Models for Discrete Epidemiological and Clinical data

A thesis presented for the degree of Doctor of Philosophy

University College London

Fiona Clare McElduff

UCL Institute of Child Health

2012

Declaration

I, Fiona Clare McElduff confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Discrete data, often known as frequency or count data, comprises of observations which can only take certain separate values, resulting in a more restricted numerical measurement than those provided by continuous data and are common in the clinical sciences and epidemiology. The Poisson distribution is the simplest and most common probability model for discrete data with observations assumed to have a constant rate of occurrence amongst individual units with the property of equal mean and variance. However, in many applications the variance is greater than the mean and overdispersion is said to be present. The application of the Poisson distribution to data exhibiting overdispersion can lead to incorrect inferences and/or inefficient analyses.

The most commonly used extension of the Poisson distribution is the negative binomial distribution which allows for unequal mean and variance, but may still be inadequate to model datasets with long tails and/or value-inflation. Further extensions such as Delaporte, Sichel, Gegenbauer and Hermite distributions, give greater flexibility than the negative binomial distribution. These models have received less interest than the Poisson and negative binomial distributions within the statistical literature and many have not been implemented in current statistical software. Also, diagnostics and goodness-of-fit statistics are seldom considered when analysing such datasets.

The aim of this thesis is to develop software for analysing discrete data which do not follow the Poisson or negative binomial distributions including component-mix and parameter-mix distributions, value-inflated models, as well as modifications for truncated distributions. The project's main goals are to create three libraries within the framework of the `R` project for statistical computing. They are:

1. `altmann`: to fit and compare a wide range of univariate discrete models

2. `discrete.diag`: to provide goodness-of-fit and outlier detection diagnostics for these models

3. `discrete.reg`: to fit regression models to discrete response variables within the `gamlss framework`

These libraries will be freely available to the clinical and scientific community to facilitate discrete data interpretation.

Acknowledgements

I would like to thank my supervisors Mario Cortina-Borja and Angie Wade for their support, guidance and invaluable advice over the term of my PhD study. I would like to thank my colleagues at the MRC Centre of Epidemiology for Child Health, UCL Institute of Child Health, in particular the past and present occupants of Room 5.09 for their support and advice.

I am grateful to the many clinicians and researchers who have provided data for this thesis: Professor Adrian Woolf, Dr Shun-Kai Chan, and Dr David long at the Centre of Nephro-eurpology, ICH; Dr Pablo Mateos and Dr James Cheshire from the UCL Department of Geography; Professor Fenella Kirkham at the Neurosciences Unit, ICH; Professor Tony Charman and Dr Greg Pasco, at the Institute of Education, Professor Pat Howlin and Dr Kate Gordon from King's College London. This project was made possible by a capacity building studentship funded by the Medical Research Council.

Finally, I would also like to thank my family for their support and understanding. Especially my parents, who are my biggest champions, Danny, Nicola and Hannah, and also to Kim, Christine and not forgetting Louie. My biggest thanks go to Michael, who has been at my side throughout this journey and whose encouragement has meant the world to me.

