Chapter 1

Introduction

1.1 Introduction

$\cdots$ simple descriptive statistics and display techniques are indispensable preliminaries to the application of even the most basic inferential statistics or tests. To my knowledge, the vast majority of statistical analyses of archaeological data, published and unpublished, have been done without adequate scrutiny of the data with such elementary display techniques and descriptive statistics. From my experience, I will be so bold as to put forward the view that this lack of adequate scrutiny of the data renders every one of these analyses, and consequently the studies and interpretations based on them, suspect a priori. *$\cdots$ However, there is another, more positive, pragmatic reason for so strongly advocating the cause of such humble displays and descriptions of one's data, which is that, in almost every instance, one can learn, more quickly, more clearly, and in more detail about one's data with these techniques than through the use of inferential statistics or tests.*

Whallon (1987: 135)

The above is from an article entitled *Simple Statistics*; there is nothing in it to take issue with. It was published more than 20 years after statistical methodology began to attract serious archaeological attention. Any statistical analysis carried out in anger should begin with 'simple descriptive and display techniques'. What has changed since Whallon wrote is what might legitimately be thought of as 'simple'. This is entirely down to the enormous increase in computing power now available. Many of the techniques discussed in these notes could not then be used easily and routinely, if at all.

Fast forward to the 2010s. Some of the software packages developed from the 1970s on survive and remain popular. This isn't necessarily a 'good thing'; practices developed then have become 'fossilized' in the software, not necessarily to

good purpose (e.g., Chapter 8 for further comment). To misquote slightly 'the past is a foreign country: they [did] things differently [then]'; they do things differently now.

Software accessibility is important. It explains the understandable widespread use (and misuse) of `Excel` – not a purpose-built statistical software package – in archaeology. There is now no longer any excuse for this. High quality, open-source software means that powerful statistical resources are available to all. The software, `R`, is used in these notes. It is often described as a 'difficult' package to learn for anyone who is not a 'sophisticated' statistician and/or computer programmer; the thinking, simply put, seems to be that potential users who have been weaned on menu-driven software find the command-driven mode that characterizes `R` daunting. Some additional comment on this is provided in Section 1.3.

My own thought is that anyone who can write a sentence with due regard to spelling, syntax and capitalization can start analyzing their own data in `R` in a matter of minutes. Methods sometimes portrayed as 'complex', such as principal component analysis (PCA), are actually quicker to implement in `R` than the menu-driven implementations of software commonly used in teaching (Chapter 2). While `R` is undoubtedly 'sophisticated', matching sophistication on the part of a potential user is not essential. That is, by imitating existing code widely available in texts and online, analytical progress can be made without any great initial understanding of how what you can achieve has been done. Such understanding and greater 'sophistication' will come with practice; the important thing is to get started. Motivation, in the form of having your own data that you want to analyze efficiently, helps.

These views rest on the premise that a lot of statistical ideas are 'simple', in a sense somewhat broader than that used by Whallon. Many methods, traditionally thought of as 'complex', are *computationally* straightforward to execute. The idea of 'simplicity' motivates much of what follows in later chapters. It can be construed in various ways, as follows

- conceptual simplicity;
- computational simplicity (i.e. simplicity of execution);
- mathematical simplicity.

Here, 'computational simplicity' is being equated with simplicity of execution. The *application* of many 'standard' (and 'non-standard') statistical methods in `R` easily meets this criterion. This is not to claim that the computational details are necessarily simple, but the necessary work on this has been carried out by experts, leaving the average end-user free to enjoy the fruits of their labor.

To take full advantage of this an understanding of what any particular method is intended to do helps; that is, is it conceptually simple? If the answer to this is

'yes' then, allied to computational simplicity, and from the end-user's viewpoint, a method may be regarded as 'simple'. A fundamental thesis of these notes, argued in more detail in individual chapters, is that the statistical methods mostly used in archaeology are simple in this sense.

A simple idea can lead to mathematically complex developments but, as with the computational detail, experts have dealt with this, so it's not usually necessary for the user to understand the details[1]. An exception to this generalization, as some might view it, is that some understanding of mathematical/statistical notation is desirable for a number of reasons, among them brevity and clarity of expression.

A fine but important distinction to make is that statistics is often simple in the sense described but this does not make it easy. Almost anyone who wishes to (and can afford it) can learn to drive a car without any deep understanding of the technology that makes this possible. The part that is less easy is learning to do this with facility and appropriately, and this only comes with practice.

