

Quantitative Methods in Archaeology Using R

Quantitative Methods in Archaeology Using R is the first hands-on guide to using the R Statistical Computing System written specifically for archaeologists. It shows how to use the system to analyze many types of archaeological data.

Part I includes tutorials on R with applications to real archaeological data showing how to compute descriptive statistics, create tables, and produce a wide variety of charts and graphs. Data transformation and missing values are also covered. A chapter on confidence intervals and hypothesis testing introduces these statistical concepts and provides examples of several approaches including bootstrapping. Associations between variables including Chi-square tests, and simple linear regression completes the section on basic statistics. Part II addresses the major multivariate approaches used by archaeologists including multiple regression (and the generalized linear model); multiple analysis of variance and discriminant analysis; principal components analysis; correspondence analysis; distances and scaling; and cluster analysis. Part III covers specialized topics in archaeology including intra-site spatial analysis, seriation, and assemblage diversity.

David L. Carlson is a Professor of Anthropology at Texas A&M University where he has been teaching quantitative methods and the R statistical system to anthropology graduate students for eight years. His research focuses on the application of quantitative methods to discover and understand patterning in the distribution of artifacts on archaeological sites. He is a co-author of *Clovis Lithic Technology* (2011).

Cambridge Manuals in Archaeology

General Editor

Graeme Barker, *University of Cambridge*

Advisory Editors

Elizabeth Slater, *University of Liverpool*

Peter Bogucki, *Princeton University*

Cambridge Manuals in Archaeology is a series of reference handbooks designed for an international audience of upper-level undergraduate and graduate students and professional archaeologists and archaeological scientists in universities, museums, research laboratories and field units. Each book includes a survey of current archaeological practice alongside essential reference material on contemporary techniques and methodology.

Books in the series

Vertebrate Taphonomy, R. LEE LYMAN

Photography in Archaeology and Conservation, 2nd edition, PETER G. DORRELL

Alluvial Geoarchaeology, A. G. BROWN

Shells, CHERYL CLAASEN

Sampling in Archaeology, CLIVE ORTON

Excavation, STEVE ROSKAMS

Teeth, 2nd edition, SIMON HILLSON

Lithics, 2nd edition, WILLIAM ANDREFSKY, JR.

Geographical Information Systems in Archaeology, JAMES CONOLLY and MARK LAKE

Demography in Archaeology, ANDREW CHAMBERLAIN

Analytical Chemistry in Archaeology, A. M. POLLARD, C.M. BATT, B. STERN and
S. M. M. YOUNG

Zooarchaeology, 2nd edition, ELIZABETH J. REITZ and ELIZABETH S. WING

Quantitative Paleozoology, R. LEE LYMAN

Paleopathology, TONY WALDRON

Fishes, ALWYNE WHEELER and ANDREW K. G. JONES

Archaeological Illustrations, LESLEY ADKINS and ROY ADKINS

Birds, DALE SERJEANTSON

Pottery in Archaeology, 2nd Edition, CLIVE ORTON and MICHAEL HUGHES

Applied Soils and Micromorphology in Archaeology, RICHARD I. MACPHAIL and PAUL
GOLDBERG

Quantitative Methods in Archaeology Using R

David L. Carlson *Texas A & M University*



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

4843/24, 2nd Floor, Ansari Road, Daryaganj, Delhi – 110002, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107040212

DOI: 10.1017/9781139628730

© David L. Carlson 2017

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2017

Printed in the United States of America by Sheridan Books, Inc.

A catalogue record for this publication is available from the British Library.

ISBN 978-1-107-04021-2 Hardback

ISBN 978-1-107-65557-7 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

CONTENTS

<i>List of Figures</i>	<i>page xi</i>
<i>List of Tables</i>	xv
<i>List of Boxes</i>	xvii
<i>Acknowledgments</i>	xix
1 Introduction	1
1.1 Organization of the Book	6
PART I R AND BASIC STATISTICS	9
2 Introduction to R	11
2.1 First Steps Using R	11
2.2 Next Steps Using R	22
2.3 Getting Your Data Into R	26
2.4 Starting and Stopping R	28
2.5 R Functions	28
2.6 Getting Help	30
2.7 Other Ways to Use R	31
2.8 Archaeological Data for Learning R	33
3 Looking at Data – Numerical Summaries	36
3.1 Arithmetic with R	38
3.2 Four Common Distributions	42
3.3 Descriptive Statistics – Numeric	49
3.4 Descriptive Statistics Using R	51
4 Looking at Data – Tables	65
4.1 Factors in R	65
4.2 Producing Simple Tables in R	68
4.3 More Than Two Variables	72
4.4 Binning Numeric Variables	77
4.5 Saving and Exporting Tables	78

