbers

Smaller number	'	Bigger number
A few	A lot	All
Uncommon	Common	Universal
Impoverished		Wealthy
Few	Many	Most

3 Problems in counting

Superficially, counting is a straightforward skill. However counting can become difficult in a number of situations.

3.1 Missing, unavailable or non-existent data

1. Data that never existed We are only able to count data that exists. We don't know what proportion of households in the UK had an inside toilet in 2001 and 2011 because this data was not collected.

2. Data that is missing or destroyed. Even if data was collected it can be lost or destroyed, deliberately or accidentally. Documents can fall victim to fires, floods, rodents as well as wear and tear damage.

3. Data that does exist, but will not be available until a future date. The original UK census returns will not be made public for 100 years after the census data. Therefore we cannot use data which depends on access to original census returns of 1921 and later. Some police, prison and legal records are also restricted. UK Government Papers are released after 30 years, but some information is restricted for reasons of national security

4. Data is missing because its subjects were deliberately or accidentally excluded, through their own actions or those of others. Even though censuses are an attempt to collect data on an entire population they are always an undercount. People may try to avoid censuses if they are in the country illegally or are concerned that the census is being used for taxation or military purposes. Many people oppose the census on the grounds that it invades their privacy

5. Data can be inaccurate. It may have been wrongly entered into a computer, mistakes may have been made in making calculations or someone might have lied when reporting data.

3.2 Non-comparable data

When we compare data taken at two time periods it may appear that we are comparing like with like. For example the question, 'how many people speak Welsh?'has been asked regularly over the years, but in different ways. There are lots of ways to asking questions to get a sense of someone's knowledge of the Welsh Language but they may lead to different answers, (See Table 2).

Comparisons across countries need to be undertaken with some degree of caution. The calculations and methods used for measuring inflation and unemployment, for example, sometimes differ between countries.

3.3 Changing geographical boundaries over time

1. Changes of borders within a country. What was the population of Oxfordshire in 1881? We need to clarify whether we are talking about Oxfordshire as it is now or Oxfordshire was it in 1881. The town of Abington was in Berkshire until 1974 when it was moved into Oxfordshire. There are numerous cases like this in the UK so they need to be checked. The exact administrative boundaries of towns and cities have also changed over time.

2. Movement of national boundaries can be a more difficult subject as useful data which was collected in the past might no longer be and vice versa. For example, the Alsace region of France has been part of both France and Germany over the past 400 years and was subject to the data collection regimes of both countries during different parts of the nineteenth and twentieth centuries.

3.4 Double (and triple) counting:

We will talk more about classifying data in the next section, but when we put data into different categories we can an end up counting the same data two or three times. Suppose you were counting the number of incidents of machine vandalism and arson which took place during the Swing Riots in Berkshire? What would you do if you came across a single incident in which protesters had vandalised a machine, then set fire to the barn in which it was housed? Is this one incident or two? Would you count it once or twice? What is being counted?

3.5 Differing definitions:

In October 2012 1.58 million people were registered as unemployed in the UK, a rate of 7.8 %. But who counts as unemployed? Possible answers include:

- Number of people who don't have jobs.
- Number of people claiming job seekers' allowance
- Number of people not working, but looking for a job.
- Number of people of working age who do not have a job.
- Number of people who could work, but are not working.
- Adults who do not have a job and are not studying

Table 2: Welsh language questions of the 1981, 1991 and 2001 censuses. [2]

1981 census For all persons aged 3 or over (born before 6 April 1978)	**1991 census For all persons aged 3 or over (born before 22 April 1988) "**	**2001 census Can you understand, speak, read or write Welsh?"**
"Does the person speak Welsh?	Speaks Welsh	Understand spoken Welsh "
If the person speaks Welsh, does he or she also:	Reads Welsh	Speak Welsh
"Speak English?	Writes Welsh	Read Welsh
Read Welsh?	Does not speak, read or write Welsh "	Write Welsh
"Write Welsh?		None of the Above

In order to make meaningful comparisons, definitions need to be agreed. The International Labour Organisation uses the following definition of unemployment:

> *"An unemployed person is defined by Eurostat, according to the guidelines of the International Labour Organization, as someone aged 15 to 74 without work during the reference week who is available to start work within the next two weeks and who has actively sought employment at some time during the last four weeks. The unemployment rate is the number of people unemployed as a percentage of the labour force."* [3]

4 Summary

The questions `How many?' and `How large?' are the starting point of any statistical analysis. Counting is not always as straight forward. We have to deal with changing questions, changing geographical boundaries, missing or inaccurate data, double counting and different definitions.

5 Exercises

Examine the Welsh language questions from the 1981, 1991 and 2001 UK censuses. How do they differ? How could they lead to different answers?

6 References

[1]C. Holden (2012), Life without numbers in the Amazon, Science 305 p.109e: An aggregate analysis, Area, 36(2), 187-201

[2]G. Higgs, C. Williams, and Dorling, D. (2004)Use of the Census of Population to discern trends in the Welsh language

[3]Eurostat: http://epp.eurostat.ec.europa.eu/statisticsexplained/index.php/Unemploym

CHAPTER 3: SUMMARISING DATA

1 Introduction

Lists of numbers and tables of data are useful, but a few statistical measures can usefully summarise a whole data set. We will read in the newspaper that the average income in the UK is £26,500. [1] The average weekly wage of an agricultural worker in 1850 was 9 shillings 3 12 pence. [2] We also use the word `average' in a qualitative sense. We might say that a student is of average academic ability or that we live in an average-sized house. When we talk about an average we are summarising a larger set of data in one figure. So when we say the average income is £26,000 it acknowledges that some people earn more than £26,000 and other people earn less, but someone on middle income earns around £26,000. We often associate the term `average' with `normal'. A person on an average income is not rich and not poor. A student of average ability is neither the one of the highest performers, nor one of the lowest performers. This chapter will show you how to calculate mode, median and mean, upper and lower quartiles and will address some of the issues surrounding the use of the mean, median and mode.

2 Mean

There are actually several types of average, but the most familiar average used is the mean average. This is calculated by adding the observations together then dividing by the number of observations. The following is a list of the heights in centimetres of 10 soldiers who enrolled in the French army in 1790. By adding all the heights together, then dividing our answer by the number of soldiers we can find the mean height: So:

$$\frac{168+165+165+168+168+165+165+173+175+165}{10} = \frac{1677}{10} = 167.7$$

When you see the mean average reported academic papers you may see it written as $\bar{x}$ and spoken as `bar x' or `x bar'. If different averages are being compared you may see the different averages written as $\bar{y}$ or $\bar{z}$

We can present the sum above as an equation where x is one observation (in this case a soldier's height), $\bar{x}$ is the mean height of the soldiers (the mean of the x's) and n is the number of observations. As we have 10 observations we can write each of these as x_1 ,x_2, x_3 … etc. So our formula for calculating the mean is

$$\bar{x} = \frac{x_1, x_2, x_3, x_4 ...etc.x_n}{n}$$

$\bar{x}$ = Mean average.
n=Number of observations.
x_1 =Soldier 1's height
x_2 = Solider 2's height etc., up to x_{10} which is soldier 10' s height.

3 Median

The median average is simply the observation which comes in the middle. The following example comes from the register of burials in Accrington, Lancashire. Five people buried in succession died at the following ages:

8,20,32,17,82

All we need to do to find the median is to put the ages in order.

8,17,20,32,85

The median average is the one in the middle. In this case the median age at burial is 20 years of age. There are five observations of which two are below the median and two above. If we have an even number of observations we have a situation where there is no single middle number. In the example below we have six observations.

1,8,17,20,32,85

To find the median we must take the middle two values (17 and 20) and divide them by 2. This will give us our median.

$$\frac{17+20}{2}=18.5$$

The median of these six observations is 18.5. Notice that there are three observations which are less than 18.5 and three observations which are more than 18.5.

4 Mode

A third type of average is the mode. The mode is simply the value which occurs most frequently. Below we have added some more burial ages from Accrington to those we used in the example for the median.

8,20,32,17,82,0,0,22

In this example we can see that the most frequently occurring value is 0. Therefore the mode of this sample of burials is zero years of age.

5 Making sense of averages

An average summarises a set of data in one number. Each type of average has its own strength and weaknesses.

The mode is the least frequently used form of average. It only uses one number from the dataset. It is mostly used for describing nominal data (that is data with names or categories). For example if we did a questionnaire which asked people to name their religion and the most commonly occurring religion was Christian we would say that the mode or modal group was `Christian'. There is not a median or mean religion, sex, race or national identity.

The advantage of a mean average is that it takes account of all the observations. However, taking account of all values can be misleading. A few very high or very low values skew the data to give a misleading view of the data as a whole. This is very common in the case of income data where a small number of wealthy people drive the mean income up to a level which does not reflect anyone's income. For example consider the following five incomes

£18,000;£22,000;£20,000;£28,000;£100,000

The mean income is

$$\frac{18,000+22,000+20,000+28,000+100,000}{5}=37,600$$

As we can see the mean income is greater than four of these five incomes. The statement that the mean average income is £37,000 would be correct but it does not summarise the data very well. The median income, £20,000, is a much more realistic reflection of the income earned by four of these five people.

6 Case study: Life expectancy

A similar issue occurs when examining life expectancy. It is common to hear that the average person living in nineteenth century England had an average life span of around 30 years. This seems to suggest that most people were dead before the age of 40. Does this mean that there were no old people in the nineteenth century? Let's return to our example from Accrington in 1838. Table 1 records the age of death of 39 people buried that year.

The mode conveys that sad reality that the most common age to die was before the age of one. However, although it was the most common age at which to die it does not mean that most people died before the age of one. The mean age of death was 25.1 and the median 20 years. Both these averages indicate that people were able to live sufficiently long enough to have children themselves and deaths of people in their late teens and twenties were clearly common. All three averages fail to reflect the fact that people did live into their seventh, eighth and ninth decades. Not very many people lived this long, but it was not impossible. So there were old people living in nineteenth century England.

This pattern is also important in understanding the age profiles today in countries with low life expectancies. Societies with high levels of infant mortality reduce the mean, median and mode age of death far below that of societies where infant mortality is very low. A closer examination of the figures suggests that anyone who manages to live beyond the age of about four has a very good chance of reaching adulthood.

Life expectancy is particularly interesting in statistical terms. It is commonly expressed in terms of life expectancy at birth, but as you get older your life expectancy continues to increase.

As early as the seventeenth century an estimated 10% of the population were over 60. [3] Additionally, your life expectancy never stops increasing, so a 100 year-old's life expectancy is clearly over 100 years, much more than a five year-old or a 65 year-old.

To find the mean average we add together the 39 numbers then divide by 39. When we do this we find that the mean average age of death was 25.1 years old.

$$\frac{\begin{array}{c}8+20+32+17+82+0+0+22+47+0+38+25+57+1+0+21+39\\+15+54+48+1+3+78+1+29+22+63+41+73+1+2\\+35+80+13+0+12+0+0+0\end{array}}{39}=25.1years$$

To find the mode average we look for the value which occurs the most frequently. In this case the most common age at which to die was 0. 8 out of the 39 people died before they were one year old.

To find the median average we put all the ages in order. The median is the value with the same number of values

Table 1 Age of death in Accrington, 1838

Age	Number	Age	Number	Age	Number
0	8	20	1	41	1
1	4	21	1	47	1
2	1	22	2	48	1
3	1	25	1	54	1
8	1	29	1	57	1
12	1	32	1	63	1
13	1	35	1	73	1
15	1	38	1	78	1
17	1	39	1	80	1

before it an after it. In this case our median is 20 years. 19 people died younger than this 20 year old and 19 people were older when they died.

To summarise, the mean average age of death was 25.1 years old. The mode average age of death was 0 years. The median average age of death was 20 years. You will notice that these three averages give us very different answers (see Figure 1). But, which is the most useful average?

7 The five figure summary

The above section has demonstrated some of the hazards of relying on the average value (whether mean, median or mode) alone. The five figure summary consists of

1. the median,

2. upper quartile

3. lower quartile

4. minimum observation

5. maximum observation.

We will calculate the five figure summary for a sample of farms in Chile. [4] The data below shows the size of 24 farms in Chile (in hectares).

It can be seen that farm sizes in Chile vary considerably.

0.5, 3.5, 15.1, 508.3, 0.7, 3.5, 13.1, 1701.7, 0.2, 7.1, 39, 10.1, 1.2, 8, 57, 19.5, 1.5, 9.9, 198.2, 4.9, 2.4, 6.2, 276.4, 3

8 Calculating the median

There are 24 farms altogether. We find which place the median is in by There are 24 farms altogether. We find which place the median is in by (

$$\frac{24+1}{2}=12.5$$

First we sort the farms out into order of size

510.2,0.5,0.7,1.2,1.5,2.4,3.0,3.5,3.5,4.9,6.2,7.1,8.0,9.9,10.1,13.1,15.1,19.5,39.0,57.0,198.2,276.4,508.5,1701.4

So if we were to place all the farms in order of size the median would be the farm in 12th and 13th place. As

Figure 1 Different types of average illustrated: Life expectancy in 1830s Accrington

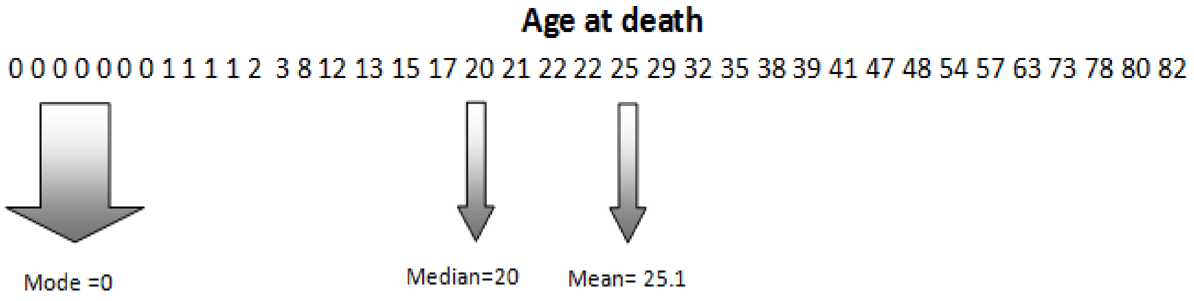


Figure 2 Boxplot: Farms in Chile

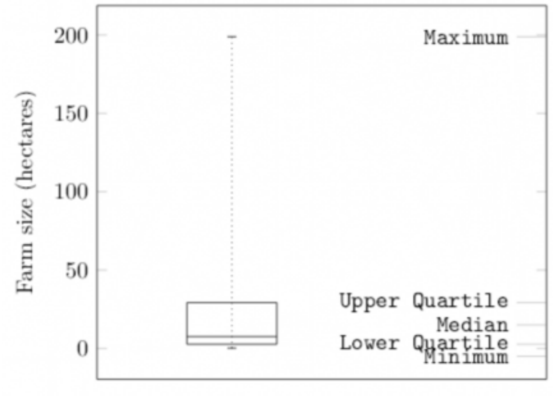
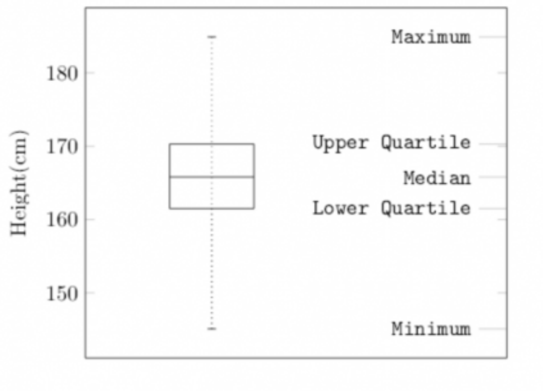


Figure 3Boxplot: Heights of Bavarian



there are an even number of farms and the median lies between two places we need to find:

$$\frac{7.1+8}{2}=7.55$$

It is instantly clear at this point that the median size farm at 7.55 hectares is considerably smaller than the mean at 120.5 hectares. This would seem to indicate that there are a high number of smaller farms and a small number of bigger farms. We can learn a bit more about the distribution of farm sizes by calculating the quartiles.

9 Upper and lower quartiles

The lower quartile lies halfway between the median and the lowest value. The upper quartile lies halfway between the median and the highest value. To find the lower quartile we need to find the median of all the values below the median of the whole dataset. There are 12 values below the median

0.2,0.5,0.7,1.2,1.5,2.4,3.0,3.5,3.5,4.9,6.2,7.1

To find where the median of these 12 values is located (12+1)2=6.5 The median lies between 6th and 7th place 0.2,0.5,0.7,1.2,1.5,2.4,3.0,3.5,3.5,4.9,6.2,7.1 (2.4+3.0)2=2.7 So our lower quartile is 2.7.

To find the upper quartile we need to find the median of all the values above the median. There are 12 values above the median: To find where the median of these 12 values is located

$$\frac{12+1}{2}=6.5$$

The median lies between 6th and 7th place

$$\frac{19.5+39.0}{2}=29.25$$

So our upper quartile is 29.25 So far we have three numbers for our five number summary:

Our median: 7.55

Our lower quartile: 2.7

Our upper quartile: 29.25

We need two more numbers for our summary, but these are easy to find. We need the smallest value and the largest value. The smallest number is 0.2 and the highest is 1701.4.

Therefore our five numbers are:

1. Minimum value: 0.2
2. Lower quartile: 2.7
3. Median 7.55
4. Upper quartile 29.25
5. Maximum value 1701.4

10 Presenting the five figure summary

This five figure summary gives us a better idea about the distribution of farm sizes. We can see that half of the farms are less than 7.55 acres, but a quarter are less than 2.7 hectares. This indicates that the vast majority of farms are small. The largest farm in the sample is huge in comparison, over 255 times the median and 58 times the size of the farm which forms the upper quartile. We can represent this graphically in the form of a box-plot . The ends of the box mark the two quartiles and the line inside the box is the median. The lines coming out of the box are known as whiskers. The whiskers go from the top of the box to the maximum value and from the bottom of the box to the minimum

Table 2: Number of booksellers registered for Value Added Tax(VAT) Turnover Size (thousands)

Turnover Size (thousands)	Number of booksellers
0-49	160
50-99	215
100-249	310
250-499	180
500-999	85
1000-4999	35
5000+	15
TOTAL	1000

Table 3 Number of booksellers registered for Value Added Tax(VAT) with midpoints calculated

Turnover Size (thousands)	Number of booksellers	Lower boundary	Upper boundary	Number of booksellers	Mid point
0-49	160	0	49	160	24.5
50-99	215	50	99	215	74.5
100-249	310	100	249	310	174.5
250-499	180	250	499	180	374.5
500-999	85	500	999	85	749.5
1000-4999	35	1000	4999	35	2999.5
5000+	15	5000	10000	15	7500
TOTAL	1000				

value. We can see from the boxplot (see Figure 2) the median and lower quartiles lines are very near the bottom of the box plot. We say the distribution is skewed.(Instructions on how to draw a boxplot can be found in Chapter 20. We can also get a sense of how skewed the data is by calculating the mean

(0.7+0.2+1.2+1.5+2.4+4.9+3.0+3.5+3.5+7.1+8.0+9.9+6.2+10.1+19.5+15.1+13.1+39+57+198.2+276.4+508.3+1707.7) ÷ 24=120

The mean farm size is 120 hectares, yet only four of the 24 farms are above the mean. A small number of large farms are so large that they are skewing the mean towards it. The boxplot in Figure 3 represents the heights of 20,000 Bavarian conscript soldiers in the early nineteenth century. It can be seen clearly that the soldier's heights are much more evenly distributed than the Chilean farms. The median height is 168cm and the mean height is 167.3 cm. If a soldier of mean height stood next to a soldier of median height it is unlikely that you would notice the difference. Although a small number of tall soldiers skew the mean slightly it is clear that most of soldiers have heights close to the mean. When most of the values are equally distributed around the mean, we might say that the values are normally distributed. We will be discussing the normal distribution in the next chapter. The five figure summary (minimum, lower quartile, median, upper quartile and maximum) is a useful way of describing our data which takes all the observations into account. Large

Table 4 Number of booksellers registered for Value Added Tax(VAT) with midpoints calculated and group turnover

Lower boundary	Upper boundary	Number of booksellers	Mid point	Group turnover
0	49	160	24.5	3920
50	99	215	74.5	16017.5
100	249	310	174.5	54095
250	499	180	374.5	67410
500	999	85	749.5	63707.5
1000	4999	35	2999.5	104982.5
5000	10000	15	7500	112500
		1000		422632.5

Table 5 Calculating the cumulative frequency Class

Class	Number of booksellers	To calculate cumulative frequency	Cumulative frequency
0-49	160	160	160
50-99	215	160+215	375
100-249	310	160+215+310	685
250-499	180	160+215+310+180	865
500-999	85	160+215+310+180+85	950
1000-4999	35	160+215+310+180+85+35	985
5000-10000	15	160+215+310+180+85+35+15	1000

Table 6 Mid points and cumulative frequency

Lower boundary	Upper boundary	Number of booksellers	Mid point	Cumulative frequency
0	49	160	24.5	160
50	99	215	74.5	375
100	249	310	174.5	685
250	499	180	374.5	865
500	999	85	749.5	950
1000	4999	35	2999.5	985
5000	10000	15	7500	1000

differences between the mean, median and mode indicate that our data is skewed. We will explore this further in the next chapter.

11 Dealing with data in classes

Sometimes we don't have access to all the data, but we are given a summary of the data classified into groups (sometimes referred to as `bins'.) The booksellers statistics are from: The Publishers Association (2011)[5] showing the turnover of 1000 booksellers is a good example of this. It tells us that that there were 160 booksellers with a turnover of between 0 and 49. (the numbers are in thousands here so 49 actually means 49,000). Although we know that 160 booksellers were making between 0 and 49,0000 we don't know the exact amount each one is making.

However, we can calculate a five figure summary for this grouped data group by calculating the midpoint for each bin. We add the lower boundary of the group to the upper boundary then divide by 2. For example to find the midpoint of 0-49 we add together 0 (the lower boundary) and 49 (the upper boundary) then divide the answer by 2.

0+49=49

49 ÷ 2 =24.5

Table 3 shows the groups with all the midpoints calculated.

When we calculate the midpoint we are effectively assuming that all the booksellers with a with a turnover of between 0 and 49,000 had an actual turnover of 24,500. It is unlikely that this is actually the case, but it is the best estimate we can make with the data we have.

Mean

To find the mean average we need to estimate the amount of money that all the booksellers turn over. In order to do this we need to add another column to Table 3 to produce Table 4 which calculates the group turnover of each `bin'. We do this by multiplying the mid-point turnover by the number of booksellers.

Now to calculate the mean average we divide the total turnover of all the book sellers, but the number of booksellers
442632.5÷1000=442.6325
As the numbers are in thousands this makes our mean 442,632.50.

Median

To work out the median, lower quartile and upper quartile we need to calculate the cumulative frequency. This means starting with the number of booksellers in the lowest group then adding on each the number of book sellers in the next group (see Table 5).

As we don't know the exact turnover figures we cannot be sure of the exact median; however, we can find out what class it is in. As we have 1000 booksellers the median lies between the 500th and 501st booksellers. If we look at our cumulative frequency we can see that the 500th and 501st booksellers lies in the 100-249 group. As the mid point of this group is 175.5 we can say that the median group is 175.5, actually 175,500 .

The upper and lower quartiles

Finding the upper and lower quartiles is straight forward from here. The lower quartile is the median of the lower half of the book sellers meaning the bookseller in 250th place marks the lower quartile. 0+500 2 =250 The 250th bookseller is in the 50-100 group. As the midpoint of this group is 74.5 we can say that the lower quartile is 74.5. Similarly the upper quartile is the median of the upper half of book sellers meaning the bookseller in 750th place marks the upper quartile

(500+1000)÷2=750
The bookseller in 750th is in the 250- 499 group. As the mid point of this group is 374.5 we can say the upper quartile is 374.5.

12 Exercises

1. Examine Table 7 . Calculate the mean, median and mode length of time each became Prime Minister. (Treat Winston Churchill's and Harold Wilson multiple times as Prime Minister as different entries).
2. Table 8 displays the turnover of UK booksellers by group. What do you think are the a) advantages and b)limitations of displaying the data in groups.
3. Table 9 shows the number and capacity (in tons) of freight wagons used on the Barbados Railway in the 1930s. [6]

a) Calculate the total capacity of the railway in tons.
b) Calculate the mean, median and modal wagon capacity.

References

1. ↑ BBC(2012) Average earnings rise by 1.4% to £26,500, says ONS. http://www.bbc.co.uk/news/business-

Table 7 British Prime Ministers since 1940

Prime Minister	Time as PM (years)
Gordon Brown	3
Tony Blair	10
John Major	7
Margaret Thatcher	11
James Callaghan	3
Harold Wilson (second term)	2
Edward Heath	4
Harold Wilson (first term)	6
Alec Douglas-Home	1
Harold Macmillan	6
Anthony Eden	2
Winston Churchill (second term)	4
Clement Atlee	6
Winston Churchill (first term)	5

Table 8 Number of booksellers registered for Value Added Tax(VAT) Market Research and Statistics Available from www.publishers.org.uk

Turnover Size (thousands)	Number of booksellers
0-49	160
50-99	215
100-249	310
250-499	180
500-999	85
1000-4999	35
5000+	15
TOTAL	1000

Table 9: Number and capacity (in tons) of freight wagons used on the Barbados Railway. Data from: Jim Horsfield (2001) From the Caribbean to the Atlantic: A Brief History of the Barbados Railway St. Austell: Paul Catchpole

Class	Number	Capacity (tons)
A	18	7
AA	31	8
B	2	6
C	10	6
D	4	8
E	2	15
AA converted	4	6
B converted	5	6
C converted	4	6
D converted	3	6
BK converted	2	6
BP converted	6	6

Chapter 4: Measuring spread

1 Mean, median and mode

Means, medians and modes provide a one number summary of a set of data, but they do not tell us anything about the distribution of the data. Suppose that five students sit three exams in three different subjects, and the marks are as follows: Table 1

But look carefully at the number of marks out of ten each student received. We can see that the distribution around the mean is different.

Table 1 Exam marks

Subject	Marks out of ten	Mean Average	Median average
French	2, 4, 5, 7, 7	5	5
Religious Studies	0, 5, 10, 7, 3	5	5
History	5, 5, 4, 6, 5	5	5

By putting the marks into dotplots we can see that the marks for Religious Studies are very spread out and the marks for History are very close together.

2 The standard deviation

The five figure summary and box plots provide a very useful way for summarising a large data set numerically and graphically. A more commonly used measure of spread is called the standard deviation. The standard deviation uses the mean rather than the median as its central point and provides a one number summary of how spread out the data is. In this book we use the abbreviation SD for standard deviation, but you may see it abbreviated to s or STDV.

Figure 1 Dotplot for Religious Studies 1

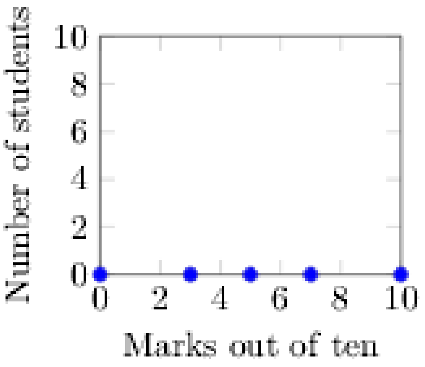


Figure 2 Dotplot for History

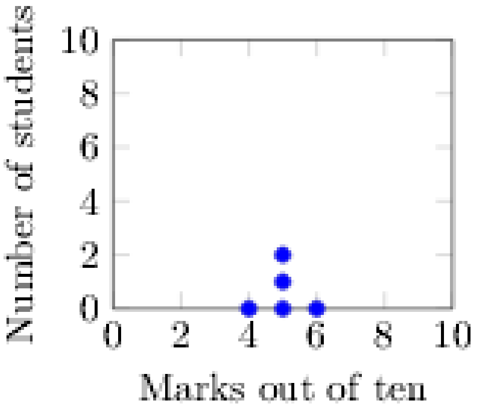


Figure 3 Dotplot for French

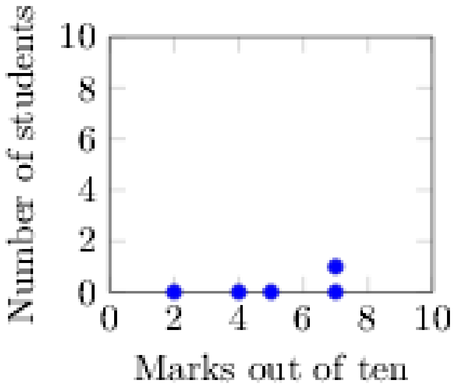


Table 2 Calculating the Standard Deviation

Observation. Each piece of data is called an observation. We write the age of each child in this column. We will call this observation	**Deviation. The deviation is the difference between the observation and mean (which in this case is 8). We take the observation then subtract the mean. This means that some of our numbers will be negative (minus) numbers.**	**Squared deviation. We take the deviation then square it (by multiplying it by itself). These numbers are always positive as a minus number multiplied by a minus number is always a positive number.**
4	4-8=-4	$-4\times-4=-4^2=16$
5	5-8=-3	$-3\times-3=-3^2=9$
8	8-8=0	$0\times0=0^2=0$
11	11-8=3	$3\times3=3^2=9$
12	12-8=4	$4\times4=4^2=16$

Example 1: Suppose a couple have five children aged 4, 5, 8, 11 and 15.

To find out the standard deviation we first need to find the mean age of the children.

$$\frac{4+5+8+11+15}{5}=8 years$$

The mean average is usually written as $\bar{x}$

Now we need to calculate the variance. The variance is a measure of how spread out the data are. This is best done by drawing a table:

2.1 Calculating the Standard Deviation

Add the squares of the deviation together to calculate

$$Variance=\frac{sum\ of\ the\ squared\ deviations}{number\ of\ observations}$$

the sum of the squares of the deviation.

16+9+0+9+16=50

Now we have the numbers we need to calculate the variance.

$$\frac{50}{5}=10$$

The standard deviation is the square root of the variance

$$\sqrt{10}=3.16$$

Standard deviation (SD) = 3.16 years

2.2 Interpreting the standard deviation

The SD is always given in the same units as your observations. In this case we have used the age of five children in years so the SD is also in years.

If we were examining heights in centimetres for our observations then the mean and SD would also be in centimetres. If we were examining weight in tonnes, the SD would be in tonnes. If we were examining bushels of wheat, the SD would be in bushels of wheat.

So what does a SD of 3.16 years actually tell us?

What we can know from this figure is that around 68% of the observations will be within one standard deviation of the mean. We will discuss where this 68% comes from in a later section. In other words we expect that 68% of the children will be been the age of 8 (the mean) and plus AND minus 3.16 years:

Therefore 8 (the mean) + 3.16 (the SD) = 11.16 years

AND 8 (the mean) -3.16 (the SD) = 4.84 years

Therefore 68% of the children will be between the ages of 4.84 and 11.16 years.

As we stated before the SD gives us an idea of how spread out our data is.

Another couple with five children who an average of eight years old, but their children are 8-year old quintuplets.

Like our first couple their children still have a mean age of 8, but their children are all exactly the same age.

If we calculate the SD we will see that the SD equals zero.

This is because there is no variation in their ages.

Figure 2 The Canadian Dionne quintuplets (born 1934), aged about 4. Mean =4, Median=4, Standard Deviation=0. A special Act of the Ontario Legislature, The Dionne Quintuplets' Guardianship Act, 1935 was passed to allow the Ontario Government to take them away from their parents and exhibited as a tourist attraction.

Example 3:

Suppose a third couple have children age 1, 2, 3, 14 and 20. Again this couple have five children with an average age of eight years but they have SD of 7.52 as their children are more spread out.

All three couples have five children with a mean age of 8, but the standard deviations are different. The SD gives us an idea of the spread of the children's ages.

2.3 Stem and leaf plots

A stem and leaf plot is not a graph as such but is a good way of checking out the shape of the distribution of a small sample. The stem and leaf plot can done by hand fairly quickly.
Table 4 records the age of death of 39 people buried in Accrington in 1839.

2.4 The Normal Distribution

If a dataset is normally distributed the mean, median and mode will coincide. Most of the observations will be near this central point with a smaller number far away from the central point.

Imagine a crowd of people and consider their heights. Most people seem to be similar height, give or take a few centimetres. A small number of the people are clearly shorter or taller than the average and an even smaller number of people seem to be very short or very tall.

Figure 3 Heights of Bavarian conscripts (1810-1840)

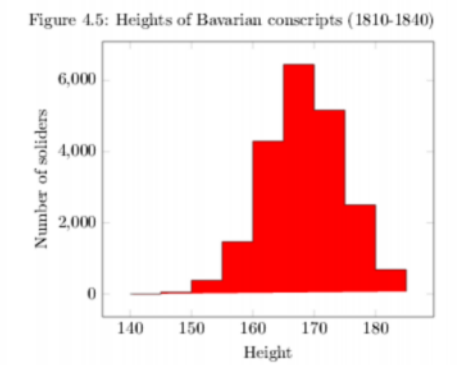


Figure 3 shows the heights of 5000 Bavarian men, conscripted between 1810 and 1840. [1] We can see from Figure 3 (called a histogram) that there were a small number of conscripts who were particularly short (less than 155 cm tall) and a small number who were very tall (taller than 180cm) but most of the conscripts were between around 160 and 172 cm.

So how do we describe the shape of this distribution?

This shape is often called the bell curve due to its resemblance to the shape of a bell. As the dataset is so large I have used a data analysis pack Minitab to calculate the mean and standard deviation of the sample.

The standard deviation is 6.439*cm* (we will say 6.4cm for convenience).

Creating a Stem and Leaf plot

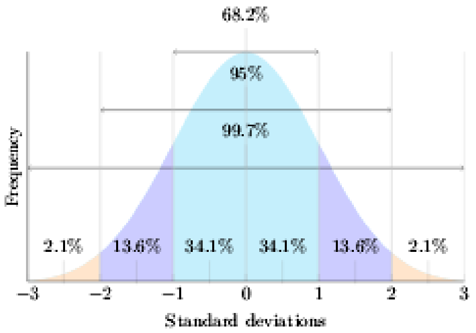
8	20	63	78	32
17	82	1	22	0
0	22	80	73	47
0	38	12	35	25
57	1	0	0	0
21	39	3	0	15
54	48	29	41	1
2	13	0	1	

0	8 0 0 2 1 1 0 3 0 0 0 1 0	Put numbers 0 to 9 in this row
1	1 2 3 7	Put the second number of 10-19 in this row
2	0 1 2 2 2 9	Put the second number of 20-29 in this row
3	5 8 9	Put the second number of 30-39 in this row
4	4 7 1 8	Put the second number of 40-49 in this row
5	7 5 5	Put the second number of 50-59 in this row
6	3	Put the second number of 60-69 in this row
7	8 3	Put the second number of 70-79 in this row
8	2 0 0	Put the second number of 80-59 in this row

STEM	LEAF
0	0 0 0 0 0 0 0 1 1 1 2 3 8
1	1 2 3 7
2	0 1 2 2 2 9
3	5 8 9
4	1 4 7 8
5	7 5 5
6	3
7	3 8
8	0 0 2

Figure 4 The normal distribution

Figure 4.6: The normal distribution



The mean is

$166.8cm$

So what are the heights for the conscripts 1 standard deviation (1 SD) from the mean?

As before we take our mean of

$168.8cm$

then add the SD.

$168.8cm+6.4cm=175.2$

Then we take our mean again and subtract the SD

$168.8cm-6.4cm=162.4$

Therefore we can state that the conscripts with heights of between 162.4 cm and 175.2 cm are within one standard deviation (1SD) of the mean. Approximately 68% of data is within 1 SD of the mean; this is a rule about the standard deviation.

Now we have calculated SD we can also identify the heights of conscripts within 2 standard deviations (2SD) of the mean. This will take into account the heights of conscripts who were taller or shorter than those within 1SD of the mean.

To find the extent of the second deviation we simply add or subtract the standard deviation twice instead of once.

$168.8cm+6.4cm+6.4cm=181.9$

$168.8cm-6.4cm-6.4cm=156.0$

Therefore we can say that conscripts between 156.0 cm and 181.9 cm are within 2 standard deviations (2SD) from the mean. Approximately 95% of data lies within 2SD of the mean.

We can keep on going to calculate the heights of conscripts within 3 standard deviations (3SD) of the mean.

$168.8cm+6.4+6.4+6.4=188.3cm$

$168.8cm-6.4-6.4-6.4=149.6cm$

Therefore we can say that conscripts between 149.6 cm and 188.3 cm are within 3 standard deviations (3SD) from the mean. Approximately 99.7% of data lies within 3SD of the mean.

We can see at this point that only 0.3% of conscripts remain outside 3 standard deviations from the mean (0.15% of whom are very very short and 0.15% are very very tall). We can keep on adding standard deviations with smaller and smaller percentages of conscripts, but by now we have a good idea of the distribution of our dataset.

3 Non-normal distributions

Not all samples are normally distributed. In the case of the burials in Accrington or the farms on Chile we can see that the mean, median and mode do not coincide; they are not even close to one another. Many of the statistical tests in Part 2 work on the assumption that the data is normally distributed– these are known as parametric tests meaning the tests are based on assumptions that the dataset lies within the parameters or expected pattern of the normal distribution. If the data is not normally distributed then these parametric tests will not be reliable. Non-parametric tests are those which do not assume a normal distribution. This will be noted in each section of the book.

3.1 Skewness

Skewness is a very important concept in dealing with social data. Normal distributions are common in natural phenomena such as the distribution of people's heights. However data from the human social world is often not normally distributed. As we have seen with the examples of the Chilean farms only four of the 24 are larger than the mean average.

In a normally distributed sample the mode, median and mean will coincide, meaning that the same number of observation will be below the mean as above the mean. In non-normal distributions, this is not the case. Non-normal distributions are particularly common in the case of income. A large number of people have below average incomes and a small number of people have very high incomes. This results in a mean which is misleading. Any report of income which states only an average should be treated with suspicion. In the above examples it is quite easy to see that the samples are skewed. The skewness test enables us to identify whether or not a distribution is skewed (not always as easy to spot as in the examples here), the size of the skew (is it near normal or far from normal) and the direction of the skew, (Is it positively skewed or negatively skewed?). A distribution which is perfectly normally distributed will have a skewness of zero. A positive skew (a number greater than zero) will occur where most of the observations are less than the mean and a negative skew (a number of less than zero) means that most of the observations are greater than the mean.

3.2 Calculating skewness

To calculate skewness we will use the same data as for calculating the standard deviation of the children's ages in Section 1. The SD for the children's ages in Example 1 was 3.16. We will need this to calculation skewness. The equation of measuring skewness is

$$Skewness = \frac{\Sigma(x-\bar{x})^3}{(n-1)SD^3}$$

n= the number of observations
SD = The standard deviation
$(x-\bar{x})^3$ = Sum of Squares to the power of 3 (see Table 7)

$$\frac{0}{(5-1)\times 3.16^3} = 0$$

$$Skewness=0$$

Interpreting the skewness

Our skewness is 0, but what does this mean? Bulmer suggests the following guidelines (see Table 8). [2]

So with a skew of 0 we can say our data is not skewed at all.

