Contents

Preface **x**

1 Introduction **1**
1.1 Introduction 1
1.2 How (not) to read these notes 5
1.3 Suggested reading 8

2 Introductory examples **11**
2.1 Introduction 11
2.2 Example – Principal component analysis 12
2.3 Example – Correspondence analysis 16
2.4 Example – Cluster analysis 18
2.5 Example – Linear discriminant analysis 21
2.6 R notes 23
2.6.1 Introduction 23
2.6.2 Aspects of labeling and presentation 23
2.6.3 Code used for analyses in the text 29

3 Continuous data **33**
3.1 Continuous data 33
3.1.1 Introduction 33
3.1.2 Histograms, dotplots and boxplots 34
3.1.3 Kernel density estimates 37
3.2 R notes 42
3.2.1 Functions 42
3.2.2 Code used for analyses in the text 45

4 Discrete data **50**
4.1 Discrete data, barplots and histograms 50
4.2 Barplots for two-way tables 54
4.3 The iniquitous pie-chart 57

4.4 R notes 59

5 Regression analysis 63
5.1 Linear regression analysis 63
5.1.1 Introduction – an example 63
5.1.2 Regression models and notation 65
5.1.3 Model checking 69
5.1.4 Inference 73
5.2 Examples 74
5.3 Non-parametric regression 79
5.4 R notes 83

6 Graphs – a miscellany 88
6.1 Introduction 88
6.2 Enhanced pairs plots 88
6.3 Graphics with more than one variable 89
6.3.1 Two-dimensional KDEs 89
6.3.2 Ellipses, convex hulls, contours – one group 91
6.3.3 Ellipses, convex hulls, contours – several groups 92
6.4 Correlation diagrams 94
6.5 Ternary diagrams 95
6.6 Chernoff faces 99
6.7 R notes 101

7 Principal component analysis 105
7.1 Introduction 105
7.2 Example 1 – Roman glass compositions 106
7.3 The idea of distance 112
7.4 Example 2 – Stone axe morphology 113
7.4.1 Definition and properties of principal components 114
7.4.2 Interrogating PCA output 115
7.5 R notes 119

8 Factor analysis and PCA 121
8.1 Factor analysis 121
8.2 Theory - a brief summary 122
8.3 Examples 124
8.3.1 PCA and rotation 124
8.3.2 Variants of factor analysis 126
8.4 Factor analysis in archaeology 129

9 Correspondence analysis **133**
9.1 Introduction 133
9.2 CA and PCA – similarities and differences 134
9.3 Romano-British glass assemblages 136
9.3.1 First- to third-century vessel glass 136
9.3.2 First- to second-century vessel glass 138
9.3.3 Second- to third-century vessel glass 139
9.4 Flavian drinking-vessels 141
9.5 Anglo-Saxon male graves and seriation 144
9.6 R Notes 147

10 Cluster analysis **148**
10.1 Introduction 148
10.1.1 Main ideas 148
10.1.2 Example – Blue medieval window glass 149
10.2 Hierarchical Clustering 152
10.2.1 The most commonly used methods 152
10.2.2 Example – Levantine glass compositions 153
10.3 Ward's method and model-based methods 158
10.3.1 Ward's method 158
10.3.2 Model-based clustering 159
10.3.3 K-means clustering 160
10.3.4 Fuzzy clustering 162
10.4 Summary 164
10.5 R notes 165

11 Discrimination and classification **169**
11.1 Introduction 169
11.2 Mahalanobis distance 170
11.2.1 MD and confidence ellipsoids 170
11.2.2 MD, outliers, and allocation to groups 171
11.3 Linear discriminant analysis – examples 172
11.3.1 Lead isotope-ratio data – three groups 172
11.3.2 Neolithic pot dimensions 174
11.3.3 Practicalities 177
11.3.4 Example - Steatite compositions 178
11.4 Classification trees 180
11.4.1 Basic ideas 180
11.4.2 Example – Steatite compositions (continued) 181
11.5 Methodology 182
11.6 Example 2 - North Apulian pottery 185