viii CONTENTS

5	Looking at Data – Graphs	85
5.1	True and False in R	86
5.2	Plotting One or Two Categorical Variables	88
5.3	One Numerical Variable	95
5.4	One Numerical Variable and One Categorical Variable	99
5.5	Two Numerical Variables	103
5.6	More Than Two Numerical Variables	109
5.7	Printing Graphs	116
6	Transformations	126
6.1	The Apply Family of Functions in R	127
6.2	Transforming Variables (Columns)	129
6.3	Transforming Observations (Rows)	136
7	Missing Values	144
7.1	Missing Values and Other Special Values in R	145
7.2	Eliminating Cases or Variables with Missing Values	147
7.3	Imputing Missing Values	150
8	Confidence Intervals and Hypothesis Testing	159
8.1	Programming R – Writing Functions	160
8.2	Confidence Intervals	162
8.3	Hypothesis Testing	169
8.4	Comparing Two Samples	171
8.5	Comparing More Than Two Samples	178
9	Relating Variables	190
9.1	Categorical Data	190
9.2	Numeric Data – Association	198
9.3	Numeric Data – Regression	204
PART II MULTIVARIATE METHODS		217
10	Multiple Regression and Generalized Linear Models	219
10.1	Multiple Regression	219
10.2	Regression with Dummy Variables	232
10.3	Generalized Linear Models – Logistic Regression	235
11	MANOVA and Discriminant Analysis	244
11.1	Hotelling's <i>T</i> and MANOVA	245
11.2	Descriptive (Canonical) Discriminant Analysis	249
11.3	Predictive Discriminant Analysis	255
12	Principal Components Analysis	265
13	Correspondence Analysis	279
14	Distances and Scaling	296
14.1	Distance, Dissimilarity, and Similarity	297
14.2	Multidimensional Scaling	303
14.3	Comparing Distance Matrices – Mantel Tests	311

15 Cluster Analysis	318
15.1 <i>K</i> -Means Partitioning	321
15.2 Hierarchical Clustering	334
15.3 Other Methods	342
PART III ARCHAEOLOGICAL APPROACHES TO DATA	347
16 Spatial Analysis	349
16.1 Circular or Directional Statistics	349
16.2 Mapping Quadrat-Based Data	358
16.3 Mapping Piece Plot Data	367
16.4 Simple Spatial Statistics	372
17 Seriation	379
17.1 Distance Matrix Ordering	381
17.2 Ordering the Data Matrix Directly	384
17.3 Detrended Correspondence Analysis	388
17.4 Principal Curves	390
18 Assemblage Diversity	397
18.1 Diversity, Ubiquity, and Evenness	399
18.2 Sample Size and Richness	403
18.3 Rarefaction Curves	408
19 Conclusions	412
<i>References</i>	415
<i>Index</i>	423

FIGURES

1	The R Project website	<i>page</i> 12
2	R Console window in Windows® using multiple document interface (MDI)	13
3	Main “HTML help” page	31
4	Four probability distributions: a. Binomial, b. Poisson, c. Normal, d. Lognormal	44
5	Darl, Ensor, Pedernales, Travis, and Wells dart points. Redrawn from Turner and Hester (1993)	53
6	Upper Paleolithic end scrapers. Redrawn from Sackett (1966)	73
7	Imported tables in Microsoft® (a) Word and (c) Excel and LibreOffice® (b) Writer and (d) Calc	80
8	Tables produced by xtable and imported as html files in Microsoft Word illustrating different design styles	81
9	Bronze fibulae. Redrawn from Doran and Hodson (1975)	88
10	Pie chart (a) and bar chart (b) showing the number of coils on bronze fibulae	90
11	Bar charts for the Upper Paleolithic end scrapers showing differing degrees of curvature: (a) stacked bars of frequency by site, (b) stacked bars of percent by site, (c) side-by-side bars of percentage grouped by site, (d) side-by-side bars of percentage grouped by curvature	92
12	Dot charts showing percentage of end scrapers with different degrees of curvature: (a) grouped by curvature, (b) grouped by curvature with group means	94
13	Histograms of fibulae lengths: (a) frequency using the default number of bars, (b) frequency setting 12 bars at 10 mm increments, (c) density setting 12 bars at 10 mm increments, (d) percentage setting 12 bars at 10 mm increments	97
14	Kernel density (a) and QQ (b) plots for fibulae lengths	98
15	Box-and-whiskers plots for dart point lengths	100
16	Box-and-whiskers (a), strip (b), violin (c), and bean (d) plots of the dart point lengths	102

