A FEEL FOR STATISTICS: ESSENTIAL CONCEPTS UNDERLYING THE CALCULATIONS

Ivan Lowe 2016b. Version 1.1
www.scientificlanguage.com

ABSTRACT

Most published courses of basic statistics aimed at beginners are at too high a level for the needs of English students and teachers in their research and reading. In addition, there is a tendency for researchers to make life more complicated than needed and to seek the prestige of using a standardised significance test, when simpler methods can be used which are usually quicker and may even be more reliable and informative. There is a need to explain how numerical data can be presented and interpreted without recourse to formal statistics, and to train a feel for numbers and experimental design.

This book presents 20 keys to successful manipulation of numbers and their presentation and use by teachers and researchers. These are some of the concepts and simpler methods which are needed before formal statistical tests are used, and which can be part of the tricks of the trade of the working teacher and researcher.

INTRODUCTION

Practice point
Two versions of the introduction are presented below. The 'Simple version' is journalistic. The 'Long version' is written in the more formal, academic style appropriate to a textbook or research article.

As an exercise in noticing the features of academic writing,
1. Compare these two versions, ie list their similarities, and their differences.
2. Apply Swales' three moves (Context – Gap – Intend to fill gap) to BOTH introductions.

Simple version
You are fortunate! Very few people have the chance to study statistics in such an easy way as in this course. In this course I have done most of the hard work for you. I understand the fear and dislike that many of you have of statistics. I used to feel that way. Although I loved mathematics I used to dislike statistics. I found it was all full of Greek letters and formulae and tests that you did on some data. The mathematicians when they taught me statistics seemed to get lost in the intricacies of the formulas. When the non-mathematicians had a go, they got me lost in the different statistical tests. This course starts with ideas and concepts that come before the mathematics. It will introduce you to the ways researchers think, and it will help you to evaluate for yourselves any report you read which involves numbers.

Long version

The concern for accurate selection, use, and presentation of statistics is widespread, and is one which crosses subject boundaries. In linguistics and language teaching it almost seems obligatory that introductory texts on research methods will include some advice. Nunun for instance (1992:28ff) has a helpful section on "The logic of statistical inference" and in an earlier book (1989), has a long appendix devoted to a summary of the basics. Brown (1991) is particularly helpful in explaining the key steps of statistical reasoning as he sees them, and in part two (1992) clearly explains two major areas: variables, and the choice of statistical method. Turning to manuals of style for science, Rubens (1994), Turabian (1987) and JR Matthews et al (1996) but not CBE (1994), to name but a few, devote significant sections to the presentation of tables and graphs.

Little has changed since then, except for easier ways to present results through programs such as Microsoft Excel.

But, much of this is still at too high a level for many of our Arts trained colleagues and students, who may well have an irrational but powerful fear of numbers, and have given up with mathematics long ago. Even if people do manage to cope with the so called elementary or introductory texts, it is rare for me to work with someone I would describe as having a 'feel for numbers'. There is a gap in the reasoning and knowledge between the level they are at, and the so called basic texts. Few books or articles come close to bridging the gap.

Huff's book "How to lie with statistics" (1954) is by now a frequently reprinted classic and is still worth reading even though I wish the language could be modernised since in places the old English style renders it more difficult to read than need be. Fitz-Gibbon & Morris (1987) also come closer to real beginners, providing ways of handling numbers with a pencil, paper, and squared paper. Scholfield's guide (1995) to quantifying language is at the same time frustrating because of the crowded poor presentation, and fascinating because of the many clear explanations and insights.

Rowntree's "Statistics without tears" (1981) comes close to being really basic and is extremely helpful as far as it goes. But, he concentrates on the 'normal' curve, with hardly a mention of the existence of other types, or of rank order. He also lacks material on the interpretation of tables of figures.

This section is an attempt to bridge the gap between basic numeracy, and introductory statistics courses. In it I explain some of the foundational concepts that are needed in any work involving numerical data. I will attempt to write down what I instinctively know, so that we can begin to bridge this gap and acquire a feel for statistics. To this end eighteen inter-related key underlying concepts will be presented, concepts which underlie the formal statistics of basic texts.

Characteristics of a good thesis

1. Testable hypotheses
2. Clearly specified population and sample.
3. Questions of validity and reliability considered
4. Variables are identified, evaluated, and balanced one against another
5. Data analysis is planned BEFORE data is collected
6. If any statistical tests are to be done, these are planned in advance, BEFORE data is collected
7. Supervisor checks each stage
8. The data collection instruments (eg questionnaire) are piloted AND the data analysis is piloted.

KEY 1. DEFINITIONS AND CIRCUMSTANCES

The most important key to statistics is the definitions, which means beginning with the local conditions and historical context. Definitions can be different to what are expected, or simply not stated (the problem of missing information, Key 2 below). They can also differ according to the author, and, in official documents such as census reports, the definitions can change between editions. In TEFL, to take only one example, Crookes and Schmidt (1991:475,497) note that the study of motivation is severely handicapped for lack of a clear definition. Clarity of definition in this case means having roughly the same definition, and having a definition that is measurable in a clear easy unambiguous way.

Extra information

To avoid confusion, let me say clearly, that this concern must not be interpreted in terms of 'Social Constructivism', or the idea that knowledge is somehow 'constructed' or agreed upon by a group of people, and therefore has no real correspondence to reality. The concern for definitions is actually the opposite of Social Constructivism.

The concern for definitions is foundational to the correct understanding of and use of statistics.

a. A concern for definitions is a concern for:

1) **operationalism** - am I specifying clearly what I am proposing to measure in a way that makes the feature precise, measurable and testable?

2) **clarity**

3) **fair judgements and comparisons,** which can only be made when we are talking about the same thing. A lot of sloppy talk and confusion happens when we are not working to the same understandings.

Example 1

A memorable example of how the change of definitions affecting the numbers happened after the first and second world war in England when the divorce rate went up. Why?

> Just as after World War 1, there was a surge of divorces after World War 2, and probably for similar reasons, including war-time adultery by wives.

In this era, as well as the usual behavioural questions, women received improved divorce rights, therefore more were able to seek a divorce.

> After World War 2, one thing that happened was more focus on "the welfare state", and eventually better state support for lone mothers. Another was that women had got used to working during WW2, and didn't just go back to not working afterwards. So families often had a working mother, although normally far less well paid (although this didn't start with this war, of course). The trend for mother-custody still appeared to make sense, although not as clear as before. And the cost to the state was increasing.
>
> As divorce got easier, it was opened up to more people - the "middle classes" and (horror!) the "working classes". Why did mothers start to get custody? The law was still biased somewhat towards custody for fathers (until the final change in the Guardianship Act 1973). But mothers were getting custody long before that. I think the answer is simple - in most of the sort of families by then able to get divorced, the tradition was for the father to work and the mother to care for the children. So after separation, the obvious answer was to continue that. He was probably working long hours (and there was little useful childcare), and she was less likely to get an adequate job. The courts reflected the realities of family & working life.

http://www.childsupportanalysis.co.uk/information_and_explanation/world/history_uk.htm accessed 2 May 2007

Notice that after both world wars both the law and the legal customs changed.

Now I do not believe we should go over the top with definitions - like some French do. They endlessly define words, and that seems to be a large part of their introductions and literature review! As long as everyone knows what you are talking about you do not have to define - though you might want to indicate a few refinements. It will be remembered that even Popper, who was very concerned for testable hypotheses, was implacably opposed to spending a lot of time on definitions: **in science we are only as precise as we need to be and as precise as the measuring tools allow us to be.** (For a good introduction to Popper, see his autobiography: Unending Quest).

b. I begin with definitions because:

1) interpretation of statistics depends on knowing what is being measured and how,

2) people knowingly or unknowingly misuse statistics by
   a) not being explicit about their definitions,
   b) using one definition when people will use another. See Key 3 on different kinds of 'average'.

3) students who are fearful of numbers are usually less fearful when they see that statistics is a lot to do with definitions, which may or may not at heart be numerical.

   As Jack Cuzick, head of epidemiology at Britain's Imperial Cancer Research Fund, puts it: "There is no such thing as context-free statistics." In other words, **there is no substitute for looking at the data and thinking about what they mean**. (Editorial, New Scientist 15 January 2000 page 3).

What I am getting at is that we always need to know what people mean, and how they measured it.

Example 2
Background
In Britain, everyone gets a State Pension, which is supposed to provide the minimum for living. Everyone is encouraged to also build up an extra private, or occupational pension.

The problem
In the 1980s in Britain, the Conservative government distorted the truth about pensions by playing with definitions. The government asserted that two-thirds of older people had an occupational pension, therefore a state pension was no longer very important. But the assertion was based on evidence that looked only at men. Yet obviously, more men than women have an occupational pension. "By excluding women and failing to indicate this, the official statements gave the impression that only a minority of older people rely on state pensions, whereas over half do so." (Ginn 1999:118)

This is an example of misleading the public by playing with definitions.

KEY 2. MISSING OR INCOMPLETE INFORMATION

Another way readers are misled is when there is important information about the numbers that is not provided. Some people call this being "selective with the truth" or being "economical with the truth".

When statistics are used by people in their arguments, and particularly in newspapers where standards are lower, it is very easy for only part of the story to be told, and to miss out inconvenient data, or data that would work against the view being defended. Missing information can be of several kinds:

a. The definitions used are not specified (Key 1)

b. The sample size is not stated
The sample size is important because it is too easy for someone to make an unjustifiable generalisation simply because the sample size was too small. Or the size is hidden by being converted to a percentage. Percentages can make the numbers appear bigger than they actually are. Five people out of ten is 50%. It is far more impressive to say 50% than five out of ten. "Fifty percent" sounds grander, more conclusive, and more generalisable.

Example 3: from published work where the sample size was not stated

This type of missing information is very important, and yet, strangely, it is even found in reputable journals. I recently saw it in World Englishes (Chen 2006). The article initially looks very impressive. The research reported was of code-mixing between Chinese and English in the advertisements placed in 43 different magazines, and a total of 64 volumes. They found 226 examples of code mixing. But how many advertisements were read? Is this 226 out of 1000, or 266 out of 10,000 adverts? This is important, because code-mixing may vary in frequency, and this context of the frequency of code mixing in advertisements is very important for assessing the significance of the findings. To put it in simple English, are we dealing with a rare (but interesting) phenomena, or is code switching in adverts so frequent that it can be considered a major feature of magazine adverts in Taiwan?

While Cheng has, impressively, looked at 43 different magazines, they also only considered a total of 64 volumes. Some magazines therefore had only one issue as an example. This means I have another question as to how representative these single volumes were.

These two questions together impose huge limitations for the outsider reading the report. In particular, the **validity** of the work is called into question.

Other questions not considered by Cheng's work:

1. The types of advertisement
2. The readerships of the magazines. Eg would there be more code-mixing in a ladies magazine, or a computer magazine?
3. The readership figures of the magazines.
4. Details about the Chinese script used in the code mixing. Was it standard, simplified, or the Latin script?
5. The formality levels of the English and Chinese.
6. Was there any diglossia (high and low forms)?

c. The kind of average used may be missing see (Key 3), and also the range of values involved (Key 5).

d. When a formula is given, the meaning of each of the letters used is often not stated
This is important because conventions do change, or may not be known to the reader.

e. Actual numbers are missing
For most statistical formulae, the actual numbers, not percentages, must be used.

f. Evidence is excluded
Evidence that is contrary to the argument being presented must also be explicitly discussed. Inconvenient and contrary data must be included.

It is not surprising then, that given the many different ways that statistical information can be twisted, these two sayings are well known in Britain:
– There are three kinds of lies: lies, damned lies, and statistics
– You can say anything with statistics

KEY 3. WHICH AVERAGE IS IT?

Huff (1954 chapter 2) helpfully explains how there are three kinds of 'average', namely, the mean, the median, and the mode. All three types can legitimately called an average, and while they are often similar, they are not always. The trivial example is of a series of numbers: 3,3,4,4,7,7,7, the mean being five, the median (middle number when the series is placed in incremental order as here) is four, and the mode, the 'most popular' number, is seven.

Definitions

Average - common use. Example, examination marks in a class. Assuming there are 20 students and each receives a mark, then the average for the class would be a total of all the marks given in the class divided by the number of students.

Average - technical use. There are at least three definitions:

Mean. The common use 'average' as above.

Mode. This is the 'most popular' result. In the lung function data given below the mode cannot be observed. But in the initial example, 3,3,3,4,4,7,7,7, the number 7 appears more often than any other number, and is therefore the mode.

Median. The middle number. To find the median, you need first to put the numbers in order, the lowest numbers on the left, and the highest numbers on the right. We call this order **'rank order**' or '**incremental order**'. You then choose the number which is exactly in the middle.

Example 4: Lung function data
A teenager once achieved the following results over one month for the lung function tests FEV1 and FVC. [it is not necessary to know what this test is - I am simply using some real data that is readily available to me].

Date	Result for FEV1	Result for FVC
13.01	96%	97%
14.01	85%	106%
15.01	81%	89%
16.01	86%	102%
22.01	90%	94%
23.01	94%	99%
30.01	90%	100%

The incremental order for FEV1 is: 81, 85, 86, 90, 90, 94, 96. There are seven percentages, therefore the fourth result in this series of numbers is the number we are interested in. It is 90%. This figure is in the middle: three figures come in front of it, and three figures come after it. This figure is called the median.

Question: what is the mean for the FEV1?
Answer: (622/7= 88.9%)

What is the median of the FVC?
To find the median [notice the language: you find the median, and you calculate the mean] you first put the numbers in order, from low to high. This is: 89, 94, 97, 99, 100, 102, 106. Once again there are seven numbers, so the median is the fourth number which is 99%. There are three numbers which are smaller and three numbers which are equal to or larger than the median in this example. The mean is 687/7 = 98.14.

Question: How do you find the median when there is an even number of examples?
Take some new data, from the month of March. The figures for the FVC were: 125, 126, 131, 120,119, 120, 123, 128. Placing these in incremental order gives: 119, 120, 120, 123, 125, 126, 128, 131. The median in this case is the average (ie mean) of the fourth and fifth results. It is the mean of 123 and 125 (ie 123 + 125 divided by 2) and = 124.

Someone who knows that the word 'average' can be determined in three different ways can choose and quote the average that suits their purpose. For example, the average level of pay can be distorted if there are a few very high wage earners, and the average (mean) can thus appear to be higher than it should be.

Example 5. There is a small factory of twenty people, including the manager.

5	earn	500 per month.	5 x 500= 2,500
10	earn	1000 per month	10 x 1000= 10,000
5	earn	1500 per month	5 x 1500= 7,500

The monthly wage bill is therefore 20,000. In this case the mean, usually written $\bar{x}$, is 20,000 divided by 20 = 1,000 per month.

Remember the convention: the symbol for the mean is usually $\bar{x}$
Mean: 1000
Median 1000
Mode: 1000

Example 6. But supposing one of the higher earners earned 6500 instead of 1500.

5	earn	500 per month.	5 x 500= 2,500
10	earn	1000 per month	10 x 1000= 10,000
4	earn	1500 per month	4 x 1500= 6,000
1	earns	6500 per month	1 x 6500= 6,500

In this case the total wage bill is 25,000, so $\bar{x}$ is 25,000 divided by 20, which equals 1250 per month.

Mean: 1250
Median 1000
Mode: 1000

Example 7. Supposing the highest wage earner earns 16,500 per month.

5	earn	500 per month.	5 x 500= 2,500
10	earn	1000 per month	10 x 1000= 10,000
4	earn	1500 per month	4 x 1500= 6,000
1	earns	16500 per month	1 x 16500= 16,500

In this case the total wage bill is 35,000, so $\bar{x}$ is 35,000 divided by 20, which equals 1750 per month.
Mean: 1750
Median 1000
Mode: 1000

In examples 6 and 7, the mode (the wage earned by the largest subgroup of workers) or the median (the wage earned by the middle worker) would be a fairer average. Both the mode and the median stay constant at 1000 per month, even though one person has got an inflated salary much higher than anyone else.

As Woods et al (1986:32) point out, **the median is particularly useful when**

a. There are a few atypical examples at one of the extremes of the range of values, since it is relatively unaffected by

them. **[atypical** means non-typical, since the prefix a- commonly means non-].

b. A large difference between the median and the mean is also a clear indication that the data is skewed.

This word **skew** is often used in statistics. In the books on elementary statistics you might find it explained as 'there is not a symmetrical distribution of scores on either side of the mean'. If you do not understand that, I have made my point that the books assume too much.

