A.1 Software for Categorical Data Analysis

The free software \( R \), for statistical computing and graphics, is of increasing popularity and usage (\( R \) web site: http://www.r-project.org/). Many researchers support their published papers with the related \( R \) code. This way, \( R \) software is continuously updated and one can find a variety of functions for basic or advanced analysis of categorical data and special types of them. \( R \) language and environment is similar to \( S \) and code written for S-Plus runs usually under \( R \) as well. Furthermore, standard statistical packages, such as SAS, SPSS, and Stata, are well supplied to treat categorical data. Especially in their updated versions, their features concerning categorical data analysis are enriched. They incorporate procedures for applying the recently developed methods and models in categorical data analysis following the new computing strategies. Briefly, one could say that their major new features concern mainly options for exact analysis and analysis of repeated categorical data. Thus, NLMIXED of SAS fits generalized linear mixed models while GEE analysis for marginal models can be performed in GENMOD. SPSS offers the “generalized estimating equations” sub-option under the “GLMs” option. The related \( R \) function is \texttt{gee()}.

For categorical data analysis with SAS, we refer to Stokes et al. (2012) while a variety of SAS codes are presented and discussed in the Appendix of Agresti (2007, 2013). Advanced models are fitted in \( R \) using special functions, developed individually, and included in different libraries. Orientated toward categorical data analysis and models for ordinal data as well are the libraries \texttt{MASS} (Venables and Replay) and \texttt{VGLM}, \texttt{VGAM} developed by Yee (2008). For example, generalized linear mixed models can be fitted through the \texttt{glmmPQL()} function of the \texttt{MASS} library.

Other software, as BMDP, Minitab, and SYSTAT, have also components for categorical data inference.
Bayesian analysis of categorical data can be carried out through WINBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), which is a free software. Another option is to perform categorical data analysis through MATLAB, as Johnson and Albert (2000). The MATLAB functions they used are described in their Appendix.

For categorical data analysis, there have been developed also some special packages. Thus, exact analysis of categorical data is performed by StatXact while exact conditional logistic regression can be fitted by LogXact. SUDAAN is specialized for analysis of mixed data from stratified multistage cluster designs. It has also the feature of analyzing marginal models for nominal and ordinal responses by GEE. Software tool for estimating marginal regression models is also MAREG.

Finally, some algorithms may be found in Fortran. For example, Haberman (1995) provided a Fortran program for fitting the association model RC($K$) by the Newton–Raphson method while Ait-Sidi-Allal et al. (2004) implemented their algorithms for estimating parameters in association and correlation models also in Fortran.

### A.2 Contingency Table Analysis with R

All procedures and models discussed in this book are worked out in R, in a fashion aiming that even readers not familiar with R will be able to apply in practice all the models discussed here, even the nontrivial ones, fast and directly. A web companion of the book serves this goal. This section of the Appendix is basically the content description of the web companion of the book, to be found under

http://cta.isw.rwth-aachen.de

#### A.2.1 R Packages for Contingency Table Analysis

An extensive list of special R packages, useful in the analysis of contingency tables, is provided in the web appendix.

#### A.2.2 Data Input in R

Alternative forms of defining contingency tables data in R are presented (matrix(), array(), and data.frame()) and transformations from one type to another are illustrated. Ways of entering or reading data are discussed.
A.3 R Functions Used

The R functions constructed for the descriptive and inferential needs of this book are given in the corresponding section of the web appendix, organized by chapter of their first use.

A.3.1 R Functions of Chap. 1

- Binomial–Normal Distribution Graph: `bin_norm()`

A.3.2 R Functions of Chap. 2

- Likelihood Ratio Statistic for Testing Independence in Two-way Contingency Tables: `G2()`
- Odds Ratio for a 2 × 2 Table: `odds.ratio()`
- Local Odds Ratios for an $I \times J$ Table: `local.odds.DM()`
- Global Odds Ratios for an $I \times J$ Table: `global.odds.DM()`
- Cumulative Odds Ratios for an $I \times J$ Table: `cum.odds.DM()`
- Continuation Odds Ratios for an $I \times J$ Table: `cont.odds.DM()`
- Linear Trend Test: `linear.trend()`
- Midrank Scores Computation: `midrank()`
- Fourfold Plots for the Local Odds Ratios of an $I \times J$ Table: `ffold.local()`

A.3.3 R Functions of Chap. 3

- Breslow–Day–Tarone Test of Homogeneous Association: `BDT()`
- Woolf’s Test of Homogeneous Association: `woolf()`