xii LIST OF FIGURES

17	Scatter plots of length by bow height for the bronze fibulae: (a) original x, y plot, (b) identifying particular specimens, (c) function scatterplot(), and (d) function scatterplot() showing outliers	104
18	Texas dart point scatter plots with regression lines by point type: (a) linear scale and (b) log scales	107
19	Ways to deal with overplotting points by jittering (a), using sunflowers (b), using size of plotting symbol (c), and using counts (d)	108
20	Scatterplot matrix for three Dart point measurements	110
21	Scatterplot matrix for three Dart point measurements by type	111
22	Using scatterplot3d() on the dart points data	113
23	A conditioning plot of the dart point length and width by thickness	114
24	Triangle graph of Olorgesailie assemblages	117
25	Transformations for right-skewed data: (a) normal, (b) transform with square root, (c) transform with cube root, and (d) transform with logarithm	130
26	Transformations for left-skewed data: (a) normal distribution, (b) transform with square, (c) transform with exponential, and (d) transform with negative reciprocal	131
27	Effect of transformations on fibulae lengths: (a) original data, (b) logarithm base 10, (c) negative reciprocal square root, and (d) negative reciprocal	133
28	Boxplots for various transformations on fibulae length	134
29	Fibula length by bow height. Raw values (left) and size corrected (right)	137
30	Assemblage sizes at Olorgesailie	139
31	Olorgesailie large cutting tools and small tools	140
32	Comparison of imputation methods	156
33	Distribution of sample means from a normal distribution: (a) sample size 10 and (b) sample size 100	164
34	Ninety-five percent confidence interval for fibulae length based on (a) a t -distribution and (b) a 1,000-sample bootstrap	166
35	Bootstrapping plots for mean length of bronze fibulae	168
36	The Snodgrass site	172
37	Multiple comparison analyses of (a) Snodgrass house areas and (b) dart point widths using Tukey honest significant difference	182
38	Plot of length and width of Darl points (a) QQ plots for length (b) and width (c) of Darl points	206
39	Diagnostic regression plots for Darl length against width	209
40	Regression of width on length for Darl points with confidence and prediction bands	210
41	Sample hand axes from Furze Platt. Redrawn from Roe (1981)	220
42	Scatterplot showing length versus breadth for hand axes	223
43	Regression diagnostic plots for hand axe length using breadth as the explanatory variable	224

44	Regression diagnostic plots for hand axe length using six explanatory variables	228
45	Box–Cox plot for hand axe regression model	229
46	Regression diagnostic plots for log ₁₀ hand axe length using six explanatory variables	231
47	Comparison of actual and fitted areas of Snodgrass houses using segment and using segment and types	234
48	Logistic regression for Snodgrass houses: (a) regression lines, (b) regression plane	240
49	Biplot of canonical variates for the Romano-British pottery by region	253
50	Plot of canonical variates for Darl, Ensor, and Travis dart points	255
51	Histograms of discriminant analysis scores for Romano-British glass from two sites	258
52	Scree plot for the hand axe measurements	271
53	Biplot for the first two principal components for the hand axe data	272
54	Biplot for the second and third principal components for the hand axe data	274
55	Hand axe shapes in the four quadrats of Figure 52	275
56	Symmetric biplot for Olorgesailie correspondence analysis	288
57	Row (a) and column (b) principal biplots for Olorgesailie correspondence analysis	289
58	Symmetric biplot of Olorgesailie data excluding L4 and M2b	291
59	Battleship plot with rows and columns ordered by the first dimension in the correspondence analysis for the Olorgesailie assemblages	293
60	Four ways of measuring distances: (a) raw Euclidean; (b) standardized Euclidean; (c) principal components Euclidean; and (d) Mahalanobis distance	301
61	Locations of African Acheulean sites	304
62	Metric (a) and non-metric (b) multidimensional scaling of Acheulean sites	307
63	Shepard plot for non-metric multidimensional scaling of Acheulean data	310
64	Plot of geographic distance and assemblage distance for the Acheulean sites	313
65	Tripartite plot of Furze Platt hand axes based on Roe (1964)	322
66	Biplot of Furze Platt hand axes with three clusters defined by <i>k</i> -means	326
67	Scree plot for Furze Platt hand axes showing original data and five sets of randomized data	328
68	Three well-separated clusters	331
69	Silhouette plot (a) and clusplot (b) for well-separated clusters	331
70	Silhouette plot (a) and clusplot (b) of Furze Platt hand axes	333
71	Hierarchical cluster analysis of Romano-British pottery	336
72	Scree plot for Romano-British Pottery data with hierarchical clustering	340
73	Silhouette plot (a) and clusplot (b) for Romano-British pottery	342