Contents

Abstract **2**

Acknowledgements **5**

Contents **6**

List of Figures **12**

List of Tables **18**

List of Listings **20**

Acronyms and abbreviations **21**

1 Introduction **23**

1.1 Discrete data 23

1.2 Examples 25

1.2.1 UK Surnames distributions 25

1.2.2 Cysts in steroid treated fetal mouse kidneys 29

1.2.3 Electroencephalographic seizures in paediatric coma patients 31

1.2.4 Picture Exchange Communication System (PECS) training in teachers of autistic children 34

1.3 Overview of Thesis 37

2 Discrete Probability Distributions **39**

2.1 Definitions 39

2.1.1 Overdispersion 39

2.1.2 Value-inflation 41

2.1.3 Long tails 42

2.1.4 Truncation 42

2.1.5 Notation 43

2.1.6 Special Functions 49

2.2 Basic Distributions 55

2.2.1 Bernoulli (p) 57

2.2.2 Binomial (p, n) 58

2.2.3 Geometric (p) 60

2.2.4 Hypergeometric (m, n, k) 62

2.2.5 Poisson (μ) 66

2.3 Parameter-Mix Distributions 68

2.3.1 Negative Binomial 69

2.3.2 Holla (α, θ) 74

2.3.3 Sichel (α, θ, γ) 76

2.3.4 Delaporte (α, β, γ) 80

2.3.5 Yule (λ) 83

2.3.6 Waring (b, n) 86

2.3.7 Beta-Binomial (a, b, n) 88

2.4 Component-Mix Distributions 93

2.4.1 Zero-inflated Poisson (ω, μ) 94

2.4.2 Zero-inflated Negative Binomial (ω, p, r) 96

2.4.3 Zero-inflated Sichel $(\omega, \alpha, \theta, \gamma)$ 99

2.4.4 2-component Poisson Mixture (ω, μ, λ) 101

2.4.5 2-component Poisson-Negative Binomial Mixture (ω, μ, r, p) . 105

2.5 Truncated Distributions 108

2.5.1 Positive Poisson (μ) 109

2.5.2 Positive Geometric (p) 111

2.5.3 Positive Negative Binomial (r, p) 113

2.5.4 Positive Holla (α, θ) 115

2.5.5 Positive Sichel (α, θ, γ) 117

2.5.6 Positive Yule (λ) 121

2.6 Lerch Family Distributions 123

2.6.1 Lerch (p, a, c) 124

2.6.2 Zipf (a, c) 128

2.6.3 Good (p, c) 130

2.6.4 Zeta (c) 132

2.7 Generalized Poisson Distributions 135

2.7.1 Neyman Type A (μ, ϕ) 136

2.7.2 Hermite (a, b) 138

2.7.3 Generalized Hermite (a, b, m) 141

2.7.4 Gegenbauer (a, b, k) 145

2.7.5 Generalized Gegenbauer (a, m, α, β) 148

3 Fitting the models **153**

3.1 Estimation methods 153

3.1.1 Rapid Estimation 153

3.1.2 Maximum Likelihood 158

3.1.3 Expectation-Maximization (EM) algorithm 162

3.2 Frameworks for model fitting 165

3.2.1 Generalized Linear Models (GLM) 165

3.2.2 Generalized Additive Models (GAM) 168

3.2.3 Generalized Additive Models for Location, Scale and Shape (GAMLSS) 171

3.3 Diagnostics 174

3.3.1 Goodness-of-fit 175

3.3.2 Model Comparisons 184

3.3.3 Outlier Detection 193

4 Software for fitting discrete probability models 198

4.1 Current Software 198

4.1.1 `PASW` 199

4.1.2 `Stata` 200

4.1.3 `SAS` 200

4.1.4 `R` 201

4.1.5 MATHEMATICA 208

4.1.6 Altmann Fitter 208

4.2 Gaps in methodology 209

4.3 Outline of software 212

5 Altmann Library 215

5.1 Datasets 216

5.2 Summary of discrete datasets 218

5.3 `pdqr` for distributions 221

5.3.1 Probability density function `d` 221

5.3.2 Cumulative density function `p` 223

5.3.3 Quantile function `q` 226

5.3.4 Random generating function `r` 227

5.4 Maximum likelihood estimation functions 229

5.4.1 Estimation of starting values 233

5.4.2 Maximum likelihood estimation using `mle` 236

5.4.3 Goodness-of-fit statistics and Output 237

5.5 Plotting mle objects 239

5.6 Model comparisons 244

5.7 Validation of the functions 249

5.8 Further Examples 252

5.8.1 Automobile accidents claims for drivers in Belgium, 1978 . . 252

5.8.2 Numbers of births occurring to HIV-infected women 258

5.9 Application to UK surnames distribution 261

6 discrete.diag Library 269

6.1 Goodness-of-fit Methods 270

6.1.1 Chi-squared Goodness-of-fit Test 270

6.1.2 Residuals 272

6.2 Model Comparison 275

6.2.1 AIC and BIC 275

6.2.2 EPGF plots 276

6.3 Outlier Detection 281

6.3.1 EPGF Outliers plot 281

6.3.2 Surprise Index plot 285

6.4 Validation of the functions 289

6.5 Application to counts of cysts in steroid treated foetal mouse kidneys . 290

6.5.1 Outlier Detection using the EPGF 291

6.5.2 Model fitting 292

6.5.3 Outlier detection using Surprise Index 294

7 discrete.reg library 299

7.1 Geometric Distribution 299

7.2 Yule Distribution 304

7.3 Waring Distribution 309

7.4 Validation of the functions 314

7.5 Application to Electroencephalographic Seizures in coma patients . . 315

8 Discussion 327

8.1 Contributions to software 327

8.1.1 `Altmann` library 327

8.1.2 `discrete.diag` library 328

8.1.3 `discrete.reg` library 328

8.2 Implications for data analysis 328

8.3 Limitations of libraries 330

8.4 Further Work 331
8.5 Conclusion 333

References 334

Appendices 348

A Distribution Moments 349

B Publications and posters arising from this research 374

B.1 List of publications 374
B.2 List of Posters 375

List of Figures

1.1 UK Surnames frequencies. 27

1.2 Histograms of counts of cysts in steroid treated and control foetal mouse kidneys. 30

1.3 Number of ES in paediatric coma patients. 32

1.4 Rate of ES in coma patients. 33

1.5 Outcome measure (frequency of initiations, PECS use and speech) as frequencies by treatment group by time period. 35

2.1 Binomial *pdf* 59

2.2 log(*SI*)'s for Binomial distributions 60

2.3 Geometric *pdf* 61

2.4 log(*SI*) for the Geometric distribution 62

2.5 Hypergeometric *pdf* 63

2.6 log(*SI*)'s for Hypergeometric distributions. 65

2.7 Poisson *pdf* 67

2.8 log(SI)'s for Poisson distribution 68

2.9 Negative Binomial *pdf* 70

2.10 log(*SI*)'s for negative binomial distributions 71

2.11 Holla *pdf* 74

2.12 log(*SI*)'s for Holla probability distributions 76

2.13 Sichel *pdf* 77

2.14 log(*SI*'s for Sichel distribution 79

2.15 Delaporte *pdf* 82

2.16 Yule *pdf* 84

2.17 SI's for the Yule distribution 85

2.18 Waring *pdf* 86

2.19 log(*SI*)'s for Waring distributions 88

2.20 Beta-Binomial *pdf* 90

2.21 log(*SI*)'s for the Beta-Binomial distributions 92

2.22 Zero-inflated Poisson *pdf* 95

2.23 log(*SI*)'s for Zero-inflated Poisson distributions 96

2.24 Zero-inflated Negative Binomial *pdf* 97

2.25 Zero-inflated Sichel *pdf* 100

2.26 Poisson-Poisson mix *pdf* 102

2.27 log(*SI*)'s for 2-component Poisson mixture distributions 104

2.28 Poisson-Negative Binomial mix *pdf* 105

2.29 log(*SI*)'s for 2-component Poisson-Negative Binomial distributions . 107

2.30 Positive Poisson probability *pdf* 109

2.31 log(*SI*)'s for the Positive Poisson distribution 110

2.32 Positive Geometric *pdf* 111

2.33 log(*SI*)'s for Positive Geometric distributions 112

2.34 Positive Negative Binomial *pdf* 113

2.35 log(*SI*)'s for Positive Negative Binomial distributions 114

2.36 Positive Holla *pdf* 115

2.37 log(*SI*)'s for Positive Holla distributions 116

2.38 Positive Sichel *pdf* 118

2.39 log(*SI*)'s for Positive Sichel distributions 119

2.40 Positive Yule *pdf* 121

2.41 log(*SI*)'s for Positive Yule distributions 122

2.42 Lerch *pdf* 124

2.43 log(*SI*)'s for Lerch distributions 126

2.44 Estoup and Lotka *pdf*'s 127

2.45 Zipf *pdf* 129

2.46 log(*SI*)'s for Zipf distributions 130

2.47 Good *pdf* 131

2.48 log(*SI*)'s for Good distributions 132

2.49 Zeta *pdf* 133

2.50 log(*SI*) for the Zeta distribution 134

2.51 Neyman Type A *pdf* 137

2.52 log(*SI*)'s for Neyman Type A distributions 138

2.53 Hermite *pdf* 140

2.54 log(*SI*)'s for Hermite distributions 141

2.55 Generalized Hermite *pdf* 142

2.56 log(*SI*)'s for Generalized Hermite distributions 144

2.57 Gegenbauer *pdf* 145

2.58 log(*SI*)'s for Gegenbauer distributions 147

2.59 Generalized Gegenbauer *pdf* 149

2.60 log(*SI*)'s for Generalized Gegenbauer distributions 151

3.1 Minus log-likelihood curve of Poisson model for counts of cysts in steroid treated kidneys 160

3.2 Maximum likelihood curve of negative binomial model for counts of cysts in steroid treated kidneys 161

3.3 Residual analysis using Randomized Quantile Residuals for a Poisson model for counts of cysts in steroid treated kidneys 182

3.4 Residual analysis using Randomized Quantile residuals for a negative binomial model for counts of seizures in steroid treated kidneys . . . 183

3.5 Plot of the log of *pgf*'s for a Poisson distribution with $\mu = 8$, Binomial with $n = 5$ and $p = 0.7$, negative binomial with $r = 8$ and $p = 0.3$ and a truncated Poisson distributions with $\mu = 8$. 190

3.6 Plots of the epgf for a) counts of yearly deaths by horse kicks and b) counts of earthquakes in Mexico. 191

3.7 EPGF plot of counts of cysts in embryonic mouse kidneys with fitted it pgf's for the Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial and Holla distributions. 192

3.8 EPGF analysis to detect outliers for frequencies of incidents of international terrorism. 195

3.9 Plots of SI for counts of cysts in steroid treated embryonic mouse kidneys for a) a Poisson distribution and b) a negative binomial distribution. 196

5.1 Plots of (a) the *pdf*, (b) *cdf*, (c) quantile and (d) a histogram of a random sample of the negative binomial distribution with parameters `r`$= 2$ and `p`$= 0.6$, created from the `dNB`, `pNB`, `qNB` and `rNB` functions. 224

5.2 Plots likelihood profiles of parameters `r` and `p` for the number of stillbirths in litters of New Zealand white rabbits for the negative binomial distribution using the function `mle.NB`. 238

5.3 Plots of observed and expected frequencies of stillbirths for the negative binomial distribution using the function `plot.mle`. 243

5.4 Profile likelihood plots for Holla model for number of automobile accidents claims for drivers in Belgium, 1978 255

5.5 Profile likelihood plots for Sichel model for number of automobile accidents claims for drivers in Belgium, 1978 256

5.6 Profile likelihood plots for Poisson-negative binomial mixture model for number of automobile accidents claims for drivers in Belgium, 1978 258

5.7 Profile likelihood plots for a positive negative binomial model fitted to numbers of births occurring in the UK and Ireland to HIV-infected women 260

5.8 UK Surname distribution. 262

5.9 Observed and fitted values of Zipf model for surname frequencies of Manchester. Observed values are black points and the fitted model is shown in red. 265

5.10 Observed and fitted values of Zipf model for surname frequencies of Hackney. Observed values are black points and the fitted model is shown in red. 266

5.11 Observed and fitted values of Zipf model for surname frequencies of Carlisle. Observed values are black points and the fitted model is shown in red. 267

6.1 Residual plots for the number of stillbirths in litters of New Zealand White Rabbits under a negative binomial model. 274

6.2 EPGF plots for the number of stillbirths in New Zealand white rabbits with a) Poisson, Geometric and Yule distributions, b) negative binomial, zero-inflated Poisson, Neyman type A and Waring distributions and c) hypergeometric, zero-inflated negative binomial and Poisson-Poisson mixture distributions. 279

6.3 EPGF outliers plot of the rabbits data, ... 284

6.4 EPGF outliers plot of the rabbits data, ... 285

6.5 *SI*'s for the number of stillbirths in New Zealand White rabbits under Poisson and Negative Binomial distributions. 288

6.6 EPGF outlier plots of counts of cysts in steroid treated foetal mouse kidneys 291

6.7 EPGF outlier plots of counts of cysts in steroid treated foetal mouse kidneys without observation of 19 cysts. 292

6.8 *SI* plots of counts of cysts in steroid treated foetal mouse kidneys for models 1) a Poisson distribution and 2) a Poisson distribution excluding the kidney with 19 cysts. 295

6.9 *SI* plots of counts of cysts in steroid treated foetal mouse kidneys for models 1) a negative binomial distribution and 2) a negative binomial distribution excluding the kidney with 19 cysts. 296

7.1 Numbers of stillbirths in New Zealand White rabbits with fitted a) Geometric b) Yule and c) Waring distributions respectively 305

7.2 Residual plot from the fitted negative binomial model 321

7.3 Worm plot from the fitted negative binomial model 322

7.4 Predictions from the fitted negative binomial model across three centres by the presence of seizures 324

List of Tables

1.1 Top surnames by country. Figures in parentheses are percentages. . . 28

1.2 Summary statistics for counts of cysts in kidneys for steroid treated and control groups. 30

1.3 The effectiveness of PECS training: Study Design 34

1.4 Medians of frequencies of initiations, PECS use and speech by treatment group by time period. IQR is given in bold and 25% and 75% quantiles are given in parenthesis. 36

2.1 Weaver (1948)'s interpretation of *SI* values. 48

3.1 Observed (o) and Expected (e) frequencies of cysts in steroid treated mouse kidneys for a Poisson model. 176

3.2 Observed (o) and Expected (e) frequencies of cysts in steroid treated mouse kidneys for a negative binomial model. 176

3.3 Frequencies of incidents of international terrorism in the United States between 1968-1974 194

4.1 Discrete distributions available as *pdf*s, *cdf*s and random generations using PASW. Function names are in parenthesis. 199

4.2 Probability distributions available in the base library of the R language 203

4.3 Discrete distributions implemented within the gamlss.dist library (Stasinopoulos and Rigby, 2007). 205

4.4 Discrete probability distributions available in VGAM library of R . . . 207

5.1 Frequency of stillbirths in litters of New Zealand white rabbits 216

5.2 Counts of morpheme length in lakota language 216

5.3 Counts of yeast cells 217

5.4 household size from Housing Allowance Demand Experiment 217

5.5 Frquency of surnames across eight non-overlappping districts 218

5.6 Probability distributions available in the Altmann library 222

5.7 Fitted values for a Gegenbauer distribution fitted to 400 haemocytometer counts of yeast cells. 251

5.8 Number of automobile accidents claims for drivers in Belgium, 1978 . 253

6.1 Counts of Cysts in steroid treated kidneys 290

6.2 Counts of Cysts in control kidneys 290

6.3 Table of *SI*'s for the four models. 294

7.1 Summary of discrete regression models resulting from stepwise model selection fitted to incidence of ES dataset. 318

Listings

5.1 Summary function for discrete datasets 218

5.2 Probability density function `d` 223

5.3 Cumulative density function `p` 225

5.4 Quantile function `q` 226

5.5 Random generating function `r` 227

5.6 Maximum likelihood estimation function for the negative binomial distribution. 230

5.7 Plot function for class ‘mle’ 240

5.8 Altmann Fitter Model Comparison Function 245

6.1 Chi-squared Goodness-of-fit Test 270

6.2 Residual Analysis 272

6.3 `AIC` function 275

6.4 EPGF Plot function 276

6.5 EPGF outliers plot 281

6.6 Surprise Index function 286

7.1 Geometric GAMLSS family distribution function. 300

7.2 Yule Family distribution function. 306

7.3 Waring family distribution function. 310

Acronyms and abbreviations

ACS – Adelaide Coma Scale

ADOS-g - Autism Diagnosis Observation Schedule- Generic Module One

AIC – Akaikes information criterion

ASD - Autism Spectrum Disorder

BIC - Bayesian Information Criterion

cdf - cumulative density function

CRAN – Comprehensive R Archive Network (CRAN)

ECM – Expectation/Conditional Maximisation

EEG – Electroencephalographic monitoring

EM - Expectation Maximisation

EPGF – Empirical Probability Generating Function

ES – Electroencephalographic Seizures

GAM - Generalized Additive Model

GAMLSS - Generalized Additive Model for Location, Scale and Shape

GLM - Generalized Linear Model

IQR – interquartile range

mgf moment generating function

MLE - Maximum Likelihood Estimate

NB – negative binomial

NBI – negative binomial type I

NB II – negative binomial type II

NVDQ - Non-verbal Developmental Quotient

OD - Overdispersion Index

PASW – Predictive Analytics SoftWare

pdf - probability density function

PECS - Picture Exchange Communication System

pgf - probability generating function

PIM - Paediatric Index of Mortality

SAS – Statistical Analysis Systems

SCQ - Social Communication Questionnaire

SD – standard deviation

SI - Surprise Index

SPSS – Statistical Package for the Social Sciences

ZI - Zero-inflation Index

ZINB - Zero-inflated negative binomial

ZIP - Zero-inflated Poisson

ZISI - Zero-inflated Sichel

Chapter 1

Introduction

1.1 Discrete data

Data is either categorical or numeric, with numeric variables further classified as continuous or discrete. Continuous variables are measured on a scale such that between any two values it is always possible to find another. Discrete variables can only take a (usually) limited number of separate values such that there are no possible realizations of the variable between any two of its consecutive values. Discrete variables are often of interest in clinical and epidemiological studies.