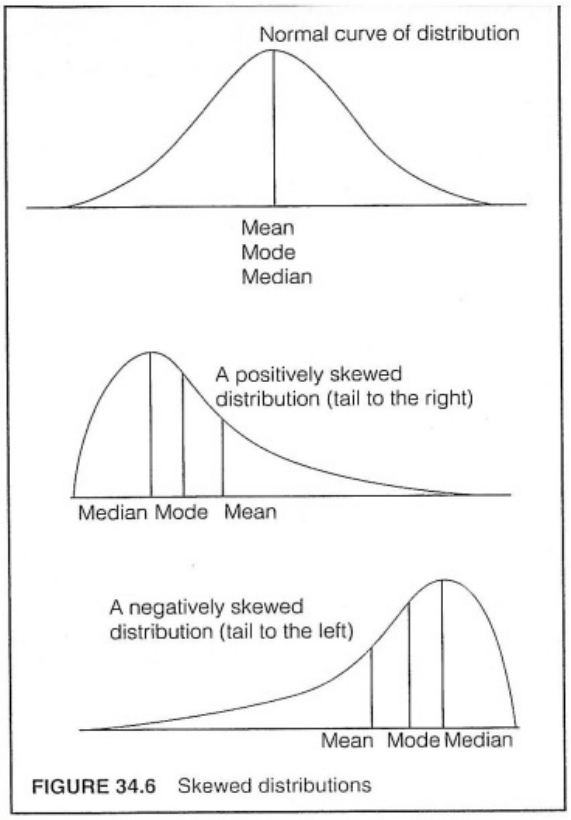


FIGURE 34.6 Skewed distributions

So let me explain.

The word **skew** is a normal English word which is rare, but not particularly technical. The Cobuild dictionary explains it well. It means "slanted when you would normally expect it to be vertical or horizontal". In the example of the one high wage earner skewing the results it means that **the results are distorted.**

(Figure 34.6 Skewed distributions comes from Cohen Manion & Morrison 2011:612).

Note, this can easily be turned on its head. Without needing to draw a curve from data, simply calculate the mean, median, and mode.

If the order is: median < mode < mean
then most of the data bunches to the left, ie to the lower figures.

If the order is: mean < mode < median
then most of the data bunches to the right ie to the higher figures.

All this is relatively easily explained. But in the TEFL context I work in, there are additional factors directly resulting from the interference of French, and this interference needs discussing. I often hear students say "**I did not get the average**". I regularly bring this example to the attention of my students, and rare it is that someone can explain why the statement is wrong, not because of the grammar, but because of the meaning of 'average'. The rough equivalent in French, 'la moyenne', includes all three averages above, with the addition of what we call in English the 'passmark'. Now, by default, in French culture the passmark is almost always fixed in advance at 10 out of 20. A student with 9/20 has 'failed to get the average'. The idea that the passmark can vary according to the subject, or, that the passmark can be fixed after exams have been set and marked, as is commonly done in Britain, could appear to students and teachers alike as a form of manipulation of the results. There is the working assumption that examination marks are objective, comparable, and reliable. Which leads us to the fourth Key, that numbers should never be reified.

Of course, in the context of International English, “I didn’t get the average” may be perfectly acceptable.

Example 8: How averages were used to distort the truth

Ginn (1999:117) gives a modern example of how different averages were used to distort the truth. In the 1980s in Britain, the Conservative government started claiming that pensioners were no longer poor, therefore had a diminishing need for state pensions.

> Older people's mean income, which was substantially higher than the median, was usually quoted, giving undue weight to the minority of well-off younger pensioners. (Ginn 1999:117).

To explain. The mean and the median were different. The government chose to use the higher figure so as to support their argument that pensioners were relatively wealthy, therefore no longer needed as much money from their state pension. The reason the mean was higher than the median was because a relatively small group of younger pensioners were well off, and this small group pushed up the mean as in Examples 3a and 3b above. In fact, as the opposition pointed out, the real solution was to increase the basic state pension, and take it back from the richer pensioners in higher taxes. This was and is an easy solution and very simple to do administratively. But, it was unacceptable politically to the Conservative government who favoured reducing taxes even for the rich. (Ginn 1999:118)

KEY 4. NUMBERS SHOULD NEVER BE REIFIED ie trusted as absolute

To avoid misunderstanding the value of numbers, several basic concepts need repeating.

a. There are limits to the accuracy of measuring tools
All measuring tools have limits as to their accuracy. I will show this by taking some common tools and looking at their limits.

Tool 1 ± 1mm	A ruler which is 30cm long. It is marked in millimetres. It is not possible to measure to a greater precision. If greater accuracy is needed, then special tools must be used.
Tool 2 ± 2-5mm	A tape measure, perhaps made of metal and rolled up. These are commonly 2-5 metres long. They are marked in metres, centimetres and millimetres, but in practice it is difficult to be accurate to the nearest millimetre once the distance gets longer than 1 metre.
Tool 3 ± 1s	An analogue watch - the type with fingers on it. If there are only two hands, then one can only read to the nearest minute. If it has a seconds hand, then time can be measured to the nearest second.
Tool 4 ± 1/100 s	A digital watch - the type with numbers on it. Commonly these measure seconds, and hundredths of a second.

b. The difference between a decimal place and a significant figure
Because of the limits of experimental accuracy in the measuring techniques, scientists usually measure to two or three significant figures (s.f.). It is rarely possible or desirable to measure accurately with a higher level of precision than two or three s.f.

Example 9: Decimal Places and Significant Figures

Number	Decimal places	Significant figures
13	0	2
13.2	1	3
13.02	2	4
0.13	2	2
0.03	2	2 (or 1 in some contexts)
133	0	3
130	0	2 (or 3 in some contexts)
Approximately 2400	0	2
Exactly 2400	0	4
2403	0	4

The concept is easily illustrated by considering the length of some furniture such as a table. With a normal tape measure, an accuracy to three significant figures is easy to obtain, ie to the nearest centimetre, but to measure to the nearest millimetre requires an accuracy of four significant figures, which, as any amateur carpenter who has tried to cut wood knows, is not as easy as it sounds. Decimal places (dp) are the number of figures beyond the decimal point. If the table of more than a metre were measured in centimetres to 3 sf then there would be zero decimal places; if the same table were measured to 4 s.f. there would be one decimal place.

Usually, significant figures are more important than decimal places.

c. The importance of calculating properly

There must be no more stated precision than the measuring tool can reasonably give, even if, due to calculations, longer sequences of numbers are actually generated.

Example 10: in the above example of lung function data, the figure 622/7 = 88.857142.

Q1. How many significant figures are there?
Q2. How many decimal places are there?
Q3. How many significant figures were there in the original data?[1].

The original data was a collection of percentages (2sf, zero dp) which indicates that the mean should be stated as 89%, **NOT** 88.857142.

Therefore there should be no more than 3 significant figures, and the long number given by the calculator must be rounded accordingly. The general rule is that **you should not finish up with an answer which is more accurate, or more detailed, than the original data.**

You should not finish up with an answer which is more accurate, or more detailed, than the original data.

A small difference is usually insignificant, and is usually caused by chance, or by factors beyond the control of the experimenter. That is why they must not be given any importance.

---

1. There are 8sf and 6dp in the answer, and there were 2sf in the original data used for the calculations.

d. Avoid 'pseudo-precision'

Pseudo-precision is where too much numerical detail is provided. It is often hard for non-scientists to appreciate that every time a measurement is made, there are margins. Using a simple 30cm ruler it is not possible to measure to fractions of a millimetre. See also, *h. Limits to precision* below. Usually scientists are happy with an accuracy of 2 or 3 significant figures. But, in biology and medicine the results can vary considerably, partly because the person changes quickly (for instance, blood pressure can change significantly from one minute to the next) and partly because the measuring instrument itself may have problems of reliability and validity. Most blood tests for instance have an error of less than 5 percent, but, some tests can have a range of 30 percent. In all these cases, it makes sense to avoid providing numbers which go into too much detail.

Pseudo-precision (JR Matthews et al. 1996:39) can be reduced by rounding. But care is needed.

Numbers which are near the cut off point need extra attention so that the rounding is done correctly. Whenever the last digit is a five, the next digit to its right needs to be identified if rounding is desired. In the above example, before 24.5 is rounded to a whole number the calculation should be done again so that if 24.5 is itself an approximation for 24.47, then to two s.f. the final figure is 24 not 25.

Example 11: pseudo-precision in distances

The feedback page of the New Scientist (4 February 2006 page 88) reported a good example of pseudo-precision. Apparently, if you go to www.memory-map.co.uk and enter your UK postcode under the "where to buy" section, you can obtain the exact distance from your home to the nearest stockist of their products. A reader of the New Scientist reports that "according to the site there is a stockist in Cardiff and it is 9.71317564206823 miles from

where he lives.". The editor comments that "evaporating a single layer of water molecules from the shop door will put these estimates out by a whole four decimal places".

Example 12: too much rounding
I once met a research candidate who collected 190 text messages. They reported the finding that there were 4000 words and 32000 characters. The candidate presented these figures as absolute - they gave no indication that they had been intentionally rounded.

Example 13: incorrect rounding
In an examination, the mean mark for a student was 9.46. One teacher saw this and rounded it to the nearest half mark, so making 9.5. The next teacher handling the marks decided that those with half a mark would go up to a whole mark, so the fortunate student went up to a mark of 10.

Extra information: rounding to the nearest half mark or the nearest quarter mark

Usually rounding follows the number system which is based on tens. Sometimes though in examinations, teachers round to the nearest half mark.

Number	Rounded figure to the nearest half mark
9.1	9.0
9.2	9.0
9.3	9.5
9.4	9.5
9.5	9.5
9.6	9.5
9.7	9.5
9.8	10.0
9.9	

It will be seen that numbers 9.3 and 9.4 are rounded up to 9.5 rather than being rounded down to 9. Similarly, numbers 9.6 and 9.7 are rounded down to 9.5 instead of being rounded up to 10.0. A similar procedure can be used to round to the nearest quarter mark.

Computer tip

Here is a trick for those using spreadsheets, which normally are difficult to set to round to the nearest half mark. Set the square to double the mark then round to one decimal place then divide by 2. eg 9.6 x 2= 19.2, which rounds to 19.0, when divided by 2 = 9.5

e. There are also **important points of style** that must not be forgotten. While it is true that whole numbers do not need the decimal point followed by a zero, if in a table of figures there are 14.1 15.3 17.4 18.1, then seventeen would be written 17.0 not 17 in order to follow the pattern and to show that the number genuinely has one decimal place, and is not a number rounded to zero decimal places. By saying 17.0 you are saying that this is genuinely 17.0 and not a rounded figure.

Since 0.5 can represent any number between 0.45 and 0.54, therefore whenever 5 is the last number, you do well to check the real number.

Wrong	**Right**
14.1	14.1
15.3	15.3
17	17.0
17.4	17.4
18.1	18.1

Extra information: Basic rules for rounding

Sometimes the details are difficult. If you are new to this concept I suggest you go on the web sometime and look for more information. There are also practical examples, with tests and answers.

There are three general rules for rounding:

Rule 1- if the remainder beyond the last digit to be reported is less than 5, drop the last digit. Rounding to one decimal place, the number 5.3467 becomes 5.3.

Rule 2-if the remainder is greater than 5, increase the final digit by 1. The number 5.798 becomes 5.8 if rounding to 1 digit.

Rule 3- To prevent rounding bias, if the remainder is exactly 5, then round the last digit to the closest even number. Thus the number 3.55 (rounded to 1 digit) would be 3.6 (rounding up) and the number 6.450 would round to 6.4 (rounding down) if rounding to 1 decimal. See Hurlburt (1994:12).

Example 14: Only round once!!

It is surprisingly easy to round more than once. This means, that when collecting and presenting data, rounding happens more than once**. If you round more than once then your figures could be incorrect**.

To illustrate this, consider the FEV1 and FVC data which were given earlier. The figures are presented as percentages and they are actually rounded figures. The reason is that the raw data is converted into a percentage (according to the height of the person). Then when we

calculated the monthly average, the mean, there was rounding on top of rounding. Let us look at this example in more detail.

FEV1			FVC		
Raw data	100% = 2.73	%	Raw data	100% = 3.22	%
2.63		96	3.14		97
2.31		85	3.41		106
2.22		81	2.88		89
2.36		86	3.29		102
2.46		90	3.03		94
2.57		94	3.2		99
2.45		90	3.22		100

	FEV1	FVC
Rounding once The sum of the raw data, divided by 7, then expressed as a percentage of the norm	88.959	98.358
Rounding twice The mean of the percentages	88.857	98.14

Even if you cannot follow the calculations, you can easily see that there is a difference between the figures when rounding happens once, and rounding happens twice. In this example, the difference is small and of no consequence. In fact, it is much smaller than the experimental error in the original medical test. Near example 10 above I said, **There must be no more stated precision than the measuring tool can reasonably give**, even if, due to calculations, longer sequences of numbers are actually generated.

f. Different versions of 'approximately' .
Scientists commonly use five different versions of approximately, the fifth one, the percentage, meriting a separate discussion below.

Firstly, there is the **rounding of numbers**, as discussed above, **to reduce decimal places.**

Secondly, there is the **rounding to appropriate significant figures**.

Thirdly, **a figure can be given with stated confidence limits**, which are commonly expressed with a plus or minus of say 5%.

Example 15: ranges in opinion polls
Opinion polls might give 60% of the British people opposed to joining the Euro. This figure might be presented as 60% ± 5%, which means that between 55% and 65% of British people are opposed to joining the Euro. Weather forecasts commonly give a minimum and a maximum day temperature, and a minimum and a maximum night temperature since actual temperatures in a day vary enormously with location.

Fourthly, another concept called **the order of magnitude is used**. This is probably the least known in the social sciences.

An order of magnitude is the power of ten. Thus, 1000 is three orders of magnitude, and 5000 is the same order of magnitude as 2000 or 8000. On the other hand, 12 000 is not the same order of magnitude as 8000. If B is 1000 times as large as A, then it is three orders of magnitude larger. Orders of magnitude are seen on special graph paper that uses logarithms, or exponential functions. They are especially useful when handling large numbers.

Example 16: Order of magnitude improvement
In the developments of the technology used to etch the circuitry of computer chips, between 1994 and 1998 there was a reported improvement of over twelve orders of magnitude (Irion 1998:30).

Example 17: The decibel scale
In hearing, the decibel scale works on orders of magnitude, such that an increase of 10 decibels is the equivalent of an increase in sound energy by one magnitude, and we perceive it as a doubling in apparent loudness. The human ear is remarkable in that it can work across a range of 12 orders of magnitude, something which the man-made measuring instruments for sound cannot do. For details on this see a basic textbook of linguistics. The human ear can detect tiny amounts of energy, and tolerate explosive amounts. This is illustrated in the table below.

Intensities of common noises and corresponding dBA values

Sound	**Sound level** Energy	decibels
	$W\ m^{-2}$	dBA
Threshold of hearing	10×10^{-12}	0
Quiet countryside	3.2×10^{-10}	25
Normal conversation	3.2×10^{-6}	65
Inside railway carriage	3.2×10^{-5}	75
Very noisy factory	0.01	100
Discotheque	0.1	110
Pain threshold (Road works, discos etc)	1.0	120
Jet aircraft at 50m	10	130
Gun shot close to ear	10,000	160

Note that the energy is presented here in the units of Watts per square metre, which is equivalent to Joules per Second per Square Metre.
Note also the enormous range of 10^{12}, which is impossible in a machine, and represents twelve orders of magnitude.

g. Use of percentages

Percentages are often deceptively difficult to handle correctly. There are four baseline principles.

1) **Specify how big the population is.** Somewhere, in the table or in the text, the divisor, which is often the population or sample size, must be specified. The larger the sample size, the more reasonable it is to present results as a percentage, or as a ratio.

2) **Actual numbers must always be presented** where the sample size, or number of subjects, is less than 50 (JR Matthews et al. 1996: 92). If need be, both a percentage and a number can be quoted eg 48 (24%). This rule exists for good reason, and fully understanding the reason may be better than memorising the rule. Whenever a percentage based on small samples is calculated, then small differences become magnified and look larger than they actually are. Using percentages increases the likelihood that a small difference will be interpreted as significant.

3) The third rule is that **if ever statistical tests are performed, then the actual absolute numbers must be used, not the percentages** (JR Matthews et al. 1996:40).

   Of course when there are variable sample sizes percentages are almost essential. For instance, when comparing the success rates in classes of different sizes, then obviously the actual numbers have to be converted into a percentage. But when this is done, the class sizes still need stating, somewhere, and small differences must NOT be treated seriously. **Do not draw conclusions from small differences**.

4) **Whenever the populations are small, comparisons become unfair**.

Example 18: success rates in two classes

Class	Success rate
A	55%
B	60%

Question: Were the results for Class B better than the results for Class A?

The five percent difference looks invitingly significant. Can we conclude that class B were better than class A? On the basis of such information certainly not, because the size of the class was not stated. If I tell you that Class A had 20 students, and B had 10 students, then for A, 11/20 (55%) passed, and for B 6/10(60%) passed then it should become obvious why we cannot say that Class B were better than Class A.

Class	Success rate	Student numbers	Number passing
A	55%	20	11
B	60%	10	6

In addition, the possibilities of experimental error in the examination are numerous, and the numbers involved so small, that even if for Class A, only 9 students out of 20 (45%) passed, then I would NOT conclude that Class B were a better class. I have seen drafts of masters level theses with similar serious errors of data presentation and interpretation, so the point needs emphasising. **Percentages can often make a small difference look bigger than it is**.

Example 19: expanding example 18

To expand the example 19. Take Class C. All of them get 11/20, which means that the class average is 11/20.

Class C	Each student gets	Total Marks	Mean score
10 students	11/20	110	110/10 = 11

But supposing one of the ten students got 1/20. What happens is that the bad mark of just one student significantly reduces the class average.

9 students 1 student	11/20 1/20	99 1	100/10 = 10

Now take similar figures, but think of 100 students.