A.3.4 R Functions of Chap. 5

- Independence (I) Model for Two-way Contingency Tables: `fit.I()`
- Quasi-Independence (QI) Model for Two-way Contingency Tables: `fit.QI()`
A.3.5 \textbf{R Functions of Chap. 6}

- Scores’ Rescaling to Obey the Weighted Constraints (6.17): \texttt{rescale( )}
- Uniform (U) Association Model: \texttt{fit.U( )}
- Row Effect (R) Association Model: \texttt{fit.R( )}
- Column Effect (C) Association Model: \texttt{fit.C( )}
- Row–Column (RC) Association Model: \texttt{fit.RC( )}
- \texttt{RC}(M) Association Model: \texttt{fit.RCm( )}
- Plotting the Row and Column Scores in Two Dimensions: \texttt{plot_2dim( )}

A.3.6 \textbf{R Functions of Chap. 9}

- \((1 - a)100\%\) Asymptotic Confidence Interval for the Difference of Correlated Proportions: \texttt{McNemar.CI( )}
- Factors Needed to Fit Symmetry Models on an \(I \times I\) Table in \texttt{glm}: \texttt{SYMV( )}
- Scores’ Rescaling to Satisfy Constraints (9.38): \texttt{rescale.square( )}

A.4 \textbf{Contingency Table Analysis with SPSS}

The association and symmetry models cannot be fitted directly in SPSS through the options of the windows commands. Association models that are GLM can be fitted through the GLM option by defining the appropriate vectors, as explained in Sect. 6.6. For all two-way association models (RC(M) included, which is nonlinear and thus cannot be fitted in GLM) and the symmetry models, we provide appropriate syntax codes to be fitted in SPSS MATRIX.

In particular, we provide MATRIX codes for:

- Independence for two-way tables using SPSS MATRIX
- Association models for two-way tables (uniform (U), row effect (R), column effect (C), and RC (M) association models)
- Symmetry models
References


M. Kateri, Contingency Table Analysis: Methods and Implementation Using R, 275
Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4811-4,
© Springer Science+Business Media New York 2014


Bhapkar, V.P.: On tests of symmetry when higher order interactions are absent. J. Ind. Stat. Assoc. 17, 17–26 (1979b)


References


Davidson, R.R., Beaver, R.J.: On extending the Bradley-Terry model to incorporate within-pair order effects. Biometrics 33, 693–702 (1977)


References

Fay, M.P.: Confidence intervals that match Fisher’s exact or Blaker’s exact tests. Biostatistics 11, 373–374 (2010a)


References


Goodman, L.A.: Total-score models and Rasch-type models for the analysis of a multidimensional contingency table, or a set of multidimensional contingency tables, with specified and/or unspecified order for response categories. Socio. Meth. 20, 249–294 (1990)


Greenacre, M.J.: Clustering the rows and columns of a contingency table. J. Classification 5, 39–51 (1988a)


Grizzle, J.E.: Continuity correction in the $X^2$ test for $2 \times 2$ tables. Am. Stat. 21, 28–32 (1967)


References


Lancaster, H.O.: The exact partition of chi-square and its application to the problem of pooling small expectations. Biometrika 37, 267–270 (1950)


Lancaster, H.O.: Canonical correlations and partitions of \( X^2 \). Quart. J. Math 14, 220–224 (1963)


References


Liang, K.Y., Self, S.G.: Tests for homogeneity of odds ratio when the data are sparse. Biometrika 72, 353–358 (1985)


References


McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157 (1947)


Pearson, E.S.: The choice of statistical tests illustrated on the interpretation of data classed in a $2 \times 2$ table. Biometrika 34, 139–167 (1947)

Pearson, K.: On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos. Mag. 5th Ser. 50, 157–175 (1900a) (Reprinted in Karl Pearson’s Early Statistical Papers ed. by E.S. Pearson. Cambridge University Press, Cambridge, 1948)


Pearson, K.: On the probable error of a coefficient of correlation found from a fourfold table. Biometrika 9, 22–27 (1913)

Pearson, K., Heron, D.: On theories of association. Biometrika 14, 186–191 (1913)


Plackett, R.L.: The continuity correction in $2 \times 2$ tables. Biometrika 51, 327–337 (1964)


296 References


References


Tarone, R.E.: On heterogeneity tests based on efficient scores. Biometrika 72, 91–95 (1985)