A discrete random variable, Y, is a function from a sample space Ω (the set of all possible outcomes of a random experiment) to a (finitely or infinitely) countable set, R_Y, known as the range of Y. Discrete data, i.e. observed values of Y, are also known as frequency or count data. Count variables are defined by Dobson (2002, pg. 151) as 'the number of times an event occurs' and the number of occurrences can originate from a finite or infinite range. An example of an infinite range is the number of complete days a patient stays in a paediatric intensive care unit, which may take integer values {0, 1, 2, ... } (Brown et al., 2003) and has no higher bound. In a finite range there is an upper limit to the number of times an event can occur, for example the number of correct responses in a test consisting of 10 questions may take values in {0, 1, 2, ..., 9, 10}. Another example are quality of life measures, which assume discrete values from a finite, ordered numeric scale and are often found in health research. For example, quality of life or overall health can be rated as an integer in the range 1 to 10, where 1 indicates a low and 10 a high quality of life or overall health (Testa and Simonson, 1996). Similarly, the Social Communication Questionnaire (SCQ) (Rutter et al., 2003) a screening tool for Autism Spectrum Disorders (ASD) takes one of the 40 discrete integer values in the range 0 to 39, where a score of less than 8 is considered a low score, 8-14 moderately low, 15-21 moderately high and greater than 22 represents a high score (Baird et al., 2006).

Note that R_Y may not contain 0. For example, consider the number of times a surname appears in a population, which can be used to study its genetic structure (Voracek and Sonneck, 2007), and the frequency of words in a text or in discourse (Monaco et al., 2007). In both cases the minimum value of Y is necessarily 1.

Rates are an instance of discrete observations which are expressed per measure of time (e.g. hours, minutes or seconds) in which the events occur. For example, in epidemiology annual incidence of a condition, and mortality rates per year or per 100 person years of follow up are often used (Kirkwood and Sterne, 2003, pg. 229). Where events are rare, rates may be multiplied by 1,000 (or even 10,000 or 100,000) and expressed per 1,000 (or 10,000 or 100,000) subjects per unit of time. Rates allow counts to be adjusted for variations in time periods where necessary. For example, the number of epileptic seizures observed in children during a specific hospital episode can be considered as a rate where the length of hospital stay will differ between patients.

In clinical and epidemiological studies data collection is crucially constrained by both ethical and financial considerations attached to the recruitment of each additional respondent. Hence it is very important that any data collected is analysed using the most appropriate methods and processed in a way that will extract the maximum information. This is a key issue for discrete variables which are often skewed and may have irregular features in their distribution (McElduff et al., 2010). Models for continuous data such as linear regression and Analysis of Variance (ANOVA) should not be directly applied to discrete response variables due to the underlying distributional assumptions required by these models for their correct application (Afifi et al., 2007).

Another approach is to separate the rates or frequencies into ordered categories and use ordinal logistic regression. For example, the categorisation of the SCQ described above into low, moderately low, moderately high and high score groups. However, information is lost and hence this approach is an inefficient use of the available data.

1.2 Examples

In this section four discrete datasets from the fields of epidemiology and child health are presented; these will be used to illustrate the statistical methods shown in this thesis.

1.2.1 UK Surnames distributions

Surnames have been used since the 19th century to understand the relationships between population subgroups (Darwin, 1875) at regional or national levels (Colantonio et al., 2003; Lasker, 1985). An established relationship exists between surname frequencies, geographic distributions and the ethnic and genetic structures in a population (Piazza et al., 1987). Surnames are used in the field of child health as indicators of ethnicity in probabilistic record linkage (Cook et al., 1972), for example in studies of childhood cancer (Rankin et al., 2008; Duncore et al., 2008). Surnames are often patrilinearly inherited so they correlate well with Y-chromosomes (Jobling, 2001) and can be used to identify genetic factors in certain diseases/conditions. For example, a study of incidence of suicide in Austria used surname frequencies to represent the genetic structure of the general population and found that differences in regional suicide rates correspond to patterns of surname distributions (Voracek and Sonneck, 2007).

The data on surnames used in this thesis is from a study on the quantitative properties of the geographic and statistical distributions of surnames in the UK (McElduff et al., 2008). The data is taken from the 2001 UK electoral register, which is a public register containing the names and addresses of all adults (over the age of 16) that are registered to vote in any type of UK elections; this includes nationals of the UK, Commonwealth countries and the European Union. In addition to registered voters, the companies which distribute the electoral register supplement it with additional residents not registered to vote which they source from commercial surveys and credit scoring databases. The resulting database is known as the 'enhanced electoral register', and the version used here was purchased by University College London (UCL) Department of Geography for research purposes. The 2001 UK surnames distribution is the last version of this dataset before opting out of the electoral register was made possible by the data protection act and is therefore the most complete data source of names and locations publicly available in recent years.

Within the UK enhanced electoral register there are 434 districts, each of which is an administrative subdivision corresponding to a Local Authority or their equivalent. These districts can be grouped into 13 regions according to the official Government Office Regions which are used by the Office for National Statistics (ONS) (http://www.statistics.gov.uk/geography/gor.asp): nine English: North East, North West, Yorkshire and Humberside, East Midlands, West Midlands, the East of England, London, South East and South West; Wales; Scotland; Northern Ireland and the Channel Islands. The dataset contains one record per person detailing their surname and location, both as a district and a region of the UK. Hence this dataset can be used to view the national distribution of surname frequencies across the UK.

There were a total of 45,690,258 people comprising the enhanced electoral register of residents in the UK in 2001 with a total of 828,130 different surnames. Figure 1.1 shows the distribution of the UK surnames frequencies on a log-log scale. The y-axis shows the number of different surnames in the UK and the x-axis their frequency in the population. For instance, of the total number of different surnames, 431,554 were unique (i.e. total frequency of one), representing 52.11% of the total surnames but only 0.94% of the population. The percentage of the population with surnames that occur only twice is 0.26%. The very long right-hand tail corresponds to surnames shared by a large number of people, for instance, the most frequent surname in the UK, Smith, is shared by 555,982 people. On average, a surname is bourne by 183.95

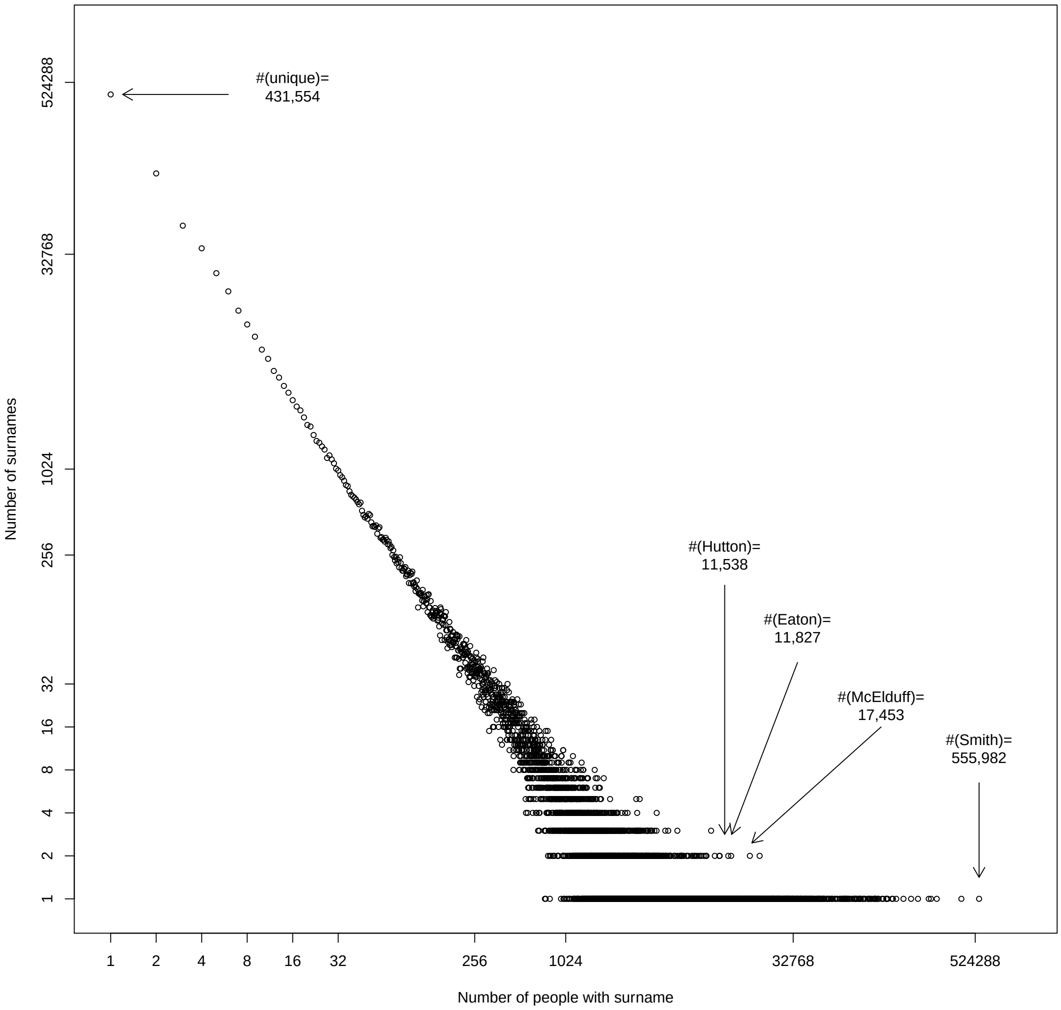


Figure 1.1: UK Surnames frequencies.

persons (median =1), though the extreme skewness of this distribution (SD=6767.775, $(Q_{0.25},Q_{0.75})$=(1,4), IQR=3) makes this figure rather meaningless. The skewness coefficient[1] value of 58.94 highlights the very large positive skew in the distribution and the kurtosis coefficient[2] value of 3691.12 reflects the peakedness due to the high frequency of unique surnames in the distribution.