99 students get 11/20
1 student gets 1/20

The mean is therefore (99 x 11 = 1089) + 1, total 1090, divided by 100 students gives the class mean of 10.9.

Class D	Each student gets	Total Marks	Mean score
99 students	11/20	1089	1089 + 1 = 1090 -------- 100
1 student	1/20	1	Mean = 10.9

Notice how in a large class, **One bad student has had little effect on the class mean, whereas in a small class, one bad student significantly affected the mean.**

There are more advanced techniques for handling small differences, or for differences when the sample is less than 30. The important point here is to develop a feel for why statistics work the way they do. A combination of a small sample, with using percentages, is a bad combination. **You should never draw a conclusion based on a small difference or a small sample**.

h. Limits to precision

Example 20: English blood pressures.

The English method of taking blood pressure gives the results to the nearest millimetre of Mercury (mmHg). But, **it is very hard to reliably get an accuracy to the nearest millimetre**, that in practice it is acceptable to seek accuracy at the half-centimetre level, or in other words, plus or minus 2mm.

Example 21: Analogue and Digital watches

A similar area of confusion surrounds the use of watches. The older style, the so called 'analogue' watches are those which use fingers on a circular 'dial'. They often only have two indicators: the hour hand and the minutes hand. This means they are accurate to the nearest minute. This was accurate enough for over a century. For skilled users, it is quicker to read a clock face and get the approximate time, eg 'almost ten past three' or 'just past ten past three' rather than be highly accurate. The conventions in the language used to describe time in Arabic and English conventions are similar. Unless catching a train, or timing a race, a higher level of accuracy is rarely needed and can be obtained on a closer look at the watch.

Digital watches though are slower to read, and allow people to only read a so-called precise time.

Despite the common image of science as the pursuit of precision, **the reality is that scientists are only as precise as:**

1) the tools they use permit
2) the need - too much precision takes up time and effort in calculations
3) the situation allows
4) in some cases, the skill of the experimenter permits.

A feel for the accuracy of the tools is an essential part of the training of a scientist. Every measurement (except some counts) is to some extent or other an approximation. But **scientists know the approximate size and direction of the approximation and how serious it is**. They are also usually trained to evaluate the factors in an experiment which makes the results probably too high or too low. Often there are several trends involved, and by identifying them, assessing their direction, size and importance, and then balancing them, a better estimate can be obtained.

i. Misleading claims and comparisons

Example 22. Consider the following statement. "In Bonwana in the last ten years there has been an increase in murders of ten percent".

But if the town had increased in size by 20% then though the numbers increased then the relative numbers, the rate (in this case, usually measured as murders per 100,000 per year) might have gone down.

j. How could they know that number?

Example 23. Consider the following statement. *"*Breast feeding went up 10% last year". How could they know? Was their sample big enough? Did the women tell the truth? How was the survey conducted? Was it done exactly the same way as in previous years? Then there is the question as to what is meant by "breast feeding". Many mothers try for a few weeks then give up. Does the woman

who tries for a week and then gives up count the same as someone who breast feeds for a year? We really must acquire a balanced appreciation for the validity of numbers.

k. Approximations are true though estimations are not necessarily true

It is easy to think that because an approximation is made, then it is not true. In other words, some think that only a highly precise measurement is true. It must be categorically stated that **an approximation is a true statement, but true within clear limits**: it is limited truth, truth as far as it goes. An approximate statement is true but not very precise. Approximate statements are true, and precise statements are true.

l. The importance of visually inspecting numbers and making sure that they are realistic

I once saw a candidate who told me that the average length of the text messages they had collected was 168 characters. Now I do not personally know much about SMS, but I had heard from others that there is a limit to the number of characters of 160, unless the local system has been modified. The candidate gave no indication in the thesis that this figure of 168 characters was unusually high and this was because of local conditions. A quick check on the internet confirmed that the normal limit applied locally.

The problem was, the candidate had done the calculations and accepted the resulting figure even though it was a wrong result. I saw that the resulting figure was impossible therefore there must have been something wrong with the calculations. This shows that it is the responsibility of the researcher to make sure that the results are realistic – that they correspond to reality, and where they do not this means there is something wrong. Also, candidates cannot rely on others to do the calculations for them. The candidate is ultimately responsible for the data presented.

KEY 5. IT IS RARELY ENOUGH TO STATE THE MEAN

Example 25. How can 20°C in winter in Cairo be considered cold, whereas in England in the summer it is warm?
Apart from the psychological attitude to temperature, the adaption of the body to different seasons, and physiological factors such as wind speed, and relative humidity, there is a problem inherent in the way the data in world temperature charts is presented. In particular, there is the question of the number of hours at which the temperature stays at 20°C, which in Cairo in winter is likely to be only two or three hours, then dropping swiftly in the early afternoon to 10°C or lower, whereas the same English summer temperature is likely to last most of the day, dropping only slightly at night.

The example nicely illustrates the usefulness of stating ranges of temperatures. It is useful for some people to know the highest and lowest expected day temperature, and the highest and lowest expected night temperature.

Stated bluntly, **a mean without an accompanying indication of the range of values, is almost always almost *mean*ingless** (excuse the play on words!!)**.** The statisticians have provided several ways of expressing that range. Conventionally they include the **Standard Deviation**, (a formal way of expressing a plus-or-minus) and, if appropriate, the Standard Error of the mean. These tests are well explained in other texts that are too advanced for this book. Other ways are possible, such as the mean with a simple plus or minus and a value, this being useful where the absolute minimum and the absolute maximum are known and are interesting and where standard deviations are not understood. Another way is to state the range of values that would cover 90% of the sample, the so called percentile lines. These are commonly seen on growth charts, for expected weights and heights for boys and girls in their first five years of life. I have found that most of my students know about these charts, and therefore they are good starting point for discussing ranges of the mean.

Extra explanation. Discussion of the ranges in weather forecasts

Weather reports have to simplify. One way they do this is to quote the highest temperatures for the day, without saying how long these high temperatures last.

In Cairo, in the winter, it is likely to be 20°C only from noon to 2pm. The night temperature may well be 10°C.

In Britain in the summer it is likely to be 20°C from 10am until 4pm, and the night temperature to be 18°C.

Therefore, to state 20°C without saying for how long it is that temperature is a little misleading. In Cairo in the winter it is 20°C for only two hours, whereas in Britain in the summer it is 20°C for six hours.

The other 'range' to notice in the above example is the difference between the day temperature and the night temperature. In the winter, this difference is likely to be greater than in the summer. In the example, the range of temperatures for Cairo in the winter is 10°C whereas the range of temperatures in Britain is 2°C.

Finally, not given above but frequently given in the more detailed weather forecasts is the minimum and maximum day temperatures, and the minimum and maximum night temperatures. This can either be expressed as

20°C minimum 18°C maximum 22°C, or

20°C ± 2°C. [Note, the symbol ± means 'plus or minus].

People who use weather forecasts regularly are able to interpret this for their local conditions. For instance, the minimum day temperature might be correspond to the coolest room in your house, and the maximum temperature might be very close to the temperatures you measure in the shade on your terrace. Similarly, in winter, of great interest to those with a garden is the predicted minimum night temperature. Some areas in gardens are more exposed than others. In other words, gardens typically have areas which are protected from cold and other areas where plants are easily damaged by cold weather. When weather forecasts give a range of night temperatures, the home gardener is able to protect some plants on cold nights.

KEY 6. CAREFULLY CHOOSE YOUR SAMPLE

The good researcher
The good researcher begins by defining a whole group – the population. The 'defining' probably means a list of features, and can mention what is NOT included**.** Then the researcher works out what a representative subgroup – the sample, would look like. **In this way, it is fair to infer that the results from the sample apply to the whole population**

Few researchers can study the whole group. Therefore they end up selecting part of the group. The good researcher begins by defining the whole group – the population, then working out what subgroup – the sample, to use in their work.

1. **Population and Sample**. Rowntree (1981:20)
   Population refers to all the cases or situations that a statistician wants his work to apply to. Often it is people, but it could be all the exam results, or all the yoghurt pots being made in a factory.
   Sample It is usually logically or practically impossible to test everyone or everything, therefore some of them are tested, and it is hoped that any inferences drawn, any conclusions, could be applied to all. In other words, the results from a sample should be generalisable to the whole population. This is the question of validity.

   How safe is it to generalise? First there is careful identification of the whole population, then, by various means, a sample is taken. In addition to careful identification and management of the variables, statistics gives us some tools. In particular, statistics enables us to quantify (put a figure on) the probability of error. "...the reliability of the generalization will depend on how well the sample mirrors the population. Is the sample truly representative of the population?" (Rowntree 1981:21).

   Example 1
   Rowntree (1981:23) gives the example of a group of 12 year old English boys who apparently learned more French when taught by the 'conversation' method than did a similar group who learned from a textbook. Then he asks, to which group

could this result be generalised to?

a) all 12 year old English children
b) all 12 year old English boys
c) all 12 year old boys
d) all boys
e) all learners.

In other words, of which population was the sample the most representative? Which population was the population the least representative of? See Answer 1 at end for a commentary.

2. **Samples choosing themselves**
Example 2
Rowntree (1981:23) gives the example of the gunners in the RAF bombers. These gunners were asked, on their return from combat missions, "From which direction were you most frequently attacked by enemy fighters?". The majority answer was "from above and behind".

From this can we therefore generalise and say that most enemy attacks on bombers came from above and behind? It is very tempting, but there is one more fact or circumstance that means we cannot generalise. Think about the answer. See Answer 2 at the end of this section.

As we see here, there is a real danger of 'opportunistic' samples as in the RAF gunners, NOT being representative.

3. **Random samples** (Rowntree 1981:24)
The researcher starts off with an idea of what his total population is. He then selects a sample he believes will fairly represent it. In order to be fully representative, the sample must be chosen randomly from the population. By random we mean that every member or item in the population should have an equal chance of being selected**.** In practice it is often very difficult to achieve. For instance, selecting people randomly in the street. The tendency is to approach people who look approachable and who are not too busy. Therefore to avoid bias some kind of mechanical selection method should be used. For instance, stopping every tenth person going through a gate, or using a table of random numbers.

4. **Stratified random sample**. (Rowntree1981:26)
   Often, this is the best method. We state in advance the number of men and women, adults and children, rich and poor etc that we need in the sample (and these numbers usually reflect their proportions in the whole population). Then we actively seek members of each subgroup, and pick them randomly.

KEY 7. USE SIMPLE PROCEDURES

To take a common procedure for a teacher, the collation and inspection of marks of several classes that have been taught the same curriculum by two or more teachers. What I do, as I mark scripts, is to build up a frequency table of how many get each mark. In Tunisia, we mark out of 20, which gives a convenient list of 20 values down the page. If working to percentages, these could be set in ranges of for instance 0-4, 5-9, 10-14, etc. I build up the numbers getting each mark, using the conventional gate system, and so build up a visual frequency table of the number of students who got each mark. On most exams I would expect to see a peak somewhere in the middle, though where it is expected to be depends on the system of teaching you work under. One year I found for instance that I was giving more even numbered marks than odd numbered, and the curve was very bumpy. This necessitated remarking, but I spotted the trend early enough to make that a feasible endeavour.

Having finished the marking, I could then enter the results into a database and analyse them. I could, but I do not waste my time. First I calculate the class mean. This is easily computed given that I already know how many students got each mark. I therefore multiply the mark by how many students get each mark, then add up the totals and divide by the number of students. I also count the number of students who have passed outright, and the number of students who got 8/20 or more. I choose 8/20 because in the system I teach under I expect around two thirds of students to get eight or higher. I then take all three measures, and compare classes, looking for anything unusual. Where double marking is not practiced, these inspections are particularly important. These calculations give very meaningful information and are speedily carried out using pen and paper mostly, with some use of a simple calculator.

Simple data exercise.

The following are actual data from an examination. There were 60 candidates, so we write n=60.

5	5	12	14	10	7	14	7	15	14
5	10	5	12	10	4	11	8	13	1
3	4	12	9	15	13	7	16	1	10
14	8	12	10	15	13	12	14	9	14
16	11	8	15	12	16	3	10	14	3
10	10	14	10	14	8	8	10	13	15

Mark	Tally system	Total number of students with this mark	Total of marks awarded
0	/	1	0
1		0	0
2		0	0
3	///	3	9
4	/	1	4
5	////	4	20
6		0	0
7	///	3	21
8	/////../	6	48
9	///	3	27
10	/////../////	10	100
11	//	2	22
12	/////../	6	72
13	////	4	52
14	/////..////	9	126
15	/////	5	75
16	///	3	48
		n=60	total = 624

From this the mean can easily be calculated, and the median and mode identified.
mean = 624/60 =10.4
median = value of 30-31 candidate. This is easily identified by counting up the total number of students from the beginning. Remember, they are by now in rank order, from least to best. The result is obvious – the 30th and 31st students each obtained 10. The mode is also obvious as 10, with another peak at the score of 14.

I then find how many passed, ie had a score of 10 or more. This is 39 students, therefore the pass rate was 39/60 = 65%. For interest, I also like to see how many obtained 8 or higher. This was 48, meaning 80%. A very satisfactory set of exam results.

Another example for practice

2	7	6	10	10	10	12	11	5	6
14	4	4	10	4	11	5	10	15	1
3	6	11	16	14	10	1	16	3	13
8	10	4	14	13	4	12	11	15	11
14	14	5	10	7	8	15	7	7	7
7	11	7	10	3	1	11	14	10	10

KEY 8. IDENTIFY ALL POSSIBLE EXPLANATIONS FOR THE RESULTS Then argue for the most probable

Example 26: A tale of four classes

It is all too easy to make hasty and quick judgements - to jump to conclusions. This human tendency must be resisted. One year, with a set of examination results I noticed there were two bad classes and two good classes. We had two teachers each teaching two classes so, could it be that one of us was significantly worse than the other? That conclusion could easily be drawn, and in some educational establishments where the pressure is on a teacher to get results, such a conclusion regretfully probably would be drawn. There are many other possible reasons for the difference, and when using the example in the course of teaching, I usually get students to brainstorm as to other **alternative scenarios**. In fact in this case, to my relief I found that each of us had a good and a bad class, but **even if one of us had had two bad classes, that would not in itself have warranted a comparative judgement of us as teachers**.

As we know from Key 1, the context is important. It turned out that the two bad classes had their lessons on Friday afternoon and Saturday morning respectively. When discussing this example I now like to ask students, **how could evidence be found that would test this idea?** It would still be too easy to explain away the results as being caused by the lesson time. The observations as they stand are suggestive, but not conclusive. One possible way would be to change the timetable so that the good classes received the Friday or Saturday time slot, and the bad classes received the time slots at the beginning of the week. If it were simply the time slot that was affecting the results, then this should show up in the next tests. The idea is attractive, but would not necessarily work, for there are still other factors that could over-rule the change. Before you read on, or before you give the game away to your students I suggest you stop and think what else could be happening.

---------- STOP. Think! ----------

As soon as you start changing classes, you introduce a new variable that was not originally there. The moment you noticed that the difference existed, then there will be a natural tendency to teach the poorer class differently, either teaching better, or teaching with less effort because of lessened motivation. Simply by changing your attitudes, the poor classes could improve or get worse. Simply by the students themselves noticing they have done badly, there could be an improvement or a further slide into mediocrity.

So what can be done? It is not ethical to leave poor classes without extra help, so the question is left open. Another method would be to look at similar scenarios in other subjects. Is there a trend for poorer results, regardless of the class and teacher, for those classes that are taught at inconvenient times? Is it a trend that repeats from year to year? It might be possible to look into historical data and come to conclusions. If it exists in other subjects, and if it could be shown to have existed, un-noticed, in the past, then the evidence would be quite strong that time of day has a significant influence on examination success. By using historical data you are removing from the picture the variable of the teacher noticing that the time of day is important.

The discussion could go on, for this example well illustrates the problems of careful inferencing from numerical data, and the discussion can all take place with very little mathematics involved. The key concept here is what the textbooks call the 'variables', ie the factors that are involved in the situation. One of the keys to designing research is to identify the variables, and account for them. Brown (1992) has a very accessible introduction to this. When teaching, I also include the related topic of the well known confusion between 'cause' and 'correlate', a subject that is usually best explained in the books on critical thinking, such as Warburton (1996), Thomson (1996) and that old classic, Thouless (1974).

A sub-theme here is that the simplest explanation is usually the most likely. Having listed the possible explanations, a scientist frequently seeks to eliminate the various possibilities. Some are quickly ruled out, others will need checking and testing.

KEY 9. FALSE POSITIVES AND FALSE NEGATIVES

Language point

Some people initially find the language of 'positive' and 'negative' to be confusing. Some examples may help.

Positive and negative

1. When you pass an examination the result you get is positive. When you fail an examination the result you get is viewed as negative.
2. When you do a blood test for a disease such as AIDS, then the laboratory report will be positive or negative. If it is positive then you have the disease; if it is negative then you do not have the disease.

False positives and false negatives

Very few tests are perfect. Sometimes the result is wrong. When the result is wrong we call it a false result. Since there are two types of result, a positive result and a negative result, it is possible to have a false positive result or a false negative result.