Kurtosis

Kurtosis is often neglected in statistics books; Kurtosis is a measure of the `peakiness' of the distribution. Some distributions have a sharp peak and others a flatter peak. A distribution can be normally distributed (has a measure of skewness close to zero), yet have a high Kurtosis. The equation for measuring kurtosis is

$$Kurtosis = \frac{\Sigma(x-\bar{x})^4}{(n-1)SD^4}$$

n = the number of observations
SD=The standard deviation
$(x-\bar{x})^4$ = The Sum of the Squares to the power of 4 (see Table 7.

$$\frac{674}{(5-1)\times 1.364} =$$

$$\frac{674}{4\times 99.71} =$$

$$\frac{674}{398.8} = 1.69$$

Table 7 Table for calculating skewness and kurtosis

Age of child	$x-\bar{x}$	$(x-\bar{x})^2$	$(x-\bar{x})^3$	$(x-\bar{x})^4$
4	-4	16	-64	256
5	-3	9	-27	81
8	0	0	0	0
11	3	9	27	81
12	4	16	64	256
Totals		50	0	674

Figure 7: Measures of skew and kurtosis are two similar tests which help us identify how biased a distribution is below or above the mean and the shape of the distribution

Figure 4.7: Distribution of skew and kurtosis

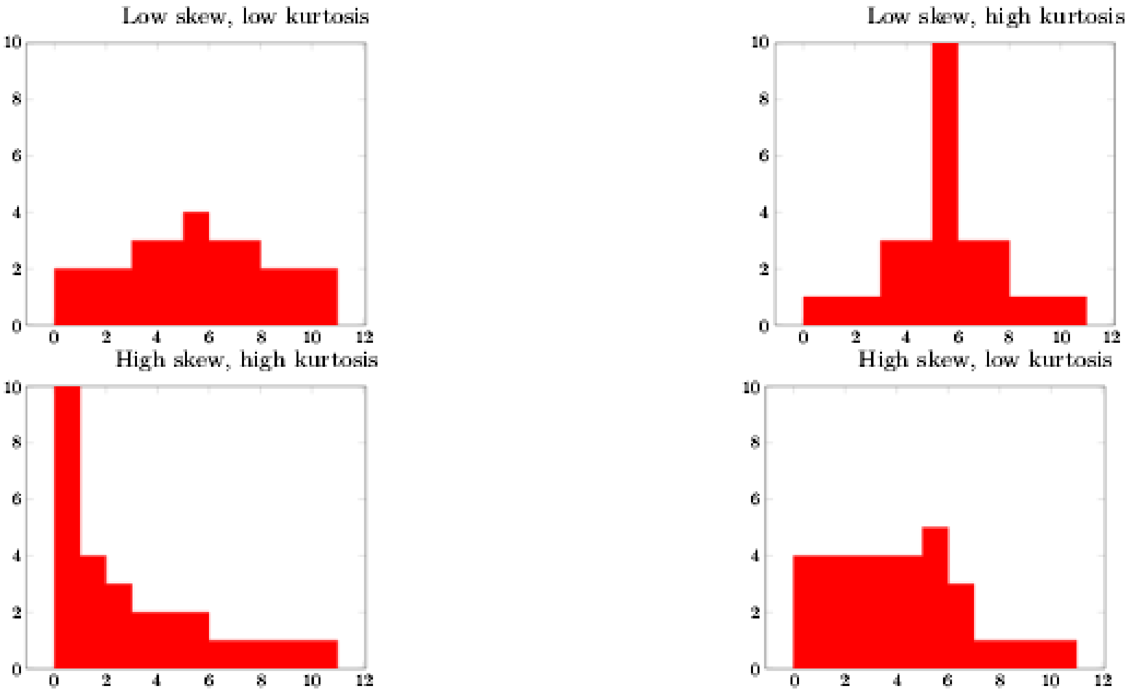


Table 8 Interpreting skewness

Description	Measure of skew
Highly skewed	More than 1 or less than -1
Moderately skewed	Between 0.5 and 1 or -0.5 and -1
Low skew	Between 0.5 and -0.5
No skew	Zero or approximately zero

Table 9: Interpreting kurtosis

Measure of kurtosis	Possible characteristics
Greater than 3	Leptokurtic distribution. `Peakier' and sharper than a normal distribution. Values concentrated around the mean and with high probability of extreme values
3	Mesokurtic. A normal distribution would have a kurtosis of 3.
Less than 3	Platykurtic distribution. Usually flatter than a normal distribution with a wider peak. The probability for extreme values is less than for a normal distribution. Values spread wider round the mean

Exercises

1. The following are results from a French speaking exam: 5, 6, 7, 10, 1, 2, 9, 8, 8
    1. Calculate the mean, median and mode.
    2. Calculate the standard deviation.
    3. Calculate the skew and kurtosis.
2. Draw a stem and leaf plot of the age at accession of English monarchs, Table 10

References

[1] Baton, J (1999) Heights of Bavarian male , 19th century, Data hub Heights and Biological Standard of Living Available from http://www.uni-tuebingen.de/uni/wwl/dhheight.html
[2] M G. Bulmer (1979) *Principles of statistics*. Dover}.
[3]These figures have been derived from those used by D. Ebdon (1985) *Statistics in Geography* (Oxford: Blackwell).
[4]Lawrence T. DeCarlo(1997) On the Meaning and Use of Kurtosis, *Psychological Methods* 2, pp.292-307 goes into some details about the complexities

Table 11: Top ten subsidised theatres (England) 2012-13: Source: Arts Council for England

Theatre name	Money each is receiving from ACE financial year 12-13, (£)
Royal National Theatre	17,462,920
Royal Shakespeare	15,675,270
Royal Exchange	2,318,609
English Stage Company (Royal Court Theatre)	2,297,916
Leicester Theatre Trust Ltd (Curve Theatre)	1,903,000
Birmingham Repertory Theatre	1,823,385
Young Vic	1,750,000
Liverpool Everyman and Playhouse	1,649,019
Chichester Festival Theatre	1,604,079
Northern Stage	1,551,976

Table 10 Age at Accession: English and (from 1603) British monarchs

Monarch	Age at Accession	Monarch	Age at Accession
King William IV	64	King Edmund II lronside	25
King Edward VII	59	Queen Elizabeth II	25
King George IV	57	King Henry V	25
King George I	54	Queen Elizabeth I	25
King James II	51	King Charles I	24
King Harold II	45	King Edward II	23
King George V	44	King Edred	22
King George II	43	King George III	22
King Edward VIII	41	King Harthacnut	21
King George VI	40	King Henry II	21
King Stephen	38	King Cnut (Canute)	21
King William I	38	King Harold I Harefoot	19
King William III and Queen Mary II	38	King Edward IV	18
King Edward The Confessor	38	King Edmund	18
Queen Mary I	37	Queen Victoria	18
Queen Anne	37	King Henry VIII	17
King James I	36	King Edgar	16
King Henry IV	33	King Edwy (Eadwig)	15
King Edward I	33	King Edward III	14
King John	32	King Edward V	12
King Henry I	31	King Edward The Martyr	12
King Richard I	31	King Richard II	10
King William II	31	King Aethelred II The Unready	9
King Richard III	30	King Edward VI	9
King Charles II	30	King Henry III	9
King Athelstan	29	King Henry VI	0
King Henry VII	28		

CHAPTER 5: SAMPLING

1 Sampling

It is rare that we are able to look at a set of data in its entirety. We do not have time to ask every Muslim in Britain to answer our questionnaire, read through the entire 1891 UK census or collect language use data from every single person who speaks English. When we are collecting data we need to take a **sample**. A sample is a part of something which gives an idea about the whole. We might buy a paint sample to get an idea of what the whole wall will look like when it has been painted. A carpet sample is a small piece of carpet which gives an idea of the colour, thickness and fluffiness of the carpet when it has been laid in a whole room.

2 Defining terms

2.1 Population

The entire group of objects about which information is wanted. See Table 1 for a list of examples.

Figure 1 A unit from a population of chickens

Table 1 Population for different research questions

What we want to find out about	Our population
Use of the Irish language amongst people living in Ireland in 1911	People living in Ireland in 1911
Use of the Irish language amongst people living in Castleffrench, Galway in 1911	People living in Castleffrench, Galway in 1911
Life expectancy of people in Accrington in the nineteenth century	People in Accrington in the nineteenth century
Heights of Bavarian soldiers in the nineteenth century	Bavarian soldiers in the nineteenth century
The occupations of people who sailed on the Titanic	People who sailed on the Titanic
Attitudes of school pupils learning French in Hampshire	School pupils learning French in Hampshire
Development of the use of the word ‘noob’ in the English language	A corpus of written English covering different years.
The cost of bags of wool in England in the 1500s	Bags of wool sold in England in the 1550s

2.2 Unit

Any individual member of the population. Depending on the population a unit could be:

- A person
- A bag of wool
- A cow or chicken
- A soldier
- An Irish speaking person

2.3 Sample

Part or sub-set of the population used to gain information about the whole. By definition a sample is not the whole population. Going back to the above examples a sample could be:

- Some, but not all, pupils studying French in Hampshire
- Some, but not all, people who sailed on the Titanic
- Some, but not all, people who lived in Accrington in the nineteenth century.
- Some, but not all, bags of wool sold in the 1500s.

2.4 Sampling frame

The source or list from which the sample is chosen. A sampling frame could be:

- The 1911 Census of Ireland
- A list of school pupils studying French in Hampshire
- The passenger list from the Titanic
- Records of wool sales in in England in the 1500s.

2.5 A variable

The characteristic of a unit, to be measured for all those units in the sample. We will be finding out specific information about the population of Ireland or the passengers on the Titanic.

Examples of variables might include:

- Sex
- Income
- Place of Birth
- Religion
- Language spoken
- The price of oats
- Number of windows
- Occupation
- Opinion on how enjoyable it is to study languages.

2.6 Census

Obtaining information about every unit in a population. The best known type of census are the population census when everybody who lives in particular country is asked questions. A census is when we look at every individual person, bag of wool, tax return, school pupil etc. By definition a census is not a sample.

3 Why we take a sample

When we take a sample we hope that our conclusions about the sample apply the whole population generally. When pollsters ask 1000 people how they plan to vote in an election their underlying purpose is to come to a conclusion about what the election result will be when everybody has voted. If we look at a sample of children learning French in Hampshire we are intending to draw conclusions about all children learning French in Hampshire, even if they were not included in the survey.

One of the most famous examples of sampling going wrong was the Chicago Tribune's headline"Dewey defeats Truman" in the 1948 US election. Sure of their sampling techniques journalists at the newspaper were confident enough to proclaim that Dewey had won. When all the votes at been counted they discovered that this was not the case. Their sample led them to believe that Dewey would win but it turned out that the electorate as a whole (the population) favoured Truman. Although this is well-known example of sampling going wrong the great thing about opinion polls of this kind is that we can see where the sampling went wrong (the sample suggested an outcome which turned out not be true). However when it comes to knowing whether the conclusions we draw from sampling bags of wool, Irish speakers or school pupils are representative of the population as a whole we are unlikely to get such quick feedback on our conclusions. So how do we go about taking a sample of our population?

4 A convenience sample

I used the example of Accrington in Chapter 3 because the Lancashire Parish Registers are available on the internet.[1] Of all the settlements in Lancashire I chose Accrington because it was near the top of the list which was in alphabetical order. 1838 was just a year that I guessed at. This is a pure convenience sample. I do not know whether the people of Accrington enjoyed a longer or shorter life than other residents of England or of Lancashire. I do not know whether or not 1838 was a year of a lot of deaths or few deaths. I do not know what diseases might have been going round Accrington in 1838 and what age of people would have been most effected. In no sense do I know if Accrington was representative of the country as a whole. To get an accurate picture of life expectancy in nineteenth century England we would need to get a lot more data. So why don't we just collect the ages of all the people who died in England between 1800 and 1899?

Put simply we do not have time to search through the records and to record the age of death of each person who died in England in the 100 years of the nineteenth century.

5 Random sampling

The advantage of selecting at random is that a random sample is free of researcher bias. If I was just picking and choosing from the 1911 Census of Ireland I might be inclined to choose individuals or households which I think might be interesting. I've just been searching the census for individuals over 100 years-old and came across a 104 year women called Ellen Hefferan working as a live —in domestic servant in County Wexford which is both unusual and interesting. I might conveniently choose to make her part of my sample.

You can take a random sample by giving each unit (person, household, bag, cow etc.) a number and use random numbers to select your sample. Traditionally you could have a used a printed page of random numbers or a random number generator on a scientific calculator. Today, however, the internet is your friend and www.random.org will generate random choices of all sorts including coin tosses, playing cards and numbers. This website even has a free list randomiser. For our Irish community example we can simply paste in the names then click the generate button and select the names at the top of the list for our sample.

6 Stratified sample

One of the problems with random samples is that some types of people (or communities or objects) might be underrepresented, overrepresented or not presented before in our sample. We might end up with a disproportionate number of men compared to women, urban residents to rural residents, English speakers to French speakers, Catholics to Protestants, or big towns to small towns. Stratified sampling ensures that our sample contains individual units (people, bags of wool, cows etc.) from different strata of the population. Suppose we want to take a sample of 100 people from the village of Castleffrench in Ireland with a population of 1000 people. We want to explore household size in the 1911 census. From the census we see that:

1. 5% of the population are Irish Speakers.
2. 2% are over 80
3. 48% are male and 52% female
4. 75% are Catholics and 25% are Protestants

We can use a stratified sample to ensure that people from each of these groups are included in the sample. A total random sample cannot guarantee that every group will be represented in the sample. We can see that just 2% are over 80. If we randomly select 100 people how many of those people will be over 80? 2% of 100 is 2 so we might say that we expect two people in the random sample to be over 80. We will explore randomness and probability later in this book, but can we be sure that we would get two people?The answer is that we can't be sure. We may get two, but we may get one or zero or six or ten. A stratified sample will ensure that we have some representatives from this group of people in our sample. If a strata of the population is small we might choose to oversample, that is to select more than 2% of the population over 80. We can then take a random sample with each section.

7 Important considerations when sampling

Samuel Johnston is reputed to have said "You don't have to eat the whole ox to know the meat is tough.". This is a good metaphor from sampling — we can draw a conclusion about the whole from a sample of the part. An ox is actually a good illustration for statistics. The toughness of the meat depends upon which part of the ox the cut came from. If our sample is a sirloin cut we will come to different conclusions about beef than if our sample comes from the shank. However carefully we select a sample there are certain issues which may arise: These are particularly common in questionnaire research, but they can appear in other situations too.

7.1 Oversampling

Units from outside the population are included These would need to be removed (if we know that they are there). For our survey of school pupils studying French in Hampshire we may find that some of our questionnaires have been filled in by pupils in Hampshire not studying French or pupils studying French in other counties.

7.2 Undersampling

Units from certain groups in the population are not included or are underrepresented. In the case of the Irish community we may find that no people over 80 are included if we took a random sample.

7.3 Non-response bias

A particular problem with questionnaires is that 100% response rates are very unusual. It is difficult to ascertain whether the people who did not answer your questionnaire have the same characteristics or opinions as the people who did respond. Even censuses have this problem to a certain extent — they are meant to survey the whole population, but some people are missed out or didn't respond for one reason or another.

7.4 Sampling frame error

A sample is only as good as the sampling frame from which the sample was derived. If the sampling frame is incomplete, inaccurate or excludes units which are present in the population you are researching then your conclusions are less likely to be true of the whole population. There were many shortcomings with polling in the 1948 "Dewey defeats Truman" election, but one was using the telephone for polling. Owners of telephones were more likely to support Dewey, whereas non-telephone owners were more likely to support Truman. However pollsters assumed that telephone owners were representative of the population as a whole, and called the election for Dewey. [2]

7.5 Voluntary response bias

People who volunteer to fill in your questionnaire or contact you to give their opinion are not necessarily representative of the whole population. If you ask people to give their views on the TV show Neighbours then the respondents are likely to be people who watch Neighbours. They might also be people who have time to respond to you or are particularly outspoken. They may not be representative of the population at large or of Neighbours fans in particular.

8 Exercises

1. A headteacher wishes to know whether or not parents of children at her school favour expanding the range of languages available at GCSE. She receives 33 letters of which 29 support an expansion of languages and four oppose. Can she assume that 87.9% of parents support expanding the range of languages? Why/ why not?
2. You are investigating the extent to which people who identify as 'Christians' believe in traditional doctrines such as the resurrection and the virgin birth. You have designed a questionnaire to find out. What sampling issues arise if you carry out your questionnaire:
    1. Face-to-face on the high street on a Saturday afternoon?
    2. Face-to-face outside a church on a Sunday morning?
    3. On your personal website?
3. You are researching life expectancy in Accrington for the whole of the nineteenth century.You don't have time to look at all burials in the town. How might you sample in order to get a picture of life expectancy over the course of the whole century?
4. You are doing a survey of television viewing habits in your neighbourhood. Who might be excluded from your survey if you carried it out:
    1. As a door-to-door survey on a Tuesday morning?
    2. As an online survey
    3. As a telephone survey on a Friday evening?

References

[1] Online Parish Clerks for the County of Lancashire http://www.lan-opc.org.uk/

[2] Polling errors in the Truman-Dewey election were not restricted to this one poll. See, for example Jeanne Curran and Susan R. Takata (2002)*Getting a Sample Isn't Always Easy* Online at http://www.csudh.edu/dearhabermas/sampling01.htm

CHAPTER 6: MEASURING CHANGE

1 Measuring change

- nflation rises to 5.1%
- Unemployment falls by 2.3%
- New drug could increase cancer survival rates by 10%.

Every day there will be a headline like one of these in the news. These statistics are measures of change, or more accurately the rate of change.

Suppose we wish to describe the change in numbers of children in Scotland being educated thought the medium of Gaelic that took place between 2005 and 2011 (see Table 1). We could describe the change in terms of absolute numbers

$$2929-2480=449$$

So in 2011 there were 449 more children studying through the medium of Gaelic than in 2005.

Table 1 Number of children in Scotland being educated in Gaelic schools.

Year	Number of children educated in Gaelic
2005	2,480
2006	2,535
2007	2,601
2008	2,766
2009	2,638
2010	2,647
2011	2,929

To calculate a percentage increase

First we need to find the difference between the two numbers as we did above

$$2929-2480=449$$

Then we divide the answer by the original number

$$\frac{449}{2480}=0.18$$

Then we multiply the answer by 100 to get the percentage

$$0.18\times100=18\%$$

Therefore between 2005 and 2011 the numbers of pupils studying through the medium of Gaelic increased by 18%.

To summarise, to calculate a percentage increase

$$\frac{New\ value-Original\ value}{Original\ value}\times100$$

To calculate a percentage decrease

To calculate a decrease we can use exactly the same formula. The only difference is our answer is a negative number:

$$\frac{New\ value-Original\ value}{Original\ value}\times100$$

To use an example: the number of students accepted to study Non-European languages fell from 1,691 in 2006 to 1,434 in 2011. First we need to find the difference between the two numbers as we did above

$$1434-1691=(-257)$$

(Note that this is a minus number)

Then we divide this answer by the original number (that is the 2006 number)

$$\frac{-257}{1691} = (-0.15)$$

(Again this is a minus number).Then we multiply the answer by 100

$-0.15\times100=(-15)$

Our final answer is also a minus number, which shows that there has been a reduction of value. Therefore between 2006 and 2011 the number of students accepted to study non-European languages fell by 15%.The same formula is used to calculate a percentage increase or decrease.

1.1 Calculate percentage change in terms of real numbers

Sometimes we will need to work out how a certain percentage change will affect the original number.Over the six years between 2005 and 2011 we saw that there was an 18% increase in number of children attending Gaelic-medium schools. How many is an 18% increase? In 2011 there were 2,929 children studying in Gaelic-medium schools. How many children will be studying in Gaelic-medium schools in 2017 if there is an 18% increase between 2011 and 2017? There are two ways to calculate this:

In 2011 the number of children was 2,929. The quickest way to calculate is to multiple the original number by 1.18%.

$2929\times1.18=3456$

Alternatively, We can find 18% of 2929 by dividing by 100 then multiplying by 18.

$$\frac{2929}{100}\times18 = 527$$

In 2011 the number of children was 2,929. The quickest way to calculate is to multiple the original number by 1.18%.

$2929\times1.18=3456$

Then we add the 527 (the 18%) onto the original number (2,929)

$2929+527=3456$

Therefore if there is an 18% increase in the number of children studying in Gaelic-medium schools between 2011 and 2017 there will be 3,456 children studying in Gaelic-medium schools in 2017.

So what if we were to predict that numbers in Gaelic-medium schools were to decrease by 18% by 2017?

The first step is the same as Method 2 above.

$$\frac{2929}{100}\times18 = 527$$

First we need to find 18% of 2,929

Then subtract this 18% (527) from 2929.

$2929-527=2402$

So if there is to be an 18% decrease between 2011 and 2017 there will be 2,402 pupils studying in Gaelic-medium schools.

1.2 Year on year changes

Our local council charges us a tax of £100 each year. They come with good news that they are going to reduce the tax by 10% each year for the next three years. They tell us "Your tax bill will be 30% lower in three years time".

Is their claim true?

After one year

$100-10\%=100-10=90$

After two years

$90-10\%=90-9=81$

After three years

$81-10\%=81-8.1=72.90$

After three years our tax bill will be £72.90

However a 30% reduction from £100 is

$100-30\%=100-30=70$

So after three years our bill has been reduced by less than 30% overall?

Why isn't the decrease 30%?

You will remember from the previous section that the formula for calculating percentages is

$$\frac{New\ value - Original\ value}{Original\ value}\times100$$

At the end of each year we have a new original value. The new value at the end of year one becomes the original value when we calculate the value at the end of year 2. So £90 is the new value at the end of Year 1 and the original value at the end of Year 2.

And what if the tax increased by 10% each year for the next three years? The same principle would apply.

- A 30% increase would equal 100+30= 130 paid at the end of the year 1
- In contrast if the rise is 10% each year:
  - £100+10%=100+10=£110
  - £110+10%=110+11=£121
  - £121+10%=121+12.1=£133.10
- So whereas a 30% increase on 100 would be 130, a 10% increase each year over three years would lead to us paying 133.10 per year at the end of year 3. We are actually paying 33.1% more.

2 Interpreting changes over long periods of time

Prices

So how do we calculate changes over long periods of time? The UK Retail Price Index upon which inflation (price rises) is measured is calculated from the price of a `basket of goods'. Changes in the price of goods not in the basket do not count towards in the inflation figure. At the time of writing canned tomatoes are included but canned peas are not. The contents of this basket change over time to reflect changes in technology and buying patterns. In 2010 music downloads were added to this basket of goods. [1] In 1956 E H Phelps Brown and Sheila Hopkins published their seminal paper on prices and wages in England since 1260.[2] Gathered from a huge range of sources this paper created what has become know as the Phelps Brown Hopkins index. It an index of wages and prices for almost every year from 1260 right up to 1954, two years before their paper was published. This index continues to inform our understanding of prices since the Middle Ages. The fact the data is year on year means that we can identify short term changes in prices and wages as well as longer term trends. As we noted in the introduction of this book we can only use the statistics we have. We cannot use figures we don't have. The Phelps Brown Hopkins index has been constructed upon information about prices which does exist. It can't take into account everything from regional variations as we don't have information on the prices of all necessities (or non-necessities) from 1260 to 1954. The index uses different products at different times in history reflecting changes in what people bought and needed, but in part based on price information which is available. This reminds us about the ways in which statistics are socially constructed.

Wealth

According to the UK National Archives website “Jane Austen left less than 800 in assets when she died in 1817.[3] The use of the term `less than' implies that 800 was not a great deal of money in 1817, but was this really the case? How rich was Jane Austen? There are two main questions here:

1. What could Jane Austen buy with her £800 in 1817?
2. How wealthy would Jane Austen be if she were alive today? £800 was worth a lot more in 1817 than it is today, but how much more?
3. We can measure today's value of her income in two ways.

   Supposing her income had risen in line with prices since 1817

   Supposing her income had risen in line with increases in earnings since 1817.

The first question relates to Jane Austen's purchasing power with her 800. To give some sort of perspective 13 per year was sufficient to feed, cloth, teach and shelter a boy at Christ's Hospital. [4] To work out the answer to the first part of the second question we can use the Retail Price Index to calculate how much prices have risen since 1817. Fortunately the www.measuringworth.com website can do the calculations for us here. If we type £800 in 1817 into the indicator tool we will get a figure of £47,000 by 2010 prices, barely a quarter of the price of an average house in the UK. This seems quite modest for a novelist who was part of the landed gentry and wrote novels based around the lives of this group of people.

However, we could also look at how much Austen's £800 would be worth were still alive today and her income (and that of everyone else living in 1817 had risen at an average rate). On this measure £800 comes to £562,000. If we look at the average value of an estate in 2008 we will see that £562,000 would put Jane Austen is well within the top 6% of estates today. [5] She may not have been a millionaire like many bestselling authors of today, but it is difficult to argue that she was poor. Also she was part of a family with substantial property which is not included in this 800. Building and maintaining great houses would run into the tens of thousands of pounds. [6]

2.1 Marriage and divorce

It is commonly heard that around 50% of marriages in the UK will end in divorce. But wait! Saying that 50% of marriages will end in divorce is predicting future trends from past data. So where does the 50% come from? Approximately 40% of couples who got

Figure 1: Chawton House, Chawton, Hampshire. Home of Jane Austen

married in 1985 are now divorced. This does not mean that of all couples who are currently married 40% (or more) will get divorced. How long then does the average marriage last?

The problem is that we only have actual data on marriages which have ended, either through the death of one of the marriage partners or through divorce. This means that any discussion of how likely any couple who are currently married will get divorced at some point in the future is, statistically speaking, an exercise in predicting future events. All couples who are married today *could* remain married until one partner dies, or all couples *could* end their marriages through divorce. Our own experience tells us that neither scenario is likely, but it does raise the question about where statistics we hear about divorce actually come from.

One measure used is the *Crude Divorce Rate*. The Crude Divorce Rate is the number of divorces per 1000 population. In 2002 the Crude Divorce Rate in the UK was 2.7. [7] In other words for every 1000 people in the UK there were 2.7 divorces. By 2010 this had fallen to 2.1. Does this mean that divorce is becoming less common?

An alternative measure is the *Divorce to Marriage ratio* which the proportion of divorces to marriage in any one year. In the UK there were 1.8 marriages for each divorce in 2002 and 2.1 marriages for each divorce in 2010.

Both the Crude Divorce Rate and Divorce to Marriage Ratio are impacted by the proportion of the population which are (or are not) of marriageable age. [8] If a society has a high proportion of elderly people (who are less likely to get married) or a high proportion of young people then divorces per 1000 population are likely to be quite low as there are fewer people relative to the total population likely to get married at all.All these measures assume that people get divorced and married in the same country– this is not always going to be the case of course. The Divorce to Marriage ratio is also problematic; the couples who get divorced in any one year are not the same people who get married in that year. Therefore the Divorce to Marriage ratio is assuming a relationship between the number of couples getting divorced today and the number of couples who will get divorced in the future.

Another way of measuring divorce is to examine *cohorts*. A cohort is a group of people who share in a particular experience at a particular time. We can refer to all the people who married in 1970 as the `1970 cohort' and all the people who married in 2005 as the the `2005 cohort'. By 2010 35% of the 1970 cohort were divorced (just over 1 in 3) and 8% of the 2005 cohort were divorced. 65% of the 1970 cohort and 92% of the 2005 cohort were still married in 2010 (minus those whose marriages ended with the death of a partner), but there is a strong indication that divorce has become more common. Of those married in 1985 40% were divorced by 2010. The statistical complexity of marriage and divorce is not merely an intellectual curiosity. Marriage carries political, sociological, cultural and religious meanings which are the subject of passionate debate. Should marriage be encouraged through the tax system? Do children thrive better if their parents are married? should marriage be available to couples who are the same sex? Is marriage an outdated convention which should be dispensed with altogether? Divorce statistics only measure relationship break-up in couples who have been married and do not account for relationship changes amongst couples who have not been married. The statistics cannot tell the public or policy makers what policies should be put in place in light of new trends. Different people will react to the statistics in different ways.

The only truly reliable way to calculate a divorce rate is when *all* the marriages of a particular cohort have ended, through death or through divorce. The assumptions under which future trends are forecast should always be made explicit.

3 Differences between countries and regions within countries

Similar questions about income and wealth arise when comparing economic data between different countries. When comparing incomes in different countries it is important to ask about purchasing power and the relative wealth of a person on a given income. For example the world poverty measure of how many people live on less than one dollar per day is not a case of converting one US dollar into a local currency and seeing how many people earn less than it. The measure creates a basket of goods which would cost one dollar in the USA and then measures the number of people in a particular country who, on average, do not earn enough each day to buy that basket of goods in their own country. The convenience of the dollar per day statistic is coincidental. In the 1980's economists at the World Bank noticed that a number of less wealthy countries defined poverty as an income of less than around $370 a year. As there are 365 days in a year this conveniently worked out at just over one dollar per day. [9] Similar questions arise about differences between difference regions within a country. In England and Wales teachers are on national pay scales, - that is they earn the same salary irrespective of where they live (though this is a small premium of teachers in London). In 2003 a report for the Joseph Rowntree Foundation [10] reported that the basic starting salary for a teacher outside London was 23,835. Imagine a situation where a teacher wishes to buy a family home in the lowest quartile near to their workplace. A teacher working in Mole Valley, Surrey would need to secure finance for 4.43 times their annual salary whereas a teacher in Hartlepool would need to secure just 2.3 times their annual income to buy the same sort of house. In income terms both teachers earn the same annual salary, but the teacher in Hartlepool is able to buy a house for a third of the price as the teacher in Mole Valley. In conclusion, when comparing between different eras or places, we must not use only absolute figures, but also look into their relative value.

4 Exercises

Examine Table 2. The Publishers Association (2011) *Market Research and Statistics* Available from www.publishers.org.uk

1.Calculate the percentage increase of Digital books sales from 2009 to 2011

2. Calculate the percentage decrease in sales of Physical books sales from 2009 to 2011 Table 3 [11] shows the prices of some typical American thanksgiving ingredients in 1911 and 2012.

3. What information would you need to know in order to find out whether these ingredients are really more expensive in 2012 than in 1911?

4. Table 4 shows the number of agricultural and forestry related accidents in Luxembourg between 1960 and 2009.[12]

5. Calculate the percentage decrease in the number of accidents between 1960 and 2009.

6. Calculate the percentage increase or decrease between:

1. 1960 and 1970

2. 1970 and 1980

3. 1980 and 1990

4.1990 and 2000

5.2000 and 2009

5 References

1. Philip Gooding (2012) *Consumer Prices Index and Retail Prices Index: The 2012 Basket of Goods and Services*(Newport: Office of National Statistics)
2. E.H. Phelps Brown and Sheila Hopkins (1956), "Seven Centuries of the Prices of Consumables, Compared with Builders' Wage-Rates," *Economica* 23:92, pp. 296-314
3. http://www.nationalarchives.gov.uk/museum/item.asp?itemid=33
4. Ralph Turvey (2010) The Cost of Living in London, 1740-1834 (London: London School of Economics). http://eprints.lse.ac.uk/29960/1/WP147.pdf
5. Law Commission (2011) *Intestacy and family provision claims after death*Consultation Paper No 191. Available from:http://lawcommission.justice.gov.uk/docs/cp191IntestacyConsultation.pdf
6. Stobart, Jon. Gentlemen and Shopkeepers: Supplying the Country House in Eighteenth-century England.â The Economic History Review 64, no. 3 (2011): 885-904..
7. The figures used in this section can be found at http://appsso.eurostat.ec.europa.eu
8. England, J. Lynn; Kunz, Phillip R. (1975), "The Application of Age-Specific Rates to Divorce", *Journal of Marriage and the Family* 37 (1), pp.40 -46.
9. BBC (2012) *Dollar benchmark: The rise of the $1-a-day statistic.* Available from http://www.bbc.co.uk/news/magazine-17312819
10. The figures used here come from: Steve Wilcox (2003)*Can work: Can't buy. Local measures of the ability of working households to become home owners*York: Joseph Rowntree Foundation
11. A. Kleinmann (2012) Thanksgiving dinner cost a LOT more 100 years ago, *Huffington Post*, 21 November. Available from http://www.huffingtonpost.com/2012/11/21/thanksgiving-cost-100-years-agon2170620.html
12. Accidents reconnus par l'Association d'assurance contre les accidents, section agricole et forestiére 1960 - 2010, *Statistiques Luxembourg.* Available from http://www.statistiques.public.lu/stat/TableViewer/tableViewHTML.aspx?sCSChosenLang=fr&ReportId=587

Table 2 Sales of digital and physical books in the UK
(2009-2011)(millions)

2009	2010	2011	
Digital	114	158	243
Physical	3053	3115	2967
TOTAL	3167	3273	3210

Table 3 Cost of Thanksgiving ingredients, 1911 and 2012 (US dollars)

	Actual price in 1911 (US dollars)	Actual price in 2012 (US dollars)
16 pounds of Turkey	4.48	22.23
1 dozen eggs	0.4	1.79
1 pound of cranberries	0.13	2.45
3 pounds of sweet potatoes	0.08	3.15
1 pound of peas	0.05	3.15
5 pounds of sugar	0.3	3.1
5 pounds of flour	0.18	2.2
Total	5.62	38.07

Table 4 Accidents in agriculture and forestry (Luxembourg)

Year	1960	1970	1980	1990	2000	2009
Number of accidents	3515	2185	1580	1676	762	289

CHAPTER 7: CONCLUSIONS THOUGH EVIDENCE

1 Conclusions through evidence

The first part of this textbook concerned descriptive statistics. The second part introduces inferential statistics. Used in everyday language inference means "The action or process of inferring; the drawing of a conclusion from known or assumed facts or statements". Statistical inference is where we draw this conclusion from quantitative facts or statements. To get an idea about inference we will start with examples of non-statistical inference.

1. All men are mortal (an observed and accepted fact)
2. Socrates was a man (an observed and accepted fact)
3. Therefore, Socrates was mortal (a conclusion based on the fact that Socrates was a man and all men are mortal).

So from two accepted facts about men (that they are mortal) and Socrates (that he was a man) we can come to the conclusion that Socrates was mortal. However sometimes correct facts or observations can lead to wrong conclusions when poor inference is made.

1. Margaret Thatcher was Prime Minister of the UK (an observed and accepted fact)
2. Margaret Thatcher was a woman (an observed and accepted fact)
3. Therefore all Prime Ministers of the UK are women (a conclusion based on the fact that Margaret Thatcher was Prime Minister of the UK and that Margaret Thatcher is a woman).

False premises can lead to false conclusions

1. Hillary Clinton is a woman (an observed and accepted fact)
2. Women are not allowed to be President of the USA (a false statement)
3. Therefore Hillary Clinton is not allowed to become President of the USA (a false conclusion from one correct and one incorrect premise).

To complicate matters further false premises can logically lead us to a true conclusion.

1. Gordon Brown was Foreign Secretary when Tony Blair was Prime Minister (a false statement)
2. Foreign Secretaries always become Prime Minister (a false statement)
3. Therefore Gordon Brown became Prime Minister (a true statement logically derived from two false statements).

Statistical inference is when we draw conclusions based on quantitative data. All the above examples can apply to quantitative data.

1. Smokers are more likely to get lung cancer than non-smokers (a true observation)
2. People die of lung cancer (a true observation)
3. Smokers are more likely to die of lung cancer than non-smokers (a true conclusion based on true premises).

With same data we might infer:

1. Smokers are more likely to get lung cancer than non-smokers (a true observation)
2. People die of lung cancer (a true observation)
3. Non-smokers are less likely to die than smokers (a false statement because everybody dies eventually whether they smoke or not).

False premises, false conclusion:

1. 93% of people in the UK speak French fluently
2. School pupils in the UK learn French at a very young age.
3. The vast majority of people in UK are fluent in French because they start at such a young age.

In order to make accurate conclusions we must understand what has gone into creating our statistics and think logically. First, we will examine the concept of probability, a key concept in drawing conclusions from statistics.

2 Introducing probability theory

The statistical tests explored in this section of the book are based on probability theory. We use the idea of probability is everyday speech, for example “It will probably snow tonight”, “Manchester United will probably win today” or “I am probably not going to be at the party tomorrow”. In this context when we say probably we mean that it is highly likely but not definite or certain. You may wake up in the morning and find that it did not snow, or that Manchester United lost or that you can be at the party after all.

We might also use numbers when we talk about probability in everyday speech. We might say, “There is an 80% chance of snow tonight”, “I'm 90% sure that Manchester United will win” (in other words it is highly likely these events will occur) or “There's only a 10% chance of me being at the party tomorrow”—it is highly unlikely that I will be at the party.

We are also familiar with probability in games of chance. For example when I toss a coin will it land on heads or tails? The coin cannot land on heads and tails. There are two possible outcomes, (we are assuming that the coin will not land on its side). The coin toss is a game of chance and heads or tails are equally likely.

Table 1 : Four possible outcomes for guessing heads or tails

The coin lands on heads	The coin lands on tails	
You call heads	You were right	You were wrong
You call tails	You were wrong.	You were right

We calculate the probability as:

Number of possible outcomes The coin has two possible outcomes (head and tails) so we take 1 and divide it by 2.

½ or 0.5 or 50% or 1 in 2.

Or say we wanted to get know the probability of throwing a six on a 6-sided die: The die has six possible outcomes (1, 2, 3, 4, 5, 6) so we take 1 and divide it by six.

=0.167 or 16.7% or 1 in 6.

⅙

If we wanted to calculate the probability of two coins both being heads we multiply the two probabilities together: c
0.5 (probability of Coin 1 being heads)
0.5 (probability of Coin 2 being heads)
= 0.25 or 25% or 1 in 4

Similarly if we want to know the probability of throwing 2 sixes on two dice. c
0.167 (probability of Die 1 being a six)
0.167 (probability of Die 2 being a six)
= 0.028 or 2.8% or 1 in 36

When the number of possible outcomes is small you can simply count the number (as we did for the two-sided coin and six-sided die). As the numbers become bigger this is more difficult to do, but it is actually easy to calculate the number of possible outcomes. How do we know what the number of possible outcomes for three heads is? We take the number of sides on each coin (2) and multiple it by itself 3 times

2×2×2=8

This will usually be written as 2^3 So with three coins the chances of us getting three heads is 1 in 8. Similarly if we want to work out the changes of getting 3 sixes with 3 dice. We take the number of sides on each die (6) and multiply it by itself three times. 6×6×6 or 6^3=216. The chance of throwing three sixes with three dice is 1 in 216. We can extend the same principle indefinitely.

What about getting all sixes on six dice?
6×6×6×6×6×6 *or* 6_6 =
46,656 The chance of throwing six sixes is 1 in 46,656.

3 Why probability theory matters

You might be wondering what applying probability to random events such as dice throws and coin tosses has got to do with the sorts of statistics we might use in humanities research. We make inferences from quantitative data on the basis of probability. So if we see that wheat prices go up when oat prices go up and wheat prices go down when oat prices go down it would suggest that there probably is some sort of relationship between the prices of these two commodities that isn't just a coincidence of down to chance. Inferential statistics enable us to explore whether there is a relationship and what the nature of this relationship is (this will be explored in later chapters).
When we perform a statistical test we are looking to see if there are differences between two or more groups of data or that there is no difference. For example we may wish to investigate whether there are differences in how men and women perform on an exam or whether soldiers from Town A are taller than soldiers from Town B. Scientists may wish to find out whether treatment A is a better treatment than treatment B.

So why do we perform statistical tests? Why don't we just compare the average height of a soldier for Town A with the average height of solider from Town B? If we are trying to see if a new treatment for an illness is better than an established one, why don't we just see how many people are cured by Treatment A and how many by treatment B?

4 Dice throws and coin tosses

This is where our discussion of probability comes in. Let us return to our dice; we established that the probability of throwing all sixes on six dice is 1 in 46,656. In other words if I throw six dice 46,656 times I would expect, on average, that I would get six sixes on just one occasion. I have no plans to throw six dice 46,656 times but if I did would I get six sixes just once? Well I don't know for sure - I might get six sixes three, four, five or even ten times. I might not get any six sixes at all. What the probability is telling us is that if I was to keep on throwing the dice millions of times six sixes would come up an average of once every 46,656 times.

To illustrate this point I have tossed a coin ten times (Figure 1).. The chances of getting heads are 1 in 2 or 50%. On average I would expect that Heads would come up six times in my twelve coin tosses.

On this occasion I have eight heads and four tails. Suppose my research question was to find out how often you get heads when you toss a coin. I've tossed my coin twelve times and have got eight heads and four tails. Can I conclude that the chances of getting a head is actually eight in twelve or 66.6% or two-thirds? No, because this is just what happened on the one occasion I tossed my coin twelve times. If I toss the coin a hundred times or a thousand times would I get an equal number of heads and tails? Probably not exactly and I may have runs where I get ten heads in a row or ten tails but the more times I toss the coin the closer I will get to 50% heads and 50% tails.

50 coin tosses resulting in 24 heads and 26 tails.

I have just been to the website www.random.org where I tossed 50 virtual coins (Figure 1).

On this occasion I have 24 heads and 26 tails, that is 48% heads and 52% tails, much closer to 50% than my throw of ten heads.

