STATISTICS BEYOND THE ABSOLUTE BASICS

Ivan Lowe 2019

www.scientificlanguage.com

Version 3.0

Reference will be made to two previous books published in 2016 by the author.

"First Textbook" *A first textbook of research methodology and thesis writeup for second language English speakers*.

"Feel for Statistics"
A feel for Statistics: Essential concepts underlying the calculations

Both are available from this site, and are assumed in this book.

Change log

2019

This is a complete re-organisation of the textbook. I no longer ride two horses. Priority is given to the New Statistics, and the old is only explained later, when needed. I have not even bothered to move old material to an appendix. Material has been moved around and there are different chapters, especially at the end.

Excluded from this book is:
- the t-test for groups
- Chi Square
- Choosing a test.

I have also improved the presentation of the appendix on validity and reliability. There are general improvements elsewhere.

2018. Added, Loess lines.

2017 edition, there have been many cosmetic improvements. I also decided that giving screen shots from programs was largely useless: either I did it more extensively, or not at all.

Chapter 3 Types of data.
I have dealt with the disagreement among statisticians as to whether you can fairly use the Mean for a Likert Scale. As always, I have cut through the discussion by reminding readers that sometimes questionnaire data shows two peaks, and that visual inspection really must precede and dominate the calculations.

Chapter 8 Correlations
The discussion has been greatly improved, especially with visuals. The Margin of Error gets a full explanation.

Chapter 14. Effect Size
This has been tidied up, and finishes with an introduction to z-scores.

Why statistics matter

Consider the following sobering tale by Geoff Cumming:

> In the late 1970s, my wife and I followed the best advice on how to reduce the risk of SIDS, or cot death, by putting our young kids to sleep face down on a sheepskin. A recent review applied meta-analysis to the evidence available at various times, and found that, by 1970, there was reasonably clear evidence that back sleeping is safer.
>
> The evidence strengthened over the years, although some parenting books still recommended front sleeping as late as 1988. The authors of the meta-analysis estimated that, **if an analysis such as theirs had been available and used in 1970** – and the recommendation for back sleeping had been widely adopted – **as many as 50,000 infant deaths may have been avoided across the Western world.**

Who says the choice of statistical technique doesn't make a difference?

Geoff Cumming. Mind your confidence interval: how statistics skew research results. 18 April 2012

https://theconversation.com/mind-your-confidence-interval-how-statistics-skew-research-results-3186

Note: you do not need to know what a 'meta-analysis' is to get the point he is making.

CONTENTS

CHAPTER 1
INTRODUCTION

1. Introduction

This section is designed to take students quickly from the essentials in the previous keys through to the major statistical tests. Reference will NOT be made to SPSS since this program is very complicated. Easier free alternatives will be used.

I often find that students are in a hurry to use one of the big statistical tests. They research them, ask questions of mathematicians, and expect that the examiner will be impressed! Here are some basic rules.

2. The foundations MUST be mastered and used first

The first foundation is that the researcher must understand the subject such as linguistics behind the numbers. The researcher must be in control. The researcher needs to be able to explain and interpret their statistics. Statisticians can be extremely helpful, but, unless they intimately understand the topic, they will focus on the mathematics and can often end up being of little use and sometimes downright unhelpful. That is why in some universities the department statistician is also a subject specialist in their own right.

The researcher must understand the subject behind the numbers. The researcher must be in control. The researcher needs to be able to explain and interpret their statistics.

3. **Manage your statistician**
   It is the responsibility of the researcher to 'manage' any advice from a statistician. Tips for this will be provided in boxes!

4. **Give priority to understanding**
   High quality understanding of the keys to statistics must precede and must dominate any work in a thesis. See my book, "Feel for Statistics".

5. **Description of this book**
   In this book, there will be some more detail about basic concepts, and an introduction to more concepts assumed in the tests, before going on to present some of the easier tests. Significance testing is examined and found to be wanting. Alternatives are then presented and encouraged.

   Most of the 'elementary' or 'basic' books are in my opinion too difficult: it is like asking a toddler to scale a two metre high wall before getting onto the first step of the staircase. This book bridges the gap and fills in many of the details which are frequently missed out.

   The book can serve as a 10 hour introduction, or as a refresher course for established researchers in the Arts who need to come to terms with new developments.

Most of the 'elementary' or 'basic' books are in my opinion too difficult: it is like asking a toddler to scale a two metre high wall before getting onto the first step of the staircase. This book bridges the gap and fills in many of the details which are frequently missed out.

6. **Importance of understanding the basics**
   I cannot stress too much the importance of knowing what is going on in the real world you are describing. Statistics are a useful tool, but they must not be allowed to control you.

   Many years ago there was a secondary school teacher working in a school for delinquent boys. Many of the boys had been

sent there by the courts. Most had come from very rough backgrounds, and had missed a lot of school. The school was only for boys evaluated to be of high intelligence.

This teacher was given the task of teaching them mathematics. The boys were 14-15 years old and had no intention of studying mathematics. Some of them had picked up some maths and could run rings round this teacher, who, in fact, was an instructor in carpentry and building. These boys found algebra easy. But, they had one gap: they had never learned to add up and subtract, in their heads, at speed. Somehow, they had missed so much school in their younger years, they had never learned the basic skills of mental arithmetic, of how to multiply and divide using a pen and paper (this was the days before electronic calculators. Students had to use a slide rule, and as is well known, slide rules could never be used for addition and subtraction).

In short, these intelligent delinquents had a major gap. They needed foundations in arithmetic before going on to algebra and calculus.

The teacher solved the problem by playing the game of darts. This game requires you to do mental arithmetic in order to know where to aim the next dart in order to win. If you scored too high, you lost. The teacher proved he could handle mental arithmetic faster and more accurately than the boys. He also drilled into them, with plenty of practice, some basic skills.

Many times students have come to me and asked me about statistical tests. They do not know about the material in this book. Sometimes they have come asking me about ANOVA and other complicated procedures.

More commonly, students come vaguely saying they are going to collect some data and analyse it with SPSS (a favourite package of statistics programs).

It seems that there is an unjustified mystique and aura about SPSS and ANOVA. I am never impressed. These same students do not know a standard deviation from an inter-quartile range. They have no idea what a normal curve looks like, or does not look like. They could not explain a correlation coefficient if they tried, and could certainly never understand how to compare data from two groups (a pre-requisite for understanding ANOVA).

I repeat. Students need to commit themselves to understanding the phenomena being studied. They need to agree with the great importance of accurate description, They must be thoroughly skeptical about statistical tests.

I firmly believe that in many cases, even published authors have little basic understanding of their data, and have hidden behind the supposed glory of complex statistics, and in doing so they have dazzled a few, and missed the point.

This book is an attempt to help students get the point.

7. Why have you abandoned traditional statistics?

I was taught classically. Over fifty years I have struggled with statistics, and boiled it down. While respecting the expertise of experts, I find that often an expert is needlessly complicated. That theme occurs over and over again in these pages. I was therefore very excited to read about the new statistics, because it is inherently much easier to understand, it is more suited to the non-mathematician, and it is a real help in the struggle to understand exactly what is going on.

I have emphasised hands-on descriptive statistics. Correlations have stood the test of time. What is different is that comparisons between groups no longer use the t-test. Instead, it has been replaced by a figure related to the standard deviation. This book presents Cohen's *d*.

The other major difference is the recognition of the problems with the Null Hypothesis routine. Now, framing a hypothesis

in terms of no change/difference, versus significant change/difference, is sometimes a helpful approach. I have dealt with it in my book on methodology. Crucially, the researcher has to set out in advance what they decide is a significant difference.

In the old statistics, the *p*-value linked to some test would be used to decide on (statistical) significance. This always was misleading. Real world significance is NOT the same as statistical significance. In the new statistics, there is a great emphasis that the researcher must decide for themselves, and openly defend, what they will accept as a significant difference. And no test from statistics will help you.

The *p*-value does have a role. Its role is in assessing the quality of the data.

The other major take-home message is that correlations, and differences between groups, must always be assessed taking into account the MoE, the Margin of Error. This leads to error bars, which are very visual and very helpful and easy to understand.

All of this, and more, is presented in this book.

Dedication

This book is dedicated to the long-suffering students who have suffered from earlier material. I am the kind of teacher who learns most when brainstorming with students struggling to understand. Many times their perceptive questions have left me saying "I do not know" and I have gone away and puzzled and tried to understand more. Many times in lessons, I have had to ask the students to wait while I write down a good idea that has come while struggling to explain or to answer a question.

Students deserve a teacher who understands better than I do. Their patience, and their joy in studying, has been a great encouragement.

Statistics is an ideas course. I have always taken the view that the foundations are more important than the more advanced material, therefore I have rarely rushed the foundations. This makes for a relaxed course compared to courses with high-content. Statistics also gives, repeatedly, those gorgeous thrilling moments when the light suddenly dawns, the penny drops, and suddenly students (one or more) sit back and say "Aha, now I get it". Those, often unpredictable, moments, make it worthwhile being a teacher.

CHAPTER 2
WEB SOURCES AND PROGRAMS

1. Preliminaries

The web is awash with many good quality sites that explain statistics, and provide various free services.

Anyone using a statistics website to do the calculations needs to cope with the question of 'copy and paste'. You will need a way to handle the data you copy, and the results that you get. Therefore, before you start, try out some of the 'Clipboard extender' programs described in Chapter 19 of "A first textbook..."

The other major preparation is to master Excel! Excel has within it many statistical functions, and very sophisticated graphing and presentation tools. for many purposes. Excel may well have all that you need and there are many free suites of extras out there which will provide more functions. Try http://chandoo.org/wp/ or http://xltoolbox.sourceforge.net/. This is regularly updated, and will now for instance handle error bars and confidence intervals.

2. Links

http://www.statsoft.com/textbook/basic-statistics/?button=1 provides a basic textbook.

For students or those who want to learn about statistics, the best places to start is with an on-line statistics books. One is HyperStatistics Online, at http://davidmlane.com/hyperstat/ This is a nice statistics book, and it is a comprehensive list of other on line statistics books. Most of these are basic to intermediate. Statsoft www.statsoft.com/textbook/ has the basics as well as fairly advanced topics. Another approach is Robert Niles' site *Statistics Every Writer Should Know* http://www.robertniles.com/stats/ with (supposedly) plain English explanations for many basic statistical concepts. Another list of online statistics books is here http://gsociology.icaap.org/methods/stat.htm

Alex Reinhart has an evolving online book called "Statistics done wrong" www.refsmmat.com/statistics/index.html which could be understood, with some work, by anyone who has mastered the material in this book. In particular, his material is available under a copy with acknowledgement principle. Reinhart (2014) refers to this book.

A site for free pdf books www.bookboon.com also has a large number of free statistics books, including some guides to SPSS. Once again though, most of the 'elementary' or 'basic' books are in my opinion too difficult: it is like asking a toddler to scale a two metre high wall before getting onto the first step of the staircase.

3. My recommendations for programs

I have several recommendations, starting with the easiest.

- **SOFA** (Statistics Open For All) – an innovative statistics, analysis, and reporting program. Available for Windows, Mac and Linux systems. Has an emphasis on ease of use, learn as you go, and beautiful output.

☺ My favourite ☺

- **Past3** Do not be put off with its origins in palaeontology, the program is powerful and fairly easy to use and comes with a detailed pdf manual. It will also handle many data file formats including xls but NOT xlsx.

- **OpenStat** Supposedly it is easy friendly, designed for educational use. The interface is clear and simple, but it needs some experimentation or knowledge. For instance, you have to know what all the labels are, and when you set up the variables you have to specify what data type it is (floating point, integer, string, date, or money). If you want to import data you first have to save it in a Tab, Space, or Comma format. Frankly, I do not call that 'user friendly' because most of the time such distinctions do not matter to you, though they matter for programming. The number of tests it can do is greater than SOFA.

- **MicroOsiris** is one of the most comprehensive and includes a guide to selection of suitable techniques. This is a free program for someone needing more than the basics.

- **SAS University Edition**
  A free, powerful, well documented suite of programs with an easy to use graphical interface. Apparently you can also use it online, if you do not want to install it.
  www.sas.com/en_us/software/university-edition.html

4. Other links

- For other excellent free statistics programs, both free and online you are encouraged to visit John Pezzzullo's excellent site at: http://statpages.org,

- the very helpful summary
  http://gsociology.icaap.org/methods/statontheweb.html

- If you are looking for a free alternative to SPSS try this site: http://alternativeto.net/software/spss/

- Cohen Manion & Morrison (2011) provide a regularly updated list of links: http://cw.routledge.com/textbooks/cohen7e/links28.asp

- The librarians Index is no longer maintained. Google scholar is worth looking at, along with alternatives to google scholar. Try finding them!

- www.quora.com/What-are-some-good-alternatives-to-Google-Scholar has some alternatives, such as http://academicindex.net/

CHAPTER 3
TYPES OF DATA

A. Introduction

In this chapter and the next ones some of the basic concepts in statistics will be explained. Please work through them carefully and make sure you understand them.

Data types are something that most people have never thought much about. The experimental sciences distinguish between different types of data. The researcher needs to know the types because data presentation analysis and interpretation, and statistical tests, depend on these types.

The information in the statistics books is often confusing. This is because they make more than one distinction, and the distinctions overlap.

B. The distinction between discrete and continuous data

There is a major difference between 'discrete' data and 'continuous' data.

1. Discrete data

Discrete means distinct. Apples, oranges, bananas, and pears, are distinct types of fruit. You cannot get anything half way between an apple and a banana. Also, you do not measure the kinds of fruit, instead, you count them.

2. Continuous data

When you have one kind of fruit, such as oranges, the weight can vary considerably, between almost nothing, and 250 grammes. You could take 50 oranges and measure the weight. Then you would need to group the data, for instance, by 50 grammes. In this way you would have four groups, and you could count how many were in each group.

Figure 3.1 Continuous data

Weight in grammes	Number of oranges
1-50	
51-100	
101-150	
151-200	
201-250	

With continuous data there is a sliding scale – there are no jumps. Any boxing, any classification, is your choice.

3. Implications for graphs

When presenting graphs this distinction is important. See Key 18 of Feel for Statistics. The bars touch only when the data is continuous. You may feel this convention is arbitrary, and only a matter of style. Your feelings are wrong. This is NOT a style question where the choice is a matter of personal preference. The convention exists to send a signal to the reader about the type of data.

Scientists get it right instinctively. Scientists pick up the visual signal instinctively. Failure here is a sign that you are only a mere journalist, or a mere business person. It is easy to get it right.

C. The distinctions between Nominal, Ordinal, Interval and Ratio data

1. **Nominal**: named data, distinct data. For instance:
    - black/white
    - brand of car
    - gender
    - nationality
    - ethnicity
    - language
    - genre
    - style,
    - biological species etc.

2. **Nominal data could be:**
    a. **Dichotomous** if there are only two categories eg male/female
    b. **Multi-category** if there are more than two categories and the categories have no inherent order eg
        married single divorced engaged widowed

3. **Data interpretation of nominal data**
    You cannot do much. There are no statistical tests. You cannot add things up. You cannot even use a mean average. All you can do is report what you have counted, and say that one is more than another.

    The only average you can use for nominal data is the mode.

4. **Ordinal data**: data which has a natural ordering
    a. It could be data which is grouped into ordered categories eg 'excellent, acceptable, poor'

    b. It could be data which is numbered in rank order eg 1st, 2nd, 3rd, 4th in a class

    c. It could be a Likert scale, sometimes called a rating scale

5. **We must NOT make interval and ratio claims about ordinal data.**

    So if the average customer satisfaction on Product A is 4.0 and the Average on B is 2.0, **we need to be careful in thinking the difference in satisfaction is twice**. We can

say there is a difference, but we are less sure if it is two times.

NB. The trouble with ordinal data is that you cannot assume the differences between each interval are equal.

6. **Visually inspect the ordinal data before using an average**

When you get data from Likert scales, you need to put it all out in a table, then visually inspect the data. Are there one peak or two?

I hope it is obvious, that if there are two peaks, then you CANNOT use any kind of average.

For instance, in a Likert scale of five points, there might be two peaks: one clustered round 2, and the other clustered round 4, which means you have two distinct groups, one at an extreme, and one somewhere in the middle.

Figure 3.2 Averages for ordinal data

Scale	**1**	**2**	**3**	**4**	**5**
Data A: one peak	5	6	8	15	6
Data B: two peaks	4	14	4	14	4

In data set A, there is clearly one peak, at scale 4. The median is 4, the mode is 4, and the mean is 131/40 = 3.3. In data set B, there are two modes, the median is 3, and the mean is 120/40 = 3.0. But, even just looking at the data ("eyeballing" it is obvious that it is unfair to talk about an average of any kind, since there are two distinct groups.

6. **Interval data**
   a. **Each step up or down is equal**
      This is an important point, and distinguishes such data from Likert scales. With interval data, each interval is the same. Therefore, simple operations such as addition and subtraction can take place.

   b. **Example**
      An increase of one degree Celsius is one degree, whether that be from 0 to 1, or 21 to 22.

7. **Data interpretation**
   a. **Ratios are not allowed**, since 20°C cannot be said to be "twice as hot" as 10°C!

   b. **Interval data allows use of parametric statistics,** which assume a normal distribution. [Questionnaires and surveys use non-parametric tests. In non-parametric work, the data is not "normal'. Experiments use parametric tests, which tend to be more powerful].

   c. The arithmetic mean can be used.

8. **Ratio data**
   a. This is interval data with a natural zero point

   b. **Examples**
      - Time is a ratio since zero time is meaningful.
      - The Kelvin temperature scale is, strictly speaking, a ratio scale since by definition 0K (note, never zero *degrees* Kelvin) is the starting point, known as absolute zero.
      - Most measurement in the physical sciences and engineering is done on ratio scales. Examples include mass, length, duration, plane angle, energy and electric charge.

9. **Interpretation of ratio data**
   Ratios have a non-arbitrary zero point. By this is meant that the zero point has a natural existence. Therefore, it is meaningful to say, for example, that one object has "twice the length" of another. Very informally, many ratio scales can be described as specifying "how much" of something (i.e. an amount or magnitude) or "how many" (a count).

Table 3.1 Scale types with their properties				
	Nominal (unordered)	***Ordinal (ordered)***	***Interval***	***Ratio***
x or ÷				✔
+ or -			✔	✔
> or <		✔	✔	✔
= or ≠	✔	✔	✔	✔
Examples	Gender Nationality	Health Truth Opinion (Likert Scales)	Date Degrees Celsius	Age
Measure of central tendency	Mode	Median	Arithmetic mean	Geometric mean
	Non-parametric **NOT normal**		Parametric IF normally distributed, otherwise use an equivalent non-parametric test	
Typical methods	Questionnaires Surveys		Experiments, and tests such as the scores in an examination	

10. Discussion of the summary table above

a. The table clearly shows that when working with scales, only the ratio scales allow you to use multiplication or division. Addition and subtraction can only be applied to ratio or interval scales.

b. Parametric statistics apply to interval and ratio data. Non-parametric statistics apply to nominal and ordinal data.

11. Can parametric tests be used for Likert Data?

There are two different answers. The standard answer is a clear No. For convenience, rating scales are often numbered. So, when on a one to five scale, people are asked to score something. Technically speaking, **the correct average for ordinal data is the median.**

Others, such as Norman (2010) disagree. He says:

> Parametric statistics can be used with Likert data, with small sample sizes, with unequal variances, and with non-normal distributions, with no fear of ''coming to the wrong conclusion''. These findings are consistent with empirical literature dating back nearly 80 years. The controversy can cease (but likely won't).
>
> In my view Norman (2010) has convincingly shown that while technically you cannot use parametric tests, and cannot use the mean, in practice, the statistical patterns are robust enough that the mean can be used.

CHAPTER 4
THE SHAPE OF DISTRIBUTIONS

1. Introduction

So you have collected your data, entered it into a spreadsheet, and graphed it. The first thing you will notice is how many peaks there are.

Not all data gives a peak in one place. Sometimes there are two peaks (bimodal) or more than one peak (multimodal). Figure 6.1 is an example from Wikipedia of bimodal results from biology.

Most of the statistics works on the assumption that your data has one peak and is symmetrical, that it approximates to a normal curve. Ideally, the shape of your curve, or 'distribution' should be close to the so called 'normal' curve. Statistics programs often give you the option to test your data for normality, and you should always to this if you can and if it is relevant.

Statistics programs often give you the option to test your data for normality, and you should always to this if you can and if it is relevant. See Chapter 5.

Figure 4 .1 Example of a curve with two peaks

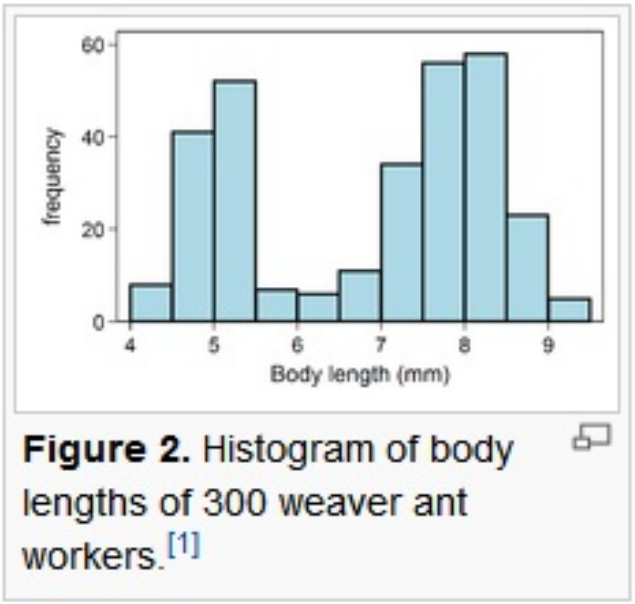


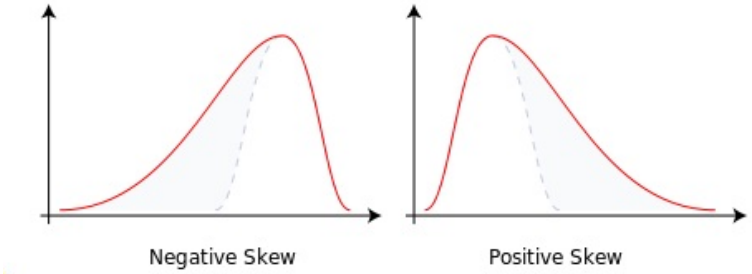
Figure 2. Histogram of body lengths of 300 weaver ant workers.[1]

If the data gives a distorted curve then there are two major types of distortion. These are Skew (also called Skewness) and Kurtosis. There are various ways of giving a number to the extent of distortion. Fortunately, the statistics program way well give you these. If not, you could always try http://www.wessa.net/skewkurt.wasp for a free skew and kurtosis analysis.