11.7 R notes 187

12 Statistical inference 190
12.1 Introduction 190
12.2 Common hypothesis tests 191
12.2.1 The normal distribution 191
12.2.2 Inference 193
12.2.3 Tests of means – t-tests 194
12.2.4 Tests of variances 196
12.2.5 Confidence intervals 197
12.3 Examples of R use 198
12.3.1 Data 198
12.3.2 t-tests 200
12.3.3 Chi-squared tests 202
12.3.4 F-tests and ANOVA 206
12.4 Some omitted topics 211
12.5 Discussion 213

References 215

A Getting R, getting started 224
A.1 Finding R 224
A.2 Data entry 224
A.3 Packages 225

B Data sets 227

C Covariance and correlation 252
C.1 Definitions and notation 252
C.2 Applications 255
C.2.1 Linear regression analysis 255
C.2.2 Principal component analysis 256
C.2.3 Mahalanobis distance and LDA 256

D More on PCA and factor analysis 259
D.1 Introduction 259
D.2 The singular value decomposition 260
D.3 The factor analysis model 261
D.3.1 The model 261
D.3.2 Rotation 263
D.3.3 Factor extraction 263
D.4 Discussion 265

List of Figures

2.1 A PCA biplot of the chemical data from Table B.1. 13
2.2 An enhanced PCA row plot of the oxide data from Table B.1. 14
2.3 Bivariate plots for variables from Table B.1. 15
2.4 Correspondence analysis plots of the data from Table B.2. 18
2.5 Ward's method cluster analysis of the data from Table B.1. 19
2.6 Single-link cluster analysis of the data from Table B.1. 21
2.7 LDA analysis plot of the data from Table B.1. 22
2.8 Enhanced LDA plot of the data from Table B.1. 24
2.9 Available symbols for plotting points in `R`. 26
2.10 Available colors for plotting in `R`. 27
2.11 Available lines for plotting in `R`. 28

3.1 Graphical displays for the lengths of early Romano-British hairpins. 35
3.2 Graphical displays for the lengths of all Romano-British hairpins. 36
3.3 Applications of KDEs to the Romano-British hairpins data. 38
3.4 KDEs for untransformed and log-transformed data to base 10. 39
3.5 Comparing two groups using histograms and KDEs. 40
3.6 Distribution of %Fe by region for the data from Table B.1. 41

4.1 Right and wrong ways of presenting a histogram. 51
4.2 Different ways of representing the data from Table 4.1. 53
4.3 Examples of barplot presentation. 55
4.4 Enhanced barplots for the data of Table 4.2. 56
4.5 Two- and three-dimensional `Excel` barplots. 56
4.6 Various forms of 'pie-chart' presentation. 58

5.1 Regression of date against body height (Table B.5). 64
5.2 Graphical displays of the stone axe distance-decay data of Table 5.1. 67
5.3 Graphical displays of the pottery distance-decay data of Table 5.2. 69
5.4 Residuals from the regression fit of Figure 5.1. 72
5.5 Diagnostics of the regression of `Date` against `BH`. 73
5.6 Post-medieval wine bottles – pairs plot. 75

5.7 Fitted regressions for the data in Table 5.1. 76
5.8 Fitted regressions for the data in Table 5.2. 77
5.9 Deviations from circularity and diameters of stone circles. 78
5.10 Deviations from circularity and diameters of stone circles with loess smooths. 80
5.11 Loess smooths for the southern circle data. 81
5.12 Non-parametric regressions for the post-medieval wine bottle data. 82

6.1 A pairs plot for the data of Table B.1. 89
6.2 One and two-dimensional KDEs for the loomweight data. 90
6.3 Confidence ellipsoid (95%), convex hull and contour plot. 91
6.4 Confidence ellipsoids for grouped data. 93
6.5 Convex hulls for grouped data. 94
6.6 Correlations represented as ellipses for stone axe dimensions. 95
6.7 Ternary plot examples – assemblage data. 97
6.8 Ternary plots for archaeozoological data. 98
6.9 Chernoff faces for the data of Table B.1. 100