I remember my grandfather talking about a contest in which members of the public would pay £1 and if they could throw six sixes on a dice they would win a brand new car. The chances of throwing six sixes are 1 in 46,656. The organisers would have been

Figure 1 Twelve coin tosses: eight heads and four tails

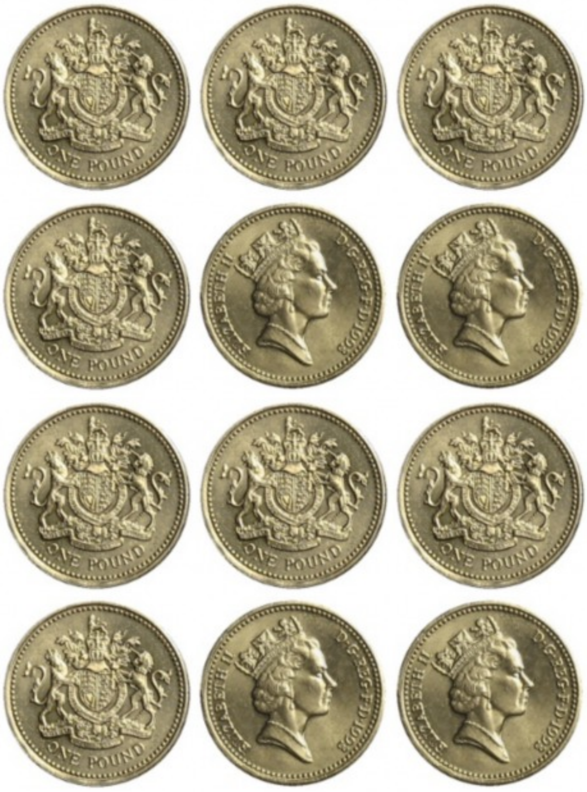
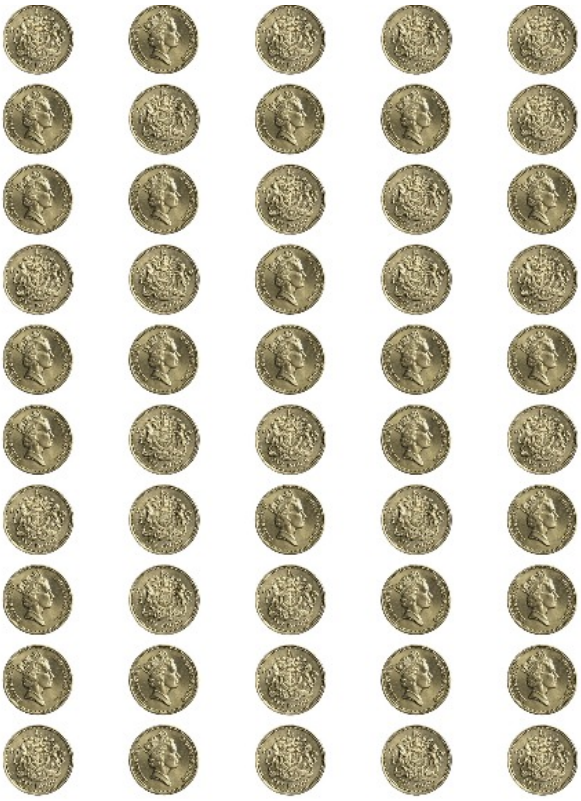


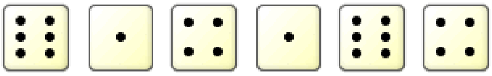
Figure 2 50 coin tosses resulting in 24 heads and 26 tails.



confident if they would have collected a lot of money and more than paid for the car before someone won the prize. However on this occasion one of the first people to throw the dice threw six sixes and the car was won leaving the organisers with a big loss. The point here is that when we collect data we don't know where our data stands. I have just visited the website www.random.org again

and this time I have rolled six dice. I rolled the numbers 6, 1, 4, 1, 6 and 1.

Figure 3 A throw of six dice



If I was now to tell you that whenever you throw six dice you will always get the numbers 6, 1, 4, 1, 6 and 1 you would object and say that I only got those numbers on the basis of one throw and that if I threw the dice again I may get a different result. When we collect our data it is a bit like one throw of a dice. We need to know how our one throw of the dice fits in with all the possible outcomes which could occur. We want to know if it is likely that any differences we see are real or are down to chance. This is the underlying principle of all the tests than follow.

5 A real life example

In one of the examples we look at the height of conscripts from three towns in Bavaria. The average heights of the men from each town vary, but is this variation down to chance or are men in some towns really taller than men from other towns? I previously introduced the normal distribution (Chapter 4). If we were to collect the heights of millions of men and plot them we would find that they would roughly follow the normal distribution. You can observe this in everyday life by looking at the heights of people around you. A small number of people are very short, a small number are very tall but a vast majority of people are around about average, some a little shorter, some a little taller maybe, but still quite average.

Suppose you came down to earth from another planet and had never seen a human being before. You capture a man and decide, amongst other things, to measure him. The human being they happen to capture is Sultan Kosen, believed to be the tallest man in the world at present. You then report back to your home planet that Earth is inhabited by life forms who are 251cm tall (8 foot 3 inches) tall. As a human, experience tells you that Sultan Kosen is an outlier, he is as far from average as exists, but as a being from another planet your sample of one has led you to conclude that humans are considerably taller than they really are. As someone who only seen one human being you don't know the real distribution of human heights.

So when you take a sample of data you don't always know for sure whether your sample mean is the same as the overall population mean, close to it, or three or four standard deviations away. You would never draw any conclusions from a sample of one of course but you feel justified to do so from a sample of 100, 1,000, 10,000 or 100,000. But you don't know how representative your sample is of the whole population from which you collected your sample. This is something we must always be aware of when designing studies which use statistical tests. We will discuss who to deal with this challenge in later chapters.

6 Exercises

1. Rank the following achievements in the order that a person born today is likely to achieve them. (It is not important to know the actual probabilities)
    1. Live until the age of 100
    2. Swim the English Channel.
    3. Go to university
    4. Be taller than their same sex parent.

CHAPTER 8: KEY CONCEPTS IN STATISTICS

1 Introduction

Statistical tests such as chi-square, t-tests, and F-tests (more about these later) enable us to make a statement about our sample based on probability. There are a number of important concepts which underlie most of these tests.

2 The null hypothesis

Each time we conduct a statistical test we start with two hypotheses:

1. Any differences we see are down to chance and there is no real difference between two or more groups of data. Statisticians call this the null hypothesis. You might see this written as H_0.
2. The other possibility is that there is really a difference between two or more groups of data. Statisticians call this the alternative hypothesis. You might see this written as H_1. When concluding that the alternative hypothesis is likely is to be correct it is said that we have rejected the null hypothesis.

3 Confidence

There is another dimension to accepting or rejecting the null hypothesis and that is the confidence limits. Are we 50% sure that we can reject the null hypothesis? 90% sure? 95% sure? 99.9% sure? When we accept or reject the null hypothesis we also state the confidence level at which this is done.

For example we might state that the null hypothesis was rejected at 95% confidence. What we are effectively saying in this case is that if we repeated our research and took a different sample from the same population, 19 times out of 20 we would find that the differences we observed were statistically significant (not down to chance). When you perform statistical tests you need to check your answer against the tables which appear in the appendices to this book and see whether your result is significant at your chosen level of confidence. Instructions on using these tables appear in the next chapters. The question of what level of confidence we should set when performing statistical tests is given relatively little attention and the answer is that it depends. 95% is commonly used by scientists and social scientists, but there is no particular reason why 95% should always be chosen. To illustrate this point think of a few possible situations from real life.

Suppose you are in a country which executes people convicted of murder. You are on a jury which has been given the task to decide whether or not the defendant is guilty of the crime. If he is found guilty he will sent for execution. How certain would you need to be that he is guilty? What about 95% sure? Are you willing to see a person executed when there is a 1 in 20 chance he might be innocent? What about 99%? Are you willing to see this man executed where there is a 1 in 100 chance he is innocent? I am assuming at this point that you want to be pretty sure he was guilty.

Take another example. You suffer from a medical condition which is not life threatening but makes your life difficult. A new treatment comes out which experiments conclude helps around 10% of people who suffer from this condition and there are no known side effects. There is a nine in ten chance that the treatment will not work for you, so would you take the treatment? There might be other considerations such as cost, but I imagine most people would see a ten percent chance of improvement would seem to be worthwhile in this situation.

So why don't we just say set our confidence level high at 99% or 99.9% so that we are always very sure our conclusion is correct? If we set confidence level close to certainty then the burden of proof is so high and our differences would need to be so big that we would hardly ever reject the null hypothesis. If we set of confidence levels very low then we will being rejecting the null hypothesis all the time and we will be saying that anything is significant.

4 Significance

Are Methodists more likely to identify themselves as 'evangelical' than Anglicans? Are people more likely to remember a list of long words than short words? Is the difference between the gross takings of the top 10 Warner Bros films and the top 10 Paramount films significant? Almost all the examples of data discussed in this book are drawn from a sample (see Chapter 5). We do not usually have the time or resources to sample a whole population. The full data we need may be unavailable to us, be restricted, lost or destroyed. Therefore when we see a pattern in any sample data, we need to ask ourselves, could this have happened by chance?

Suppose we wish to know the life expectancy of people living in England in the 1830s. We do not have time to go through every single burial record in England and work out what the life expectancy is. If we wished to know the age at which people got married in 1830 we do not have the time or resources to go through all the records that exist.
If we wish to do a survey of British people's attitudes to learning languages we cannot ask every person in the UK. When a newspaper commissions an opinion poll asking people how they would vote if there were a general election tomorrow the pollsters cannot ask everybody of voting age how they will vote- they have to take a sample.

The question therefore arises, how reliable are our findings? Are the relationships, patterns and observations we make representative of the population as whole? If we repeated the same survey or took a different sample of historical data would we get the same outcomes again? And what would happen if we repeated our research ten, fifty, one hundred times?

Figure 1: A lottery ticket. The odds of picking the correct six numbers from 49 in the UK's National Lottery are 13,983,815 to 1

In statistics *significance* means that the observations we have made are *unlikely* to have occurred by chance. We need to note the following:

- Unlike in everyday speech, in statistics the word significance is not the same thing as importance. Just because a piece of data is statistically significant does not mean that it is important. We have to decide this for ourselves.
- The word 'unlikely' is important and we need to remember that unlikely events do occur. If you go out today and buy a ticket for the UK National Lottery you have about a 1 in 14,000,000 chance of winning. It is *unlikely* that you will win, but it is *possible* that you will win. And as unlikely as it is somebody wins almost every week. Unlikely is not the same thing as never.
- Identifying a relationship as statistically significant does not explain how that relationship has come about or why.

5 Critical values

So how do we know whether we should accept or reject the null hypothesis after we have performed a particular statistical test? As you will see, each calculation you do leads to a number. For example in the chi-square example (see chapter 10) the value of Chi-square comes out at 875.1. What does 875.1 mean? 875.1 what? Alone this statistic doesn't really mean anything. The final step of most statistical tests is to look up our answer in the critical values tables. There are different critical values for different tests. Essentially the 'critical value' marks the boundary between significance and non-significance at a chosen level of confidence. If the answer exceeds the critical value then we can reject the null hypothesis and conclude that the relationship has not occurred by chance.

6 Degrees of freedom

The 'partner' of critical values is Degrees of Freedom, sometimes abbreviated to DF. The degrees of freedom is easy to calculate, but quite difficult to understand. Degrees of freedom are essentially the number of values which are free to vary.

Suppose we have a dataset with just five numbers. In this case we only know four of them and that the overall mean is 20. The numbers we know.

10, 25, 30, 15, ?

We will refer to this '?' As 'x'.

So to have a mean of 20

$$\frac{10+25+30+15+x}{5}=20$$

We can add together the numbers in our dataset and call this y.

So 10+25+30+15+x = y

As we know out mean is 20 and we have numbers in our dataset his means that $y \div 5 = 20$

So what is the value of y? We know that an unknown number divided by 5=20.

We can rearrange the equation here

$$20 \times 5 = 100$$

So 10 + 25 + 30 + 15 + x = 100

To find x here we can take all the other numbers away from 100

$$100 - 10 - 25 - 30 - 15 = 20$$

Therefore now we know that $x = 20$ as

$$\frac{10+25+30+15+20}{5} = \frac{100}{5} = 20$$

Now suppose we only have the following information about the above dataset:

1. We have 5 numbers n = 5

2. One of these numbers is 20

3. The mean of the 5 numbers is 20, so

$$\frac{x_1 + x_2 + x_3 + x_4 + 20}{5} = \frac{100}{5} = 20$$

So what is the value of x_1, x_2, x_3 and x_4? On the basis of the information we have we don't know but
$x_1 + x_2 + x_3 + x_4 = 80$ because:

$x_1 + x_2 + x_3 + x_4 + 20 = 100$

And

100 − 20 = 80.

Additionally 100 ÷ 5 is not the only way we can get a mean of 20 from five numbers.

Although we know for sure that one of the numbers is 20 we cannot be sure what the other four numbers are: For example all of the following are possible:

$$\frac{15+15+25+25+20}{5} = \frac{100}{5} = 20$$

$$\frac{20+20+20+20+20}{5} = \frac{100}{5} = 20$$

$$\frac{30+10+30+10+20}{5} = \frac{100}{5} = 20$$

Of all our numbers four are free to vary and one (20) is not free to vary. Therefore we can say that we have four degrees of freedom. [1]

7 Another illustration

Suppose you have an ordinary pack of 52 playing cards. You pick one at random and it happens to be the Ace of Spades. There are now 51 playing cards left in the pack. If you pick another card at random there are now 51 possibilities. It cannot be the Ace of Spades as you have already picked it. So we could say that the pack of cards has 51 possible opportunities to vary or 51 degrees of freedom. Next we might pick out the two of diamonds. now there are only 50 cards left in the pack and 50 opportunities to vary. By the time we have picked the 51st card there will be just one card left and by process of deduction we know which card it is, and there will be no freedom to vary (Degrees of Freedom=0). The calculations of the Degrees of Freedom will be discussed in each example.

8 One and two tail tests

If you look at the critical values table (Appendices) you will see reference to one tail tests and two tails test. Going back to our normal distribution diagram we can see that 95% of normally distributed data lies within 2 standards deviations of the mean. In the case of a two-tailed test this means that the critical value is taking into account the 2.5% at each end of the normal distribution (Figure 2). However, in a one-tailed test (Figure 3) the 5% is all put into one tail (in this case the left tail).

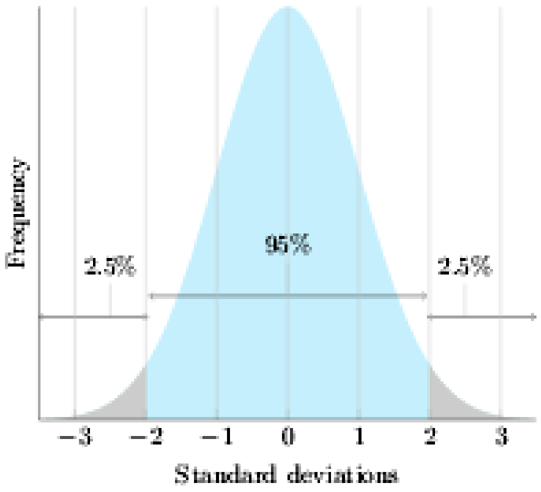
So when do we use a one-tail test and when do we use a two-tail test? Suppose that you want to find out whether 6 year olds will do better than 5-year-olds on a particular test. You get two classes of children, one of 6 year-olds (Blue class) and one of 5 year-olds (Green class) to sit the same examination. Blue class get better marks than Green class and you want to see if this difference is significant- in other words, could it have occurred by chance? Critical values for a two-tailed test take into account the possibility that 5 year-olds might perform better than 6-year olds. However, as it would be logical to expect that 6-year-olds would perform better than 5-year olds you might decide to do a one tail test instead. With a one-tailed test all the probability of non-significance is loaded into one side of the curve. By doing this you are more likely to reject your null hypothesis, but at the expense of ruling out the possibility that 5-year-olds perform better than six-year-olds.

9 Two types of error

As we are dealing with probabilities there is always a chance that our conclusion will be wrong. If you were to tell me that you had just bought a ticket for this week's UK National Lottery I can say with a very strong degree of confidence that you won't win the jackpot and that you have a 14 million to one chance of winning. But despite my confidence that you won't win the jackpot there is still a chance that you could. Winning the lottery is very unlikely but almost every week someone does win. *Unlikely* is not the same thing as *impossible.* Whatever levels of confidence we set for ourselves there are two possible errors we could make:

1. We could reject the null hypothesis when it is actually true. In other words we have said a relationship exists when it does not. Statisticians call this a Type I error. If you are being tested for a medical condition and the test reveals that you have the condition, but in reality you do not have the condition, this is a Type I error. You may have heard of this being called a false positive.

Figure 2 Two-tail test



2. We could accept the null hypothesis when it is actually false. In this case we have said that there is no relationship when in fact there is a relationship. Statisticians called this a Type II error. In the medical test scenario this means you are told you have not got the medical condition when in fact you have, a false negative.

The consequences of these errors can be very serious. You will probably be very worried if you are told you have a medical condition when you don't (a Type I error) or even worse, you don't have a medical condition when you do (Type II error). In most of the cases we will be examining the consequences of a Type I or Type II error are not so serious, but we must be aware of the possibility that they may happen.

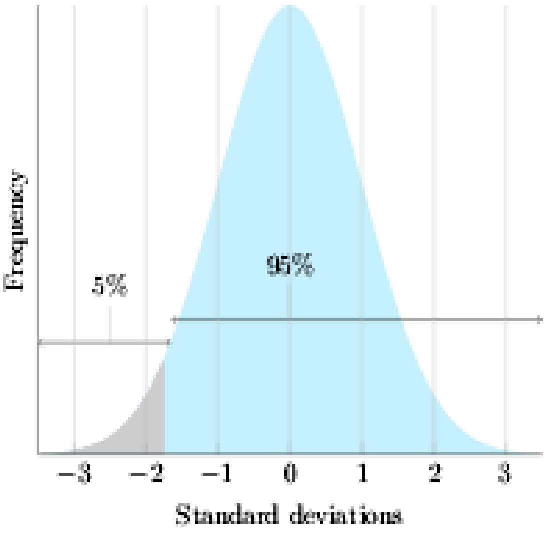
10 p-values, alpha, beta

After performing a statistical test it is helpful to present a summarising table of results. You will often see in these a p-value, an alpha value (*α*) and sometimes a beta value (*β*). Put simply a statistical test can tell us whether a result is statistically significant or not. p-values, alpha and beta indicate the scale of the significance.

11 p-values

A p-value may be reported as an exact figure, e.g. 0.6751 or a more rounded value alongside a 'less than' (<) sign, e.g. < 0.05. The confusing aspect about p-values are that they are related to confidence levels (e.g. 95%) to determine statistical significance, but are not the same thing as confidence levels. For example, if I wish to determine whether or not a result is significant at 95% my starting point is the 95%. There are two possibilities: 1) accept the null hypothesis, or 2) reject the null hypothesis. So the answer to the question, "Is the result statistically significant?" the answer is either *No* (accept null hypothesis) or *Yes* (reject null hypothesis). Therefore deciding whether a particular test has given a statistical significant answer puts every test into one of two categories, either 'yes' or 'no'. In contrast the p-value gives us an idea of the scale of the statistical significance. There are multiple ways of calculating a p-value. Exact p-values can calculated, but this is beyond the scope of this book.

Figure 3 One tail test



The simplest way to calculate a p-value is to look at the critical values tables of the appropriate test. Table 4 shows the critical values of chi-square at p-values of 0.1, .05 and 0.01. Notice that these correspond to confidence levels of 90%, 95% and 99%, but they are not the same thing. Suppose that my Chi-square statistic is 2.706 with 1 degree of freedom and I wish to know whether this is statistically significant at a 95% confidence level. I look at the table under 95% confidence and find the that the critical value of Chi-square at 1 Degree or Freedom is 3.841 which is a higher value than 2.706. My results for the Chi-square would need to be higher than the critical value in order to accept it. Therefore the answer to the question "Is a chi-square statistic of 2.706 with 1 degree of freedom significant at 95%?", the answer is 'no'. In another case I find that my Chi-square result comes out at 11.345 with 1 degree of freedom. I ask the question "Is a chi-square statistics of 11.345 with 1 degree of freedom significant at 95%?", the answer is 'Yes' as this figure exceeds the critical value required to be significant at 95%. However, if we have a chi-square statistic of 12.76 with 1 degree of freedom was I could look on the critical values table under 1 degree of freedom and find that this value is more than the 11.345 required for a p-value of 0.01. Therefore, instead of the having to go through a complicated process of finding an exact p-value I can write p-value< 0.01.

Table 1 Advantages and disadvantages of one-tail and two-tail tests

'	Advantages	Disadvantages
One-tailed test	A positive effect is more likely to be identified	One possible direction of relationship is not tested, e.g. that 5 year olds might perform better than 6-years olds.
Two-tailed test	Allows for the possibility that the relationship could be in either direction.	Where the direction of the relationship is known with some certainty half the possible scope of significance is 'wasted'.

Table 2 Type I and Type II errors and the null hypothesis

Example of Type I and Type II errors (medical test)		
	In reality you don't have the disease	In reality you do have the disease
Tests shows you have the disease	Accurate	A Type I error
Tests shows you don't have the disease	A Type II error	Accurate

Table 3 Type I and Type II errors

Your research says the null hypothesis is true	Accurate	A Type I error
Your research says the null hypothesis is true	A Type II error	Accurate

12 Caution about p-values

It seems natural to suppose that the closer your p-value is to zero the lower the probability that you have rejected the null hypothesis when you should have accepted it. However, the size of your sample impacts on the p-value. P decreases as sample size increases. Whilst this might appear to be sensible to be more confident with a larger sample size, if a sample size is large enough almost anything will be statistically significant.

13 Alpha and Beta

Alpha (*α*) and Beta (*β*) are measures of the probability of Type I and Type II errors. The probability of a Type I error is referred to as Alpha and a Type II error as Beta.

Alpha and beta relate to the power of a statistic and address some of the concerns about the p-value, by accounting for either a Type I error or a Type II error. Most scientific experiments set alpha at 0.8 and beta at 0.2. Alpha plus beta adds up to 1.

14 Exercises

1. A survey of 100 history graduates showed that 59% of them earned more than £25,000 per year one year after graduation. Give a 95% confidence interval for the percentage of graduates who earn more than £25,000 per year.

15 References

[1] This example comes from 'ed1234ize' on youtube. http://www.youtube.com/watch?v=09fBxrzwUb8

Table 4 Critical values of Chi-Square at different p-values

Confidence	**90%**	**95%**	**99%**
p value	0.1	0.05	0.01
Degrees of freedom			
1	2.706	3.841	11.345
2	4.605	5.991	15.086
3	6.251	7.815	18.475
4	7.779	9.488	21.666
5	9.236	11.07	24.725

CHAPTER 9: USING STATISTICS INTELLIGENTLY

With the advent of computer software which can perform tests on thousands of pieces of data in seconds, the need to think about using statistical tests intelligently is greater than ever. A software package like SPSS or Minitab or even Microsoft Excel will undertake all manner of calculations not telling you whether your test is appropriate or what conclusions you might draw from your data.

There are thousands of statistical tests available for every situation. It is also useful to remember that Statistics itself is an academic discipline with its own debates, schools of thought and discussions. Not all statisticians will agree on the best test to use in every situation and similarly not everyone who uses tests, whether linguists, historians or religious studies practitioners will always agree either. In many cases there are different formulae for doing the same tests which may lead to slightly different answers (if you run the tests here through a computer programme you may find that you get a slightly different answer to the manual calculation).

1 Try to understand the purpose of the test

In the tests in this book I have outlined when it might be appropriate to use this test, when this test should definitely not be used, the assumptions of the test and whether there is anything in particular we need to be careful about.

2 Work through the tests manually

It seems slightly unfair to suggest that you should spend a half-an-hour working through an example manually rather than just putting the data into a software programme and getting the answer in seconds, but the manual approach will help you to see how the test works and, hopefully, help you make better sense of interpreting your example.

3 Don't suspend your understanding of your subject

Your studies of history, linguistics, religion, music, archaeology, film (or other discipline) are not suspended just because there are statistics involved. Statistics is as much an art as a science and the evidence needs to be weighed up just like any other kind of evidence. Like any evidence, quantitative data is subject to bias, can be inaccurate and can be incomplete which can lead us to wrong conclusions. The data we have is not always the best available. Also just like any other evidence, statistical evidence needs to be situated in the broader context of the area of study.

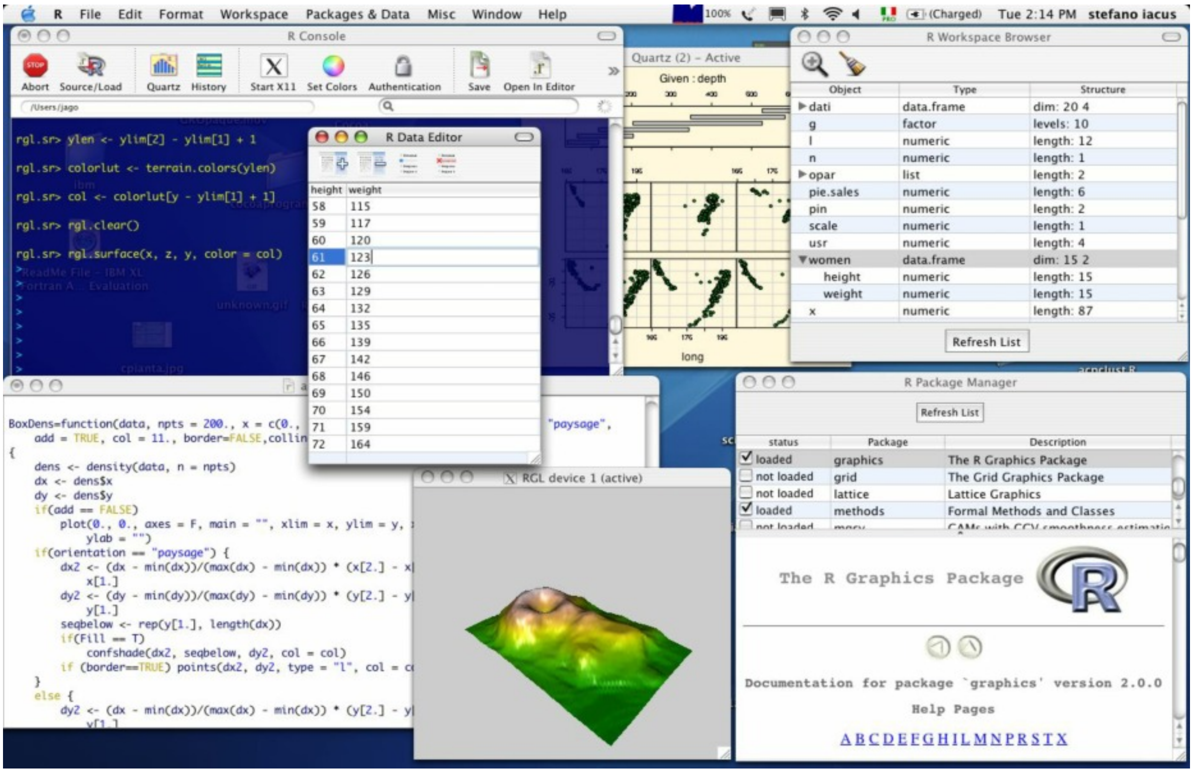
4 Don't confuse significance with importance

This is a key lesson — just because a relationship is found to be significant does not mean that it is important. This point has been made very strongly Stephen Ziliak and Deirdre McCloskey in their fascinating book *The Cult of Statistical Significance* [1] where they critique the obsession of economists with significance without any thought to the importance of the argument or the actual size (or power) of the relationship. Does X have a big effect on Y or a small one? Does this medicine make you a lot better or just a little bit better? Is Beethoven a lot more popular than Mozart or just a little bit more popular? I have used a number of different examples in this book, but there is no statistical test to work out whether anyone cares how tall Bavarian soldiers were in the early nineteenth century, whether Beethoven is more popular than Mozart with Classic FM listeners or whether it is really important that people can remember short words more easily than long words. Whether or not these possible differences are important or not is a matter for human judgement, not for statistics.

5 Garbage In Garbage Out (GIGO)

GIGO was a phrase coined early in the development of computers. It served as a warning that a computer will do what you tell it to do. A computer won't refuse to perform a Mann-Whitney U test on the

Figure 1: A screenshot showing the some of the capabilities of the open source statistical analysis package R. http://www.r-project.org/



```
rgl.sr> ylen <- ylim[2] - ylim[1] + 1

rgl.sr> colorlut <- terrain.colors(ylen)

rgl.sr> col <- colorlut[y - ylim[1] + 1]

rgl.sr> rgl.clear()

rgl.sr> rgl.surface(x, z, y, color = col)
>
>
>
>
>
>
>
>
```

height	weight
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164

```
BoxDens=function(data, npts = 200., x = c(0.,                "paysage",
    add = TRUE, col = 11., border=FALSE,collin
{
    dens <- density(data, n = npts)
    dx <- dens$x
    dy <- dens$y
    if(add == FALSE)
        plot(0., 0., axes = F, main = "", xlim = x, ylim = y,
            ylab = "")
    if(orientation == "paysage") {
        dx2 <- (dx - min(dx))/(max(dx) - min(dx)) * (x[2.] - x
            x[1.]
        dy2 <- (dy - min(dy))/(max(dy) - min(dy)) * (y[2.] - y
            y[1.]
        seqbelow <- rep(y[1.], length(dx))
        if(Fill == T)
            confshade(dx2, seqbelow, dy2, col = col)
        if (border==TRUE) points(dx2, dy2, type = "l", col = c
    }
    else {
        dy2 <- (dx - min(dx))/(max(dx) - min(dx)) * (y[2.] - y
            y[1.]
```

Object	Type	Structure
▶ dati	data.frame	dim: 20 4
g	factor	levels: 10
l	numeric	length: 12
n	numeric	length: 1
▶ opar	list	length: 2
pie.sales	numeric	length: 6
pin	numeric	length: 2
scale	numeric	length: 1
usr	numeric	length: 4
▼ women	data.frame	dim: 15 2
height	numeric	length: 15
weight	numeric	length: 15
x	numeric	length: 87

status	Package	Description
☑ loaded	graphics	The R Graphics Package
□ not loaded	grid	The Grid Graphics Package
□ not loaded	lattice	Lattice Graphics
☑ loaded	methods	Formal Methods and Classes
□ not loaded	mgcv	GAMs with GCV smoothness estimatio

Documentation for package `graphics' version 2.0.0

grounds that it is not an appropriate test for your data, it won't warn you that the agricultural statistics from twelfth century England probably vary regionally or that the census enumerator put the decimal point in the wrong place. Statistical tests do not bring new evidence into being —comes out of them depends on what goes in. If you put garbage in you get garbage out whether you are using a computer or calculating manually.

6 Using data imaginatively

If you can't find any data available for what you want to measure think about what might be a substitute. We might not be able to get data on the price of bread, but there is data on the price of wheat. If we wanted to investigate health in a particular place or particular point in history we may not be able to find direct data but knowing the heights of conscripts might be useful. Entitlement to free school meals is often used to identify deprivation in schools. Numbers of windows or hearths declared for window tax or hearth tax might give an idea about the size of a house or the wealth of its occupants in the absence of more direct data. These are all examples of proxies. The concept of a 'proxy variable' can be very useful in statistics as they help us when no actual data for the variable is available.

7 Exercises

From 1662 to 1689 each household in England had to pay tax according to the number of hearths (fireplaces) they had in their house. What information might we be able to deduce about households from the Hearth Tax returns apart from the number of hearths they have in their houses?

The proportion of children receiving free school meals (FSM) is often used as an inductor of deprivation for a school or a geographical area. Outline the possible advantages and limitations of FSM as a measure of deprivation.

8 References

[1] Ziliak,S. and D. McCloskey (2007) The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives(University of Michigan Press)

CHAPTER 10: COMPARING GROUPS: THE CHI-SQUARE TEST

Different groups, difference choices?

A primary school offers its pupils the opportunity to learn either the flute or the clarinet. In this ideal situation pupils have a free choice of either flute or clarinet and no pupil is made to play the flute when they want to play the clarinet or vice versa. The music teacher notices that girls seem to choose the two instruments in equal numbers. However the boys seem to have a preference for the clarinet. She wants to know if difference in choice is i) real and ii) significant.

In this example there are two variables. The first variable is the choice of musical instrument. There are two possible options a) flute or b) clarinet. The second variable is the gender of the pupil. Again, there are two possible options: a) boys and b) girls. As there are two variables and two possible options for each variable there are four possible combinations (2 x 2= 4). These groups are:

1. Boys who choose the flute. The music teacher finds that there are 10 pupils in this group.
2. Boys who choose the clarinet. The music teacher finds that there are 15 pupils in this group.
3. Girls who choose the flute. The music teacher finds that there are 20 pupils in this group.
4. Girls who choose the clarinet. The music teacher finds that there are 20 pupils in this group.

We can represent this as a 2 x 2 table (Table 1).

Table 1: Choices of flute or clarinet made by 25 boys and 40 girls

	Choose flute	Choose Clarinet
Boys	10	15
Girls	20	20

So is there a gender difference in the choice of instrument? We can see that girls chose between the two instruments in equal numbers: 20 girls chose each instrument. However, only ten out of the 25 boys (40%) chose to play the flute. All things being equal we would expect boys and girls to choose the flute and clarinet in equal numbers. From our data we might conclude that boys are more likely to choose the clarinet than the flute. But is this a satisfactory conclusion? Would this same pattern being seen again in the future? Is it relevant that there are more girls than boys choosing to play an instrument and more clarinettists than flautists.
The next section will be an introduction to the Chi-square test.

The Chi-square test

The Chi-square (or χ^2) test is used to identify whether or not the distributions of categorical variables are different. A chi-square test will reveal whether or not the choices of flute and clarinet in Table 1 are independent of gender or not. The test does this by comparing the actual observations (numbers) with those we would expect if boys and girls were equally likely to choose flute and clarinet. The 1989 church census explored how Methodists and Anglicans described their `churchmanship' (their theological beliefs), either as liberal or evangelical. Therefore there are four possible groups of clergy.

1. Methodists who identify as evangelical
2. Anglicans who identify as evangelical
3. Methodists who identify as liberal.
4. Anglicans who identify as liberal.

But is denomination significant to whether a churchgoer identifies as evangelical or liberal or is it likely to be a matter of chance? To find out we can use the Chi-square statistic. It is called after the Greek letter χ`or chi' (pronounced `kai'). It is important not to confuse this with the Roman letter x . The following is the formula for the Chi-square statistic.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

How to calculate the observed values

The observed values (*Oi*) are simply the values we have observed – in this case the values reported in Table 2

Table 2 Anglican and Methodist clergy and their identification as liberal or evangelical.

	Methodist	Anglican
Evangelical	125900	292200
Liberal	81900	223300

1. A is the Total number of Evangelicals
2. B is the Total number of Liberals
3. C is the Total number of Methodists
4. D is the Total number of Anglicans
5. E is the total number of clergy altogether

Calculating the chi-square

Expected values of Evangelical Methodists

Which is

$$\frac{\text{Total number of Liberals}(B) \times \text{Total number of Methodists } (C)}{\text{Total number of clergy}(E)}$$

$$\frac{418100 \times 207800}{723300} = 120{,}117.8$$

Table 3 Table showing total numbers of Anglicans, Methodists, Liberals and Evangelicals

	Methodist	Anglican	
Evangelical	125900	292200	418100 (A)
Liberal	81900	223300	305200 (B)
	207800 (C)	515500 (D)	723300 (E)

Expected values of Evangelical Anglicans

$$\frac{\text{Total number of Evangelicals } (A) \times \text{Total number of Anglicans } (D)}{\text{Total number of clergy } (E)}$$

Which is

$$\frac{418100 \times 515500}{723300} = 297{,}982.2$$

Expected values of Liberal Methodists

$$\frac{\text{Total number of Liberals}(B) \times \text{Total number of Methodists } (C)}{\text{Total number of clergy}(E)}$$

$$\frac{305200 \times 207800}{723300} = 87{,}682.2$$

Which is

$$\frac{\text{Total number of Liberals}(B) \times \text{Total number of Anglicans } (D)}{\text{Total number of clergy}(E)}$$

Expected values of Liberal Anglicans

Which is

$$\frac{305200 \times 515500}{723300} = 217{,}517.7$$

Table 4: Chi-square table

'	Observed values (O)	Expected Values (E)	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Evangelical Methodists	125900	120117.8	5782.2	33433836.84	278.3421
Evangelical Anglicans	292200	287982.2	4217.8	17789836.84	61.77408
Liberal Methodists	81900	87682.2	-5782.2	33433836.84	381.307
Liberal Anglicans	223300	217517.8	5782.2	33433836.84	153.7062
TOTALS			10000	118091347.400000006	875.1294

$$\frac{(O-E)^2}{E}$$

Then we needed to added up the four values of to get our chi-square statistic.
278.3+61.8+381.3+153.7=875.1

Chi-square = 875.1

Calculating the degrees of freedom

On its own a Chi statistic does not mean anything. We have two further tasks to undertake before we can judge whether similarities and differences in churchmanship in our observed sample are real or are down to chance.
First we need to calculate the Degrees of Freedom (DF). Fortunately this is easy to do:

We have two columns (Methodist and Anglican) and two rows (Evangelical and Liberal, so (2−1)×(2−1)=1

$$(\text{Number of columns - 1}) \times (\text{Number of rows - 1})$$

Looking up the critical values

Secondly we need to look up in the critical values for Chi-square which is on our table (see Table 5). We will choose a confidence level of 95%). The critical Chi-square value for 1 DF at 0.05 is 3.841. You can find more critical values of Chi-square in appendix H.

Presenting the results

Our chi-square is well above this figure so we can be confident that there is an association between domination (Anglican or Methodist) and churchmanship (Liberal or Evangelical).

Advantages of using chi-square

The method for calculating chi-square can be used for any number of categories and groups. In this example we used a 2 x 2 table (two denominations and two forms of churchmanship), making four squares or **cells**. If there were two forms of churchmanship and five denominations there would be ten cells (2 × 5) The method can be used for any number of cells. The chi-square method is a good way of identifying association between numerical variables.

Table 5 Table of critical values for Chi-square at 95%. The full table can be found in the appendices

Chi-square	Degrees of freedom	p-value
875.1	1	0.001

Table 6 Summary of Chi-square results

Degrees of freedom	Critical Value
2	5.99146
3	7.81473
4	9.48773
5	11.0705
6	12.59159
7	14.06714
8	15.50731

Table 8 Deaths by accidents in coal mines 1851-53: by region

	Explosions	Roof-fall	Other accidents	Total accident
Scotland	26	36	33	96
Northumberland. Durham, Cumberland	38	47	64	149
Yorkshire, Derbyshire, Nottingham, Warwickshire, Leicestershire	36	33	43	112
Lancashire, Cheshire, North Wales	86	58	72	216
Staffordshire, Shropshire, Yorkshire	36	120	75	231
South Wales, Monmouthshire, Gloucestershire, Somerset	43	59	66	168
All	266	352	353	971

Cautions when using chi-square

If more than 20% of the cells return expected values below 5 then the chi-square statistic may not be valid. Actual numbers must be used. Do not use percentages, proportions, ratios or averages. The chi-square can identify association, but it does not indicate the strength of the association or the reasons for the association. The same individual observation cannot be in more than one group. For example for the purpose of this test no clergyman can be both liberal AND evangelical or Anglican AND Methodist.

Exercises

1. Calculate the χ^2 statistic for the boys and girls choosing flute and clarinet in Table 1.
2. Examine Table 8 showing deaths in coals mines in different regions of Britain. [1]
3. Calculate the χ^2 statistic for the different regions and accident types. What can you infer from the Chi-square statistic?
4. Table 7 shows the number of acting roles for men and for women according to the gender of the playwright. [2] Are female playwrights more likely to write parts for women actors than male playwrights?

Table 7 Gender of contemporary playwrights and

Roles for women	Roles for men	
Women playwrights	75	78
Men playwrights	128	216

References

1. P E H Hair (1968)Mortality from Violence in British Coal-Mines, 1800-50, *Economic History Review*, New Series 21, pp. 545-561

2. Elizabeth Freestone (2012) Women in theatre: how the '2:1 problem' breaks down. *The Guardian* http://www.guardian.co.uk/news/datablog/2012/dec/10/women-in-theatre-research-full-results

CHAPTER 11: COMPARING TWO GROUPS: THE STUDENT'S T-TEST

1 Comparing two groups: the Student's t-test

In the King James translation of the Bible, the fourth chapter of the book of Malachi has an average sentence length of 20.3 words. The New International Version (NIV) translation of the same passage of has an average sentence length of 16.1 words. In other words the average sentence length of the King James version is 4.2 words longer than NIV. But are these sentence length differences statistically different? Alone the difference in the means does not tell us very much. Figure 1, Figure 2, Figure 3 show that distributions with the same differences in means can be very different. Figure 2 has low variability. The distributions hardly overlap at all. On the other hand Figure 4 has high variability. The distributions overlap considerably. In the case of Figure 3 about half of the distributions over lap each other. Therefore in order to make any statement about the statistical difference we need to consider , not only the means, but also the distributions.

The Student's t-test was devised by William Sealy Gossett, an employee of the Guinness brewery in Dublin. Guinness was so concerned about its employees giving away trade secrets that his employer only allowed him to publish his findings anonymously – all his writings appear under his pen name 'Student'. [1] As an employee of the brewery Gossett had to make statistical judgements based on very small samples.

In the introduction to the this section I noted that most statistical tests are based on where your observed statistics sit in relation to the overall population, which is presumed to fit a normal distribution. Gossett saw two main problems with these sorts of statistical tests when your sample is very small.

If your sample is small then you have no way of knowing how widely your sample differs from the population as a whole. Whatever experiments Gossett was doing at Guinness he was not able to repeat his experiments thousands of times in order to get a sense of what the true mean, median and mode of his experiments might have been. As well as being unable to get a sense of the true mean, median and mode, he was also unable to know for sure what the overall distribution looked like. He didn't know for sure whether his population

Figure 1 Medium variability

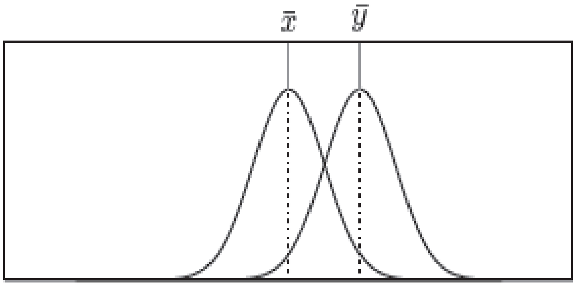


Figure 2 Low variability

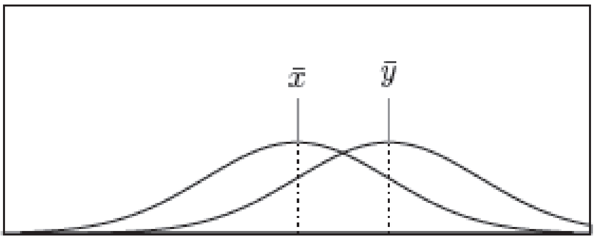


Figure 3 High variability

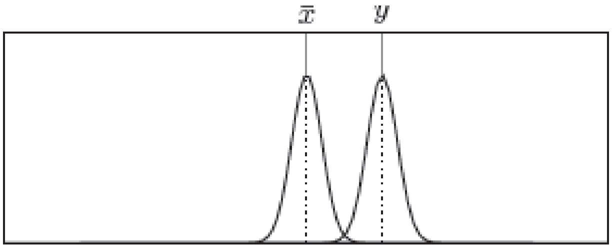


Figure 4: William Sealey Gossett (1876-1937)

was normally distributed, skewed positively, skewed negatively, or had some other distribution.