Example 27: Aids tests

The concepts are probably familiar to students from what they know about diagnostic testing for Aids, or screening tests for cancer. The challenge is to encourage students to apply the concepts to their own situations and studies. **With disease, a test result that is positive when it should not be (a false positive) is alarming, but rarely fatal**, since people testing positive are usually re-tested to 'confirm' the result, the word 'confirm' being used in the unusual sense of 'check'. The second test and maybe even more tests are used to replicate and verify the results. On the other hand, a result that falsely says there is no disease, that falsely says there is no danger, (a false negative) could be fatal. A false negative in cancer screening could mean that the cancer goes undetected for a long time, and could mean the difference between early and late treatment, and even between many more years of life, or an early death. Screening programmes seek to reduce both

types of error, and especially to reduce as much as possible the number of false negatives rather than the false positives, since it is the false negatives that have potentially fatal consequences.

		Reality	**Treatment**	**Consequence**
T E S T	**Pos-itive** (Type I or Alpha error)	True	Treatment that is needed	Probable life
		False positive	Treatment given needlessly	Possible death due to the medicines or surgery. All surgery has a risk to it.
	OR	True	Do nothing	No danger. Life
	Neg-ative (Type II or Beta error)	False negative	Do nothing	Disease gets worse and remains undetected for a long time. Possible death due to undetected disease. If another test is done later, it might then be too late to treat the disease

Extra information
A false positive is also known as a type one error and is sometimes called Alpha. A false negative is known as a type two error which is called Beta.

Example 28: Appendicitis
As you would expect, real situations are often far more complicated than this. Doctors are frequently in a position when they have to evaluate the dangers of their decisions. Take for example the question of appendicitis. This is when the appendix becomes inflamed for various reasons. If it is not dealt with then it can burst thus spreading bacteria from the intestines all over the abdomen. The bacteria are safe and usually beneficial when they are in the intestines, but once into the usually sterile abdomen they can cause horrific problems and even death. Therefore, appendicitis is a potentially fatal problem.

The trouble comes in the diagnosis. Fortunately in many people the symptoms are clear and with a good conscience a surgeon will operate - often within a few hours of first seeing the patient. The trouble is that the symptoms presented by a patient are not always clear: it is not always obvious that appendicitis is a problem, or it is something more banal such as soreness or even nervousness that will go away without treatment.

There is always a risk whenever someone has general anaesthetic and has surgery. In an otherwise fit young person who has no other medical problems then this risk is relatively small, but it still exists and is significant. Obviously surgery should be avoided unless it is really needed. Therefore there is a slight risk that a false positive will lead to treatment that is fatal. At best a false positive leads to discomfort and pain due to surgery.

In the case of genuine appendicitis, surgery is the treatment of choice. There is a risk that something will go wrong, but this risk is small compared to the risk of not operating, because left without treatment the appendix could burst and lead to death.

I have by no means outlined all the risks and factors to be considered with appendicitis, but I have shown enough to illustrate how a doctor has to take the risk factors for each possible choice into conclusion.

Example 29: Examinations
What happens in examinations? Something similar happens. Students pass or fail. Some students deserve to pass and some students deserve to fail. The problem comes with students who pass who do not deserve to (false positives) and students who fail

who really should have passed (false negatives).

First Session results			**Consequences**
Pass	Deservedly	Positive	Go to next year and do well
	Undeservedly (Pass due to luck)	False positive	Go to next year and do badly
Fail	Deservedly	Negative	Go to resit and fail again
	Undeservedly (Fail due to luck)	False negative	Go to resit and pass

Example 30: Language testing
This is a good time to ask the question, **in language testing, which error is more significant, and to who? The false positive or the false negative?** Is it better to let pass those who should fail, (the false positives) or to fail those who should pass? The answer involves applying Key 1 again, for **the circumstances of the test are critical**. Usually, there is concern not to fail students without good reason, therefore the false negatives must be worked against. But in the case where the main examinations are followed by resits, a few weeks or a few months later, then failure in the first session is not a disaster – the consequences are that they are given a second chance to succeed. On the other hand, those who succeed are not retested. Therefore students who are allowed to go up a year before they are ready (false positives) may well suffer badly, and it would have been better for them if they had been failed in the first session and obliged to do some more work and resit the examination.

When examinations are set in two sessions, with those who fail at the first attempt being able to resit within days or weeks, the concern in the first session will be to minimise the number of students who undeservedly pass. The worst that can happen in this scenario is that a student has the pain of re-sitting. But if poor students are allowed to pass, then that is serious. In the resit session, the concern must be the reverse, there must be no false negatives. The consequences for failure in the second session

(resit) are that a student will have to wait another year before being allowed to try again, and it may well be that they have to repeat the whole year. Because this consequence is so large, examiners must make all reasonable attempts to make sure that they do not fail someone incorrectly.

But even that is not the whole story. Once again there are other factors. In some universities in Britain I have heard the reasoning that since students are being given a second chance, and since they have more time to prepare for the examination, and since they have already had some examination practice, then the passmark will be raised, for instance from 10/20 to 12/20. But even if this reasoning is applied, there will still be a concern to avoid incorrectly failing students who deserve to pass.

Scenarios where false positives are undesirable

There are scenarios where it is more important to fail a large number of people than to accept anyone who is remotely lacking in competence. There are two obvious examples.

Example 31: The driving test
The biggest concern of an examiner is that the candidate is safe enough to be allowed to drive a car without supervision. Failure for the candidate means more training, more practice, and another examination. All this takes time and costs money, but is better than wrecked cars and injured people. Therefore, false positives must be avoided.

Example 32: Appointment of senior staff
In an institution there might be 20 vacancies and over 50 applicants. Each applicant is assessed on their merits. Only if there are a lot of good candidates will it be necessary to choose the best among them. It sometimes happens that only some of the vacancies, maybe 15 of them are filled, and people ask why, given that there are 20 vacancies, only 15 were filled. The answer is simple once the idea of false positives and false negatives is grasped. In an institution, senior staff rarely become junior staff. If someone becomes a senior member of staff and in the end they are not competent, then it is very difficult to return them to a lower level. Therefore those doing the selection must make sure that their decisions are right: no false positives can be allowed, though false negatives can be accepted.

KEY 10. THE MOST DEMANDING METHOD SHOULD BE USED

One of the most neglected principles of research is that **the design of research should be such that the data is collected in such a way that the hypothesis is worked against.** I have never seen this principle stated, but it is common enough in the hard sciences where I began my career. It is in fact another way to avoid a false positive. One must make a reasonable attempt to show that the results obtained genuinely did happen due to the reason you are studying. Therefore, you arrange the data collection in such a way that the other factors work against you. If you still find an effect, having worked hard to make sure you would not, then the effect you find is highly likely to exist and be significant.

Example 33: Recognition of names of chemicals
The first example involves the problem that **the experimental procedure sometimes sensitises people to the problem being studied**. When you give people a test or a questionnaire or some other item to do or think about, then the next time they do something similar they will be better prepared, they will have had more practice. This is like pre-warning, pre-arming, or sensitising people.

In my research I had already established that the morpheme order within names in organic chemistry in 1989 was different in English and in French. For instance,

English	1,1-dichloro-4-methylepenta-2-yne
French	dichloro-1,1méthyl-4 pentyne-2

I was interested not only in documenting the differences, but in making some assessment of how significant they were to students. I wanted to see if students of chemistry in French could understand the English names. I therefore wrote a recognition test.

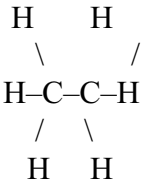
The question was, which test should I administer first? Whichever was given first would give extra practice to the students, and would therefore sensitise them to the second version. To complicate matters, I only received permission from the

teachers to administer the questionnaires in one time slot, not several weeks apart as I wanted. My hypothesis was that English names were sufficiently different that they would prove more difficult to interpret than French names, despite their French counterparts. I decided I must use the order that worked against my hypothesis, and which would sensitise the students. If despite the sensitisation there were differences, then the results would be more convincing. Therefore I administered the French version first. No one taught me to do this, and I did not consult with my supervisor. I simply applied common sense reasoning.

The problem of indirect measurement – the intervening variable

In the example above I decided to test for recognition by asking students to draw the structure. This needs some explaining. With chemical formulae there is a code which is in three parts. There is the name, in words, which can be written as a formula, or drawn as a shape.

Note that, conveniently, the formula and the drawing stayed the same across languages, but the name changed.

Name	Formula	Shape
Ethane	C_2H_6	

I chose to test the recognition by asking them to draw the shapes. I was a former teacher of chemistry, and had observed these students in class, and judged that the students would have no problems at all drawing the shape. This means that the ability to draw I thought was good, therefore did not interfere with the recognition. This is an example of how in research you sometimes have to measure something indirectly.

I made two versions of the test, one using French names, the other using English names but presented in a different order. The **protocol** (instructions and how the test was administered) was in French in both cases.

Example 34: Lister and antiseptics

Fowler (2003) reports on the early use of this principle. Lister (1827-1912) is famous as the man who discovered antiseptics. He found that by liberal [generous] use of 'carbolic acid' on the hands of surgeons, and on the equipment and dressings used in surgery, that he significantly reduced infection rates in hospital. He obtained good results "in spite of the fact that, in order to show his confidence in antisepsis, he continued wearing dirty operating clothes and refused to wear gloves". Lister deliberately only used antiseptics, he did not use other methods like cleaner clothes. He made it difficult to report a result, therefore when he did, it was in spite of the factors against him, therefore was extremely significant.

Example 35: Questionnaire question

In an MA thesis questionnaire, the following question was asked to teachers.

> If you were not trained in the use of the Internet in ELT, you would probably: (tick as appropriate)
>
> Not bother knowing about it at all
>
> Request the inspector to organise training sessions on the topic in question
>
> None of these (please explain).

In the thesis defence, I accused the candidate of having used a 'leading question'. In particular, I disliked the option Request the inspector to organise training sessions on the topic in question. I as the examiner, as will be explained, was wrong, and the candidate to his credit successfully defended himself against the accusation. My reasoning was that the first two possible answers were not the only ones - the candidate could have offered more choices. Most people will take the easy path in a questionnaire or interview, therefore are likely to be influenced by and opt for one of the two easy choices given. The third choice, "None of these (please explain)" is a difficult one to take. Therefore, either the question should have been entirely free for response, for instance, "If you were not trained in the use of the Internet in ELT what would you do?" or, at least two other options should have been provided, including one referring to self study.

The candidate replied that he had deliberately NOT presented 'self study' as an option, because he did not want to suggest this answer! Instead, following this Key 9, he wanted to make it hard for the 184 teachers who did the questionnaire to come up with the answer that self study was an option. This was a very successful

defence. He showed that he was in control of such factors, and was able to balance them and choose how best to obtain the data he wanted. The example is a good one for this book.

Example 36 - see Example 26: a story of four classes
In Key 8 the question of classes taught on Friday afternoons and Saturday mornings was discussed. If the time of day is a really significant factor, then it will be bigger than other factors. One way to show the importance of this factor would be to announce to classes held at those times that the factor existed, and that the whole class would make an extra effort to do well. If, despite noticing, and despite making more effort, the results were consistently and significantly lower than other classes, then this factor would be acceptable as very significant. Note, I would be very cautious about accepting results from only two classes in one year. I would want data from at least 10 classes over several years.

On the other hand, if these classes [Friday afternoon and Saturday morning, and informed of the problem] were to do consistently better than the other classes, then it would convincingly show that the extra motivation (which is linked with noticing the time of day as unhelpful) is a more important factor. It would show that despite the unfavourable lesson times, good motivation can be more powerful.

KEY 11. COMPARE LIKE WITH LIKE ie make fair comparisons

Example 37: Excluding abortions from the "deaths at birth" statistics leads to misleading comparisons between countries
An example of unfair comparisons happens all the time, when the number of children who die at birth is presented. The data excludes abortions. If abortions were included, then probably Europe and America would have a much higher death rate than poor countries in sub-saharan Africa.

Rough figures will do. The figures vary somewhat due to the source, the year, and the problems of collecting accurate information. Europe has around 190 abortions per 1000 live births. In the Eastern Europe and Russia, the abortion rate is over 1600 abortions per 1000 live births, ie the abortions outnumber the live births. The estimated abortion rate in 2008 was 38 per 1000 in Eastern Africa, 36 in Middle Africa, and 28 in Western Africa. Compare this with the infant mortality rate per 1000 live births. In Africa this is around 86, and Europe it is around 12.

rates are per 1000 of live births	**Abortion rate**	**Infant mortality rate**	**Deaths at or before birth**
Europe	190	12	202
Africa	38	86	124

This means that abortion is a bigger cause of death in Europe than the infant mortality rate, and that Africa, for all its poverty and high death rate of babies, is far ahead of Europe because the abortion rate is low.

Example 38: The native speaker standard
It is all too easy to compare the incomparable, or to make an unfair comparison. Therefore whenever comparisons are made, it is a good idea to first of all make a list of the similarities, and a list of the differences.

For years, the standard of comparison for L2 has been the Native Speaker without any L2. The comparison has now been

acknowledged as being unfair. (Widdowson 1994, Wiley & Lukes 1996). A better comparison is other multi-linguals, since multi-lingualism is probably more widespread than the monolingual NS. The use of controls (See later Keys) is based upon fair comparison linked with a full inventory and appraisal of the variables.

Qualitative investigations and studies of the situation have an important role in enabling fair comparisons to be made, so that a) the way variables are inter-related is understood, b) the experimental data are interpreted in their context, and c) valid comparisons and generalisations are made. Burgess is but one good example of combining the qualitative with the quantitative, and he has published both textbooks of methodology, and provided case studies (1982, 1983, 1984).

Commonly we teach students who are doing their first piece of serious research, to specify the question, then to ask what data could be collected to answer the question. Another, more time consuming but rewarding approach is to study a situation well using methods and approaches from ethnography, and obtain a thorough detailed grasp of the populations being studied before proceeding to study one question in greater depth. In this approach, a situation is studied, initially, with a fairly open mind, until it is known well. Then more directed research can take place, and can be informed and directed by an intimate knowledge of all the many factors in the situation.

Whether one begins with a theory, or begins with a situation, this statement is true. The variables, the factors influencing the data, must be identified and accounted for, so that like can be compared with like and the data collected can be interpreted in context. **Early on in any research students should then make a list of all the variables involved in their research which could influence the results. The aim is to show that your result is caused only by the factor you have identified.**

Summary tables are useful, but the writer should always introduce them, then comment on the more interesting points in the table.

Example 39: To show the amount of detail required for fair comparisons
Similarities between the English and the French schools (Lowe 1992 p2.7-9)
In view of the comparisons I will make between French and English, and in particular between pupils in the two schools it is important to identify the similarities and the differences between the two groups. The similarities are:

1) All pupils had a similar primary school education, similar in the sense that no English was taught, and they all followed courses for six years in Tunisian government primary schools. Unlike Britain, this meant a common syllabus with state textbooks.
2) All students had, in accordance with the regulations for entry into a pilot school, never repeated a year, had a maximum age of 13 in October of the year of entry, and had succeeded in the 'sixième' with a mark of at least 80%. (Lycée Ariana 1984). Students admitted to the French school in 1988 had a mark of 19.2/20 to 17.8/20 (96 to 89%) and to the English school of 18.55/20 to 17.6/20 (92.75% to 89%). (La Presse 1988).
3) The students came from all over the country and were at least 80% boarders.
4) The maternal language was the Tunisian dialect of Arabic.
5) The syllabi for all subjects in the lycée were the same. This also includes the subjects taught. The only exceptions to this were:
    a) Students at the English school had a term of intensive English before starting biology and mathematics in English. Students at the French school, having had French in the primary school were able to continue mathematics and start biology in the first term. By the time of the fifth and sixth year, where most of my research took place, these initial starting differences had probably evened out.
    b) In the fourth year, one class of students were allowed at the French school to proceed towards the Arts baccalaureate. This option did not exist at the English school: right from the first year of entry all students were oriented towards the Maths-Science baccalaureate, and the few students who needed to do the Arts baccalaureate were encouraged to leave and go to a normal lycée.

Example 40: To show the amount of detail required for fair comparisons

c) At the English school French was taught as language only, ie no other subject used French as the 'vehiculing' language. whereas in the French school, English was taught as language only.

6) The same subjects were taught in Arabic ie history, geography, art, music, home economics, religious and civil instruction, and games.
7) All students were taught classical Arabic.
8) Students studied mathematics, physical science (from now on referred to as physics or chemistry as appropriate, as the lessons were divided up this way, but taught by the same teacher), natural science (biology), technology and computer studies in the given vehiculing language of the school, ie French at the French school and English at the English school.
9) The school routines were similar, for instance testing five times a year.
10) All students sat the same baccalaureate examinations, the first ones for the schools being in June 1990. The examinations were national ones, and an English translation was prepared for the students at the English school. 1
11) Both schools had non-Tunisian teachers, and received foreign aid that included books and computers.
12) Science subjects were divided up into time for practicals, in which half the class was taught, and theory lessons to the whole class.
13) No fees were payable, and poor students received financial help to enable them to stay in the state provided hostel.
14) Students came from all over the country. The information in Appendix 15 shows this in detail for the English school and there is no reason to doubt that the French school was not similar.
15) There was a high teacher to pupil ratio, especially when it is known that much of the time in science was spent in half classes, the full class being 30 students or less.