2. Skew (See also Key 3 of Feel for Statistics)

a. The skew is the distortion due to unusually high or low figures.

Figure 4.2. Skews



b. Traditionally, textbooks of statistics teach a rule of thumb stating that the mean is to the right of the median under rightskew, and to the left of the median under leftskew. But this rule fails with surprising frequency. It can fail in multimodal distributions, or in distributions where one tail is long but the other is fat. Most commonly, though, the rule fails in discrete distributions where the areas to the left and right of the median are not equal.

c. The best way to consider skew is to actually use your statistics program to give you a graph, and to visually inspect the curve.

d. If a curve is perfectly symmetrical there is no skew. Therefore, skewness is a measure of how far a curve deviates from perfect symmetry. It is a measure of how symmetrical the results are.

e. There are also some common numerical measures of skewness. Some authors favour one, some favour another.

3. Kurtosis

This is the flatness of the graph, sometimes called a 'distribution'. The easiest way to see it is to look at three symmetrical curves.

All three of these distributions have:

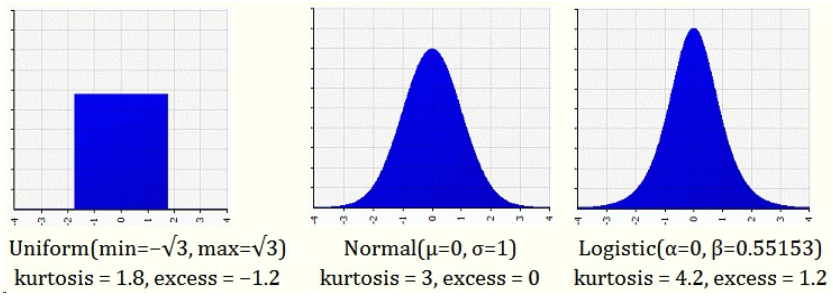
- mean = 0
- standard deviation = 1
- skewness = 0, which means there is no skew whatsoever.

All are plotted on the same horizontal and vertical scale.

So, in theory, all three curves should be identical. But, as you can see from the three curves below, they are distinctly different! There is another feature.

Look at the progression from left to right, as **kurtosis** increases.

Figure 4.3 Kurtosis



(From https://tc3.edu/instruct/sbrown/stat/shape.htm)

As you can see, a flat curve has a low kurtosis, and a sharply peaked curve has a high kurtosis.

Kurtosis is something to be aware of. But tests of normality, and the implications for which test you use, are far more important.

4. Messages to learn

You need to observe your data. That means you need to plot your data, and have a visual display. Then, you need to visually inspect your data. Especially when comparing two groups, this visual inspection is extremely important.

5. Example

Imagine for instance, two teachers were double marking some examinations.

The means for both teachers were identical. Some teachers would conclude that checking the double marking was not needed. They would be wrong, because they had not considered the range. Some teachers mark close to the average, say a small spread of 9-11 marks. Other teachers might use a wider range, say 7-13. In which case, the marking would NOT be comparable.

But, supposing the teachers were smart enough to put their marks on a spreadsheet and they had identical means, and identical standard deviations.

They would have again missed a feature. The two sets of marking were NOT similar, just as curves two and three of figure 6.3 above are NOT identical.

Now, you could use descriptive statistics. You could make sure your statistics program provides the kurtosis. But, why bother? It is much quicker, simpler, and easier to understand, if you simply sketch and look at the two graphs. You can either use a computer, or, often, use a piece of paper to speedily sketch the curves. Then visually inspect them. How similar are they?

In fact, the marking of Teacher A, resulted in a curve like the second curve of figure 6.3, and the marking of Teacher B resulted in a curve like the third curve.

You also need to know if there is skew or kurtosis involved. Then you use this information when you comment on your results, interpret them, and reason with your findings.

Then you can ask the question, **are they similar enough?** And for that, it will be your judgement. Note well, there is no statistical test which will help you. You cannot do a t-test and then look at *p* values, and conclude that there is 'no statistical difference'. That, as we will see later, is a misuse of statistics. It is a very common misuse, but that does not make it right.

6. **Graphics can reveal patterns that are hard to see in data summaries**. In particular:
   a. Does the data have one peak or two? In examinations, in many of them there is a peak, hopefully at over 50% so that must students pass. But a good examination clearly separates out the failures from those who deserve to pass, therefore a bimodal shape is excellent. Simply averaging the results hides this shape, ie information is lost.

      So, an examination which has the lowest pass rate could in fact have a large number of those who pass at a high score.

   b. Is the correlation data linear, so that there is ONE correlation? Why should linearity be assumed?

Reference

Larson-Hall J 2016. *Moving beyond the bar plot and the line graph to create informative and attractive graphics*. Prepublication version.

CHAPTER 5
SUMMARISING THE DATA – DESCRIPTIVE STATISTICS

A. Introduction

1. Descriptive statistics are just that: you present some basic statistical facts about the data. This can include:
   - mean, median, and mode
   - minimum and maximum scores
   - one or more type of range
   - standard deviation and maybe the standard error
   - skewness
   - kurtosis – a measure of how peaked the curve is, how steep is the slope

2. The most basic summary is known as a **frequency table.** The easiest way to imagine this is to take out some money from your pocket, then sort it, and put the smallest coins on the left and the largest on the right.

   From this you can add up how much money you have.

Example 5:1. A frequency table for coins in Tunisia

Coins (denomination)	Number	Total value
5	6	030
10	3	030
20	9	180
50	3	150
100	5	500
200	8	1600
1000	4	4000
2000	2	4000
5000	1	5000
	Total coins = 30	Total money = 15,490 dinars

3. **This material can be visualised on a graph**
   Which type of bars are used, those which are connected, or those with gaps between them? See Answer 1 at the end of the chapter.

4. Example of pulse rates (Rowntree 1981:43) arranged in increasing order

Example 5:2. Pulse rates in increasing order
The table below presents the pulse rates of 50 people, arranged in increasing order.

62	64	65	66	68	70	71	71	72	72
73	74	74	75	75	76	77	77	77	78
78	78	79	79	79	80	80	80	80	81
81	81	81	82	82	82	83	83	85	85
86	87	87	88	89	90	90	92	94	96

Q. What can this tell us?
a. minimum
b. maximum
c. median
See Answer 2. This data could easily be plotted on a graph, but a better way is to group the results as below.

5. **Grouping data**
Commonly such data are grouped. This is not just a matter of convenience which it is. It also recognises that the measuring technique is not that accurate, and approximating to the nearest half centimetre (in this case) is legitimate.

Example 5:3. Grouping the pulse data

Pulse rate (beats per minute)	Number of students (frequency)
60-64	2
65-69	3
70-74	8
75-79	12
80-84	13
85-89	7
90-94	4
95-99	1
	total = 50

Question. How would the graph be drawn of this data? Would the columns be touching or would there be a space between them? See Answer 3 for a commentary.

B. Ranges

1. Introduction

Students often ask a teacher after an examination, as a measure of how hard the teacher has been, what was the lowest mark, and what was the highest mark. This is quick and easy to identify. The gap between them is known as a range: the minimum to the maximum. The trouble is this relies on the outliers, ie the two most extreme cases. These can easily distort the overall pattern. These values may be 'atypical' (which is the technical word for 'not typical').

If there are 99 students, and the marks are put in increasing order, then the range is between the lowest mark and the highest mark, ie the mark of the first student and the mark of the last student. These marks are easily misleading, since one bad mark, or one really high mark, does not present what the majority of students were awarded.

Therefore, there are several different ways of expressing a range.

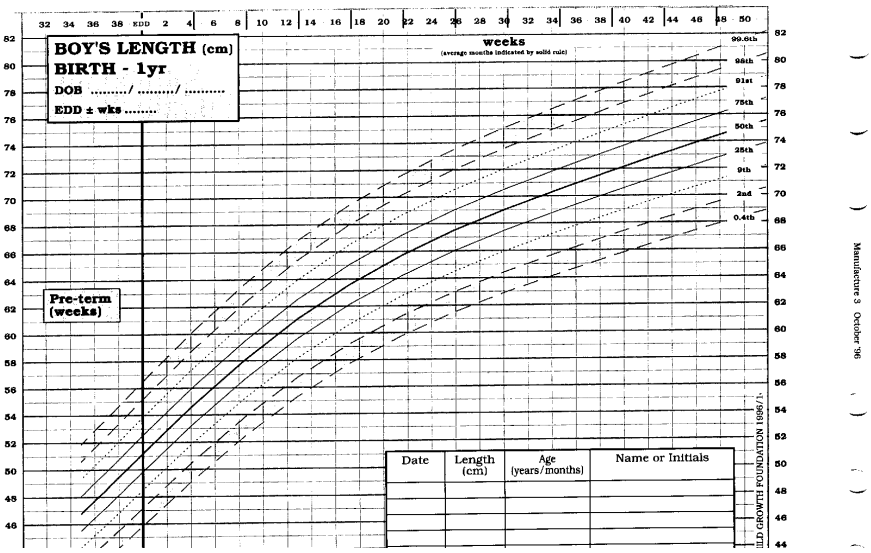
2. The minimum and the maximum

This is by far the easiest. Often, these values can be found by simply inspecting the data. If using a spreadsheet there are functions, which have the advantage of being flexible, and changing whenever required by modifications to the data

3. The inter-quartile range.

If there are 99 students, we put the marks in increasing order. Then we find the marks of the 25^{th}, the 50^{th} and the 75th students. The 50^{th} mark is our old friend, the median. The other two marks together divide the marks into quarters. The inter-quartile range is the difference between the 25^{th} and the 75^{th} mark. It has the great advantage of not being affected by extremes of high or low.

Example 5:4 The ranges for the increase in length of baby boys from birth to 12 months



(Scanned from a free booklet given to parents in Britain).

The chart above is for the growth of young male babies. The percentile lines are clearly drawn. By regular measurement it can be seen for instance if the baby is growing too fast, or too little. What matters in this chart is to see if the position of the baby on the percentile line is changing. For instance, if a baby starts at the 50th percentile, then suddenly shoots up to the 95th, then they are probably eating too much. If they start at the 50th and go down to the 5th then there is probably something wrong. But a baby at the 10th percentile who stayed there would have started low but experienced normal progress after that.

4. **The Standard deviation** (Rowntree 1981:53ff.)
   The standard deviation is a more sophisticated measure of the range. When a teacher gives mainly marks in the range 9-11 out of 20, we notice that the range is very small, the teacher is being cautious. In this case the standard deviation would be

small. But some teachers deliberately use a mark scheme that enables a wider spread of marks to be awarded. Some teachers then may have an average mark of 10, but give more low marks, and more high marks: they mark in the range 4-16. This is (arguably) a fairer way of marking, since it is clearer who has passed and who has failed. When a smaller range is used, the element of change is more likely to intervene and give false positives and false negatives. (Rowntree 1981 p54). The standard deviation is a way of indicating a kind of 'average amount' by which all the values deviate from the mean. The greater the spread, the bigger the deviations, and the bigger the Standard Deviation.

The standard deviation is the average deviation from the mean. In the examples below you will work through how it is calculated.

Example 5:5a. For practice

Though the computer, and even a hand calculator, can easily calculate a mean and a standard deviation, it is helpful to do some worked examples first. In this way you will get a hands on feel for what is actually happening. Probably the easiest way to work these examples is to use a table. In the following data the previously calculated Mean=116. In words, we say:

1) calculate the deviation from 116
2) square these deviations
3) add them up and divide by the number, ie take the average of them.
4) take the square root of the average. This is to maintain consistency with units.

I strongly encourage the reader to work through these examples manually so that you get a feel for what is going on. Otherwise much of statistics will just be like playing with magic numbers.

Presented as a table, a method I encourage people to use, especially at the beginning, we have the following:

Example 5:5b. Showing how to calculate a standard deviation

Previously calculated mean = 116	111	114	117	118	120
deviation from mean	-5	-2	+1	+2	+4
deviation squared	25	4	1	4	16

Now to take the average of these squared deviations:

$$\frac{25 + 4 + 1 + 4 + 16}{5} = \frac{50}{5} = 10$$

Then, because all the deviations were squared to get rid of the problem of negative numbers, we take the square root, which makes 3.16.

We write: mean = 116, SD = 3.16

6. What is the normal curve?

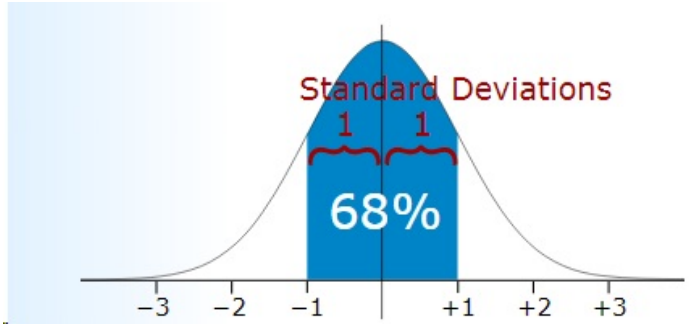
This is a bell shaped curve, sometimes called a Gaussian curve. It is very common in social sciences and in nature. For a true normal curve:

- Mean = Median = Mode
- It is symmetrical
- 50% of the values are less than the mean
- 50% of the values are more than the mean.

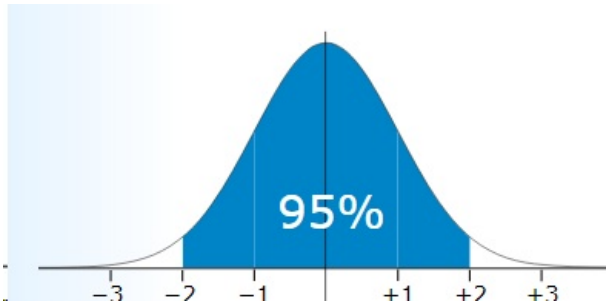
7. What is a Standard Deviation?

It is a measure of how spread out the numbers are.

68% of the values are within 1 standard deviation of the mean.



98% of the values are within 2 standard deviations of the mean.



Pierce, Rod. (12 Jan 2018). "Normal Distribution". Math Is Fun. Retrieved 24 May 2018 from www.mathsisfun.com/data/standard-normal-distribution.html

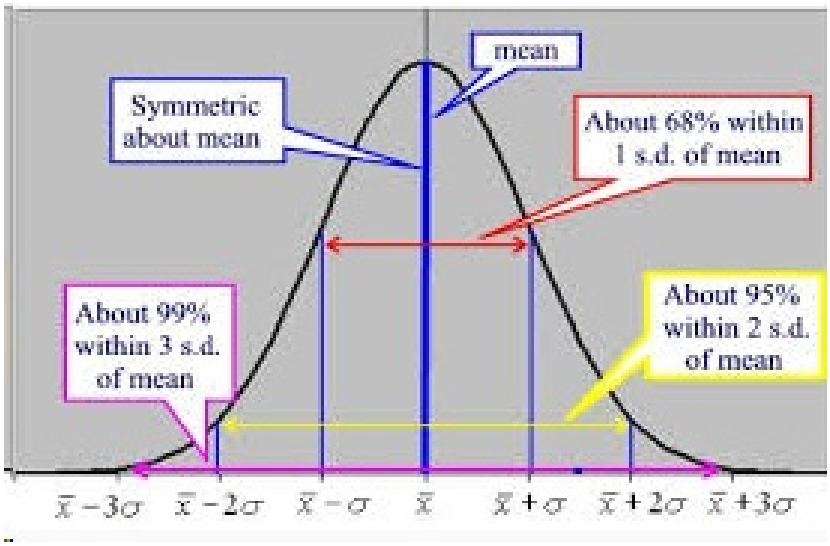
Results tend to organise themselves in the so called 'normal' curve. In which case we find that:

a. A Mean ± 1 SD covers 68%, ie 2/3. This means that 68% of all the observations in a normal distribution lie within 1 SD either side of the mean (Rowntree p72).
b. 95% of observations lie within 2 SD (actually 1.96 SD)
c. 99% of observations lie within 2.5 SD.

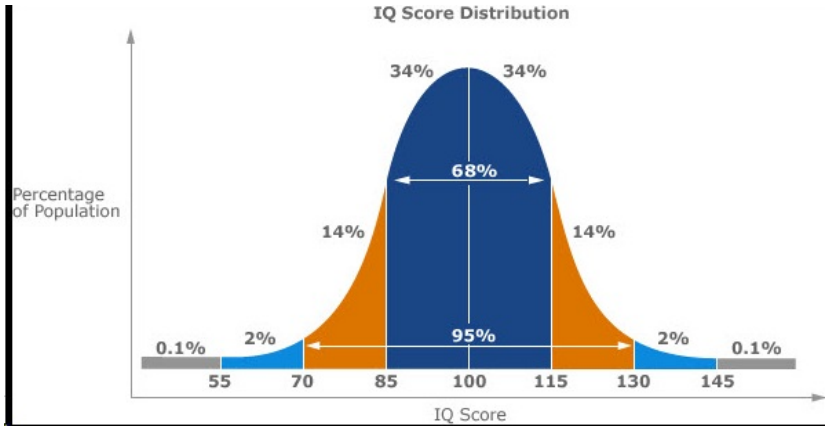
Given the mean and the standard deviation we can say the following:

68% of the data lies between: mean minus 1SD to mean plus 1SD
95% of the data lies between: mean minus 2SD to mean plus 2SD
99% of the data lies between: the mean plus or minus 2.5SD.

For instance, in an examination where
the mean = 10
the SD = 2,
then 68% of the students received between 8 and 12
95% of the students obtained between 6 and 14



Applied to intelligence, you get the following graph.



Here you can see that 68% of people have an IQ between 85 and 115, while 95% are between 70 and 130.

8. Which Standard Deviation?

There are two important ways of calculating SD. You do not need to know the details to be able to decide which one to use. Excel calls them STDEVPA AND STDEVA. The version STDEVPA assumes that you are using the entire population. If your data represents a sample of the population, you must compute the standard deviation by using STDEVA.

For large sample sizes, STDEVA and STDEVPA give approximately the same results.

If in doubt, use STDEVA, as Cumming (2012 Ch4) does in most of his book.

Answer 1

The bar lines must not touch, since the coin sizes are distinct – often called 'discrete' in technical language.

Answer 2

The minimum pulse is 62, and the maximum is 96. The median is the mean of the scores 25 and 26. The 25^{th} pulse is 79 and the 26^{th} is 80 therefore the median is 79.5 (the number between 25 and 26).

Answer 3

In this case the bar lines would be touching since the scale goes from 60 to 100, and the scale is continuous.

CHAPTER 6
DEPENDENT, INDEPENDENT, AND INTERVENING VARIABLES

1. Introduction

In science, **the quantity that you fix goes on the x-axis (the horizontal axis) and the data you collect by experiment goes on the vertical y-axis.** This is a well established custom that a researcher ignores at their peril. Most scientists are so habituated to this that they draw their graphs with scarcely a thought to the distinction.

The problem is that among students in social sciences the distinction is not always innate. In addition, computer statistics packages make it harder, because when working with them the menus will often ask you to specify which is the dependent variable and which is the independent variable, instead of asking you the questions: what do you want to go on the x-axis? What do you want to go on the y-axis?

2. Dependent versus independent variables

A dependent variable is the outcome variable – it is what you measure, it is the data you collect. Y-axis.

The independent variable is the settings and categories you make. X-axis. Here are some common independent variables:

1. Male/Female
2. Age
3. Languages spoken
4. Profession
5. Level of education
6. Grew up in a city or a town/village

If you think the variable will be an explanatory factor for some of the other results, then it is an independent variable.

Handling experts tip 2

- Ask the expert to check that which variables are dependent and which are independent.
- When you see the graph, check it by eye. Make sure the experimentally collected data is on the y-axis. If not, then backtrack and swap around the settings and you will see the same graph where the x and y are changed round.

Example 6:1

Is there a relationship between doing more practical exercises for homework and the weekly test in mathematics?

In this case the hours of homework go on the x-axis therefore they are the independent variable, and the test results obviously belong to the y-axis therefore they are the dependent variable.

Notice how I have understood the question. Results data – the observations and findings, belong to the y-axis, and this is called the dependent variable. So, whenever you see a choice box in a statistical program, just speed translate the misleading terms:

dependent variable (Stats package) = y-axis results (common sense).
independent variable (Stats package) = x-axis (common sense)

3. Intervening variables

If the results showed that more homework time meant higher test scores, can we assume that homework caused the better performance?

Maybe the threat of a test increased the amount of time spent on homework! Maybe lower marks motivated students to do more homework. Maybe the threat of punishment due to low marks influenced the time spent on homework.

There are several points here.

a. Did increased homework increase the test scores, or did the test scores encourage more homework? The direction of causality is not always clear and may be bi-directional.

b. Proven association (correlation) does NOT mean there is a link of cause.

c. The link may be a curve: it could be that too much homework and too little homework is linked with poor test results.

CHAPTER 7
CONFIDENCE INTERVALS AND CONFIDENCE LEVELS

NB they are often quoted together and they are NOT the same!

1. **The confidence interval** (also called margin of error)
   This is the plus-or-minus figure usually reported in newspaper or television opinion poll results. For example, if you use a confidence interval of 4%, and 47% percent of your sample picks an answer, you can be "sure" that if you had asked the question of the entire relevant population between 43% and 51% would have picked that answer.

   Higher power in a study will result in smaller confidence intervals ie the gap between the two lines or points will be smaller.

   Note, Confidence Intervals are often disconcertingly large. This is in fact healthy: they show the how wide the variation is, and help us to face up to realities sooner rather than later. Natural variation is often large, and drowns any effect or variable being studied.

   The problem is, how is this confidence interval to be defined? Is it to be the interval between the minimum and the maximum? Or does it have something to do with Standard Deviations?

2. The confidence level

Statisticians love to use the normal curve. Where to draw the line is known as specifying the confidence level. For convenience, this drawing the line is usually done in terms of standard deviations. Where to draw the line is a matter of convention, and, right now there is no widely agreed convention. Therefore, you have to state which convention you are using.