Therefore he needed to devise a reliable test for small samples which was not based on assumptions governing large samples. He devised two forms of what became known as the student t-tests, a non-paired test and a paired test.

Both tests are asking one main question. Are the means of two samples significantly different from each other or could any differences have occurred by chance?

1.1 When to use a t-test

A t-test is a useful test when you have two samples and want to know if the differences between the means are significant. It is useful for small samples. If you have a sample larger than about 30 you should use an alternative such as the F-test. If you have a small sample, but your data are ranked rather than absolute, use the Mann-Whitney U test.

1.2 Assumptions made

All the observations in your sample must be independent of one another. In other words, t-test should not be used where the same observations appear in both the samples or in which the data from Sample A are derived from Sample B or vice versa.

Below, in Table 1 and Table 2 are listed the top-grossing 2011 films from two distributors, Paramount and Warner Bros. The top Warner Bros film grossed more than the top Paramount film and the second placed Warner Bros film grosses more than the second placed Paramount film. However the tenth placed Paramount film grossed more than the sixth place Warner Bros. But are the mean gross takings of the two companies significantly different? Calculating the t-test statistics involves a lot of steps which are outlined here.

1.3 Calculating an independent sample t-test: worked example

Stage 1: Calculate the mean gross takings of the Paramount films by adding them together and dividing by the number of observations (in this case 10)

$$\frac{350.6+180.1+175.8+164.4+145.0+126.4+122.9+103.5+84.1}{10}=159.3$$

Now do the same for the Warner Bros films

$$\frac{371.1+253.2+136.2+116.9+116.0+83.9+75.3+71.7+63.3+60.7}{10}=135.6$$

Stage 2: Now square each of the Paramount gross takings (multiply each film's taking by itself), then add them together:

$$350.6^2+180.1^2+175.8^2+164.4^2+145.0^2+$$
$$126.4^2+122.9^2+103.5^2+84.1^2=179{,}976.9$$

Then the same for the Warner Bros films:

$$371.1^2+253.2^2+136.2^2+116.9^2+116.0^2+$$
$$8.9^2+75.3^2+71.7^2+63.3^2+60.7^2$$
$$=279{,}039.1$$

Stage 3: Square the total scores for each of the film companies:

For Paramount:

$$350.6+180.1+175.8+164.4+145.0$$
$$+126.4+122.9+103.5+84.1=1593$$

$$1593^2=2{,}538{,}397$$

Table 1: Top ten grossing films (Paramount)

Rank	Film	Millions gross (US dollars)
1	Transformers: Dark of the Moon	350.6
2	Thor	180.1
3	Captain America: The First Avenger	175.8
4	Kung Fu Panda 2	164.4
5	Puss in Boots	145
6	Mission: Impossible - Ghost Protocol	140.5
7	Super 8	126.4
8	Rango	122.9
9	Paranormal Activity 3	103.5
10	True Grit	84.1

Then for Warner Bros

71.1+253.2+136.2+116.9+ 116.0 +83.9 +75.3+71.7+63.3+60.7 = 1365

$$1365^2 = 1,839,622$$

Stage 4: Divide the square totals by the number of films in each group

For Paramount

$$\frac{2,538,622}{10} = 253,839.7$$

For Warner Bros

$$\frac{1,839,622}{10} = 183,992.2$$

Table 2: Top ten grossing films (Warner Bros)

Rank	Film	Millions gross (US dollars)
1	Harry Potter and the Deathly Hallows: Part II	379.1
2	The Hangover Part II	253.2
3	Sherlock Holmes: A Game of Shadows	136.2
4	Horrible Bosses	116.9
5	Green Lantern	116
6	Crazy, Stupid, Love	83.9
7	Contagion	75.3
8	Dolphin Tale	71.7
9	Unknown	63.4
10	Happy Feet Two	60.7

Stage 5: For each group subtract the Stage 4 result from the Stage 2 result.

For Paramount1

$$279,039.1 - 183962.2 = 95,076$$

For Warner Bros

$$179,976.9 - 253,839.7 = -73,862.8$$

Stage 6: Add together the two results from stage 5

$$95,076.9 + \ -73,862.8 = 21214.1$$

Stage 7: Subtract 1 from each of the number of films in each group

For Paramount 10-1=9

For Warner Bros 10-1=9

Stage 8: Add the two results from stage 7 together

9+9=18

Stage 9: Divide the result from Stage 6 by the result from Stage 8

$$\frac{21214.1}{18}=1178.6$$

Stage 10: Find the reciprocal of the number of scores for Paramount and Warner Bros.

The reciprocal is 1 divided by a number. We have ten scores for both Paramount and Warner Bros for we need to divide 1 by 10.

For Paramount

$$\frac{1}{10}=0.1$$

For Warner Bros

$$\frac{1}{10}=0.1$$

Stage 11: Add the score from stage 10 together

0.1+0.1=0.2

Stage 12: multiply the result for stage 11 by the result for Stage 9

$$0.2\times 1178=5893$$

Stage 13: Calculate the square root of stage 12

$$\sqrt{5}893=76.8$$

Stage 14: Calculate the difference between the mean of Paramount Takings and the mean of the Warner Bros Takings.

159.3-135.6=23.7

Stage 15: Take the result of Stage 13 and divide it by the result of Stage 14 to find your t-statistic.

$$\frac{76.8}{23.7}$$

Therefore *t =3.27*

We now need to work out whether the difference in the average gross takings from the two companies are significant or not.

First we need to calculate the degrees of freedom:

Therefore

DF=(10-1) + (10-1)=18

At 18 DF our critical value for t at 0.05 level of significance is 2.101. Our t result is 3.27 which is greater than 2.101. Therefore we can conclude that there is a significant difference between the gross takings of the two film companies. Our results are summarised in Figure 5.

Figure 5 Results of t-test

Degrees of freedom	t	Significant at 95%?	p-value
18	3.27	Yes	<0.005

In the example above both groups contained ten films. It is not necessary for both groups to contain the same number of items for the t-test to work.

2 A paired t-test

Another version of the t-test is a **paired t-test**. In the case of paired t-test we look at pairs of observations which are somehow connected with each other. We might be comparing the results the same students got on different exams or whether the income of the group of graduates measured today and five years ago. The paired t-test is often used to assess change in some form.

In a cognitive linguistics experiment participants are given a list of words to memorise. We can use a paired t-test to see whether each person is more likely to remember long or short words.

Stage 1: Draw up Table 3 to calculate the differences and the differences squared.

Table 3 List of words to memorise

Person number	Short words	Long words	Difference	Difference squared
1	4	4	0	0
2	8	5	3	9
3	9	6	3	9
4	6	4	2	4
5	6	5	1	1
6	9	6	3	9
			Total differences 12	Total of differences squared 32

Stage 2: Now we square the total difference

$$12^2 = 144$$

Stage 3: Divide the total difference squared by the number of pairs

$$\frac{144}{6} = 24$$

Stage 4: Subtract the Stage 3 result from the Total of the differences squared.

32-24=8

Stage 5: Multiply the number of pairs by the number of pairs -1

$$6 \times (6-1) = 6 \times 5 = 30$$

Stage 6 Divide the result of Stage 4 by the result of stage 5

$$\frac{8}{30} = 0.267$$

Stage 7: Calculate the square root for Stage 6

$$\sqrt{0.267} = 0.516$$

Stage 8: Divide the result of Stage 3 by the result of Stage 7

Therefore t = 46.5

Calculating the Degrees of Freedom

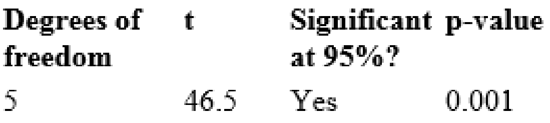
$$\frac{24}{0.516} = 46.5$$

Degrees of freedom = numbers of pairs -1

6-1=5

At 0.5 significance the critical value is 2.571. We can see that our t value exceeds that, therefore, we can say there is a significant difference between the subjects' ability to remember long words and short words. The results are summarised in Figure 6.

Figure 6 Results of paired t-test



Degrees of freedom	t	Significant at 95%?	p-value
5	46.5	Yes	0.001

Table 4 Number of words per sentence in two bible translations of Malachi 4

Sentence	New International Version	King James Version
1	5	11
2	6	13
3	24	23
4	18	31
5	11	7
6	7	24
7	20	23
8	19	20
9	19	31
10	32	

Table 5 Earnings at graduation and after five years for 7 individuals

Graduate	Earnings at graduation (thousands)	Earnings 5 years after graduation (thousands)
1	20	25
2	32	35
3	33	40
4	42	42
5	15	19
6	9	16
7	18	24

Table 6 Ages of bride and groom marrying in St Elizabeth of Hungary, District of Aspell, Wigan, 1883

Groom	Bride
24	21
25	18
20	21
25	22
31	32
27	22
22	27
22	28
35	32
25	21
21	25
23	24
20	18
25	22
22	21

Table 7 Levels of Strontium in parts per million by Gender

Male	Female
121	114
93	133
52	76
57	113
72	104
51	130
102	66
84	114
67	104
	105

Excercises

Use an independent t-test to find the t-statistic of the data in Table 4.

Use a paired t-test to find the t-statistic for the data in Table 5.

Explain why a paired t-test is unsuitable for the data in Table 4.

A university course is assessed by 50% exam and 50% coursework. You have each student's mark for the coursework and their final mark overall. Is this data suitable for any sort of t-test? Explain your answer.

Use a paired t-test to calculate the t-statistic for the ages of the bride and grooms in Table 6.[2]

Use an independent t-test to find the t-statistic of the data in Table 7. [3]

References

1. Student (1908), The Probable Error of a Mean *Biometrika* 6, 1–25.
2. Available online at http://www.lan-opc.org.uk/Wigan/Aspull/stelizabeth/index.html
3. C. Chenery et al (2010)Strontium and stable isotype evidence for diet and mobility in Roman Gloucester, UK. Journal of Archaeological Science 37, pp.150-163

CHAPTER 12: ANALYSIS OF VARIANCE

1 Introduction

Analysis of Variance (or ANOVA) compares the means of two or more samples. It is called Analysis of Variance rather that Analysis of Means because it analyses variance in order to make inferences about the mean. ANOVA examines both differences within samples and difference between samples. ANOVA leads to a single number known as F. If F is found to be significant we then we need to undertake a post-hoc test such as the Tukey Honestly Significant Difference (or Tukey HSD) test. Where there are two fairly small sample samples (fewer than 30 observation a T-Test may be used instead.) Where there are more than three samples ANOVA must used. You should never use multiple t-tests where there are three or more samples.

2 Calculating the Analysis of Variance (ANOVA): A worked example of the F test

For this test we will be examining the heights of 30 conscript soldiers from three different towns in Barvaria —Toelz, Reichenhall and Friedberg (see Table 2).

1. The null hypothesis is that the samples are taken from a common population.
2. The alternative hypothesis is that the samples were taken from populations with different means.
3. If the samples come from the same population we would expect that there would be more variation within the samples than between them. In other words we would expect to find the same sorts of differences in each of the three towns.
4. However if there is more variability between the towns than there is within the soldier heights within each town this would indicate significant differences between the population

The ANOVA involves three main calculations

1. The estimate of the variance within each sample
2. The estimate of variance between samples
3. The estimate of the variance between each sample divided by the estimate of variance within samples gives us the F ratio.

Additionally we need to know the following for both the estimates between and within samples:

1. Total number of samples which will be labelled as k. k = 3 (because there are three towns)
2. Total number of individuals which will be labelled as N. There are 10 individuals in each of the three towns. 10+10+10=30 so N=30

Table 1: Heights of 30 conscripts (in cm) from Toelz, Reichenhall and Friedberg

Toelz	Reichenhall	Friedberg
165.4	168	148.4
178.6	180.8	158.1
178.8	166.4	171.3
177.6	166.8	155.7
171.1	163	166.6
170.5	155.9	175.1
166.4	177.6	172.9
169.8	163	153.2
165.4	172.5	158.1
181	160.7	163

Table 2: Table to calculate estimates within sample

Toelz x̄ =172.46			**Reichenhall x̄ =167.47**			**Friedberg x̄ =162.24**		
x	$x-\bar{x}$	$(\bar{x}-x)^2$	x	$x-\bar{x}$	$(\bar{x}-x)^2$	x	$x-\bar{x}$	$(\bar{x}-x)^2$
165.4	-7.06	49.84	168	0.53	0.28	148.4	-13.84	191.55
178.6	6.14	37.7	180.8	13.33	177.69	158.1	-4.14	17.14
178.8	6.34	40.2	166.4	-1.07	1.14	171.3	9.06	82.08
177.6	5.14	26.42	166.8	-0.67	0.45	155.7	-6.54	42.77
171.1	-1.36	1.85	163	-4.47	19.98	166.6	4.36	19.01
170.5	-1.96	3.84	155.9	-11.57	133.86	175.1	12.86	165.38
166.4	-6.06	36.72	177.6	10.13	102.62	172.9	10.66	113.64
169.8	-2.66	7.08	163	-4.47	19.98	153.2	-9.04	81.72□
165.4	-7.06	49.84	172.5	5.03	25.3	158.1	-4.14	17.14
181	8.54	72.93	160.7	-6.77	45.83	163	0.76	0.58
		326.424			527.14			731

1. The mean of each of the three samples (towns).
2. Grand mean which is the mean height of all the individuals in the three samples. We will label this as $\bar{x}_G$

$$\bar{x}_G(Grand\ mean)=\frac{1724.6+1674.7+1622.39}{30}=167.39$$

3. The ANOVA involves three main calculations:

For a refresher on the concept of variance and the standard deviation, see Chapter 3.

Additionally we need to know the following for both the estimates between and within samples:

1. Total number of samples which will be labelled as *k*. k = 3 (because there are three towns)

1. Total number of individuals which will be labelled as *N*. There are 10 individuals in each of the three towns. 10+10+10=30 so *N*=30

1. The mean of each of the three samples (towns).

Toelz

$$\frac{Sum\ of\ heights=1724.6}{10}=172.46$$

Reichenhall

We call the calculation here an **estimate** because we do not know the height distribution of the whole population of the soldiers in the three towns. Instead we make an estimate of the mean and variance (which is the standard deviation squared) of the whole population, based on the sample we have here. The symbol for the standard deviation of the population rather than a sample of is the lower case Greek letter sigma *σ*. We use a `hat' over the *σ*, ($\hat{\sigma}$) to show that

this is a estimate and not the known standard deviation of the population as a whole. As the variance is the standard deviation squared the symbol for the estimate variance of the population is $\hat{\sigma}^2$.

Friedberg

$$\frac{Sum\ of\ \ heights = 1622.4}{10} = 162.24$$

1. Grand mean which is the mean height of all the individuals in the three samples. We will label this as $\bar{x}_G$

$$\frac{1724.6+1674.7+1622.39}{30} = 167.39(\bar{x}_G)$$

3 Stage 1: Find the estimated variance from within samples

The equation for this is

$$\hat{\sigma}_W^2 = \frac{\sum^{k}\sum^{n}(x-\bar{x})^2}{N-k}$$

k = the number of samples
n = Number of observations in each sample
x̄ The mean of each sample.

$\bar{x}_G$The Grand mean. (The mean of all the observations in all the the samples)

$\hat{\sigma}_W^2$ Estimated variance within samples

The best way to calculate the variance within the samples is to draw a table (see Table 2). We need to subtract the mean of each sample from each observation $x-\bar{x}$ then square the result $(x-\bar{x})^2$.

Now calculate the Sum of the Squares.

$$\frac{326.40+527.14+731.00}{30-3} = \frac{1584.57}{27} = 58.69$$

Stage 2: Estimate the variance between samples

$$\hat{\sigma}_B^2 = \frac{\sum^{k} n(\bar{x}-\bar{x}_G)^2}{k-1}$$

σ_W^2 = Estimated variance between samples

k = the number of samples
n = Number of observations in each sample

x̄ The mean of each sample.

$\bar{x}_G$ The Grand mean. (The mean of all the observations in all the samples)

Toelz

x̄=172.46

n=10

10(172.46−167.39)²=257.05

Reichenhall

x̄=167.47

n=10

10(167.47−167.39)²=0.06

Friedberg

x̄=172.46

n=10

$10(162.24-167.39)^2=257.05$

Stage 3: We then insert these figures into the equation:

$$\frac{257.05+0.06+265.23}{3-1}=\frac{522.23}{2}=261.17$$

Stage 4: Calculate the degrees of freedom.

Degrees of freedom for between samples variance estimate = 2

Degrees of freedom for within samples variance estimate = 27

We can record all these calculations in table form (see

Table 3 Results summary table

'	Variance estimate	Degrees of freedom
Between samples	$\hat{\sigma}_B^2=261.17$	K-1=2
Within samples	$\hat{\sigma}_W^2=58.69$	N-k=27

Table 3)

Stage 5: Now we can calculate the F-ratio.

$$F\ ratio=\frac{between\ samples\ variance\ estimate}{within\ samples\ variance\ estimate}$$

$$=\frac{261.17}{58.69}=4.45$$

We now need to look up our 4.45 on our critical values table.

At the 0.01 Level the critical value is 5.49. This is more than 4.45 so we do not reject the null hypothesis that there is no significant difference between the three samples. However at 0.05 significance level the critical value of F is 3.35. At this choice of significance level we reject the null hypothesis and accept that there seems to be some difference between the three groups.

Tukey Honestly Significant Difference (or Tukey HSD)

Having rejected our null hypothesis and accepted that there is some variation between the groups we need to do some further exploration. We have variation between groups, but which variations are actually significant? Are the differences between Toelz and Reichenhall more variable than the differences between Toelz and Friedberg between Reichenhall and Friedberg? Moreover, a non-significant F-value can disguise significant relationships between individual samples.

We can use the Tukey Honestly Significant Difference (or Tukey HSD) test to explore the differences between the towns:

We need the following data from our Analysis of Variance:

1. The means from each of the groups (towns)
   1. Toelz: $\bar{x}=172.46$
   2. Reichenhall: $\bar{x}=167.47$
   3. Friedberg: $\bar{x}=162.16$
2. The Within Sample variance estimate (also known as Mean Squares Within) = 58.69
3. The number of individuals in each group = 10

Table 4 Results summary

Source	Degrees of Freedom	Sum of Squares	Mean square	Value of F	Probability	Interpretation
Three towns	2	522.3	261.15	4.45	1	Significant, at p 0.05
Residual	27	1584.6	58.689			
Total (N = 30)	29	2106.9				

4. The number of degrees of freedom N-k. N= Total number of individuals - Number of Groups k): 30 individuals - 3 groups = 27.

$$TukeyHSD = q \times \sqrt{\frac{MSw}{n}}$$

MSw Within Sample variance estimate (also known as Mean Squares Within (or Residual Mean Square)
n Number of observations in each sample.
q The Critical value of Q.

$$TukeyHSD = 3.48 \times \sqrt{\frac{58.689}{10}}$$

The formula for the Tukey HSD test is

$$TukeyHSD = 3.48 \times \sqrt{5.8689} = 8.431$$

To calculate q:

We need to refer to our Critical values for the Tukey Q test (Table 5). The critical values of Q for 27 degrees of freedom (N- k) (at a significance of 0.05) and 3 groups is between 3.48 (30 degrees of freedom) and 3.58 (20 degrees of freedom). We will use 3.48 as our q value as 30 is fairly close to 27.

Our HSD value is 8.431.

We now compare the differences between each pair of means. If the differences between any two means are greater than the HSD value of 8.431 we can conclude that these differences are significant at 95%.

The differences between the mean heights of conscripts from Toelz and Reichenhall and between Reichenhall and Toelz are less than our HSD value of 8.431, but the difference between the means of Toelz and Freidberg is 10.22cm so we should reject the null hypothesis. Rather than finding significance `between the towns' we have found a significant difference between two of the towns Toelz and Freidberg.

Four or more samples

There is no limit to the number of samples for which AVOVA can be used. If there are four samples there are five possible combinations of sample.

1. Sample 1 and Sample 2
2. Sample 1 and Sample 3
3. Sample 1 and Sample 4
4. Sample 2 and Sample 3
5. Sample 3 and Sample 4

Discussion

The ANOVA and the subsequent Tukey HSD Test take a long time to calculate manually, but a statistical analysis software package will perform these calculations in a matter of seconds.

The ANOVA F-test can be used for any number of groups and not all the groups have to have the same number of observations, though the Tukey HSD does require equal group sizes. The F-test can tell us whether or not there are differences between the group.

If our ANOVA of three or more samples reveals that the differences are significant we need to perform a post-hoc test (such as Tukey HSD) to explore which differences are significant. Conversely a non-

Table 5: Critical values of Q (extract)

Degrees of freedom	Number of means	
	2	3
1	18	27
2	6.09	8.33
3	4.5	5.91
4	3.93	5.04
5	3.64	4.6
6	3.46	4.34
7	3.34	4.16
8	3.26	4.04
9	3.2	3.95
10	3.15	3.88
11	3.11	3.82
12	3.08	3.77
13	3.06	3.73
14	3.03	3.7
15	3.01	3.67
16	3	3.65
17	2.98	3.62
18	2.97	3.61
19	2.96	3.59
20	2.95	3.58
30	2.89	3.48
40	2.86	3.44

Table 5 Bestselling fiction books of all time (UK sales only)

Book title	Author	Sales	Genre
Da Vinci Code,The	Brown, Dan	5,094,805	Crime, Thriller and Adventure
Harry Potter and the Deathly Hallows	Rowling, J.K.	4,475,152	Children's Fiction
Harry Potter and the Philosopher's Stone	Rowling, J.K.	4,200,654	Children's Fiction
Harry Potter and the Order of the Phoenix	Rowling, J.K.	4,179,479	Children's Fiction
Fifty Shades of Grey	James, E. L.	3,758,936	Romance and Sagas
Harry Potter and the Goblet of Fire	Rowling, J.K.	3,583,215	Children's Fiction
Harry Potter and the Chamber of Secrets	Rowling, J.K.	3,484,047	Children's Fiction
Harry Potter and the Prisoner of Azkaban	Rowling, J.K.	3,377,906	Children's Fiction
Angels and Demons	Brown, Dan	3,193,946	Crime, Thriller and Adventure
Harry Potter and the Half-blood Prince:Children's Edition	Rowling, J.K.	2,950,264	Children's Fiction
Fifty Shades Darker	James, E. L.	2,479,784	Romance and Sagas
Twilight	Meyer, Stephenie	2,315,405	Young Adult Fiction
Girl with the Dragon Tattoo,The:Millennium Trilogy	Larsson, Stieg	2,233,570	Crime, Thriller and Adventure
Fifty Shades Freed	James, E. L.	2,193,928	Romance and Sagas
Lost Symbol,The	Brown, Dan	2,183,031	Crime, Thriller and Adventure
New Moon	Meyer, Stephenie	2,152,737	Young Adult Fiction
Deception Point	Brown, Dan	2,062,145	Crime, Thriller and Adventure
Eclipse	Meyer, Stephenie	2,052,876	Young Adult Fiction
Lovely Bones,The	Sebold, Alice	2,005,598	General and Literary Fiction
Curious Incident of the Dog in the Night-time,The	Haddon, Mark	1,979,552	General and Literary Fiction
Digital Fortress	Brown, Dan	1,928,900	Crime, Thriller and Adventure
Girl Who Played with Fire,The:Millennium Trilogy	Larsson, Stieg	1,814,784	Crime, Thriller and Adventure
Breaking Dawn	Meyer, Stephenie	1,787,118	Young Adult Fiction
Kite Runner,The	Hosseini, Khaled	1,629,119	General and Literary Fiction
One Day	Nicholls, David	1,616,068	General and Literary Fiction
Thousand Splendid Suns,A	Hosseini, Khaled	1,583,992	General and Literary Fiction
Girl Who Kicked the Hornets' Nest,The:Millennium Trilogy	Larsson, Stieg	1,555,135	Crime, Thriller and Adventure
Time Traveler's Wife,The	Niffenegger, Audrey	1,546,886	General and Literary Fiction
Atonement	McEwan, Ian	1,539,428	General and Literary Fiction
Bridget Jones's Diary:A Novel	Fielding, Helen	1,508,205	General and Literary Fiction

significant F-value can disguise significant relationship between some of the samples. However, if we do find that there are differences we must perform further investigations using the Tukey HSD test or an alternative.

The F-test is much less likely to result in the null hypothesis being rejected when it is actually true (what statisticians call by Type I error) than performing multiple Student t-tests on the data.

Exercises

Table 5 displays the bestselling books of all time (UK sales only) by genre.[1] Perform an ANOVA of the data by genre.

References

[1] Derived from the Guardian datablog (2012) http://www.guardian.co.uk/news/datablog/2012/aug/09/best-selling-books-all-time-fifty-shades-grey-compare#data

CHAPTER 13: UNDERSTANDING RELATIONSHIPS

1 Making connections

When we use quantitative data we are often seeking to demonstrate that there is a link between one set of data and other. We might want to investigate what effect a major historical event had on the price of food or whether married men more use more words on a daily basis than their wives.

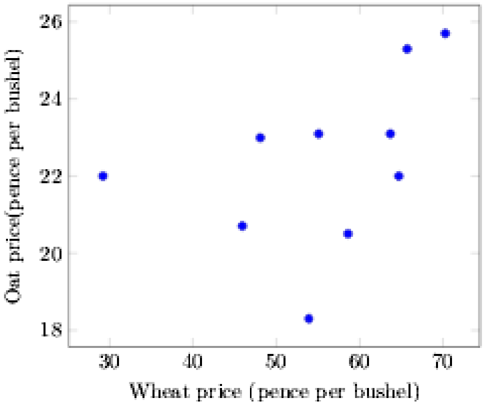
In Table 1 we have data about the price of wheat and the price of oats between 1830 and 1839. The prices were originally in pounds and shillings, but have been decimalised here for clarity. What is the relationship between wheat prices and oats prices. When wheat prices are high are oat prices high too? From the data alone is difficult to see for sure.

Table 1: Wheat and oats prices

Year	Wheat	Oats
1830	63.7	23.1
1831	65.7	25.3
1832	58.6	20.5
1833	53.9	18.3
1834	45.9	20.7
1835	29.2	22
1836	48.1	23
1837	55.1	23.1
1838	64.7	22.4
1839	70.3	25.7

On the graph in Figure 1 we have plotted the wheat price against the barley price for each year. We are see that there is a pattern of sorts as the numbers plots sort of line up. But how do we describe this pattern in more detail? One way is to calculate the Pearson product-moment correlation coefficient of the data. The Pearson product-moment correlation coefficient is a number between 1 and -1. This number is referred to as r or `Pearson's r'.

Figure 1: Wheat and oats prices, 1830-1839



2 Calculating the Pearson product-moment correlation coefficient

The method used to calculate the correlation coefficient [1])There are other correlation coefficients which are not Pearson product-moment correlation coefficients, e.g. The Spearman's rank). It is useful to create a scatterplot so that you have a rough idea of what the correlation co-efficient might be. You will also be able to identify outliers, that is an observation which does not seem to fit the overall pattern. Additionally you may find that your relationship is U-shaped or S shaped in which case there may be a relationship but the calculation of r will not reveal this. See Anscombe's Quartet for more about this

STAGE 1: Notice that we have calculated the mean of both the wheat prices (x) and the oat prices (y).

STAGE 2: We need to take the variances of wheat (column 7) and oats (column 8) and divide them by the number of observations:

Wheat

$$\frac{1326}{10} = 132.6$$

Oats

$$\frac{44.1}{10} = 4.41$$

STAGE 3: We then take the square roots to find the Standard Deviations

Wheat: = √218.8=11.52

Oats: = √44.1=2.10

We now have the four numbers we need to calculate r.

1. The SD of the wheat: 11.52

r = correlation coefficient. This will be number between -1 and +1.

x = the price of wheat

y = the price of oats

n = number of observations. In this case n=10 because there are ten pairs-- wheat and oat prices were observed each year.

SDx The standard deviation of the wheat

Sdy The standard deviation of oats

2. The SD of the oats: 2.10
3. The number of observations: 10 (note that this is 10 and not 20 as there are ten years of paired data).
4. The Sum of Column 6: 107.7

STAGE 4

We now have a value of r of 0.442. Your value of r will be between -1 and +1. If you have a value of r which is more than 1 or less than -1 you have made an error in your calculation. The value of the r helps us

$$r = \frac{\frac{1}{n}(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y}) + \ldots + \ldots (x_n - \bar{x})(y_n - \bar{y})}{SD_x SD_y}$$

Table 2: Table to calculate correlation coefficient

Year	x	x²	y	y²	xy
1830	63.7	4057.69	23.1	533.61	1471.47
1831	65.7	4316.49	25.3	640.09	1662.21
1832	58.6	3433.96	20.5	420.25	1201.3
1833	53.9	2905.21	18.3	334.89	986.37
1834	45.9	2106.81	20.7	428.49	950.13
1835	29.2	852.64	22	484	642.4
1836	48.1	2313.61	23	529	1106.3
1837	55.1	3036.01	23.1	533.61	1272.81
1838	64.7	4186.09	22.4	501.76	1449.28
1839	70.3	4942.09	25.7	660.49	1806.71
Sum of columns	ΣX= 555.2	ΣX²=32150.6	ΣY= 224.1	ΣY²=5066.19	ΣXY=12548.9

Figure 2 Karl Pearson 1857-1936. Pearson developed numerous statistical tests including the Pearson Product Moment Correlation Coefficient

to make a judgement about the strength of the association between wheat prices and oat prices.

2.2 Interpreting r

Notice that I am careful to use the word association rather than relationship. What we are looking for is to find the nature of the cause and effect in any association we might observe. It seems reasonable to conclude that there is a moderate association between wheat prices and oat prices. Table 3 is a useful guide to interpreting the r value.

2.3 Things to remember

Fortunately you can use a spreadsheet or a statistical package to work out the correlation co-efficient for a large amount of data.

The Correlation co-efficient is a bivariate test. This means that each observation has two parts. In this case each year has two (a pair of) prices -the price for wheat and the price for oats.

Very importantly the correlation measures association and not causation. We have been able to demonstrate a moderately weak association between wheat prices and oat prices but we cannot say that a rise in oat

Table 3 Interpreting value of r

r		How the association between would work between wheat prices and oat prices.
1	Perfect positive association	If r =1 Wheat and oat prices go up and down together in the same direction.
0.8	Strong positive association	
0.6		
0.4	Moderately weak positive association	The association between oat and wheat prices is r = 0.442
0.2		
0	No association	If r= 0 There is no association between wheat prices and oat prices.
-0.2		
-0.4	Moderately weak negative association	
-0.6		
-0.8	Strong negative association	
-1	Perfect negative association	If r = -1 Wheat prices go up when oat prices go down and oat prices go up when wheat prices go down.

prices is caused by a rise in wheat prices based on the correlation co-efficient alone. We will be addressing the issues of causation in Chapter 18.

3 Exercises

Calculate the correlation coefficient for the UK price and US price for the best-selling books in Table 4 .

Calculate the correlation coefficient for the distance and times taken in days. [2]

4 References

[1]The Pearson product-moment correlation coefficient is often known simply as the `Correlation coefficient' and is referred to as such in this chapter.

[2]This data comes from From Pliny the Elder (AD 23-AD 79), cited in Lionel Casson (1951) Speed of the sail of ancient ships Transactions of the American Philological Association 82, pp. 136-148

Table 4 Comparison of US and UK prices of best-selling fiction on major retailer's website

Title	Author	UK price	US price (converted to £)
Lover reborn	JR Ward	8.4	13.4
I've got you number	Sophie Kinsella	8	12.8
Betrayal	Danielle Steel	8	12.8
The Patchwork Heart	Jane Green	8.4	13.5
Lone Wolf	Jodi Picoult	9.2	14.7
The Thief	Clive Cussler	9.6	15.4
Death comes to Pemberly	P D James	10.3	16.5

Table 5 Distance and voyage length in the Roman Empire

Voyage	Distance (Nautical Miles)	Length of Voyage
Ostia-Africa	270	2
Messina-Alexandria	830	6
Ostia-Gibraltar	935	7
Ostia-Hispania Citerior	510	4
Messina-Alexandria	830	7
Ostia-Provincia Narbonensis	345	3
Puteoli-Alexandria	1000	9

CHAPTER 14: Predicting new observations from known data

1 Introduction

When it gets cold outside I turn on the central heating in my house. If I am using the heating then I am using more gas. Therefore there is a relationship between the temperature outside and the amount of gas I use. Suppose that each day I collect two sets of data: 1) the outside temperature and 2) the amount of gas I use. Over a period of time I will be able to predict the amount of gas I use just by taking the temperature.

Notice that this only works one way. A change the outdoor temperature affects the amount of gas I use, but using more or less gas will not increase or decrease the outside temperature.

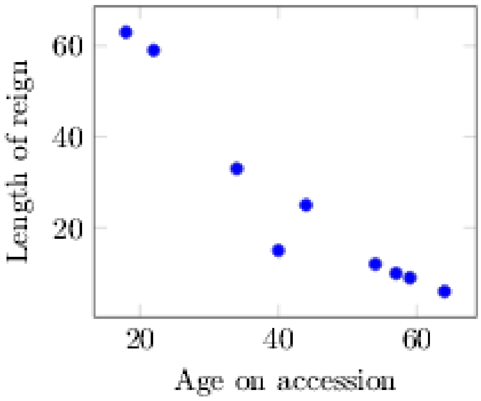


Figure 1 Relationship between age on accession and length of reign

2 Predicting reign length of British monarchs

shows the age that British monarchs were when they came to the throne and how long they reigned for. It would be reasonable to suppose that monarchs who came to the throne younger would have reigned longer and this seems to be the case. Only monarchs since George I in 1714 are included as no kings or queens since this time have been deposed or died violently. Edward VIII is excluded as he abdicated his throne and the present Queen Elizabeth II is excluded on the grounds that she is still alive.

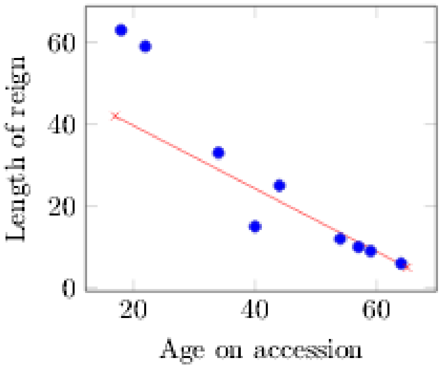
Table 1 Reign of British Monarchs 1714-1952

'	Age on Accession to throne	Reign
George I	54	12
George II	34	33
George III	22	59
George IV	57	10
William IV	64	6
Victoria	18	63
Edward VII	59	9
George V	44	25
George VI	40	15

'

If we plot these figures onto a scatter graph (Figure 1) we can see the relationship between age on accession and reign is negatively correlated. The older a person becomes king or queen the shorter their reign.

Figure 2 Relationship between age on accession and length of reign with approximate `good fit' line drawn by eye



We can see the rough pattern of the dots and draw a 'good fit' line. I have drawn such a line to create Figure 2.

3 Prediction

We can use the line to make a *prediction* about how long a monarch will reign if the only information we have is their age. We can do this by finding the age on the *x* axis, finding where it intersects the best-fit line, and reading off answer off the *y* axis. For example, We can see that a monarch who becomes king or queen at the age of 50 can expect to reign for 20 years. So by looking at a known pattern of data we can predict the value of one variable simply by knowing the value of one other variable. To use another example, if I have lots of data on the relationship between exam performance at school and exam performance at university I can use this data to predict the university exam performance of an individual student, simply by knowing about their performance at school.

4 Simple linear regression

I have drawn the line in Figure 2 by hand. If you were to perform this exercise you might put the line in a slightly different place. And because the line is in a different place you will get a slightly different answer when you made your prediction about the length of time a monarch will reign for. This section introduces a technique called **simple linear regression.** This technique is called simple linear regression to distinguish it from other forms of regression analysis. This is the only form of regression covered in this book. The simple regression analysis uses a mathematical technique to predict instead of trying to predict by eye.

5 Finding the regression line

The regression line is not a number, but a simple equation which describes the line of the best fit. It is most easily done using computer software, but it is useful to work through it manually to see how it works. Importantly it essential to understand the following:

We have designated age of accession to the throne as *x* and the length of reign as *y*. This is not a co-incidence. X is the information we have and Y is what we are trying to predict. It is important to get these the right way round. If we were looking at predicting how much electricity we are going use today based on our observation of the weather, the weather is our *x* variable and the gas bill (which we are trying to predict) is the *y*.

We will use this formula for simple linear regression.

$$y=mx+b$$

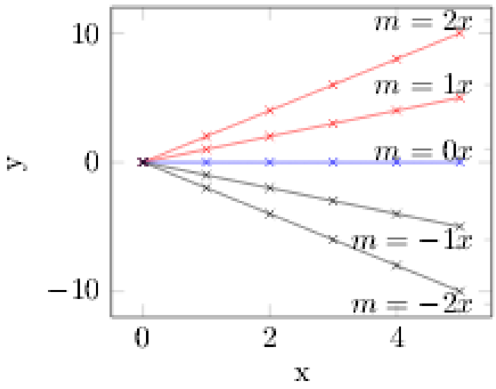
y is what we wish to find out. In the this case, the length of time a monarch reigned for.
m is the **gradient** of the slope. The gradient is simply how steep a slope is. *m*=2 is steeper than *m*=1. If *m*=0 there is no slope. If *m* is a minus number then the slope is negative. You may find Figure 3 a helpful illustration

You may sometimes find the formula presented with different letter, e.g. *y*=*ax*+*b* or with Greek letters Alpha *a* and Beta *β*. *e.g. y*=*ax*+*β*

b is the **intercept**. This is the value of *y* when x is zero. In this case this is how long we would expect a monarch to reign if they came to the throne at the age of zero. You may find that Figure 4 is a helpful illustration of this.

x is our predictor variable. This is the variable we are using to explain *y*. This is why the equation takes the form *y*=*mx*+*b*. You may find that Figure 5 is a helpful illustration of this. In order to find the value of *y* we need to find *m* and *b*.
The best way to start is by drawing a table (see Table 2). It will give us the numbers we need to calculate *y* and *m*.

Figure 3Various values of *mx*



Firstly to find the gradient of the slope *m*

$$m=\frac{n\Sigma(xy)-\Sigma x\Sigma y}{n\Sigma(x^2)-(\Sigma x)^2}$$

Figure 4 Various values of b(the intercept)

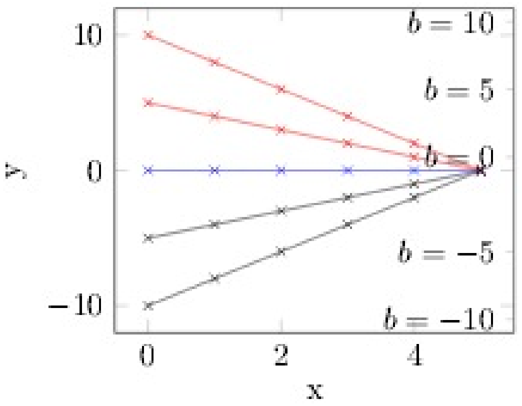
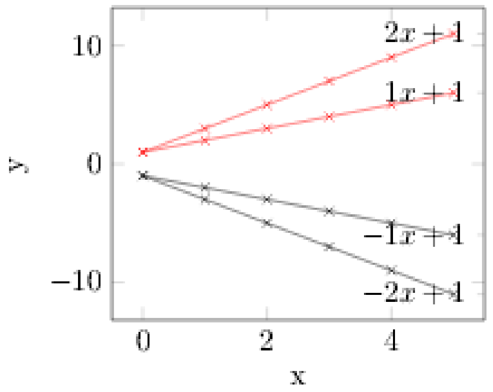


Figure [5](#)Various values of $mx+b$



n is the number of observations. In this case we have nine monarchs so

Therefore:

$n=9$

$\Sigma(xy)=7684$

$\Sigma x=401$

$\Sigma y=232$

Σ^2

$$m=\frac{9\times 7684-(401\times 232)}{9\times 19935-(401\times 401)}$$

$$=\frac{-23{,}876}{18{,}614}=-1.28$$

Therefore $m= -1.28$.