7.1 PC plots of the standardized Romano-British glass data. 107
7.2 PC plots of the unstandardized log-transformed Romano-British glass data. 108
7.3 PC plots of the log-ratio transformed Romano-British glass data. 111
7.4 Score plot from a PCA of the stone axe data (standardized). 113
7.5 Coefficient plots for components 1 and 2, and 2 and 3, for the stone axe data. 117
7.6 Scree plots for the PCA of the stone axe data. 118
7.7 Pairs plots for components 2-4 from a PCA of the stone axe data. 119

9.1 Correspondence analysis of 1st==3rd century glass assemblages. 137
9.2 Correspondence analysis of 1st–2nd century glass assemblages. 138
9.3 Correspondence analysis of 2nd–3rd century glass assemblages. 140
9.4 Correspondence analysis of Flavian drinking-vessel glass assemblages. 142

10.1 Average-link cluster analysis for the medieval glass compositions of Table B.16. 149
10.2 Pairs plot from a PCA of the glass compositions of Table B.16. 151
10.3 Enhanced average-link cluster analysis for the medieval glass compositions data of Table B.16. 152
10.4 A Ward's method cluster analysis for the Levantine glass data. 154
10.5 A single-link cluster analysis for the Levantine glass data. 155
10.6 An average-link cluster analysis for the Levantine glass data. 156
10.7 PCA plots for the Levantine glass data. 157

10.8 A scree plot for the Levantine data k-means cluster analysis. 161

11.1 Confidence ellipsoids for two lead isotope ratios from two fields. 170
11.2 PCA and LDA plots for lead isotope-ratio data from three fields. 172
11.3 A pairs plot for lead isotope ratio data from three fields. 173
11.4 Measurement points of Danish Neolithic pot profiles. 174
11.5 PCA and LDA plots of Neolithic pot dimensions – three types. 175
11.6 PCA and LDA plots of Neolithic pot dimensions – two types. 176
11.7 LDA of the steatite compositional data. 179
11.8 Classification tree for steatite compositional data. 181
11.9 Error-rate plotted against tree size using `cp` for the steatite data. 184
11.10 Classification tree for northern Apulian fine ware pottery data. 186

12.1 Stripcharts of area and log-transformed area of *marae* enclosures. 199
12.2 Boxplots of maximum flake lengths by material for Table 12.5. 209

C.1 Artificial data showing a positive correlation. 253

List of Tables

2.1 Summary statistics for K from Table B.1. 16

3.1 Lengths of Romano-British copper alloy hairpins. 34

4.1 Chronology of Roman glass pillar-moulded bowls. 53
4.2 Artifact classes by strata from Bailey *et al.* (1983). 54

5.1 Frequency of Neolithic stone axes at different distances from a distribution center. 67
5.2 Frequency of Middle-Late Iron Age pottery at different distances from a source. 68

6.1 Assemblage counts from Ksar Akil. 96

7.1 Variable means (%) for the Romano-British glass and the two sites. 108
7.2 Correlations for the Romano-British glass data. 110
7.3 PC coefficients from the PCA of the stone axe data. 116
7.4 Variances and cumulative variances for the PCs for the stone axe data. 117

8.1 Varimax rotated coefficients from PCAs of the stone axe data. 124
8.2 Factor analyses of the Bronze Age fibulae data of Table B.15 – four factors. 127
8.3 Factor analyses of the Bronze Age fibulae data of Table B.15 - three factors. 129

9.1 Inertias from a correspondence analysis of 1st to 3rd century vessel glass assemblages. 137
9.2 Diagnostic statistics (columns) for the CA of 1st to 2nd century AD vessel glass assemblages. 139
9.3 Diagnostic statistics (rows) for the CA of 2nd to 3rd century AD vessel glass assemblages. 141