There are two popular conventions.

- Plus or minus one standard deviation = 68% of the data.
- Plus or minus two standard deviations = 95% of the data.

The MOST popular convention is the second one. But because the researcher can choose whatever levels they want, this means that the researcher has to state the confidence levels, every time. Researchers can if they want, for different parts of their results, use different levels.

3. Combining Confidence Intervals, and Confidence Levels

When you put the confidence level and the confidence interval together in the example given above, you can say that you are 95% sure that the true percentage of the population is between 43% and 51%.

With a 95% confidence level, if the study were repeated 100 times then 95% of the time the result would be found within the stated confidence intervals.

The wider the confidence interval you are willing to accept, the more certain you can be that the whole population answers would be within that range.

4. Online calculators

- GraphPad QuickCalcs

http://www.graphpad.com/quickcalcs/index.cfm

- Measuring

http://www.measuringusability.com/ci-calc.php
Very helpful short description
http://www.measuringusability.com/blog/ci-10things.php

- Survey system

http://www.surveysystem.com/sscalc.htm
Provides a calculator and some helpful notes.

The one I use is:

- For correlations: http://www.vassarstats.net/rho.html
- For other examples, search the site

NB 5. Factors that affect Confidence Intervals

There are three factors that determine the size of the confidence interval for a given confidence level:

- Sample size
- Percentage
- Population size

6. Sample Size

The larger your sample size, the more sure you can be that their answers truly reflect the population. This indicates that for a given confidence level, the larger your sample size, the smaller your confidence interval. However, the relationship is not linear (ie doubling the sample size does not halve the confidence interval).

7. Percentage

In polls, multiple choice answers, Likert scales etc, your accuracy also depends on the percentage of your sample that picks a particular answer. If 99% of your sample said "Yes"

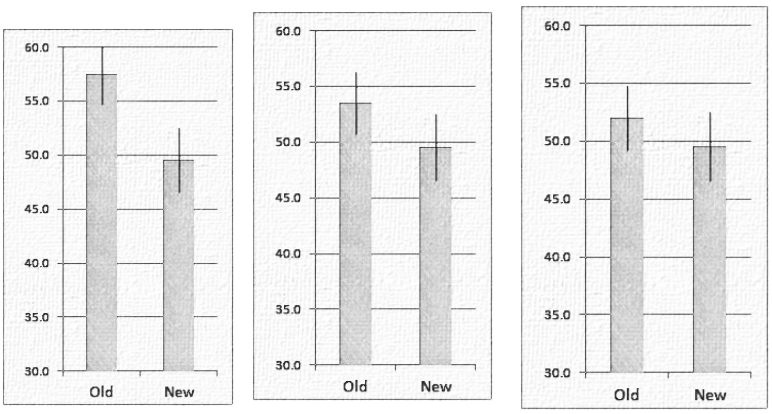
and 1% said "No," the chances of error are remote, irrespective of sample size. However, if the percentages are 51% and 49% the chances of error are much greater. **It is easier to be sure of extreme answers than of middle-of-the-road ones.**

When determining the sample size needed for a given level of accuracy you must use the worst case percentage (50%). You should also use this percentage if you want to determine a general level of accuracy for a sample you already have.

ompare with Chapter 15 point 9, where it is stressed that a smaller CI implies higher precision, ie LESS margin of error.

8. Population Size

Figure 7.1 Illustration of Confidence Intervals applied to groups



From: http://www.measuringusability.com/ci-calc.php . Note, the diagrams have been modified using portablefotosketcher, a free program for adjusting photos and documents.

Population size is only likely to be a factor when it is small.

The confidence interval calculations assume you have a genuine random sample of the relevant population. If your sample is not truly random, you cannot rely on the intervals. Non-random samples usually result from some flaw in the sampling procedure.

9. In practice, using confidence interval graphics as a rough and ready check for a difference

In the first graph there is clearly no overlap. In the second graph there is some overlap, therefore decisions will be difficult. In the third graph there is a lot of overlap, and the results are probably very similar.

Notice how the confidence intervals are shown on the graphs by vertical black lines at the top of each column. When there is no overlap the difference is significant, and you do not need a statistical test to show that. When there is a large overlap the difference is not significant and no more statistical tests are needed.

Brief note on old statistics

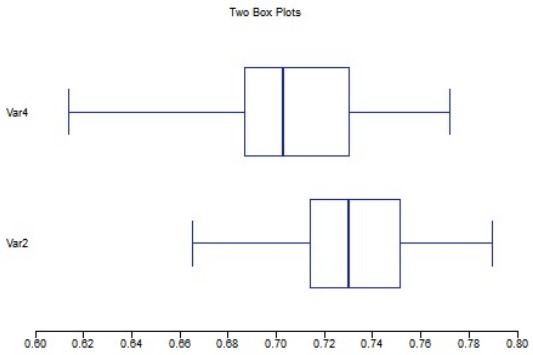
In the new statistics, that is it. You must decide, based on the situation, if the difference is significant. And 'significant' will be how you define it. Preferably, you should define in advance what you will accept as significant.

For the sake of compatibility with the old statistics, When there is some overlap you need to use a 2 sample t-test. In traditional statistics, you can use the overlap in confidence intervals as a quick way to check for statistical significance. If the intervals do not overlap then you can be at least 95% confident there is a difference (for 95% confidence intervals). If there is a large overlap, then the difference is not significant (at the $p < 0.05$ level). The intervals can actually overlap by as much as 25% and still be statistically significant, so when there is some overlap, it's best to conduct the 2-sample t-test and find the *p*-value. The three graphs above illustrate this.

9. Another way of visually inspecting variance

Another way is to look for Box Plots. Two of them from the free programs mentioned are illustrated below.

Figure 13.2 Box plots



What the different parts mean will depend on the context. A common meaning is that left and right are the two extremes, the box represents 68% confidence intervals, and the vertical center line is the mean.

These figures provide some other important information:

Look carefully: the two halves are not equal. This means that the data is skewed, is not normally distributed, and that the confidence interval to the left of the mean is NOT the same distance from the mean as the confidence interval to the right.

Therefore, it would be completely wrong to state the confidence interval in these cases as the mean ± a figure.

10. Presentation of confidence intervals

APA style (2009) requires that the confidence intervals be placed in square brackets after the figure being qualified. This applies to text and tables. In tables, there is also the option of providing a separate column for them. Note, since it is a matter of free choice which confidence levels are chosen, then these must be stated, and it helps if you are consistent throughout your work. The common confidence levels are 68% and 95%. **For the exact format, see the latest APA guidance.**

Some examples:

- *M* = 30.5cm, 95% CI [18.0, 43.0]

Endnote

11. Confidence Intervals and Standard Error bars

Many conventional textbooks of statistics will talk about Standard Error. Unfortunately, though SE has an equation, it is not at all clear to me why SE is used and even less clear how to interpret it. Therefore, following the advice of Cumming (2012 Ch4) I advise:

- You can set the CI at any value. The most common level is 95%, but you could set it at 68% which is one Standard Deviation.
- For most purposes, one SE = 68%.
- If you see a SE, double the length of the bars and you will get 95% CI.
- CI at 95% is far less misleading, and is much easier to read, since 95% of the results are within this range. This is the easiest range to manipulate and assess.

CCJ (2017) say "The SE formula is vital". The most common equation is easy enough. The SE is the standard deviation divided by the square root of the sample or population.

If you are confused, then you are not alone. I have yet to find one who can explain a SE. I have been struggling with statistics for over forty years and teaching them for over thirty years. It is highly unlikely that I am right. But.

Since SE is derived from two useful scores, the SD and the Mean, I cannot see why I need to bother.

I can see that when N (sample size) is small, there is wide sampling variability. That means, the chances that two samples are identical is gloriously small, and the chances that my sample is close to the true population is also small, therefore be cautious in extending the results to a wider group.

I think SE is the fancy name given to this. But since the underlying reality is so obvious, I fail to see the need for the SE.

My take home message is this. Set the CI to 95%, use graphs, and follow your nose. The SE will take care of itself.

CHAPTER 8
THE CORRELATION COEFFICIENT R

1. Use

To measure the strength of correlations, linkage, between two variables.

2. General situation

In science we commonly want to know what is the link between two variables. We usually assume in basic statistics that the relationship is **linear**, which means that there is a straight line relationship, not something more complicated like exponential or logarithmic. $y = mx + c$, or $a = bx + c$ is the relationship we are testing. There are more complicated tests to cope with non-linear relationships. Some of the common alternatives include semi-logarithmic, logarithmic, quadratic, and exponential curves.

- Non-linear is quite common.
- In phonetics, the decibel scale is semi-logarithmic, so that an increase of 10 decibels is equivalent to a doubling of loudness.
- The weight of a person is related to the square of their height. This is used in the famous BMI, the Body Mass Index, which is a reasonable indicator of body fat.

NB

A relationship does not have to be linear. There are many other alternatives. A good question to ask about your data, at an early stage is this simple question: Is the relationship between the two variables linear? If YES, proceed with caution. If NO, or NOT SURE, then stop, and seek urgent skilled advice.

3. Example situations:

a. The link between the baccalaureate score in English and the reading comprehension, writing. and grammar marks in the first year of university.

b. The link between the overall English mark and your own test.

c. The link between intelligence and manual dexterity. It could well be negative, and popular believe is often that intelligent people are not practical.

d. The link between crime rate and unemployment rate for each of the countries in Europe.

e. The clear link between radius and circumference of a circle.

4. The link between the theory marks and the practical marks

Rowntree (1981:158) gives the following example of theory and practical marks for ten students.

Figure 8.1: Data example to link theory and practical marks

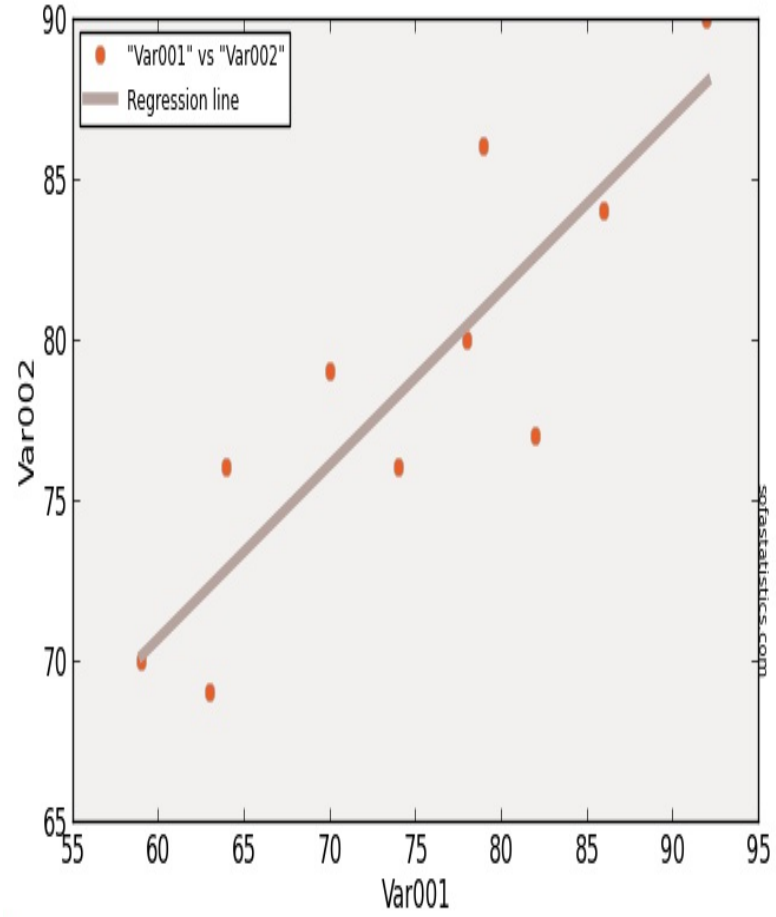
Student	Theory test (%)	Practical test (%)
A	59	70
B	63	69
C	64	76
D	70	79
E	74	76
F	78	80
G	79	86
H	82	77
I	86	84
J	92	90

These marks can be plotted on a graph, presented below using SOFA. The graph shows that there is a relationship between the two examinations: higher scores on one are linked with higher scores on the other. But there are exceptions.

Note, the easiest way to put Data into SOFA is to create an Excel 2003 file ie a file with xls as the suffix NOT xlsx. If you are using Excel 2010 simply use File|Option|Save As.

What should be noted is Pearson's r which is 0.872. The relationship is positive, and approaching the maximum value of 1.0.

Figure 8.2: A correlation graph
Note: both variables are continuous data, therefore it does not matter which axis is used.



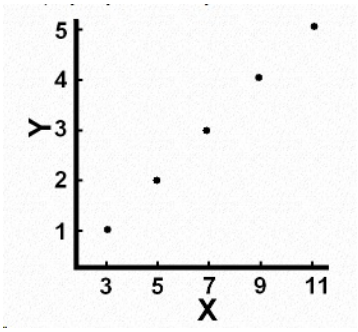
5. **Describing and interpreting the results**
   In order to interpret the results, you need to know what a perfect correlation looks like. These are presented below.

Remember, a correlation can be **negative**. Another way of saying this is that the relationship is **inverse**. Variable A is inversely related to Variable B.

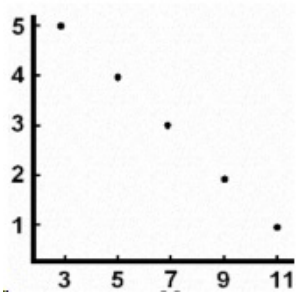
Mathematically this would be written:

$$A \propto 1/B, \text{ or } A \propto \frac{1}{B}$$

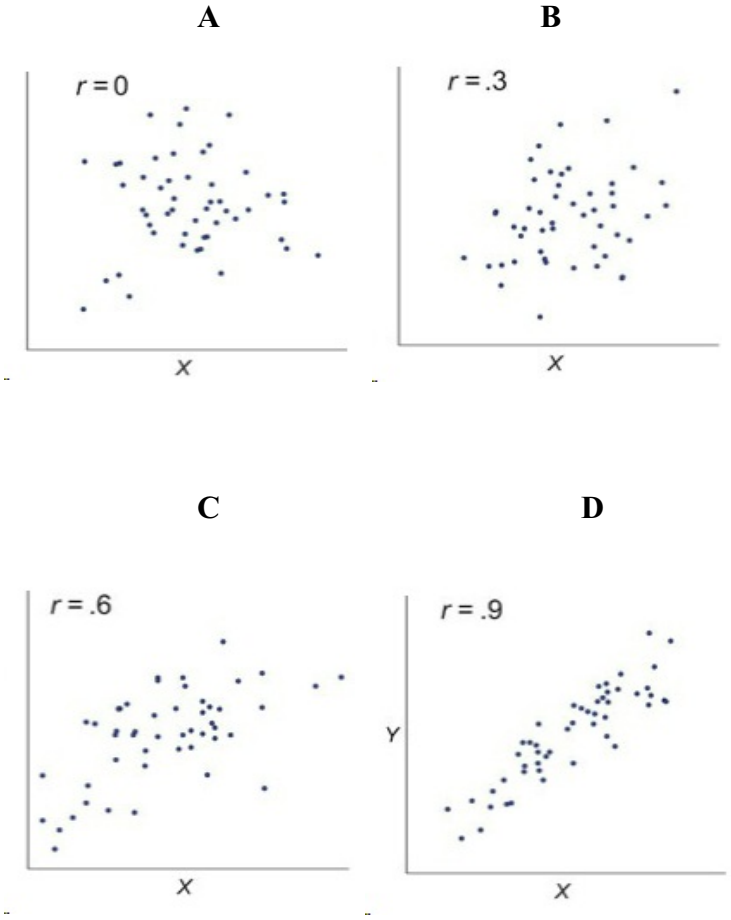
a. **Figure 8.3: Perfect positive correlation, r = +1**



b. Figure 8.4: Perfect negative correlation, r = -1, for instance, when air pressure increases, volume decreases



c. Figure 8.5: Some other scattergrams, for eyeballing
(From CCJ Chapter 11).



Why is it important to study these graphs?
Plot A shows what zero correlation looks like. It is a picture of randomness. Remember, r = 0 means no correlation, and r = 1 or -1 is perfect correlation.

Plot D shows very high correlation. The real world is rarely so high, get it is still messy compared to physics.

Plot D shows what many would call a 'high' correlation, with a 'marked' relationship, but it is still messy. How messy is explained when Confidence Intervals are applied.

Plot B shows a very common correlation level in social sciences. Visual inspection and comparison with Plot A show that the correlation is little better than none at all.

6. The words you should use

Rowntree (1981:170) has the following list of how to describe a correlation.

Figure 8.6: Correlation and interpretation

correlation	interpretation
0.0 to 0.2 very weak, slight	relationship is so small as to be negligible
0.2 to 0.4 weak, low	weak relationship
0.4 to 0.6 moderate	substantial relationship
0.6 to 0.8 high	marked relationship
0.9 to 1.0 very high correlation	very strong relationship

Cohen Manion & Morrison (2011:637) divide things up differently.

Figure 8.7: Another correlation and interpretation

0.2 to 0.35	A correlation of only 0.2 shows that only 4% (0.2 x 0.2) of the variance is common to the two measures.
0.35 to 0.65	Within this range correlations are statistically significant at 1% level. Cautious and rough prediction for a group is possible though not for an individual case.
0.65 to 0.85	Correlations within this range can make possible some accurate group predictions.
0.85 to 1.0	Correlations as high as this indicate a close relationship between the two variables.

All this means, once again, **take great care in interpreting the results. Context is everything.**

Choose ONE style of words, and stick to it. Examiners tend to be fussy about these matters. See the end of the chapter for more, and new material on this question.

7. Writing in the report

In a report you could write something like this. "There is a significant positive relationship between arithmetic scores and English scores r = + 0.75, 95% CI [0.6, 0.8] which states the Confidence Intervals of the correlation.
(See also the end of chapter 7).

8. Commentary on an example

Now imagine you have done the computing, and you are in the proud possession of a set of figures such as those relating theory and practical (Rowntree1981:158), and reported in full above in Figure 8.2. You will see that r = 0.87, which looks wonderfully scientific and a figure that will amaze you examiners!!

But actually, if a student only does this part and presents the figures they have miserably failed to present and interpret the statistics!!

I insist on this point. Even if you get someone else to do the calculations for you, it is your responsibility to plan the data collection and interpret the data and the results of the statistical tests. Planning experiments and interpreting the results is well within the capabilities of MA students. No one says it will be easy. Students must learn to enjoy struggling to understand, and be prepared for the exhilaration of the sweat and tension as they grapple with such topics.

What is the problem? The correlation coefficient was correctly calculated, and it is high. Wow! A professor of mathematics even confirmed that the maths were fantastically correct. What could be the problem?

Even if you get someone else to do the calculations for you, it is your responsibility to plan the data collection and interpret the data and the results of the statistical tests.

The data concerned ten students, and the relationship between the theory mark and the practical mark. There are only ten pairs of observations. It was not possible to test ALL the students. It is quite possible, quite conceivable, that if you had tested ALL the students you would have got a correlation of zero, or even a negative correlation.

Therefore there are **two possible explanations** for your correlation coefficient r:

a. The correlation is real, and it showed up fairly in your sample, and would apply to all the students if all their marks could have been used.

b. There really is no relationship at all between the theory and practical marks. But accidents do happen, especially when you take only ten students.

There is no statistical test that will decide between the two alternatives. None whatsoever. Even if the *p* level is good, this would mean nothing. Only experience, clear reasoning, and evidence, can decide.

In this case, experience predicts that the correlation coefficient was a fluke, ie pure chance. No one in his right mind would then go on to present the result as a strong fact. In simple terms, small samples have extremely low validity. And that should be the end of it.

9. General comments on interpreting correlation coefficients (Burns 2000:248ff)

a. The inherent relationship between two variables may vary with the circumstances and the population.
Example 8.2 Among children aged 10-16 there is a strong link between physical prowess and chronological age

It would be easy to extend the interpretation, and say that this continues, at least till the age of forty. In fact, among adults of 20-26 there is no such link.

Example 8.3 Among children, the variables 'mental age' and 'chronological age' are positively correlated.

Again, in middle age, and old age, you might expect a similar correlation. In fact, among the middle aged there is no correlation, and in the elderly they are somewhat negatively correlated, ie inversely correlated (older people are more likely to be less alert, and have poorer memories).

b. Be very very careful in extending results to the whole population, or the other way round. For instance, if we correlate creativity with IQ scores for the whole population, there is a strong association. But if we look at only university lecturers, there Burns (2000:249) reports that the correlation is zero. In the population as a whole there is a strong association, but not in this particular subgroup. For University professors, creativity is NOT linked with intelligence.

The problem can exist the other way round. **It is quite possible that a strong association exists with a subgroup, but does not exist in the population as a whole.**

Reasoning from the whole to the subgroup is dangerous.
Reasoning from the subgroup to the whole is dangerous.

c. **The problem of the intervening (third) variable**
Two variables may be linked. We may have shown a strong link between them. But they are not linked causally: one does not cause the other. Burns (2000:250) gives the example of the seaside. It could probably be shown that there is a strong association between seaside accidents and drownings, and sale of ice-cream. But this does not mean to say that somehow accidents cause the sale of ice-cream, or that ice-cream causes accidents. There would be other

variables involved. For instance, when it is very hot, more people go to the seaside, more people buy ice-cream, and there are more accidents.

> **"While we can predict the likely occurrence of one event from another event, we cannot say that one event is the cause of the other. This statement cannot be over-emphasised"** Burns (2000:250)

To put it another way. The fact that you always find two things going together, does not mean that one causes the other.

d. Some correlations are meaningless

Some correlations, while being mathematically correct, are in fact meaningless. Therefore all results must be looked at in their context. The classic example is the following, as reported in the New Scientist, 15 Sept 1990. "Two countries that head the world's longevity tables are Iceland and Japan. In Japan, women live an average of 82.4 years and in Iceland they live 81.5 years. What do these two countries have in common? They are the only two developed countries that do not put their clocks back in winter". Obviously in this case there is no meaningful correlation at all between changing the clocks and longevity of life. The correlation though nicely illustrates the need to be careful.