References


Index

A
adjacent categories odds logit model, 225
connection to R model, 225
AIC, 132, 135, 137, 140
association graphs, see hierarchical log-linear models
association model
and bivariate normal distribution, 166
and statistical evidence, 212
and stochastic ordering, 165
ANOAS, 164, 191
Bayesian approach, 267
column effect (C), 157–158
connection to CA, 207
correlation to stereotype model, 231
for global odds ratios, 199, 197–199
for multi-way tables, 181–187
homogeneous U, 187–191
for square tables (homogeneous), 248–249
generalized by $\phi$-divergence, 207–208
in R, 171–180
on local odds ratios, 180–181
linear by linear (LL), 154–155
logit equivalent, 224, 225, 231
merging categories, 211, 208–211
RC(M), 166–171
role of weights used, 165
row effect (R), 157–158
row-column (RC), 158–159
uniform (U), 154
BIC, 133, 135, 137, 138
binomial distribution, 3
in R, 4
likelihood, 9
Bowker test of symmetry, 237
Bradley-Terry model, 255–256
Breslow-Day test, 72, 140
in R, 74, 74
Breslow-Day-Tarone test, 72, 103

C
canonical correlation model, 212
collapsed categorical data, 152, 258–259
Cohen’s kappa, see rater agreement
collapsibility in multi-way tables, 113–116
complementary log-log link, 127, 229
conditional independence, 70, 76–77
and collapsing in multi-way tables, 114
for $I \times J \times K$ tables, 82–83
in graphical models, 111
log-linear models, 95, 98
Mantel-Haenszel test, 71
stratified $2 \times 2$ tables, 81–82
test conditional on homogeneous association, 102–103
conditional independence graphs, 112, 118, 121
conditional symmetry model, 240–241
in R, 242
modeling agreement, 255, 258
conditional testing, 101, 163, 164, 166, 222
confidence intervals (CI), 9
Wald CI in R, 10
confounding, 113

B
barplots, 50–52
baseline category logit model, 223
Bayesian analysis of contingency tables, 267–269

continuation odds ratios, 43
  in R, 46
continuation ratio logit model, 226
correlation model, 206–207
  and statistical evidence, 212
correspondence analysis (CA), 199–205
  connection to association models, 207
  and canonical correlation model, 212
Cressie-Read divergence, see power divergence
cumulative logit model, 223
  ordinal explanatory variables, 224
  in R, 226–227
continuous odds ratios, 42, 45
  in R, 46
  positive regression dependence, 60

decomposable models, see hierarchical log-linear models
deviance, 132–134
diagonal symmetry model, 241
  in R, 242
  modeling agreement, 255, 258
dissimilarity index, 89, 138, 183
  in R, 92
  multi-way tables, 100

ϕ-divergence, 124, 208
  association model, 207–210
  logistic model, 229
  quasi symmetry model, 257
  Fisher scoring algorithm, 130–131, 163
  Fisher’s exact test, 29–31, 57, 119, 265
  in R, 32–33
fourfold plots, 53, 83
  for 2 × 2 × K tables, 78
  for local odds ratios, 53

global odds ratios, 41, 45
  in R, 46
  modeling of, 148, 197–199
  positive quadrant dependence, 60
  graphical models, 99, 110–113, 264, 266, 269

hierarchical GLM, 135
hierarchical log-linear models, 96–98, 119
  association graphs for, 113
  conditional independence graphs for, 110, 112
  decomposable, 99, 112, 118, 119
  graphical, 111–113
  nested, 101, 105
high dimensional contingency tables, 269
homogeneity analysis, 212
hypergeometric distribution, 8, 29, 71
  non-central, 30, 58, 72

incomplete tables, see quasi independence
independence graph, see conditional independence graphs

kappa, Cohen, see rater agreement
Kullback-Leibler divergence, 124, 207

latent class models, 191, 212, 258, 264
linear trend test, 47, 83, 178, 192
  and the uniform correlation model, 207
  in R, 49–50
LL model, see association model
local odds ratios, 41, 45
  conditional for three-way tables, 67
  fourfold plots, 53
  in R, 46
  independence in terms of, 43
  marginal for three-way tables, 67
  modeling of, 147, 154, 158, 159, 168, 180–181, 196, 238, 249
  positive likelihood ratio dependence, 60
log-linear models
  Bayesian approach, 267
  connection to logit model, 216
  for multi-way tables, 97–98
  for three-way tables, 94–97
  for two-way tables, 85–88
logit model, 215–217
  connection to LL, U models, 217
  in R, 221
  connection to log-linear model, 216
  in R, 219–220
  ordinal explanatory variables, 217
longitudinal categorical data, see clustered categorical data
LR statistic $G^2$, 11, 36, 100, 123, 132