The archaeological trowel provides another analogy. It's a simple tool and what you do with it is simply explained; using it well does not, for most diggers, come immediately, effective use comes with practice. Most aspirant diggers can arrive on a site and be on their knees and using a trowel within a short space of time, but will do so under supervision or relegated to an area where they can do no damage. To be realistic, some remain on these margins or are transferred to other duties more suited to their aptitudes but, with practice and motivation, most graduate to a position where they can operate with minimal supervision. Some diggers will be better than others; it's a shibboleth that 'practice makes perfect' but it certainly does no harm[2].

To summarize, employing statistics usefully can start from quite a limited knowledge base. You do eventually need to learn more about methods you find useful, a process of accretion, and be aware of when their use is appropriate. Statistics is no different, in this respect, from other subjects worthy of study; but, contrary to occasional misconceptions, you don't need to be especially gifted, with

[1]Mathematics is not avoided in these notes but, for the most part, is quarantined so that it can be ignored by the reader if wished. Some appreciation of the mathematical distinction between principal component and factor analysis (Chapters 7. 8 and Appendix D) helps to appreciate the way in which they have been used and confused but, unless you are attracted by what factor analysis appears to offer, the mathematics can be ignored.

[2]When I did this sort of thing seasoned diggers, armed with their 4-inch cast-steel WHS pointing trowels (`R`) would look pityingly on aspirant excavators arriving with inappropriate over-sized and over-flexible welded (they break easily) plasterers trowels (`Excel`) and condescendingly point them to the path of righteousness. The overly-seasoned digger would sport a second trowel, discreetly but visibly displayed and well-worn, testifying to their vast experience. These were very functional but equally important as symbols of superiority to be mutely admired by the cognoscenti. Except in the pub when conversation sometimes descended to the 'mine is smaller that yours' kind of boasting.

a strangely configured intellect, to engage in productive statistical analysis.

Chapter 2 illustrates by example what is intended by the term 'simple'. The standard methods of multivariate analysis used in archaeology are each executed in one line of `R` code, and in most cases immediately useful and interpretable graphical output is produced. That's the computationally simple bit. Conceptually, the methods are intended to reduce a large table of data to a picture, usually two-dimensional and often interpretable as a map showing the relationship between the rows and/or columns of the data table. The output can be examined for archaeological patterns in the data. Also a simple idea.

The simplicity conceals specific differences that distinguish between the methods that are discussed in the individual chapters devoted to them; Chapter 7 for principal component analysis (PCA), Chapter 9 for correspondence analysis (CA), Chapter 10 for cluster analysis and Chapter 11 for linear discriminant analysis (LDA). The chapter on LDA includes a brief treatment of classification trees, an attractive alternative to LDA that is non-linear and computationally intensive, but underpinned by a very simple idea.

It is worth emphasizing the important idea of 'mapping' the data in these methods. There's an analogy with map projections in cartography where different projections can be chosen to emphasize particular features of interest. Methods such as PCA, CA and LDA measure the 'distance' between the rows of a data matrix in different ways (i.e. they project the data differently). The choice depends on the nature of the data available and the aim of an analysis. Results are 'distorted' in the sense that only a visually accessible *approximation* to the 'reality' is obtained, and the quality of approximation needs to be assessed. The niceties of this are discussed in the relevant chapters; other issues that require attention in practice are also covered, such as the choice of data transformation or specific methods (e.g. in cluster analysis) to use, and the many different ways of interrogating the output of an analysis that are available.

Chapter 8 on factor analysis follows that on PCA fairly naturally, since the two methods are often confused. I've suggested elsewhere that, despite the 'historical' importance of factor analysis in the development of quantitative archaeology it is past its 'sell-by' date. That I've devoted space to the subject is because of what I read as misleading advocacy of the method in a recent quantitative archaeology text. Anyone who thinks factor analysis might be for them ought to be acquainted with its 'problematic' aspects, even if they disagree with my take on the subject.

A systematic discussion of descriptive statistics, such as the mean, median, standard deviation etc. is not attempted. Read (almost) any introductory text on statistics for this kind of thing[3]. Chapters 3 and 4 look very selectively at

[3]Those that shy away from notation can get definitionally confused, particular in distinguishing between population parameters and their estimates – so read more than one text.

some of the more commonly used methods of graphical display for continuous and discrete data. Kernel density estimates, as an alternative to the histogram that is sometimes more useful, is given more space in Chapter 3 than is common in quantitative archaeology texts

Chapter 4 is primarily concerned with bar-charts (or barplots) and pie-charts. The former are sometimes confused with histograms or inappropriately presented as three-dimensional constructs. Pie-charts are often similarly misrepresented and, depite their popularity, a good case can be made for not using them at all. The chapter is more opinionated than I might allow myself if writing a textbook, but the opinions are not uniquely mine.

