

# Index

- analysis of variance (ANOVA), *see* hypothesis tests
- Bayesian, 145, 160, 165, 211
- biplots, *see* graphical methods
- bootstrapping, 92, 99, 141
- boxplots, *see* graphical methods
- chi-squared tests, 196
  - examples
    - contingency tables, 202–206
  - for small samples
    - Fisher’s test as alternative, 205
  - of no association, 202
  - of variance, 196
  - Yates continuity correction, 205
- classification trees, **186**
  - binary splitting, 182
  - cost complexity pruning, 184
  - nodes
    - purity of – Gini index, 183
    - types of, 182
- cluster analysis, 18–21, 91, **148–165**
  - average link, 153
  - Bayesian, 165
  - dendrogram, 19, 150, 153
    - manipulation in R, 167
  - fuzzy clustering, 162–164
  - hierarchical, 152–153
    - average-link, 150–151, 154
    - complete link, 153
    - single link, 152
    - single-linkage, 153
    - Ward’s method, 153
  - k-means, 160–162
    - scree plots, 161
  - model-based, 158, 159
    - and Ward’s method, 159
  - similarity, 152
  - single link, 20
  - Ward’s method, 19, 92, 158
- confidence ellipsoids, *see* graphical methods
- confidence intervals, 197–198
  - in *t*-tests, 200
  - relation to hypothesis tests, 198
- contingency tables
  - analysis of, 202–206
- correlation, *see also* covariance and regression analysis, 70
  - biplot approximation, 13
  - definition of, 254
  - diagrams, 94, 115
- correspondence analysis, 16–18, **133–147**
  - and PCA, 134–136
  - biplots, 17–18
  - convex hulls, 92
  - history of, 133
  - inertia, 135
  - mass, 135
  - seriation, 17, 31, 133
    - chronological, 144–147
    - spatial, 141–144
- covariance, *see also* correlation
  - definition of, 252
  - matrix, 254
  - matrix, 261

- cross-validation
  - in classification trees, 184
  - in linear discriminant analysis, 172, 176, 179
- data set analyses
  - Marae* enclosure size
    - ANOVA, 207
    - F-test of variances, 202
    - one-sample *t*-test, 199
    - stripcharts, 198
    - two-sample *t*-test, 201
  - assemblages from Ksar Akil
    - ternary diagrams, 96–97
  - Bronze Age fibulae
    - factor analysis, **126–129**
  - flake frequencies by material and site
    - chi-squared, 204–206
  - flake length by material
    - ANOVA, 208–211
    - boxplots, 208
    - Shapiro-Wilk normality test, 210
  - Flavian drinking vessels
    - correspondence analysis, 142–144
  - glass vessel assemblages
    - correspondence analysis, 136–141
  - hairpin lengths
    - boxplots, 34
    - dotplots, 34
    - histograms, 34
    - kernel density estimates, 37–42
  - Kastritsa assemblages
    - barplots, 54
  - lead isotope ratios
    - confidence ellipsoids, 170
    - linear discriminant analysis, 172–173
    - pairs plot, 172
  - Levantine glass compositions
    - cluster analysis, 153–158
    - fuzzy clustering, 162
  - k-means clustering, 160–162
  - loomweight dimensions
    - barplots, 51
    - confidence ellipsoids, 91–93
    - contour plots, 91–93
    - convex hulls, 91–93
    - histograms, 51
    - kernel density estimates, 89–90
  - Medieval glass compositions
    - cluster analysis, 149–151
    - fuzzy clustering, 163–164
  - Neolithic pot dimensions
    - discriminant analysis, 175–176
    - principal component analysis, 175
  - Philistine tomb assemblages
    - correspondence analysis, 16–18
  - pillar-moulded bowl chronology
    - barplots, 52
  - polished stone axe dimension
    - rotated PCA, 124–126
  - polished stone axe dimensions
    - correlation diagrams, 94
    - PCA, 113–119
  - pot frequency – distance decay
    - linear regression, 68, 76–77
  - regional distribution of monuments
    - pie-charts, 57
  - Roman glass compositions
    - PCA, 106–112
  - Roman pottery compositions
    - biplot, 12
    - Chernoff faces, 99
    - cluster analysis, 18–21
    - linear discriminant analysis, 21–22
    - pairs plots, 88
    - principal component analysis, 12–16
    - summary statistics of, 16
  - steatite source characterization
    - classification trees, 181–182