	England	Northern Ireland	Scotland	Wales
1	Smith (1.26)	Wilson (0.75)	Smith (1.28)	Jones (5.75)
2	Jones (0.75)	Campbell (0.75)	Brown (0.94)	Williams (3.72)
3	Taylor (0.59)	Kelly (0.74)	Wilson (0.89)	Davies (3.72)
4	Brown (0.56)	Johnston (0.69)	Robertson (0.78)	Evans (2.47)
5	Williams (0.39)	Moore (0.62)	Thompson (0.78)	Thomas (2.43)
6	Wilson (0.39)	Thompson (0.61)	Campbell (0.77)	Roberts (1.53)
7	Johnson (0.37)	Smyth (0.60)	Stewart (0.73)	Lewis (1.53)
8	Davies (0.34)	Brown (0.59)	Anderson (0.70)	Hughes (1.23)
9	Robinson (0.32)	O'Neill (0.57)	Scott (0.55)	Morgan (1.16)
10	Wright (0.32)	Doherty (0.54)	Murray (0.53)	Griffiths (0.96)
11	Thompson (0.31)	Stewart (0.54)	MacDonald (0.52)	Edwards (0.93)
12	Evans (0.30)	Quinn (0.51)	Reid (0.52)	Smith (0.85)
13	Walker (0.30)	Robinson (0.50)	Taylor (0.49)	James (0.82)
14	White (0.30)	Murphy (0.49)	Clark (0.47)	Rees (0.81)
15	Roberts (0.28)	Graham (0.48)	Ross (0.43)	Jenkins (0.69)
16	Green (0.28)	Martin (0.45)	Young (0.42)	Owen (0.67)
17	Hall (0.28)	McLaughlin (0.45)	Mitchell (0.41)	Price (0.67)
18	Wood (0.27)	Hamilton (0.44)	Watson (0.41)	Phillips (0.65)
19	Jackson (0.27)	Murray (0.43)	Paterson (0.40)	Morris (0.63)
20	Clarke (0.26)	Hughes (0.41)	Morrison (0.40)	Richards (0.55)

Table 1.1: Top surnames by country. Figures in parentheses are percentages.

Table 1.1 gives the percentages of the population with the top 20 ranked surnames for each country in the UK. Wales has the highest cumulative percentage for these 20 surnames (31.6%) followed by Scotland, Northern Ireland and England (12.40%, 11.51% and 8.30%, respectively.) These figures highlight the much lower diversity

[1] The skewness coefficient is given by the standardized third central moment of a distribution and is a measure of symmetry (Groeneveld and Meeden, 1984). A positive skewness coefficient indicates a distribution with a long right tail, whilst a negative skewness indicates a distribution with a long left tail; zero corresponds to symmetric distributions.

[2] Similarly, the kurtosis coefficient is the standardized fourth central moment of a distribution (Groeneveld and Meeden, 1984). The kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution, i.e. datasets with high kurtosis (leptokurtic) tend to have a distinct peak near the mean and have heavy tails. Datasets with low kurtosis (platykurtic) tend to have a flat top near the mean rather than a sharp peak. This coefficient is often expressed with respect to 3, which is its value for the (mesokurtic) Normal distribution.

of surnames in Wales. Examination of the 20 most common surnames reveals that all of them originate from the British Isles. For example Jones, Williams and Evans are considered to be of Welsh origin, whereas the names Robertson, Thomson and Campbell are considered to be of Scottish derivation. Similarly, Irish surnames (e.g. Kelly, ONeill and Doherty) and Scottish surnames (e.g. Campbell, Johnston and Thompson) rank highly in the top surnames in Northern Ireland. English origin surnames, however, occur in all four UK countries: Smith (or its variant Smyth) and Brown arise in the top 20 surnames for all countries. Some surnames of Scottish or Welsh origin also appear in the top 20 surnames of England and, more markedly, of Northern Ireland.

1.2.2 Cysts in steroid treated fetal mouse kidneys

A pregnant woman's diet may affect kidney development of her unborn child and may lead to the infant developing kidney problems. It has been shown that when a pregnant mother eats a low protein diet, cell survival and gene expression are altered during kidney development (Welham et al., 2002, 2005). Other studies indicate that a low protein diet causes a higher proportion of the mother's corticosteroids to be exposed to the foetus (Langley-Evans et al., 1996) and it is thought that this increase of corticosteroid exposure may directly influence fetal kidney development.

In a study by Chan et al. (2010), developing embryonic mouse kidneys were cultured in different steroids to help understand how a mother's diet could lead to the offspring developing kidney problems in later life. Cultured embryonic mouse kidneys were subjected to steroids and the number of cysts counted after six days, a high number of cysts indicating abnormal kidney growth. The analysis compared counts of cysts from $n = 111$ steroid treated kidneys and $n = 103$ untreated (control) kidneys. Figure 1.2 gives the distribution of the number of cysts in kidneys in the steroid-treated and control groups.

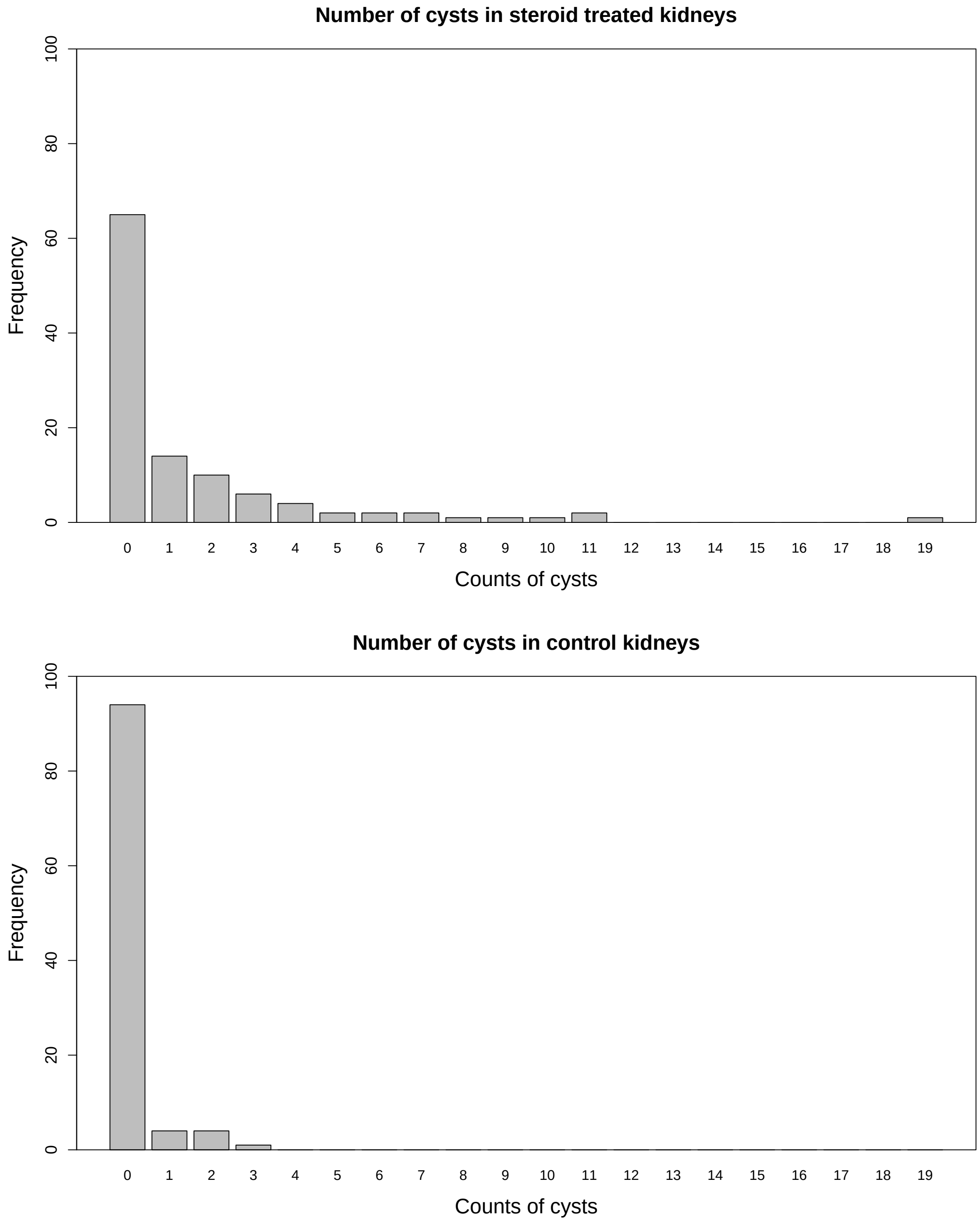


Figure 1.2: Histograms of counts of cysts in steroid treated and control foetal mouse kidneys.