French and English school differences. (Lowe 1992 p2.10-11)

	French school	**English school**
French language	Taught by French nationals. Host language in which sciences and mathematics were taught	Taught by Tunisians Taught as language only
English language	Taught by Tunisians as language only	Taught by Tunisians, but was the host language for mathematics and sciences
Sciences/Maths	Taught in French, up to 14 teachers from France involved	Taught in English, up to 5 teachers from Britain involved
Numbers of classes 1987-8 (Classes were a maximum of 30 pupils)	Year one: 3 two: 3 three: 6 four: 4 + 1 Arts five: 5 + 1 Arts Total: 21 + 2 Arts	Year one: 3 two: 3 three: 5 four: 4 five: 4 Total: 19
Resources	School given to Tunisia in its entirety by France Aid from France	New School, built by Tunisia Aid from Britain and America
School location	Center of Tunis, near French Culture Centre	In Ariana, 5km North of center of Tunis and British Council

Science Textbooks	National, in French	In English, translations/adaptations of national French texts
Other books	Bookshops have limited but existing stocks of non-official textbooks in French	Few books available in English outside school library
Tunisian Science Teachers	Degree, through French	Degree through French + one year in UK learning to teach in English (one teacher had been to America)
Milieu outside schools	Arabic, French	no English
Approaches	Tuniso-French	Tuniso-French, and Anglo-Saxon
Inspectors	NOT necessarily the same set for each school as the two schools were in different administrative regions	
Studies after baccalaureate	The majority will probably do degrees in Tunisia through French. A select few will get scholarships abroad.	

Above I give an example from my own research. The similarities list and differences list come in the context of a chapter that explained the broad lines of the Tunisian Education system, and some of the history of the pilot schools. Extra detail was footnoted and is not included here. I also had several paragraphs on differences between the hours of study, the detail going into an appendix, and a summary going into the comparisons. The example above shows the sort of detail that must be documented, and documented in a systematic way. The documentation provides the basis for fair comparisons, and enables the reader to check if you have fairly interpreted your data. They provide a context, which needs to be stated explicitly.

Introduction to keys 12-19

In keys 12-19 I present a series of over-lapping and interconnected ideas. You may need to read these keys several times in order to understand them, since each one contains several complex ideas, and is related to the other keys.

The core problem is that in studying human phenomena we are dealing with complex situations and trends and people. There are always many variables which interact. Yet, we usually want to focus on one variable, and to describe and measure it.

In conducting our investigations we want to make sure we are concentrating only on the factor that interests us. The trouble is, the other factors will not go away – they are still there, and they must somehow be accounted for and separated out. At all costs, any difference we see really must be attributed to the correct variable. In addition, there is often variability within an individual as well as variability between individuals.

KEY 12. MANAGE THE PROBLEM OF NATURAL VARIABILITY BETWEEN PEOPLE

One of the complicating factors in doing research on people is that there is often a wide range of individual variation. This statement applies to almost anything, from physical items such as shoe size, height, and weight, to more complicated factors such as motivation. Whenever there is a group there will naturally be a range within the group.

What makes this so important and so complicated is that this natural variability may be quite large: so large that changes due to an experiment may not be detected. The problem is made even worse when this variability is added to the problems of selecting a small group which is fully representative of the larger population.

Take a large group of people. Take any natural feature, such as height. In adults, there will be a range of values. Some adults will be very small, some adults will be very tall, and most will be somewhere in the middle. Statisticians call this "the normal curve".

Similarly, in a fair examination with a lot of students, there will be a range of marks. Some students will do extremely badly. Some students will do extremely well. Most will come somewhere in the middle.

Various methods are used in an attempt to take control over a situation with wide individual variation, and these will be discussed below.

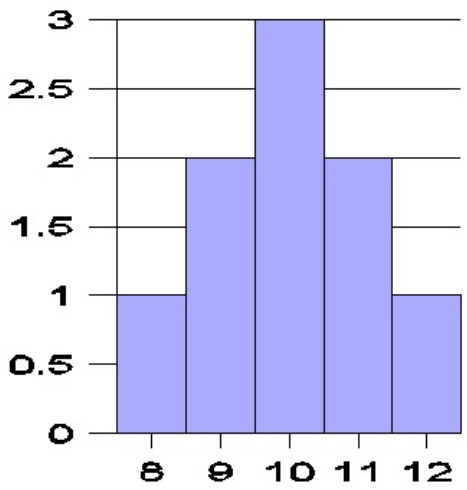
The problem that any given group or sub-group may not be fully representative of the whole population, and the variation within that population. Representativeness as a problem often exists because of the use of small groups. A small group is more likely (though not always) to be unrepresentative than a large group.

Whenever small groups are used there will always be the problem of the unrepresentativeness of the group, therefore the findings from a small group can only with difficulty be fairly extended to the whole group.Students need to appreciate the following.

Example 41: Nine classes of fifty

If you take an intake of 450 new students into a university to study English, then divide them up into nine groups of 50, and test them on any feature of English you like, preferably using some form of multiple-choice examination (so as to increase objectivity - the reliability of a test), then one group will do extremely well and one group will do badly. In fact, the results should look something like this:

Class average	Number of classes with this average
8	1
9	2
10	3
11	2
12	1



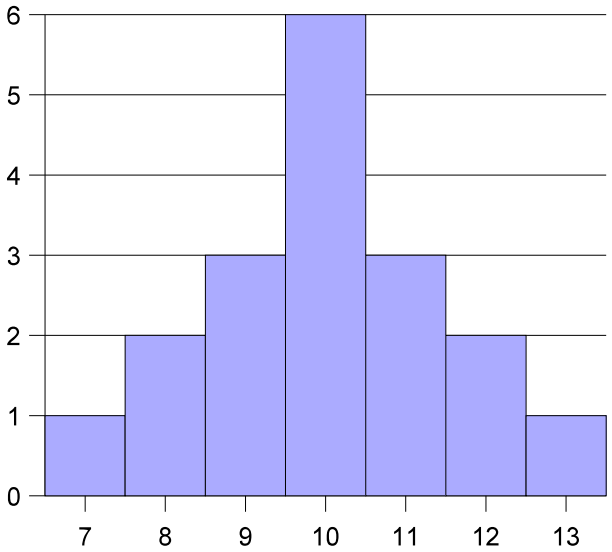
The gap between the mean of the lowest group and the highest group may be 4/20. It would be tempting to look for a reason, but there is no need. Statisticians would assure us that this is to be expected. In fact, if there was no such gap: if all of them were there same, I would be highly suspicious. Groups of 50 will show variability.

Of course I have deliberately made one mistake in presentation in drawing the graph. What is it?

Answer: the axes of the graph are not labelled. **It is a fundamental rule of presentation that axes must be labelled, and the units must be specified – usually with appropriate abbreviations**

Example 42: Eighteen classes of twenty-five
Groups of 25 might show even greater differences.

Class average	Number of classes with this average
7	1
8	2
9	3
10	6
11	3
12	2
13	1



This is known as the **normal** curve. It is what you would expect from grouping people randomly. it means that any one group is totally **unrepresentative** of the whole large group. **Therefore to study only one group means the results cannot easily be interpreted**.

This problem of the representativeness of a group is a given, that means **it is built in to any situation** where we use groups of people. Any experimentation needs therefore to take it into account, and that might mean measuring it. We describe this taking into account of the natural variability as *controlling*, or *the natural variability was controlled for*. How this is done will be explained below. But at this point notice how the variability often increases when the groups are small. In the groups of 50 the variability was 4, or ± 2 (this is read as 'plus or minus'). In the groups of 25 the variability was 6, or ± 3. In each case the average (any type) stayed the same.

Example 43: Real life example to show the existence of large natural variability, and how one sub-group may not be representative of the whole

Many years ago some students in the second year at University were given a graded vocabulary test. Common words were tested first, then less common, then rare words. Almost all the classes were tested, and almost all the students took the test. The testing was for vocabulary at several levels:

a) The first 2000 most frequent word families
b) Word families in the 2000-3000 word frequency
c) Word families belonging to the academic vocabulary used at university.
d) Rarer words, from the 3000-5000 word frequency

Table to show the results at the 5000 Word level arranged by increasing order of class mean

Class	Number in class	Mean	Standard Deviation
1	24	7.08	2.91
2	16	7.13	3.14
3	20	7.35	2.71
4	18	8.56	3.20
5	26	8.65	2.33
6	18	9.06	2.55
7	19	9.68	2.62
8	20	9.75	2.75

If an experimenter had only taken results from one class, they could easily, just by chance have chosen the weakest class. They could easily, by chance, have chosen the strongest class. That is why results from one class are difficult to interpret.

Obviously there is another problem that the numbers in the class vary, and are much smaller than the numbers of students enrolled. There are therefore supplementary questions as to how representative the students were who took the test compared to the students who did not take the test. It is likely that both the really good students and the really bad students did not bother to come - in which case the two extremes would cancel each other out. Because of the missing students, this data is already questionable, but it is useful data for teaching purposes.

Another figure to note is **the Standard deviation**. I have already said that the Standard Deviation is a measure of the range of variability around a mean. The higher the Standard Deviation, the greater the variability within the group. It is quite obvious that

Class 4 had the highest variability (3.2) and Class 5 had the least (2.33).

Statistically, to compare two groups we must look for similar means and similar Standard Deviations.

This data therefore also shows the great importance of visually examining the results and not relying on statistical tests. It also confirms that while there were reasonable differences between the classes, somewhat confirming teacher impressions, these differences were not as significant as the teachers thought. This is in fact reassuring: if for practical reasons the experimenter had only chosen two, three, or four classes, then the selection would have been reasonably representative of the whole population. It also shows clearly the dangers of only choosing one group for data collection purposes.

To summarise

- Wide variations between classes are normal
- Choosing one class randomly or by convenience, is highly problematic
- Know the meaning of the Standard Deviation: the greater the SD, the greater the natural range of results within a group.

KEY 13. MANAGE THE PROBLEM OF VARIATION WITHIN PEOPLE

1. Introduction

One of the great assumptions in social sciences is that people will behave consistently. If they are measured once as being highly motivated then we can assume that they will always be highly motivated. If they express a preference for a learning style, then they will always and consistently prefer this learning style. If people show a preference one day they will express the same preference on another day.

The assumption is highly questionable. There are various methods for managing the problem, and all of them involve being sceptical of one set of data collected at one time. Motivation, learning styles, opinions, need studying several times in order to confirm presumed consistency.

In my own research, using semi-structured interviews, I asked the opinions of several teachers. Then, I typed a summary and presented it to them for comment a few weeks later. I sometimes found that the teachers 'corrected' me even though I was confident my summary was correct. Of course, I never insisted I was right!! What I learned from this was that people sometimes forget they have said something, they forget the view they had expressed, or they change/develop their view. Therefore I had to be sceptical of the information from just one interview.

People are often inconsistent, therefore there is a range of results they might give you depending on a whole host of factors. Most people are not aware that they change a lot.

But you as researchers know this now, and can plan on people being mixed up and inconsistent. In fact, people being highly variable within themselves should be the default assumption. So, look out for it, and include it in your analysis.

2. People are NOT average

a. Reference: Todd Ross, in 2016, wrote a book called, "The end of average"**.** (Penguin books, UK). He confirmed what I had known for some time, and provided some interesting evidence which is presented below.

b. Robert L. Dickinson

In the 1940s in America, he and others collected detailed body measurements from 15000 young adult women. This was averaged to produce the fictitious ideal average woman called Norma. This became an idealised form to which women aspired. In 1945 there was a competition to find an actual woman who had, or came close to, these average measurements. The reality was different. Less than 40 of the 3864 contestants were average on five dimensions, and no one came close on all nine dimensions. The average sized woman did not exist.

Most doctors though bought into the ideal, and believed that most women were unhealthy and out of shape. A man called Daniels disagreed. He was interested in redesigning equipment for the military. He concluded that averages did not fit reality, therefore environments such as aircraft cockpits needed to fit the individual rather than the average.

c. The science of the individual

The field rejects the average as a primary tool for understanding individuals, arguing instead that we only understand individuals by focussing on individuality in its own right. The fatal flaw of averagarianism is the assumption that you can understand individuals by ignoring their individuality.

The primary research method of averagarianism is *aggregate then analyse.* Get a lot of data about individuals, combine it all, then use the group patterns to analyse and model individuals. The science of the individual works the other way round. First, look for patterns within each individual. Then look for ways to combine these individual patterns into collective insight.

d. The stepping reflex

With infant development, from the 1930s to the 1980s there was a puzzle. With an infant of 2 months, held upright they begin moving their legs in a way that resemble walking. But then it disappeared until just before they begin to walk. Using the averaging method, the stepping reflex became linked with presumed neural development, and the 'myelination' theory. Then Ester Thelen studied 40 babies over two years. She photographed and measured each baby, almost daily. Eventually she formulated a new hypothesis: chubby thighs. She noticed the babies who gained weight the slowest had the stepping reflex for the longest period of time. It was not the absolute chubbiness of the thighs that mattered. It was the rate of physical growth. What mattered was the amount of body fat relative to the development of muscle strength. Previous scientists using the aggregate then analyse approach could not see this kind of detail. Thelen proved her theory by placing babies in water, and, surprise surprise, the vanished stepping reflex reappeared.

e. Character traits are a myth

(Based on Todd 2016, chapter 5)

There is a whole industry to do with personality testing. The most famous are the Myers-Briggs test, and the Enneagram personality test. This vested interest does not take kindly to being questioned. In addition, they seem to have achieved popular support, so those who challenge it appear to be ridiculous. Tests which score people on a set of traits touch a deep rooted sense that a few traits define the essence of a personality. It is a form of essentialist thinking. If we can identify and establish a personality type then we can form conclusions about their personality and behaviour.

But, correlations between personality traits and behaviours that should be related are rarely stronger than 0.3, which is extremely weak, and means the traits explain at most 9% of your behaviour. So, traits are there, but are not very important. There is another angle: situation. Thus for instance, the question "are you an introvert" can have the fair answer, when

I am tired, yes, but most of the time, professionally, I am an extrovert and love the thrill of teaching and interacting with people.

The third angle, is not just the context, it is the interaction between trait and context. Also, it must never be forgotten that traits are not necessarily biological in origin, they can be learned, and unlearned. Traits are a convenient shorthand, a convenient automation for the first response, but they need not dominate, and need not be slaves, and the automation can be changed.

In a massive study of 84 children over six weeks, Shoda found that each child exhibited different personalities in different situations. He documented that children could be BOTH introverted and extroverted.

Shoda did find predictability. He found there was predictability in that people are consistent within a given context. His book is "The person in context: building a science of the individual".

If the personalities of other people appear to be static, remember, we interact with most people within a narrow range of contexts.

f. Crawling

There is no such thing as a normal pathway of stages towards crawling. There are at least 25 distinct routes. In addition, crawling itself may not always happen. The idea that crawling is the stage that precedes walking is the result of averaging a highly unusual sample of children – those from industrialised western countries.

It would be difficult to find nowadays the confirmed existence of any single normal pathway for any type of human development. In all aspects of our lives, there are many ways to get there, and the optimum pathway for you depends on your own individuality.

KEY 14. MANAGE THE PROBLEM OF CONTAMINATION

One way to explain the problem of contamination is to begin with a discussion of simple science experiments.

Example 44: Contamination in testing bacteria
Whenever you do a test of the bacteria growing for instance in your sputum, or at the back of your throat, you place a specimen on a sterile enclosed Petri Dish, which has within it some sterile growth media.

The trouble is that the growth media – usually known as 'agar jelly' might not be sterile ie the dish might be contaminated. But the whole point of doing the test is to grow the bacteria, and see which antibiotics work against them. If the dish is contaminated then the wrong bacteria will also grow and the results will not be at all clear.

This is one example where there has to be a major effort to check that the test has not been contaminated. Whenever such testing is done, there is always the possibility that the tubes or plates of growth media were never completely sterilised, therefore **any growth could be due to contamination.** As a check for this, usually at least one tube or plate is left, unopened, and placed in the warm incubator, along with the tests. The assumption is that if there was imperfect preparation techniques, then all the prepared media will be contaminated, and this will show up in a positive culture even in the control tube, thus leading to a checking and repeat of the tests. The language used is that what is being *controlled for* is the presence of contamination. The tube or plate that did not have any extra substances inserted later, is described as the 'control', and becomes the standard for comparison against the 'test' tubes or plates.

In other areas of research, the main way that contamination is *controlled for* is to use a careful experimental design and reasoning that identifies the variables, assesses their size and

direction, and tries to either measure several variables, or to fix them so that they have no effect. This will be explained later.

But such an approach often seems a long way from a language classroom or the study of linguists, and students often find it hard to apply the principles to their area of study, therefore, discussing case histories such as those above can help. At the design stage the question must be considered as to how to control for contamination of the results. Possible contaminating factors need identifying and listing. **The aim is that only the experiment causes the measured change**. **In a control, we expect no change.**

KEY 15. MANAGE THE PROBLEM OF SAMPLES AND NUMBERS TESTED

It is rarely possible to test everyone, so a few are chosen. The few who are chosen are supposed to be *representative* of the whole *population*, therefore a result from a part can be generalised, said to be true of the whole. A sample is representative if no one subgroup of the whole population is represented more than its proportion in the population. This means that it represents in miniature the whole population.

Example 45 for discussion

Assume that the total number of students studying for a Bachelor's degree, in all subjects is 100,000.