Secondly we need to find the intercept. The formula for the intercept is

$$b=\Sigma y-m(\Sigma x)n$$

Using the numbers from our table this makes

$$b=(232)-(-1.28)\times(401)\times 9=232--153.39=745.39=82.8$$

Now we have our m and b values we are able to predict how long a king or queen will reign for when coming to the throne.

Remember:

Table [2](#) Reign of British Monarchs

	Age on Accession to throne (x)	Reign (y)	xy	x2
George I	54	12	648	2916
George II	34	33	1419	1849
George III	22	59	1298	484
George IV	57	10	570	3249
William IV	64	6	384	4096
Victoria	18	63	1134	324
Edward VII	59	9	351	3481
George V	44	25	1100	1936
George VI	40	15	600	1600
Totals	$\Sigma x=401$	$\Sigma y=232$	$\Sigma xy=7684$	$\Sigma x^2=19935$

$$y = mx + b$$

Now replace the letters with the values we have

$\mathit{Length\ of\ reign} = -1.28 \times \mathit{age\ at\ accession} + 82.8$

So how can we this data to predict future events? If someone comes to the throne age 50

$$-1.28 \times 50 + 82.8 = 18.8 \mathit{years}$$

If at the age 20

$$-1.28 \times 20 + 82.8 = 57.2 \mathit{years}$$

This regression line has been plotted in Figure 6.

6 The limitations of simple linear regression

The main limitation of the regression equation is that it is a generalisation derived from a lot of data points. It can shed light on a general pattern, but it will never be 100% accurate. To use the example again of predicting university exam performance from school exam performance some students will do as well as predicted, some better and some less well. A small number will do a lot better or a lot worse than predicted.

6.1 Predicting the correct way round

There are important cautions to be aware about when using any form of regression analysis. Firstly unlike the correlation analysis we looked at in Chapter 13 the regression equation only works one way. We can use the same regression equation to calculate how long we might expect a monarch to reign when we know their age on coming to the throne. In other cases it is obvious from cause and effect that the equation can only work one way round. We use our central heating more when the temperature decreases but using the central heating does not make the outside temperature decrease. In the case of Table 6" we can calculate a regression equation for the relationship between the distance from London to a particular city and the price. However it is not possible to predict the city from the price. Even if I form a regression equation to predict the distance from the price this does not tell us what direction from London the price refers to. Paris is a similar distance from London as York to the north, Amsterdam to the East and Swansea to the west.

6.2 Outliers

Secondly outliers are an important issue in regression analysis. If you are using a statistical analysis software package it will probably alert you any outliers, but they are usually easy to spot on the scatterplot like in Figure 6. Outliers are distant from the regression slope and may have a disproportionate influence on the slope. Looking back at our scatterplot you might have noticed that one of the observations is quite a bit further away from the line that the others. The observation marked in Figure 7 is that of George VI who became king at the age of 40, but reigned for just 15 years. We can read off on the graph how long we might have expected George VI to reign. If we follow age 40 up to the regression line we can see that a monarch who comes to the throne age 40 could usually expect a reign of over 30 years. How best to deal with outliers is a matter of judgement. Sometimes researchers remove outliers from the analysis in order to make their equation more reliable and increase the value of r^2.

This is sometimes appropriate, but it does need to be justified. Generally speaking we do well not to remove outliers unless we know why they are outliers or that they are sufficiently rare to be able to dismiss them as `freak' observations.

6.3 Scope

Life expectancy is increasing so using the life expectancy of eighteenth and nineteenth century kings and queens is likely to underestimate future life expectancy.

It is not advisable to use the regression equation when the *x* values are outside the scope of the original observations. William IV was 64 years only when he became king. The regression line `predicts' that any future monarch who came to the throne in their seventies would actually have a negative reign which is of course impossible.

7 How reliable is our equation? Calculating r^2

It is all very well to use the regression equation to predict how long a monarch might reign, but how reliable is it? Does our regression line equation explain the relationship between a monarch's age on accession and the length of the their reign, perfectly? Reasonably well? Hardly at all? In a similar way to the correlation co-efficient we can use an *r*2 calculation to find out how well the regression equation explains the relationship.

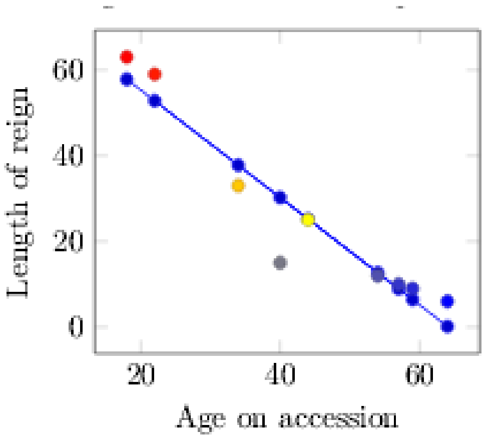
The *r*2 value is between 0 and 1. A value of 1 indicates that the regression equation explains the relationship perfectly. A value of zero indicates that the regression equation does not explain the relationship at all.

R^2 is calculated as follows

$$r^2 = \frac{\mathit{Sum\ of\ squares} - \mathit{Estimated\ Sum\ of\ Squares}}{\mathit{Sum\ of\ squares}}$$

Stage 1:

Figure [6](#) Relationship between age on accession and length of reign with



First we need to to calculate the sum of squares (see Table [3](#)).

1. Calculate the mean of the *y* values. *y* is the length of a monarch's reign. We call this ȳ.

   ȳ= 25.78

2. Calculate the difference between each *y* and .
3. Then square the differences
4. Add together the squared differences.This gives us the total Sum of Squares, which is 3769.56.

Stage 2: Secondly we need to calculate the Estimated Sum of Squares (ESS) (see Table [4](#)).

1. Find the estimated *y* values. These are the values of *y* (the length of reign) which would be the case if the regression equation was perfect. We call this ŷ or y-hat.
2. To so this we need to take each *x* value (that is the age on accession to the throne) and use the regression equation to find out what the *y* *would be* if the regression equation worked. For example George I was 54 when he became king.
3. Our regression equation is

$$\textit{Length of reign} = (-1.28 \times \textit{age at accession}) + 82.8$$

So for the case of George I the equation is

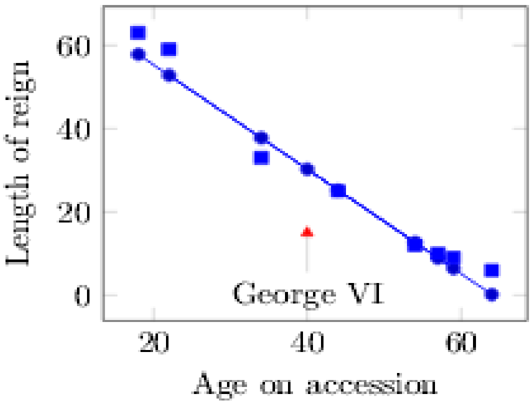
$$\textit{length of reign} = -1.28 \times 54 + 82.8 = 13.68$$

So ŷ = 13.68

In other words we would expect George I to reign for 13.68 years based in his accession at the age of 54.

1. Do the same for each king and queen.
2. Find the differences between the actual reign (*y*) and the estimated length of reign. (ŷ)

Figure [7](#) Relationship between age on accession and length of reign with regression



3. Square all these differences.
4. Now add together all the squares of the differences. This comes to 378.72.

$$r^2 = \frac{3769.56 - 378.72}{3769.56} = 0.90$$

Now we calculate the r^2

0.9 is near to 1 which shows that the regression equation is a very strong, though not perfect predictor of reign length. Sometimes r^2 is expressed at a percentage, so we would express 0.9 as 90%. Put in simple terms this indicates that 90% of reign length can be explained by the age at which a monarch came to the throne.

Results summary for predicting monarch reign

We can put a summary of results into a table (Table [5](#)).

8 Exercises

1. Table [6](#)" shows the lowest available prices to travel to different world cities from London and the approximate distance in kilometres. Longer distance flights are more expensive that shorter flights. Create a scatterplot for cheapest price (on the y vertical axis) and distance (on the x horizontal axis)
   1. Calculate the best fit regression line.
   2. Compare your regression line with that of the length of monarchs'

    reigns. Which line reflects the data better?

2. Study Table 7 of British Prime ministers since 1940. The table shows the age at which each Prime Minister came into office and how long they were in office for.
    1. Draw a scatterplot with the age the person became Prime minister on the x axis and the length of time they served as Prime Minister on the y axis. Compare the Prime Minister's scatterplot to the scatterplot of the King and Queens. How do the plots differ? Why do the plots differ?
3. Table 9 shows the distance between Bridgetown, Barbados and the Third class train fare in 1910. [1]
    1. Calculate the regression equation for the relationship between distance (the predictor variable) and the cost in cents (the variable to be predicted).
    2. Calculate r2
4. Examine the data in Table 8
    1. Draw a scatterplot of the data with price discount on the y axis anrom Rome on the x axis.
    2. Calculate the regression equation for Wheat price discount and Distance from Rome. (Note that the numbers are all negative). [2]
    3. Peter Temin [3] reports that this regression equation was rejected by reviewers of Roman history journals as a fluke. Why might experts in Roman history be sceptical of this data?

Table 3 Table for calculating the Sums of Squares

y	ȳ	y−ȳ	(y−ȳ)²
12	25.78	-13.78	189.83
33	25.78	7.22	52.16
59	25.78	33.22	1103.71
10	25.78	-15.78	248.93
6	25.78	-19.78	391.16
63	25.78	37.22	1385.5
9	25.78	-16.78	281.5
25	25.78	-0.78	0.6
15	25.78	-10.78	116.16
	Sum of squares (SS)		
			3769.56

References

1. Data from: Jim Horsfield (2001) *From the Caribbean to the Atlantic: A Brief History of the Barbados Railway,* St. Austell: Paul Catchpole}
2. These figures were published are from Kessler D. and P. Temin. 2008. 'Money and Prices in the Early Roman Empire' in William V. Harris (ed.) The Monetary Systems of the Greeks and Romans (Oxford).
3. Peter Temin (2006) The Economy of the Early Roman Empire, The Journal of Economic Perspectives 21, pp. 133.151

Table 4 Table for calculating Estimated Sum of Squares (ESS)

x	y	ŷ	y−ŷ	(y−ŷ)²
54	12	13.68	-1.68	2.82
34	33	39.28	-6.28	39.44
22	59	54.64	4.36	19.01
57	10	9.84	0.16	0.03
64	6	0.88	5.12	26.21
18	63	59.76	3.24	10.5
59	9	7.28	1.72	2.96
44	25	26.48	-1.48	2.19
40	15	31.6	-16.6	275.56
		Estimated Sum of squares		378.72

Table 5 Summary

Equation	r	r²
-1.28+ age at accession+ 82.8	0.95	0.9

Table 6 Airfares and distances from London to selected destinations worldwide

City	Cheapest price from London £ (y)	Distance from London (km) (x)
Paris	37	341
New York	354	5586
Las Vegas	495	8423
Mexico City	645	8943
Doha	369	5219
Johannesberg	524	9039
Sydney	761	16991
Auckland	765	18329
Hong Kong	504	9646
Barbados	558	6771
Amsterdam	63	359

Table 7: British Prime Ministers since 1940

Prime Minister	Age became Prime Minister	Time as PM (years)
Gordon Brown	56	3
Tony Blair	44	10
John Major	47	7
Margaret Thatcher	54	11
James Callaghan	64	3
Harold Wilson (second term)	58	2
Edward Heath	54	4
Harold Wilson (first term)	48	6
Alec Douglas-Home	60	1
Harold Macmillan	63	6
Anthony Eden	58	2
Winston Churchill (second term)	77	4
Clement Atlee	62	6
Winston Churchill (first term)	66	5

Table 9 Distances and third class train fares to from Bridgetown, Barbados c.1910

Miles from Bridgetown	Station	Third class fare (cents)
2.5	Rouen	4
5.5	Bulkeley	6
7	Windsor	8
9	Carrington	12
10	Sunbury	14
11	Bushby Park	16
13	Three Houses	16
16	Bath	20
20	Bathsheba	24
24	St Andrews	24

Table 8 Relationship between distance from Rome and wheat prices. (c.150BC to AD 80)

Region	Distance from Rome (km)	Distance from Rome "discount"
Sicily	427	-1.5
Spain (Lusitania)	1363	-2.5
Po Valley	1510	-3
Asia Minor (Pisidian Antioch)	1724	-3.13
Egypt (Fayum)	1953	-4
Palestine	2298	-3.25

CHAPTER 15: RANKING DATA

1 Ranking data

Sometimes the rank or order of data is more important to us than the actual data. The person who comes first in the 100 metres final in the Olympics will win the gold medal. They don't need to have run a certain speed to get the medal. They don't have to run faster than anybody has ever done. They just need to run the race faster than everyone else in the same race. A sports team with the most points over the course of a season will be crowned champions. It does not matter how many points they got, only that they got more than any other team.

Examples include

1. Exam results for best to worse
2. Places in a race
3. People in order of income
4. Farms in order of size
5. Books or music in order of sales

2 The Mann-Whitney U-test

The Mann-Whitney U-test is often considered to be a ranking test equivalent of the Student's t-tests. Like the t-tests it is suitable for small samples. The radio station Classic FM asks listeners to vote for their favourite classical pieces and the station plays the top 300 over an Easter weekend. Some composers only appear once in the top 300 whereas others appear several times. [1] In 2011 the top 300 included 16 pieces by Beethoven and 23 by Mozart. Of these pieces both had eight in the top 100. For the purposes of illustration we will only consider the pieces in the top 100 and we have taken out composers other than Beethoven and Mozart.

Although we know that both composers had eight pieces in the top 100 how can we say which composer was the more popular with Classic FM listeners? Beethoven has three pieces in the top four, but he also has two pieces in the bottom three.

Our null hypothesis is both of the composers are equally popular with Classic FM listeners. We can use a test called the Mann-Whitney U-test to find out whether listeners preferred one composer to another. This test does not require that data be paired (e.g. like the correlation co-efficient) or that the two groups contain the same number of observations. The Mann-Whitney test compares the median for two samples.

The method

Our first task is to rank the 16 pieces in order of popularity.

Beethoven Mozart Beethoven Beethoven Mozart Beethoven Beethoven Beethoven Mozart Mozart Mozart Mozart Mozart Beethoven Mozart Beethoven

As there are two variables we need to get two numbers. One for Beethoven (U_1) , and one for Mozart (U_2).

$$U_1 = R_1 - n_1(n_1+1)^2$$

Stage 1: First we need to add up the ranks for Beethoven

1+3+4+6+7+8+14+16=59 (R_1)

Putting this into the (U_1) equation

$$59 - 8(8+1)^2 = 59 - 722 = 59 - 36 = 23$$

Therefore U_1 =23

Stage 2: Secondly we do the same for Mozart

$$2+5+9+10+11+12+13+15=77(R_2)$$

Putting this into the equation

$$U_2 = R_2 - n_2(n_2+1)^2$$

$$77-8(8+1)^2 = 77-722 = 77-36 = 41$$

$U_1 =$ *U value for Beethoven.*

$U_2 =$ *U value for Mozart.*

$R_1 =$ *Sum of Ranks for Beethoven*

$R_2 =$ *Sum of Ranks for Mozart*

$n_1 =$ *Number of Beethoven pieces*

$n_2 =$ *Number of Mozart pieces*

Table 1 Critical values of U at 95%

N1 (Larger sample)	**'**	**1**	**2 3**	**4**	**5**	**6**	**7**	**8**
n2 (Smaller sample)	1							
	2							0
	3				0	1	1	2
	4			0	1	2	3	4
	5		0	1	2	3	5	6
	6		1	2	3	5	6	8
	7		1	3	5	6	8	10
	8		0 2	4	6	8	10	13

Table 2: Report results of the Mann-Whitney-U-test

Total number of observations	**U_1**	**U_2**	**Significant at 95%**	**p-value**
16	23	41	No	>0.1

Figure 1 Mangolds or mangelwurzel

Therefore U_2=41

As the U-value for Beethoven is 23 and the U-value for Mozart is 41 it appears that Beethoven is more popular that Mozart (remember that these are based on ranking data, so smaller numbers are better (1= the best).

However this could be down to chance so we have one more task to perform. We need to take a look at our ``Critical Values of `U chart. An extract is in Table 1.

Both Beethoven and Mozart contain 8 observations so we can see that the critical value of U for two samples of 8 is at 95% is 13 (see Table 1). We take the lower U value (the 23 for Beethoven) which we can see is greater than the critical value of 13. With critical values of U we need the lower value of U to be less than or equal to the critical value. In this case we must accept the null hypothesis that there is no difference between the preferences. The results of the Mann-Whitney test we have just performed are in Table 2.

2.1 Notes and cautions

The Mann-Whitney U-test can be used with small samples (though you need at least seven observations altogether.)

This test is ideal for ranked data. You can use it for non-ranked data, but the student's t-test would normally be preferable.

This test is based on ordinal (ranked data). It does not tell us how much more popular Beethoven might be than Mozart.

3 Calculating the Spearman's rank correlation coefficient

The Spearman's Rank correlation coefficient might be seen as a ranking equivalent of Pearson's Product-Moment Correlation Coefficient Chapter 13. The data in Table 3 shows the number of acres used for growing different crops and grasses in Shropshire in 1870-79 and 1970-79. [2] Just by glancing at the data we can see that there are some changes between the two periods. Wheat took up almost twice as much land in the 1870s as in the 1970s, but twice as much land was given over to barley in the 1970s than in the 1870s. Sugar beet wasn't grown at all in the 1870s and turnips and swedes took up almost 10 times as much space in the 1870s as in the 1970s. Therefore this test is looking at the rankings (popularity) of the different crops and not the actual data. This test can help us to see if there is an association between the number of acres given over to each crop in the 1870s and in the 1970s. In case you were wondering mangolds (also known as mangelwurzel among other names- see Figure 1) are root vegetables grown mainly to feed livestock.

The formula for the Spearman's Rank Correlation Coefficient is

$$r_s = 1 - \frac{6 \Sigma d^2}{n^3 - n}$$

r_s is the Spearman Rank Co-efficient

d is the difference in the ranks (rank of 1870s minus rank of 1970s)

d^2 is the differences squared (NB this value will be positive).

Σd^2 is the sum of the differences squared (the differences added together).

n is the number of observations, in this case 10 as we have 10 different crops.

n^3 is the number of observations cubed. As we have 10 crops this is 10×10×10

The answer will be between -1 and +1. A score of 0 would indicate that there is no correlation between the crops grown in the 1970s and in the 1870s. A score of 1 would indicate perfect positive correlation. A score of -1 would indicate perfect negative correlation (in other words the ranks are totally reversed). An answer of 1 would indicate perfect positive correlation - that the ranks matched perfectly between the 1870s and 1970s. For the test to be reasonably effective at least 10 pairs of observations are required.

Table 3: Crops and grasses grown in Shropshire (acres)

Crop	01/09/1870	01/09/1970
Wheat	81,074	44,862
Barley	53,832	128,385
Oats	25,623	15,572
Other Corn	12,503	4,481
Potatoes	5,553	12,104
Turnips and Swedes	49,856	5,264
Mangolds	4,804	670
Sugar Beet	0	22,776
Other Green Crops	3,272	6,351
Clover and Rotation Grass	77,292	121,434

STEP 1 The first thing we need to do is rank the data from both sets of years. The crop using most acres is given the rank 1; the crop using fewest acres is given the rank 10.

STEP 2 Calculate the difference in the ranks then square the difference (see Table 5)

STEP 3: Calculate the sum of the difference the ranks squared

1+9+0+9+1+25+4+36+4+1=90

STEP 4: Put the numbers into the equation

n represents the number of crops and grasses, which in this case is 10 so we can put numbers into replace $\sum$ d² and *n*

Table 4: Crops and grasses grown in Shropshire (acres)

Crop	01/09/1870	Rank of 1870s	01/09/1970	Rank of 1970s
Wheat	81,074	2	44,862	3
Barley	53,832	4	128,385	1
Oats	25,623	5	15,572	5
Other Corn	12,503	6	4,481	9
Potatoes	5,553	7	12,104	6
Turnips and Swedes	49,856	3	5,264	8
Mangolds	4,804	8	670	10
Sugar Beet	0	10	22,776	4
Other Green Crops	3,272	9	6,351	7
Clover and Rotation Grass	77,292	1	121,434	2

Table 5: Crops and grasses grown in Shropshire (acres) Rank differences

'	01/09/1870	Rank of 1870s	01/09/1970	Rank of 1970s	Differences in ranks (d)	Different in ranks squared
Wheat	81,074	2	44,862	3	-1	1
Barley	53,832	4	128,385	1	3	9
Oats	25,623	5	15,572	5	0	0
Other Corn	12,503	6	4,481	9	-3	9
Potatoes	5,553	7	12,104	6	1	1
Turnips and Swedes	49,856	3	5,264	8	-5	25
Mangolds	4,804	8	670	10	-2	4
Sugar Beet	0	10	22,776	4	6	36
Other Green Crops	3,272	9	6,351	7	2	4
Clover and Rotation Grass	77,292	1	121,434	2	-1	1

Our value of is 0.45. This indicates a moderate positive correlation between the 1870s ranks and the 1970s ranks.

$$r_s = 1 - \frac{6\times 90}{10^3 - 10} =$$

$$1 - \frac{540}{10\times 10\times 10 - 10} =$$

$$1 - \frac{540}{1000 - 10} =$$

STEP 5: Calculate the number of degrees of freedom.

$$1 - \frac{540}{990} =$$

$$1 - 0.55 = 0.45$$

Number of pairs of data−2

Therefore

10−2=8

STEP 6: Look up the critical value. The critical value of r 95% at 8 degrees of freedom is 0.738. This is

Table 6 Critical values for Spearman's Rank

Degrees of freedom	Alpha= 0.10	α=0.05	α=0.01
1			
2			
3			
4	1		
5	0.9	1	
6	0.829	0.886	1
7	0.714	0.786	0.929
8	0.643	0.738	0.881

Table 7 Results of Spearman's rank test

Rs	Degrees of freedom	Significant at 95%	p-value
0.45	8	No	>0.1

more than our value of r at 0.45 so we can accept the null hypothesis that there is no difference between the ranks.

3.1 Notes on Spearman's Rank

Although there is a certain amount of consensus that at least 10 pairs are necessary to make the Spearman's Rank test effective, it is frequently used for fewer pairs.

Sometimes the lower case Greek letter (rho) is used instead of rs. (It is important not to ρconfuse this with the Roman letter p.)

4 Cautions with with ranking data

Ranking data can tell us the order of data, but it does not tell us anything about the distribution of the data. A website about the Romans in Britain tells us that the following towns were the ten largest (in area) in Roman Britain. (Table 8) [3]

Table 8 Ten largest towns in Roman Britain, by area

Rank	Town
1	London
2	Cirencester
3	St Albans
4	Wroxeter
5	Canterbury
6	Winchester
7	Leicester
8	Silchester
9	Chichester
10	Colchester

There is no data here about which how large the towns actually are and how they differ from each other in size. London was bigger than Cirencester, but how much bigger? Was Cirencester just a little bit smaller? Half the size? A quarter of the size? The ranks alone do not provide any of this data. It is like knowing who won a marathon but not the time they did it in or whether the person who came second was half a second behind the winner or ten minutes behind. In the Classic FM favourite music chart we are not told how many people voted for each piece of music. The top rated Beethoven piece could have got one more vote than the piece one place behind it or half a million more votes.

5 Combining ranking and non-ranking data

In Chapter 3 we examined means, median and modes. We observed that in many cases the mean average is not a good summary of the data as a whole. We mentioned that distribution of income was one such phenomenon where a small number of people earn more than the mean and a lot of people earn less than the mean-- therefore the mean average is not a good overall summary of the data. These observations have led users of statistics to explore these distributions and discover ways to make comparisons and predictions based on them.

6 Gini coefficient of income inequality

In a society where income is perfectly distributed everybody would have the same amount of income. In a very unequal society a small number of people would have nearly all the society's income and the rest of the population very little at all. The Gini coefficient is a way of measuring income inequality in society. The Gini coefficient gives a number between 0% (total equality) and 100% (total inequality). Table 9 shows the income distribution of a country we will just call Country A. The people of Country A have

Figure 2 Protests against income inequality in the USA. The 99% refers to the population which is not part of a very wealthy 1% elite.

been placed into one of five groups depending on their income. Each group makes up 20% of the population (20%×5=100) The Line of Equality is a straight line of 45 degrees (Figure 3). When we plot the data from Country A we can see that the line does not match the Line of Equality. On Figure 4 I have shaded area between the line of equality and Country A data (this is called the Lorenz curve). If we calculate the size of the this shaded area we will have our Gini coefficient. The more unequal the income distribution the larger the shaded area. Figure 5 shows a country with high inequality and Figure 6 shows a country with little inequality.

The actual mathematical formula for the Gini co-efficient is fairly complex. We can however calculate the Gini coefficient by measuring the area of each of the shaded sections (Figure 7): Roughly we could write the formula of the Gini coefficient as

$$\Sigma 0.5\times(a+b)\times c\times 2$$

Area 1= $0.5\times(0+2)\times 02.=0.2$

Area 2= $0.5\times(2+10)\times 0.2=3$

Area 3= $0.5\times(10+25)\times 0.2=3.5$

Area 4= $0.5\times(25+50)\times 0.2=7.5$

Area 5= $0.5\times(50+100)\times 0.2=15$

a= lower boundary

b= upper boundary

c= proportion of the total population in each class 0.2 is used here for 20%

Population rank	**Percentage of income**	**Cumulative percentage of income (prefect equality)**	**'**
Poorest 20%	2	2	The poorest 20% of Country A have just 2% of the income
Next richest 20%	8	10	The poorest 40% of Country A have just 10% of the income
Next richest 20%	15	25	
Next richest 20%	25	50	
Next richest 20%	50	100	The wealthiest 20% have 50% of the income

We then we add the areas results together to get a total

$0.2+3+3.5+7.5+15=29$

We then subtract this total from 50 (50 is the total area under the line of equality)

$50-29.2=20.8$

We then multiply by 2

$20.8\times 2=41.6\%$

So the Gini co-efficient for Country A is 41.6%.

Figure 3 Gini coefficient: Line of equality

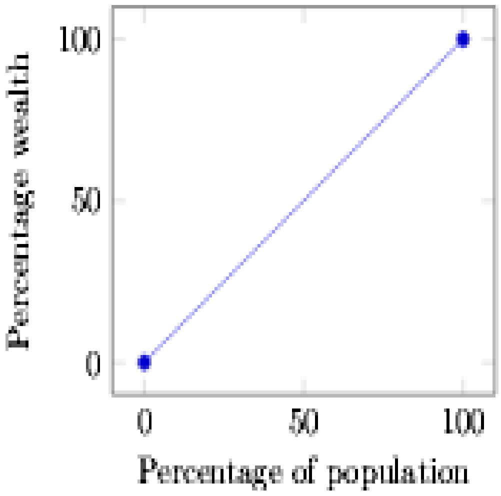


Figure 4 Gini coefficient: Shaded plot for Country A

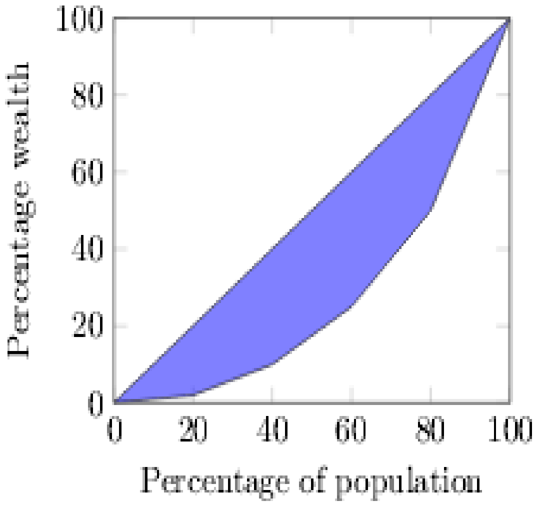


Figure 5 Gini coefficient: Shaded plot for Country B

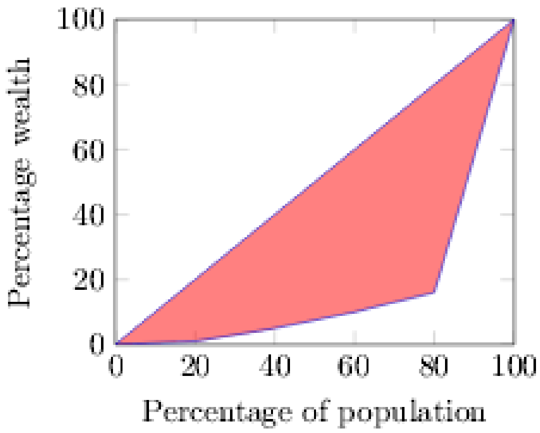


Figure 6 Gini coefficient: Shaded plot for Country C

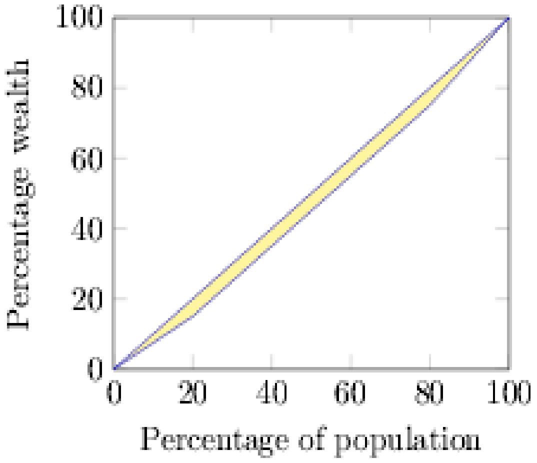
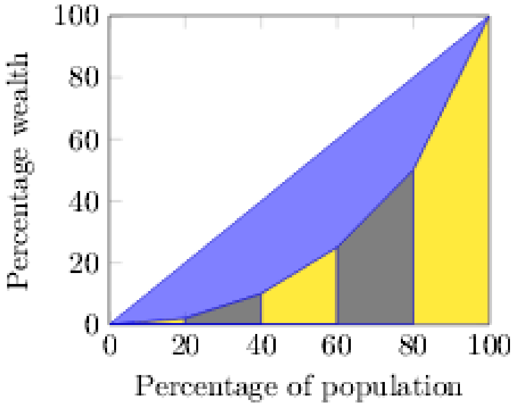


Figure 7: Calculating the Gini co-efficient with areas



6.1 Cautions about the Gini co-efficient

The Gini co-efficient is a measure of income inequality and not standard of living. According to recent World Bank figures the two of the world's most equal countries are Afghanistan with a Gini Coefficient of 27.8% and Finland with 26.9%. Whilst the two countries have a similar income distribution there is no suggestion that the people in Afghanistan and Finland enjoy a similar standard of living. To give two more unequal examples the USA has a Gini co-efficient of 45%, similar to South Sudan with 45.5%.

7 Exercises

Table 10 shows estimated migration to Glamorgan from other Welsh counties between 1861-1871 and 1901-1911.[4] Calculate the Spearman's rank correlation co-efficient for 1861-71 and 1901-1911.

Table 11 shows UK sales of the 30 bestselling books of all time.[5] Perform a Mann-Whitney U-test to work out whether there is the difference in the ranks between

Table 10 Migration to Glamorgan from other parts of Wales.

'	1861-71	01/11/1901
Monmouthshire	2300	11600
Carmarthen	5700	6800
Pembrokeshire	3700	5500
Cardiganshire	1500	2900
Brecknockshire	2100	3800
Radnorshire	400	1300
Montgomeryshire	450	2200
Flintshire	75	300
Denbeighshire	100	1100
Menionethshire	100	2500
Carnaravonshire	50	3900
Anglesey	0	1000

8 References

http://halloffame2011.classicfm.co.uk/

Domesday Book: 1875-1985', A History of the County of Shropshire: Volume 4: Agriculture (1989), pp. 232-269. URL: http://www.british-history.ac.uk/report.aspx?compid=22845 Date accessed: 20 May 2013

http://www.roman-britain.org/romano-british-towns.htm

Brinley Thomas (1930) The Migration of Labour into the Glamorganshire Coalfield (1861-1911) Economica 30, pp. 275-294.

↑ Derived from the Guardian datablog (2012)http://www.guardian.co.uk/news/datablog/2012/aug/09/best-selling-books-all-time-fifty-shades-grey-

Table 11: Bestselling books of all time (UK sales only)

Book title	Author	Sales	Gender of Author
Da Vinci Code,The	Brown, Dan	5,094,805	Male
Harry Potter and the Deathly Hallows	Rowling, J.K.	4,475,152	Female
Harry Potter and the Philosopher's Stone	Rowling, J.K.	4,200,654	Female
Harry Potter and the Order of the Phoenix	Rowling, J.K.	4,179,479	Female
Fifty Shades of Grey	James, E. L.	3,758,936	Female
Harry Potter and the Goblet of Fire	Rowling, J.K.	3,583,215	Female
Harry Potter and the Chamber of Secrets	Rowling, J.K.	3,484,047	Female
Harry Potter and the Prisoner of Azkaban	Rowling, J.K.	3,377,906	Female
Angels and Demons	Brown, Dan	3,193,946	Male
Harry Potter and the Half-blood Prince:Children's Edition	Rowling, J.K.	2,950,264	Female
Fifty Shades Darker	James, E. L.	2,479,784	Female
Twilight	Meyer, Stephenie	2,315,405	Female
Girl with the Dragon Tattoo,The:Millennium Trilogy	Larsson, Stieg	2,233,570	Male
Fifty Shades Freed	James, E. L.	2,193,928	Female
Lost Symbol,The	Brown, Dan	2,183,031	Male
New Moon	Meyer, Stephenie	2,152,737	Female
Deception Point	Brown, Dan	2,062,145	Male
Eclipse	Meyer, Stephenie	2,052,876	Female
Lovely Bones,The	Sebold, Alice	2,005,598	Female
Curious Incident of the Dog in the Night-time,	Haddon, Mark	1,979,552	Male

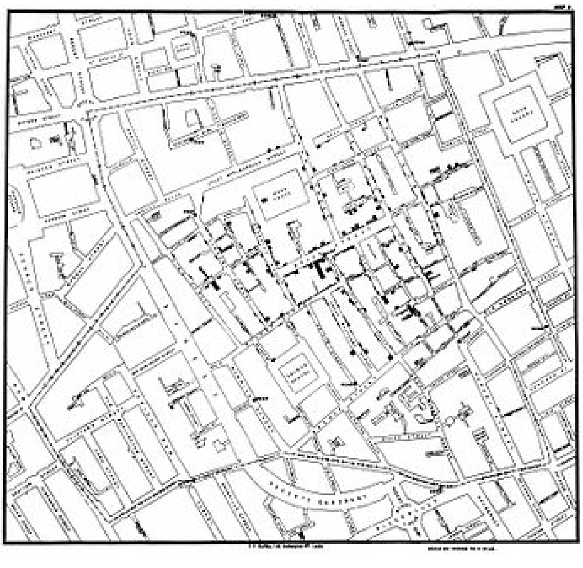
Chapter 16: Everything happens somewhere: Spatial data

1 Geographical aspects of data

Most statistical tests are based on the premise that each observation is independent. In other words data you collect on each person (or ship or plane or cow or bushel of wheat) is independent of other people, ships, planes, cows or wheat) in the sample. When we identify relationships we are looking for common characteristics between individuals.

In contrast geographical data analysis takes into account that relationships can be spatial. For example is there a relationship between how we vote and how our neighbours vote? Do people of the same race, ethnicity or religion tend to live near each other in a particular city? Are people on high incomes living in different places to people on lower incomes? If you think of a town or city you know you'll probably be able to think of poor areas or better-off areas. In many UK cities there are Chinatowns and Italian Quarters or areas where people from a particular religious group live. We can see many of these patterns, but we don't always feel we have ways of describing them. Spatial data analysis can also help us to see how places have changed over the course of time.

Figure 1 John Snow's map of the 1854 Broad Street Cholera epidemic



John Snow's maps of the 1854 Broad Street Cholera epidemic have become very well known (Figure 1). Most of Snow's contemporaries believed that cholera was spread through the air. If this was the case then there would be no particular reason why some parts of London should have had much worse cholera outbreaks than others. Snow marked where cholera victims lived and discovered the existence of obvious clusters of victims. Why did the people of Broad Street suffer so badly whilst neighbours in nearby streets were not infected? Snow found the one thing that all the victims had in common - they were using the same water pump. Although initially not all Snow's colleagues were convinced of his theory that cholera was spread through water and not the air, his discovery led to improvements in the treatment and prevention of the disease.

2 Spatial Patterns

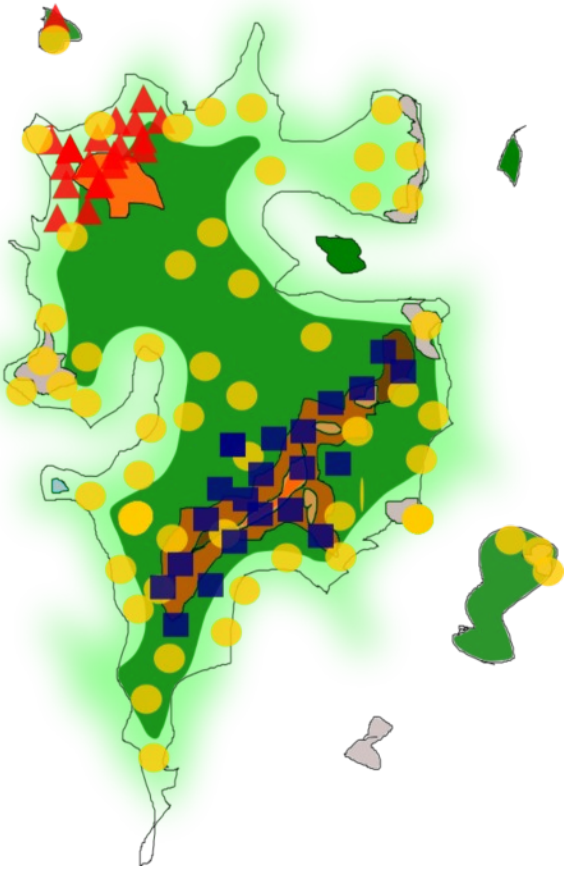
When we put observations onto a map we can sometimes spot patterns just by sight. Figure 2 shows the location of different ethnic groups on a fictional island.

The Red triangle group are clustered in the north-west of the island.

The Blue Square group are spread out in a diagonal line from the South-West of the island to the north-east.

The Yellow circle group are spread randomly throughout the island. There does not appear to be any particular pattern.

Figure 2 Distribution of three groups of people on a fictional island



3 Nearest Neighbour Analysis:a worked example

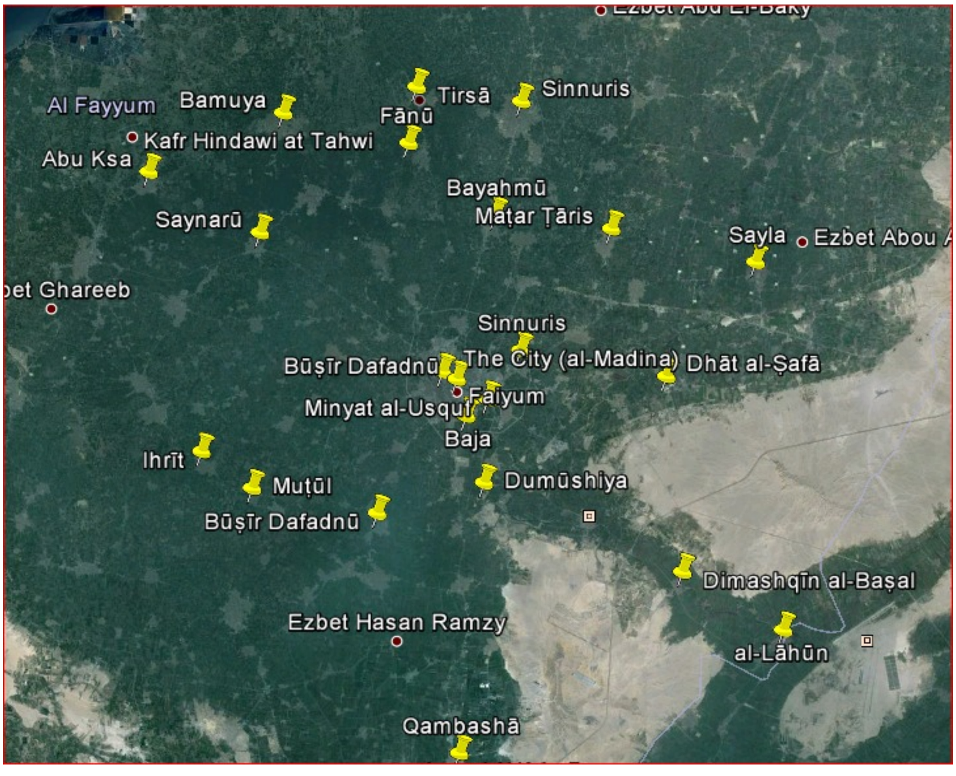
On the previous page we saw how we could describe patterns by sight. However, in real life it is often difficult to spot patterns so easily. Nearest neighbour analysis provides a basic test to identify whether there are spatial patterns in the data and what the nature of these patterns might be.

al-Nābulusī was a high ranking official Ayyubid administration of Egypt. In 1245 he was ordered to travel to the Fayuum region of Egypt to collect information on agriculture and taxes. He collected information on everything from chicken breeding to land taxes to sugar cultivation. Overall he surveyed around 130 towns and villages in the Fayuum region. One of the interesting things about the time was that non-Muslims had to pay a special poll tax. In the Fayyum there were 21 settlements where at least one person paid the poll tax for being a non-Muslim. [1]

To do a Nearest Neighbour Analysis we need:

1. A list of the all the settlements where non-Muslims were living
2. How far each settlement with non-Muslims is from nearest other settlement with non-Muslims (we we do this in kilometres here).