	Steroid	**Control**
Mean	1.55	0.15
(SD)	(2.98)	(0.51)
Median	0	0
Interquartile range	2	0
(lower quartile, upper quartile)	(0,2)	(0,0)
Minimum, Maximum	0, 19	0, 3
Percentage of zeroes	58.56	91.26

Table 1.2: Summary statistics for counts of cysts in kidneys for steroid treated and control groups.

Summary statistics for the steroid treated and control kidney groups are shown in Table 1.2. The mean number of cysts was 0.15 for controls and 1.55 for treated mice kidneys, although the medians are both equal with values of 0. Summary statistics for the dispersion in the steroid treated group (SD=2.98, ($Q_{0.25}$,$Q_{0.75}$)=(0,2), IQR=2) are much higher than that of the control group (SD=0.51, ($Q_{0.25}$,$Q_{0.75}$)=(0,0), IQR=0). In each group, the majority of kidneys had no cysts (58.56% in the steroid groups and 91.26% in the control), although there were a few kidneys in the steroid-treated group with large cysts counts. In the steroid-treated group, one kidney had a value of 19 cysts which was much higher than the maximum number of cysts found in the control group kidneys (maximum=3).

1.2.3 Electroencephalographic seizures in paediatric coma patients

Around 0.5% of patients admitted to paediatric intensive care units are recognised to have clinical seizures (Valencia et al., 2006). There are few studies on the incidence of clinical or electroencephalographic seizures (ES) or status epilepticus (a condition in which the brain is in a state of persistent seizure) in acute paediatric encephalopathies (a dysfunction in the central nervous system). Mortality and morbidity for status epilepticus in children is known to be related to aetiology (Raspall-Chaure et al., 2006). The availability of electroencephalographic (EEG) monitoring has enabled the detection of ES in comatose patients. The dataset presented in this section forms part of a study using continuous EEG monitoring to document the incidence of ES in unconscious children.

Data from 184 patients was collected from three centres. There were 141 children who were treated in two UK paediatric intensive care units between 1982 and 1990 (comprising 15 neonates with cardiac disease plus 126 children recruited for a study on hypoxic-ischemic encephalopathy), together with 43 patients entering a high dependency hospital unit in Kenya in 1990. Children were monitored continuously using EEG machines and the number of ES was recorded until movement was detected or their condition became fatal. The duration of monitoring was recorded for each patient.

There are nine explanatory variables associated with the number of ES included in the dataset: The Paediatric Index of Mortality (PIM) (mean=45.47, SD=26.95); Adelaide Coma Scale scores (mean=5.43, SD=2.54); temperature on admission (mean=37.31, SD=1.92); centre (UK=126, UK neonates=15, Kenya=43); aetiology (Encephalitis=9, Head Injury=11, Hypoxic-ischaemic=108, Malaria=42, Meningitis=4, Reyes=3, Other=7); EEG classification (Burst suppression=15, Diffuse slowing=29, diffuse slowing with some fast activity=68, Isoelectric=7, Low amplitude=30, Normal=35); presence of clinical seizures (yes=86, no=98); the use of the drugs benzodiazepine (yes=27, no=157) or phenytoin/phenobarbitone (yes=19, no=165) to terminate seizures before EEG monitoring occurred.

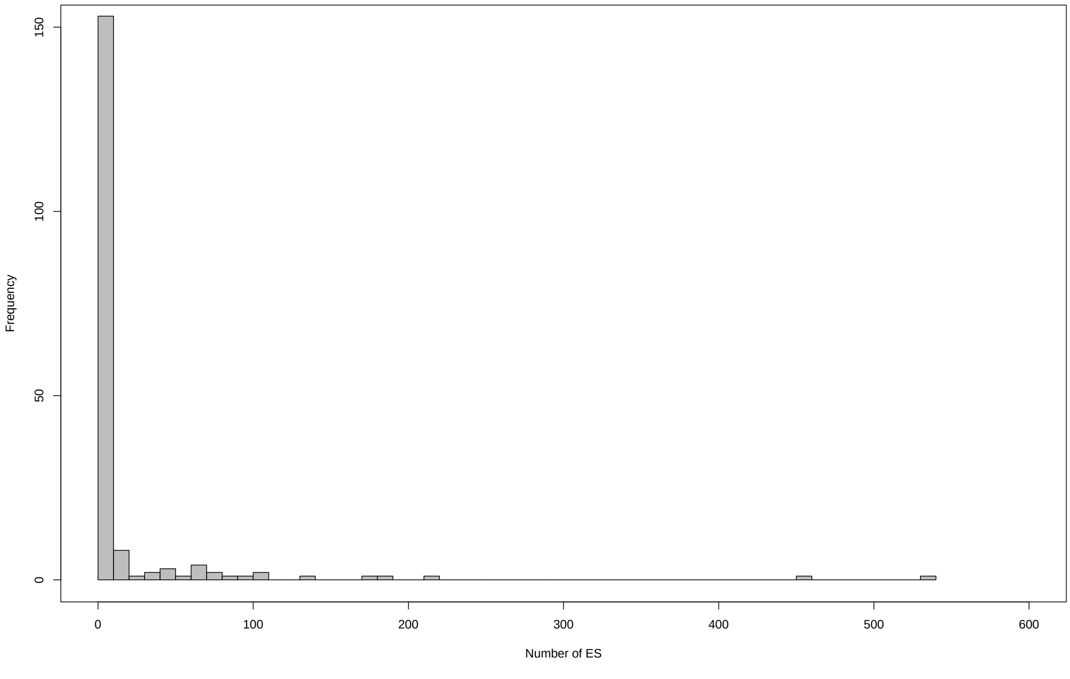


Figure 1.3: Number of ES in paediatric coma patients.

Figure 1.3 shows a histogram of the distribution of the number of ES in the complete dataset. The mean average number of ES is 16.58 (SD=59.87) with median 0 (($Q_{0.25}$,$Q_{0.75}$)=(0,3), IQR=3) and a large right skew in the distribution. The skew is due to a small number of severely ill patients who have a very high number of ES. Most of the counts of ES range between 0 and 218, with two extreme observations at 458 and 531. There is a high proportion of zeros present in the data (63.2%) and 87.5% of patients have numbers of ES between 0 and 20. Both the skewness and kurtosis coefficients for the number of ES are very high at 6.22 and 47.71. The duration of monitoring (hours) (mean=61.65, SD=76.23) ranges between 2 and 630. The number of ES provides an example of a discrete variable that can be modelled as rates by dividing the number of ES by the duration of monitoring so that analyses can be adjusted for the variable durations of monitoring.

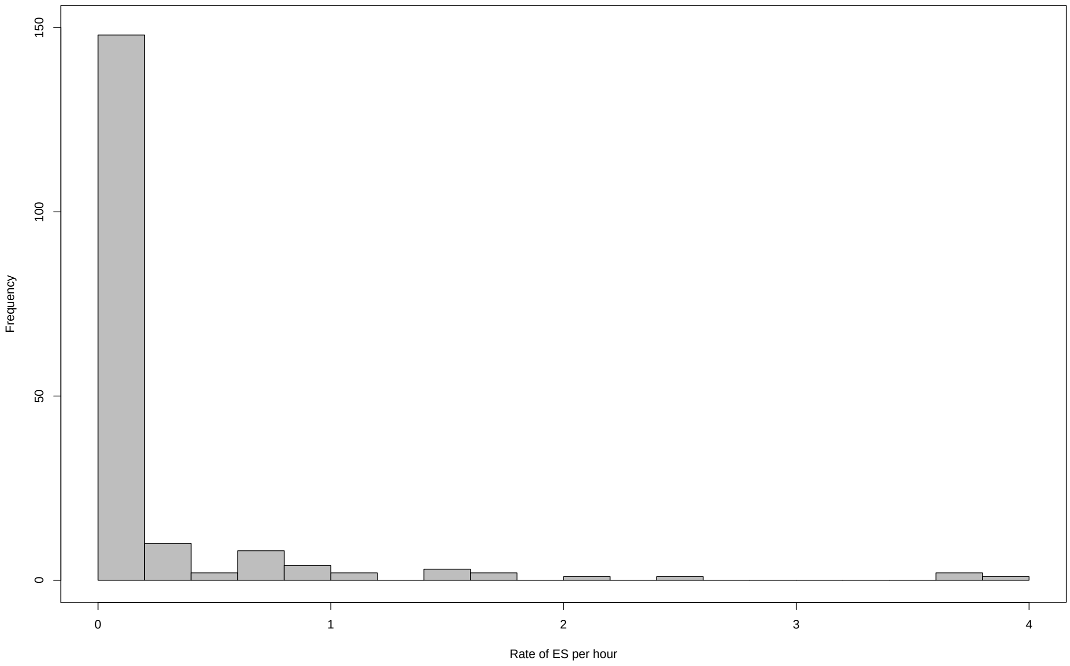


Figure 1.4: Rate of ES in coma patients.

Figure 1.4 gives a histogram of the rates of ES in coma patients. Rates of ES vary between 0 and 3.82 ES per hour with the median rate of ES is 0 ES per hour (($Q_{0.25}$,$Q_{0.75}$)=(0,0.08), IQR=0.08). The rates of ES are again highly skewed, with a high number of patients having low rates of ES per hour. A high proportion of the rates are zero (63.2%) due to the highly skew distribution of the counts of ES. Note the loss of information when analysing this dataset as rates, since a zero count translates to a zero rate irrespective of length of monitoring, yet zero events in 1 hour, for example, is clearly different to zero events in 100 hours. Hence the preferable way of adjusting for length of observation is to model the counts with a fixed-coefficient variable to adjust for the rate of duration (known as an offset) (Hilbe, 2007, P.45).