Chapter 6, which can be thought of as a continuation of Chapters 3 and 4, is a pot-pourri of graphical methods of analysis. Some are fairly standard; some are little-used but deserve more attention; and some are probably neglected for good reasons.

Chapter 5, on regression analysis, covers a topic invariably dealt with in texts on quantitative archaeology. At its simplest – fitting a straight line through a scatter of points – the ideas are straightforward. The chapter strays beyond this into the realms of non-parametric regression, not a topic much mentioned in archaeological textbooks. More notation is needed here than in most other chapters, to distinguish adequately between different models, and between models and their estimates.

I had second thoughts about including Chapter 12, on statistical inference. It's arguable that the development of formal theories of statistical inference was one of the greatest intellectual achievements of the first half of the twentieth century; unarguable that these ideas motivated the promotion of quantitative methodology in archaeology; and arguable that *formal* inferential methods have delivered much less than was originally promised. Quantitative archaeology texts sometimes present the methodology with enthusiasm but I suspect equally often out of a sense of duty. Having included the subject I've been equally dutiful and illustrated the use of `R` with sometimes quite extensive examples, but with added opinion that the reader is free to disagree with.

1.2 How (not) to read these notes

The point made in the preface that this is not intended as a textbook, should be reiterated. It's meant more as a 'dipping into' kind of text – as the use of the term *Notes* in the title indicates. It also indicates that coverage is selective, including methods I have practical experience of and excluding those where this is not the case. Some methods are excluded for other reasons (see below). It's assumed that anyone reading this will have been exposed to a systematic introductory treatment

of statistics.

The material might be regarded as a kind of practical supplement to an introductory textbook development. The 'standard' texts for archaeological purposes will be taken to be Shennan (1997) and Drennan (2009) (see Section 1.3). Other than their treatment of sampling methods in archaeology, omitted here, much the same kind of methodology is covered, but with rather different emphases. As already noted some of the commoner descriptive and graphical methods used arise in context, rather than being systematically presented. There is considerably more emphasis on undertaking analysis than to be found elsewhere, with more analyses of real, large and, in some cases, quite complex data sets.

Some topics not usually covered in quantititative archaeology texts, or accorded only brief treatment, are introduced, including kernel density estimates (Chapter 3), non-parametric regression (Chapter 5), fuzzy cluster analysis (Chapter 10), classification trees (Chapter 11) and some of the graphical methods of Chapter 6 . These are present because they are useful and fulfil the criteria for 'simplicity' that is being applied.

In a few sections a more critical view is taken of the way some commonly used methods have been applied. This includes comment on the misuse of bar- and pie-charts (Chapter 4) and the real need for much of the standard inferential methodology of hypothesis tests (Chapter 12). Factor analysis is usually mentioned in the standard texts, but often in a cursory way or, where treated at any length, in a manner I think is unsatisfactory. It will be clear from Chapter 8 that I have my doubts about the value of the method for archaeological purposes and that others would disagree entirely; the chapter and associated Appendix D is possibly one of the fuller *critical* appraisals in the archaeological literature.

Faced with an academic text of any kind my normal practice is to skim it to see what's there then ignore the bits which don't interest me or are not suited to immediate purposes (which may be the entire text). Some experience is necessary to get away with this; teaching texts are usually intended to be read in a linear and systematics fashion, so you have to wade through some fairly boring, if essential, bits to get to the parts that might interest you. I'd never rely on a single text; just because something is published in some form doesn't make it sacrosanct (even if written by your instructor). It's as well to be aware that there's more than one way of tackling a subject, and it can come as surprise to learn that statisticians can have different views, sometimes vehemently expressed, about what does and doesn't constitute valid and useful methodology.

Having said this, a systematic reading of parts of these notes may be useful. Anyone entirely new to `R` would need to start with Appendix A and Section 2.6 and is then advised to do a few data analyses as soon as possible, either reproducing the analyses in the notes (e.g., those in Chapter 2) or using analyses found elsewhere

in the notes as templates for looking at your own data. Chapter 7 can be usefully read as a prelude to those which follow, on other multivariate methods, as it includes discussion of data-analytic concerns of general relevance. If it seems like regular use of R is an attractive proposition an early acquaintance with user-defined functions (Section 3.2.1) is helpful.

