BRIEF NOTES ON STATISTICS
Introduction – why, what and how?

Michael Wood

6 April 2013

There are links to electronic versions of this document, and the other parts, at http://woodm.myweb.port.ac.uk/stats. For a rough, but still useful, understanding, you can ignore the numbered endnotes, which provide extra detail.

Why study statistics? Because it's useful. The basic idea is to take a lot of data and analyze it to show patterns that may not be clear from a superficial glance. This way we get to know about the dangers of smoking (http://woodm.myweb.port.ac.uk/stats/StatUses.pdf) and asbestos, predict the quality of wine and the best design for a website (Ayres, 2009), assess the effectiveness of innovations in various domains, assess the effectiveness of a marketing campaign or the risk of a financial investment, understand the risks and benefits of cycling in Barcelona (http://www.bmj.com/content/343/bmj.d4521), learn that in 2011 there were 7.8 road deaths per 100,000 of population in Belgium compared to the UK average of 3.1 (FCO, 2012), and so on.

Statistical analysis is indispensable in most walks of life. And where it is not considered indispensable, it perhaps ought to be – arguably many lives would be saved if doctors had a better appreciation of the role of statistics, and products and services would be of higher quality if those responsible for them made better use of statistics. Ayres (2007) gives an evangelical account of these, and many other, examples.

In my opinion, the core concepts of statistics are *averages and probabilities,* and *randomization*. These are what give the statistical approach its power. It is impossible to predict with certainty whether a particular person who smokes will develop lung cancer or exactly how many more years they will live, but the statistics can estimate the probability of developing lung cancer, and an average life expectancy for people in a similar position. It typically does this by taking a sample of data, and it is obviously important that this sample represents the whole population as accurately as possible. In practice the easiest way of doing this is often to use a *random sample* – this is chosen so that every member of the population is equally likely to be selected so there shouldn't be any consistent bias (see http://woodm.myweb.port.ac.uk/rm/u3critic.pdf). Randomization is also important in experiments or trials such as drug trials. Suppose you wanted to compare a drug treatment for a disease with a placebo, and you decided to ask patients whether they wanted the drug or the placebo. The difficulty would be that the two groups would be different in ways that would almost certainly affect the result (e.g. those who were

more seriously ill might opt for the drug). The solution is to allocate patients to the drug or placebo group at random.

Quick question 1. Suppose you wanted to investigate whether cycling has any health benefits, and if so what they are. How would you design your study? (Pay particular attention to how you would choose samples, whether you would use averages, and whether you would make any random choices.) What problems do you foresee?

Statistics is an extensive and complicated discipline. What can you do if you don't want to become an expert in statistics but simply appreciate its use in disciplines such as management, medicine or mountaineering?

The first possibility is to stick to the easy bits – averages, percentages, bar charts, pie charts and so on. However, this won't really do because you are likely to need some of the more difficult bits – e.g. correlation coefficients, standard deviations, regression analyses, hypothesis tests, and so on. But it is important to think hard about the easy bits – it is possible to use averages, percentages, bar charts, pie charts in very silly ways, so be careful!

The second possibility is to use a computer for the difficult bits. Sometimes this may be OK, but the difficulty is that you may not know what to ask the computer to do, and even if you do, you may not understand what the computer's answer means. The same problem may arise if you are reading an article which gives the results of some sophisticated techniques – you may not know how to interpret these results.

The third possibility, of course, is to study statistics and become an expert yourself. The difficulty here, of course, is that you may not have the time, or the academic background to follow, for example, some of the mathematical aspects.

In practice we need all three tactics to some degree. I will assume you know the easy bits – but don't under-estimate the necessity to think carefully about things that may seem trivial! I want to be as brief as possible, and to confine the coverage to those aspects of statistics that are important, or widely used (not always the same!). I have considered carefully every topic covered, and only included it if it satisfies these criteria. The basic approach follows my book (Wood, 2003), so you can refer to this for more detail.