Male	60,000	Female	40,000
Married	5,000	Single	95,000

If you were to take a representative sample of 1000 students then you would get:

Male	600	Female	400
Married	50	Single	950

I could go into further detail, but the example should be enough to illustrate what is meant by a representative sample. It is important that the sub-group, the sample, is a scaled down version of the larger group or population.

A sample is **biassed** if it is not representative.

Example 46: Examinations as a sample of the knowledge and skills of students

An easy way to consider this question is to look at examinations. It is not possible to test all the knowledge and all the skills that a student is supposed to learn in a course. Therefore the examination tests part of the course. The assumption is that the questions are representative of the whole.

Now this is easy to do with something like testing the ability of a student to write a sentence in the phonemic script. All that is needed is two or more sentences. Asking for more than two sentences will probably not give higher performance: two sentences are representative of the overall ability to write in the phonemic script. The problem though becomes more problematic with long answer questions. Many teachers therefore use short-answer questions, and set more than one essay, in an attempt to get a representative sample that can be unambiguously and fairly marked. By the same argument some teachers reason that long examinations, perhaps setting four essays, are not needed. A few short questions are a better sample of student knowledge and ability, as well as being easier and quicker to mark.

We need to distinguish between a **haphazard/convenience** sample and a **random** sample. Interviewing students as the come out of college would be haphazard, but it is NOT random. The selection would have no intentional bias, but could still be biassed by other circumstances. Haphazard samples may be fully representative, or they may not.

A **random** sample is if everyone chosen has an equal chance of being chosen. Most spreadsheets now can generate a list of random numbers, and these can be used to pick out students to be interviewed. A sample chosen randomly is representative of the whole population, and there is only a very small chance that the results could be biased.

In practice it is often very difficult to get a truly random sample, so the researcher aims to get numbers in proportion to the whole population.

Example 47: Critique of published work

One of the biggest limitations of research in linguistics is small sample size. This is obvious when one reads the research skeptically. For instance, Bowles (2006) studied telephone calls. The topic is very interesting, and he makes many thought provoking points. The introduction is a model of clarity, and he makes a persuasive case for not ignoring Conversation Analysis. However, his corpus of data was extremely limited. He took 40 calls by English Native Speakers to English bookshops, and used only eight subjects. He recorded 50 calls to the same bookshops by nine Italians, all with B2 level of English (as measured on the Common European Framework of Reference scale). There were a maximum of six calls per person recorded. Several bookshops were used, and there was no description whatsoever of the native speakers.

This is a mess of variables, and MA students of mine would not be permitted to work this way. Notice that the data collected was not natural, in that the subjects knew they were being recorded.

What bothers me most though is that the data was too small for any valid conclusions to be drawn. In this text I have already explained the importance of considering variations in individuals, and between individuals. While individuals have patterns of language use, they are not always consistent. Eight native speakers and nine non-native speakers can hardly be considered representative of all, or even for the Italians, of all Italians at B2 level in English.

The last straw so to speak though is that the author proceeded to look in detail at the 40 and 50 calls recorded. [By the way, there is even a mistake of addition on page 350: the NNS TCU total is 22, not 23]. There was a lot of variability in the phrases used on the telephone, such that 22 out of 28 for NNS, and 19 out of 25 for NS were single examples. All the data shows is that there is wide variety of phrases used, by Native Speakers and by Non-Native Speakers. It would take a much larger sample to begin to spot the patterns, and to group patterns together for the purposes of analysis.

The rest of the article goes on, mostly unacceptably, to draw implications for textbooks, so going way beyond the data. The ideas may be good, but they are unsupported by the research presented. The apparent strong point of the data was the variety of phrases. But, given the small sample, and the way the many variables were not controlled, it does NOT necessarily mean that there is a large variety of phrases that are commonly used. That may turn out to be true, but is not a valid deduction from the data. All we can say is that for this tiny sample, with so many uncontrolled variables, this data was collected.

The only redeeming feature of the article is that it provides proof in principle that the methods of conversation analysis can be used. As such, the research is the report of a pilot of the method. The work needs repeating in a much more controlled way, and with careful specification and management of variables.

KEY 16. MANAGE THE PROBLEM OF SIZE

Example 48: The problem of size
A random occurrence of five students having a sleepless night because of road works, in a group of 20, could put the school at the bottom of the league of examination results, but is less likely to do so for a group of 100.

The smaller the group, the greater the possibility that it is unusual and unrepresentative.

Example 49: A large subgroup which does well can hide no change in a smaller subgroup
Imagine a class of 30 pupils, 10 male, and 20 female. At the beginning of the experiment, they all got 10/20 on a test. After the experiment, the 10 males got 11/20 each, and the 20 females got 14/20 each. Simple calculations show that the average mark for the whole class was 13/20. Now if the results were published, without paying any attention to gender, then the experiment would show a good result. But on closer examination, when the subgroups are considered, then one of them made insignificant progress. Therefore **the 13/20 is extremely misleading, because a significant subgroup did not get anywhere near this high score.**

Students		**Test 1**	**Test 2**	**Total Marks**	**Mean**
10	Male	10 each	11 each	110	11
20	Female	10 each	14 each	280	14
30 Male and Female				390	13

Example 50 of a large subgroup which hides a significant change in another subgroup

Imagine a class of 30 pupils, 10 male. and 20 female. At the beginning of the experiment they all got 10/20 on a test, as in the first example. This time, after the experiment, the mean score was 11/20, which is an insignificant difference. But on closer examination of the data, the 10 males got 13/20 and the 20 females got 10/20 again. This means that the males did significantly well, whereas the females made no change. Here we have an example where **a larger subgroup in which did badly can hide the good results of a smaller subgroup.**

Students		**Test 1**	**Test 2**	**Total Marks**	**Mean**
10	Male	10 each	13 each	130	13
20	Female	10 each	10 each	200	10
30 Male and Female					11

The larger group of females have so to speak, watered down the significant result of the males.

It is always possible that a significant result of a small group will be covered up within a large group. Averaging between two subgroups such as Males and Females is always dangerous. This is important because conceivably, one subgroup might do well, and the other show no effect, therefore averaging over the male/female sample could have hidden an interesting result. If you mix up males and females, and only have a small number of females, and then there is no effect on the males but a significant effect on the females, then when males and females groups are merged, the large effect in the females will be swallowed up in the averaging with the many males. In this case a significant difference no longer shows up, because of the dilution in the larger group.

In addition, if any researcher plans to do statistical tests, then each subgroup needs to be at least 30 in size. Every additional variable considered demands an increase of the sample size. In the chemical formulae recognition experiment (Example 34 above), there were over one hundred students in the school, with an almost 1:1 ratio of males to females. Therefore the question could be answered for male and female students separately, as well as for the whole group.

Example 51: league tables of school performance

In 1994, the British government introduced league tables for the examination results of schools. The idea was that by publishing these figures then poor schools would be motivated to do better, and that parents could choose the better schools for their children.

In 1996, the New Scientist warned that, according to two leading statisticians, the figures were fatally flawed. Firstly, the figures failed to take into account factors such as family circumstances. In a deprived area of a city for instance, a low score for a school might actually be very good considering the circumstances. More fundamentally, the publication ignores the problem of small sample size, which renders rankings all but meaningless.

"Small samples are intrinsically less reliable than large ones, because random effects have a disproportionate influence on the overall figure".(Matthews 2004:5) Most schools have only 50-100 pupils taking the examinations. For fair comparisons, the size would need to be at least 300. Most schools are simply too small to give reliable rankings.

The meaninglessness of the results is so great that the only clear difference is between the top 15% and the bottom 15%. Of all the schools, the 70% in the middle, there is no way they can be reliably ranked. Note what this is saying. It is NOT saying that of those schools in the top 15%, then a reliable ranking can be obtained. It is only saying that a valid comparison can only be made if one of the two schools you want to compare is in the top 15%, and the other is in the bottom 15%**.**

KEY 17. CONTROLS AND CONTROLLED EXPERIMENTS

The problem introduced

A very useful site is http://skepdic.com. The author is readable, well referenced, and credible.

Supposing a teacher has a new class, and tests them for their ability to recall the irregular verbs in English. Then the teacher proceeds to teach and revise the verbs systematically, and then gives another test some time later. **Can it be presumed from this that the teaching has had any effect?** No, it cannot be presumed that the teaching had been worthwhile. Maybe there was some other factor causing the improvement. This actually happened to me. A new French teacher said to the class that from now on no mistakes in irregular French verbs would be tolerated. To my knowledge, none of us from then on made any mistake: if we had have done, the teacher, in true French style, would have been sure to pounce upon us publicly. The story illustrates that sometimes factors other than direct teaching can influence a teaching outcome. The same teacher could also have gone ahead and revised the French verbs. In that case, would it have been the revision that caused the learning, or the expectation?

Incidentally. There is a lot of work suggesting that most people work best when the anxiety level is kept low. In this case, fear, even for adults, was a major helpful factor in helping us students to put aside easy errors and move on. Sometimes fear can be a helpful motivator.

To avoid problems like this, **controls** are commonly used.

'Controls' defined

A control is where the experiment is divided into two parts or sections. In the first part, some experimental work is done. In the second part, nothing new is done. This second part/section/group or groups is referred to as the 'control' or the 'control group(s)'.

Controls can be used for one or more of these reasons.

1) To reduce error in the results due to **contamination** from factors that are not being measured. There may be an 'intervening variable' which is influencing the results.

2) They assist in the making of fair comparisons. They help to establish a baseline.

3) They help us to manage the natural variability within a population.

4) They help us to manage the problem that a group may not be representative of the whole population and we do not always know how representative a group is.

5) They help us to deal with the fact that a small group will have problems of representativeness just because it is small in size, and one unusual result within the group can distort the group results.

6) They help us to cope with the very real problem that factors such as natural variability, and small groups, may actually be bigger than the effect being measured, therefore they are likely to swamp the results and a small change due to the experiment may not be noticeable.

For some reason, the students I teach find this concept of a control hard to grasp. While the concept should have been grasped as part of their general school education, this cannot be assumed. In addition, the concept and use of controls lies at the heart of the methodology of systematic investigation, therefore it is a vital idea. Some textbooks of methodology seem to give this very little attention: Nunan (1989 and 1992) for instance has no mention, Scholfield (1995) is brief. Fortunately Cohen & Manion (1985:187ff and updates) have a helpful and clear chapter with examples.

Part of the problem is that even when controls are obviously needed, how to set then up is not always obvious. **Often great**

imagination, creativity, and ingenuity is required, and students who are expecting recipes, or a series of rules, are going to be disappointed.

Some of the reasons for controls have been dealt with in the previous keys. First I will discuss the factors that have not so far been discussed, then move on to show how controlled experiments can overcome some of the problems mentioned above.

Another related problem is that a change resulting only from the experiment may actually be a small change. This small change may in fact be smaller than the natural variability. Therefore the following scenario is not good enough.

Example 52: Simplified experimental procedure

Stage	**Experimental group**
Stage 1	Test
Stage 2	Experiment
Stage 3	Retest

Let us imagine that in stage 1, the class average was 60% on a feature of grammar, then, after formal teaching, the class averaged 70%. Can we conclude therefore that the teaching was effective? The answer is a clear no, because other factors were at work which (together) could have had an even greater influence than the actual teaching.

Other factors, due to the school, the classroom organisation, the materials studied etc could have affected the results. These other factors are outside the control of the experimenter, and these factors may change, for instance bad classes noticing they are bad and changing their study habits.

One of the most common design procedures in educational research is to take two groups, do a pre-test on both, expose one group to an experimental procedure, then retest both groups. The

group which had no extra experimental procedure is the control.

That is why the following method is frequently used. It is seen as a 'true' experimental design.

Example 55: Experiment with controls

Stage	**Many Experimental group(s)**	**One or more Control groups**
Stage 1	test	test
Stage 2	experiment	nothing
Stage 3	test	test

But, this method only works if the differences between the experimental group and the control group are due to chance alone. In simple terms, the aim is to get similar mixtures. What students need to learn, and learn well, is that **classes, as organised in most schools and universities, are NOT similar enough for scientists**. Therefore, taking two classes, probably because of convenience, and doing some research then comparing the results is a dangerous piece of research. The results obtained are so likely to be subject to errors due to small sizes of groups and errors due to the groups being untypical, that fair comparisons cannot be made. Ideally, this means that each person in the group must be randomly assigned. In practice this is rarely possible in the educational settings. Therefore the problem of small groups being unrepresentative still remains.

The example given above of classes 1 and 8 nicely illustrates this point. If the vocabulary test had been done only on class 1, or only on class 8, then extended to the whole year, then the results would have been misleadingly high or misleadingly low respectively. **How does one know that one has got a class that represents the whole year?** Now, if class 1 had been picked as the experimental group, and class 8 as the control, then very unreliable results would have been obtained. If any two classes had been

compared with two other classes, then something approaching a fair comparison would have been made.

Ideally, allocation to the control group or the experimental group should be made by some kind of randomisation, or by using matched pairs, both of which are unrealistic in the face of the educational realities of most linguistics research. Where matching is not possible, the experimental and control group must be as alike as possible. (Cohen & Manion 1985:193, also Key 11 Compare like with like).

Therefore, the safest approach is to make half the classes the experimental classes, and the other half the control classes. To improve the situation, unlike natural sciences, where one test-tube is used as the control, and the other tubes are used for the experiments, half the groups must be used as control groups and half for experiment. If there are four our more groups in the experiment (say 100-200 students) and another four are used as controls, the experimental design is improved.

The natural variation needs to be 'controlled for'. This also means that **variables that could interfere with the study need to be identified and assessed.** The most common ones that are so treated in linguistics are age, gender, and social background. Either the variable is made constant, for instance by restricting the sample to females only, or, the experimental and control groups should include equal proportions of each variable (Scholfield 1995:28).

If ever someone does research on University students, then one obvious factor that needs to be controlled for is students who are repeating the year. Another common factor in social sciences is the genders of the students.

More reasons why controls are needed

Smyth (2004) comments on controls in the context of what it is like when a doctor prescribes treatment. In a sense, every patient is like an experiment, in which there is a problem, a solution (treatment) is proposed, and then evaluated. There are many

factors that can bias the conclusions.

a. The patient may have improved anyway

b. New treatments are likely to start when the disease is at its worst. We know from basic probability about 'regression towards the mean', once an extreme has been reached, the next reading is highly likely to be a more middle reading. Therefore, once an extreme has been reached, any treatment will appear to be effective. Statistically speaking, extremes are rare, therefore after an extreme, it is likely that a more normal event or result will take place.

   There is also the natural body mechanisms which work against extremes and pull the body back to normal. In flu for instance, once someone has reached a higher temperature for a day or so, the most likely sequence of events is that the temperature will go down. If one treats at this point, is the reduction in temperature caused by the treatment, or is it caused by the natural tendency for the body to recover and go back to normal?

c. **There is the placebo effect**
   A placebo is similar treatment, such as a sugar pill that looks and tastes identically to the real thing. Many people who are ill improve, even when the treatment is just sugar pills. There are various explanations, and you will find them explained at http://skepdic.com/placebo.html. The most common one is 'psychological'.

   The main method to avoid the placebo effect is to divide people into two groups: one group gets the medicine and another group gets the treatment. A new treatment should be even more effective than the placebo effect. If the new treatment is not significantly more effective than a placebo then it is an expensive waste of time and money. But, sometimes the doctor giving the medicine knows what is real and what is placebo, therefore in practice the double-blind system is usually used (see below).

Language point
The Hawthorne effect - an increase in worker productivity produced by the psychological stimulus of being singled out and made to feel important.
http://www.nwlink.com/~donclark/hrd/history/hawthorne.html accessed 3 May 2007

This effect is often confused with the placebo effect, and is probably a special name given to one part of the placebo effect.

d. **Expectations may influence the results.** Many people view expectations as a type of placebo effect.

The main way of designing experiments in medicine to overcome the placebo effect is to use a methodology called **double blind controlled trials**. Firstly, two groups are set up, experimental, and control. Secondly, both groups receive the same treatment, be it injections, tablets, or whatever, but, neither the doctors working with the patients or the patients themselves know if they are receiving a placebo or the medicine. Only later the code is broken, once the results are fully assembled.

Sometimes three groups are used for comparison purposes:
** normal treatment
** placebo
** new medicine

Double blind trials are vital for the assessment of new medicines. They often take several years to complete, and they need to be carried out on a large enough group of similarly ill patients. Sometimes a new treatment appears and somehow patients get to hear of it and start the new treatment before full double blind trials have been done. This can happen for instance when a medicine that is readily available and used for other purposes is found to have a desirable effect on the disease in question. With the rise of the internet, and the rise

of self-help and support groups, more and more patients are able to try out new medicines before they have been thoroughly and rigorously tested. It is very frustrating for patients, who may well be suffering from a terminal illness to read reports of a new treatment that appears to be effective, yet, unless the double blind trials are done adequately, the new treatment will never receive the thorough testing it deserves, and therefore it will be harder in the future for doctors to know if it works or not. My daughter has Cystic Fibrosis - a terminal chronic genetic problem which, unless there is some significant progress in the next few years to limit damage especially to the lungs but also to other organs, then her likely lifespan could well be only 40-50 years. At the time of writing, I see several potential new treatments that are ready for double blind trials, but I know it will be several years before we know which one of them really works, and in the meantime we must use existing tried and tested methods to delay the lung damage. I would dearly love to get hold of some of those new treatments, but cannot. And even if I could, the wider good - which means thinking of other people in her position, means that it would not be a good idea to rush. I add this personal note to show how methodology can be a life or death affair when it comes to medicine.