- linear discriminant analysis, 178–180
- stone axe frequency – distance decay
  - linear regression, 67, 75–76
- stone circle diameters
  - linear regression, 77–79
  - non-parametric regression, 80–81
- wine bottle dimensions
  - non-parametric regression, 81–82
  - pairs plots, 74
  - regression analysis, 64–65, 72–74
- zooarchaeology species assemblages
  - ternary diagrams, 97–99
- data sets
  - Marae* enclosure size
    - prehistoric Polynesia, **198**
  - analyses of, *see* data set analyses
  - assemblages from Ksar Akil
    - Palaeolithic – Lebanon, 96
  - Bronze Age fibulae
    - Bronze Age Switzerland, **241**
  - fineware compositions
    - late-antique/early-medieval Italy, **250**
  - flake frequencies by material and site
    - pre-Hispanic Mexico, **203**
  - flake length by material
    - Late Archaic Mexico, **207**
  - Flavian drinking vessel assemblages
    - Romano-British, **240**
  - Flavian drinking-vessels
    - Romano-British, **142**
  - glass vessel assemblages
    - Romano-British, **239**
  - hairpin lengths
    - Romano-British, 34
  - Kastritsa assemblages
    - Palaeolithic, Greece, **54**
  - lead isotope ratios
    - Bronze Age Aegean, **244**
  - Levantine glass compositions
    - Roman Levant, **243**
  - loomweight dimensions
    - Roman Italy (Pompeii), **229–230**
  - Medieval glass compositions
    - Medieval British, **242**
  - Neolithic pot dimensions
    - Neolithic Denmark, **245–246**
  - Philistine tomb assemblages
    - Early Iron Age, **228**
  - pillar-moulded bowl chronology
    - Romano-British, **53**
  - polished stone axe dimensions
    - Neolithic Italy, **235–237**
  - pot frequency – distance decay
    - Iron Age British, **68**
  - regional distribution of , **57**
  - Roman pottery compositions
    - Romano-British, **227**
  - Roman waste glass compositions
    - Romano-British, **233–234**
  - steatite source characterization
    - Prehistoric North America, **247–249**
  - stone axe frequency – distance decay
    - Neolithic British, **67**
  - stone circle diameters
    - British Neolithic, **232**
  - wine bottle dimensions
    - Post-medieval British, **231**
  - zooarchaeology species assemblages
    - Roman - various countries, **238**
- data sets analysis
  - Anglo-Saxon burials
    - correspondence analysis, 144–147
- data transformation
  - centered, 106
  - in cluster analysis, 20, 150
  - in PCA, 12, 106–107
  - log-ratios, 109–111
  - logarithmic, 38