My approach to the use of a computer versus "using your head" is as follows.

When the concept is simple enough to be understood in common sense terms, I'll get you to do it using your head so you know what it means, and then use a computer for large scale work. Computers are quicker and less likely to make mistakes, but you do need to understand what they are doing. This is the approach the notes adopt for medians, quartiles, standard deviation, etc.

When the underlying mathematics is too complicated to be appreciated in common sense terms, the approach will be to try to provide an intuitive account but without going through the detailed mathematics. It is usually possible to find such an intuitive approach. This is the approach taken to the normal distribution.

There is a third option. This is when the standard approach is too complicated but there is an alternative approach which can be appreciated in common sense terms. The standard correlation coefficient (Pearson's) is mathematically complicated, but there is another (Kendall's) which is relatively transparent – so we will focus on the latter which should provide an insight into how the former works. A large part of conventional statistics uses regression and hypothesis testing. The standard method for regression involves some formulae derived by calculus (a branch of mathematics) which are difficult for the uninitiated to follow. However, it is possible to use computer assisted trial and error to bypass these formulae (http://woodm.myweb.port.ac.uk/stats/LeastSquares.pdf). Similarly, most hypothesis tests are mathematically complex, so we will look at a method called resampling which makes the rationale far more transparent (http://woodm.myweb.port.ac.uk/stats/ShuffleTest.pdf)[1].

The reason I have included computer assisted trial and error for regression, and resampling methods for hypothesis testing, is to clarify the rationale behind the techniques. Understanding the rationale behind statistical methods is vital, but it is also difficult and, for many students of statistics, unfortunately often not achieved. However, in practice, if you go on to use regression or hypothesis testing you are likely to use the conventional methods implemented by statistical software packages, so it's more important to understand the spirit of the trial and error and resampling methods than it is to master the detail.

It is also important to work through the exercises and questions in the notes. Understanding how statistics can be applied means you must practice applying it!

Finally, I ought to confess that I have difficulties with the way statistical techniques are often used and presented in management research (Wood, 2013; http://tinyurl.com/dyyzawc). This extends to the whole idea of null hypothesis testing which is the main subject of Part 3 of the notes. If I had my way, null hypothesis tests would only be used in very special and rare circumstances. Usually, other methods are more useful (Wood, 2012). But null hypothesis tests are very widely used, so you need to know about them. Sorry!

References

Ayres, Ian. (2007). *Super crunchers: how anything can be predicted.* London: John Murray.

FCO (2012). Foreign and Commonwealth Office travel advice accessed on 14 October, 2012 from http://www.fco.gov.uk/en/travel-and-living-abroad/travel-advice-by-country/europe/belgium.

Wood, Michael. (2001). The case for crunchy methods in practical mathematics. *Philosophy of Mathematics Education Journal, 14* (a web journal at http://www.ex.ac.uk/~PErnest).

Wood, Michael. (2003). *Making sense of statistics: a non-mathematical approach*. Basingstoke: Palgrave.

Wood, M. (2012). P values, confidence intervals, or confidence levels for hypotheses? Retrieved from http://arxiv.org/abs/0912.3878v4 .

Wood, Michael. (2013). Making statistical methods more useful: some suggestions from a case study. *Sage Open, vol. 3*, no. 1 (available from http://sgo.sagepub.com/content/3/1/2158244013476873).

---

[1] Both of these approaches use the power of computers to do simple repetitive processes as a substitute for clever mathematics (Wood, 2001). They provide a complete, and mathematically rigorous, explanation of what is going on – which no other elementary account can do. Conventional courses (and text books) try to clarify the rationale of techniques by making you do some of the arithmetic – plugging numbers into formulae, etc. There are two difficulties with this. First, it's the rationale behind the formula, where it comes from, that you need to understand, not the mechanics of how to use it. And second, you often end up having to use statistical tables – which are another unexplained black box! For this reason I have largely avoided formulae.