Figure 3 The Fayyum region and settlements plotted on Google Earth.



3. The total area of the Fayuum region

The settlements where non-Muslims were present are shown on the map above (produced using Google Earth). Is there any way we might describe this pattern? Are the communities in which non-Muslims live close together or are they scattered throughout the region? Are non-Muslims likely to live in settlements close to other settlements where non-Muslims live? Are they scattered evenly though the region or is the location of these villages totally random?

Tip: To calculate distances and area you can use the measurement tools on the freely available Google Earth

Nearest neighbour analysis allows us identify whether the pattern might be described as clustered, regularly spaced or completely random.

The formula is:

$$R_n = 2\bar{d}\sqrt{\frac{n}{a}}$$

R

R_n =Value of the Nearest Neighbour Analysis

$\bar{d}$ =Average distance

n =Number of settlements

a =Area of the region

Table 1 Distances between recorded settlements in the Fayuum

Settlement	Nearest neighbour	Distance
al-Lāhūn	Dimashqīn al-Baṣal	5.2
Tirsā	Fānū	2.6
Maqṭūl	Ihrīt	2.6
Qambashā	Buljusūq	6
Dumūshiya	Minyat al-Usquf	3.8
Buljusūq	Qambashā	6
Saynarū	Bamuya	5
Sayla	Dhāt al-Ṣafā	5.9
Bayahmū	Fānū	4.9
Abū Ksā	Saynarū	5.4
Būṣīr Dafadnū	The City (al-Madina)	0.8
Maṭar Ṭāris	Bayahmū	4.8
Ihrīt	Muṭūl	2.6
Fānū	Tirsā	2.6
Minyat al-Usquf	Bāja	1.2
Dimashqīn al-Baṣal	al-Lāhūn	5.2
Dhāt al-Ṣafā	Sayla	5.9
Bāja	Minyat al-Usquf	1.2
Sinnuris	Tirsā	4.1
Bamuya	Saynarū	5
The City (al-Madina)	Būṣīr Dafadnū	0.8

STEP 1: Firstly we need to identify each of the settlements in which non-Muslims lived and see how far each one is from it's the nearest in which other non-Muslims lived. (Remember, as the data is spatial, Settlement A's nearest neighbour may be Settlement B, but Settlement B's nearest neighbour is not necessarily Settlement A).

STEP 2: Calculate the mean distance (add together all the distances and divide by 21 - the number of settlements).

This comes to 3.89 km.

STEP 3: Next we need to find out how big our region is. The Fayuum region is approximately 1800km².

STEP 4: Now can use our formula to find out the **Rn** value.

$$\bar{d} \times 2 = 3.89 \times 2 = 7.77$$

$$\sqrt{\frac{21}{1800}} = 0.108$$

$$0.1808 \times 7.77 = 0.83$$

Rn=0.83

So the description of distribution (**Rn**) equals 0.83.

STEP 5: Now how do we interpret this?

The **Rn** value will be between 0 and 2.15.

With our **Rn** value 0.83 we can see that this quite close to 1. This indicates that the distribution of settlements lived in by non-Muslims is close to random and that there is no particular pattern in the spatial data. Therefore we can see that there is not much clustering in Fayuum region, but more a tendency toward randomness.

Figure 4: Interpreting the Nearest Neighbour Analysis

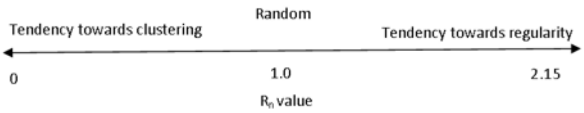
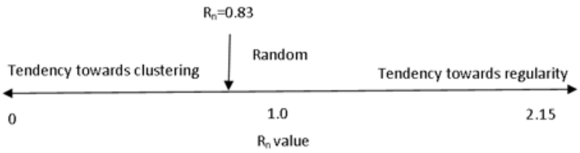


Figure 5: Interpreting the Nearest Neighbour Analysis of 0.83



4 Cautions with Nearest Neighbour Analysis

1. The calculation for Nearest Neighbour Analysis is very sensitive to changes in size of the geographical area covered.
2. Valleys, seas, hills, rivers, lakes, forests, deserts and other geographical phenomena can be problematic to Nearest Neighbour Analysis. For example, areas which are not inhabitable for humans might be included in the area covered, for example in the Fayuumn. This can lead to apparent randomness where clustering might be found if only habitable areas are considered. The area covered could be modified to account for this.
3. The area of interest might be seen as artificially constructed. Question 1, at the end of this chapter concerns seven National Trust properties in Gloucestershire, UK. However the National Trust also has properties in neighbouring counties which are actually nearer to some of the Gloucestershire properties than those actually in Gloucestershire.

5 Exercises

Table 2 shows a list of the seven properties in Gloucestershire owned by the National Trust.

Calculate the value of **Rn**. The area of Gloucestershire is 3,150km²

Is the pattern of National Trust properties closer to being random, clustered or regular?

6 References

[1] The data here is derived from Yossi Rapoport's website Rural society in Medieval Islam: History of the Fayyum http://www.history.qmul.ac.uk/ruralsocietyislam/

Table 2 National Trust Properties in Gloucestershire, UK

Property	Nearest Neighbour	Distance (km)
Chedworth Roman Villa	Lodge Park and Sherbourne Estate	10
Hailes Abbey	Snowshill Manor	6
Hidcote Manor Garden	Snowshill Manor	12
Lodge Park and Sherbourne Estate	Chedworth Roman Villa	10
Newark Park	Westbury Park	22
Snowshill Manor	Hailes Abbey	6
Westbury Park	Newark Park	22

CHAPTER 17: HAVING CONFIDENCE IN DATA

1 Introducing confidence intervals

In Chapter 4 we calculated the median, mean and mode of a sample of burial ages in Accrington in 1839. In Chapter 5 we recognised that the characteristics of our sample may be different from those of the population as a whole. Confidence intervals are a way of expressing how much the samples might differ from the population as a whole.

A common use of confidence intervals is used in the opinion polls which take place before elections. Suppose an election is taking place between Ms Smith and Mr Jones and an opinion poll of 1000 people finds that 51% will vote for Ms Smith and 49% for Mr Jones. Does that mean that Ms Smith is going to win? However well the sample was taken, there is a possibility of error. We can take the possibility of error into account, by recording not only the figures, but also the confidence intervals and the confidence levels.

We have already explored the confidence level elsewhere. At 95% confidence we can saying that 95% of the time our poll results will be within the calculated confidence intervals.

The confidence interval is expressed using the plus or minus sign ±. So instead of stating only that 51% of people will vote for Ms Smith we state that Ms Smith will receive 51%±3.1%. This means that Ms Smith could receive as many as 51%+3.1% of the vote or as few as 51%−3.1%. (Don't worry for the time being about where this 3.1% comes from).

Putting the confidence level and confidence interval together we say that at 95% confidence Ms Smith will receive 51%±3.1% of the vote.

If we increase the confidence level our confidence interval will be bigger. If we increase the confidence level to 99% our confidence interval will go up to ±4.08% (again don't worry where this figure comes from for the time being.)

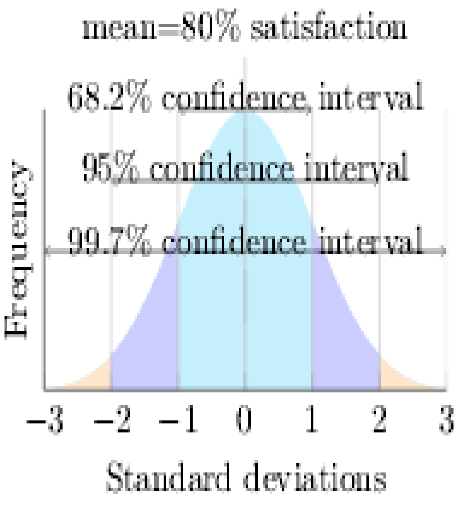
Estimating population characteristics from a sample

Each year the UK National Student Survey asks every final year student what they thought about their course. The results for each university and each subject are published in places such as newspaper league tables and in official Key Information Sets (KIS). Although every student is asked for their opinion, not every student responds to the questionnaire. Suppose that there are 100 students studying French at University A. All 100 students respond to the questionnaire and 80 (80\%) of them say they are satisfied with the course. So out of the 100 students we can say with total certainty that 80 of them satisfied and 20 are not.

However, suppose that there are also 100 students studying French at University B. At University B only 50 students respond to the questionnaire. 40 of them (80\%) say that they are satisfied with the course and 10 (20%) indicate that they are not satisfied.

Both University A and University B have satisfaction rates of 80%. In the case of University A all the students have responded so we know that 80% of ALL students are satisfied. But what about those 50 students at University B who did not respond to the survey? What percentage of those students would

Figure 1 The normal distribution



have said they satisfied had they responded? Is it fair to assume that 80% of non-respondents were also satisfied? The truth is we cannot know for sure. If all 50 of the non-respondents were satisfied that would mean that 90 out of 100 (90%) were satisfied. If all 50 were not satisfied then only 40 out of the 100 (40%) were satisfied. So the actual figure could be as low as 40% or as high as 90%.

2 Confidence intervals and the standard error

The idea that the true figure could be as low as 40% or as high as 90% is not particularly helpful to us. Satisfaction at University B could be higher than at University A or worse, or much worse. If we want to compare University A and B then what do we need to do?

To calculate our confidence intervals we need to calculate the standard error. Using our normal distribution curve we can `assume' that if we were able to repeat our survey over and over again the mean average rate of satisfaction would be 80%.

To calculate the standard error (SE)

In our example p=0.8 That is the 80% who said they were satisfied. We should use proportions (Numbers between 0 and 1 rather than percentages).

1−p=1−0.8=0.2

p×(1−p)=0.8×0.2=0.16

n=50 as there are 50 students in the sample.

0.16÷50=0.0032

$$\sqrt{0.0032} = 0.056$$

So the SE=0.056. Therefore 1SE=± (plus or minus)5.6

2.1 Interpreting the Standard Error

To interpret the Standard Error we can invoke the 68:95:99 rule. You may remember that 68% of a normally distributed sample is 1 standard deviation either side of the mean. 95% of observations are within 2 standard deviations and 99.7% within 3 standard deviations. Our SE is 5.6% so at a confidence level of 68% we can say the true mean lies between 80±5.6% So at a confidence level of 68% we can say the actual figure lies between 80−5.6=74.4 and 80+5.6=85.6

To use a 95% confidence interval we need to multiply our Standard Error by 1.96 (95% is actually 1.96 standard deviations rather than 2). So 5.6×1.96=10.976. At a confidence level of 95% we can say the true mean lies between 80±10.976 which is 80−10.976=69.02 and 80+10.976=90.98

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

SE= Standard error

p=proportion

n=number in the sample

2.2 Correction for finite population

You may have noticed that upper limit of 90.98% for the confidence interval of 95% exceeds the known maximum of 90% we discussed in the section above. In many cases in which confidence intervals are known we either don't know the population size or our sample is a very small proportion of the population. In cases where the sample size exceeds 5% it is advisable to make a small correction to the formula. In our National Student Survey example our sample covers 50% of the population (the population being students studying French in University B). Although it may seem a little inconvenient to undertake another calculation the correction is taking into account the fact we have a sizeable to sample in order to make our confidence limits more accurate.

$$FPC = \sqrt{\frac{N-n}{N-1}}$$

The formula for the correction is

FPC= Finite Population Correction

N=Population

n=number in the sample

STEP 1: For our example N=100 (the total number of students studying French at University B) and n=50 (the number of students who responded).

N−n=100−50=50

N−1=100−1=99

50÷99=0.505

FPC=0.71

$$\sqrt{0.505} = 0.71$$

STEP 2: Now we need to multiply our FPC by the Standard Error.

0.71×0.505=3.98

So 3.98×1.96=7.8.

At a confidence level of 95% we can say the true mean lies between 80±7.8 which is 80−7.8=72.2 and 80+7.8=87.8

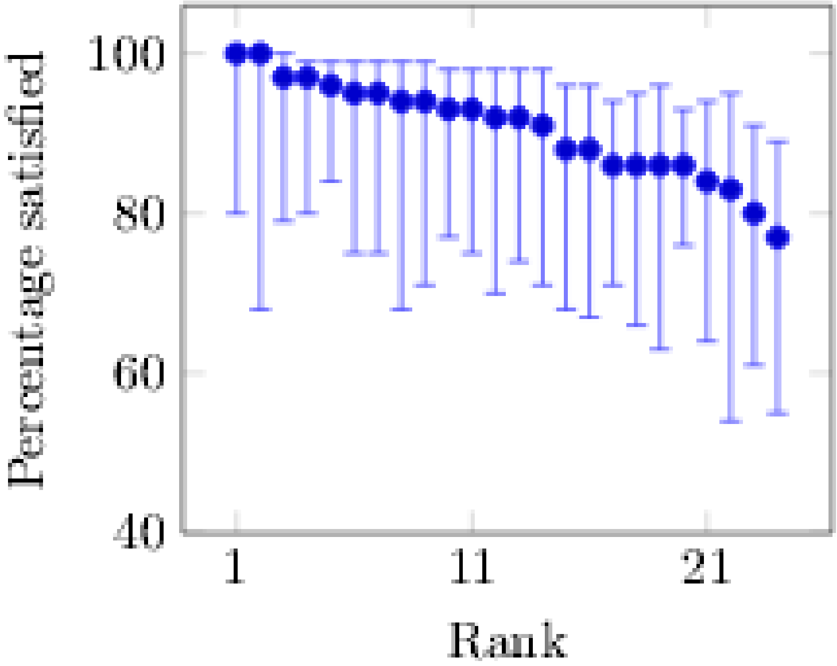
2.3 Presenting confidence limits

In Figure 2 t he black dots represent the percentage of satisfied students in 23 UK departments of American Studies. The lines coming off the black dots are the error bars. In the department ranked 23rd (last) 78% of students were satisfied. The error bar reveals that the real figure within a 95% confidence interval could have been as high as 86%, but as low as 58%. One of the interesting things about this graph is that it reveals that if the full extent of the confidence limits are considered, the lowest ranked department with 78% might actually have a higher level of satisfaction than the first ranked department which could have a satisfaction rate as low as 80%.

2.4 Cautions with confidence intervals

The confidence intervals only make sense if the sample is random. If we have reasons to believe that the characteristics of those in the sample are different to those of the population as whole the picture becomes more complex. In the student survey example, you may have reason to believe that people who respond to the survey are more positive or more negative about their course than those who do not respond.

Figure 2 National Student Survey: Percentage of graduates satisfied with their course in American Studies (2009), with confidence limits at 95%



In some circumstances confidence intervals can exceed 100% or go below 0% of the population, which is clearly not possible. This is an issue with the National Student Survey where some courses score close to 100% satisfaction. There are various advanced techniques for adjusting confidence intervals under these circumstances, some of which lead to asymmetrical confidence limits. The publishers of the National Student Survey have done this, hence the error bars in Figure 2 are not the same length each side of the data actually observed.

3 Exercises

1. An opinion poll for an election next week reports that Candidate A will get 47% of the vote and Candidate B will get 53% of the vote. A footnote states that the margin of error is ± 3% at 95%. Imagine you are explaining this to a friend who does not understand statistics. What would you say to him/her.

2. Taking 95% as your confidence level calculate the margin of error for the above election with a sample size of:

200

1,000

10,000

3. In a recent study Cheshire et al (2011) recorded and compared the speech of teenagers and elderly people living in two areas of London. [1] In total they collected around 2.8 million words. They recorded the speech of 1052 teenagers. 219 (20.8%) used a phrase using the verb ``to be with like, e.g. ``I'm like..., ``She's like.... Calculate the lower and upper confidence limits with a 95% confidence rate.

4 References

[1] J. Cheshire et al (2001) Ethnicity, friendship network and social practices as the motor of dialect change: Linguistic innovation in London. Sociolinguistica 28

CHAPTER 18: ASSOCIATION CAUSATION AND EFFECT

1 Association, causation and effect

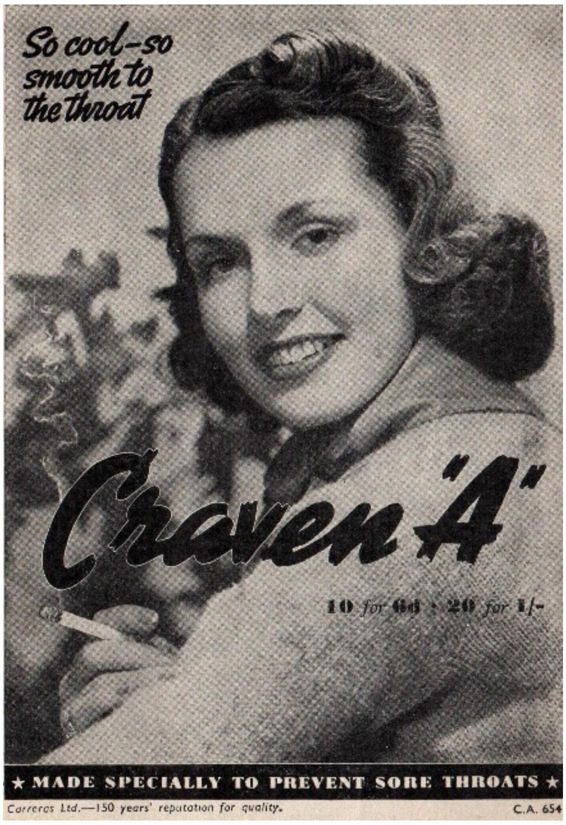
1.1 Smoking and causes of lung cancer

It would be difficult today to find anyone who would deny that smoking causes lung cancer, but until the 1950s many scientists and medics maintained that there was insufficient evidence to consider the association between smoking and lung cancer as causation - coal smoke and exhaust fumes were among other suggestions. Of 40,000 doctors questioned about their smoking habits in 1951, 789 had died by early 1954. 36 had of these had died of lung cancer; all were smokers.[1] In 1957 the Medical Research Council published a statement confirming the links between lung cancer and smoking. Further research had revealed smoker deaths from lung cancer were over 40 times those of non-smokers. This was found to be the case worldwide, in both men and women, and, crucially, the more people smoked, the more likely they were to die of lung cancer. [2] Despite what now appears to be overwhelming evidence, not everyone was convinced. Later in 1957 the statistician, Ronald Fisher wrote a letter to the British Medical Journal in which he stated:

> *A common ``device is to point to a real cause for alarm, such as the increased incidence of lung cancer, and to ascribe it urgent terms to what is possibly an entirely imaginary cause. Another, also illustrated in your annotation, is to ignore the extent to which the claims in question have aroused rational scepticism. [3]*

Saying that smoking causes lung cancer is not the same thing as saying that everyone who smokes will get lung cancer or everybody who gets lung cancer is a smoker. However, smoking can be said to be a major contributing cause of lung cancer. Exposure to asbestos and passive smoking can also cause lung cancer. Smoking can also cause heart disease and other forms of cancer. How, therefore, can we say with certainty that smoking causes lung cancer?

Figure 1 Adverts associated smoking with glamour and good health, as this advert from 1939 shows. Tobacco companies frequently accused their rivals of making products which caused sore throats.


So cool-so smooth to the throat
Craven "A"
10 for 6d · 20 for 1/-
★ MADE SPECIALLY TO PREVENT SORE THROATS ★
Carreras Ltd.—150 years' reputation for quality.
C.A. 654

In sciences the most effective way of establishing cause and effect is through controlled experimentation. If we want to find out whether a new medicine (Pill A) is more effective a treatment for Condition X than the currently used medicine (Pill B) a scientist would take two groups of people with Condition X. [4] One group would receive Pill A and a second group would receive Pill B. The health of the two groups would be monitored and side effects noted then. If after the trial period the health of the group taking Pill A is significantly better than that of the group taking Pill B we might conclude that Pill A is a better treatment for Condition X than Pill B. In the humanities this sort of experimentation

Figure 2 Ronald Aylmer Fisher (1890-1962)Fisher was a statistician and biologist who devised many statistical tests, including the F test. However, he rejected the claim that smoking causes lung cancer

is not generally feasible, and in the sciences it is not always ethical (asking a group people to smoke for 30 years to see if they get lung cancer would not constitute ethical behaviour.)

2 Association, causation and effect in the humanities

Establishing causation and effect is a complex undertaking. It is very rare that one outcome is caused by just one factor, and in this sense the smoking-lung cancer link is unusually conclusive. With social and cultural data the relationship between association, causes and effect is very messy. Many debates in the humanities are essentially discussions about cause and effect. What were the causes of the Second World War? What is the impact of the work of the 1960s Civil Rights movement on the lives of African-Americans in 2012? What was the impact of the assassination of President Kennedy on US-Soviet relations? We cannot bring Kennedy back to life and see how US Soviet relations would have been different had he lived. We cannot go back in time and change a factor believed to have contributed to the start of Second World War and see if things work out differently. These examples are not statistical examples, but the same principles apply to establishing cause and effect when we use quantitative data.

However just because we now have the data in the form of numbers does not mean that cause and effect can be established objectively without any doubt. In the previous chapter we observed that there was a moderate correlation between the price of wheat and the price of oats. But does this correlation mean that a rise in the price of oats causes a rise in the price of wheat? Or does it mean that a rise in the price of the wheat causes a rise in the price of oats? Maybe the price of wheat and oats is caused by some third factor we haven't considered. Perhaps the price of wheat is caused by the price of oats AND some other factors. Or perhaps the association is a pure coincidence.

Or what about the observation that Methodists are more likely to identify as Evangelicals than Anglican? Do people go to Methodist churches because they are Evangelicals? Or are people who go to Methodist churches more likely to see themselves as Evangelical? This does not get away from the fact that lots of Anglicans identified themselves as evangelical and Methodists as liberals. But how do we know which one of these possibilities is the case? Our statistical tests have identified associations between different factors, but they serve to raise questions as much as they provide answers.

3 Identifying cause and effect

David Moore (1991) [6] Moore lays down three tests to see if our quantitative data can strengthen (not) the possibility than an association is due to causation.

1. Firstly does the association between two variables occur in a variety of different circumstances (e.g in different periods of history, in all countries, in all towns and villages?). If so there is less chance that third (or fourth or fifth or sixth) factors may explain the association. In the case of our observed association between oat prices and wheat prices we don't know the answer to this question. We have data from a particular period of time in a particular country. We can't be sure if the same association would be observed in other countries or other periods of history -then again we can't dismiss the possibility either. On this basis we do not have evidence that there is a cause and effect association between wheat and oat prices.

Figure 3 Identifying cause and effect: Many of the different possible ways in which we might explain cause and effect. Figures a,b, and c derive from David Moore (1991)[5]

Figure 3a A causes B

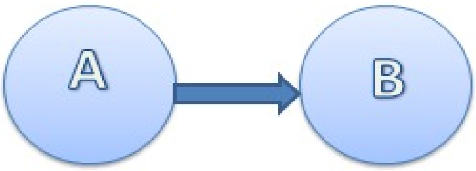


Figure 3c: Changes in B are caused both by A and third variable C.

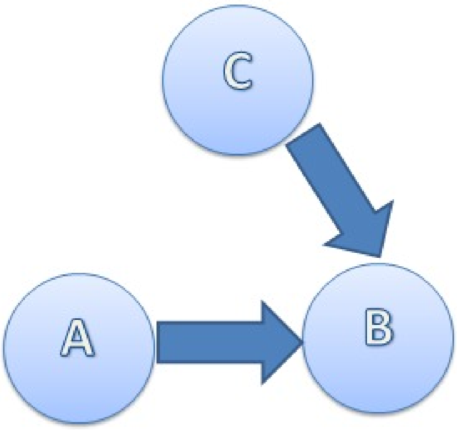


Figure 3e: A has multiple causes

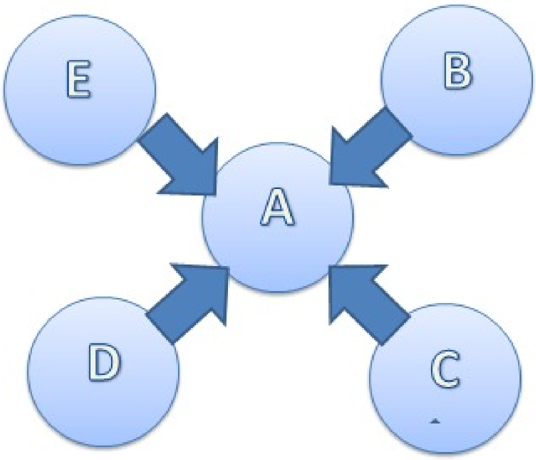


Figure 3b Common response. A and B are caused by a third variable C

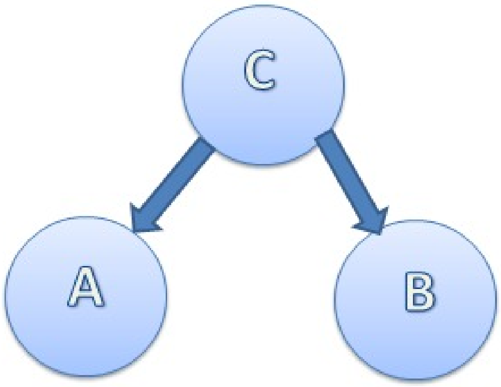


Figure 3d: Apparent relationship has no causality

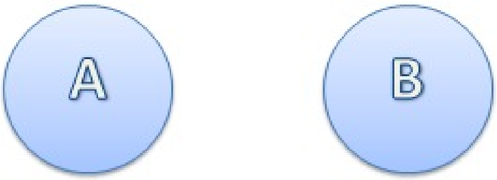
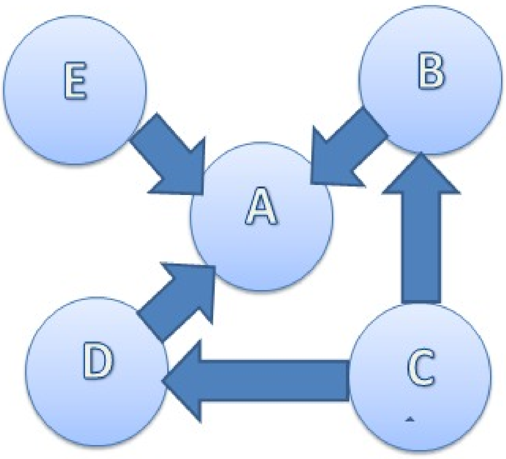


Figure 3f: Complex interaction between different variables.



> only cause of changes in oat prices). `Third' factors which could be as important (or more important) might include the weather (and thereby the harvest) in a particular year. Other factors might include economic policy at the time, war, disease among the human population or agricultural pests. We can see at this point that there are a whole range of directions for investigating the reasons for this association, so our wheat-oats price relationship may be explained by other `third' factors. Notice that I am careful to use the word `association' rather than relationship. What we are looking for is to find the nature of the cause and effect in any association we might observe. It seems reasonable to conclude that there is a relationship of some sort between wheat prices and oat prices.

4 Cautions of cause and effect

Therefore there is much more to cause and effect than getting a high correlation co-efficient or being able to construct a plausible narrative about the data. Here are a few more cautions of cause and effect:

4.1 Direction of cause and effect

When we identify an association between two variables is not necessarily clear in which way any causation is likely to work. From our correlation between wheat prices and oat prices it is plausible that higher wheat prices cause lead to higher oat prices or that higher oat prices lead to high wheat prices. However if we were to look at the relationship between oat prices and weather it is plausible that weather could affect oat prices but it is not plausible that higher oat prices cause changes in the weather.

4.2 Co-incidence

Sometimes the association between two variables is nothing more than a coincidence. This is especially true in the case of time-series data where two unrelated variables have a similar pattern which results in high correlation scores. Yule [7] calculates a correlation coefficient of 0.9512 between the decline in the proportion of all marriages which took place in the Church of England and the decline in the death rate. Any attempt to suggest any direct or indirect reason is nonsense.

4.3 Missing variables/ Simpson's paradox

Table 1 Support for the 1964 Civil Rights Act by party and US region House

House	Democrat	Republican
Northern	94% (145/154)	85% (138/162)
Southern	7% (7/94)	0% (0/10)
Both	61% (152/248)	80% (138/172)

Sometimes conclusions from data can be reversed when groups are merged together or different categories are used. This is sometimes known as Simpson's paradox.

In 1964 61% of Democrat Representatives and 85% of Republican represented voted in favour of the Civil Rights Act. It is clear from these figures that Republicans were much more supportive of the Act than Democrats. Or were they? If we divide the two parties' representatives between those from the North and those from the South we find that 94% of Northern Democrats supported the Act compared to 85% of Northern Republicans. Does this then mean that more Southern Republicans supported the Act that Southern Democrats? It sounds plausible but is not actually true. 7% of Southern Democrats supported the Act compared to 0% of Republicans. On this data it seems that Democrats were more supportive of the Act.

How can this be possible? Take a look at Table 1.

So were Republicans more supportive of the Act or were Democrats more supportive? In a Simpson's paradox case this is really the wrong question. The significant factor was not party affiliation but whether the representatives came from the North or the South.

4.4 Fallacies

Fallacies in statistics can occur when we derive general rules about individuals from group data or general rules about groups from individual data. [8]

4.4.1 The Ecological fallacy

Another form of confounding which can lead to wrong conclusions is the ecological fallacy. An ecological fallacy occurs when we make assumptions about individuals in a group of people based on the

Table 2 Internal competition results for Schools A and B.

School X	'	School Y	'
Harry	9	Lily	10
Amelia	9	Charlie	7
Oliver	9	Jessica	7
Sophie	8	Alfie	7
Jack	8	Emily	5

characteristics of the group as a whole. This phenomenon was first described by W S Robinson who identified that strong correlations between race and illiteracy at the national level were weaker or non-existent in smaller geographical areas.[9]

Think for a moment about two schools. At School A 89% of pupils leave school with five or more GCSE's at grade C or above. At School B just 35% of pupils leave schools with these qualifications. Does this mean that pupils in School A perform better than pupils from school B? More pupils from school A get their GCSE's than from School B, but not every pupil in School A gets better results than every pupil in School B. If we pick a random pupil from each school it is a fallacy to assume that the pupil in school B will not have 5 GCSE's, but the pupil for School A will.

Suppose now that School A and School B are going to compete in a general knowledge quiz. Each school has to choose their five best pupils. Assuming that the five top pupils are likely to be among those with five or more GCSE's at grade C or above, which school is more likely to win? The average pupil in School B is likely to be less able than the average pupil in School A. But the school will not be choosing pupils of average ability for the quiz team - they will each choose their five best.

If having five or more GCSE's is a good predictor of how the pupils would perform on the quiz there is no reason to assume that School A will definitely beat School B on the quiz - in fact their chances are equal. In a different case School X and School Y decide to participate in a French language spelling competition. Each school ran an internal competition using spellings supplied by the organisers and selected the top best scoring pupils. The pupils at School X scored 8.6 on average and School Y 7.2.

Assuming that the internal competition is a reliable indicator of who is most likely to win which pupil is most likely to win?

School X has the higher average score, but the highest score came from School Y. So in an individual competition the pupil with a score of 10 would be most likely to win, even though all his or her fellow pupils scored lower than anyone from School X. However, if the competition is a team competition it is likely that School X will win. The ecological fallacy occurs when the average is taken to be representative of the whole population. This is another reason why averages can be misleading. The schools and their performance at GCSE are descriptions of groups, not of individuals.

4.4.2 The individualistic fallacy

The individualistic fallacy is the opposite of the ecological fallacy. In the ecological fallacy we generalise about the behaviour of individuals from group data. In the individualistic fallacy we generalise about a group from observing individuals in that group. From time to time there will be a story in UK newspapers about a non-working couple with 11 children whose income, derived entirely from the state, is three or four times the national average income of a working household. If the family are shown to have nice possessions, live in a large (taxpayer-funded) house or enjoy foreign holidays, even better. Such cases are very rare -- after all that is why they are the subject of newspaper articles, but the individualistic fallacy allows the case to be extended to the groups sharing some of the same characteristics such as people in the same neighbourhood as the case study family, other families with a lot of children or other non-working families.

4.4.3 Other fallacies

There are various other types of fallacy which can occur. Like the individualistic fallacy and the ecological fallacy these involve making general `rules' either from individual case studies, individual subjects, particular sub-samples of a population or presenting commonly observed behaviours as universal. Outliers and exceptions can always be found, but it is a fallacy to turn them into a general rule. There are 100 year-old smokers, billionaires who failed to finish school, and people have survived horrendous accidents that should have killed them, but such individuals get noticed because they are unusual, not because they conform to a general pattern of experience.

5 Exercises

Table 3 shows the number of agricultural and forestry related accidents in Luxembourg between 1960 and 2009.[10] Think of possible reasons why the number of agriculture and forestry accidents may have decreased.

Writing in 1848 J T Danson [11] reported that between 1840 and 1847 prices of beef, mutton, barley, oats and peas in the United Kingdom increased, but prices of tea, sugar and tobacco decreased during the same period. What possible reasons are there for these differences?

6 References

[1] Doll, R.; Hill, A. B. (1954). "The mortality of doctors in relation to their smoking habits; a preliminary report". British Medical Journal 1 (4877): 1451-1455.

[2] Tobacco Smoking and Lung Cancer Br Med J 1957; 1 doi: http://dx.doi.org/10.1136/bmj.1.5034.1523 (Published 29 June 1957)

[4]R.A. Fisher: Letter in British Medical Journal., vol. II, p. 43, 6 July 1957

[5] Another possibility is a control with Condition X, but no treatment. Apart from being potentially unethical, the real question is whether the new treatment works better than an existing treatment, not that it works better than no treatment.

[6]David Moore (1991) Statistics: concepts and controversies.

[6] David Moore (1991) Statistics:concepts and controversies

[7]Yule, G. U. (1926) Why do we sometimes get nonsense-correlations between time series? A study in sampling and the nature of time series, Journal of the Royal Statistical Society 89, pp.1-16.

[8]A more detailed analysis of different types of fallacies can be found in H. Alker (1969) A Typology of Ecological Fallacies. in M. Dogan and S. Rokkan (ed.). Quantitative Ecological Analysis in the Social Sciences. Cambridge: MIT Press.

[9]Robinson, W.S. (1950) Ecological correlations and the behaviour of individuals. American Sociological Review 15:351-57

[10]Accidents reconnus par l'Association d'assurance contre les accidents, section agricole et foresti\'ere 1960 - 2010, Statistiques Luxembourg. Available from http://www.statistiques.public.lu/stat/TableViewer/tableViewHTML.aspx?sCS_ChosenLang=fr&ReportId=587

[11]J T Danson (1848) A Contribution Towards an Investigation of the Changes which have Taken Place in the Condition of the People of the United Kingdom During the Eight Years Extending from the Harvest of 1839 to the Harvest of 1847; and An Attempt to Develope the Connexion (if any), Between the Changes Observed and the Variations Occurring During the Same Period in the Prices of the Most Necessary Articles of Food, Journal of the Royal Statistical Society of London 11.2, pp. 101--140

Table 3 Number of agricultural and forestry related accidents in Luxembourg between 1960 and 2009.

Year	1960	1970	1980	1990	2000	2009
Number of accidents	3515	2185	1580	1676	762	289

CHAPTER 19: COLLECTING YOUR OWN DATA

1 Designing a questionnaire

On occasion you will want to, or need to collect your own quantitative data in the form of a questionnaire. Despite their familiarity, designing a questionnaire is quite a challenging task. A poorly designed questionnaire will lead to poor quality data and any conclusions you draw on the basis of it will be of poor quality, irrespective of how good you are at statistical analysis.

1.1 Why do we use questionnaires?

To find out people's opinions and/or attitudes.

1. To collect factual data

2. To show relationships between variables. We may be wishing to find out whether men have different opinions or experiences to women, French teachers from German Teachers, footballers from cricketers.

3. To evaluate a product or service. E.g. customer satisfaction surveys. & The findings of our questionnaire may lead us to change our practices.

1.2 Types of questionnaire

1. Face to face. The interviewer is present and fills in the questionnaire in accordance with the respondent's answers.

2. On the telephone. The interviewer telephones the respondent and fills in the questionnaire in accordance with the respondent's answers.

3. A questionnaire where the participant fills in the questionnaire themselves. Traditionally this may have been a questionnaire handed to the participant in person or sent through the mail, but nowadays online questionnaires are increasingly common.

1.3 Issues with questionnaires

1. Response rate

The response rate is the percentage of people targeted to fill in your questionnaire who actually filled it in and returned it. This is often a cause of anxiety. How many returned questionnaires is enough? Is a survey in which 1500 questionnaires are returned from a sample of 70,000 (2.1% response rate) as good as 30 questionnaires returned from a sample of 100 (30% response rate)?

2. Sample size

If a questionnaire is put online where anybody can fill it in it might be argued that the sample size is the same size, or near to the same size as the population. The number of people who actually become aware of your survey will depend on how you publicise it, or invite people to respond, and you will exclude those who do not have access to a computer.

3. Sample population and representativeness

Is your sample representative of the population as a whole? By population here we mean the people who are eligible to fill in your questionnaires. If your questionnaire was aimed at languages undergraduates the population is languages undergraduates (not all the people in the UK/ the world). How confident can we be our sample population is representative of the overall population? There are statistical techniques based on probability theory. You can read more about these in Chapter [chap:sampling]

4. Non-response bias

Non-response bias (not to be confused response bias discussed in Section [sec:response]) is when the people who do respond have characteristics which are different to the people who do not respond. For example it might be discovered that people of a certain gender, ethnic group or social class disproportionately do not respond to your questionnaire.

1.4 Things to think about when designing a questionnaire

Firstly we need to think about what we want to find out. Many projects run into difficulties because the researcher has designed a questionnaire without thinking about what they want to find out first. They design a questionnaire, gather lots of answers then sit there staring at their data hoping that something will interesting will reveal itself. If we don't know what we are trying to find out, we cannot know what questions to ask.

Think also about how you going to analyse the responses? If you have a small number of questionnaires, you may be able to do this manually, but you may need a software programme like SPSS or Minitab or R for a larger dataset. You can also use a database or Excel for small-scale questionnaires.

Have you piloted the questionnaire? Problems in the questionnaire design are most likely to come to light if you pilot the questionnaire with some people in your target audience.

1.5 Advantages of questionnaires

1. Views and experiences of a large number of people can be obtained quickly and efficiently.
2. Questionnaires can be filled in relatively quickly meaning less time commitment from the respondent. Communicating statistics derived from questionnaires can be very powerful in lobbying and reporting to others. In-depth interviews and focus groups usually involve small numbers of people. Some policy makers (and indeed academics) are quite suspicious of the opinions of a small number of people.
3. It is sometimes possible to identify important relationships between variables such as how attitudes vary between gender, language studies, social class etc.
4. It may be possible to identify groups of people who are more likely to act in a certain way. For example, we can identify the characteristics of students who are most likely to drop out of university.
5. Identifying issues which could be addressed. I recently undertook a survey of language learners which revealed that 8% of learners were learning a language which was essential for their job. [1] However only 2% of learners were having their course paid for by their employer. This might suggest that employers are not investing their own money in the skills they need. It is also possible that they got their job by misrepresenting their ability in a that language to their employer.

1.6 Questions about questions

1. Language is very complex and when you are designing a questionnaire you want to be sure that you are communicating as clearly as possible. A question which can seem straight forward in ordinary conversation can be problematic on a questionnaire. If different respondents interpret the same question in different ways, then your data is unlikely to be useful. This is why piloting your questionnaire can be illuminating. When designing your questionnaire think about the following:
2. Does the target population have the knowledge to answer the question? If the people you are asking do not have the required knowledge or insight necessary to make considered judgement or any judgement at all, your data will be of little value. For example if might make sense to ask school teachers which textbook they think is the best for Year 7 beginners' German, but it would not be a sensible question to ask the pupil's parents in a questionnaire about learning languages.
3. Does the question lead to or suggest a particular answer? The UK Census asks "What is your religion?" Residents are invited to select from arrange of choices including `No religion" or `other'. Some commentators, for example those for the National Secular Society favoured the questions "Which of the following best describes you?" with a list of the same answers. On the face of it these seem to be the same question, but some critics claim that the questions are different. If I ask you "What is your religion", there is a possible underlying assumption that you have a religion. If I instead ask "Which of the following best describes you?", you may give a different answer, especially if "No religion" is the first option I put you.
4. Is the question vague? How often do you read La Monde? How often do you attend church? How often do you read novels? Frequently ...Occasionally ... These terms can often be ambiguous what do Frequently ...Occasionally ...etc. mean? Any question with the word `satisfied' or `satisfactory' in it can enable ambiguity to effect your response and be problematic. Does `satisfied' mean the same thing as `adequate' or `fulfilment? Can one be "mostly satisfied"?