If you are interested in this, then have fun viewing the old and new versions of this website: www.tylervigen.com/

e. Beware the ecological fallacy

Even when a relationship, the *r*-value is strong, this does NOT tell you the relationship is causal. Just because a town has a large number of unemployed people and a very high crime rate does not mean that unemployment causes crime, or that crime causes unemployment. Cause requires much greater proof than an association. In particular, cause requires a proven mechanism. (Greenhalgh 2007:83).

For decades the cigarette industry was able to avoid blame because everyone knew that association was not causation.

It took years to actually show how smoke caused the damage in cells, and to provide direct evidence, and, crucially, to prove a mechanism for the association.

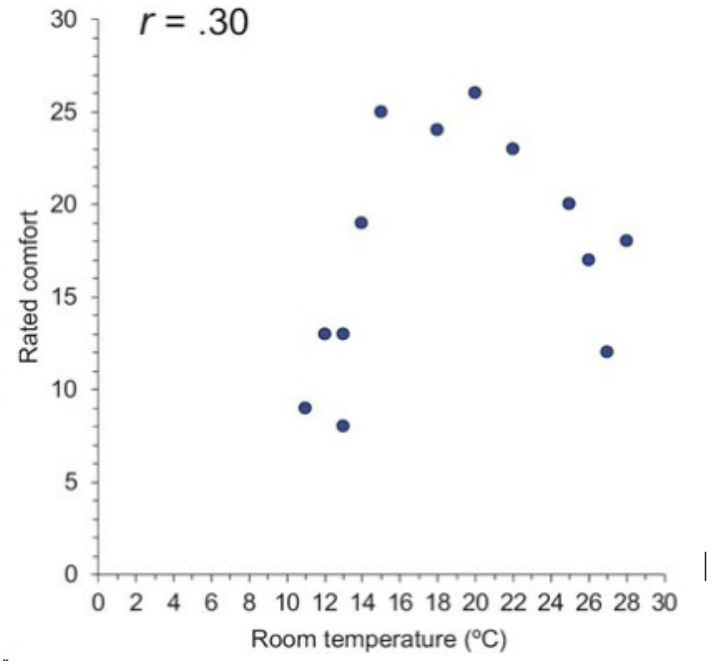
10. What to lo look for in a scatterplot

(These points are based on Cummins and Calin-Jageman (2017) chapter 11. **I have abbreviated this to CCJ.** I cannot give page numbers since I am using an electronic edition whose page numbers bear no relation to the original printed version.)

a. Pearson correlation r is a measure of the strength of the linear component of the (X, Y) relationship. This is important, because when you look at a scatterplot, a curve would make a better line for the data! But, all too often, we do not look, or we decide to try to fit a straight line to the data.

For instance, a curve would best suit this data, than calculating the r-value. When you calculate it, r - .3, because r values assume a linear relationship.

Figure 8.8 Example of a low correlation that is probably a curve or two lines



(Taken from CCJ, ch 11. Figure 11.4a.)

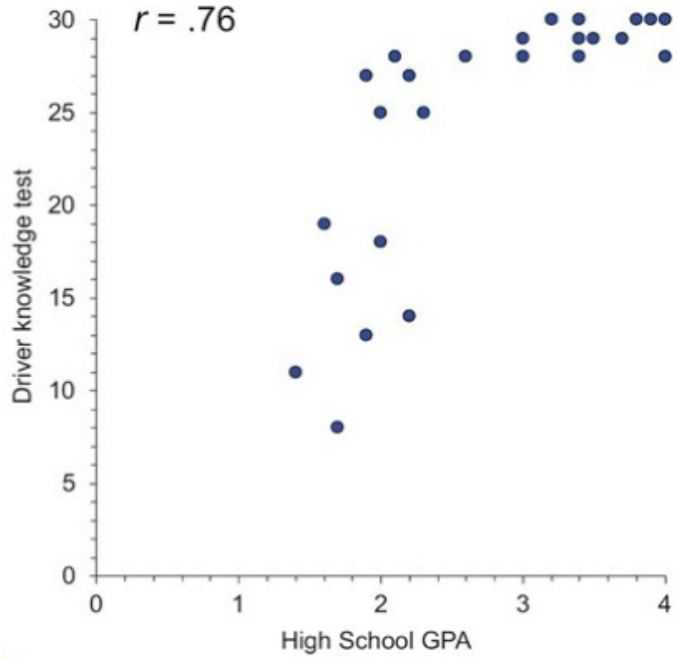
Interpreting the data, it is clear that for the group of people tested, there is an optimum temperature for comfort of around 16-20 degrees. Below this plateau, and above this plateau, comfort level falls off.

It would be interesting to do this experiment in Tunisia, and to do it at various times of the year, and to relate it to plumpness, or BMI, and to the question of heating/air conditioning, and age. Think: What predictions could you make?

Possible predictions:

- There will be a distinct seasonable adjustment. The optimum in winter will be several degrees lower than the optimum in summer.
- Those with heating/air conditioning are more likely to maintain similar optimums, regardless of the season
- Thin people will prefer a higher value than 16-20 degrees, even in winter
- Many people in June would state that their optimum was 25 degrees, or 25-30 degrees.

Figure 8.9 . Correlation between a driver knowledge test, and general school knowledge.



High school GPA is a general knowledge test used in America, in part, instead of baccalaureate examinations.

(Taken from CCJ Ch 11 Figure 11.4b). As CCJ say, the r value does not represent the relationship well. In fact, the curve represents reality. There is a steep beginning, in which a small change in general ability is linked with a large change in driver knowledge. Then, the relationship is almost flat.

b. Even for linear relationships, when *r* is around .3 or less, the scatterplot is close to a shotgun blast. Even for larger *r* there is considerable scatter. This is explained further when I talk about margins of error for correlation coefficients.

c. For large *r,* tightness to the line is helpful for eyeballing the value of *r*.

d. Outliers can have an enormous influence on *r*, which is especially sensitive to points far from the means of X and Y.

e. Examine the scales on the X and Y axes for any sign of a range restriction. For instance, on the X the data may start at zero, but not on the Y, which will give an impression of a tight fit.

f. If you see a correlation, but the scatterplot is not provided, then you need to take care. It is quite possible that the simple scatterplot would reveal crucial information.

In writing theses. When presenting correlations. Do not be afraid or ashamed to present simple scatterplots. They are easily done, take little space, and can be very informative.

11. Margins of error in correlations

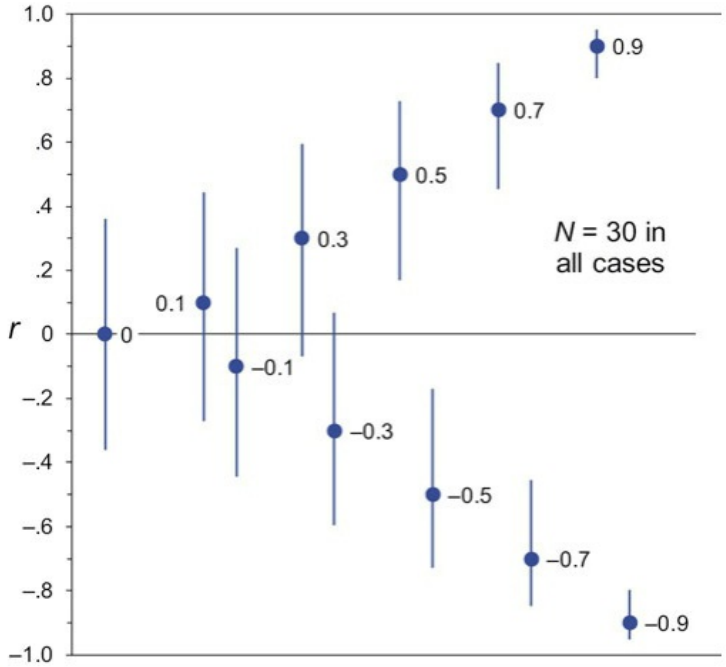
First, some basic predictions:

- High correlations, both positive and negative, will have low margins for error
- Low correlations have the most margin for error

That is in fact exactly what we find.

CCJ chapter 11 have produced a very helpful set of figures, which ought to be widely studied and used. They use 95% Confidence Intervals, which basically state where 95% of the time, the data point is likely to be. In effect, they are the Margin of Error using the convention, 95% Confidence Interval. The first example is when N = 30 ie the sample size. This is a very common number used in research.

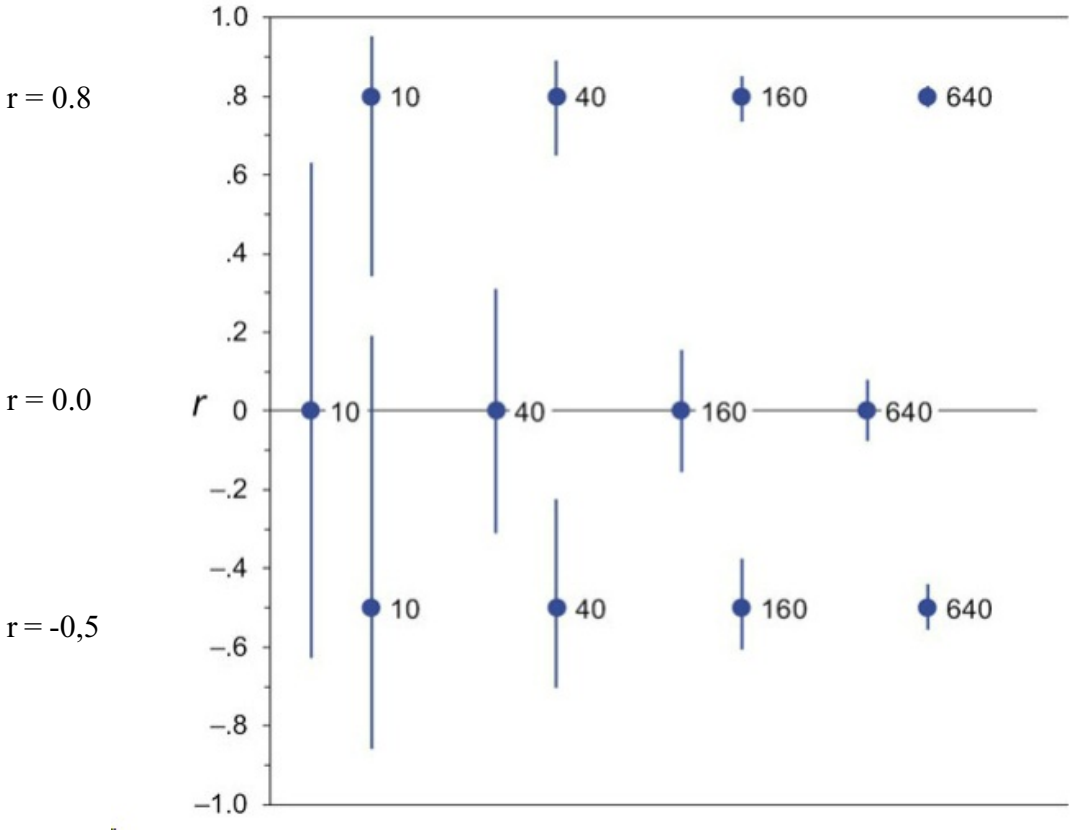
Figure 8.10 Margins of error of correlation coefficients for samples of 30.



Points to note

- The margins of error are not symmetrical. This is a fact of life that is difficult to explain.
- Even zero correlation could be as high as + 0.274 or - 0.274
- Only r=0.4 is to be high enough that you are reasonably confident the correlation is not zero.
- Weak correlations have a huge MoE.

Figure 8.11 Margins of error of correlation coefficients 0.8. 0, and -0.5, for different sample sizes.



The graph clearly shows the effect of increasing sample size, as well as the previous effect of higher correlations. Small samples have a huge MoE.

Finally, because most of the correlations are below 0.5 the following graph is useful. It shows the margins of error given 95% Confidence Intervals, for three common correlation coefficients: 0.1, 0.3, and 0.5.

Figure 8.12 Correlation coefficients and margins of error for different sample sizes

r= 0.5

r= 0.3

r= 0.1

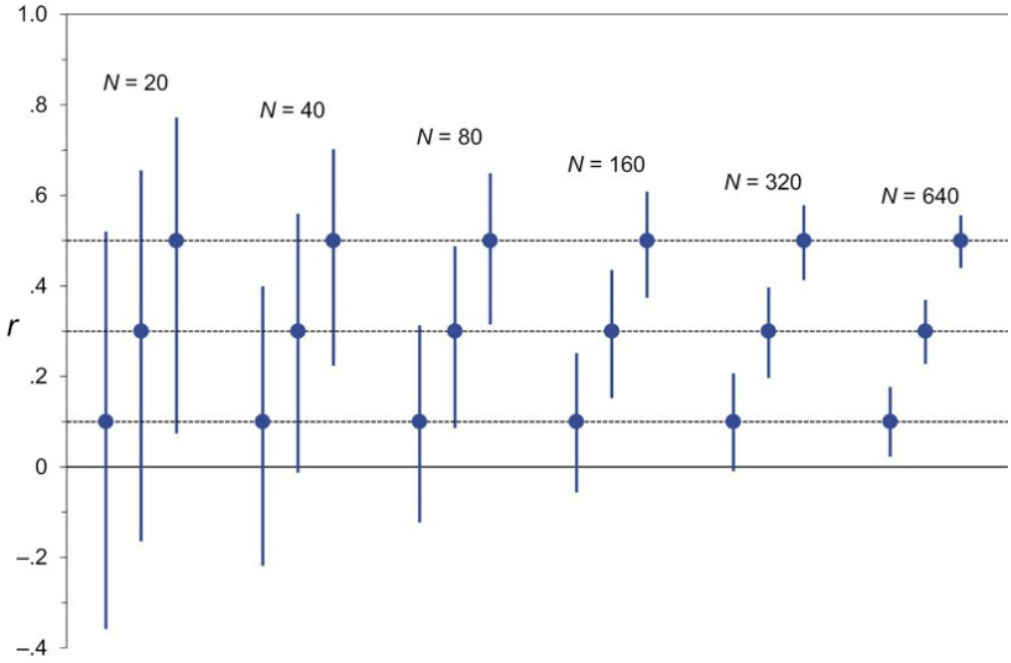


Figure 11.19. Examples of 95% CIs for r = .1, .3, and .5, respectively, left to right, for each of various values of N, as indicated at the top.

These figures should also be used to think very hard about comparing correlations. It frequently happens that there is a series of correlations and the researcher is tempted to say that one correlation is stronger than another.

- For N = 40, comparing $r = 0.3$ and $r = 0.5$, look at the overlap of the error bars. It is over 50%. Therefore it would be very unreasonable to argue that $r = 0.5$ is stronger than $r = 0.3$. To be confidence of a genuine difference you would need a much larger sample size.
- In figure 8.10 only a correlation as high as 0.9 can with reasonable confidence be taken as stronger than a correlation of 0.5.
- For n = 40, a correlation of 0.3 could be zero
- For n = 20, a correlation of 0.5 could be as low as 0.1, and a correlation of + 0.3 could in fact be negative.

Concluding wise words from CCJ

To interpret a value of r, consider also the CI, and any correlations reported by related past research. Also have in mind scatterplots. I am always struck by how widely scattered the points are, even for r as large as 0.6. It's sobering to learn that many researchers are studying relationships between variables that have small values of *r* **with scatterplots that look like shotgun blasts**. Such relationships may be interesting and important—or they might not be—but, either way, it's still a shotgun blast.

Loess lines

12. The problem of non-linearity

It is all too easy to assume that an association is linear. Sometimes just eyeballing the data is enough. Then for instance, knowing the context, the data can be divided into two or more sections, each directly describing part of the phenomena being studied.

But sometimes it is not obvious. So a line called a Loess line, also known as Cleveland's smoother, can be drawn by the computer.

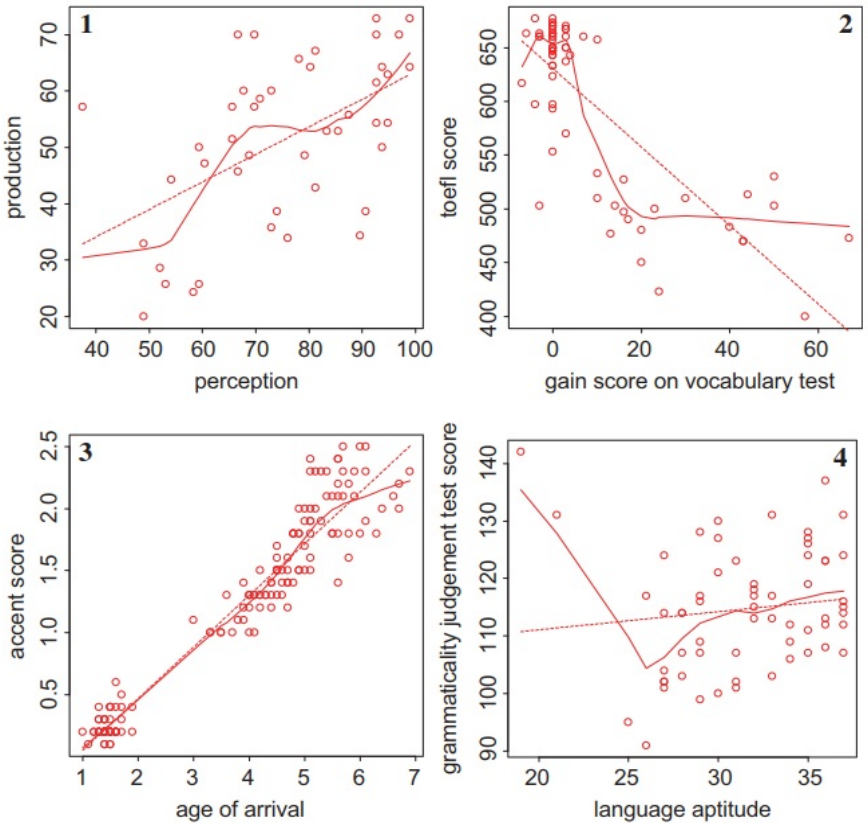
What the Loess line does is to draw the regression line based on a small part of the data, then repeat it several times.

The Loess line is shown on top of the scattergram with regression line. In general, the closer the two lines are, the more likely it is that the data is linear.

In figure 8.13 the Loess lines in graphs one and three are close enough to be considered linear. Graph two shows two distinct groups, which should be analysed separately. In graph four, two outliers (extremes) at the far left of the graph have skewed the regression line, and in effect made it flat. There is clearly a sharper angle in the non-outlier data.

Figure 8.13 Four scatterplots with superimposed regression (dotted) and Loess lines (solid)

From: Larson-Hall J & Herrington R 2009. Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics* 31/3:368-390.



- Plots 1 and 3 are linear
- Plot 2 could be mistaken as linear. But when Loess lines are drawn then a curve appears. The shape of the curve suggests two distinct lines, therefore two distinct stages which should be considered separately.
- Plot 4 again, in the traditional plotting, produces a straight line, and this line is almost flat. But, the reason for this flatness is the widely differing points to the left. These points are known as 'outliers' ie they are on the extremes. It is these outliers that have modified the line. The Loess line shows this, and enables more interesting data interpretation.

13. Example: Morphological awareness and reading comprehension

a. See Zhang D & Koda K 2013. Morphological awareness and reading comprehension in a foreign language: a study of young Chines EFL learners. *System* 41:901-913
They present the following correlations e, and draw the following conclusions.

Figure 8.14 Correlations

	1	2	3	4	5	6	7	8
1	–							
2	.193**	–						
3	.237***	.277***	–					
4	.139*	.173**	.290***	–				
5	.162*	.059	.194**	.150*	–			
6	.092	.345***	.243***	.128*	.206***	–		
7	.034	.204***	.225***	.179**	.175**	.176**	–	
8	.272***	.394***	.407***	.263***	.231***	.349***	.154*	–
9	.171**	.238***	.358***	.297***	.265***	.249***	.262***	.431***

1	Nonverbal intelligence	–
2	Morphological relation (inflection)	.193**
3	Morphological relation (derivation)	.237***
4	Affix choice	.139*
5	Compound structure	.162*
6	Morpheme discrimination	.092
7	Grammatical knowledge	.034
8	Vocabulary knowledge	.272***
9	Reading comprehension	.171**

The statistical significance is indicated by

$*p < .01$ $**p < .05$ $***p < .001$.

Statistical significance, the p-values, will be dealt with later. Right now, you need to know that they are routinely misunderstood. They are used to assess how good a correlation is. We know that a correlation is assessed by its margin of error: in other words, its CI.

Note, there was no mention of CI, and ALL the 95% CIs should have been presented.

b. Here are the conclusions drawn by the authors.

> ... all morphological awareness measures [2-6] correlated significantly with grammatical knowledge [7], vocabulary knowledge [8], and reading comprehension [9]; and the correlations between the morphological awareness measures themselves were almost all significant. Grammatical knowledge, vocabulary knowledge, and reading comprehension significantly correlated with each other, too. Overall, the correlations suggest a close relationship between different types/facets of morphological awareness and reading comprehension. (p908-909)

c. Student action

From your knowledge in this chapter, by now you will be able to assess these statements. Do so now, before reading my comments.

d. Now, go over to http://www.vassarstats.net/rho.html and type in some data. Elsewhere in the article, n=245 ie their sample was 245.

Take the statement that grammatical knowledge (item 7) correlated with items 2-6, which are all measures of morphology.

Figure 8.15 Grammar knowledge and:

item	r	Lower CI	Upper CI
2	.204	.081	.321
3	.225	.103	.340
4	.179	.055	.297
5	.175	.051	.293
6	.176	.052	.294

Note that the strongest correlation, for item 3, is still only 0.225 and the best value, given 95% CI is 0.340.

Figure 8.16 Vocabulary knowledge and:

item	r	Lower CI	Upper CI
2	.394	.283	.494
3	.407	.297	.506
4	.263	.143	.297
5	.231	.109	.346
6	.349	.234	.485

e. Interpretation commentary

... all morphological awareness measures [2-6] correlated significantly with grammatical knowledge [7], vocabulary knowledge [8], and reading comprehension [9]; and the correlations between the morphological awareness measures themselves were almost all significant. Grammatical knowledge, vocabulary knowledge, and reading comprehension significantly correlated with each other, too. Overall, the correlations suggest a close relationship between different types/facets of morphological awareness and reading comprehension. (p908-909)	This is misleading. Strength of relationship, and statistical significance of the data, must not be confused. Even using 'old' statistics, the correlations are 'weak' indicating 'weak relationship' (see figure 8.6). 'Significantly correlated' implies a real and strong correlation. They do not exist, especially for grammar and morphology. Even for vocabulary and morphology, and taking into account the upper CI, there are only a few which could possibly be moderate relationships.

