

Contents

| | |
|--|-----------|
| List of Figures | ix |
| List of Tables | xix |
| Introduction | xxi |
| 1 Word Frequencies | 1 |
| 1.1 Introduction | 2 |
| 1.2 The frequency spectrum | 8 |
| 1.3 Zipf | 13 |
| 1.4 The quest for characteristic constants | 24 |
| 1.5 The lognormal distribution | 32 |
| 1.6 Discussion | 34 |
| 1.7 Bibliographical Comments | 35 |
| 1.8 Questions | 35 |
| 2 Non-parametric models | 39 |
| 2.1 Basic concepts | 39 |
| 2.2 The Urn model | 42 |
| 2.3 The Structural Type Distribution | 47 |
| 2.4 The LNRE zone | 51 |
| 2.5 Good-Turing estimates | 57 |
| 2.6 Interpolation and Extrapolation | 63 |
| 2.6.1 Interpolation | 64 |
| 2.6.2 Extrapolation | 69 |
| 2.7 Discussion | 76 |
| 2.8 Bibliographical Comments | 76 |
| 2.9 Questions | 77 |
| 3 Parametric models | 79 |
| 3.1 Introduction | 79 |
| 3.2 LNRE models | 82 |
| 3.2.1 The Lognormal Structural Type Distribution | 82 |
| 3.2.2 The Generalized Inverse Gauss-Poisson Structural Type Distribution | 89 |
| 3.2.3 The Zipfian Family of LNRE Models | 93 |
| 3.3 Evaluating Goodness of Fit | 118 |
| 3.4 Parameter estimation | 122 |
| 3.5 A comparative study | 124 |
| 3.6 Comparing Lexical Measures Across Texts | 132 |
| 3.7 Discussion | 132 |
| 3.8 Bibliographical Comments | 133 |

| | | |
|----------|---|------------|
| 3.9 | Questions | 133 |
| 4 | Mixture distributions | 135 |
| 4.1 | Introduction | 135 |
| 4.2 | Expectations, variances, and covariances | 139 |
| 4.3 | Examples of mixture distributions | 142 |
| 4.3.1 | A text-level mixture model | 142 |
| 4.3.2 | Morphological mixtures | 145 |
| 4.4 | Morphological Productivity | 154 |
| 4.5 | Discussion | 158 |
| 4.6 | Bibliographical Comments | 160 |
| 4.7 | Questions | 160 |
| 5 | The Randomness Assumption | 161 |
| 5.1 | The Randomness Assumption | 161 |
| 5.1.1 | Non-randomness and lexical specialization | 162 |
| 5.1.2 | Consequences of non-randomness | 167 |
| 5.2 | Adjusted LNRE models | 173 |
| 5.2.1 | Partition-based adjustment | 174 |
| 5.2.2 | Parameter-based adjustment | 179 |
| 5.3 | Discussion | 192 |
| 5.4 | Bibliographical Comments | 193 |
| 6 | Examples of Applications | 195 |
| 6.1 | Distributional properties of the lexicon | 195 |
| 6.1.1 | Word length and sample size | 195 |
| 6.1.2 | Matching reliability across corpora | 199 |
| 6.2 | Morphological productivity | 203 |
| 6.2.1 | Global analyses | 203 |
| 6.2.2 | Productivity and register | 208 |
| 6.3 | Authorship and Style | 211 |
| 6.4 | Beyond word frequency distributions | 214 |
| 6.4.1 | Counts of filarial worms on mites on rats | 214 |
| 6.4.2 | Year references | 215 |
| 6.4.3 | CV-structures | 218 |
| 6.4.4 | Word pairs | 221 |
| 6.4.5 | Discussion | 221 |
| 6.5 | Some practical guidelines | 223 |
| A | List of Symbols | 237 |
| B | Solutions to the exercises | 241 |
| C | Software | 251 |
| D | Data sets | 289 |
| | Bibliography | 321 |
| | Index | 329 |