9.4 Diagnostic statistics from a CA of Flavian drinking-vessel glass assemblages. 143

10.1 Ward's method and k-means clustering compared. 161
10.2 Results from a c-means clustering of the Levantine data. 163
10.3 A c-means clustering of the York medieval glass data. 164

11.1 Cross-validated classification for the Danish Neolithic pot data. 177
11.2 The 'confusion' table for LDA of the steatite data. 180
11.3 Classifications from LDAs of the steatite compositional data. 180

12.1 Area of *marae* ceremonial enclosures from two regions. 199
12.2 Observed and expected frequencies of flaked stone artifacts. 203
12.3 Residuals from a chi-squared test of Table 12.2. 204
12.4 Artifical data - small samples and chi-squared tests. 205
12.5 Maximum lengths of unbroken flakes for four raw material types. 208

B.1 Romano-British pottery chemical compositions. 227
B.2 Counts of pottery types in Early Iron Age tombs. 228
B.3 Dimensions of loomweights from Pompeii I. 229
B.4 Dimensions of loomweights from Pompeii II. 230
B.5 Post-medieval wine bottle dimensions. 231
B.6 Neolithic stone 'circle' diameters. 232
B.7 Romano-British waste glass compositions I 233
B.8 Romano-British waste glass compositions II 234
B.9 Neolithic stone axe dimensions I. 235
B.10 Neolithic stone axe dimensions II. 236
B.11 Neolithic stone axe dimensions III. 237
B.12 Species percentages by site and region from Roman sites. 238
B.13 Site by type data for Romano-British vessel glass. 239
B.14 Assemblage profiles for Flavian drinking-vessel glass. 240
B.15 Measurements on Bronze Age fibulae from Münsingen, Switzerland. 241
B.16 Medieval glass chemical compositions. 242
B.17 Roman Levantine glass compositions. 243
B.18 Lead isotope-ratio data for three sources in the Aegean. 244
B.19 Dimensions of Early and early Middle Neolithic pot vessels I. 245
B.20 Dimensions of Early and early Middle Neolithic pot vessels II. 246
B.21 Steatite compositional data I. 247
B.22 Steatite compositional data II. 248
B.23 Steatite compositional data III. 249
B.24 North Apulian fineware compositions I. 250
B.25 North Apulian fineware compositions II. 251

Preface

These notes were started during an enforced period of idleness in late 2011. They were abandoned when I recovered enough to move on to other things (i.e. my day job). What was written up to that point, roughly Chapter 7 in the present format, was tidied up and made available on my academia.edu and web sites. It was always the intention to return to them, if only to complete and add a chapter on cluster analysis. This has only happened in the last year or so.

Things didn't go quite according to plan. To begin with, I was dissatisfied with aspects of what had been written and the existing chapters were subjected to major revision and/or expansion. Much of the `R` code for undertaking the analyses was also rewritten. Apart from the chapter on cluster analysis and some rearrangement, chapters on discriminant analysis, factor analysis and statistical inference (mainly hypothesis tests) were added. For reasons explained in the text – chiefly because I think they've been oversold and not delivered what was promised – I had second thoughts about including the last two but they have survived.

The original and rather unworthy genesis of these notes lay in dissatisfaction, possibly bordering on the irrational at times, with what I viewed as the widespread misuse of software such as `Excel` to produce inadequate graphs that disfigure archaeological publication. This remains manifest in the treatment of discrete data. The original intention, abandoned almost immediately, was to keep things short and restrict attention to a few topics traditionally regarded as 'simple'.