xiv LIST OF FIGURES

74	Plot of Ernest Witte burial direction (a) and looking (b) showing circular mean (black) and the arithmetic mean (gray)	354
75	Wind rose plots of Ernest Witte burial direction and looking by group	357
76	Plot of debitage counts from the Barmose I site: (a) unit boundaries and axes included; (b) block boundary only	361
77	Choropleth map for Barmose I debitage: (a) equal count bins; (b) equal range bins	363
78	Dot density map of the Barmose I debitage	365
79	Simple contour maps of Barmose I debitage: (a) simple contour map; (b) density map; (c) 3-D perspective map	366
80	Piece plot map of Barmose I artifacts	369
81	Piece plot map of Barmose I artifacts by type	370
82	Kernel density map	371
83	Plot of the nearest neighbor distribution function (G)	373
84	K -means clusters identified with convex hulls	376
85	Battleship curve for Pueblo San Christobal	382
86	Nelson data with re-ordered rows and types	385
87	Bertin plot of Hodson's Munsingen data	387
88	Detrended correspondence analysis of Nelson data	389
89	Detrended correspondence analysis of the Munsingen data	391
90	Correspondence analysis of the Nelson data with principal curve showing missed horseshoe (a) and corrected version by identifying three points (b). Gray circles show the points used to adjust the principle curve	392
91	Correspondence analysis of the Munsingen data with principal curve showing missed horseshoe (a) and corrected version by identifying three points (b). Gray circles show the points used to adjust the principle curve	394
92	Summary of Early Stone Age site assemblage diversity: Pielou's J index of evenness by the Shannon diversity index (H)	403
93	Richness (S) by sample size (N) with logarithmic, power, and asymptotic curve fits	405
94	Rarefaction curves for Early Stone Age Sites based on combined (a) and largest assemblage (b)	410

TABLES

1	Functions introduced in Chapter 2	<i>page</i> 34
2	Functions introduced in Chapter 3	63
3	Percentage of end scraper varieties in each site	83
4	Functions introduced in Chapter 4	84
5	Functions introduced in Chapter 5	124
6	Functions introduced in Chapter 6	143
7	Functions introduced in Chapter 7	157
8	Types of errors in statistical inference	171
9	Functions introduced in Chapter 8	189
10	Functions introduced in Chapter 9	214
11	Functions introduced in Chapter 10	243
12	Functions introduced in Chapter 11	264
13	Functions introduced in Chapter 12	278
14	Functions introduced in Chapter 13	295
15	Rank order of distances for different measures of distance	302
16	Functions introduced in Chapter 14	317
17	Functions introduced in Chapter 15	345
18	Functions introduced in Chapter 16	377
19	Functions introduced in Chapter 17	396
20	Functions introduced in Chapter 18	411