Summary

In experiments, the measured result must NOT be due to:

interference from another factor - contamination

natural variation within a population.

A measured result must only be due to the variable you are interested in.

KEY 18. EVALUATE THE INTERACTION OF VARIABLES

Behind the concept of controls is the concept of variables, and the necessity of evaluating all possible explanations for any data collected. That is why all good experimentation involving people usually begins with a careful study of the situation, and a specification of the characteristics of the population.

The basic aim in many experiments is that when two groups are compared, they should be identical in every way except for one feature, which is the experimental, or dependent variable (Scholfied 1995:30). Ideally only one variable at a time should be changed, except in large sophisticated experiments where several variables can be studied and accounted for simultaneously, which is a subject for advanced texts on methodology and statistics.

But in social sciences, we are coping with multiple variables, just like in physical and biological sciences. Students are probably not aware of how complicated the realms of physics and biology are, since many experiments done in schools are greatly simplified, or consist in the early experiments which did not have to consider the huge number of other variables. The closest students may have come to scenarios with many variables in them is when they study ecology, for instance the interacting balanced communities of a forest, desert, or pond.

Our students could profitably receive training not just in naming variables but also in evaluating and balancing them. Variables need to be identified, evaluated in terms of their effects and their interactions, and in addition, the results need to be interpreted in the context of these variables. Not every variable can be measured. In which case its effect needs estimating (in other words, 'accounted for').

The table below shows how two variables can interact. Any given effect, such as a difference or a correlation, can be strongly negative, weakly negative, or neutral. It can also be weakly

positive, or strongly positive. The extra variable in turn could be neutral, or strongly or weakly negative or positive, which means that one factor can have one of several different effects. Sometimes effects cancel each other out: one trend increases the result, another decreases it, and in the table this is represented by a zero.

TABLE: THE EFFECT OF A VARIABLE ON A RESULT

RESULT						
Factor ONE		**-2**	**-1**	**0**	**1**	**2**
Factor TWO						
-2		-4	-3	-2	-1	0
-1		-3	-2	-1	0	1
0		-2	-1	0	1	2
1		-1	0	1	2	3
2		0	1	2	3	4

Example 56: teacher and pupil motivation
An obvious example to consider is in a classroom, the interaction of teacher ability with the motivation of the students. The best scenario is high teacher ability with high student motivation. The worst scenario is low teacher ability with low student motivation.

Assuming a neutral teacher, then student motivation can either compensate for this (+2) or when there is lack of motivation, result in failure.

Assuming average neutral students, then a teacher can motivate them or discourage them.

The point is that these two factors alone interact with each other, and compensate or reinforce each other.

Example 57 Effect of medicines
With any medicine, there are many other variables also influencing the result. A rough table can be drawn up listing them with the estimated size and direction of the effect. In this case + refers to positive, and - refers to minus.

placebo	+ +
medicine	+
nursing care	+ +
tiredness	**- - -**
natural improvement	+ +
motivation	**- - - or** + + +

From this, some easy conclusions can be drawn. The tiredness factor can easily be as important as the combined effects of the medicine and the placebo. Motivation is a powerful influence, and it can go either way.

Example 58: Variability in teacher marking
Another example is the variability in teacher marking. It could be hypothesised that teachers mark lower when the discomfort level is higher. Therefore hot weather would decrease scores by 1, as would humidity. On the other hand, a fresh breeze might increase scores by 2. Therefore these factors could balance each other out.

Example 59: Testing organic chemistry
In my own research, one part of comparing French and English for the language of science was the field of organic chemistry. Now a key area of organic chemistry is the names for the chemicals. The names are organised systematically, so that someone who knows the code, who understands the language, can read the name and be able to write out the formula, and draw the overall structure of the compound. In the early stages of my research I had documented the two different patterns of writing naming these chemicals.

I wished to measure their ability to read the formulae. How could I measure this ability? I did not mean the ability to read aloud the names: I meant the ability to understand and interpret the names. I decided, based on professional experience of teaching organic chemistry in Britain, plus my own observations, plus my own feel for how difficult this part of chemistry was, that actually drawing a formula was easy. In other words, if a student could interpret the formula, they could easily draw it.

This ability to draw a formula is here an 'intervening variable'. It is a variable in its own right. As such, if the students were weak in this skill, then they would have been less accurate than I expected (ie its effect would have been weakly or strongly negative). I actually evaluated this intervening variable as being neutral (or similar in both languages, similar in size and direction). I could therefore ignore it.

Also, the students did not need more practice in drawing formulae, therefore asking them to draw English version formulae and later the French version formulae did not mean they were being sensitised to the method in the first test.

Just to make sure though, by applying Key 10, I arranged the different circumstances and variables such that it was difficult to get a result due only to the effect I was studying, therefore any difference would have shown despite a difference being unlikely, therefore the result was more convincing.

Even if an intervening or ignored variable cannot be pinpointed, or quantified, its effect (both in magnitude and direction) on the result should be assessed and estimated, and the results interpreted accordingly. In the case of the chemistry terminology research, there were three intervening variables.

- The first one was the skill of drawing the structure of the chemical, which I assessed as being neutral because the students had had a lot of practice at this basic skill.
- The second one was any sensitisation due to doing the same questions (different questions would have introduced the more complicated variable of the level of difficulty of the questions).

- The third variable was the sensitisation due to doing both tests on the same day. I planned the experiment so that the intervening variables were strongly stacked against measuring a difference, therefore any difference that did appear could be taken seriously.

Notice how I had identified and balanced the variables.

Viewing a formula, and drawing it, was a skill I evaluated the students as having mastered. Every time they saw a formula they could draw it, with 100% accuracy, and always achieving 100% accuracy.

Therefore
1. Students did NOT need more practice in this skill.
2. Any use of this in one language would NOT influence them in another.

But, just to make sure there was no sensitisation effect, I planned it so that any sensitisation would work AGAINST me.

KEY 19 UNDERSTAND THE ROLE OF COINCIDENCE

1. Introduction

A popular book on mathematics is called "Why do buses come in threes?" (Eastaway and Wyndham 1998). Why do buses often come in threes? Is it coincidence, or is there another explanation?

Several people have assured me that they had a dream, and it came true. The dream was about a specific marriage breaking up, or about the death of someone. Commonly this is viewed as evidence that God is at work, but is this a fair view?

Before we tackle such coincidences that 'must' have a supernatural explanation, let us take a step back.

2. If you heard that in a school class, two pupils had the same birthday. Coincidence?

Surprising as it may seem, statisticians assure us that provided the class size is at least 23, then there is at least a 50:50 chance that in any given class, there will be two pupils with the same birthday. (Paulos 1988:27. Note, I am NOT saying that if we take someone in the class, there is a 50% chance they will find someone else in the class with the same birthday. "It takes twenty-three people to be 50 percent certain there is some birthday in common, not any particular birthday ....It requires a large number of people, 253 to be exact, to be 50 percent certain that someone in the group" has exactly the same birthday as any previously chosen individual. (Paulos 1988:27). I am saying that in a class, when all the birthdays are compared, on average, half or more of the classes will have this happy coincidence**.**

Proof of the statement that there is a 50:50 chance that in a class of 23 there will be two people with exactly the same birthday.

The easiest way to do this is to calculate the chance that all the children have different birthdays. Imagine that one person enters the class, and has a birthday on 1 June. The second person has a 364/365 chance of having a different date. The third child entering has a 363/365 chance of having a different birthday. And so it continues until we arrive at the 23rd child who has a 343/365 chance of having a different birthday. When all these probabilities are multiplied together we get 0.49, or a 49% chance. See also Eastaway & Wyndham 1998: 49-50.

There is a simple formula for working out how many people you need to meet, for a given chance this will happen. Suppose that the individual has a 1 in C chance of having this specific characteristic. To have a 50% chance of a match in a group of N people, then $N = 1.2\sqrt{C}$. For a birthday match, the formula is $1.2\ \sqrt{365} = 23$ people. For a 95% match the formula is $2.5\ \sqrt{C}$. (Blastland & Spiegelhalter 2013: 80).

3. Dreams and so called correct predictions by psychics and others

a. The first comment to make is that there are thousands of people in the world making predictions, and most of them do not come true. Sometimes someone famous makes a correct prediction and the world thinks they are someone special. The fact is that sooner or later someone will make a correct prediction. If someone makes a lot of predictions, it is quite likely that a few of them will be right.

We tend to remember the right predictions and to forget the false predictions.

Coincidence can be normal

b. The second point relates to the fact that the human brain is hard wired, is programmed to notice and give importance to coincidence and to ignore non-coincidence. Most of the time our lives are lived without coincidence, but sooner or later two events will happen together. But in advance we do not predict which two events. If there are 10,000 independent events during the week, then sooner or later two of them will come together.

c. Thirdly, there is the phenomena of the sixth sense, dealt with below, where subconsciously we have picked up information that gets into our dreams, and this information is the foundation for them.

d. Fourthly, the vaguer the dream, the greater number of possibilities exist for its fulfilment. For instance, dreaming that one of the many people known to you know will die in the next week is quite a normal event. Each person may know up to a thousand or more people. Sooner or later such a dream will come true for someone.

e. These four points can be used to explain predicative dreams. Sooner or later someone will have a dream which comes true in the next few days (the longer the time frame the more likely it is that it will come true). We tend though to ignore the dreams that do not come true. Sooner or later many people will have a dream that comes true. Many people have sensed by intuition that something may happen.

All these points and others explain why horoscopes seem to work.

We must resist our feelings at this point and calmly follow the reasoning. At best a dream can be used to ask questions. Other than that, we should be very cautious of accepting the explanation of special intervention from the other world,

without also considering the normal rules of coincidence.

It is important to remember that dreams do not cause events, they do not cause coincidence. Thus, dreaming of someone dying does not CAUSE them to die. Obviously, if someone believes dreams cause something to happen, then bad dreams will be even more terrifying than they already are.

3. **Runs of coincidence**
   As Paulos (1988:44ff) points out, "Most people don't realize that random events generally can seem quite ordered". The following is a computer printout of a random sequence of o's and p's, each with a probability of a half. Imagine it as one long line.

A series of random events

```
opppoooppppoppppopppppooppoppopoopopoooopoppoooppppopopppppp-
pppppoppppopopppppopooppppooopppppppooppoooppoooooppoopppppppo-
pppppoopppppooppoppoopppopopoopppppoppoppppppoppoppppppppppoppp-
ppooooopoopppppoppppoopoopopppppoppppppooooopopoppppoppppooppooo-
opppppooooppppppoppoopppppppppoppooooooopopppppoooppopppppoooo+
popopoopppppopopppppoppoopppopoopoopppppopp
```

Note how often there is a run of three or more p or o. In fact, when I did a rough count I got the following table:

Long and short runs

	Size of run								
number	1	2	3	4	5	6	7	8	9
	p	pp	ppp	pppp	ppppp	pppppp	ppppppp	pppppppp	ppppppppp
p	21	20	13	9	6	1	0	0	2
	o	oo	ooo	oooo	ooooo	oooooo	ooooooo		
o	37	20	5	4	3	6	1	0	0

Runs of three or more are very common, and in fact accounted for 75 out of the 142 'turns' or changes from head to tail or tail to head. And that is the point here, that 'runs' of possibility are very common. If we noticed a series of coincidences in normal life and felt compelled to give an explanation for them, we might end up seeking an explanation when none need exist, except that **runs of coincidence are normal events**. We need to be careful here, careful not to jump to conclusions, and careful not to choose only one possible conclusion. It is possible that there is a reason for a run of coincidences, and that there is also the statistical fact that **runs are common: distinguishing between a coincidental run and a genuine causal reason may not be easy or possible.** And that is partly why you have scientists.

Hence the folk lore that 'accidents always happen in threes' is not far off the mark. Statistically a lot of accidents are grouped in this way. Gilovich (1991:3) points out that humans tend to notice the remarkable, and the problem is that people while being rational are not rational enough - in short, their rationality needs educating. “Human nature abhors lack of predictability and the absence of meaning. As a consequence, we tend to 'see' order where there is none, and we spot meaningful patterns where only the vagaries of chance are operating”(p9). Human intuitions about what is to be expected are frequently wrong. “People expect sequences of coin flips, for example, to alternate between heads and tails more than they actually do.” (p15) therefore people see order when in fact there is none.

4. General and particular probability

Paulos (1988:28) points out that if we make a general prediction it is highly likely to come true some time some where. But if we are specific, then the likelyhood is small. For instance, if there is a special roulette wheel, which as the 26 letters of the alphabet on it, and this wheel is spun 100 times and the letter it falls on is recorded. The probability that among those random hundred letters there is one or more recognisable words is high. But the probability that the words CAT and WARM may appear, is exceedingly small.

If you take the first letters of the months of the year, there are several words in them. JFMAMJJASOND - gives the word JASON, and the planets gives MVEMJSUNP, with the word SUN in it. Is this coincidence? Yes, entirely coincidence, because any recognisable word is allowable as a result: it would only be interesting if a specific word was predicted.

5. Intuition

There is a growing body of evidence in science that people are able to perceive and sense things at the subconscious level. Phillips (2004 p 14) for instance provides a news item in the New Scientist reporting recent work suggesting that Some people may be aware that a scene they are looking at has changed without being able to identify what that change is. "Our visual system can produce a strong gut feeling that something has changed", Rensink says, "even if we cannot visualise that change in our minds and can't say what was altered or where the alteration occurred". The phenomena has been labelled as 'mindsight' by Rensink. Phillips goes on to say,

> Mindsight may also be at work when someone goes into a room and senses something is different but can't put their finger on what. 'It could well be an alerting system,' he [Rensink] says...'.Knowing someone is behind you may be the auditory equivalent'.

Sub-conscious perception could therefore account for a significant part of coincidental thinking and dreaming.

6. Why then do horoscopes seem to work?

a. Astrologers emphasize the importance of the positions of the Sun, Moon, planets, etc., at the time of birth. However, there is no single moment that a person is born. Is it the moment the water breaks? Is it when the umbilical cord is cut? Is it when the first breath is taken? Or does birth occur at the moment a physician or nurse looks at a clock to note the time of birth? (skeptics dictionary).

b. Astrology "works," it is said, but what does that mean? Basically, to say astrology works means that there are a lot of satisfied customers. There are a lot of satisfied customers thanks to "subjective validation" To say astrology "works" does not mean that astrology is accurate in predicting human behavior or events to a degree significantly greater than mere chance. There are many satisfied customers who believe that their horoscope accurately describes them and that their astrologer has given them good advice. Such evidence does not prove astrology so much as it demonstrates the Forer effect and confirmation bias. (skeptics dictionary).

7. The Barnum effect (also called Forer effect)

a. This is the observation that individuals will give high accuracy ratings to descriptions of their personality that supposedly are tailored specifically for them but are, in fact, vague and general enough to apply to a wide range of people. This effect can provide a partial explanation for the widespread acceptance of some beliefs and practices, such as astrology, fortune telling, graphology, aura reading and some types of personality tests.

b. A related and more general phenomenon is that of subjective validation. Subjective validation occurs when two unrelated or even random events are perceived to be related because a belief, expectation, or hypothesis demands a relationship. For example, while reading it, people actively seek a correspondence between their perception of their personality and the contents of a horoscope.

c. In 1948, psychologist Bertram R. Forer gave a psychology test—his Diagnostic Interest Blank—to a group of his psychology students who were told that they would each receive a brief personality vignette or sketch based on their test results. One week later Forer gave each student a purportedly individualized sketch and asked each of them to rate it on how well it applied. In reality, each student received the same sketch, consisting of the following items:
   - You have a great need for other people to like and admire you.
   - You have a tendency to be critical of yourself.
   - You have a great deal of unused capacity which you have not turned to your advantage.
   - While you have some personality weaknesses, you are generally able to compensate for them.
   - Your sexual adjustment has presented problems for you.
   - Disciplined and self-controlled outside, you tend to be worrisome and insecure inside.
   - At times you have serious doubts as to whether you have made the right decision or done the right thing.
   - You prefer a certain amount of change and variety and become dissatisfied when hemmed in by restrictions and limitations.
   - You pride yourself as an independent thinker and do not accept others' statements without satisfactory proof.
   - You have found it unwise to be too frank in revealing yourself to others.
   - At times you are extroverted, affable, sociable, while at other times you are introverted, wary, reserved.
   - Some of your aspirations tend to be pretty unrealistic.
   - Security is one of your major goals in life.

   On average, the students rated its accuracy as 4.26 on a scale of 0 (very poor) to 5 (excellent). Only after the ratings were turned in was it revealed that each student had received an identical sketch assembled by Forer from a newsstand astrology book. The sketch contains statements that are vague and general enough to apply to most people.

d. Two factors are important in producing the effect, according to the findings of replication studies. The content of the description offered is important, with specific emphasis on the ratio of positive to negative trait assessments. The other important factor is that the subject trusts the person giving feedback to give them feedback based on honest assessment.

