Table of contents

Acknowledgements xi

Introduction 1

Chapter 1. What is statistics? Main statistical notions and principles 7
1.1 Statistics and statistics 7
1.2 How to formulate and test your hypotheses 8
1.3 What statistics cannot do for you 15
1.4 Types of variables 16
1.5 Summary 18

Chapter 2. Introduction to R 21
2.1 Why use R? 21
2.2 Installation of the basic distribution and add-on packages 22
2.3 First steps with R 24
2.4 Main types of R objects 30
2.5 RStudio 32
2.6 Importing and exporting your data and saving your graphs 32
2.7 Summary 40

Chapter 3. Descriptive statistics for quantitative variables 41
3.1 Analysing the distribution of word lengths: Basic descriptive statistics 41
3.2 Bad times, good times: Visualization of a distribution and detection of outliers 50
3.3 Zipf's law and word frequency: Transformation of quantitative variables 62
3.4 Summary 68

Chapter 4. How to explore qualitative variables: Proportions and their visualizations 69
4.1 Frequency tables, proportions and percentages 69
4.2 Visualization of categorical data 72
4.3 Basic Colour Terms: Deviations of Proportions in subcorpora 80
4.4 Summary 85

Chapter 5. Comparing two groups: *t*-test and Wilcoxon and Mann-Whitney tests for independent and dependent samples 87
5.1 Comparing group means or medians: An overview of the tests 87
5.2 Comparing the number of associations triggered by high- and low-frequency nouns with the help of the independent *t*-test 89
5.3 Comparing concreteness scores of high- and low-frequency nouns with the help of a two-tailed Wilcoxon test 104

5.4 Comparing associations produced by native and non-native speakers: The dependent one-tailed *t*-test 110
5.5 Summary 113

Chapter 6. Relationships between two quantitative variables: Correlation analysis with elements of linear regression modelling 115
6.1 What is correlation? 115
6.2 Word length and word recognition: The Pearson product-moment correlation coefficient 116
6.3 Emergence of grammar from lexicon: Spearman's ρ and Kendall's τ 130
6.4 Visualization of correlations between more than two variables with the help of correlograms 134
6.5 Summary 137

Chapter 7. More on frequencies and reaction times: Linear regression 139
7.1 The basic principles of linear regression analysis 139
7.2 Putting several explanatory variables together: Predicting reaction times in a lexical decision task 142
7.3 Summary 169

Chapter 8. Finding differences between several groups: Sign language, linguistic relativity and ANOVA 171
8.1 What is ANOVA? 171
8.2 Motion events in Nicaraguan Sign Language: Independent one-way ANOVA 173
8.3 Development of spatial modulations in Nicaraguan Sign Language: Independent factorial (two-way) ANOVA 182
8.4 Do native English and native Mandarin Chinese speakers conceptualize time differently? Repeated-measures and mixed ANOVA (mixed GLM method) 189
8.5 Summary 196

Chapter 9. Measuring associations between two categorical variables: Conceptual metaphors and tests of independence 199
9.1 Testing independence 199
9.2 The story of *over* is not over: Metaphoric and non-metaphoric uses in two registers (analysis of a 2-by-2 contingency table) 200
9.3 Metaphorical and non-metaphorical uses of *see* in four registers (analysis of a 4-by-2 table) 215
9.4 Summary 221

Chapter 10. Association measures: Collocations and collostructions 223
10.1 Measures of association: A brief typology 223
10.2 Case study: The Russian ditransitive construction and its collexemes 226
10.3 Summary 238

Chapter 11. Geographic variation of *quite*: Distinctive collexeme analysis 241
11.1 Introduction to distinctive collexeme analysis 241
11.2 Distinctive collexeme analysis of *quite* + ADJ in different varieties of English: A unified approach 242
11.3 Summary 251

Chapter 12. Probabilistic multifactorial grammar and lexicology: Binomial logistic regression 253
12.1 Introduction to logistic regression 253
12.2 Logistic regression model of Dutch causative auxiliaries *doen* and *laten* 254
12.3 Summary 275

Chapter 13. Multinomial (polytomous) logistic regression models of three and more near synonyms 277
13.1 What is multinomial regression? 277
13.2 Multinomial models of English permissive constructions 277
13.3 Summary 288

Chapter 14. Conditional inference trees and random forests 291
14.1 Conditional inference trees and random forests 291
14.2 Conditional inference trees and random forests of three English causative constructions 292
14.3 Summary 299

Chapter 15. Behavioural profiles, distance metrics and cluster analysis 301
15.1 What are Behavioural Profiles? 301
15.2 Behavioural Profiles of English analytic causatives 302
15.3 Summary 320

Chapter 16. Introduction to Semantic Vector Spaces: Cosine as a measure of semantic similarity 323
16.1 Distributional models of semantics and Semantic Vector Space models 323
16.2 A Semantic Vector Space model of English verbs of cooking 324
16.3 Summary 331

Chapter 17. Language and space: Dialects, maps and Multidimensional Scaling 333
17.1 Making maps with R 333
17.2 What is Multidimensional Scaling? 336
17.3 Computation and representation of geographical distances 338
17.4 Computation and representation of linguistic distances: The Kruskal non-metric MDS 342
17.5 The Mantel test for distance matrices 348
17.6 Summary 349

Chapter 18. Multidimensional analysis of register variation: Principal Components Analysis and Factor Analysis 351
18.1 Multidimensional analysis of register variation 351
18.2 Case study: Register variation in the British National Corpus 352
18.3 Summary 365

Chapter 19. Exemplars, categories, prototypes: Simple and multiple correspondence analysis 367
19.1 Register variation of Basic Colour Terms: Simple Correspondence Analysis 367
19.2 Visualization of exemplars and prototypes of lexical categories: Multiple Correspondence Analysis of *Stuhl* and *Sessel* 373
19.3 Summary 384

Chapter 20. Constructional change and motion charts 387
20.1 The past and present of the future: Diachronic motion charts of *be going to* and *will* 387
20.2 Summary 393

Epilogue 395

Appendix 1. The most important R objects and basic operations with them 397

Appendix 2. Main plotting functions and graphical parameters in R 409

References 425

Subject Index 433

Index of R functions and packages 441

The companion website can be found at:
http://dx.doi.org/10.1075/z.195.website