A main reason for expanding the coverage is that, in a sense I try to explain in the introduction, I'd regard the vast majority of statistical methods that have been found useful in archaeology as 'simple'. At the conceptual level, and regardless of any mathematical complexity, it is easy to explain in ordinary language what most methods are trying to do. Similarly, the compuations associated with the mathematics can be complex, and computational statistics is a field of study in its own right. However, those who have engaged in this kind of study have made the fruits of their labor widely available, and free, in software such as `R` so, from the point of view of the end-user, the execution of an analysis is also simple. Chapter 2 illustrates this by taking some of the standard methods of multivariate analysis used in archaeology – traditionally often presented as 'complex' – and shows how

usefully interpretatble output can be produced in one line of code in R. The idea underlying most of these methods is the simple one of reducing a (possibly large) table of data to a two-dimensional, archaeologically interpretable, 'map'.

This is all made possible by the availability of powerful, open-source software such as R. This is often presented as something that is 'difficult' or 'forbidding' for users reared on menu-driven software, and if potential users do find R 'difficult' then I suppose this is true, though it's also a product of perceptions fuelled by some of the literature, or preconceptions. Writing introductory textbooks on R is something of a growth industry (and there's plenty of free material) so there is considerable assistance out there for aspirant users. Worth mentioning is the fact that it's possible to do some things more rapidly in R than in some popular statistical software such as SPSS; you don't have to wade through vast amounts of irrelevant and sometimes borderline incomprehensible output; and can customize your output (i.e. graphs) to your heart's content rather than being restricted to formats determined by anonymous programmers in the dim and distant past.

I want to emphasize that these notes are *not* intended as a textbook, either as an introduction to statistics or to R. I understand a textbook (as opposed to text) to be a piece of work written, at some length, for didactic purposes, on a topic systematically developed, intended to be read in a linear fashion, and as comprehensive as possible within its chosen remit. The present offering fails, I think, on every count.

The use of the word *Notes* in the title is meant to indicate this. As discussed in the introductory chapter, one or two sections need to be taken early and in some sort of sequence, but mostly the idea is that the text can be dipped into for code that enables you to get going with the kind of data you have to hand. This is, I'm claiming, 'simple', but this is not to be equated with the idea that statistics is 'easy'. Like any other subject worth studying effective use needs to be learned, and this occurs over time with experience and by accretion. The idea here is that you need to start somewhere, and a good case can be made for learning by doing something first and worrying about what you've done later[1].

On the 'learning-by-doing' principle the notes emphazise the importance of real data analysis, and to this end a lot of data sets from different archaeological specializations are analyzed, often in more than one way. These should be available on my academia.edu pages and website[2]. These provide a useful starting point for

[1] In *La Peste* Albert Camus creates a character, Joseph Grand, who is trying to write a novel but doesn't progress beyond the first sentence because he can't perfect it. I had friends at university who never completed, or even started writing, their PhDs for similar reasons. Much better is to get something written down and tidy it up later, possibly discarding much of it in the process. The important thing is to get started.

[2] https://nottinghamtrent.academia.edu/MikeBaxter and http://www.mikemetrics.com/ respectively.

beginning, though if your own data are immediately available in a suitable format analyzing them might be more fun.

Since R has become something of an industry standard in writings on applied statistics that has penetrated many areas of application I'm rather surprised it seems not to have gained much traction in archaeology, individual endeavour and possibly what's invisible and below the surface of publication apart. I don't know of any book length introduction to R for archaeologists, though David Carlson's website promises one[3] and includes R accompaniments to the examples in the standard introductory quantitative archaeology texts of Shennan (1997) and Drennan (2009). These texts together with Carlson's work provide a good entrèe to what is attempted here, which should be seen as complementary rather than a 'competitive' approach to the subject.

One final word of warning is that I don't regard computer programming as among my competencies – it was evident from my undergraduate days that I lacked any aptitude for it. This should encourage those with a similar view of their abilities; those who do have such an aptitude will find some of the code presented inelegant, inefficient, etc. and will be able to improve it. For my own part I have a pragmatic attitude and am usually happy if something works that does what I've asked for and I understand what's happening. – that is, I treat R as an extremely useful practical tool for data analysis.

Mike Baxter, Nottingham, May 2015

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