1.2.4 Picture Exchange Communication System (PECS) training in teachers of autistic children

The prevalence of autism spectrum disorder (ASD) is around 1% of the childhood population aged 9-10 in the South Thames area of the UK which comprises of inner and outer South London, Kent, East and West Sussex and Surrey (Baird et al., 2006). Around 25% of individuals with ASD are without functional speech (Volkmar et al., 2004). The Picture Exchange Communication System (PECS) aims to teach spontaneous social communication skills by means of symbols or pictures. Teaching relies on behavioural principles, particularly reinforcement techniques. A study on the effectiveness of PECS training for teachers of children with autism (Howlin et al., 2007) measured the frequency of initiations, speech, and pictures/symbols use in a longitudinal study of 84 children (average age 6.8 years) identified from specialist education schools located in Greater London and South East England. Teachers and parents of the children from the 18 participating classes received formal training in the use of PECS. The study consisted of three groups of children, assessed over 3 time periods (Table 1.3) with each group having a different treatment schedule. Observations are clustered within individuals (i.e. measurements were taken for each child at time periods 1, 2 and 3) which are clustered within class groups within three treatment arms.

	Time 1 (Baseline)	**Time 2** (Treatment Period One)	**Time 3** (Treatment Period Two)
Immediate Treatment Group	No treatment	Treatment	No treatment
Delayed Treatment Group	No treatment	No treatment	Treatment
No Treatment Group	No treatmen	No treatment	No treatment

Table 1.3: The effectiveness of PECS training: Study Design

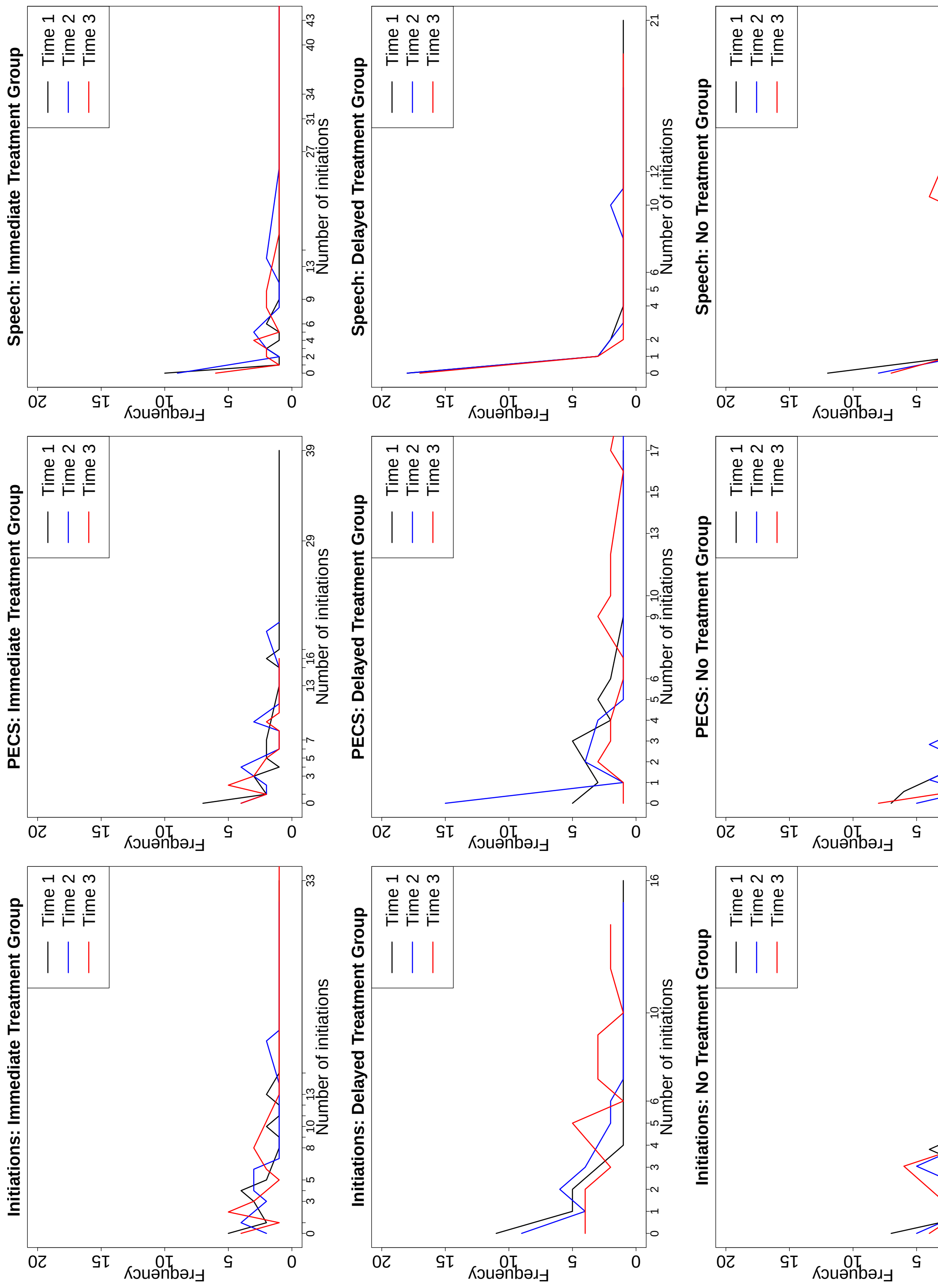


Figure 1.5: Outcome measure (frequency of initiations, PECS use and speech) as frequencies by treatment group by time period.

Time Period 1	Time Period 2	Time Period 3
Frequency of initiations		
Immediate Treatment Group		
4	5	3
8.25 (1.75,10)	**10.5** (3, 13.5)	**6** (2,8)
Delayed Treatment Group		
1	2	5.5
3 (0,3)	**3** (0, 3)	**6.5** (2.25, 8.75)
No Treatment Group		
3.5	4	4
3 (0,3)	**4.5** (1.75, 6.25)	**5.5** (2, 7.5)
Frequency of PECS use		
Immediate Treatment Group		
4.5	6.5	3
11.25 (0.25,11.5)	**8** (2, 10)	**6** (2, 8)
Delayed Treatment Group		
3	0.5	9
4 (1,5)	**4** (0, 4)	**11** (4, 15)
No Treatment Group		
2	4.5	3.5
3.5 (0.75,4.25)	**5.25** (2, 7.25)	**7.25** (0, 7.25)
Frequency of speech		
Immediate Treatment Group		
3	5	4
12 (0, 12)	**13.25** (0, 13.25)	**9** (1, 10)
Delayed Treatment Group		
0	0	0
2 (0, 2)	**2** (0, 2)	**3.75** (0, 3.75)
No Treatment Group		
3	5	4.5
8.25 (0, 8.25)	**8.25** (0, 8.25)	**9.25** (0.75, 10)

Table 1.4: Medians of frequencies of initiations, PECS use and speech by treatment group by time period. IQR is given in bold and 25% and 75% quantiles are given in parenthesis.

The outcome measures recorded included the frequency of initiations, pictures/symbols use and speech during snack time (mean length=11.1 mins, sd=3.4 mins). Baseline measures were also recorded for each child: the ADOS-g language rating on the Autism Diagnosis Observation Schedule-Generic Module One (Lord et al., 1999) was used as an index of expressive ability, the Visual Reception and Fine Motor sub-scales of the Mullen scales of Early Learning (Mullen, 1999) provides a measure of non-verbal developmental quotient (NVDQ) and also age at baseline (time period one).

Figure 1.5 plots the frequencies of initiations, pictures/symbols use and speech (columns) for the three treatment schedules: immediate, delayed and no treatment groups (rows). For each outcome measure and treatment schedule, the frequencies are plotted for the three time periods in black, blue and red, respectively. The distributions of the frequencies of initiations, use of PECS and speech across all treatment groups and time periods are skewed, i.e. there are high proportions of students achieving low counts and low probabilities of those with high counts. The medians, 25% and 75% quantiles and the IQR of the frequencies (Table 1.4) show variations in the distributions across treatment schedules across time periods.

Frequencies were transformed into rates by dividing the frequencies by length of snack break time. For the initial published analyses, the rates were divided into four ordered categories (zero, 0.01 to 0.5 per minute, 0.5 to 1 per minute and >1 per minute.) and analysed using multilevel ordinal logistic regression (Howlin et al., 2007) to allow for within-child and within-class correlations to be accounted for in the model. Such categorisation obviously reduces the information and the three discrete variables could be directly modelled as rates to provide a better description of the data.

1.3 Overview of Thesis

The examples given in Section 1.2 illustrate the type of discrete datasets that occur in epidemiological and child health research. A wide range of probability models for discrete data exist; however many are not readily available in statistical software packages. The aim of this thesis is to develop software to analyse count data; this will provide a tool kit of methods for clinicians and statisticians in order to facilitate data interpretation.

Chapter 2 explores discrete probability models, including commonly found features of discrete data, modifications to discrete probability distributions and families of distributions. A selection of discrete probability distributions is then presented. Estimation methods and frameworks for fitting these models are described in the first two sections of Chapter 3, followed by a discussion on diagnostic methods for goodness-of-fit, model comparisons and outlier detection. A review of methods currently available for discrete models in statistical software packages are detailed in the first part of chapter 4. The aim of the second part of that chapter is to identify gaps in the software currently provided for discrete models.

The project's main goals are to create three libraries within the framework of the `R` project for statistical computing. Chapter 5 presents the first of these libraries, called the `Altmann` library (named after Gabriel Altmann one of the authors of the Thesaurus of Discrete Probability Distributions (Wimmer and Altmann, 1999)), which fits and compares a wide range of univariate discrete models. The second library developed in this thesis is the `discrete.diag` library which provides goodness-of-fit and outlier detection diagnostics for these models and is described in chapter 6. The final library `discrete.reg` given in chapter 7, fits regression models to discrete response variables following the Generalized Additive Models for Location, Scale and Shape framework (Stasinopoulos and Rigby, 2007) and is available as the `gamlss` add-on package in `R` (Stasinopoulos and Rigby, 2008). Applications of the tools provided by the `R` libraries are presented in the final sections of chapters 5-7 for the discrete datasets given as examples in Section 1.2. Chapter 8 has a discussion of the libraries and statistical methods presented, and concludes by outlining the scope for further work in this area.