CHAPTER 9
SIGNIFICANCE

1. Introduction

This chapter explains the whole question of significance testing in statistics. There is a terminology to be learned. There is also the fact that many statistics books are inaccurate, and the concepts are slippery so that it is possible for an author to slip into bad ways of thinking.

I will present the correct view. In later chapters I will present the older viewpoints, and attempt to immunise the reader against them.

Recall that in Chapter 8, some data from morphology was presented. The author had assessed the correlations using statistical significance. For this they used the *p*-value. They had used it wrongly. They had used it as a means of assessing the correlations. This is incorrect. The *p*-value is a means of assessing the quality of the data. Read on.

For many years people have been protesting about significance testing. Note, they are not against the use of statistics, and not against the use of most statistical tests. They are against the use of the *p*-value. In their place, there should be confidence intervals, effect size, and so called "point estimates".

Confidence intervals are self explanatory. Most statistics programs will provide them wherever possible. They are some indication of a range, which has already been dealt with. **Point**

estimates are single numbers, such as the mean, or a correlation.

In a nutshell, in the New Statistics, *p*-values are totally abandoned. Concepts like Power, Effect Size, and Confidence Intervals, are very important.

It should be reassuring to know that APA style for papers no longer requires significance tests. Also, many famous names never used them. Names such as Ebbinghaus (famous for his research on forgetting, and his forgetting curve), Skinner famous for behaviourism, and Piaget famous for work on child development. (Cummings 2012 Ch15).

Language point

p sometimes written *P* (note the italics) is the convention used to express the level of trust in the data. In this case *p* refers to the probability that there is a false result.
$p < 0.05$ means the level of distrust is less than 0.05.

The language used here comes from basic probability theory. For instance, take any coin. A coin has two sides, known in English as the head and the tail. When a coin is thrown in the air what is the probability that when it falls you will see the side known as the head? The answer is $p = 0.5$. The probability that you will see a tail is also $p = 0.5$

Large and small numbers

Some students find it difficult to convert between decimals and fractions. But the effort is usually worthwhile because fractions seem to be easier to understand. The following table may help you.

$p = 0.5$	$p = ½$	large figure
$p = 0.05$	$p = 1/20$	
$p = 0.01$	$p = 1/100$	
$p = 0.005$	$p = 1/200$	
$p = 0.001$	$p = 1/1000$	small figure

Some students do not instantly see that 0.001 is a small figure and 0.5 is a big figure. I suggest if this is a problem that you look at the figures, play with them and similar figures until you are sure.

2. Setting significant

Three common cut-off points that are used are:

1. 95% ie a *p*-value of 0.05
2. 99% ie a *p*-value of 0.01
3. 99.9% ie a *p*-value of 0.001

Some argue that in social sciences, this should be set at 90% ie a *p*-value of 0.10.

Basically, when the *p*-value is small, eg 0.001, the results are good. The values are used in the null-hypothesis reasoning, explained in the next chapter. They are used to see if you can reject the null hypothesis that there is no change, and accept the **alternative hypothesis** that there is a change.

3. What does the *p*-value really mean?

This is a very good question. Readers should know that they are now reading about a hot potato in statistics. They should be ready to change their minds, and to do some work to understand what is going on.

I said that it was a hot potato. This means that there is more than one viewpoint, and only one viewpoint is right. In addition, many traditional authors, frankly, got it wrong. Despite constant complaints in textbooks and journals, even top medical and science journals continue to publish work that at best is incomplete or ambiguous, and at worst dangerously misleading and wrong. There are still writers and teachers in recent years who are perpetuating the old system.

For the record, I have consulted several sources, including Larson-Hall (2010), Ellis (2010), Cumming (2012) Kline (2004) and Cumming (2018). Note, I was obliged to purchase a Kindle version of this book, therefore page references cannot be supplied and I will therefore indicate chapter numbers. I have found Larson-Hall (2010) and Reinhart (2014) to be the clearest and even then I had to struggle, and to consult several sections of their books. Ellis (2010) is thorough, and Schmidt & Hunter (1997) are clear.

4. Modern definitions

The p-value is a way of evaluating the quality of the data. The higher the level (ie the lower the decimal) then the greater the confidence you can place in your data.

5. The following are true:

a. The *p*-value is the conditional probability of the *data* (ie the test statistic value that you calculate) given the null hypothesis.

b. If we were to repeat the experiment many times, and if it were true that there was *no difference* between the two groups or *NO relationship* between the two factors, then what is the probability that we would get this set of data? For a *p*-value of 0.05 that probability is 5%. Notice the negative reasoning.

c. It is the probability of the data given the null hypothesis.

"The ***p*-value** is the probability that we would find a statistic as large as the one we found if the null hypothesis were true" (Larson-Hall 2010:49).

a It refers to the probability you would get exactly these results/data given the hypothesis.

b. This is written in formal terms as $p(D \mid H_0)$. In words, this is **the probability of the Data given the null Hypothesis.** NB, it is NOT the other way round! One of the major problems with *p*-values is that they must NOT be used to assess the hypotheses.

***P*-values must NOT be used to assess hypotheses**

c. Or, you could read the statement from right to left! Use your Arabic background skills to follow the logic! If we start with the null hypothesis and get some data, what is the probability that you would get this data?

d. Larson-Hall (2010:49) proposes that you memorise this phrase in order to understand *p*-values the correct way:
The probability of finding a [insert statistic name here] this large or larger if the null hypothesis were true is [insert *p*-value].

6. **The following are false:**
   a. "...the *p*-value is the probability that the null hypothesis is true" (Larson-Hall 2010:99).
   b. With a *p*-value of 0.05 there is a 95% certainty that a difference exists and that it cannot be due to chance.
   c. The p-value is the probability that the results are due to chance. (Cumming 2012 Ch2)

7. **All this means that:**
   a. You cannot say anything about the probability that the hypothesis is true or not.
   b. For assessing hypotheses, Confidence intervals are more informative. These are dealt with later.

8. **The *p*-value does NOT indicate the importance or size of a difference or relation**

   These factors must be decided by other reasoning. Unfortunately, statistics cannot answer these questions.

Question:
Is a result with a significance of $p = 0.001$ stronger than a result with a significance of $p = 0.05$?
Information:
Remember that a *p* value of 0.05 expresses MORE doubt than a *p* value of 0.001.
Answer:
No. *P* values are estimates of the confidence we can place in the data. They have nothing to do with deciding if the hypothesis is a good one or not.

It would be easy to conclude that a result with a significance of 0.001 is stronger than one at the 0.05 level. This is a false conclusion. One major reason for it being false is the well known fact that the larger the sample the higher the significance level. With a large enough sample almost any association in a sample will be statistically significant, because as size increases, random effects are likely to cancel out, and even weak associations will surface. (Lempert 2009)

For an extremely readable and informative introduction, see the article by Gigerenzer (2004) called "Mindless statistics".

9. Key reasons to abandon significance testing

a. They are not used in physical sciences, the so called 'hard' sciences

In the hard sciences, numbers are obtained, and an estimate of the error is provided. Most of the time, results of three significant figures are acceptable. In some cases, results of the same order of magnitude to the real figure are acceptable. In the hard sciences, pseudo-precision is avoided. (Schmidt & Hunter 1997:7)

b. Significance tests are logically indefensible and are NOT needed (Schmidt & Hunter 1997:2ff)

It would be great to have a simple procedure to decide if the data is real, or just due to chance. A common misunderstanding is that null hypothesis significance testing can do this job. Unfortunately, no known technique can do that.

Power:

Is the study big enough to detect a difference that is real?

Does the study include enough people/observations or measurements to fairly reject the null hypothesis?

Power is all about having a big enough study to mean confident conclusions.

c. With a power of 0.5, half the tests will be non-significant

It is well recognised that the average power of null hypothesis significance tests is between 0.4 and 0.6. With a power of 0.5, half the tests will be non-significant. This means that real associations will be rejected 50% of the time, and false associations will be accepted 50% of the time.

d. Significance tests work against replication research

1) The majority of published work in the humanities lacks power. This means, **the sampling and the methods are too small to reliably detect the effect.** If you like, **the background noise drowns the speech**. Schmidt & Hunter (1997) and others have shown that the power of most research is usually around 0.5, which means that half the time the effect is incorrectly identified, and half the time the effect is incorrectly rejected.

2) Replication does not help. Supposing you have a series of two successful experiments, each with a power of 0.5. The probability of two positive experiments is (0.5)(0.5) = 0.25. The error rate is now 75% instead of 50%. If we have a series of 5 experiments that is ½ x ½ x ½ x ½ x ½ =1/32 which is roughly 3%. Therefore, a possible error rate of 97% exists.

3) For replication to work, then non-significant findings must also be replicated, with similar problems with a series of replications.

4) Acceptance of a Research Article is biased in favour of results which are 'significant'. Negative results, or inconclusive results, are less often reported. This of course is a big scandal in medicine because drug companies can conveniently not publish inconclusive research, and so leave the field open to only the positive results. If you are not sure of this then google "Ben Goldacre" and "Bad Pharma". (See Goldacre 2012). Goldacre is a doctor who campaigns against flawed clinical trials, suppression of unfavourable results etc. Even when there is good will and honesty, there is a natural human tendency to avoid publishing negative findings.

10. What should replace significance testing?

There must be much more focus on the numbers, for correlations, or for the ranges. There must also be a statement of the margin of error – the so called Confidence Interval. You can also consider power, and effect size.

Point estimates provide an estimation of the size of the effect or relation, so there is little indication of whether the effect is small or large. Confidence intervals provide a measure of the uncertainty, which is often quite large.

Point Estimates

You will see this term used often in the New Statistics. It is simply a group word for simply calculated figures such as:

- The mean
- The median
- The mode
- The Standard Deviation

11. How do I cope writing a thesis, with examiners and supervisors who do not understand?

Welcome to the real world of research, which is all about relationships. Now is the time to develop some tact. Let me ask you a similar question to which you already know the answer.

a. **Question.** What are the principles for using a quotation, rather than paraphrasing?

   You will recall that there are reasons for a quotation:
   - The actual words as well as the idea are important
   - The idea is unusual or surprising

b. When you have an idea that you, being reader centred in your writing, are confident that it is unusual to your

predicted reader, then you need to give extra information and explanation. Support can be provided in three ways:

- Appeal to authority
- Providing evidence
- Providing careful and strong reasoning

c. Given that many thesis examiners are weak on statistics, and if they are knowledgeable, they are probably out of date, therefore, modern significance tests need explaining. The methodology chapter is an excellent place for this.

CHAPTER 10
THE old STATISTICS

1. Introduction

The old statistics is based on two foundations:

- The disproval of the Null Hypothesis
- The Base Rate Fallacy

So to these we shall now turn.

A. The Null Hypothesis

1. The classical version of the null hypothesis

Using the null hypothesis reasoning, we start out by assuming that there is no real difference, ie any differences seen are due to chance. Then we try and disprove it.

The null hypothesis refers to a general or default position: that there is no relationship between two measured phenomena, or that a potential medical treatment has no effect. Rejecting or disproving the null hypothesis – and thus concluding that there are grounds for believing that there is a relationship between two phenomena or that a potential treatment has a measurable effect – is a central task in the modern practice of science, and gives a precise sense in which a claim is capable of being proven false.

A null hypothesis is contrasted with an alternative hypothesis, and these are decided between on the basis of data, with certain error rates.

2. **The null hypothesis and the alternative hypothesis**

The traditional way tries to use *p*-values to decide between different hypotheses.

Example 10:1

Let us assume you do an experiment with the opinions people have over the taste of diet coke and coke. The experiment has been carried out in an attempt to disprove or reject a particular hypothesis, the null hypothesis. We can write:

H_0 there is no difference in taste between coke and diet coke.
H_1 there is a difference.

Example 10:2

In a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write:

H_0 there is no difference between the two drugs on average.
H_1 there is a difference.

We give special consideration to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if and when the null is rejected.

The alternative hypothesis, H_1, is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write

H_1: the two drugs have different effects, on average.

The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write

H_1: the new drug is better than the current drug, on average.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "Reject H0 in favour of H_1" or "Do not reject H0.
We never conclude "Reject H_1", or even "Accept H_1".

NB. If we conclude "Do not reject H_0", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H0 in favour of H_1. Rejecting the null hypothesis then, only suggests that the alternative hypothesis may be true.

This is important. Rejecting the null hypothesis only suggests that the alternative hypothesis may be true.

Rejecting the null hypothesis does NOT mean you can accept the alternative hypothesis.

The values are used in the null-hypothesis reasoning. Tests are more a acceptable when The bigger the difference, the more likely we can reject the null hypothesis that there is no change, ie in rejecting this, we are saying we accept the alternative hypothesis that there is a change.

3. Time to reject the null ritual

a. The null ritual summarised

1) Set up a null hypothesis of no difference, or no correlation. Never specify the alternative hypotheses
2) Use 5% as a convention for rejecting the null. If significant, then present your result as $p < 0.05$, or $p < 0.01$, or $p < 0.001$.
3) Always perform this procedure.

b. Steps advised by Fisher

Fisher is the man who was a leading figure in the thinking behind testing the null hypothesis. He as often referred to, and often appealed to. Yet, Fisher rejected the steps above and proposed something more sophisticated. Fisher is usually blamed for the null ritual. But, towards the end of his life Fisher (1955, 1956) rejected each step.

1) "Null" does not mean no effect at all, or zero correlation. An improvement of vocabulary from 1000 to 1500 words, or a correlation of 0.5, could be a null hypothesis.
2) It is mindless to be fixated on 5% significance level, since this, one shoe fits all, totally neglects the very varied experimental situations. Therefore state the exact level of significance.
3) The primitive version should only be used for cases where we have very little knowledge.

c. **Neyman and Pearson reject these three steps**

These two very important statisticians rejected the three steps of the null ritual for different reasons. They favoured competitive testing between two or more hypotheses. The Type two error is also important (ie false negatives) which is ignored in the null ritual.

4. **Neyman-Pearson decision theory**
   a. **The theory is:**
      1) Set up two statistical hypotheses H1 and H2, decide about alpha, beta, and sample size before the experiment. These define a rejection region for each hypothesis.
      2) If the data leads to a rejection of H1 then accept H2, otherwise accept H1. NB accepting a hypothesis does NOT mean it is true. It means that you now act as if it is true.

   b. **Example: quality control**
      As Gigerenzer (2004:591) explains, this works well with quality control in a factory, when for instance regular samples from the production line are tested. At a certain point, the production line is stopped. This does not mean the controller is sure there are many mistakes, only that there is enough evidence to double check.

   c. The problem is, while this works well for quality control, this is not always the case in research.

5. **Testing Meehl's conjecture**
   All these arguments about half of the research being wrong, has been formally tested. The arguments can be summarised as: "In nonexperimental settings with large sample sizes, the probability of rejecting the null hypothesis of nil differences in favor of a directional alternative is 0.5" (Gigerenzer 2004:601).

   There is a way of testing this. Waller (2004) had access to the data of 80,000 people who had completed a 567 item questionnaire known as the Minnesota Multiphase Personality Inventory. He used computer simulation methods to test some linkages, determined randomly the direction of the alternative hypothesis and computed the significance levels. In all, he had 511 predictions, and 46% of them were confirmed, some of them with very impressive *p*-values.

The point is this. A large sample led to 46% of statistically significant results.

There is little more to be said. A strong case has been made to reject the old statistics. Many experts have made the case better than I have in this book, though, I hope, my explanations here are simpler than the others.

6. If we reject significance testing, what remains?

Quite a lot. The hard sciences manage quite well without it. That should in itself merit major consideration. So, what is left?

- Descriptive statistics. Averages. Shapes. Noting exceptions, nothing the unusual
- Strength of correlations
- Confidence intervals

These alone are worth a lot. In addition, we will go on to study:

- Power
- Effect Size

7. An important digression

Before you go any further, make sure you are comfortable with the question below. From now on, fluency in moving between ratios, percentages, and actual figures, will be assumed. Be assured that the material is extremely easy, but it is one more example of an area where some people have major gaps in their lower high school education.

Question

Ratios, decimals, fractions, and percentages

Fractions, decimals, and percentages, can easily be interchanged. It is assumed in this book that students can easily convert between them, at high speed.

Question: which statement is true?

A. 20/100 = 20% = 0.2 = 20:100 = 2:10 = 1:5

B. 20/100 = 20% = 0.2 = 20:80 = 2:8 = 1:4

Answer

The problem comes with a ratio. When you cut a cake into two equal pieces, the ratio is 1:1, NOT 1:2, therefore B is the correct answer.

This small point is often forgotten, and leads to confusion. The distinction matters when small numbers are involved. In practice, because of rounding, the difference between 1:99 and 1:100 is usually insignificant.

B. The Base Rate = Real World Rate

1. Introduction to the base rate – when false positives usually outnumber true positives

Imagine the case where a blood test has a false positive rate of 10% and a false negative rate of 10%. There are 100 people with the disease and 900 who are without disease. The easiest way to visualise this is to imagine this is to think of a big box of coloured balls: 100 are red (danger) and 900 are green (safe).

Take the red balls representing those who have the disease. Ten will falsely be identified as green leaving only 90 to be correctly identified as red. For the green balls, 90 will be falsely identified as red, and 810 identified as green. This means that the real world false positives rate is 50% not 10% and the real world false negatives is 1.2% not 10%. This discrepancy takes a bit of getting used to at first, so I suggest you study Figure 10:1 carefully.

The mixture in the true population is also called the 'base rate'. Reinhart (2014) has an interesting chapter on it and leads to a discussion of the base rate fallacy. The base rate fallacy shows us that false positives are much more likely than you would expect from a $p < 0.05$ criterion.

Figure 10:1 Base rate 1:9

True population	**Results** ie identifications	
	red	**green**
100 red **False -ve rate = 10%**	90	10 [false]
900 green **False +ve rate = 10%**	90 [false]	810
Real world rates	90/180 = 50% false positives	10/820 = 1.2% false negatives

Note. For students who have problems converting between numbers, percentages, and ratios, the ratio 1:9 is the equivalent of saying 10:90 or 100:900. or 10% versus 90%.

In this case, 180 reds are identified, half of them incorrectly, and 10 reds are not identified.

This is how the calculations work.

There exist 100 reds. But only 90 will be detected. The other 10 will (falsely) be identified as greens.

There exist 900 greens. But only 810 (ie 90% of 900) are identified. The other 90 are, (incorrectly) identified as reds.

Therefore, 90 reds are correctly identified, and there are another 90 false-reds. This totals 180 reds identified, and 90 of those 180 are falsely identified, therefore the true false positive rate is NOT 10%. The true false positive rate is 90/180 = 50%.

Notice how, when the false positives is 10% and the false negatives is 10% for a population where only 10% (100 out of 1000) have the problem, then the effective false positive rate is 50%. No one in research would accept a $p = 0.5$ level as acceptable for published research.

Now, let us change one of the figures and see how this changes the scenario.

Imagine the case where the test has a false positive rate of 10% and a false negative rate of 20%. There are 100 people with the disease and 900 who are without disease.

Figure 10:2 Doubling the false negatives rate, base rate = 1:9

True population	Results	
	red	**green**
100 red **false -ve rate = 20%**	80	20 [false]
900 green **false +ve rate = 10%**	90 [false]	810
Real world rates	90/170 = 53% false positives, NOT 10%!!	20/830 = 2.4% false negatives, NOT 20%!!

In this case, doubling the false negative rate has increased the false positives.

Imagine the case where the test has a false positive rate of 10% and a false negative rate of 5%. There are 100 people with the disease and 900 who are without disease.

Figure 10:3 Interaction of false positives and false negatives. Base rate = 1:9

True population	Results	
	red ie positive	**green ie negative**
100 red **False -ve rate = 10%**	90	10
900 green **False +ve rate = 20%**	45	855
Real world rates	45/135 = 33% false positives, NOT 10%!!	10/855 = 1.17% false negatives, NOT 20%!!

In this case, doubling the false negative rate has significantly decreased the false positives, but not in a simple way. Note well, the real world false positives has gone down! It has also only slightly decreased the real world false negatives.

6. Changing the base line

In order to get a feel for the way things work, I have set out a series of tables for when the original base line changes. To make things easier I have also chosen a **false positive of 10% and a false negative of 10%.**

Figure 10:4 Base rate of 1% ie 1:99
NB. This base rate, or smaller, is more realistic than those considered above.

True population	**Results**	
	red ie positive	**green ie negative**
10 red **False -ve rate = 10%**	9	1 [false]
990 green **False +ve rate = 10%**	99 [false]	891
Real world detection rate	9/108 = 8.3%	1/892 = 0.11%
Real world rates	99/108 = 91.7% false positives, NOT 10%!!	891/892 = real world false negatives 99.89% false negatives, NOT 10%!!

The table above is equivalent to discussing a disease which exists at 1% in the population. It is diseases like cancers, and screening tests for cancers, that attempt to detect cancer at this level or smaller. In this case only 8.3% of those who have positive tests will actually have cancer.

Blastland & Spiegelhalter (2013:262) report that mammography tests are quite good compared with other screening tests. The true incidence in the population is around

1% and false positives are 10%. This means that over 9/10 women diagnosed with cancer are falsely diagnosed, with all the attendant risks of surgery, radiation, and medicines.

Figure 10:5 Base rate of 5% ie 1:19

True population	**Results**	
	red ie positive	**green ie negative**
10 red **False -ve rate = 10%**	45	5
950 green **False +ve rate = 10%**	95	855
Real world detection rate	45/140 = 32.1% false positives, NOT 10%!!	5/860 =0.6% false negatives, NOT 10%!!

This means that the real world false positives is 67.9%

Figure 10:6. Base rate of 10% ie 1:9

True population	**Results**	
	red ie positive	**green ie negative**
100 red **False -ve rate - 10%**	90	10
900 green **False +ve rate = 10%**	90	810
Real world detection rate	90/180=50% false positives, NOT 10%!!	10/820 = 1.22% false negatives, NOT 10%!!