M
Mantel-Haenszel test, 71, 103
generalized for $I \times J \times K$ tables, 82–83
in R, 73
marginal homogeneity, 237–239, 241, 256
in R, 243
marginal independence, 77
in stratified $2 \times 2$ tables, 69
marginal models, 145, 151, 238, 242, 244, 259
McNemar test, 234
in R, 235
relation to Bowker test, 237
measures of association, 262–263
merging categories, see association model
MLE
association models, 161–162, 169
GLM, 129
log-linear models, 88–89, 98–99, 130
logit models, 218–219
mobility tables, 143, 150, 236, 246, 257
mosaic plots, 55–56, 83
for log-linear models, 120
for multi-way tables, 81
two-way independence model example, 138
visualizing log-linear model fit of conditional independence, 106–110
Mover-Stayer model, 257
multinomial distribution, 4–6
relation to binomial, 5
relation to Poisson, 7
multinomial-Poisson homogeneous (MPH) model, 146
in R, 146–149
multiple correspondence analysis, see homogeneity analysis

N
Newton’s unidimensional method, 162–163
Newton-Raphson algorithm, 130–131, 163
nominal variables, 1

O
odds ratio, 25–29
plots for, see fourfold plots
exact CI, 30
generalized for $I \times J$ tables, 40–44
Mantel-Haenszel estimate, 70
ordinal variables, 1
orthogonal polynomials, 192, 193, 212, 231
outliers, 121, 123
Bayesian analysis, 268

P
Pearson’s $X^2$ statistic, 11, 36, 100, 123, 132
Pearson’s divergence, 124, 207
Poisson distribution, 6–8
relation to multinomal, 7
positive likelihood ratio dependence, 60
positive quadrant dependence, 60
positive regression dependence, 60
power divergence, 123
association model, 208
statistic, 123
probit link, 127, 229
proportional logit model
adjacent categories odds, 225
connection to U model, 225
baseline category, 223
continuation ratio, 226
cumulative odds, 224
connection to cumulative R model, 224
proportional odds model (Cox’s), 224
connection to cumulative U model, 224
in R, 227–228

Q
quasi independence, 145, 150–151
for square tables, 246–248
in R, 247
modeling agreement, 255, 258
quasi symmetry model, 238–240, 256, 258
and graphical models, 264
and homogeneous association models, 248
connection to Bradley-Terry model, 255
in R, 242
modeling agreement, 255, 258
ordinal, 249

R
R model, see association model
Rasch model, 230–231, 256
rater agreement, 252–255, 257
in R, 254
on ordinal rating scales, 253–254
RC model, see association model
RC(M) model, see association model
repeated categorical data, see clustered categorical data
residuals, 38–40, 120–121, 133
   Anscombe, 150
deviance, 39, 133, 135
   in mosaic plots, 55–56, 106, 138
   in R, 39, 137
Pearsonian, 38, 92, 133, 200
   standardized, 39, 133

S
   sampling zeros, see zeros
   scores, 46–47, 83
      and stochastic ordering, 165, 195–196
      choice of, 47–48
      in association models, 154–156, 159
         graphs of, 170
      in CA, 199–202
         graphs of, 205
      in square tables, 248–252
      mid-rank scores in R, 49
      role in merging categories, 208–210
   sieve diagrams, 54, 83
      for multi-way tables, 78, 79
Simpson’s paradox, 66, 70, 114, 117–118
small samples, 29, 57, 58, 194, 229, 265–267
smoothing categorical data, 265
sparse tables, 71, 82, 119–120, 150, 194, 209,
   265–267
square tables, 233–258
   exact inference, 266
statistical evidence, 212
stereotype model, 231
stochastic ordering
   and association models, 165, 194
   and generalized odds ratios, 60–61
      Bayesian approach, 268
      in $2 \times K$ tables, 195
stratified $2 \times 2$ tables, 69–75, 81–82
      and log-linear models, 103
conditional odds ratios, 66
fourfold plots, 78
homogeneous association, 70, 72–75, 149
logit analysis, 222, 229
marginal odds ratios, 66
structural zeros, see zeros
symmetry model, 236–237
   in R, 242

T
   triangular symmetry model, see conditional symmetry model
triangular tables, 151, 246–248

U
U model, see association model

W
Woolf test, 72, 140
   in R, 74

Z
zeros
   sampling, 26, 119, 120, 145, 150, 181, 266
   structural, 119–121, 127, 142–145, 150,
   246, 266, 267