Chapter 2

Discrete Probability Distributions

This chapter provides an overview of discrete probability models. In the first section, common features of discrete data are described, including overdispersion, value-inflation, long tails and truncation. Notation and special mathematical functions used throughout this thesis are then presented. In the following sections, a selection of discrete probability distributions is given together with descriptions of modifications to discrete probability distributions and details of families. For each model, the distribution is defined and properties are specified. Raw and central moments for the distributions are additionally presented in Appendix A. The reader may omit Sections 2.1.6-2.7 without any loss of continuity and use these sections as a reference for the following chapters, if desired.

2.1 Definitions

In this section, several common features of skew discrete random variables are defined including overdispersion, value-inflation, long tails and truncation. The notation used throughout this thesis is detailed and the basis of statistical concepts of probability distributions and measures used to characterise discrete models are explained.

2.1.1 Overdispersion

Overdispersion occurs where there is greater variability in a dataset than expected under a simple statistical model (normally Poisson), i.e. the variance in a dataset is greater than the mean (Cox, 1986; Dobson, 2002). The presence of overdispersion (also known as extra-variation) in discrete data causes summary statistics resulting from a simple statistical model to be larger than anticipated and can lead to incorrect inferences under such a simple hypothesis. For example, a covariate may seem to be a significant predictor in an analysis when it is not (Hilbe, 2007). There are many causes of overdispersion in data; Hilbe (2007) identifies two approaches to dealing with overdispersion, where causes may be categorized as either apparent or real.

Apparent overdispersion occurs where the source of extra-variation results from the data's structure, sampling or methods of analysis used (Hilbe, 2007). Such cases of overdispersion can be removed by adjusting the model's structure to account for the extra-variation in the dataset. Multilevel experiments often yield repeated measurements which are highly correlated and may lead to overdispersion (Hilbe, 2007). For example, the study of Picture Exchange Communication System (PECS) training in Autistic children in Section 1.2.4 in Chapter 1 presents a multilevel experiment which yields repeated outcome measures across three treatments schedules, over three time periods. The frequency of initiations, PECS use and speech each have overall mean frequencies of 5.39, 6.27 and 6.26, with variances 39.73, 69.62 and 108.88, indicating a large amount of overdispersion. In this example, children are clustered within classes and within treatment schedules. Overdispersion present in the outcome variables of multilevel datasets may be accounted for by incorporating random effects terms into the model. Other cases which may cause overdispersion to be apparent in the data, include outlying observations, incorrectly specified models such as incorrect parametrizations in analyses (omitting important explanatory variables or interaction terms), or erroneously specifying the relationship between the observed counts and explanatory variables (Hilbe, 2007).

Real causes of overdispersion occur where extra-variation in an explanatory variable exists but cannot be accounted for in the structure of the model. Such causes of overdispersion are due to the underlying data-generating mechanism and therefore cannot be accounted for solely by adjusting the model structure but through the use of models specifically designed for overdispersed count data. For instance, the mean number of counts of cysts in foetal mouse kidneys (Section 1.2) in the group of kidneys subjected to steroids is 1.55, with variance 8.88. Since no covariates have been recorded or multilevel structure observed for this dataset, this example illustrates a discrete variable where the overdispsersion present cannot be accounted for in the experiment's design.

2.1.2 Value-inflation

Many epidemiological or clinical datasets exhibit value-inflation i.e. an excess number of observations of a particular value. Value-inflation occurs when a population actually consists of two latent sub populations, with observations from one population only taking a certain value whereas observations from the other population can take any value on a discrete scale. This leads to a distribution with an excess of observations at one value, which is not easily analysed using standard models. An example of extreme value-inflation can be seen in the UK surnames frequencies (Section 1.1) (McElduff et al., 2008). In this distribution, most surnames occur relatively few times with the majority of surnames occurring only once (52.11%) resulting in a distribution that is value-inflated at one.

The most common type of value inflation is zero-inflation in which there is a sub population in the dataset that always take the value zero, whilst the remainder of the dataset can take any integer value from zero upwards. Zero-inflated datasets are often heavily weighted to zero and lower values with an upper tail. The distribution of the number of cysts in steroid treated and control embryonic mouse kidneys is shown in Section 1.2 of Chapter 1 (Chan et al., 2010; McElduff et al., 2010). In this dataset, an excess number of kidneys were recorded with zero cysts (74.3%) , suggesting two sub populations of mice: those with kidneys which can/do produce cysts and those which cannot and therefore always have zero cysts. Zero-inflated datasets are commonly found in epidemiology and child health.

The inherent data-generating mechanism underlying the populations behind the distribution of a dataset may not always be obvious. In the cysts example, it is clear that two sub populations exist (kidneys that sometimes produce cysts and kidneys that cannot), however in the case of the surname frequencies there are not any obvious mechanisms motivating potential sub populations to generate the value-inflation exhibited apart from the excess of unique surnames. Although the sub populations within the surnames distribution may not be obvious they do exist resulting from, for example, immigration or social mobility. However, since the value-inflation occurs due to surname diversity, it is very difficult to characterise this in terms of sub populations.

2.1.3 Long tails

Data with long tails may also be a feature of many epidemiological and clinical datasets. Long-tailed distributions occur in discrete datasets where the majority of the population take values around the average whilst a few large, often sparsely distributed values occur. Although relatively few, these observations may be crucial in analyses and may provide valuable information. However, such datasets cannot be easily modelled using standard discrete distributions.

An example of a distribution with a long tail is the counts of electroencephalographic seizures (ES) in coma patients as described in Section 1.3. The distribution of counts of ES provides an example where there is a large amount of variation due to a small number of severely ill patients which admit a higher number of ES than the majority of patients. This results in the highly skew distribution with a long tail seen in Figure 1.3, where most of the counts of ES ranges between 0 and 218, with two extreme observations with frequencies of ES of 458 and 531. This dataset is also zero-inflated due to the high proportion of zeros present (63.2%).

2.1.4 Truncation

A distribution is truncated if the range of possible values that observations can take is bounded, due to either being impossible to observe or to those values being ignored (Johnson et al., 2005). Distributions can be truncated from below, resulting in left-truncation where observations cannot occur below a certain value or truncated above, known as right-truncation, where above a certain range values are not present (Rose and Smith, 2002). Doubly-truncated distributions occur where the range of observations are both left and right truncated.

An example of a truncated data set is the numbers of births occurring in the UK and Ireland to HIV-infected women reported to the National Study of HIV in Pregnancy and Childhood, between 2000 and 2010 (French, 2011). In this dataset we only observe data from women who have given birth in the UK and this is a design condition which truncates the distribution of number of children born to HIV infected women to values of above 0.

2.1.5 Notation

Random variables are denoted by capital letters and their observed values by lower case letters. A random variable Y is said to be discrete if its realizations come from a finite sample space or are countable in an infinite sample space (Horgan, 2009, p. 133). A discrete random variable can be defined as a function,

$$Y : \Omega \longmapsto R_Y \ , \tag{2.1}$$

where $R_Y \subset \mathbb{Z}^t$ gives the range of the values of Y and Ω denotes the sample space. An example of a discrete variable in a infinite sample space is the length of stay of patients (in whole days) in hospital which may take values in the range $R_Y = \{0, 1, \ldots\}$. On the other hand, the number of correct responses on a test consisting of 10 questions is an example of a discrete variable with a finite sample space, where $R_y = \{0, 1, \ldots, 10\}$.

The *probability density function* (*pdf*) also frequently known as the *probability mass function* (*pmf*) of a discrete random variable Y is a function $f_Y : R_Y \mapsto [0, 1]$ defined as:

$$f_Y(y; \theta) = \mathrm{P}(Y = y) = p_y \ , \tag{2.2}$$

where θ is the set of parameters for the model and

$\sum_{y \in R_Y} P(Y = y) = 1$ (Zelterman, 2004).

The *cumulative distribution function* (*cdf*) or the *cumulative mass function* (*cmf*) for a discrete variable Y is defined by:

$$F_Y(y) = \mathrm{P}(Y \leq y) = \sum_{j=0}^{y} p_j \ , \tag{2.3}$$

(Horgan, 2009; Zelterman, 2004).

The *mean* of a discrete random variable, Y, is defined as the weighted average across all possible values,

$$\mu = \mathrm{E}(Y) = \sum_{y \in R_Y} y \, \mathrm{P}(Y = y) \tag{2.4}$$

(Rose and Smith, 2002). In general, the *expected-value*, denoted by $\mathrm{E}()$, of a function $g(Y)$, of a random variable Y is the weighted sum of its values,

$$\mathrm{E}(g(Y)) = \sum_{y \in R_Y} g(y) \, \mathrm{P}(Y = y) \tag{2.5}$$

(Horgan, 2009).

The *probability generating function* (*pgf*) gives an alternative representation of the *pdf* and provides a smooth transformation of the probabilities. The *pgf*, denoted by $G_Y(t)$, is:

$$G_Y(t) = \mathrm{E}[t^Y] = \sum_{y \in R_Y} t^y \, \mathrm{P}(Y = y) \ , \tag{2.6}$$

The *pgf* is a useful tool for analysing discrete distributions, as it is often easier to manipulate than the *pdf* for many models.

A *moment* provides a quantitative measure of the shape of *pdf*. The j^{th} order raw moments of a random variable Y with *pdf* f_Y, given by μ'_j is:

$$\mu'_j = \mathrm{E}[Y^j] = \sum_{y \in R_Y} y^j \, \mathrm{P}(Y = y) \ , \tag{2.7}$$

and is also known as the j^{th} moment or the j^{th} *moment about zero* (Zelterman, 2004).