This means that the real world false positives is 50%

Figure 10:7. Base rate of 20% ie 1:4

True population	Results	
	red ie positive	**green ie negative**
200 red **False -ve rate = 10%**	180	20
800 green **False +ve rate = 10%**	80	720
Real world detection rate	180/260 = 69.2% false positives, NOT 10%!!	20/740 = 2.78% false negatives, NOT 10%!!

This means that the real world false positives is 30.8%

Figure 10:8. Base rate of 50% ie 1:1

True population	Results	
	red ie positive	**green ie negative**
500 red **False -ve rate = 10%**	450	50
500 green **False +ve rate = 10%**	50	450
Real world detection rate	450/500 = 90% false positives, NOT 10%!!	50/500 = 10% Correct false positives of 10%

This means that the real world false positives is 10% and is a very interesting scenario which you might want to come back to once you have learned about 'effect size' below. Only when the population is half-half do you get a low false positive. **When the base rate is small, then false positives are highly likely.**

Figure 10:9. Base rate of 80% ie 4:1

True population	Results	
	red ie positive	**green ie negative**
800 red **False -ve rate = 10%**	720	80
200 green **False +ve rate = 10%**	20	180
Real world detection rate	720/740 = 2.7% false positives, NOT 10%!!	80/260 = 30.8% false negatives, NOT 10%!!

This means that the real world false positives is 97.3% and is similar to when the 'power' is set at 0.80.

Figure 10:10. Summary Base rate table

false -ve and false +ve rates %	**Proportions of red/green detected**					
	10/990	50/950	100/900	200/800	500/500	800/200
false positives	92%	68%	50%	31%	10%	2.7%
false negatives	0.1%	0.58%	1.2%	2.7%	10%	31%

∑ Figure 10:10 clearly shows that when an event is rare, such as 10/990, the real world false positives are high(and false negatives are low). It is only when an event exists around 50% of the time that the base rate is close to the quoted false positives and false negatives rates.

7. General and special probability revisited

Remember Key 19. Chance is everywhere. Coincidence is more likely than you think. Often there is a change simply due to the mathematics, or due to the way several factors interact.

When the letters of the month are arranged in a line there is a word: JFMAMJJASOND and there is the word JASON in the letters. This is not at all surprising because the prediction was

too vague: predicting that you would find one recognisable word in these twelve letters is banal. The outcome would only be spectacular if someone had predicted that this word JASON would be in the letters before they looked.

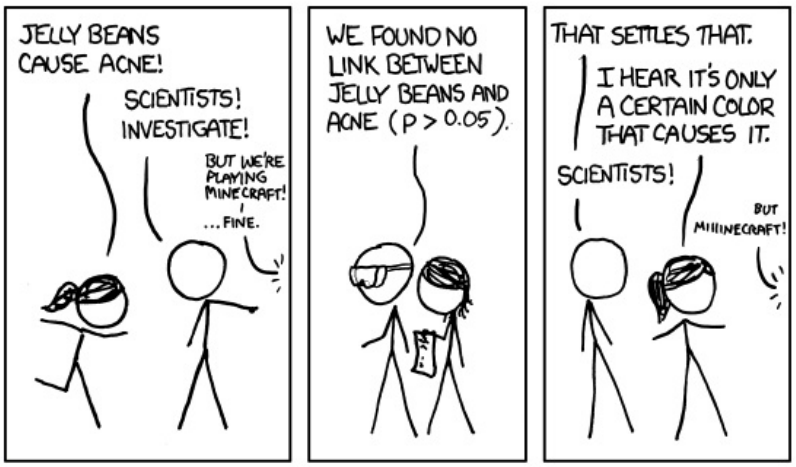
Yet many researchers fall into the same trap – of looking at the results THEN making their predictions.

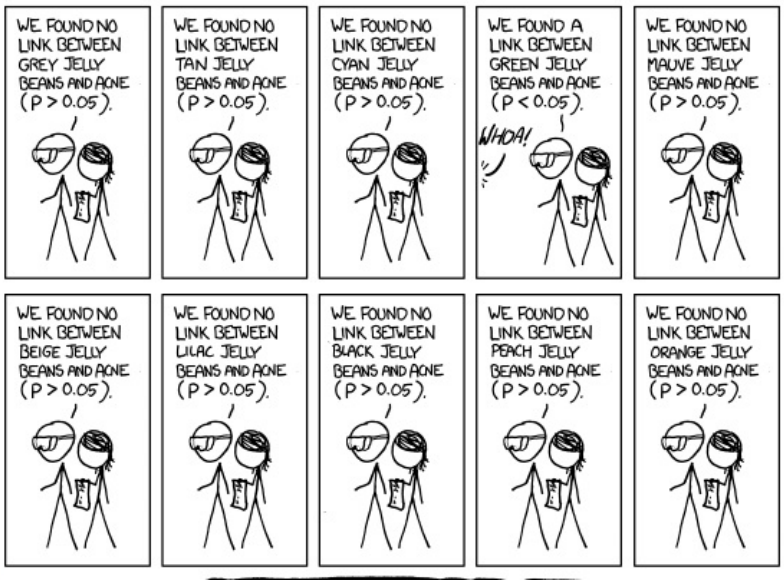
8. If at first you don't succeed, try, try again
(Slightly edited version of Reinhart (2014)
OR: Cherry picking gone mad

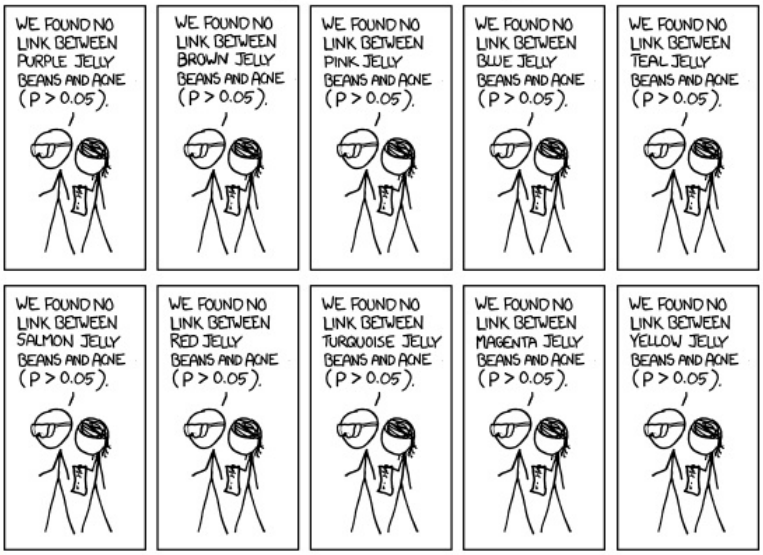
The base rate fallacy shows us that false positives are much more likely than you'd expect from a $p < 0.05$ criterion for significance. So, when something only exists 1/20 times, false positives are common.

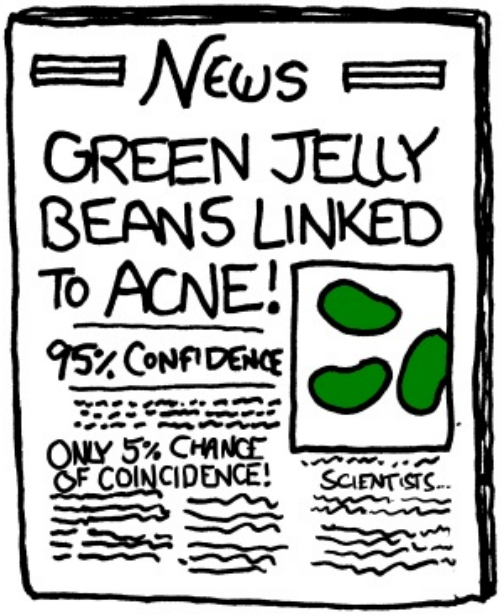
Most modern research does not make one significance test, however; modern studies compare the effects of a variety of factors, seeking to find those with the most significant effects.

For example, imagine testing whether jelly beans cause acne by testing the effect of every single jelly bean colour on acne:




WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).



Cartoon from xkcd, by Randall Munroe. http://xkcd.com/882/ (Author gave permission for non commercial use and use in a book, provided citation is provided).

As you can see, making multiple comparisons means multiple chances for a false positive. For example, if I test 20 jelly bean flavours which do not cause acne at all, and look for a correlation at $p < 0.05$ significance, I have a 68% chance of a false positive result. If I test 45 materials, the chance of false positive is as high as 90%.

It is easy to make multiple comparisons, and it does not have to be as obvious as testing twenty potential medicines. Track the symptoms of a dozen patients for a dozen weeks and test for significant benefits during any of those weeks: bam, that's twelve comparisons. Check for the occurrence of twenty-three potential dangerous side effects: alas, you have sinned. Send out a ten-page survey asking about nuclear power plant

proximity, milk consumption, age, number of male cousins, favourite pizza topping, current sock colour, and a few dozen other factors for good measure, and you'll find that something causes cancer. Ask enough questions and it's inevitable.

A survey of medical trials in the 1980s found that the average trial made 30 therapeutic comparisons. In more than half of the trials, the researchers had made so many comparisons that a false positive was highly likely, and the statistically significant results they did report were cast into doubt: they may have found a statistically significant effect, but it could just have easily been a false positive.

There exist techniques to correct for multiple comparisons. For example, the Bonferroni correction method says that if you make n comparisons in the trial, your criterion for significance should be $p < 0.05/n$. This lowers the chances of a false positive to what you'd see from making only one comparison at $p < 0.05$. However, as you can imagine, this reduces statistical power, since you're demanding much stronger correlations before you conclude they are statistically significant. It's a difficult tradeoff, and tragically few papers even consider it.

9. Another viewpoint about mindless correlations

Cumming (2012 Ch15) argues that we should not be too quick to dismiss these correlations. I think he is arguing that interesting correlations should not be taken as definitive or confirmed. Instead, the correlations can serve as a basis for clear hypotheses, and further experimental, data driven research. But, on their own, they are suspect: they must NEVER be taken as definitive. In this I think he is right, and I suspect Reinhart would agree with this clarification.

In the case of multiple correlations, these are NOT results. At best they serve as hypotheses for further investigation.

10. The statistical toolbox

Gingerenzer (2004) argues that instead of significance testing, textbooks of statistics should concentrate on teaching a variety of tools. He calls these the statistical toolbox. These include:

- Descriptive statistics
- Tukey's exploratory methods
- Bayesian statistics
- Neyman-Pearson decision theory
- Wald's sequential analysis.

Real statistical thinking means the art of choosing a good tool for a given problem. Fortunately, this book will continue to focus on the basics.

CHAPTER 11
EFFECT SIZE: AN ALTERNATIVE TO THE t-TEST

1. **What is 'significant'**?
   Traditionally, differences between groups have been assessed using the t-test. The aim is to find a way to be confident that two groups are different.

   But they have failed. Significance tests say nothing about whether or not the effect is big enough to mean something.

   You have to be careful about the concept of real world significance. Even in everyday life this is true. You may have 1000 dinars in a savings account earning 5% interest per year, which means 50 dinars per year. In another bank, the interest rate may be 6% which makes 60 dinars per year. Is this difference significant? It certainly exists, but is the difference significant ie important? And that raises the question as to what you mean by important, and how do you measure it. If you have the good fortune to have 100,000 dinars in savings, then the difference between 5% and 6% is a whopping 1000 dinars. But if you have only 1000 dinars then a 1% difference in interest rates is small.

2. **High statistical significance can in fact be attained in two ways:**
   1. Small Sample + Large difference
   2. Large Sample + Small difference

   The trouble is that conventional statistics do not provide the tools and reasoning to deduce which effect is at work. It is

important to be able to objectively decide whether it is the sample size or the coefficient that is making the difference.

In the 1970s in Britain, when inflation was high, and strikes for extra pay were common, the government and employers would often use this idea of a 'significant difference'. Is it better to give everyone a 5% payrise, or a fixed amount for everyone regardless of their basic pay? If the lowest worker is earning 500 per month and the highest paid worker earning 5000. A 5% increase for the low paid worker is only 25 per month, and this works out as 250 per month for the highest paid. Therefore, in order to fairly help everyone, it would be better to give the workforce a 'fixed' rise of 50 per person per month. The sum of 50 works out to be 10% for the low paid and 1% for the high paid. The 50 extra is a meaningful extra for all.

It is a similar question in statistics. A small difference may exist, to a high degree of probability, say $p<0.001$. But the small (statistically significant) difference may be trivial in practice.

Context is vital. I present below an example of a real-world significant difference. The example shows the importance of context.

3. **Example 11:1 A real significant difference**

Statistics mirrors normal research. In my own doctoral work, I established the existence of some faux-amis for science in French and English. My supervisor had insisted that I repeatedly asked the question 'so what?' right throughout the thesis, from the planning, through initial enquiries, through hypothesis formulation, designing data collection, analysing data, and interpreting the findings. So, having established a list of real differences, I then collected data as to how significant the differences were.

This meant answering the question: what is a significant difference? The answer I chose was that a significant difference would hinder communication, and would lead to language errors. So I gave the Lycée students two types of question:

Q1. Which of the following faux amis cause you problems?

Almost all the students rated the words as 'no problem'.

Q2. I then tested these students. I grouped the words, and wrote 'fill in the gaps' exercises where students had to choose the correct word. I found that most of them were inaccurate half the time.

My conclusion was that the faux amis were a significant area where the students were overconfident and underskilled. The differences in fact had significance for advanced second language speakers of English.

As Cumming (2012) says, damning critiques of significance testing and its pernicious effects have been published over more than half a century and Kline (2004) provides an excellent review. For further good reviews of the problem, and the an introduction to the new statistics see Gigerenzer et al (2004).

Welcome a new player in the statistical field: Effect size.

4. **The crucial importance of effect size**
   Effect size is increasingly being seen by journals as more important than significance, to the point where some journals will not accept significance testing and instead insist on a statement and full discussion of effect size.

5. **Effect Size – simply stated**
   Effect Size is the size of the effect. Effect Size is a way of quantifying the difference between two groups: it tells the reader how big the effect is. For instance, in the standard two groups method where one has a control group and an experimental group, and the control group receives no treatment and the experimental group receives treatment. Providing the *p* value for the difference between the two groups tells you nothing about how large the difference is. Effect Size tells you the size of this difference.

   The Effect Size gives you an insight into the size of the difference.

   Effect Size also, does NOT change no matter how many participants there are (Larson-Hall 2010:114).

   Effect size is measured using means and standard deviations. There are several different formulae, each with their merits and disadvantages. See below.

6. **Three key questions**
   When comparing two groups, there are three key questions:
   - How big is the effect? This is the most important question to most readers. For instance, does a new treatment work, and how well does it work?
   - How precise is the estimate/measurement?
   - Two tailed or one tailed? Is there a directional relationship between the two groups (one tailed) or could the difference go either way (two tailed)?

7. Introduction to z-scores

Remember the graphs of standard deviation. An individual can be the average, the top, or the bottom, or somewhere in between. Sometimes I hear a teacher describe a pupil as 'below average'. But what does this mean? It means that their marks are less than the mean. But how much less? And is this a big gap or a small one?

Instead of saying that someone is for instance, two marks below the average, what if we could say how much below they are in terms of standard deviations? Well, we can.

I will assume you have two groups of data, for instance, two examinations, reading and writing. For writing, the examination was scored out of 30, and the second examination was scored out of 45.

Now, this is common enough in Britain, but in Tunisia might seem very strange. In Tunisia, tests are almost always arranged so that the score is out of 20. Sometimes a teacher might set a test out of 40, then halve the marks, but that is all.

Now, the question is, how can the marks be compared? The hypothesis is that someone who is good in writing will be good in reading.

The simplest way would be to convert all the marks to a percentage. And that solution is frequently used. However, there is a more sophisticated way.

In your spreadsheet of data, calculate the mean and standard deviation for each test. This will be done in two boxes per test.

Then, create another column for each test. Compute the z-score. This gives, for each mark, the deviation from the mean, as measured in the units: Standard Deviation for that test.

Now you can do the correlations, by inputting the z-scores. Does someone who is -1SD on writing correlate with -1SD mark on reading? If writing and reading are similar, and if the

tests are working well, then the correlation will be strong.

So, there is nothing mysterious about z-scores. They are a sophisticated form of percentage. They enable fair comparisons and correlations between groups of different size.

8. **Cohen's *d***
   As expected, there are various ways of comparing groups, in terms of the standard deviations. The most commonly used method is **Cohen's *d***, and the computers will calculate it for you.

9. **Online effect size calculators**
   - http://www.uccs.edu/lbecker/index.html
     This one is easy to use. The lecture notes are also interesting. All you need are the mean and standard deviation for each of the two groups.

   - http://www.cedu.niu.edu/~walker/calculators/effect.asp
     David Walker's Effect Size Calculator. Extremely easy to use if you know the mean, SD and sample size of each group.

   - http://www.latrobe.edu.au/psy/research/projects/esci
     This site has some interesting free software linked with Excel which does some simulations of the problem. The video demonstration is also free, fun, and informative. http://tinyurl.com/danceptrial2

- The book which helped me most to understand this question is Ellis (2010) which I found by googling the title, "The Essential guide to effect sizes" .

- http://www.cognitiveflexibility.org/effectsize/
  Nicholas Cepeda offers two ways of inputting data. Either using mean and SD, or using the *t* score. He also states a preference for using the average of each mean's individual SD, as opposed to pooled or control condition SD.

- http://davidmlane.com/hyperstat/effect_size.html
  This site has links to several calculators and excellent information and explanation.

- http://danielsoper.com/statcalc3/default.aspx
  Different to the David M Lane site above, this site provides a large number of free online statistics calculators, divided into 29 categories, including effect size. He offers advanced tests but do not be put off: the basic tests are here and are easy to use.

- www.clintools.com/victims/resources/software/effectsize/effect_size_generator.html

  ☺ **Recommended** ☺
  Effect Size Generator is able to compute effect size estimates for use in Meta-analyses. It will compute the Cohen's *d* effect size estimate, apply Hedges Adjustment for sample size (to Cohen's *d*) and also provide Hedges g effect size estimate. NB **It will also provide 95% confidence intervals for the derived effect sizes** and conduct a t-test on the data. It contains a full help file and is really quite self explanatory. This programme will also print out a report and save files! Note, it is a little hard to find, so you might want to google it. Version 2.3 is free and the name is esgfree2-3.exe. I have tried it, and it really is easy to use.

- www.cem.org/evidence-based-education/research-toolbox
  This is another and perhaps better place to find the Effect Size Generator. Documentation and explanation links are provided from this page. Click on the link to the Effect Size Calculator and you will get the following links to the xls and pdf instructions. This is a nice little Excel program, and on the site you will find some easy to understand information about effect size and what it means in practice.
  www.cem.org/evidence-based-education/effect-size-calculator
  www.cem.org/attachments/EBE/EffectSizeCalculator.xls
  www.cem.org/attachments/EBE/ESCalcGuide.pdf

- http://freewareapp.com/alphan_download/
  alphaN is a standalone Windows program that estimates the sample size needed for a specified coefficient alpha, given the Type I error rate and effect size.

10. Interpreting effect size differences

NB. Effect size numbers MUST MUST MUST be discussed separately from tests of statistical significance.

This is EXTREMELY important, and is the reason why I am daring to shout by using capital letters.

For *d* (comparing two groups)

- An effect size of 0.50 means that the difference between the two groups is equivalent to one-half of a standard deviation (Ellis 2010:11)
- An effect size of 0.8 means that the score of the average person in the experimental group exceeds the scores of 79% of the control group.
- An effect size of 1.0 means that the difference is equal to one standard deviation.
- A *d* can range from negative infinity to positive infinity.

Figure 11.1 Interpreting effect size

Interpreting effect size: Cohen's *d* for two groups	
0--0.20	weak effect
0.21--0.50	modest effect
0.51--1.00	moderate effect
> 1.00	strong effect

(Adapted from Cohen Manion & Morrison 2011:617)

NB. These interpretations are a LAST RESORT. Cohen saw them as simplifications, and should be used when you really have no other choice. **The key is the choice of the Standard Deviation, and the context.**

Think of *d* as a ratio: the observed effect (numerator)
---
a specified sd (denominator)

NB. For interpretation purposes, the number is extremely sensitive to the denominator (standardiser) ie what are you comparing it with? If the SD is small, then people do not vary much, and even a small improvement will lead to a large *d*. Conversely, when people vary greatly, the SD will be greater, and it may be difficult to get a large *d*.

11. Non-normal data

NB Effect sizes ONLY apply to normally distributed statistics ie numbers that are a close fit to the normal curve. Therefore:

- Make sure you test for normality. The free program: SOFA will do this.
- If you are not sure, this is a good question to ask a statistician.
- If the data is NOT normal, or you think it might not be normal, make sure you use non-parametric statistics.

12. Importance of stating the units of Effect Size

Cohen's *d* needs the units in order to be understood and interpreted. Often, several are possible, therefore careful choice is needed.

When you see *d* appearing in an article, it's essential to know how the author calculated that *d* – otherwise the values are not interpretable. (Cumming 2012 Ch11).

Example 11:2 Effectiveness of new numeracy training and Cohen's *d*

In measuring reading ability, it is quite common in education to talk about 'reading age'. There are various ways of defining and measuring it, and it is used by educators as a rough way of making sure that reading material matches the 'reading age' of the child.

There is a similar concept known as 'numeracy age'. This of course is very cultural, since children start school at different ages in different cultures.

Suppose you do some training, and the average score in a class rises by 5 points, as measured on some standard numeracy scale. A conversion table is also available, and this translates to an equivalence of an increase of 3 months of numeracy age. This gain, expressed as numeracy age, would probably be widely understood.

Suppose the Standard Deviation of the point system is 15, then the progress could be expressed as 5/15 = 0.33, or 1/3 of a standard deviation. This way of putting it would be more meaningful to researchers and academics. You would not need to know exactly what the numeracy test was, or exactly what is a numeracy age. A gain of one third of a Standard Deviation actually makes sense, and has the advantage that it is independent of the actual test, the actual way of measuring.

At this point, you have two choices. You could compare it to the reference values above. Another way would be to use your own judgement, taking into account the circumstances. For instance, such an improvement in a short time would be more significant than if it happened over a year or more. NB, before you get too excited about an impressive result, consider the Confidence Intervals, which will show you what the weakest students achieved, and what the strongest students achieved. It could be that this method had the most effect on the weakest students, and confused the strong students so that they regressed (went backwards)! (Adapted from Cummings 2012 Ch11).