- and normality, 107
- in regression analysis, 67–68, 75–77
- to improve symmetry, 36, 198
- rank, 109
- square-root, 199
- terminological confusion, 107
- to [0,1] range, 109
- data types
  - abundance matrix, 144
  - compositional
    - sub-compositional, 110
  - contingency tables, 50
  - continuous
    - interval-scaled, 33
    - or qualitative, 33
    - ratio-scaled, 33
  - cross-classified, 50
  - cross-tabulated, 50
  - discrete, 33, 50
  - incidence matrix, 144
  - matrix
    - dimensionality, 12
    - terminology, 12
  - ordinal, 52
- dendrogram, *see* cluster analysis
- descriptive statistics
  - geometric mean, 110
  - interquartile range (IQR), 16, 36
    - in boxplots, 35
  - mean, 192
    - limitations, 105
  - mean (arithmetic), 16, 42
  - measures of dispersion, 16
  - measures of location, 16
  - median, 16, 42
    - in boxplots, 35
  - modes, 89
    - bimodality, 37
    - multimodality, 38
    - unimodal, 34
  - standard deviation, 16, 192
  - standard deviation., 42
  - standard error, 192
- discriminant analysis, 21–22, **169–189**
  - classification rule
    - leave-one-out method, 175, 257
    - resubstitution, 175
  - quadratic, 178
  - variable selection, 173
- distance
  - and multivariate methods, 169
  - chi-squared, 135
    - in correspondence analysis, 169
    - weighted Euclidean distance, 135
  - Euclidean, 112, 150, 171
    - and principal components, 169
    - squared, 158
  - Mahalanobis, 170–172, 256–257
    - confidence ellipsoids, 171
    - definition of, 256
    - in linear discriminant analysis, 169, 257–258
    - outliers, 171–172
- distributions
  - t* distribution, 194
  - chi-square ( $\chi^2$ ), 196
  - F, 196
  - normal, 34
    - assumptions in linear discriminant analysis, 177
    - bivariate, 92
    - in cluster analysis, 159
    - inference, 193
    - properties of, 191–192
    - standard, 192
- dotplots, *see* graphical methods
- eigenvalues, 261
  - of principal components, 114
- F-tests

- comparison of means
  - analysis of variance (ANOVA), 206–211
- examples
  - equality of variances, 202
  - of variances, 196–197
- factor analysis, **121–132**
  - component selection
    - Kaiser’s rule, 127
  - distributional assumptions, 264
  - factor extraction, 263–265
    - least squares methods, 265
    - maximum-likelihood, 127, 264
    - minres, 128
    - principal axis, 127, 264
  - factor selection
    - Kaiser’s rule, 266
  - model, 261–262
  - rotation, 263
    - oblimin, 127
    - oblique, 127, 263
    - orthogonal, 127, 263
- functions, **42–45**
  - kmeans, 161
  - IQR, 30
  - TukeyHSD, 209
  - abline, 28, 32, 64
  - aov, 208
  - apply, 44
  - as.dendrogram, 166
  - axis, 189
  - barplot, 52, 59–61
    - arguments to, 61
    - legend, 62
    - used in `screeplot`, 120
  - bartlett.test, 210
  - biplot, 12, 14, 225
  - boxplot, 46, 48
  - ca, 18, 136
  - cbind, 42, 44, 188
  - chisq.test, 204
  - chull, 92, 101
  - clr, 120
  - cmeans, 162, 168
  - colors, 26
  - colours, 26
  - contour, 101
  - cooks.distance, 83
  - corresp, 17, 31
  - cutree, 150, 166
  - c – combine, 25
  - dataEllipse, 187
  - dendrapply, 167
  - density, 37, 47
  - dim, 30
  - dist, 20
  - dotplot.mtb, 46
  - edit, 42
  - eqscplot – equal scaling, 15, 21, 30, 32
  - exp, 84
  - faces, 99
  - fanny, 162
  - fa, 127, 260
    - defaults, 265
  - fisher.test, 205
  - for, 44
  - hclust, 20, 166
  - hist, 46, 51, 59
  - ifelse, 189
  - influence.measures, 83
  - jitter, 83
  - kde2d, 91
  - kmeans, 162, 168
  - kruskal.test, 212
  - lda, 21, 178, 187
    - with cross-validation, 189
  - legend, 28
    - in `barplot`, 62
    - placement of, 166