BOXES

1	The <code>read.csv()</code> function in the base package	<i>page</i> 33
2	The <code>numSummary()</code> function in the <code>RcmdrMisc</code> package	61
3	The <code>Desc()</code> function in the <code>DescTools</code> package	62
4	The <code>xtabs()</code> function in the <code>stats</code> package	81
5	The <code>fTable()</code> function in the <code>stats</code> package	82
6	Graphics parameters, the <code>par()</code> function	119
7	Functions for plotting	122
8	One-sample, two-sample, and paired <i>t</i> -tests using function <code>t.test()</code>	187
9	Analysis of variance using <code>aov()</code> and multiple comparisons using <code>TukeyHSD()</code> and <code>glht()</code>	188
10	Chi-square tests, the <code>chisq.test()</code> function	212
11	Linear least squares regression, the <code>lm()</code> function	213
12	The generalized linear model function, <code>glm()</code>	242
13	Linear discriminant analysis using <code>lda()</code>	263
14	Principal components with <code>prcomp()</code>	277
15	Correspondence analysis with <code>ca()</code> in package <code>ca</code>	294
16	Multidimensional scaling with <code>cmdscale()</code> in package <code>stats</code> and <code>isoMDS()</code> in package <code>MASS</code>	315
17	Mantel tests using <code>mantel()</code> in package <code>vegan</code>	316
18	<i>K</i> -means clustering with <code>kmeans()</code>	343
19	Hierarchical clustering with <code>hclust()</code>	344

ACKNOWLEDGMENTS

I first began teaching anthropology students how to use the R statistical computing system in my graduate quantitative methods class in 2008, having used SAS, SPSS, and various other programs before. As open source and freely available worldwide for Windows, MacOS, and Linux, R was an obvious alternative. It is also the only software that readily accomplishes all of the quantitative analyses that archaeologists routinely employ. Anything that is not included in the base R system is covered in one of several thousand packages.

My interest in quantitative methods spans several decades beginning in a seminar taught by J. Ned Woodall at Wake Forest University. We were required to read a minimum number of articles or book chapters each week and I ran across a large book in the library that I thought would keep me occupied for several weeks. *Analytical Archaeology* by David L. Clarke did exactly that and illustrated many ways that quantitative approaches could facilitate the analysis of archaeological data. Later at Northwestern University, I studied with Bob Vierra, James A. Brown, and Stuart Struever and began applying quantitative methods to the Koster site for my dissertation research along with fellow students including John Hewitt, Sarah Neusius, and Mike Wiant.

At Illinois State University, I worked with Ed Jelks and Fred Lange learning to apply quantitative approaches to historic and prehistoric sites investigated during cultural resources investigations. That research continued at Texas A&M with Harry Shafer, Vaughn Bryant, Jr., and D. Bruce Dickson. It also involved a wide range of projects ranging from reservoir surveys and excavations in Texas (in collaboration with Kate Mueller Wille, Joe Saunders, and Alston Thoms) to a nineteenth-century sugar plantation in Mexico (with students Alan Meyers and Sam Sweitz). With Ron Bishop, W. Dennis James, M James Blackman, and Shawn Carlson, I was able to learn more about the analysis of ceramic compositions.

XX ACKNOWLEDGMENTS

More recently, I collaborated with Michael Waters and Charlotte Pevny in analyzing the Clovis component in an excavation at the Gault Site in *Clovis Lithic Technology*. I am currently working with Michael Alvard on Dominican fishing activities using R to identify different fishing techniques from global positioning system (GPS) locational data.

While applying R to archaeological problems in my research, I also began to develop teaching materials for other archaeologists in the form of companion guides for introductory statistics texts for archaeologists, including Stephen Shennan's *Quantifying Archaeology* and Robert Drennan's *Statistics for Archaeologists*. I also began pulling data sets included in Michael Baxter's *Exploratory Multivariate Analysis in Archaeology* and *Statistics in Archaeology* to make them more easily available to my students. Those data sets and others are incorporated into the R package, `archdata`, that is used throughout the book. I have also benefited by participating in the `r-help` mailing list where an amazing range of people ask and answer questions about how to use R in their research.

I am indebted to all of these people for stimulating my interest in quantitative methods and suggesting interesting ways to apply them to archaeological data. I also appreciate the patience of my wife, Shawn, over the last several years that this book has been in development and my repeated promises that it was “almost” done.

The anonymous reviewers of the original manuscript lead to substantial improvements in the final product and the staff at Cambridge University Press, including Beatrice Rehl, Asya Graf, and Edgar Mendez, have been patient and supportive during the process.

My thanks also to Marion Coe who created Figures 5, 6, 9, and 41.