Example 11:3 Marathons and Cohen's d

Suppose that a friend announces they improved their time in a marathon by $d = 0.2$. What would you think?

- Applying Cohen's criteria, you would say the improvement was 'weak'.
- If the standardiser is the SD of everyone who completes large marathons, is 40 minutes, then an improvement of 0.2 is 0.2 x 40 = 8 minutes. Is this an impressive improvement? It depends on the context, and for instance whether the person is male or female. A standard marathon is about 42 km, and in the Olympics is completed in just over 2 hours. But in regular racing, times of less than four hours are good times.
- The difference between the top 5, and the top 20 runners may only be a few seconds. Therefore, if the person is a top runner, then this improvement is suspiciously high. To repeat some information:

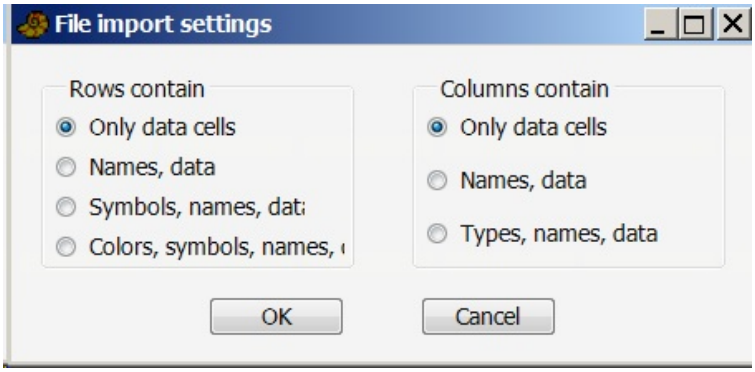
  Think of *d* as a ratio: the observed effect (numerator)
  ---
  a specified SD (denominator)

  For interpretation purposes, the number is extremely sensitive to the denominator (standardiser) ie what are you comparing it with. If the SD is small, then people do not vary much, and even a small improvement will lead to a large *d*. Conversely, when people vary greatly, the SD will be greater, and it may be difficult to get a large *d*.

Example 11:4 Differences between vocabulary in Research Articles

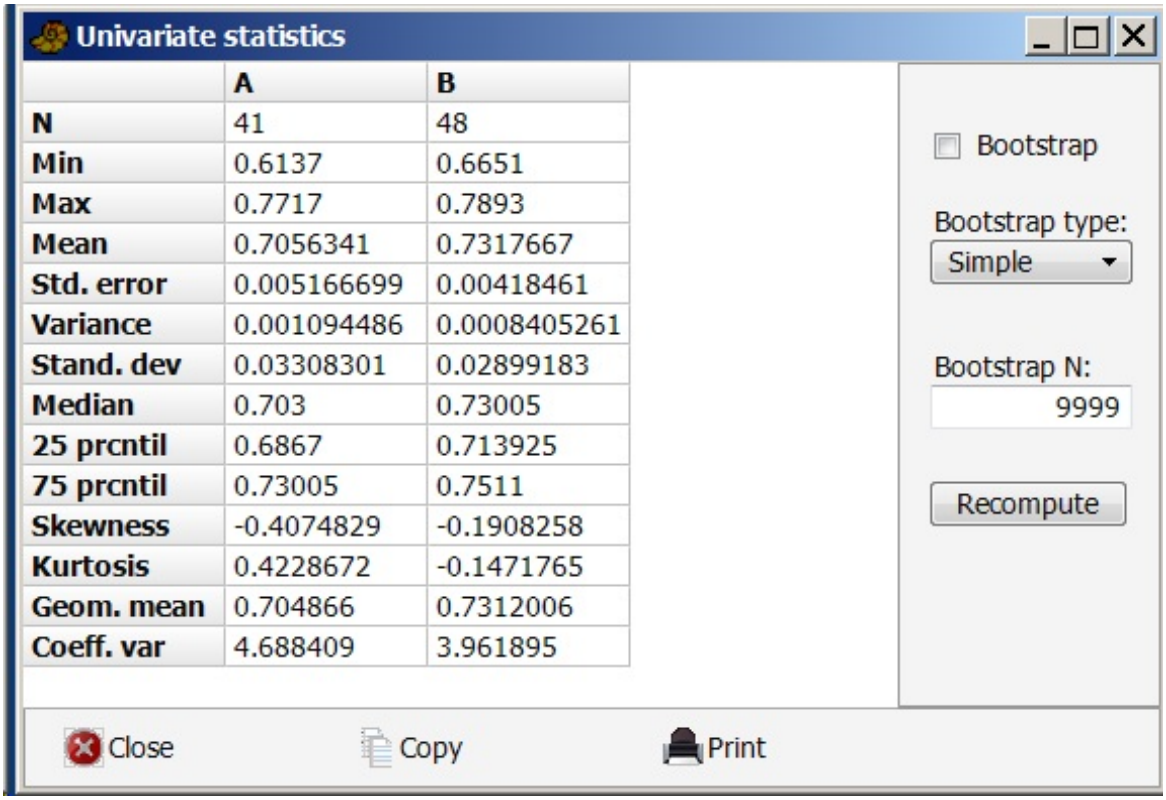
All the clearly identifiable Research Articles from the ESP journal were downloaded, converted to txt format, and analysed using AntWordProfiler to establish the percentage coverage of the K1 words. [The K1 words are those in the first 1000 most common word families]. The student was interested in the variation between articles, and the variation between the two years.

Using the program past3, the data was imported from an xls file using File|Open menu. This then gave a question screen, and the data only options were chosen.



Once imported, you have to mark both columns, by holding down the Shift key and clicking on each column in turn.

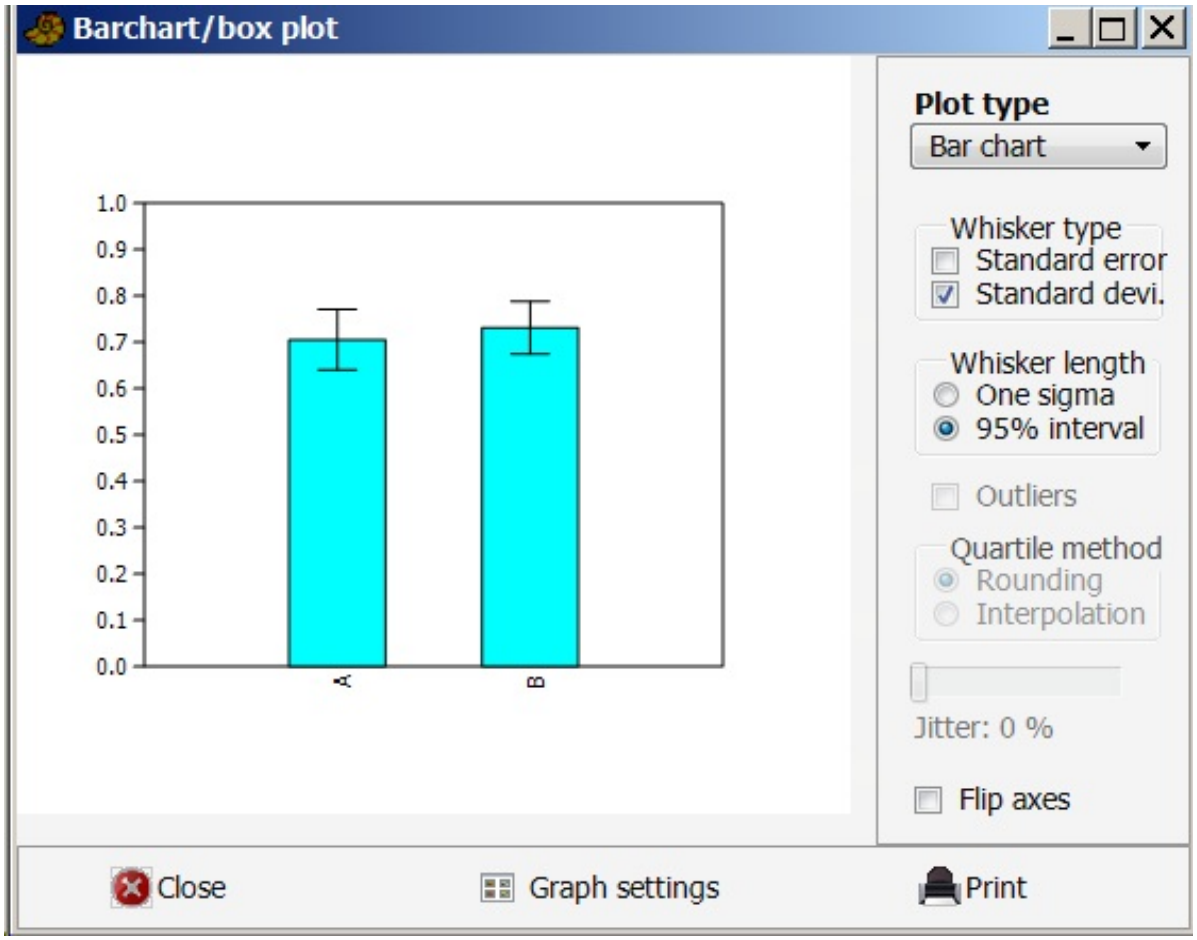
From there you can explore the menus. Univariate is the menu you want, and you see there the choice of "Summary Statistics". Click on that and you get the following data.



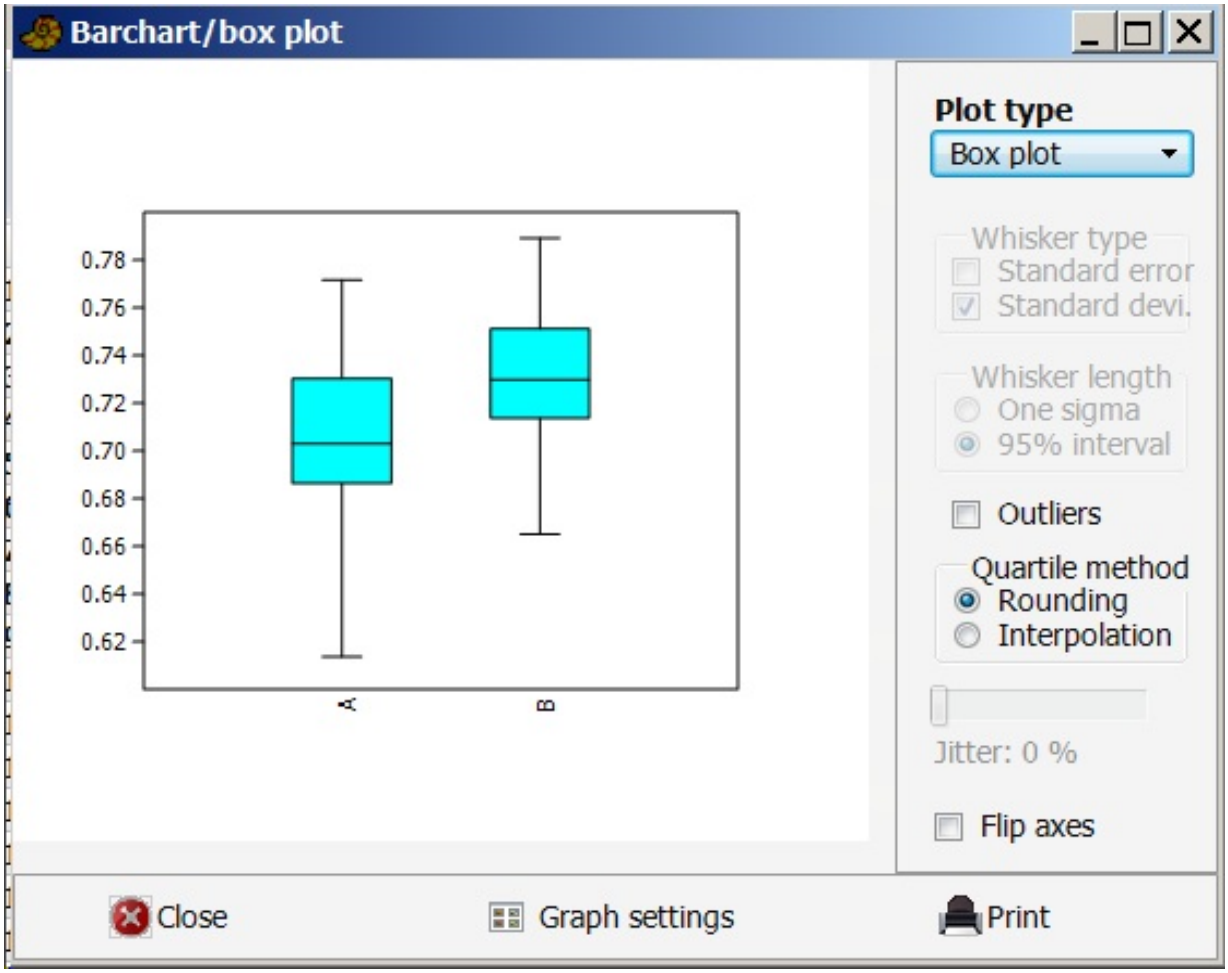
	A	B
N	41	48
Min	0.6137	0.6651
Max	0.7717	0.7893
Mean	0.7056341	0.7317667
Std. error	0.005166699	0.00418461
Variance	0.001094486	0.0008405261
Stand. dev	0.03308301	0.02899183
Median	0.703	0.73005
25 prcntil	0.6867	0.713925
75 prcntil	0.73005	0.7511
Skewness	-0.4074829	-0.1908258
Kurtosis	0.4228672	-0.1471765
Geom. mean	0.704866	0.7312006
Coeff. var	4.688409	3.961895

The important information here is N (number of articles in each year) mean, and standard deviation. You will need these figures for the effect size calculations. But, you do not need the calculations. **Just look at the graphs.**

The barchart with 95% interval looks like this:


Barchart/box plot
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
A
B
Plot type
Bar chart
Whisker type
Standard error
Standard devi.
Whisker length
One sigma
95% interval
Outliers
Quartile method
Rounding
Interpolation
Jitter: 0 %
Flip axes
Close
Graph settings
Print

The boxplot for 95% intervals looks like this:



This graph clearly shows the minimum, maximum, mean, and the 95% confidence intervals, ie, 95% of the results are in the box. It also shows, especially for group A, that the data is NOT normally distributed, since within the boxes, above the mean is greater than below the mean.

NB. It is from graphs like this that you can then go on to decide, is there an interesting difference. It is your decision. You cannot use a ‘statistical test’ to take the decision for you.

Example of a real situation with similar means and different standard deviations
(based on a true story)
Two teachers agreed to share the marking of an examination. They each took about 200 scripts each and agreed the mark scheme. Both teachers were under high pressure to mark the papers quickly and wanted to avoid 'double marking' if possible. Therefore when each of them had finished their marking, the mean scores were quickly calculated. They did this by listing and counting how many students got 1/20, how many got 2/20 and so on. In this way they swiftly had some frequency data from which the mean was easily calculated.

The mean scores were almost identical.

Could they therefore conclude that double marking was not needed? Afterall, the sample, the amount of copies for each teacher was large. Double marking would have meant at least another 10 hours work plus the time for 'confrontation' when widely differing marks were discussed.

At this point by now, you should be asking, is the mean enough information for comparing two groups? Obviously it is not. Something like the standard deviation needs to be calculated. This was simply done by observing the frequency data. It became quite obvious that the marks of one teacher were bunched around the mean, whereas the other teacher had a wider range of marks. The second teacher was giving more high marks and compensating by giving more low marks.

So, time pressed, what could be done? They agreed not to penalise anyone who had been given a high mark. They looked at the low marks of the second teacher and in many of them added two marks.

This example shows that simple methods can be used to apply statistical reasoning to a real world situation.

See also the comments on kurtosis! At the time, this was one factor that was not considered.

CHAPTER 12
POWER

A. What is statistical power?

1. Power refers to the ability of a study to detect a difference that is real. Do you have a big enough sample to credibly detect and study the effect or association you are interested in.

2. Power is an estimate of the ability of the test to separate the 'effect size' (see below) from random variation

3. Power refers to the likelihood of *avoiding* a false negative.

4. The power of a statistical hypothesis test measures the ability of the test to reject the null hypothesis when it is actually false – ie to make a correct decision.

5. The power of a hypothesis test is the probability of not committing a type II error also known as a beta error. It is calculated by subtracting the probability of a type II error from 1, usually expressed as:

   Power = 1 - P(type II error)

   The maximum power a test can have is 1, the minimum is 0. Ideally we want a test to have high power, close to 1.

6. Statistical power is the probability of rejecting the null hypothesis if there is a real effect in the population.

7. In general, larger *N* (samples) give higher power, and smaller alpha (false positive) demands lower power.

2. The components of statistical power

There are several possible reasons why some research will fail to detect a real difference. The figure below illustrates some of the likely scenarios.

3. For any power calculation, you will need to know:

8. What type of test you plan to use (e.g., independent t-test, paired t-test, ANOVA, regression, etc.
9. The expected effect size ie size of the effect you are measuring
10. The standard deviation
11. The sample size you are planning to use

4. To estimate how large a sample you will need for a study

This requires an estimate of the true difference (eg between experimental and control groups) that you are trying to detect, the associated SD, and the level of power you wish to achieve (perhaps 85 or 90%).

Ellis (2010:62) has a convenient summary for comparing two groups using a two-tailed test which I have modified to make it clearer. Ellis provides the total minimum number and assumes the reader will divide this total equally in two. I have done this step for the reader.

Ellis also provides a similar table for ES = *r*, ie the ES needed when doing correlations. But this is less common, and if the reader needs this table they can find it for themself.

Figure 12:1 Minimum sample sizes for comparison of groups

Comparing two groups: minimum sample sizes per group			
Desired ES as measured by *d* or similar	**Power = 0.70**	**Power = 0.80**	**Power = 0.90**
0.10	1236	1571	2116
0.20	310	394	527
0.30	139	176	235
0.40	79	100	133
0.50	52	64	86
0.60	36	45	60
0.70	27	34	44
0.80	21	26	34
0.90	17	21	27
1.00	14	17	23

(After Ellis 2010:62)
ES refers to Effect size, measured by Cohen's *d* or similar.

This table enables you to relate the sample size to the effect size. When the effect is small, you need high power to detect it. Power of 90% is an expression of confidence you have got the power right.

When the effect is large, then you only need low power.
As a rule of thumb, you can read Effect Size as the difference in standard deviations between two groups.

NB, the table shows the minimum number per group. In practice you will have two groups. Elsewhere I show you that

effect size and power are crucially determined by the smaller of the two groups.

See Ellis (2010:139, 140) for more detailed tables.

Ellis (2010:64) has provided a potentially very useful table in which, for conventional requirements of Power = 0.08 and alpha = 0.05, he presents the minimum predicted *r* or *d* that, for a given sample size, will be required to realistically establish a difference. The reader is well advised to obtain Ellis (2010) and to play with these tables and perform thought experiments along the lines of "what if I have a two samples of 40 what is the minimum *d* for a one-tailed test? Two tailed test?

Figure 12:2 Minimum sample sizes for correlations

Comparing two variables: minimum sample sizes per variable			
Desired ES as measured by *r* or similar	**Power = 0.70**	**Power = 0.80**	**Power = 0.90**
0.10	1233	1569	2099
0.20	308	391	523
0.30	137	173	231
0.40	77	97	129
0.50	49	62	82
0.60	34	41	56
0.70	25	33	42
0.80	19	23	32
0.90	15	18	24
1.00	12	15	19

(After Ellis 2010:62)

To understand this you need to remember correlations. Remember that perfect correlation is 1.0 and zero correlation is 0.0. Also, it is the strength of the correlation which concerns us.

A very acceptable correlation of 0.7 only needs 42 examples for the high power of 0.9, but, you cannot predict in advance what correlation you will get! So, low correlations need higher power.

5. **How do I estimate effect size for calculating power?**
   Because effect size can only be calculated after you collect data from program participants, you will have to use an estimate for the power analysis. Common practice is to use a value of 0.5 as it indicates a moderate to large difference.
   This page has some interesting clear advice:
   http://meera.snre.umich.edu/plan-an-evaluation/related-topics/power-analysis-statistical-significance-effect-size

   For a free program and other information on calculating power see:
   - **G Power**. This is a free online power analysis software program. It can perform power analysis tests for all of the most common statistical tests in behavioral research. If you want to avoid the trial-and-error process of finding a sufficient sample size, **G Power will allow you to input the desired power (e.g., 0.8) along with your statistical test type, alpha value, and expected effect size to generate the minimum sample size needed.**

   http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/download-and-register/index_html

   - **Optimal Design.** This is a more advanced, but free tool for power analyses.

   http://sitemaker.umich.edu/group-based/home

6. **The importance of power analysis**
   The importance has been slow to catch on. Larson-Hall is able to go back to 1971 in which Tversky and Kahneman argue that studies fail and data is rejected, not because of reality, but because the experiment was not big enough to show a difference or correlation.

   When researchers fail to find something they are expecting, they often go back for an explanation to the theory, or to the circumstances. This is commendable! But, often, the real problem is that the sample sizes were too small in the first place.

The real problem is that researchers did not ask the question about power before they started collecting data. I have shown you above that it is possible to ask power questions BEFORE collecting data, so that you know how big your sample should be in order to be likely to find a meaningful result.

On the other hand, Cumming (2012) sees Power as a distraction and a reversion to null hypothesis significance testing. **Confidence Intervals, with attendant attention to precision is more important, and can replace power.**

7. Maximising power

a. Power is greatest when there are two equal groups

An experiment with 10 in the control group and 30 in the experimental group is much less powerful than an experiment with 20 in each group.

Traditionally, you might assign one group as control, and three groups as experimental. The statistics would work best if the control group was as big as the experimental group. And in medical research, this is a deal killer. It is often hard to find enough people to take part in the control group.

b. Whenever the two groups are unequal, then calculations need to be based on the so called 'harmonic mean' of the two groups.

c. Burns (2000:185) gives the example of:

control group	6
test group	34
Total participants	40
The harmonic mean	10

So, even though there are 40 participants, the study has the power of a study with a control group of 10 and a test group of 10, and a total of only 20 participants.

To say it another way, a study of 20 participants divided equally, has the same power as a study of 40 participants where only 10 are in the control group. 20 participants divided equally with a large effect size of 0.80 has a power of only 0.39, whereas 40 participants increases the power to 0.69 and 60 participants divided equally increases the power to a respectable 0.86.

d. Power is greatly influenced by sample size, but it is even more influenced by alpha level (false positive) and the ES in the target population, also known as δ delta.