5. Questions which it is difficult to disagree with or suggest an action which cannot adversely impact on the respondent. Be careful about the ways you ask people for their views on something it would be illogical to disagree with. For example "Would it be helpful if the library was open in the evenings?" is an easy question to agree with because it costs the respondent nothing. This is a different question to "Would you use the library more, less or the same if it was open in the evenings?"

1.7 Possible limitations of questionnaires

In the opening chapter we commented on how quantitative data is socially constructed as the questions we ask reflect our values and priorities. This is equally true of data we collect ourselves.

1. Categorising:

Limitations include: putting people into categories: The English Church Census asked clergy whether they identify as liberal, evangelical, or broad in their churchmanship, but it is not necessarily clear whether each of the respondents understand the exactly the same thing by these different terms. Some people feel unable to identify either as male or female. For equal opportunities purposes some employers monitor whether their employees identify as heterosexual, homosexual or bisexual, but some people may feel that none of these categories adequately describe their sexual identity.

2. Question wording

If a respondent or group of respondents has misunderstood the question, you may never find this out, especially if you are not present when the questionnaire was filled in.

3. Not telling the truth

- Deliberate lying is particularly an issue where the interviewer is present, but not telling the truth can be done accidentally too.
- Sensitive topics: such as sex, attitudes to race, gambling or income are particular areas in which respondents may be disinclined to tell the truth, even anonymously.
- Response bias: People may answer `yes' to the question "Would you be interested in doing a month long intensive online French course?" because it seems like a good idea at the time of questioning, but when faced with the actual possibility of doing such a course realise they have other priorities. Situations in which the interviewer is present can cause what is known as response bias. People will often say what they think the interviewer wants them to say. This can be a particular issue if the respondent holds opinions which they think are unpopular or offensive or different from their own assessment of the interviewer's opinions.

4. Complexity of individual experiences

Individual experiences are very complex. It is not really possible to get into the depths and complexities of individual experiences in questionnaires. In-depth interviews are much better for this although it is possible to gain some useful insights from open-ended items in a questionnaire.

5. Questionnaires in support of or opposition to certain moral positions

Questionnaires can shed light on the popularity or unpopularity of certain opinions which can form the basis for action. However, a majority opinion does not justify a particular view or action in moral terms.

6. Exclusion of certain groups of people

When interviewing by telephone or online think about what sorts of people might be excluded from the process.

Think about who answers the phone in any given household.

Is everyone in your target group computer literate? An online questionnaire is unlikely to be the most appropriate way of questioning elderly people or very young people.

1.8 Costs and logistics

If you wish to administer your questionnaire in person there are costs in actual money (e.g if you have to pay someone to do the interviews or in time (other things you might be doing, including earning money). If you are sending out questionnaires by post the cost of sending out and organising the return of postal questionnaires needs to be considered. One of the great advantages of online questionnaires is that most of the these costs are minimised.

1.9 Ethical considerations

Collecting your own data raises ethical issues. People who respond to your questionnaire have a right to know what the research is about and what you are going to do with their data. If you are studying at a university or college there may be forms you need to fill in. At the very least you need to write a sentence or two at the top of the questionnaire which explains what the research is about and how they can contact you.

2 Exercises

1. What do you think the advantages and limitations of each of the ways of administering a questionnaire?

A)Online

B)By telephone

C) Face to face

Handing out questionnaires for people to fill in and return to you.

2. Read through these questions. Which questions are good questions? How might they be improved? Each question is an example of the sort of question that you might see on a questionnaire and they are not meant to be taken together.

Should languages be compulsory at GCSE in order to increase numbers of students studying languages in higher education? Yes/ No/ Don't know

What do you think of language learning in the UK?.................................. ....

Do you agree or disagree that the oral abilities of students at your institution's evening language programme are better than those of your competitors. YES/ NO/ DON'T KNOW

Do you agree or disagree that UK students speak better Spanish now than they did twenty years ago? YES/ NO/ DON'T KNOW

3 References

[1] Canning, J. (2011) Survey of non-specialist language learners Southampton: UCML.

CHAPTER 20: PRESENTING DATA

1 Everyday data

Every day we come across numbers on the news, on advertisements, in newspapers, on signposts, on packets and boxes. These numbers tell us how many grams of cornflakes are in the packet or how far it is to London, at what time the train to Southampton will be leaving and how many grams of calcium and fat are found in a serving of yogurt. This numerical data helps us to decide which way to go, to make decisions about what to buy or what to eat.

In England the Government publishes data about schools and how many of their pupils are getting the top grades in their exams. We read in the newspaper that inflation has gone up by 2.1% or that 30,000 fewer people are unemployed this month than last month. The way we present numerical data is so important to avoid misunderstanding, misinterpretation and confusion.

Imagine that the road signs were so small that you had to stop your car in order to find out how far away it is to London. Maybe you have been at a train station where the information screens have broken down are there is no information about what time the trains leave and what platform the train leaves from. Likewise, the way we present numerical data in our humanities research is very important if other people are to understand our findings and how we interpret our findings. It is often said that a picture is worth 1000 words —the same can be said for a table or a graph. A table or graph is often a better way of communicating a large amount of numbers than hundreds of words of text.

2 Data Cleaning

An often neglected issue in data analysis and presentation is that our data does not always come to us in good form. In order to present data well we need to read through it to look for possible mistakes or errors in the data.

In I have attempted to produce a chart of literacy in Castelffrench from the 1911 Census of Ireland using Minitab, a statistical analysis package. I copied and pasted the data from the website into Minitab, but something went wrong. Those filling in the questionnaire were asked "State whether he or she can `Read and write', `Read only' or `Cannot read"'. There were only three possibilities. However through misspellings and inconsistencies the residents of Castleffrench have managed to come up with a total of 15 different answers (Table 1).

The online version of the census is simply a reproduction of what the householder or census enumerator wrote on the form. The website reproduced inconsistent ways of recording that someone could read and/or write (e.g. Read and write, read/ write, read only, can read). Spelling mistakes are also reproduced, e.g. Cannor read, read right and connot. The software package does not know how to make sense of this data and simply treats each one as a different category. In fact there are only three actual possibilities: Read only, read and write, and cannot read or write - there is no evidence here that anyone claimed to be able to write but not read.

After cleaning my data I was able to produce a revised version of my graph (Figure 2). There are a number of possible issues to look out for prior to using data. Some examples include:

Table 1 Actual census entries have been placed into the recognised categories

Can read and write	Read only	Cannot read or write
Can read and write	Can read	Cannor read
Read and write	Read	Cannot read
Read right	Read and only	Cannot Read
Read write	Read only	Cannot read and write
	read only	Connot

Figure 1 Minitab output of literacy question on 1911 Census of Ireland

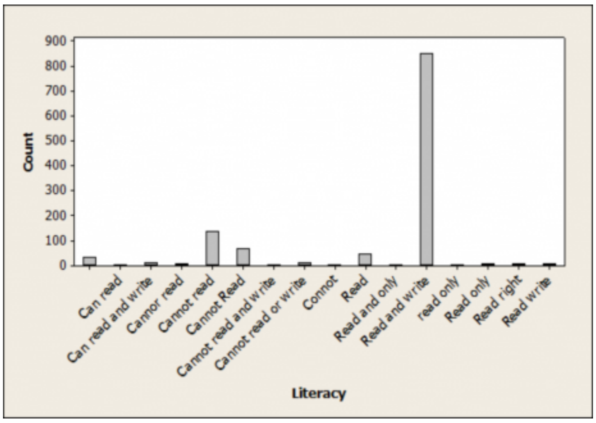


Figure 2 Minitab output of literacy question on 1911 Census of Ireland: Cleaned data

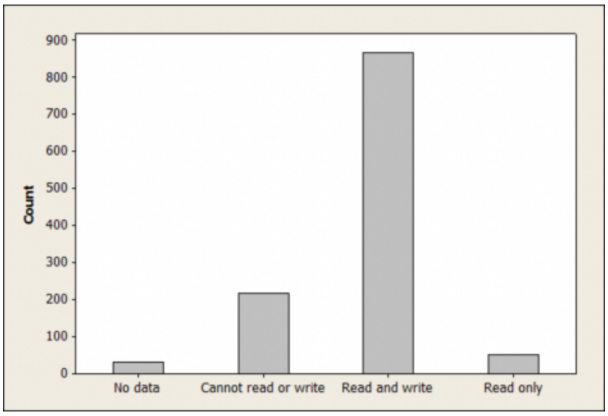


Figure 3 Principal occupations of Castleffrench residents, 1911(Source: 1911 Census of Ireland)

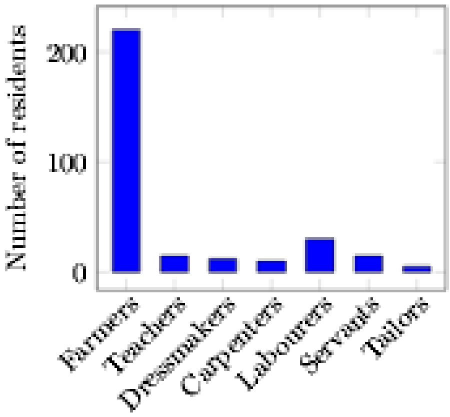
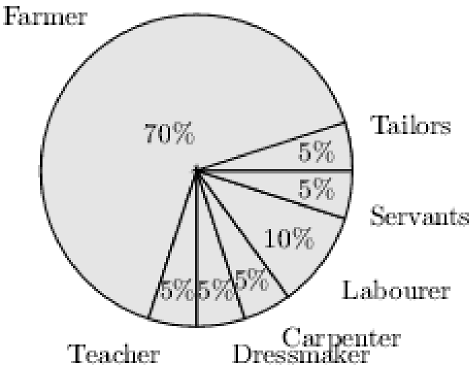


Figure 4 Principal occupations of Castleffrench residents, 1911(Source: 1911 Census of Ireland)



Spelling mistakes or inconsistent spelling: This can come from autocorrect in software.

- Different captitalisation use: A statistical analysis software package may treat `French', `FRENCH' and `french' as different categories.
- Instructions not followed by person filling in the form or entering the data (as in Table 1).
- Data entered into the wrong columns or in the wrong order: For example writing `John, Smith' where `Smith, John' should have been entered
- Unlikely answers: An unusually high (or low) income may have a decimal point in the wrong place.
- Impossible answers: Men recorded as having given birth.
- Mixing of different units of measurements: Mixture of pounds, stones, kilograms etc.

3 Ways of presenting data

There are several ways of presenting numerical data.

3.1 Text

We can simply present our numerical data in the form of text. For example:

> According to the 1911 Census 1,167 people were living in Castleffrench, 624 of who were males and 543 were females. 396 (34%) reported that they spoke Irish.

Reporting through text is perfectly appropriate in many cases. If you are writing a report or paper where you are not using much numerical data and the data is not especially complex then presenting data as part of an ordinary paragraph is probably going to be fine. If there are more details then it is likely that a table or graph will be more useful.

3.2 Tables

Tables are a good way of presenting small amounts or aggregated data clearly and concisely. They can be used for presenting simple information. Table 2 shows the population by sex and Table 3 by language spoken. Table 4 shows two a way of conveying these two sets of information together.

Note that the tables are numbered, have a title and the source of the data is referenced.

4 Representing data graphically

There is a common saying that a picture is worth 1000 words. Done well, graphical representations of data can convey a huge amount of information more clearly and quickly than tables or text. Done poorly graphical representations of data can be confusing, misleading or simply pointless.

Table 2 Population of Castleffrench by sex (Source 1911 Census of Ireland)

Female	Male	Total
543	624	1,167

Table 3 Population of Castleffrench by language spoken (Source 1991 Census of Ireland)

Irish speakers	Non-Irish speakers	Total
396	771	1,167

Table 4 Population of Castleffrench in 1911 by sex and language spoken (from 1911 Census of Ireland)

Language	Female	Male	Total
Irish speakers	193	203	396
Non-Irish speakers	436	446	882
Total	629	649	1278

Bar charts

There are several different types of chart. The most familiar is probably the bar chart (see Figure 3).

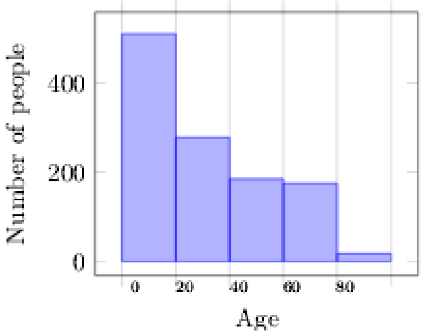
Pie charts

The same data can also be presented as a pie chart (Figure 4). Pie charts show each piece of data as a proportion of the total. They can be made colourful and/or 3D. Some writers cautions against pie charts as they are easier manipulated. [1] This sound advice. Alternative ways of presenting data such as bar charts or tables convey the same information far more clearly.

Histograms

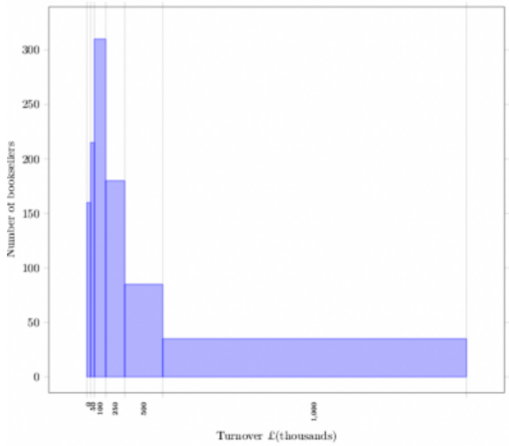
A histogram is a form of bar chart which shows the frequency and distribution of continuous data. Notice how the bars are right next to each other in Figure 5.

Figure 5 Ages of Castleffrench residents, 1911(Source: 1911 Census of Ireland)



This is because age 1 comes after zero, 2 after 1, 3 after 2 etc. Another example of continuous data is exam results: 1 comes after 2 followed by 3 or Grade A is followed by Grade B, followed by Grade C etc. This is an important distinction from categorical data such as occupations. Farmer, teacher and dressmaker are discrete variables. The grade A student has performed better than a grade B student on the exam and if the grade B student might do better next time and get an A or worse and get a C. However male students and female students are discrete variables. Doing better or worse on a test might change the

Figure 6 Area histogram: Bookseller turnover)



student's grade, but not their sex. Getting older will change a person's age but not their occupation.

Figure Figure 6 is a slightly different sort of histogram called an area histogram. The area histogram shows the number of booksellers at each level of annual turnover. This special type of histogram uses a scale on the x axis as well as the y axis, so not only do the bars vary by height but also by width. Therefore it is possible to compare the area of the bars. In the case of Figure 6 we can see that although the bar for the booksellers with a turnover of over £1,000,000 per year is short it is very wide. By comparing the areas of the bars we can see that the small number of very large booksellers generate more revenue that the larger number of booksellers with a lower turnover.

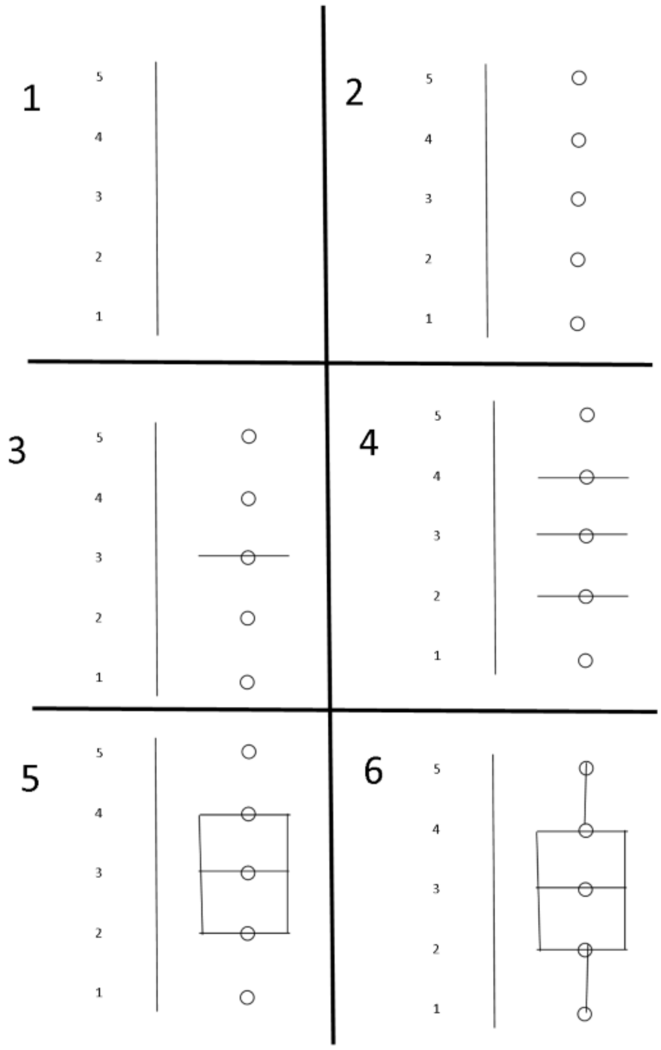
Boxplots

Another way of displaying the frequency of data is to draw a boxplot. The 'box' in the middle is drawn between the upper quartile and the lower quartile. The line through the middle of the box is the median (though the mean is sometimes used). The line coming out of the box joins the box to the highest observed value. The line coming out of the bottom of the box joins the box to the lowest observed value.

Drawing a boxplot

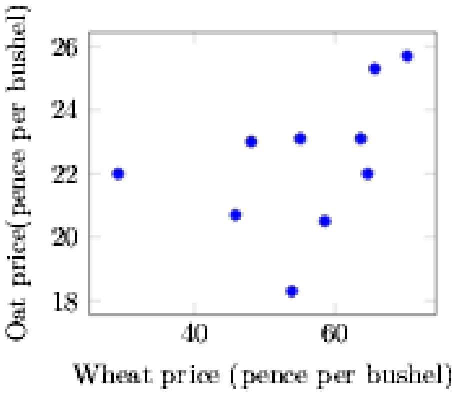
Boxplots can sometimes look confusing, but they are very simple to draw. Suppose that our data has a median of 3, an upper quartile of 4, a lower quartile of 2, a maximum of 5 and a minimum of 1.

Figure 7 How to draw a box plot



1. Draw a vertical line and mark the numbers 1,2,3,4,5 (with the 5 at the top and the 1 at the bottom).
2. Mark the median, upper quartile, lower quartile, maximum and minimum with small dots.
3. Draw a horizontal line through the median.
4. Draw horizontal lines through the upper quartile and lower quartile.
5. Draw two vertical lines to join up the left and right hand ends of the three lines to make a box.
6. Finally draw a vertical lie from the upper quartile mark to the maximum and the lower quartile to the minimum. You now have finished drawing your boxplot.

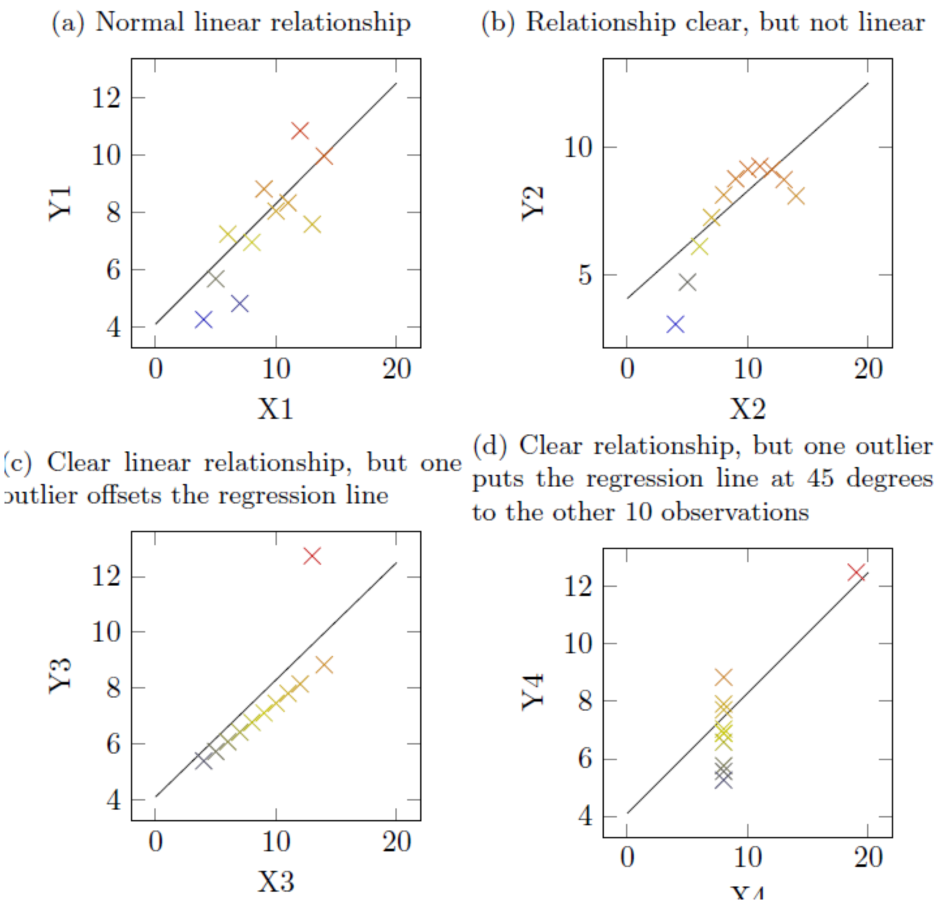
Figure 8 Wheat and oats prices



Scatterplot

A scatter plot can help to identify possible relationships between variables. Figure 8 shows the plots the price of oats on the vertical (y) axis and the price of wheat on the horizontal (x) axis. From the plot below we can see that the price of oats remained fairly constant but the price of wheat varied a lot more.

Figure 9 Anscombe's quartet [2] is a good demonstration why a scatterplot is so valuable, prior to calculating regression equations and correlation coefficients. In all four cases the x's have a mean of 9, and variance of 11. The mean of all the y's is 7.5, and a variance 4.125. The correlation co-efficient of each is 0.816 and the linear regression line is y=3+0.5x


(a) Normal linear relationship

(b) Relationship clear, but not linear

(c) Clear linear relationship, but one outlier offsets the regression line

(d) Clear relationship, but one outlier puts the regression line at 45 degrees to the other 10 observations

Figure 10 Estimated copper emissions linked with copper production: 5000 years ago to present (linear scale)

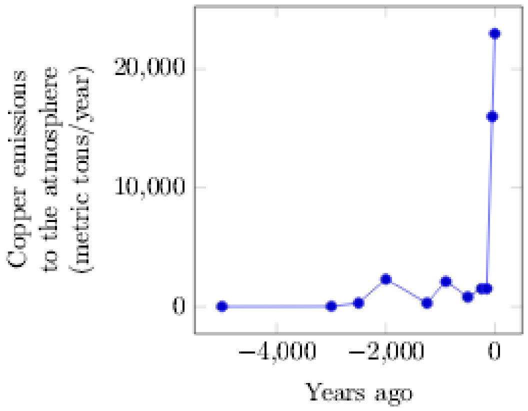


Figure 12 Frequency (Hertz) of nine C notes (linear)

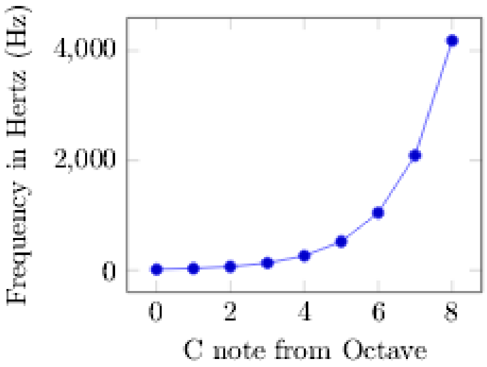


Table 5 Estimated copper emissions to the atmosphere linked with copper production (5000BP-5BP)

Date (years ago)	Copper emission to the atmosphere (tonnes/year)
5000	5
3000	20
2500	300
2000	2300
1250	300
900	2100
500	800
250	1500
150	1500
50	16000
5	23000

Figure 11 Estimated copper emissions linked with copper production: 5000 years ago to present (logarithmic scale)

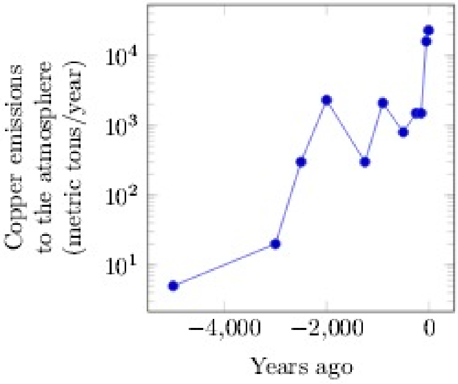
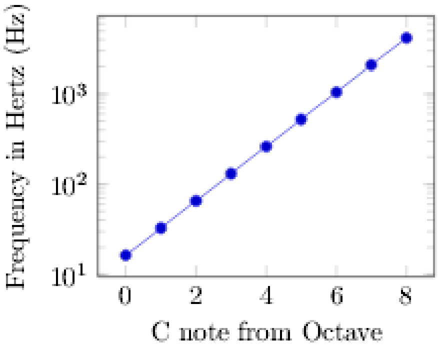


Figure 13 Frequency (Hertz) of nine C notes



Logarithmic scales

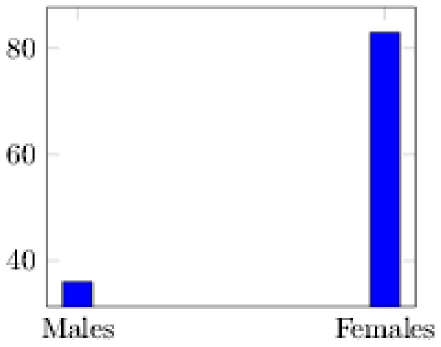
Table 5 shows estimated global copper emissions over the past 5000 years. In the original paper the authors present this data in table form. [3] If we look closely that the data we can see that there have been some big proportional increases over the past 5000 years. Emissions 3000 years ago were four times the level they were 5000 years ago. By 2500 ago emissions were 300 metric tons per year 15 times the levels 500 years previously. By looking at the table closely we can see the that the growth as not been consistent and there were peaks in emissions 2000 years ago and 900 years ago emissions actually declined before huge rises in the past 50 years.

Figure 10 shows a simple line graph of the data. The huge rise of the past 50 years is clear, but the peaks of 2000 and 900 years ago are not easy to see. The pattens of increases and decreases is hard to see and most of the graph space is unused. Figure 11 is a logarithmic version of the same graph. Instead of having a y axis going from 0 to 23,000 in equal proportion the scale is logarithmic. Instead of an equal distance between 0 and 1, 1 and 2, 2 and 3 etc., all the

Table 6 Frequencies (Hertz) of C notes

Notes	Frequencies
C0	16.35
C1	32.7
C2	65.41
C3	130.81
C4	261.63
C5	523.25
C6	1046.5
C7	2093
C8	4186

Figure 14 Number of respondents: by sex



way up to 23,000 the equal spaces are in order of magnitude. The first number on the y axis is 101 which is 10. The second number which is the same distance as the gap between 0 and 101 is 102 which is the same as 10×10 which equals 100.So there is a gap of 10 between the 0 and first point of 101 but the difference between 101 and 102 is 90. The third number is 103 (or 10×10×10) which equals 1000. The fourth number is 104 (or 10×10×10×10) which equals 10,000. On the logarithmic version of the graph the ups and downs are clearer and the trends are easier to spot. On a logarithmic graph equally spaced divisions on the y axis now mean multiplication rather than addition.

Logarithmic scales are frequency found in natural phenomena. An example which might be of interest here is the relationship between musical notes the their frequency in Hertz.[4]

A full size piano has nine C notes. The lowest C note (*C*0) has a frequency of 16.35 Hertz (Hz). The C note one octave higher (*C*1) has a frequency of 32.7 Hz twice that of *C*0. *C*2 is one octave higher than *C*1 and has a frequency of 65.41 Hz. If we examine Figure 13 we can see that this pattern continues{ each C note has a frequency twice the value of the C note one octave below it and half the value of the C note above it.

We can see the clear that logarithmic graphs make when we plot Table 6 as a graph. Figure 12 is a simple linear graph. When the dots are joined together an upward curve is created. When we plot the same data onto a logarithmic scale we can see that the upward curve has become a perfect straight line. (This pattern also applies to the notes A, B, D, E, F, G and all the semitones in between).

Issues in data presentation

Irrelevant and pointless graphs

Figure 14 is a pointless graph. It displays the very basic information that there were 36 males and 83 females in the sample. There is no need for this graph.

Manipulating axis

Figure 15 and Figure 16 are actually the same graph. They contain the same data over the same period of time. The dip is clear in Figure 15 but you can only just make out the dip in 1835 in Figure Figure 16 if you look very carefully, but otherwise it looks like prices hardly changed between 1830 and 1839. So what has changed? All I have done is stretched out the price axis up to 1000. The highest datapoint is 70.3 so there is no need for a vertical axis going up to 1000. It

Figure 15 Wheat price 1830-1839

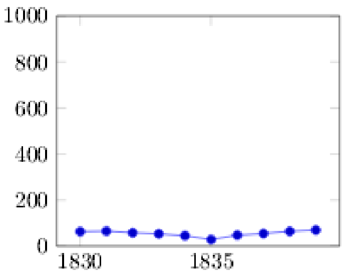


Figure 16 Wheat price 1830-1839

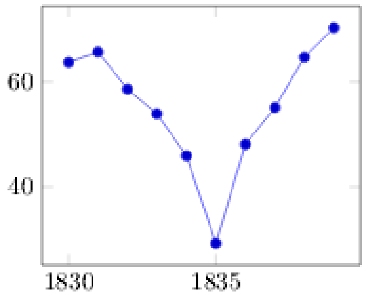


Figure 17 Gross Domestic Product in Cyprus between 2000 and 2006

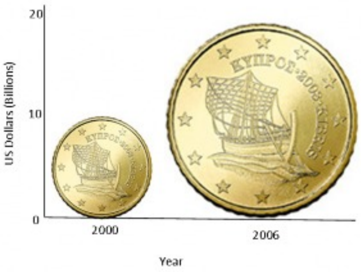


Figure 18 Chartjunk: The chart uses lots of colours, patterns and grid lines to the point at which it becomes virtually impossible to read the graph. The use of red and green is a particular challenge for those with colour-blindness.

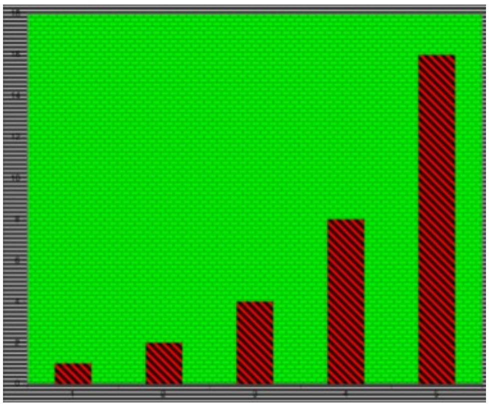
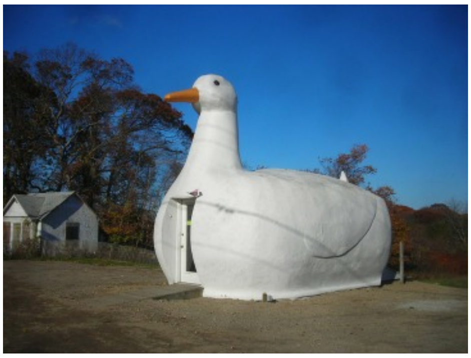


Figure 19 Big Duck in Flanders, New York



has created a lot of empty space and made it more difficult to spot the trend.

Misleading graphics

Picture graphics are often used to show impact. Figure 20.20 shows the growth of Gross Domestic Product in Cyprus between 2000 and 2006. In 2000 GDP was $ 9.31bn, but by 2006 it had grown to $18.44bn, twice the 2000 level. As you can see the Cypriot Euro coin for 2006 is about twice the size of the coin for 2000. But wait! Is the coin really twice as big? The 2006 coin is twice as tall and twice as wide as the 2000 coin. If we think in terms of the area of the coin, the 2006 coin is actually four times the size of the 2000 coin. This use of graphics is commonplace in the graphs that appear in popular newspapers and magazines. It is liable to mislead the unwary or inexperienced user of graphs, hopefully not deliberately though.

Chartjunk

`Chartjunk' is a term coined by Edward R Tufte [5] to describe unnecessary `decoration' which appears in graphs. Sometimes the decoration is used to to make a graph look pretty by using various different patterns and colours.

Other forms of chartjunk include unnecessary grid lines, self promoting graphics, and any use of ink which does not directly contribute to the reader's understanding of the graph. Figure 18 is a deliberate example of chartjunk. Tufte calls elaborate statistical graphics `ducks'. The term duck is inspired by Big Duck in Flanders, New York (see Figure 19). The whole structure of Big Duck is simply decoration and its design does not serve any useful purpose. David McCandless, the owner of the website http://www.informationisbeautiful.net describes his passion as “visualizing information -- facts, data, ideas, subjects, issues, statistics, questions - all with the minimum of words.[6] His graphs are undoubtedly beautiful, but are not necessarily the best way of communicating the data.

References

1. See Edward R Tufte (2001) The Visual Display of Quantitative Information Cheshire CT: Graphics Presse p.178 or Stephen Few (2007) Save the Pies for Dessert Visual Business Intelligence Newsletter http://www.perceptualedge.com/articles/08-21-07.pdf}
2. F. J. Anscombe (1973) Graphs in Statistical Analysis, The American Statistician, 27 pp. 17-21
3. Hong, Sungmin, Jean-Pierre Candelone, Michel Soutif, and Claude F. Boutron. “A Reconstruction of Changes in Copper

CHAPTER 21: THE NEXT STEPS: DEVELOPING UNDERSTANDING IN STATISTICS

As I noted in the preface, statistics courses are often seen by humanities and social science students as a hurdle to be overcome, rather than the beginning of developing a lifelong skill. For too many students passing a compulsory statistics course signifies an end rather a beginning. If statistical skills are required in the future they have to be re-learnt.

A good cook understands that a good stew needs to be cooked for a long time if the meat is to become tender and the meal enjoyed. If not given long enough to cook the meat is chewy or uncooked and the flavours remain separate and unfused. Although you can eat such a meal you are unlikely to enjoy it very much and you won't be in a hurry to cook the same meal again. Just because you can eat such a meal does not mean that you should. Rushing through a book on statistics may get you through a beginners statistics course, but little else. For statistics to become a lifelong skill you need to return to again and again and slowly digest.

1 Software

Where you go next will depend on what you want to get out of learning statistics. It is definitely worth gaining some familiarity with at least one statistical analysis software package (a brief introduction is in Appendix M). If you are studying at a college or university you may be taught a specific package, so this choice will be made for you. These packages are very powerful and if used well will be very helpful.

2 Further reading in statistics

2.1 Discipline specific books

This books has been written as a general introduction for humanities students. If you are an undergraduate student of history, archaeology, linguistics etc., your lecturer may recommend particular texts. Examples include Pat Hudson,*History by Numbers: An Introduction to Quantitative Approaches.* (Bloomsbury Academic, 2000, Zolt\'an D\"ornyei, *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies* (OUP Oxford, 2007) and Stephen Shennan, *Quantifying Archaeology* 2nd Edition (Edinburgh: Edinburgh University Press, 1997). Some statistics books make reference to software packages in introducing statistics, but technology and software develop so quickly that these books quickly look very dated.

2.2 General books and popular books

In a large university library and you will find hundreds, possibly thousands of statistics books. Some of these are very specialist, and some described as introductions as not particularly accessible. There are books aimed at people in different disciplines as well as more general texts. The advantage of going to a library is that you can browse through the books to see if they are likely to meet your needs. 'Popular' books can also be valuable in develop the skills required to think critically about statistics. Darryll Huff's book, *How to lie with statistics* [1] is both informative and entertaining. More recent books such as Levitt and Dubner's *Freakonomics* [2] and Ben Goldacre's *Bad Science* [3] and The Tiger That Isn't: Seeing Through a World of Numbers [4]can also help you to think about statistics in a critical, yet accessible way.

2.3 Journals

Statistics journals tend to specialise technical matters and debates about statistics. However Radical Statistics is an open access journal with very accessible articles which use statistics and statistical reasoning.

2.4 TV and Radio

More or Less on BBC Radio 4 (UK) provides a weekly look at the statistics behind the news.

3 Getting help online

Finding suitable help with statistics online can be surprisingly difficult for the beginner. A simple internet search on any statistical test will bring up thousands of websites, but these vary considerably in terms of quality and the level they are pitched at. Many of these webpages have been put up by university lecturers in support of face-to-face or on-line courses so don't always work well as independent resources, though some can be very helpful. Additionally, most of these materials are aimed at science or social science students rather than humanities students.

Wikipedia has very little to offer the beginner. Most articles about the statistical tests covered here are written by statisticians and go into a lot of theory behind the tests.

YouTube is a good source of material for the beginner, though again quality and level of previous knowledge expect varies considerably. Daniel Judge's statistics lectures are particularly valuable for the beginner. His approach is very engaging and the pace is appropriate for those who are new to statistics.

William M.K. Trochim's Research Methods Knowledge Base is a useful follow up to this website, introducing new terms and concepts. The nature of the web is such that new material is being added all the time, so its always worth looking out for new materials.

4 References

1. Darrell Huff, How to Lie with Statistics, W.W. Norton & Co., New York, 1954
2. Steven Levitt and Stephen J. Dubner (2005). Freakonomics: A Rogue Economist Explores the Hidden Side of Everything. William Morrow/HarperCollins
3. Goldacre, Ben (2008). Bad Science. London: Fourth Estate
4. Andrew Dilnot and Michael Blastland (2008)The Tiger That Isn't: Seeing Through a World of Numbers (London :Profile)

APPENDIX: DEALING WITH NON-DECIMAL UNITS.

Metric units have been used through this book. You may come across measurements in non-metric (non-base 10) measurements, e.g. feet, inches and pre-decimal (pre-1971) UK currency.

Non-base 10 units difficult to work with for statistical analysis unless you either convert to metric or convert the overall measurement into a single unit (e.g. convert pounds, shilling and pence to just pence or feet and inches to just inches. If you are comparing a dataset with metric measurements with a dataset with imperial measurements it is preferable to convert to metric for your comparisons.

1 Feet and inches

There are 12 inches in 1 foot. If we want to perform statistical tests on measurements in feet and inches we need to convert to just inches.

To convert height in feet and inches to just inches

For example:

If a man is 5 foot 8 inches (5' 8") tall, how tall is he in centimetres?

First we need to find the man's height in inches.

We start with the number of feet. There are 12 inches in a foot, so we multiple 5 by 12.

5×12=60 (5 feet=60 inches)

Now we need to add on the 8 inches

60inches+8inches=68inches

Convert to centimetres

We may now wish to convert our inches into centimetre. To convert height in inches to centimetres. You may also need top do this if you have a dataset which is a mix of decimal and on-decimal measurements.

Multiple the number of inches by 2.54 (the number of centimetres in an inch)

68 inches x 2.54 = 172.7 cm

The a man who is 5 foot, 8 inches tall is 172.7 centimetres tall.

2 Stones and pounds

There are 14 stones in one stone. In the UK weights are often given in stone (st) and pounds (lb)

To convert stones and pounds to pounds

If a man weighs 10 stone, 12 pounds,

We start with converting the weight to pounds.

There are 14 pounds in a stone.

10 stone×14=140(10 stone=140 pounds)

Now add on the twelve pounds: 140+12=152. The man weighs 152 pounds.

To convert pounds to kilograms

To convert 152 pounds to kilograms we need to multiple by 0.454 as there are approximately 0.454 kilograms in a pound.

152 pounds×0.454=69.01 kilograms

3 Pre-decimal currency

Prior to 1971 the UK used a non-decimal currency. Documents concerned with money before 1971 (e.g. wages, taxes, prices) will be expressed in in pre-decimal currency Pounds (£), Shillings (s) and Pence (d). Money values would be written in the following format: £ 5 7s 4d. `d' was used for pence from the Latin `denarius'.

It will be easier to manipulate the data if you convert to pence first.

Firstly we need to know that:

240 pence = 1 pound

12 pence = 1 shilling

20 shillings = 1 pound

So how many pence is £ 5 7s 4d?

First find how many pence in 5 pounds.

5×240=1200 (1200 pence in 5 pounds)

Secondly find how many pence in 7 shillings

7×12=84 pence

Thirdly find add the answer above together in.

1200 pence + 84 pence = 1284 pence (5 pounds and seven shillings)

Finally add on the 4 pence.

1284+4=1288 pence

When reporting your answers you will usually want to convert the pence back to pound shillings and pence to report the mean, median etc.

What is 1288 pence in pounds, shillings and pence?

To find the number of pounds

1288÷240=5.367

This shows us the number of whole pounds. Take the number of whole pounds and ignore the numbers after the decimal point.