length, 44  
letters, 32  
library, 17, 225  
lines, 28, 83  
list, 43, 61  
lm.influence, 83  
lm, 64  
loadings, 128  
loess.smooth, 86  
loess, 86, 87  
log10, 47, 84  
log, 47  
mean, 30  
median, 30  
names, 44  
oneway.test, 207  
order.dendrogram, 167  
p.chisq, 204  
pairs, 15, 74, 166  
par, 48, 60  
  oma argument, 166  
  xpd argument, 166  
pf, 197  
pie3D, 62  
pie, 62  
plot.acomp, 97, 104  
plotcp, 184  
plot, 25  
  in cluster analysis, 150  
  using hclust, 20  
  using rpart, 183  
pnorm, 192  
points, 83, 85  
prcomp, 12, 29, 187, 261  
predict, 87  
  using lda, 22, 24, 178, 187  
principal, 125  
print, 30  
  for loadings, 128  
pt, 195  
qchisq, 204  
qf, 197  
qt, 195  
rbind, 44, 187  
read.csv, 225  
read.delim, 225  
read.table, 225  
rep, 17, 25  
round, 30, 43  
row.names, 44  
rpart.control, 183  
rpart, 183  
  complexity parameter cp, 183  
  scale, 20  
scatterplot3d, 97, 104  
scatterplotMatrix, 15  
scatterplotmatrix, 88  
screeplot, 120  
sd, 30  
seq, 44  
shapiro.test, 210  
sm.density, 101  
stripchart, 46, 48  
studres, 83  
summary, 64  
t.test, 200  
table, 44  
text, 30, 31  
  using rpart, 183  
triangle.plot, 97, 98, 104  
triax.plot, 97, 104  
truehist, 47  
t – for ‘transpose’, 61  
var.test, 200  
varimax, 125  
vioplot, 37, 48  
wilcox.test, 212  
win.graph, 46  
graphical methods  
  barplots, 54–57

- and ternary diagrams, 99
- difference from histograms, 51
- biplots
  - interpretation of, 134
- boxplots, 41
  - definition, 35
  - outliers, 35
  - unsuited to multimodal data, 35
- Chernoff faces, 99
- contour plots, 91–93
- convex hulls, 91–93
  - peeling, 93
- correlation diagrams, 94
- discrete data, 57
- dotplots, 34
- ellipses, 91–93
  - confidence ellipsoids, 91–93, 170–171, 258
  - correlation diagrams, 94
  - normality assumptions, 171
- histograms, 34–37
  - bin-widths, 34
  - definition, 34
  - difference from barplots, 51
  - issues of comparison, 39
  - probability scale, 34
  - unequal bin-widths, 52
- kernel density estimates
  - bandwidth, 37
  - multi-group comparison, 39–42
  - smoothing issues, 37
  - univariate, 37–42
- pairs plots, 88–89
  - in PCA, 118
- pie-charts, 57–59
  - three-dimensional, 58
- scree plots, 117, 161
- stripcharts, 41, 198
- ternary diagrams, 95–99
  - seriation, 96
- violin plots, 36, 41
- histograms, *see* graphical methods
- hypothesis tests, **190–214**
  - $p$ -value, 192
  - $t$ -tests, 194–196
  - alternative hypothesis, 193
  - and confidence intervals, *see* confidence intervals
  - chi-squared, *see* chi-squared tests
  - critical values, 194
  - F-tests, *see* F-tests
  - Fisher’s exact test, 205
  - null hypothesis, 193
  - one-sided, 193
  - power, 194
  - significance, 193
  - two-sided, 193
  - type I error, 194
  - type II error, 194
- Kaiser’s rule, 117, 125, 266
- kernel density estimates, *see also* graphical methods
  - 2-dimensional, 89–90
    - contour plots, 90
    - perspective plots, 90
  - and pairs plots, 88
  - confidence bands, 90
- lead isotope ratios, 172–173
- leave-one-out methods, 171
- maximum-likelihood estimation, 159
- measure of dispersion, 16
- measure of location, 16
- multivariate methods, 11
  - Chernoff faces, 99
  - classification trees, *see* classification trees
  - cluster analysis, *see* cluster analysis