8. Informativeness

Cumming (2012 Ch12 and Ch13) argues that concern for Power is important, but often misses more fundamental concerns. These are:

- Representativeness of our samples
- Quality of our measuring tools
- Informativeness

So, even if you do not understand the rest of this chapter, you should understand these three basic points, and be able to work in your own research to make sure they are of high quality.

8. Examples for interpretation

Nosek Spies and Motyl (2012) did an experiment to see how well moderates and extremists from the political right, left, and center, perceived shades of grey. “The results were stunning. Moderates perceived the shades of gray more accurately than extremists on the left and right ($p = .01$)”. (p3). This was publishable and interesting, but something made them pause before publishing.

Writing note

The authors use American spelling and I have NOT changed this when quoting, though I have used normal British spelling in my commentary. They also use a slightly different style for reporting the *p*-value and again I have reproduced their style.

The authors paused, and because replication was easy, they conducted a direct replication. They reported (p4) "We ran 1300 participants, giving us .995 power to detect an effect of the original effect size at alpha = .05. The effect vanished (p = .59)."

Greater power meant that the effect vanished. The original results were an artefact of low power.

9. Margin of error

We have already seen that CI give a clear indication for the margin of error. A wide CI indicates low precision, and a narrow CI indicates high precision. Cummings (2012 Ch13) suggests that where the MoE (Margin Of Error) is less than half a Standard Deviation then you need at least 18 subjects. This fits fine with the general rule of thumb advice to have at least 30, per variable.

In fact, there is a new field opening up, known as "accuracy in power estimation (AIPE) which may be worth watching for the future.

When planning research, precision is better than power. This is where the real work and creativity takes place. It is also closer to the actual real situation being studied, and avoids completely the need for significance testing.

Figure 12:3 Approximate power for studies using the t-test for independent means, at the 0.05 significance level
(After Burns 2000:186)

Number of participants in each group	POWER		
One tailed	**ES=0.20**	**ES= 0.50**	**ES= 0.80**
10	.11	.29	.53
20	.15	.46	.80
30	.19	.61	.92
40	.22	.72	.97
50	.26	.80	.99
100	.41	.97	---
Two tailed			
10	.07	.19	.39
20	.09	.33	.69
30	.12	.47	.86
40	.14	.60	.94
50	.17	.70	.98
100	.29	.94	--

When the ES is large ie the observed difference is large, (equivalent to 0.8 SD) then groups of 30 will give good results, especially if it is one tailed. On the other hand, if the effect is small then even 100 in each group is too small for reliable detection of the effect.

Remember, the higher the power, the more likely the data is real. The effect size refers to the size of the effect being

measured. When the effect is small, then you need high powered studies to detect it, which basically means you need large ones. On the other hand, when the size is large, then you can cope with smaller samples.

12 Power 12

CHAPTER 13
CONCLUSIONS

1. Introduction

In this chapter I want to sum up, and discuss the perception of statistics. I will also suggest some more reasonable directions for research using statistics.

I have already stressed that a good researcher describes a population well, and lists the variables. It is time now to talk about rival hypotheses. The addiction to null hypothesis testing must be broken: it oversimplifies a complicated situation. There are usually more factors than one influencing an outcome. These factors may even be more important than the effect being studied. For instance, it is plausible that learning style significantly affects the speed of learning and the size of learning, and that this is always positive. But, reviews of learning style usually conclude that learning style is only one of many factors, and is dwarfed by motivation (which in itself is a megacluster of related and interacting variables).

Instead of trying to disprove a null hypothesis that no one believes, it would be more realistic to concentrate on stating the rival hypotheses in detail, and designing experiments that distinguish between them. One of these rival hypotheses is chance, but it is only one of them, and probably the least likely!

Or, try stating the main possible explanations/hypotheses, then trying to look for data that will favour one, and work against the other.

There is a time and a place for ruling out chance as an explanation. This is when chance is the most plausible alternative explanation. This is the case with studies where people are randomly assigned to the groups. Since in the social sciences such random assignment rarely happens, chance should not be the major concern of hypothesis testing. In many cases, chance is only one of several rival hypotheses, and it is often not the most likely (plausible) option.

It is all too easy to conduct work and assume that disproving the null hypothesis means accepting the favoured alternative!

As Stinchcombe (1968:13) says in an oft quoted passage "A student who has difficulty thinking of at least three sensible explanations for any correlation that he is really interested in should probably choose another profession."

2. Perception of numbers

This theme keeps popping up. I first covered it in my book, A feel for statistics, where I explained the difference between numbers and percentages, and how percentages often exaggerate a difference.

How people interpret numbers is strongly influenced by how they are presented. Apparently, (Cummings 2012) this is even true for academics, so how much more so for lesser mortals like students of statistics!

This is not just academic games – there can be serious consequences for health, life, and even death. In 1995 news media reported that taking a new third-generation contraceptive pill increased the risk of a dangerous blood clot by 100%. Wow, that looks impressive. As a result, many women stopped taking contraception, and chose to have an abortion, which is in itself a risky surgical procedure and probably these risks were much greater than the clotting risk, and probably resulted in an extra 13000 abortions until the panic was over.

Exposing this sort of scare is the stuff of a BBC podcast and website called 'moreorless with statistics'
http://www.bbc.co.uk/podcasts/series/moreorless
http://www.bbc.co.uk/programmes/b006qshd

If you have never listened to these podcasts, I suggest you do so. They regularly explain and question numbers in the news, in a fun way. Anyone who has got this far in this book should easily be able to understand them.

Back to the clotting scare. The actual increase in blood clots was from 1 in 7000 to 2 in 7000. That is a 100% increase in risk, but put this way, in terms of a ratio that everyone can understand and evaluate for themselves, this risk is tiny.

Natural frequencies are the easiest to understand for everyone. So if you see risk being expressed any other way, the first step to calm evaluation is to convert it into natural frequencies.

There is an operation called the Coronary artery bypass graft (CABG, pronounced like 'cabbage'). Mortality in the USA was down to 3.9% in 1990, and to 3% in 1999. The UK now reports a 98.4% survival rate' in 2008. Note the difference in framing. The USA reports in terms of death rates, and the UK reports in terms of survival rates. This change of framing makes the UK performance appear better, and obscures differences. For instance, a difference in survival between two hospitals of 96% or 98% looks negligible. In the USA, a difference in death rate of 2% or 4% looks like doubling the rate.

An increase from 2% to 4% looks serious. A decrease from 98% to 96% does not look serious.

REFERENCES

Note: CCJ (2017) refers to Cumming G & Colin-Jageman.

APA (2009). *Publication manual of the American Psychological Association*. 6th edition. American Psychological Association, Washington DC.

Blastland M & Spiegelhalter D (2013). *The Norm chronicles*. Profile books, UK

Burns RB (2000). *Introduction to research methods*. Sage publications, London.

Calder J (1996). Statistical techniques. In Sapsford R & Jupp V (eds) *Data collection and analysis* p226-261. Sage Publications, Open University Press, UK.

Cohen L Manion L & Morrison K (2011). *Research methods in education* (7th edition). Croom Helm: London.

Cumming G (2012). *Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge

Cumming G & Colin-Jageman R (2017). *Introduction to the new statistics: estimation, open science, and beyond.* Routledge New York.

Ellis PD (2010) *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. CUP.

Fisher RA (1955), Statistical methods and scientific induction. *Journal of the Royal Statistical Society* (B) 17, 69-77

Fisher RA (1956) *Statistical methods and scientific inference*. Oliver and Boyd. Edinburgh

Gigerenzer G (2004) Mindless statistics. *The Journal of Socio-Economics* vol 33, pp 587-606

Goldacre B (2012). *Bad Pharma: How drug companies mislead doctors and harm patients*. Fourth Estate, London.

Greenhalgh T (2007). *How to read a research paper: the basics of evidence-based medicine.* (3rd Ed). Blackwell UK.

Kline RB (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research.* APA books, USA

Larson-Hall J (2010). *A guide to doing statistics in second language learning research using SPSS*. Routledge UK

Lempert, R. (2009), The Significance of Statistical Significance:. *Law & Social Inquiry*, 34: 225–249.

McAleer S (1990). Twelve tips for using statistics. *Medical Teacher* 12(2)127-130.

Nosek BA, Spies JR & Motyl M (2012). Scientific Utopia: II. *Restructuring incentives and practices to promote truth over publishability*. Cornell University Library.

Reinhart A (2014). *Statistics done wrong*. Internet: www.refsmmat.com/statistics/index.html

Rowntree D (1981). *Statistics without tears: A primer for non-mathematicians.* Penguin, UK.

Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In Lisa A. Harlow, Stanley A. Mulaik, and James H. Steiger (Eds.) *What if there were no significance tests*? (pp. 37-64). Mahwah, NJ: Lawrence Erlbaum Associates.

Stinchcombe AL (1968) *Constructing social theories*. Harcourt Brace & World, New York.

Waller, N.G., (2004). The fallacy of the null hypothesis in soft psychology. *Applied and Preventive Psychology* (11) 83-86

APPENDIX 1: VALIDITY AND RELIABILITY

A. GENERAL

1. Introduction

Every thesis should involve an accurate assessment of validity and reliability. Considerations of validity and reliability are involved in the early stages of planning the research, the methods, data collection, and also in presenting the results, discussion and conclusions. Therefore it is vital that this subject become so well known that it is instinctive to all researchers.

In this chapter I present the basics. Types of validity and reliability seem endless, but those presented here are sufficient.

Any student of literature who thinks this chapter is irrelevant to them, should think again. This chapter is highly relevant, and I have a special section at the end, citing Literature experts, to defend my case. It is also relevant to them for teaching, and, indeed, for normal life. In particular, the distinction between validity and reliability is frequently needed, not least in the field of personal health, so that we can begin to understand and sensibly interact with doctors. Read on.

2. The difference between validity and reliability

In simple terms, reliability is how well you can trust your measuring tools so that they always give the same result when measuring the same object. Validity refers to how fairly you can generalise your findings to other groups or other situations.

Burns (2002:350ff) explains it by asking students to imagine a factory which produces 30cm plastic rulers (Burns, being American, uses the antique unit inches, here I have changed it to modern units). Now it is well known that it is easy to make a mistake in the mixture of the plastic, so that sometimes a batch of rulers is produced that is, in practice 30.3cm long. Now the question is this. Are the rulers reliable? Are they valid?

The rulers are quite reliable, because they produce consistent results. Drawing a line of 30cm will always mean drawing a line of 30.3cm. But the rulers are not a valid measuring tool, since though they consistently give the same result, they are not actually measuring the 30cm accurately.

A similar question arises with other measuring tools, that often need calibrating. A blood pressure meter will always give consistent results, so that a change of blood pressure would be reliably detected. An increase of 10mmHg would probably be measured accurately on most instruments, but the absolute value, be it 140mmHg or 145mmHg for instance, will vary with the instrument. The newer electronic blood pressure meters need calibrating against the more valid (consistently maintaining standards) mercury barometer machines. If an electronic machine were used to establish data that were to be internationally acceptable, it would need to be calibrated, checked, and adjusted regularly, so that the readings would be valid and comparable with other meters. Remember, mercury instruments are simpler, easier to calibrate, and are known to be valid and reliable, unlike most electronic meters.

In logic, a valid argument usually means an argument that is coherent and relevant.

Figure A1.1, Summary of validity and reliability (Derived from LeCompte & Goetz 1982 p32)

	RELIABILITY Replicability, especially in interpretation. Repeatability. Trustworthy. Consistent	**VALIDITY** Accuracy. Measures what you set out to measure. Calibrated.
INTERNAL Agree on findings	Degree to which other researchers would match generated constructs with the given data	Observations are authentic representations of reality
EXTERNAL Selectivity problem	Independent researchers **discover the same phenomenon** or generate the same constructs	**Degree of generalisation** possible and how comparable the results are across groups

B. INTERNAL RELIABILITY

Internal reliability

This refers to the degree to which other researchers, given a set of previously generated constructs, would match them with the data in the same way as did the original researcher. It is concerned with the accuracy of scientific findings. Would other researchers match the data with the theory in the same way you have. This is a key concern for ethnographers and LeCompte and Goetz (1982) go into detail on how to enhance the probability that within a single study, several observers would agree on the findings, and would agree on how the theory fits with the data.

C. EXTERNAL RELIABILITY

1. External reliability

This addresses the issue of whether independent researchers would discover the same phenomena or generate the same constructs in the same or similar settings. LeCompte & Goetz (1982) say that external reliability is enhanced by being explicit about five major problems.

2. Specific problems with external reliability

a. Researcher status position

"Research reports must clearly identify the researcher's role and status within the group investigated". (p38).

b. Informant choices

"External reliability requires both careful delineation of the types of people who served as informants and the decision process invoked in their choice". (p38).

c. Social situations and conditions

"Delineation of the physical, social and interpersonal contexts within which data are gathered enhances the replicability of ethnographic studies". (p39).

d. Analytic constructs and premises

Be explicit about the theories used. "Replication requires explicit identification of the assumptions and metatheories that underlie choice of terminology and methods of analysis". (p39). Recognised frameworks and classifications have the advantage of helping the research to be understood and making the results more comparable, but they may hinder, in that the categorisation may be made prematurely, and the data may be made to fit the headings thus misrepresenting the data.

NB **The biggest danger** pointed out by LeCompte & Goetz (1982) **is selectivity of informants.**

Informants tend to tell you only part of what you want to know. This selectivity can be minimised by seeking corroborating evidence – triangulating and checking so that you do not rely on just one informant, or you check using for instance written information.

e. Methods of data collection and analysis

"Ideally ethnographers strive to present their methods so clearly that other researchers can use the original report as an operating manual by which to replicate the study." (p40). The authors argue that shorthand designations for methods are inappropriate, since there is no commonly understood set of descriptors for the many methods that can be used in ethnography. They also give an admonition that replicability is impossible without precise identification and thorough description of the methods used to collect and especially analyse data. (p40).

D. INTERNAL VALIDITY

1. Internal validity

This refers to the extent to which scientific observations and measurements are authentic representations of some reality. It concerns how closely the theories match the situation, is often a major strength of an ethnograph in that unlike surveys and other quantitative techniques, the ethnographer often lives in a situation over an extended period of time, which gives the opportunity for refinement, and continual re-evaluation of the research.

2. The main threats to internal validity
See also Burns (2002:357-360 and Cohen & Manion 1984:194-195)

a. History ie other events
Sometimes in research we do a test, then after doing something, some time later, we retest. In theory the changes noticed are due to what you did. But, in this time something else might have happened and the change might be due to these other things, other variables, not due to what you planned. Time is a threat to internal validity.

b. Maturation
Subjects mature over time, and the result may be due to these maturation factors rather than your experiment.

NB c. Regression towards the mean
There is a statistical fact of life that is worth knowing about even if you do not understand.

- Over a series of tests, people often do not score consistently,
- Results of frequent tests and measurements tend to average out near the group average.
- Subjects scoring high on a pre-test are likely to score lower on a post-test.
- Subjects scoring low on a pre-test are likely to score high on a post-test.

That makes it difficult to explain gains and losses in the results. Statistical regression happens because of the unreliability of the measuring instruments, the many extra variables that can intervene and affect people, and the way that in many phenomena there are natural swings.

d. Testing effects
Whenever you give a test, you give the students practice in what you are testing, and you may sensitise them to the purposes of the test. Once sensitised the students will do

better, just because they have been tested, not because they have had for instance extra teaching. This is the old story in science that when you measure something you change it, and maybe change it irreversibly.

e. Instrumentation
The measuring instrument itself may be suspect. With human observers or judges error can result from changes in their skills and levels of concentration over the course of the experiment.

f. Selection bias
Bias may be introduced by the way groups are selected. In addition, selection bias may interact with other factors such as history and maturation. Selection bias makes valid comparisons and valid conclusions difficult.

g. Dropout
In long running experiments, some people may drop out, so the final group will be select, and therefore different in composition to the original group.

3. Example of research with questionable internal validity
In an investigation of three different methods of teaching grammatical structure, three teachers in three different schools are each trained in one of the methods and apply it to their classes. One teacher has three mixed ability classes, another has four mixed ability classes, and the third has two homogeneous groups of fast track learners. Each group is administered a test devised by their teacher. Group means for each group are computed and compared.

Critique. The results are uninterpretable. It is impossible to say whether the results are due to the method, the proficiency of the students, the skill of the teacher, or the ease of the test.

People in Tunisia at the end of each term happily ask for the average of different children and compare them, even though these averages come from different schools with different

teachers and different tests. People still think the marks are comparable. The British government commits a similar mistake when they insist that schools publish their success rates, and then the government draws up league tables of schools, and attempts to argue that schools with higher examination results are actually better schools. The science press, scientists, and educators have repeatedly discussed and explained the errors. This government mistake has been widely repeated over many years and it is sad that intelligent officials and politicians continue to publicly make basic mistakes in statistics and even make policy decisions based on such known mistakes.

E. EXTERNAL VALIDITY

1. **External validity** addresses the degree to which such representations may be compared legitimately across groups. (Le Compte & Goetz 1982 p32). The validity of the research is a question as to how closely the propositions generated, refined and tested match the reality of a situation in everyday life. How easily can the findings can be generalised to other situations? **The common way of enhancing external validity is to establish how typical a phenomenon is**, ie the extent to which it is typical compared and contrasted with other known phenomena. This means for instance the clear identification, specification and evidence for distinct characteristics of what is being investigated. (p51).

   Where it is not possible to use techniques of random sampling and statistical analysis, the characteristics of the group studied must be spelled out clearly. The results can then be compared with others and hence have a wider applicability.

It is a basic early step in research to carefully describe your group or groups. You need to carefully list all the different factors and variables. This is where ethnographic work – knowing the local context and knowing the main players and how the institutions work is so important. Explicit and systematically stated knowledge is foundational to planning data collection and interpreting the findings.

In linguistics it can often be difficult to measure what you want to measure. A test of reading comprehension for instance may in fact only be a measure of general intelligence. A test of achievement may in fact be measuring general test-taking ability. It is very important to make sure you are measuring what you set out to measure.

2. The main threats to external validity

See Burns (2002:358-60), Cohen & Manion (1984:196)

Threats to external validity are likely to limit the degree to which generalisations can be made and the way your findings can be extended and applied to other circumstances.

a. Failure to describe the independent variables

Remember, independent variables are the ones you have no control over, you can only describe them and account for them. When doing human research you really must describe all the factors in the situation, so that when someone tries to replicate the work, these factors are either kept the same or at least taken into account.

b. Lack of representativeness of subjects

While your subjects may be representative of the local population, they may not be representative of another situation which you are trying to apply to. The TEFL/TESL distinction is one example where problems can exist. Learning English in Tunisia is totally different to immigrants learning English in London. Comparing Second Language contexts with Foreign Language contexts is possible, but great care is needed.

c. Hawthorne (Placebo) effect

The mere fact of taking part in an experiment may mean a change. In medicine, so powerful are the psychological effects, that when a new medicine is tested, it is usual for special 'double-blind' tests to be set up. Volunteers are randomly assigned to one of two groups: medicine, or placebo. Then the doctor issuing the medicine does not know which patient is getting which type of pill. If the doctor knows, then very subtly their interactions with the patient will vary and this can effect the results. Commonly, even when taking a placebo, patients improve. The question is does the value of the treatment exceed the value of the placebo?

3. Example of research with questionable external validity (The generalisability of the findings is doubtful.)

A study investigated the effect of length of visual exposure on the ability to memorise and recall nonsense words. Subjects were ten postgraduate students who were undertaking a master of arts program in psychology. There were five different lengths of exposure, so five groups of two volunteers each receive different lengths of exposure. A volunteer participated in the study by being exposed to 20 nonsense words individually. After each exposure, the volunteer had to reproduce the nonsense word.

***Critique*.** Assuming that the performance scores generally increase with increased length of exposure, the question remains: **To which populations and conditions can the results be generalised?** Can they be generalised to primary and secondary students learning meaningful material? Can they be generalised to young adults working on meaningful tasks in a highly structured situation? The answer to both questions is no. The results may not even be generalisable to the graduate student population, since the participants were volunteers. (Nunan 1992:16).

F. TYPES OF VALIDITY

Types of validity
This subject has received too much attention in research. There are many types of validity, and many ways and labels. Here are some of them.

1. **Content validity**
   Consider an examination. An exam has content validity if it examines the content and the skills that have been taught, and fairly tests some or all of the course. The question then is do the questions fairly assess the whole course? Would a similar set of questions get similar results?

2. **Predictive validity**
   On the basis of these results, can we make a statement about future performance? For instance, does success in the *sixième* or the *neuvième* reliably predict that these students will successfully go on to succeed in the Baccalaureate?

3. **Concurrent validity**
   Will a low score in the CCG (reading Comprehension, Composition, and Grammar) paper also mean that someone will get a low score in the laboratory examinations? Will a high score in the written paper be followed by a high score in the orals on the same subject?

4. **Construct validity**
   "A construct is a psychological quality, such as intelligence, proficiency, motivation, or aptitude, that we cannot directly observe but that we assume to exist in order to explain behaviour we can observe (such as speaking ability, or the ability to solve problems)." (Nunan 1992:15). Constructs need careful description in any research. Example: if a study investigates 'listening comprehension' and tests it using a written cloze test, then by default, the assumed understanding of 'listening comprehension' becomes 'the ability to complete a written cloze passage'. If we find such a definition unacceptable, then we are questioning the construct validity of

the study. It needs to be shown that a given test measures a certain construct.

G. ADVICE

1. Questions to ask about evidence are:

a. What are its sources?

b. Are those sources legitimate, or how much are they legitimate?

c. Are those sources reliable, or how reliable are they?

d. How selective is the data?

e. Is the evidence relevant?

f. Who has an opinion about your data or sources? Have you accounted for their viewpoint or argument?

g. What biases are there in your evidence?

2. The options you have when you write up your work are:

a. Ignore validity because it is not a problem

b. Declare the limitations of reliability

c. Use established validation procedures

d. Discuss the problems of reliability and validity, and assess your own work

e. Declare the assumptions on which your work rests

f. Include and criticise the rationale for your procedures

g. Evaluate the paradigm from which you are working (After Barnes 1992 p161-2).