Work out how many pence in 5 pounds

5×240=1200

Subtract the 1200 pence from 1288

1288−1200=88pence

Now we have 88 pence left. To find the number of shillings in 88 pence:

8812=7.33 shillings. Take the number of whole shillings and ignore the number after the decimal point.

Work out how many pence in 7 shillings.

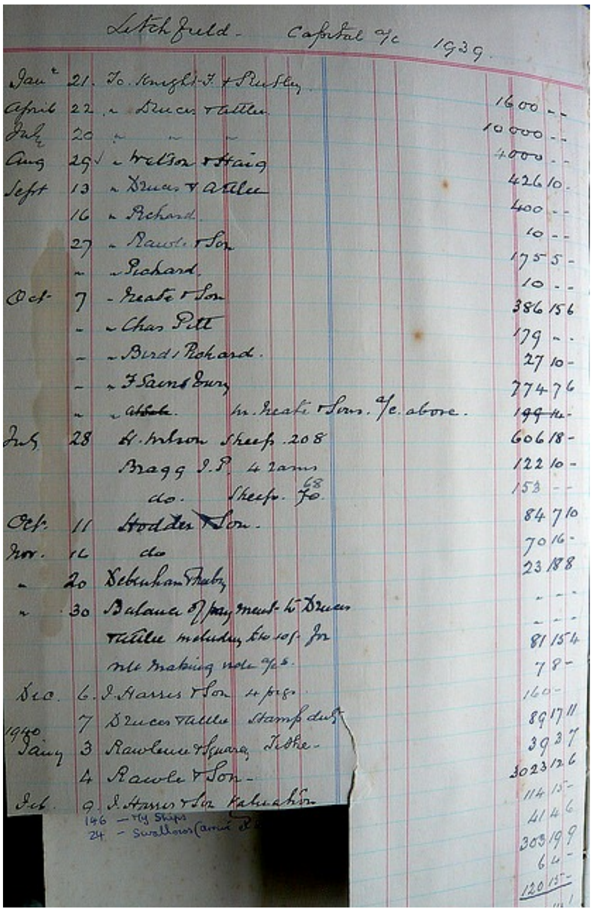
7 shillings×12 pence=84 pence

Subtract the 84 pence from the 88 pence

88 pence−84 pence=4 pence

Therefore 1288 pence equals 5 pounds, 7 shillings and 4 pence or £5, 7s, 4d.

Figure 1: A page from Patrick's Litchfield account book 1939 showing prices in pounds, shillings and pence.



Litchfield. Capital a/c 1939.

Janʸ	21.	To Knight J. & Sudley	1600	-	-
April	22.	~ Druce & Tattle	10000	-	-
July	20	~ ~ ~	4000	-	-
Aug	29✓	~ Wilson & Haig	426	10	-
Sept	13	~ Druces & Attlee	400	-	-
	16	~ Richard	10	-	-
	27	~ Rawle & Son	175	5	-
	~	~ Richard.	10	-	-
Oct	7	~ Keate & Son	386	15	6
	~	~ Chas Pitt	179	-	-
	~	~ Bird / Richard.	27	10	-
	~	~ J Sainsbury	774	7	6
	~	~ ~~abbot~~. in Keate & Sons. a/c above.	~~199~~	~~14~~	-
July	28	H. Wilson Sheep. 208	606	18	-
		Bragg J.P. 4 rams	122	10	-
		do. Sheep. 68 70	153	-	-
Oct	11	Hodder & Son.	84	7	10
Nov.	12	do	70	16	-
~	20	Debenham & Tuby	23	18	8
~	30	Balance of payment to Druces	-	-	-
		Tattle including the 10/- for	-	-	-
		rate making note a/c.	81	15	4
Dec.	6.	J. Harris & Son 4 pigs	7	8	-
	7	Druces Tattle Stamp duty	160	-	
1940 Jany	3	Rawlence & Squarey Tithe.	89	17	11
	4	Rawle & Son.	39	3	7
Feb.	9.	J. Harris & Son Valuation	3023	12	6
		146 — Fly Ships	114	15	-
		24 - Swallows (ann	41	4	6
			303	19	9
			6	4	-
			120	15	-

Figure 2: Shilling from 1958

Online resource

An excel file to assist in performing calculations using
pre-decimal currency is available from
http://www.statisticsforhumanities.net

Critical values of Chi square

Confidence	0.9	0.95	0.99		0.9	0.95	0.99
Degrees of freedom				Degrees of freedom			
1	2.705543454	3.841458821	11.34486673	26	35.56317127	38.88513866	61.16208676
2	4.605170186	5.991464547	15.08627247	27	36.74121675	40.11327207	63.69073975
3	6.251388631	7.814727903	18.47530691	28	37.91592254	41.33713815	64.95007134
4	7.77944034	9.487729037	21.66599433	29	39.08746977	42.5569678	66.20623628
5	9.2363569	11.07049769	24.72497031	30	40.25602374	43.77297183	67.45934792
6	10.64464068	12.59158724	26.21696731	31	41.42173583	44.98534328	68.70951297
7	12.01703662	14.06714045	29.14123774	32	42.58474508	46.19425952	71.20140025
8	13.36156614	15.50731306	30.57791417	33	43.74517956	47.39988392	72.44330738
9	14.68365657	16.9189776	31.99992691	34	44.90315752	48.60236737	73.68263852
10	15.98717917	18.30703805	34.80530573	35	46.05878844	49.80184957	74.91947431
11	17.27500852	19.67513757	36.19086913	36	47.21217389	50.99846017	76.15389125
12	18.54934779	21.02606982	38.93217268	37	48.36340835	52.19231973	78.61575572
13	19.81192931	22.36203249	40.28936044	38	49.51257983	53.38354062	79.84333812
14	21.06414421	23.6847913	41.63839812	39	50.65977049	54.57222776	81.06877191
15	22.30712958	24.99579014	42.97982014	40	51.80505721	55.75847928	82.29211683
16	23.54182892	26.2962276	45.64168267	41	52.948512	56.94238715	83.51342993
17	24.76903534	27.58711164	46.96294212	42	54.09020245	58.12403768	85.95017625
18	25.98942308	28.86929943	48.27823577	43	55.23019209	59.30351203	87.1657114
19	27.20357103	30.14352721	50.89218131	44	56.36854073	60.48088658	88.3794189
20	28.41198058	31.41043284	52.19139483	45	57.50530474	61.65623338	89.59134449
21	29.61508944	32.67057334	53.48577184	46	58.64053738	62.82962041	90.80153203
22	30.81328234	33.92443847	54.77553976	47	59.77428893	64.00111197	93.21685966
23	32.00689968	35.17246163	57.34207343	48	60.90660703	65.1707689	94.42207901
24	33.19624429	36.4150285	58.6192145	49	62.03753679	66.33864886	95.625719
25	34.38158702	37.65248413	59.89250005	50	63.16712101	67.50480655	96.82781556

Critical values of t (2 tail test)

Confidence level	**90%**	**95%**	**99%**	**99.90%**
Probability	0.1	0.05	0.01	0.001
Degrees of freedom				
1	6.313751515	12.70620474	63.65674116	636.6192488
2	2.91998558	4.30265273	9.924843201	31.59905458
3	2.353363435	3.182446305	5.84090931	12.92397864
4	2.131846786	2.776445105	4.604094871	8.610301581
5	2.015048373	2.570581836	4.032142984	6.868826626
6	1.943180281	2.446911851	3.707428021	5.958816179
7	1.894578605	2.364624252	3.499483297	5.407882521
8	1.859548038	2.306004135	3.355387331	5.041305433
9	1.833112933	2.262157163	3.249835542	4.780912586
10	1.812461123	2.228138852	3.169272673	4.586893859
11	1.795884819	2.20098516	3.105806516	4.436979338
12	1.782287556	2.17881283	3.054539589	4.317791284
13	1.770933396	2.160368656	3.012275839	4.220831728
14	1.761310136	2.144786688	2.976842734	4.140454113
15	1.753050356	2.131449546	2.946712883	4.072765196
16	1.745883676	2.119905299	2.920781622	4.014996327
17	1.739606726	2.109815578	2.89823052	3.965126272
18	1.734063607	2.10092204	2.878440473	3.921645825
19	1.729132812	2.093024054	2.860934606	3.883405853
20	1.724718243	2.085963447	2.84533971	3.849516275
21	1.720742903	2.079613845	2.831359558	3.819277164
22	1.717144374	2.073873068	2.818756061	3.792130672
23	1.713871528	2.06865761	2.807335684	3.767626804
24	1.71088208	2.063898562	2.796939505	3.745398619
25	1.708140761	2.059538553	2.787435814	3.72514395
26	1.70561792	2.055529439	2.778714533	3.706611743
27	1.703288446	2.051830516	2.770682957	3.689591713
28	1.701130934	2.048407142	2.763262455	3.673906401
29	1.699127027	2.045229642	2.756385904	3.659405019
30	1.697260887	2.042272456	2.749995654	3.645958635

Critical values of t (1 tail test)

Confidence level	90%	95%	99%	99.90%
Probability	0.1	0.05	0.01	0.001
Degrees of freedom				
1	3.077683537	6.313751515	12.70620474	63.65674116
2	1.885618083	2.91998558	4.30265273	9.924843201
3	1.637744354	2.353363435	3.182446305	5.84090931
4	1.533206274	2.131846786	2.776445105	4.604094871
5	1.475884049	2.015048373	2.570581836	4.032142984
6	1.439755747	1.943180281	2.446911851	3.707428021
7	1.414923928	1.894578605	2.364624252	3.499483297
8	1.39681531	1.859548038	2.306004135	3.355387331
9	1.383028738	1.833112933	2.262157163	3.249835542
10	1.372183641	1.812461123	2.228138852	3.169272673
11	1.363430318	1.795884819	2.20098516	3.105806516
12	1.356217334	1.782287556	2.17881283	3.054539589
13	1.350171289	1.770933396	2.160368656	3.012275839
14	1.345030374	1.761310136	2.144786688	2.976842734
15	1.340605608	1.753050356	2.131449546	2.946712883
16	1.336757167	1.745883676	2.119905299	2.920781622
17	1.33337939	1.739606726	2.109815578	2.89823052
18	1.330390944	1.734063607	2.10092204	2.878440473
19	1.327728209	1.729132812	2.093024054	2.860934606
20	1.325340707	1.724718243	2.085963447	2.84533971
21	1.323187874	1.720742903	2.079613845	2.831359558
22	1.321236742	1.717144374	2.073873068	2.818756061
23	1.31946024	1.713871528	2.06865761	2.807335684
24	1.317835934	1.71088208	2.063898562	2.796939505
25	1.316345073	1.708140761	2.059538553	2.787435814
26	1.314971864	1.70561792	2.055529439	2.778714533
27	1.313702913	1.703288446	2.051830516	2.770682957
28	1.312526782	1.701130934	2.048407142	2.763262455
29	1.311433647	1.699127027	2.045229642	2.756385904
30	1.310415025	1.697260887	2.042272456	2.749995654

Critical values of F

Confidence =95%
alpha=0.05

Degrees of freedom Between samples	1	2	3	4	5	6	7	8	9	10	11	12	13
1	161.448	199.5	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882	242.983	243.906	244.69
2	18.513	19.491	19.491	19.491	19.491	19.491	19.491	19.492	19.492	19.492	19.492	19.492	19.492
3	10.128	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667
4	7.709	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041
5	6.608	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95
6	5.987	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534
7	5.591	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12
8	5.318	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838
9	5.117	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863
10	4.965	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708
11	4.844	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587
12	4.747	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49
13	4.667	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411
14	4.6	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344
15	4.543	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287
16	4.494	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239
17	4.451	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197
18	4.414	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16
19	4.381	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127
20	4.351	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098
25	4.242	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991
30	4.171	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316
40	4.085	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839
50	4.034	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183
60	4.001	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758
120	3.92	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072

14	15	16	17	18	19	20	25	30	40	50	60	120
245.364	245.95	246.464	246.918	247.323	247.686	248.013	249.26	250.095	251.143	251.774	252.196	253.253
19.492	19.492	19.492	19.492	19.492	19.492	19.492	19.492	19.492	19.492	19.492	19.492	19.492
8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667	8.667
6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041	6.041
4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95	4.95
4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534	4.534
4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12	4.12
3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838	3.838
3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863	3.863
3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708	3.708
3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587	3.587
3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49	3.49
3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411	3.411
3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344	3.344
3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287	3.287
3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239
3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197	3.197
3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16
3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127	3.127
3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098	3.098
2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991	2.991
3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316	3.316
2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839	2.839
3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183	3.183
2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758	2.758
3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072	3.072

Critical values of F

Confidence =90%
alpha=0.1

	Degrees of freedom within samples										
Degrees of freedom Between samples	1	2	3	4	5	6	7	8	9	10	11	12
1	39.863	49.5	53.593	55.833	57.24	58.204	58.906	59.439	59.858	60.195	60.473	60.705
2	8.526	9	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.401	9.408
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.24	5.23	5.222	5.216
4	4.545	4.325	4.191	4.107	4.051	4.01	3.979	3.955	3.936	3.92	3.907	3.896
5	4.06	3.78	3.619	3.52	3.453	3.405	3.368	3.339	3.316	3.297	3.282	3.268
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.92	2.905
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.684	2.668
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.519	2.502
9	3.36	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.44	2.416	2.396	2.379
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.302	2.284
11	3.225	2.86	2.66	2.536	2.451	2.389	2.342	2.304	2.274	2.248	2.227	2.209
12	3.177	2.807	2.606	2.48	2.394	2.331	2.283	2.245	2.214	2.188	2.166	2.147
13	3.136	2.763	2.56	2.434	2.347	2.283	2.234	2.195	2.164	2.138	2.116	2.097
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095	2.073	2.054
15	3.073	2.695	2.49	2.361	2.273	2.208	2.158	2.119	2.086	2.059	2.037	2.017
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028	2.005	1.985
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001	1.978	1.958
18	3.007	2.624	2.416	2.286	2.196	2.13	2.079	2.038	2.005	1.977	1.954	1.933
19	2.99	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956	1.932	1.912
20	2.975	2.589	2.38	2.249	2.158	2.091	2.04	1.999	1.965	1.937	1.913	1.892
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866	1.841	1.82
30	2.881	2.489	2.276	2.142	2.049	1.98	1.927	1.884	1.849	1.819	1.794	1.773
40	2.835	2.44	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763	1.737	1.715
50	2.809	2.412	2.197	2.061	1.966	1.895	1.84	1.796	1.76	1.729	1.703	1.68
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707	1.68	1.657
120	2.748	2.347	2.13	1.992	1.896	1.824	1.767	1.722	1.684	1.652	1.625	1.601

13	14	15	16	17	18	19	20	25	30	40	50	60	120
60.903	61.073	61.22	61.35	61.464	61.566	61.658	61.74	62.055	62.265	62.529	62.688	62.794	63.061
9.415	9.42	9.425	9.429	9.433	9.436	9.439	9.441	9.451	9.458	9.466	9.471	9.475	9.483
5.21	5.205	5.2	5.196	5.193	5.19	5.187	5.184	5.175	5.168	5.16	5.155	5.151	5.143
3.886	3.878	3.87	3.864	3.858	3.853	3.849	3.844	3.828	3.817	3.804	3.795	3.79	3.775
3.257	3.247	3.238	3.23	3.223	3.217	3.212	3.207	3.187	3.174	3.157	3.147	3.14	3.123
2.892	2.881	2.871	2.863	2.855	2.848	2.842	2.836	2.815	2.8	2.781	2.77	2.762	2.742
2.654	2.643	2.632	2.623	2.615	2.607	2.601	2.595	2.571	2.555	2.535	2.523	2.514	2.493
2.488	2.475	2.464	2.455	2.446	2.438	2.431	2.425	2.4	2.383	2.361	2.348	2.339	2.316
2.364	2.351	2.34	2.329	2.32	2.312	2.305	2.298	2.272	2.255	2.232	2.218	2.208	2.184
2.269	2.255	2.244	2.233	2.224	2.215	2.208	2.201	2.174	2.155	2.132	2.117	2.107	2.082
2.193	2.179	2.167	2.156	2.147	2.138	2.13	2.123	2.095	2.076	2.052	2.036	2.026	2
2.131	2.117	2.105	2.094	2.084	2.075	2.067	2.06	2.031	2.011	1.986	1.97	1.96	1.932
2.08	2.066	2.053	2.042	2.032	2.023	2.014	2.007	1.978	1.958	1.931	1.915	1.904	1.876
2.037	2.022	2.01	1.998	1.988	1.978	1.97	1.962	1.933	1.912	1.885	1.869	1.857	1.828
2	1.985	1.972	1.961	1.95	1.941	1.932	1.924	1.894	1.873	1.845	1.828	1.817	1.787
1.968	1.953	1.94	1.928	1.917	1.908	1.899	1.891	1.86	1.839	1.811	1.793	1.782	1.751
1.94	1.925	1.912	1.9	1.889	1.879	1.87	1.862	1.831	1.809	1.781	1.763	1.751	1.719
1.916	1.9	1.887	1.875	1.864	1.854	1.845	1.837	1.805	1.783	1.754	1.736	1.723	1.691
1.894	1.878	1.865	1.852	1.841	1.831	1.822	1.814	1.782	1.759	1.73	1.711	1.699	1.666
1.875	1.859	1.845	1.833	1.821	1.811	1.802	1.794	1.761	1.738	1.708	1.69	1.677	1.643
1.802	1.785	1.771	1.758	1.746	1.736	1.726	1.718	1.683	1.659	1.627	1.607	1.593	1.557
1.754	1.737	1.722	1.709	1.697	1.686	1.676	1.667	1.632	1.606	1.573	1.552	1.538	1.499
1.695	1.678	1.662	1.649	1.636	1.625	1.615	1.605	1.568	1.541	1.506	1.483	1.467	1.425
1.66	1.643	1.627	1.613	1.6	1.588	1.578	1.568	1.529	1.502	1.465	1.441	1.424	1.379
1.637	1.619	1.603	1.589	1.576	1.564	1.553	1.543	1.504	1.476	1.437	1.413	1.395	1.348
1.58	1.562	1.545	1.53	1.516	1.504	1.493	1.482	1.44	1.409	1.368	1.34	1.32	1.265

Critical values of F

Confidence =99%
alpha=0.01

	Degrees of freedom within samples											
Degrees of freedom Between samples	1	2	3	4	5	6	7	8	9	10	11	12
1	4052.181	4999.5	5403.352	5624.583	5763.65	5858.986	5928.356	5981.07	6022.473	6055.847	6083.317	6106.321
2	98.503	99	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.408	99.416
3	34.116	30.817	29.457	28.71	28.237	27.911	27.672	27.489	27.345	27.229	27.133	27.052
4	21.198	18	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.452	14.374
5	16.258	13.274	12.06	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.963	9.888
6	13.745	10.925	9.78	9.148	8.746	8.466	8.26	8.102	7.976	7.874	7.79	7.718
7	12.246	9.547	8.451	7.847	7.46	7.191	6.993	6.84	6.719	6.62	6.538	6.469
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.734	5.667
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.178	5.111
10	10.044	7.559	6.552	5.994	5.636	5.386	5.2	5.057	4.942	4.849	4.772	4.706
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.462	4.397
12	9.33	6.927	5.953	5.412	5.064	4.821	4.64	4.499	4.388	4.296	4.22	4.155
13	9.074	6.701	5.739	5.205	4.862	4.62	4.441	4.302	4.191	4.1	4.025	3.96
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.14	4.03	3.939	3.864	3.8
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.73	3.666
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.89	3.78	3.691	3.616	3.553
17	8.4	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.519	3.455
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.434	3.371
19	8.185	5.926	5.01	4.5	4.171	3.939	3.765	3.631	3.523	3.434	3.36	3.297
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.294	3.231
25	7.77	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	3.056	2.993
30	7.562	5.39	4.51	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.906	2.843
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.727	2.665
50	7.171	5.057	4.199	3.72	3.408	3.186	3.02	2.89	2.785	2.698	2.625	2.562
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.559	2.496
120	6.851	4.787	3.949	3.48	3.174	2.956	2.792	2.663	2.559	2.472	2.399	2.336

13	14	15	16	17	18	19	20	25	30	40	50	60	120
6125.865	6142.674	6157.285	6170.101	6181.435	6191.529	6200.576	6208.73	6239.825	6260.649	6286.782	6302.517	6313.03	6339.391
99.422	99.428	99.433	99.437	99.44	99.444	99.447	99.449	99.459	99.466	99.474	99.479	99.482	99.491
26.983	26.924	26.872	26.827	26.787	26.751	26.719	26.69	26.579	26.505	26.411	26.354	26.316	26.221
14.307	14.249	14.198	14.154	14.115	14.08	14.048	14.02	13.911	13.838	13.745	13.69	13.652	13.558
9.825	9.77	9.722	9.68	9.643	9.61	9.58	9.553	9.449	9.379	9.291	9.238	9.202	9.112
7.657	7.605	7.559	7.519	7.483	7.451	7.422	7.396	7.296	7.229	7.143	7.091	7.057	6.969
6.41	6.359	6.314	6.275	6.24	6.209	6.181	6.155	6.058	5.992	5.908	5.858	5.824	5.737
5.609	5.559	5.515	5.477	5.442	5.412	5.384	5.359	5.263	5.198	5.116	5.065	5.032	4.946
5.055	5.005	4.962	4.924	4.89	4.86	4.833	4.808	4.713	4.649	4.567	4.517	4.483	4.398
4.65	4.601	4.558	4.52	4.487	4.457	4.43	4.405	4.311	4.247	4.165	4.115	4.082	3.996
4.342	4.293	4.251	4.213	4.18	4.15	4.123	4.099	4.005	3.941	3.86	3.81	3.776	3.69
4.1	4.052	4.01	3.972	3.939	3.909	3.883	3.858	3.765	3.701	3.619	3.569	3.535	3.449
3.905	3.857	3.815	3.778	3.745	3.716	3.689	3.665	3.571	3.507	3.425	3.375	3.341	3.255
3.745	3.698	3.656	3.619	3.586	3.556	3.529	3.505	3.412	3.348	3.266	3.215	3.181	3.094
3.612	3.564	3.522	3.485	3.452	3.423	3.396	3.372	3.278	3.214	3.132	3.081	3.047	2.959
3.498	3.451	3.409	3.372	3.339	3.31	3.283	3.259	3.165	3.101	3.018	2.967	2.933	2.845
3.401	3.353	3.312	3.275	3.242	3.212	3.186	3.162	3.068	3.003	2.92	2.869	2.835	2.746
3.316	3.269	3.227	3.19	3.158	3.128	3.101	3.077	2.983	2.919	2.835	2.784	2.749	2.66
3.242	3.195	3.153	3.116	3.084	3.054	3.027	3.003	2.909	2.844	2.761	2.709	2.674	2.584
3.177	3.13	3.088	3.051	3.018	2.989	2.962	2.938	2.843	2.778	2.695	2.643	2.608	2.517
2.939	2.892	2.85	2.813	2.78	2.751	2.724	2.699	2.604	2.538	2.453	2.4	2.364	2.27
2.789	2.742	2.7	2.663	2.63	2.6	2.573	2.549	2.453	2.386	2.299	2.245	2.208	2.111
2.611	2.563	2.522	2.484	2.451	2.421	2.394	2.369	2.271	2.203	2.114	2.058	2.019	1.917
2.508	2.461	2.419	2.382	2.348	2.318	2.29	2.265	2.167	2.098	2.007	1.949	1.909	1.803
2.442	2.394	2.352	2.315	2.281	2.251	2.223	2.198	2.098	2.028	1.936	1.877	1.836	1.726
2.282	2.234	2.192	2.154	2.119	2.089	2.06	2.035	1.932	1.86	1.763	1.7	1.656	1.533

Critical values of Q. Alpha=0.5

Degrees of freedom	Number of means								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	18	27	32.8	37.2	40.5	43.1	45.4	47.3	49.1
2	6.09	8.33	9.8	10.89	11.73	12.43	13.03	13.54	13.99
3	4.5	5.91	6.83	7.51	8.04	8.47	8.85	9.18	9.46
4	3.93	5.04	5.76	6.29	6.71	7.06	7.35	7.6	7.83
5	3.64	4.6	5.22	5.67	6.03	6.33	6.58	6.8	6.99
6	3.46	4.34	4.9	5.31	5.63	5.89	6.12	6.32	6.49
7	3.34	4.16	4.68	5.06	5.35	5.59	5.8	5.99	6.15
8	3.26	4.04	4.53	4.89	5.17	5.4	5.6	5.77	5.92
9	3.2	3.95	4.42	4.76	5.02	5.24	5.43	5.6	5.74
10	3.15	3.88	4.33	4.66	4.91	5.12	5.3	5.46	5.6
11	3.11	3.82	4.26	5.58	4.82	5.03	5.2	5.35	5.49
12	3.08	3.77	4.2	4.51	4.75	4.95	5.12	5.27	5.4
13	3.06	3.73	4.15	4.46	4.69	4.88	5.05	5.19	5.32
14	3.03	3.7	4.11	4.41	4.64	4.83	4.99	5.13	5.25
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.2
16	3	3.65	4.05	4.34	4.56	4.74	4.9	5.03	5.15
17	2.98	3.62	4.02	4.31	4.52	4.7	4.86	4.99	5.11
18	2.97	3.61	4	4.28	4.49	4.67	4.83	4.96	5.07
19	2.96	3.59	3.98	4.26	4.47	4.64	4.79	4.92	5.04
20	2.95	3.58	3.96	4.24	4.45	4.62	4.77	4.9	5.01
30	2.89	3.48	3.84	4.11	4.3	4.46	4.6	4.72	4.83
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.74
120	2.8	3.6	3.69	3.92	4.1	4.24	4.36	4.47	4.56
∞	2.77	3.32	3.63	3.86	4.03	4.17	4.29	4.39	4.47

Degrees of freedom	Number of means									
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	50.6	51.9	53.2	54.3	55.4	56.3	57.2	58	58.8	59.6
2	14.39	14.75	15.08	15.38	15.65	15.91	16.14	16.36	16.57	16.77
3	9.72	9.95	10.16	10.35	10.52	10.69	10.84	10.98	11.12	12.24
4	8.03	8.21	8.37	8.52	8.67	8.8	8.92	9.03	9.14	9.24
5	7.17	7.32	7.47	7.6	7.72	7.83	7.93	8.03	8.12	8.21
6	6.65	6.79	6.92	7.04	7.14	7.24	7.34	7.43	7.51	7.59
7	6.29	6.42	6.54	6.65	6.75	6.84	6.93	7.01	7.08	7.16
8	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.8	6.87
9	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.65
10	5.72	5.83	5.93	6.03	6.12	6.2	6.27	6.34	6.41	6.47
11	5.61	5.71	5.81	5.9	5.98	6.06	6.14	6.2	6.27	6.33
12	5.51	5.61	5.71	5.8	5.88	5.95	6.02	6.09	6.15	6.21
13	5.43	5.53	5.63	5.71	5.79	5.86	5.93	6	6.06	6.11
14	5.36	5.46	5.56	5.64	5.72	5.79	5.86	5.92	5.98	6.03
15	5.31	5.4	5.49	5.57	5.65	5.72	5.79	5.85	5.91	5.96
16	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.9
17	5.21	5.31	5.39	5.47	5.55	5.61	5.68	5.74	5.79	5.84
18	5.17	5.27	5.35	5.43	5.5	5.57	5.63	5.69	5.74	5.79
19	5.14	5.23	5.32	5.39	5.46	5.53	5.59	5.65	5.7	5.75
20	5.11	5.2	5.28	5.36	5.43	5.5	5.56	5.61	5.66	5.71
30	4.92	5	5.08	5.15	5.21	5.27	5.33	5.38	3.43	5.48
40	4.82	4.9	4.98	5.05	5.11	5.17	5.22	5.27	5.32	5.36
120	4.64	4.71	4.78	4.84	4.9	4.95	5	5.04	5.09	5.13
∞	4.55	4.62	4.68	4.74	4.8	4.84	4.89	4.93	4.97	5.01

0.01 critical values

k = sample size for the range = number of groups

Degrees of Freedom	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1	90.03	135	164.3	185.6	202.2	215.8	227.2	237	245.6
2	14.04	19.02	22.29	24.72	26.63	28.2	29.53	30.68	31.69
3	8.26	10.62	12.17	13.33	14.24	15	15.64	16.2	16.69
4	6.51	8.12	9.17	9.96	10.58	11.1	11.55	11.93	12.27
5	5.7	6.98	7.8	8.42	8.91	9.32	9.67	9.97	10.24
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.1
7	4.95	5.92	6.54	7.01	7.37	7.68	7.94	8.17	8.37
8	4.75	5.64	6.2	6.62	6.96	7.24	7.47	7.68	7.86
9	4.6	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21
11	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99
12	4.32	5.05	5.5	5.84	6.1	6.32	6.51	6.67	6.81
13	4.26	4.96	5.4	5.73	5.98	6.19	6.37	6.53	6.67
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54
15	4.17	4.84	5.25	5.56	5.8	5.99	6.16	6.31	6.44
16	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35
17	4.1	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27
18	4.07	4.7	5.09	5.38	5.6	5.79	5.94	6.08	6.2
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09
24	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92
30	3.89	4.45	4.8	5.05	5.24	5.4	5.54	5.65	5.76
40	3.82	4.37	4.7	4.93	5.11	5.26	5.39	5.5	5.6
60	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45
120	3.7	4.2	4.5	4.71	4.87	5.01	5.12	5.21	5.3
∞	3.64	4.12	4.4	4.6	4.76	4.88	4.99	5.08	5.16

Random numbers: 00-99

16	44	11	61	90	16	90	16	88	19
31	34	7	77	37	18	72	10	15	97
10	25	76	31	17	47	96	42	88	55
22	32	14	49	92	83	82	98	22	10
7	87	24	72	91	67	89	55	22	45
56	3	17	47	32	97	90	31	20	84
54	55	62	38	62	8	39	69	93	95
87	48	41	59	0	27	59	71	66	0
6	29	90	24	52	43	35	87	11	17
73	13	70	61	77	92	33	56	11	48
18	47	71	90	31	81	95	1	53	59
32	89	22	77	26	24	68	67	12	6
4	16	50	77	89	98	99	70	58	35
84	24	70	24	22	61	81	73	52	28
39	65	31	88	50	79	10	98	56	20
70	75	98	76	53	55	20	87	76	7
38	54	11	78	32	6	77	21	94	85
76	10	48	27	45	81	33	36	45	79
73	57	10	31	34	71	18	55	68	67
39	87	52	61	72	71	87	46	60	83
4	61	44	37	74	66	56	65	85	79
8	74	40	85	51	15	46	55	67	65
9	78	48	3	27	35	31	49	71	45
8	24	67	3	90	85	21	61	32	89
20	36	39	58	45	3	2	84	71	82
41	93	58	93	7	12	84	74	18	1
15	33	43	11	17	4	56	24	59	89
98	21	78	73	3	7	20	9	91	86
28	27	48	48	78	45	88	6	13	13
7	64	2	56	53	76	47	92	41	35

Random numbers 1-6

3	2	1	5	2	1	5	1	4
3	4	6	6	4	2	3	6	5
1	4	2	5	4	2	5	6	1
1	1	2	2	6	2	6	3	4
5	2	5	6	2	3	2	1	5
6	2	2	6	4	3	1	6	6
3	4	4	6	2	5	2	5	2
2	3	3	4	1	1	5	1	1
3	3	1	5	1	5	5	3	3
2	1	4	4	6	6	5	5	4
3	5	1	1	1	3	5	1	2
5	4	5	1	5	6	3	5	1
6	3	1	5	1	2	6	1	6
1	2	3	4	3	6	6	6	1
4	1	3	5	4	2	2	4	2
1	1	1	1	6	1	3	3	3
5	2	4	5	4	3	5	2	1
6	5	3	5	4	4	1	6	5
4	4	4	1	3	2	1	3	3
6	4	5	2	6	2	1	1	3
4	6	4	2	2	6	2	1	3
5	1	3	3	4	6	5	4	2
5	2	2	3	6	1	3	3	3
2	6	3	5	1	2	5	3	6
3	6	5	1	2	5	6	6	5
4	2	6	6	5	1	6	3	4
3	4	1	6	5	6	5	6	3
4	4	4	2	2	1	6	6	5
6	3	3	1	2	5	4	6	4
4	5	5	2	1	1	1	3	3

Critical values of R 2-tail

95%		90%		99%	
DF	R	DF	R	DF	R
3	0.878339	3	0.805384	3	0.958735
4	0.811401	4	0.729299	4	0.9172
5	0.754492	5	0.669439	5	0.874526
6	0.706734	6	0.621489	6	0.834342
7	0.666384	7	0.582206	7	0.797681
8	0.631897	8	0.549357	8	0.764592
9	0.602069	9	0.521404	9	0.734786
10	0.575983	10	0.497265	10	0.707888
11	0.552943	11	0.476156	11	0.683528
12	0.532413	12	0.4575	12	0.661376
13	0.513977	13	0.440861	13	0.641145
14	0.497309	14	0.425902	14	0.622591
15	0.482146	15	0.41236	15	0.605506
16	0.468277	16	0.400027	16	0.589714
17	0.455531	17	0.388733	17	0.575067
18	0.443763	18	0.378341	18	0.561435
19	0.432858	19	0.368737	19	0.548711
20	0.422714	20	0.359827	20	0.5368
22	0.404386	22	0.343783	22	0.515101
24	0.388244	24	0.329705	24	0.495808
26	0.373886	26	0.317223	26	0.478511
28	0.361007	28	0.306057	28	0.462892
30	0.34937	30	0.295991	30	0.448699
35	0.324573	35	0.274611	35	0.418211
40	0.304396	40	0.257278	40	0.393174
50	0.273243	50	0.23062	50	0.354153
55	0.260869	55	0.220062	55	0.338538
60	0.250035	60	0.210832	60	0.324818
70	0.231883	70	0.195394	70	0.301734
80	0.217185	80	0.182916	80	0.282958
90	0.204968	90	0.172558	90	0.267298
100	0.194604	100	0.163782	100	0.253979
150	0.159273	150	0.133919	150	0.208349
200	0.138098	200	0.11606	200	0.18086
400	0.097824	400	0.082155	400	0.128339
1000	0.061935	1000	0.051993	1000	0.08134

Critical values of R: 1 tail

1 tail					
95%		90%		99%	
DF	R	DF	R	DF	R
3	0.805384	3	0.687049	3	0.985926
4	0.729299	4	0.6084	4	0.963259
5	0.669439	5	0.550863	5	0.934964
6	0.621489	6	0.506727	6	0.904896
7	0.582206	7	0.471589	7	0.875145
8	0.549357	8	0.442796	8	0.846691
9	0.521404	9	0.418662	9	0.819927
10	0.497265	10	0.398062	10	0.794953
11	0.476156	11	0.380216	11	0.771726
12	0.4575	12	0.364562	12	0.750143
13	0.440861	13	0.350688	13	0.730074
14	0.425902	14	0.338282	14	0.711389
15	0.41236	15	0.327101	15	0.693959
16	0.400027	16	0.316958	16	0.677669
17	0.388733	17	0.307702	17	0.662411
18	0.378341	18	0.29921	18	0.64809
19	0.368737	19	0.291384	19	0.63462
20	0.359827	20	0.28414	20	0.621926
22	0.343783	22	0.271137	22	0.598598
24	0.329705	24	0.259768	24	0.577647
26	0.317223	26	0.249717	26	0.558708
28	0.306057	28	0.240749	28	0.541485
30	0.295991	30	0.232681	30	0.525739
35	0.274611	35	0.215598	35	0.49163
40	0.257278	40	0.201796	40	0.463349
50	0.23062	50	0.180644	50	0.418829
55	0.220062	55	0.17229	55	0.400877
60	0.210832	60	0.164997	60	0.385044
70	0.195394	70	0.152818	70	0.358285
80	0.182916	80	0.14299	80	0.336418
90	0.172558	90	0.134844	90	0.318115
100	0.163782	100	0.127947	100	0.302504
150	0.133919	150	0.104525	150	0.248752
200	0.11606	200	0.090546	200	0.216192
400	0.082155	400	0.064052	400	0.153689
1000	0.051993	1000	0.04052	1000	0.097513

Critical values of chi

Confidence:	0.9	0.95	0.99		0.9	0.95	0.99
Degrees of freedom				Degrees of freedom			
1	2.705543	3.841459	11.34487	26	35.56317	38.88514	61.16209
2	4.60517	5.991465	15.08627	27	36.74122	40.11327	63.69074
3	6.251389	7.814728	18.47531	28	37.91592	41.33714	64.95007
4	7.77944	9.487729	21.66599	29	39.08747	42.55697	66.20624
5	9.236357	11.0705	24.72497	30	40.25602	43.77297	67.45935
6	10.64464	12.59159	26.21697	31	41.42174	44.98534	68.70951
7	12.01704	14.06714	29.14124	32	42.58475	46.19426	71.2014
8	13.36157	15.50731	30.57791	33	43.74518	47.39988	72.44331
9	14.68366	16.91898	31.99993	34	44.90316	48.60237	73.68264
10	15.98718	18.30704	34.80531	35	46.05879	49.80185	74.91947
11	17.27501	19.67514	36.19087	36	47.21217	50.99846	76.15389
12	18.54935	21.02607	38.93217	37	48.36341	52.19232	78.61576
13	19.81193	22.36203	40.28936	38	49.51258	53.38354	79.84334
14	21.06414	23.68479	41.6384	39	50.65977	54.57223	81.06877
15	22.30713	24.99579	42.97982	40	51.80506	55.75848	82.29212
16	23.54183	26.29623	45.64168	41	52.94851	56.94239	83.51343
17	24.76904	27.58711	46.96294	42	54.0902	58.12404	85.95018
18	25.98942	28.8693	48.27824	43	55.23019	59.30351	87.16571
19	27.20357	30.14353	50.89218	44	56.36854	60.48089	88.37942
20	28.41198	31.41043	52.19139	45	57.5053	61.65623	89.59134
21	29.61509	32.67057	53.48577	46	58.64054	62.82962	90.80153
22	30.81328	33.92444	54.77554	47	59.77429	64.00111	93.21686
23	32.0069	35.17246	57.34207	48	60.90661	65.17077	94.42208
24	33.19624	36.41503	58.61921	49	62.03754	66.33865	95.62572
25	34.38159	37.65248	59.8925	50	63.16712	67.50481	96.82782

Critical values of U

N1→		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
***n*2↓**	**1**										
	2								0	0	0
	3					0	1	1	2	2	3
	4				0	1	2	3	4	4	5
	5			0	1	2	3	5	6	7	8
	6			1	2	3	5	6	8	10	11
	7			1	3	5	6	8	10	12	14
	8		0	2	4	6	8	10	13	15	17
	9		0	2	4	7	10	12	15	17	20
	10		0	3	5	8	11	14	17	20	23
	11		0	3	6	9	13	16	19	23	26
	12		1	4	7	11	14	18	22	26	29
	13		1	4	8	12	16	20	24	28	33
	14		1	5	9	13	17	22	26	31	36
	15		1	5	10	14	19	24	29	34	39
	16		1	6	11	15	21	26	31	37	42
	17		2	6	11	17	22	28	34	39	45
	18		2	7	12	18	24	30	36	42	48
	19		2	7	13	19	25	32	38	45	52
	20		2	8	13	20	27	34	41	48	55

Critical values of U (continued).

N1→		**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
***n*2↓**	**1**										
	2	0	1	1	1	1	1	2	2	2	2
	3	3	4	4	5	5	6	6	7	7	8
	4	6	7	8	9	10	11	11	12	13	13
	5	9	11	12	13	14	15	17	18	19	20
	6	13	14	16	17	19	21	22	24	25	27
	7	16	18	20	22	24	26	28	30	32	34
	8	19	22	24	26	29	31	34	36	38	41
	9	23	26	28	31	34	37	39	42	45	48
	10	26	29	33	36	39	42	45	48	52	55
	11	30	33	37	40	44	47	51	55	58	62
	12	33	37	41	45	49	53	57	61	65	69
	13	37	41	45	50	54	59	63	67	72	76
	14	40	45	50	55	59	64	67	74	78	83
	15	44	49	54	59	64	70	75	80	85	90
	16	47	53	59	64	70	75	81	86	92	98
	17	51	57	63	67	75	81	87	93	99	105
	18	55	61	67	74	80	86	93	99	106	112
	19	58	65	72	78	85	92	99	106	113	119
	20	62	69	76	83	90	98	105	112	119	127

Just 15% of students in England study mathematics beyond GCSE level. However, many of this non-mathematics studying majority find that they need for mathematical skills for the advanced study of other subjects, including humanities and social science subjects at school or university or in their job. AWithout mathematical, and in particular statistical skills whole areas of the social sciences and humanities are inaccessible to research students and future academics. This book aims to fill this gap.

Cover Illustration: Altar of Domitius Ahenobarbus, Louvre

Photo: Marie-Lan Nguyen (2007) Public domain

http://commons.wikimedia.org/wiki/File%3AAltar_Domitius_Ahenobarbus_Louvre_n2.jpg

www.statisticsforhumanities.net

@statistics4hums