   The effect is consistently found when the assessment statements are vague. People are able to read their own meaning into the statements they receive, and thus, the statement becomes "personal" to them. The most effective statements include the phrase "at times", such as "At times you feel very sure of yourself, while at other times you are not as confident." This phrase can apply to almost anyone, and thus each person can read a "personal" meaning into it. Keeping statements vague in this manner ensures observing the Forer effect in replication studies.

e. Later studies have found that subjects give higher accuracy ratings if the following are true:
   - the subject believes that the analysis applies only to them, and thus applies their own meaning to the statements.
   - the subject believes in the authority of the evaluator.
   - the analysis lists mainly positive traits.

KEY 20. CORRECTLY DRAW GRAPHS TABLES AND CHARTS

A. GENERAL COMMENTS

Unlike the material presented so far, which is sparsely documented, advice on the visual presentation of material is readily available. When writing for publication, each editor will have small differences of style, but regardless of opinions about the details, some basic points can be made and are worth reiterating.

1. **Every illustration should have a numbered descriptive heading**

2. **Sometimes very wide illustrations need presenting**
   In this case it is permissible to turn the page, but this should always be in a clockwise direction.

3. **As with statistical calculations, percentages should be avoided**
   Percentages should only be used if the numbers are greater than 50, unless there is need to compare trends across different scales. In addition, as Woods et al (1986:36) point out, it is a serious error in presentation to present percentages alone without the original numbers, since it often prevents later analysis of the data by another researcher.

4. **For graphs and bar charts***:*
   a. ***The independent variable is always shown on the x-axis*****,** and the dependent variable on the y-axis. This statement, so easy to make, is one that some students find difficult. The dependent variable is the readings or measurements taken, sometimes called the response. The independent variable or stimulus is the pre-set values, chosen by the experimenter. For instance, in a bar chart of success rate per class, the class is the pre-set variable, and the exam results are the values determined by measurement, therefore the exam results belong on the y axis, and the

classes belong on the x-axis (IOB 1989:5). In other words, **the results or response should always be on the vertical axis, the x-axis.**

b. ***Both axes need a label indicating the nature of what is being measured*****,** the units of measurement, and appropriate graduation marks. The units wherever possible should be standard, conforming to the SI system where appropriate. Ideally the label on the y-axis should be written along the axis, proceeding along the vertical line from the bottom towards the top of the paper.

c. ***The actual numbers should be clear.*** **In cases where it is not possible, because of scale, to determine the values used directly from the graph, then the actual numbers must be presented in one way or another**.

This point must be insisted upon. I have seen far too many nice looking graphs where it is impossible to decide what the actual figures are.

On bar charts this can be conveniently done by placing the data values just above the bars (Rubens 1992:401, 402, 406), a practice which makes for great clarity but which is surprisingly rarely used. Another method is to present the table of results directly above and aligned with the graph. This might be thought of as repetition, therefore bad style, and for reasons of space in an article might be excluded, But in the interests of precision a table of results needs to be present, the graph being there to bring the data to life and to make perception of the trends more obvious. Both are needed, and both have their place.

d. ***The bars or columns of a graph should not touch when*****:**

- the items on the x-axis are not numerical (eg test results per school, each school being one separate column), or when
- the numerical values are distinct and cannot be subdivided: these are known as 'discrete' values.

When plotting exam results per class, and the classes are numbered, then clearly it is not possible to have class 3½, therefore the bar for class 3 must be kept separate from the bar for class 4.

Only when 'continuous' data are used for the x-axis is it permissible to draw a histogram with the blocks touching. An example of this would be if the number of students who achieved exam scores were plotted on a frequency against score graph, and some grouping was done, for instance, by ten percent, so that the number who obtained 0-9, 10-19, 20-29 percent etc are plotted. All values from 0 to 100 percent are possible, therefore the bars should be touching. (IOB 1989:8-10).

5. ***Whenever line graphs are drawn*****,** a smooth curve, or a completely straight line, should only be drawn if there is reason to believe that the intermediate values would fall on that line, otherwise, a straight line should be drawn between each pair of points, so producing an irregular shape, but indicating that one does not know how the values vary between the recorded points (IOB 1989:5).

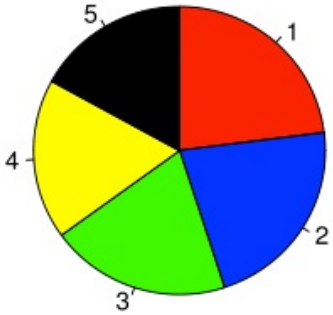
6. ***Several texts give invaluable advice for the presentation of tables.*** Rubens (1992) is very detailed, and Turabian (1987) and APA (1994) have the advantage of showing actual typed examples. A writer cannot go far wrong if these guidelines are followed.

7. ***All illustrations need a commentary.*** Illustrations are like quotations: they need to be prepared for and commented upon. In your commentary, you do not have to repeat all the information, but you should highlight what you want the reader to notice. Eg "As can be seen from Figure 3.20, the results are only significant for..."

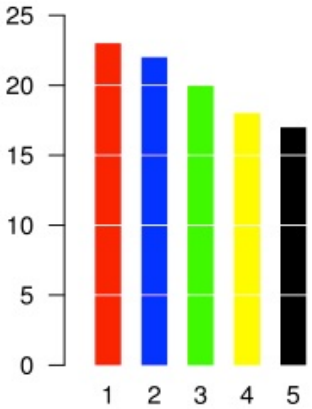
8. **For further information on graphical presentation of data I recommend the following:**
   a. Various publications at http://gsociology.icaap.org/methods/presenting.htm
   b. A real gem, which is humorous and easy to understand is Klass G 2011 on how to construct bad charts and graphs. http://pol.illinoisstate.edu/jpda/charting_data/badcharts.shtml
   c. For those who use Excel for data presentation, this site is a useful one in that it gives links to many other sites which provide tutorials and advice on how to use Excel. http://processtrends.com/TOC_links.htm

B. PIE CHARTS

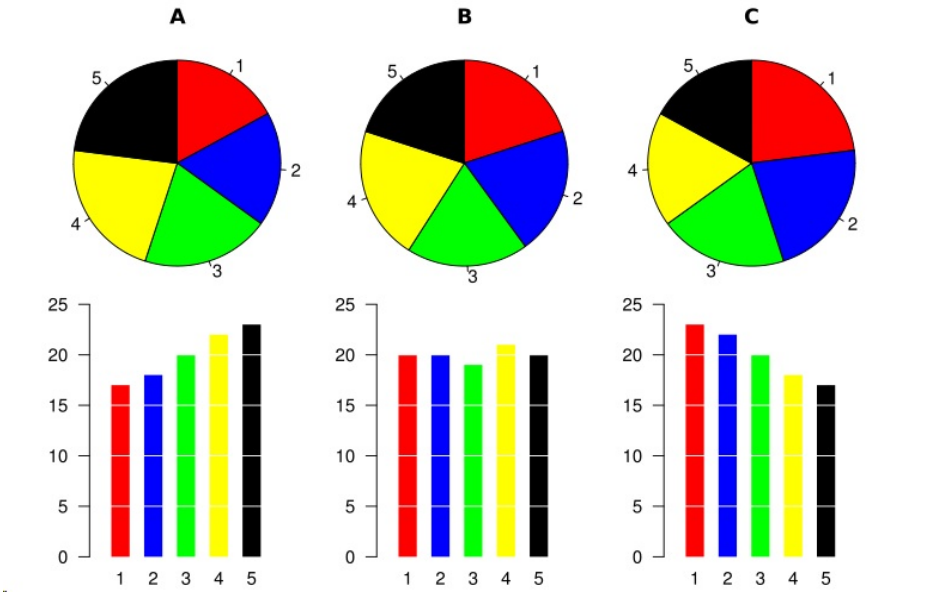
1. **They cannot be trusted (see www.eagereyes.org)**
   Which is larger in the following chart, the black or the yellow slice? What about black and green? How much larger? How sure are you? And where do you look to compare?



Now compare this to the bar chart. There is no doubt which is larger, or by how much. It's also easy to add a scale, gridlines, etc. When you care about comparing the parts, a bar chart is clearly superior. What the bar chart does not convey, though, is the part-whole relationship: changing the size of the green bar does not necessarily mean that any other bar has to change.



2. **They are not clear (see Wikipedia: pie charts)**
   Consider the following three pie charts, which all look similar, but when bar charts are used, there are clear differences.



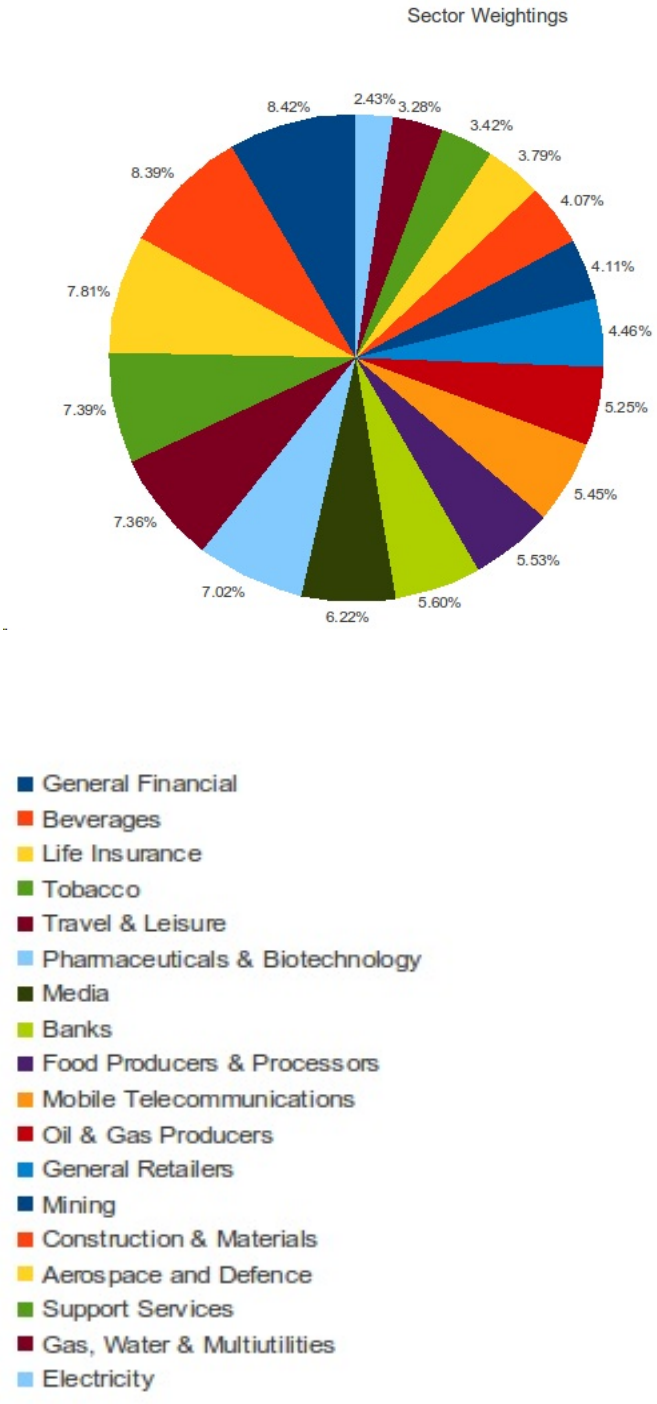
An obvious flaw exhibited by pie charts is that they cannot show more than a few values without separating the visual encoding (the "slices") from the data they represent (typically percentages). When slices become too small, pie charts have to rely on colors, textures or arrows so the reader can understand them. This makes them unsuitable for use with larger amounts of data. Pie charts also take up a larger amount of space on the page compared to the more flexible bar charts, which do not need to have separate legends, and can display other values such as averages or targets at the same time.

Statisticians generally regard pie charts as a poor method of displaying information, and they are uncommon in scientific literature. One reason is that it is more difficult for comparisons to be made between the size of items in a chart when area is used instead of length and when different items are shown as different shapes.

Three sets of percentages, plotted as both piecharts and barcharts. Comparing the data on barcharts is generally easier.

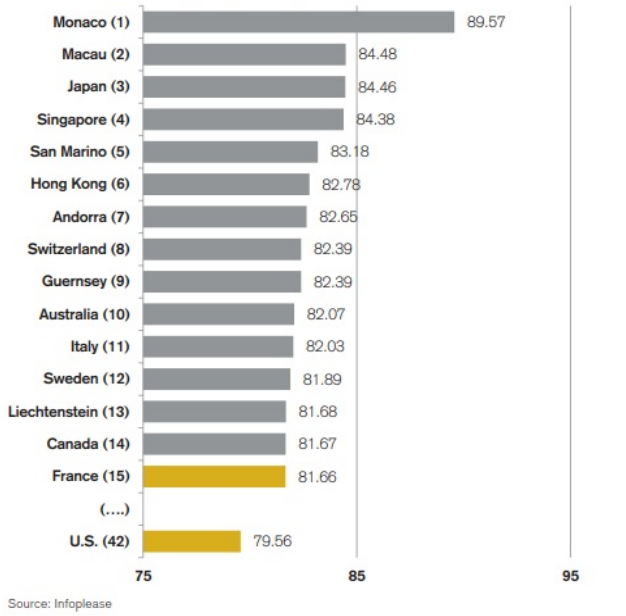
Further, in research performed at AT&T Bell Laboratories, it was shown that comparison by angle was less accurate than comparison by length. This can be illustrated with the diagram to the right, showing three pie charts, and, below each of them, the corresponding bar chart representing the same data. Most subjects have difficulty ordering the slices in the pie chart by size; when the bar chart is used the comparison is much easier. Similarly, comparisons between data sets are easier using the bar chart. However, if the goal is to compare a given category (a slice of the pie) with the total (the whole pie) in a single chart and the multiple is close to 25 or 50 percent, then a pie chart can often be more effective than a bar graph.

This is a pie chart which is impossible to read, especially for anyone who is in the least bit insensitive to small changes in colour.
The pie chart below requires extra labels, and a simple table would have been much easier to read. As presented below, it is difficult to link the key to the chart.


Sector Weightings
2.43%
3.28%
3.42%
3.79%
4.07%
4.11%
4.46%
5.25%
5.45%
5.53%
5.60%
6.22%
7.02%
7.36%
7.39%
7.81%
8.39%
8.42%
General Financial
Beverages
Life Insurance
Tobacco
Travel & Leisure
Pharmaceuticals & Biotechnology
Media
Banks
Food Producers & Processors
Mobile Telecommunications
Oil & Gas Producers
General Retailers
Mining
Construction & Materials
Aerospace and Defence
Support Services
Gas, Water & Multiutilities
Electricity

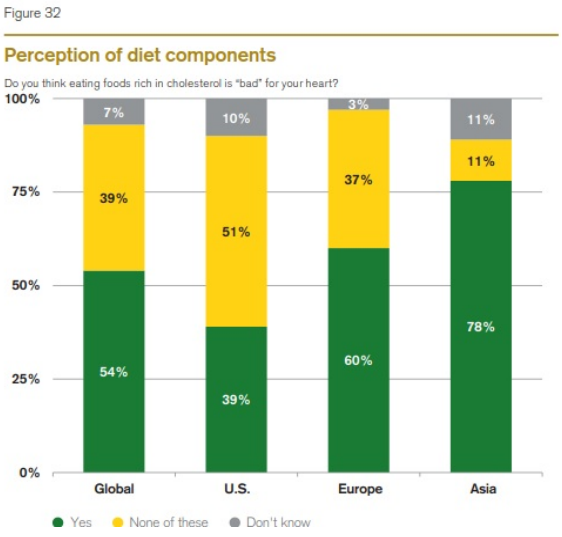
Compare those bad examples with some extremely good examples. These are taken from the Credit Suisse 2015 Research Institute report, "Fat: the new health paradigm". The report is full of clear charts and tables. Here are just a few.

Figure 14



The four bar charts below are much clearer than four pie charts

Figure 32



CONCLUSIONS AND RECOMMENDED READING

Simplicity and elegance of reasoning or experiment are highly desirable attributes of science, and should be emulated in language research. Any manipulation of data should highlight the subject matter, not the mathematics or technique. It is all too tempting to try to impress by using statistical tests, when the reality is that more work on careful interpretation of the data, using some of the keys mentioned, would give more useful information.

There is pressure to use statistical analysis in papers (Henning 1986), but such analysis does not have to be very sophisticated to be meaningful. Using numbers is part of the wider question of arguing fairly, and therefore is the concern of all. The techniques in books on statistics are a very useful extension of the methods of argument, but they should not detract from the simpler methods. A grasp of the concepts outlined here should enable anyone to handle quantitative data, and to benefit from one of the many introductory basic coursebooks in statistics. I finish with two quotations which invite us to think why we are using numerical data.

> **The most important maxim for data analysis to heed, and one which many statisticians seemed to have shunned is this: 'Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise**.' Data analysis must progress by approximate answers, at best, since its knowledge of what the problem really is will at best be approximate (Tukey 1962:13-14).
>
> A difference is only a difference if it makes a difference (Huff 1954:56).

This site contains a collection of jokes about the world of statistics and statisticians. It was presented the Golden Web Award 2003-2004 by the International Association of Web Masters and Designers.
http://my.ilstu.edu/~gcramsey/Gallery.html

I have been very impressed by this site:
http://skepdic.com