I've suggested at various points in the notes that doing things first and then worrying about what you've done afterwards is a viable way of learning about R, and the statistics involved. This is not a flippant suggestion. Jump, or ease yourself, in at the shallow end and immerse yourself a little. Any discomfort induced by the temperature will rapidly disappear, and once you are convinced you are still breathing you can think about learning to swim. You may even come to enjoy it.

What has been omitted from these notes could, as they say, fill a book. Sampling methodology has already been mentioned. It's important and more-than-adequately covered in Orton (2000). It's not a topic I've had to engage with in recent years and I'm not sure how useful R might be for acquiring and analyzing data in a realistic archaeological setting.

Spatial analysis is an interesting area also largely ignored. It's fairly astonishing that Hodder and Orton (1976) still seems to be referred to as the standard archaeological text on the subject. Archaeologists don't, of course, ignore spatial analysis but it is, I suspect, largely associated with the use of geographical information systems (GIS). I know almost nothing about GIS so am not competent to write on the subject and probably shouldn't comment either. The subject is undoubtedly important and has been around for over 20 years; it was 'oversold' in its early days and, when I last looked, made surprisingly little effective use of what I'd regard as 'proper' statistics – things have probably changed without impinging on me.

Computer-intensive methodologies are important and will become increasingly so, but are not dealt with systematically. As mentioned in several places a review of the use of such methods, for statistical purposes, in archaeology would be useful, but is not attempted here. Archaeological simulation, thoroughly reviewed recently by Lake (2014), is computationally intensive but makes limited use of statistical methodology as I understand it. I have more in mind resampling methodologies such as bootstrapping, noted at several points in what follows.

A systematic treatment of non-parametric methodology is similary not attempted, though kernel desnity estimates fall into this class of methods. Had more detailed treatment been provided it would have been in Chapter 12, where the methods, can be viewed as more 'robust' alternatives to the hypothesis tests usually discussed in introductory texts. Given they are intended to avoid assumpions that trouble some users of statistics I was surprised, on reviewing the textbook

literature, to see how little attention they have received.

Robust methods which, in some senses, are similarly motivated are similarly not much mentioned here. They are designed to minimize the effect on analytical outcomes of the violation of assumptions that underpin 'standard' methodologies, the deleterious impact of 'unusual' data on analyses having received particular attention. More exploration and exploitation of the use of the methods with archaeological data is probably needed to assess their merits. Almost needless to say, several `R` packages are available devoted entirely to robust methods.

It would have been possible to extend the treatment of regression analysis – more generally 'linear models' – to generalized linear models. These include log-linear models for contingency table analysis and logistic regression models. The beauty of the idea is that a lot of different models, some looking very complicated and non-linear, can be shown to have a common underlying structure that admits a common algorithmic approach to their estimation. Conceptually, the more 'complex' models inherit much of the interpretive machinery of linear models and – to the extent that the latter are 'simple' – it might be argued that generalized linear models are also 'simple'. In practice I don't find them as straightforward as other methods covered and have not attempted an exposition. Log-linear models had a brief vogue in archaeology in the 1980s and 1990s but I don't think are now widely used; logistic regression acquired something of a niche role in predictive modeling. Baxter (2003) provides examples of their use.

1.3 Suggested reading

As already mentioned, Shennan (1997) and Drennan (2009) are useful introductions to quantitative archaeology, referenced a lot in what follows. They can be viewed as complementary; both are worth looking at. I've seen nothing published since, in a similar vein, that I'd recommend. Don't neglect introductory statistics texts in subjects other than archaeology; they may suit your 'learning style' better.

In the past I'd recommend students to spend an hour or so browsing texts on the shelves of a university library if they had access to one. I retain an old-fasioned preference for written texts, finding them easier to dip into or exploit in a serendipitous fashion. The modern reality is to make a beeline for a search engine to see what can be found on the web, which can be very good but also highly variable in quality and reliability.

The first editions of Shennan and Drennan appeared in the late-1980s and mid-1990s and the current editions, with a few extra topics, don't differ in their conception. The style reflects their genesis; both authors are aware of the value of computational software but neither devote much space to practical implementation with specific software. This is perfectly understandable; both books were intended

to sell and the software access available to purchasers was highly variable. This has remained the case until fairly recently, but the advent and development of R has effected a radical change in the publication of applied statistics texts, as well as practice. It can seem that every other statistics text now published has a title along the lines *Statistics for Something using* R[4].