- correspondence analysis, *see* correspondence analysis
- discriminant analysis, *see* discriminant analysis
- factor analysis, *see* factor analysis, *see* principal component analysis
- principal component analysis, *see* factor analysis, *see* principal component analysis
- outliers, 34
  - and Mahalanobis distance, 171–172
  - in ANOVA, 210
  - in boxplots, 35, 36
  - in cluster analysis, 163–164
  - in dotplots, 34
  - in histograms, 34
  - in loomweight dimensions, 89
  - in regression analysis, 65, 68
    - misuse of terminology, 71
  - in stripcharts, 199
- p*-value, *see* hypothesis tests
- packages
  - Hotelling, 120
  - MASS, 17, 21, 47, 83, 91, 101, 102, 147, 178, 187, 225
  - Rcmdr, 9
  - ade4, 97, 98, 104
  - aplpack, 99
  - car, 88, 118, 187
  - ca, 136, 147
  - cluster, 162
  - compositions, 97, 104
  - dendextend, 167
  - dendroextras, 167
  - e1071, 162
  - ellipse, 95, 101, 102
  - mclust, 159
  - plotrix, 46, 62, 97, 104
  - psych, 125, 260
  - rpart, 183
  - scatterplot3d, 104
  - sm, 90, 101
  - stats, 225
  - vioplot, 37
- pairs plots, *see* graphical methods
  - also called scatterplot matrices, 88
- pie-charts, *see* graphical methods, 59
- principal component analysis, 12–16, 91, **105–120**, 256
  - and cluster validation, 150, 156–157
  - and correspondence analysis, 134–136
  - and factor analysis, 115, 121–132
  - biplots, 12
    - interpretation of, 13, 109
  - coefficients
    - loadings* as alternative term, 114
    - constraints on, 114
  - component selection, 118
    - Kaiser’s rule, 117, 125
  - data standardization, 12
  - definition and properties of, 114–115
  - distributional assumptions, 264
  - equal scaling of axes, 14
  - interpretation of PCs, 115
  - outliers, 15
  - rotation, 115, 124–126
    - varimax, 125
  - size and shape interpretation, 94, 115–116, 125
- provenance studies, 91
- R, *see* software
- rants
  - don’t use pie-charts, 59
  - misnaming chart types, 51
  - misuse of 3-D plots, 56
- regression analysis, **63–87**
  - coefficient of determination, 70
  - inference, 73–74



- least squares estimates, 69
- linear models, 63–66, 74, 83, 255–256
  - dummy variable use, 77–79
- linearizable models, 66–68, 75–77
  - exponential, 68, 84
  - power law, 67, 84
- model checking, 69–73
- non-parametric, 79–82, 85–87
  - loess smoothing, 80–82, 86–87
- residuals
  - properties of, 71
  - terminology varies, 71
- scree plots
  - in k-means cluster analysis, 161, 168
  - in PCA, 117
  - limitations of, 162
- seriation, 141–147
  - using ternary diagrams, 96
- singular value decomposition, 123, 260
- software
  - CLUSTAN, 158
  - Excel, 56
    - barplots – problems with, 56
    - import data from, 224
  - MINITAB, 46, 71
- R
  - access to, 224
  - data entry, 224
  - functions, 225, *see* functions
  - initial examples, 11
  - missing data – use of NA, 225
  - packages, 225, *see* packages
  - using functions, 42–45
- R graphics
  - labels in plots, 25
  - legend construction, 28
  - line types in plots, 28
  - plotting symbols, 25
  - use of color, 26
- SAS, 173
- SPSS, 173
  - factor analysis and PCA, 114, 122, 266
- S, 9
- spatial clustering, 160
- supervised learning methods
  - classification trees, 180
  - discriminant analysis, 169
- t*-tests, *see also* hypothesis tests
  - examples, 200–202
- ternary diagrams, *see* graphical methods
- unsupervised learning methods, 169
  - cluster analysis, 169
  - correspondence analysis, 169
  - principal component analysis, 169