Archaeology, unless I've missed something, stills need to catch-up with statistical developments. David Carlson's web pages promise a book on R and archaeology, and his notes on the subject provide a useful introduction, designed as R-based computational accompaniments to the texts of Shennan and Drennan[5]. His site also lists some of the free introductory material that is available.

Carlson makes use of the Rcmdr package, a graphical user interface (GUI) to R, designed to make R 'easier' to use and an entrée to the full command-driven version. The development of GUIs for R seems to be something of a growth industry with none having yet achieved market dominance (Valero-Mora and Ledesma, 2012). Within the command- or script-driven paradigm, much of fairly recent origin and likely to grow in importance, are developments that simplify the writing and use of code[6]. That is, R is developing at a rapid pace so any work produced over a period of time – this one included – is going to be 'out-of-date' by the time you see it.

That said, there's nothing wrong with older texts on the subject – the basics haven't changed, just the resources in terms of user-written contributions. As far as R goes, and its 'ancestor' S, I'm 'self-taught' and much of what I know was learned from Venables and Ripley (2002) and earlier editions. Its age notwithstanding, if I could only own one text on R this would be it; I still consult it regularly. In fact I own several; Dalgaard's introduction, now in its second edition (Dalgaard, 2008), was also useful. A number of, to some extent 'competing', texts emphasizing the powerful graphical facilities in R that develop the 'traditional' graphics that come with it are also available (e.g., Murrell, 2011; Sarkar, 2008; Wickham, 2009).

As with introductory statistics texts I think it's useful to own, or at least have access to, more than one treatment of R at introductory and intermediate levels. As I don't believe in the 'one-size-fits-all' principle, and if its not a requirement for a taught course, I'd hesitate to make definitive recommendations; intelligent 'shopping around' to suit your own requirements is what's needed[7].

[4]The site http://www.r-project.org/doc/bib/R-books.html lists about 150 texts on statistics and R, about a sixth at an introductory level. New texts appear on a regular basis and the list does not include free web-based material in plentiful supply.

[5]http://people.tamu.edu/ dcarlson/quant/index.html (Accessed May 2015)

[6]See Ben Marwick's brief review at http://www.r-bloggers.com/doing-quantitative-archaeology-with-open-source-software/ (accessed May 2015) and links provided there.

[7]Apart from anything else, I should confess to not having read a lot of what's available; it's nothing to be ashamed of.

On more specific statistical topics, the multivariate methods of Chapters 7 to 11, are covered in an archaeological context and in what is intended to be an accessible manner in Baxter (1994a) (reprinted as Baxter, 2015). The style of presentation, that is the graphics, shows its age but the basics haven't changed. Baxter (2003) is a wider-ranging, more recent, survey at an intermediate level, of archaeological usage of most of the methods covered in these notes, with the exception of the material in Chapter 12. In particular a fairly detailed review of the use of regression methods in archaeology is provided; Shennan (1997) includes a more conventional and quite lengthy introductory treatment that goes somewhat beyond what is often attempted in introductions, including a brief account of log-linear models avoided here.

More topic-specific reading is suggested in the relevant chapters; there are a lot of good texts out there on specific multivariate methods and regression analysis. As with other recommendations, that you may have noticed I'm not making, I hesitate to be too specific and, once you have gained a toe-hold in a subject of particular interest, read around to see what's there and suits you. Subject-specific statistical texts abound but sometimes need to be approached warily. Some domains of study have developed their own statistical idioms (I have in mind some of the psychological, social and life sciences, among others) that, beyond the introductory level, can result in texts with methodological biases and emphases not readily transferable to or appropriate for other domains. Quantitative archaeology publications are largely exempt from this comment since archaeological use of statistics is, for the most part, within the 'statistical mainstream'[8].

[8]This is notwithstanding the 'philosophical baggage' that came with some of the early promotion of statistics in archaeology, and a later mild 1980s obsession with the need for 'concordance' between archaeological theory and statistical method. The latter is a worthy aim that resulted in a body of sometimes impenetrable writing that reinvented existing statistical wheels in shapes not necessarily fit for purpose. Clive Orton (1992: 137), in the context of methods developed for intra-site spatial analysis, memorably characterized this as 'the Audrey syndrome ("an ill-favoured thing, sir, but mine own". *As You Like It*, Act V, Scene iv' (Baxter, 2003: 7–8). This has more general application – don't ignore statistical publications just because they have been produced by statisticians.