Distributions can be characterized by a number of statistics such as the mean, variance or skewness, and the *central moments* of Y denoted by μ_j define them. Central moments or moments about the mean are:

$$\mu_j = \mathrm{E}\left[(Y - \mathrm{E}\,(Y))^j\right] \;. \tag{2.8}$$

The first-order moment (i.e. $j = 1$) gives the *mean* or *expected-value* of Y,

$$\mu = \mathrm{E}[Y] = \sum_{y \in R_Y} y\,\mathrm{P}(Y = y) \;. \tag{2.9}$$

The variance of Y is defined as:

$$\mathrm{Var}[Y] = \mathrm{E}[(Y - \mu)^2] = \mathrm{E}[Y^2] - \mu^2 = \mu_2' - \mu^2 = \mu_2 \;, \tag{2.10}$$

i.e. the second moment about μ. The skewness coefficient γ_1 can be calculated from the standardized third central moment as follows,

$$\gamma_1 = \mathrm{E}\left[\left(\frac{Y - \mu}{\sigma}\right)^3\right] = \frac{\mathrm{E}[(Y - \mu)^3]}{\mathrm{Var}[Y]^{\frac{3}{2}}} = \frac{\mu_3}{\sigma^3} \;, \tag{2.11}$$

where μ_3 is the third moment about μ and σ is the standard deviation. The standardized fourth central moment can be used to calculate the kurtosis coefficient, γ_2, given by,

$$\gamma_2 = \mathrm{E}\left[\left(\frac{Y - \mu}{\sigma}\right)^4\right] = \frac{\mathrm{E}[(Y - \mu)^4]}{\mathrm{Var}[Y]^2} = \frac{\mu_4}{\sigma^4} \;, \tag{2.12}$$

where μ_4 is the fourth moment about μ and σ again is the standard deviation.

The *moment generating function* (*mgf*) provides an alternative definition of a distribution's *pgf*. The general form for the *mgf* $M_Y(t)$ for a discrete random variable is,

$$M_Y(t) = \mathrm{E}[\mathrm{e}^{tY}] = \sum_{y \in R_Y} \mathrm{e}^{ty}\,\mathrm{P}(Y = y) \;. \tag{2.13}$$

The j^{th} raw moment of a distribution can be found by differentiating the *mgf* and solving it at zero, as follows

$$\mathrm{E}\left[Y^j\right] = M_Y^{(j)}\left(0\right) = \left.\frac{\partial^j\, M_Y(t)}{\partial t^j}\right|_{t=0} \ . \tag{2.14}$$

The *mgf* and *pgf* are related as follows,

$$M_Y(t) = G_Y(\mathrm{e}^t) \ . \tag{2.15}$$

Indices for discrete distributions

This section details measures which describe discrete distributions: the overdispersion, zero-inflation, third central moment inflation indices, Gini's coefficient and the surprise index.

Overdispersion Index

A useful quantity to characterize dispersion in a distribution is the index of dispersion, or *overdispersion index* given by,

$$OD = \frac{\sigma^2}{\mu} \ , \tag{2.16}$$

(Nikoloulopoulos and Karlis, 2008a). When $OD = 1$ the mean and variance are equal, and there is therefore no overdispersion. Values of $OD > 1$ indicate overdispersion is present in the model, whilst where $OD < 1$ the model is underdispersed. The index is widely used in the field of ecology as a measure of clustering (overdispersion) or repulsion (underdispersion) (Johnson et al., 2005).

Zero-inflation Index

The *zero-inflation index* introduced by Puig (2003) concerns the shape of the head of the distribution relative to its mean and is defined as,

$$ZI = 1 + \frac{\log(p_0)}{\mu} \ . \tag{2.17}$$

where p_0 is the probability of a value of zero. When Y follows a Poisson distribution then $ZI = 0$ and if the distribution is zero-inflated the index is $ZI > 1$ (Nikoloulopoulos and Karlis, 2008a).

Third central moment inflation index

The third central moment inflation index was introduced by Puig and Valero (2006) and provides another measure of skewness in the data. Denoted by κ_3 the index is given by

$$\kappa_3 = \frac{\mu_3}{\mu} - 1 \tag{2.18}$$

(Nikoloulopoulos and Karlis, 2008b). For the Poisson distribution κ_3 equals 0, larger values indicates a higher skew in the distribution.

Gini's coefficient

Gini's coefficient measures how large differences between observations are and therefore provides a measure of variability. The coefficient is given by,

$$\text{gini}(y) = 1 - \frac{\sum_{t=0}^{\infty} S_y(t)^2}{E(Y)} \ , \tag{2.19}$$

where $S_y^t(t) = P(Y > t)$ is the survival distribution of a discrete random variable Y (Nikoloulopoulos and Karlis, 2008b). Values of Gini's coefficient are between 0 and 1, with large values of Gini's coefficient indicating shorter tails.

Surprise Index

The *Surprise Index (SI)* is an empirical measure of how unexpected a value of a random variable is. An event with low probability is considered to be 'rare'; but Weaver questioned whether a rare event is always surprising. For example, winning the lottery (an interesting event) involves choosing the correct combination of a small set of numbers out of larger set of possible numbers. Whilst winning the lottery is certainly a rare event, it is not 'surprising' that somebody wins the lottery as each combination has an equal probability of occurring. Another well-known experiment, tossing a coin, may result in three possible outcomes heads, tails and edge and these occur with probabilities $\left\{\frac{1-\epsilon}{2}, \frac{1-\epsilon}{2}, \epsilon\right\}$, where ϵ, the probability a coin lands on its edge, is very small. A coin landing on its edge would thus be both a 'rare' and 'surprising' event as the probability of this occurring is very small compared with the other possible alternatives.

Let V_n, be a random variable representing the result of an experiment resulting in one of n possible outcomes, with probabilities of occurrence $p_1, p_2, \ldots, p_n$. Supposing the event V_i with probability p_i actually occurred, Weaver (1948) defined an index to measure how surprising the event V_i is as:

$$SI_i = \frac{\mathrm{E}(p)}{p_i} = \frac{p_1^2 + p_2^2 + \ldots + p_n^2}{p_i} \ . \tag{2.20}$$

The *SI* compares the probability of V_i occurring with the expected value of the model's probability. Thus a $SI_i = k$ means that the probability of E_i is k times smaller than the probability of all outcomes that the model refers to. If the *SI* is large then it can be considered as 'surprising' and therefore the *SI* measures whether the probability p_i is small compared with its expected probability $\mathrm{E}(p)$. Weaver (1948) suggested the somewhat arbitrary categories shown in Table 2.1 to determine if a value of SI may be considered as 'large' enough to correspond to a surprising event.

< 5	Not surprising
10	Begins to be surprising
1,000	Definitely surprising
1,000,000	Very surprising
10^{12}	Miracle!

Table 2.1: Weaver (1948)'s interpretation of *SI* values.

The SI can be used to assess whether a particular observation can be considered surprising assuming that the data come from a particular discrete probability model (Weaver, 1948; Redheffer, 1951). We can calculate the *SI* for a discrete distribution with *pdf* $f_Y(y;\theta)$, as follows,

$$SI_i = \frac{\mathrm{E}\left(f_Y(y;\theta)\right)}{f_Y(y;\theta)} = \frac{\sum_Y f_Y(y;\theta)\, f_Y(y;\theta)}{f_Y(y;\theta)} \tag{2.21}$$

where θ denotes the model's parameters. So far, for discrete distributions, analytical expressions of Surprise Indices have only been published for the Binomial and Poisson distributions, obtained by Redheffer (1951).

2.1.6 Special Functions

A large number of formulas and results are featured in this thesis, many of which contain special functions, which are defined in this section.

Binomial coefficient $\binom{n}{m}$

The binomial coefficient $\binom{n}{m}$ gives the number of different possible combinations of m items from n different items:

$$\binom{n}{m} = {}_nC_m = \frac{n!}{m!(n-m)!} = \frac{\Gamma(n+1)}{\Gamma(m+1)\,\Gamma(n-m+1)} \tag{2.22}$$

(Zelterman, 2004; Wimmer and Altmann, 1999; Johnson et al., 2005).

Binomial expansion ($(a+b)^n$)

The binomial theorem describes the algebraic expansion of powers of a binomial $(a+b)^n$ for a positive integer n as follows,

$$(a+b)^n = \sum_{j=0}^{n} \binom{n}{j}\; a^{n-j}\, b^j \tag{2.23}$$

(Johnson et al., 2005).

Polylogarithm ($Li_s(z)$)

The polylogarithm (also known as Jonqui's function) is a special function $Li_s(z)$ that is defined by the infinite sum or power series,

$$Li_s(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^s} \tag{2.24}$$

(Wimmer and Altmann, 1999).

Unit step function (U_x)

The Unit step function, also called the Heaviside step function is denoted by U_x, given by,

$$U_x = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{2.25}$$

The function is equal to 0 when $x < 0$ and 1 when $x \geq 0$.

Floor and Ceiling functions, ($\lfloor x \rfloor$) **and** ($\lceil x \rceil$)

Floor and ceiling functions map a real number to the largest previous or the smallest following integer, respectively. The floor function, $\lfloor x \rfloor$, is the largest integer not greater than x and the ceiling function, $\lceil x \rceil$ is the smallest integer not less than x. For example, $\lfloor 5.7 \rfloor = 5$ and the ceiling $\lceil 5.7 \rceil = 6$.

Bell polynomials ($Bl_n(x)$)

The Bell number of order n, Bl_n, is the number of ways to partition a set of n objects and can be calculated using the recursion formula,

$$Bl_{n+1} = \sum_{k=0}^{n} \binom{n}{k} \, Bl_k \tag{2.26}$$

where $Bl_0 = Bl_1 = 1$ (Johnson et al., 2005). The Bell polynomial of order n , $Bl_n(x)$, satisfies the following the generating function relation,

$$e^{(e^t-1)x} = \sum_{n=0}^{\infty} \frac{Bl_n(x)t^n}{n!} \tag{2.27}$$

which enables the Bell polynomial to be calculated (Johnson et al., 2005).

Pochammer Symbol ($(a)_j$)

Pochammer's Symbol, $(a)_j$ is used to denote ascending (or rising) factorials as follows,

$$(a)_j = a(a+1)\dots(a+j-1) \tag{2.28}$$

(Johnson et al., 2005).

Hermite polynomial ($H_y(n)$)

The Hermite polynomial, $H_y(n)$, is given by,

$$H_y(n) = \sum_{j=0}^{\lceil n/2 \rceil} \frac{n!y^{n-2j}}{(n-2j)!j!2^j} \tag{2.29}$$

(Johnson et al., 2005).

Gamma and Beta functions

Gamma function ($\Gamma(x)$)

The Gamma function is given by,

$$\Gamma(x) = \int_0^{\infty} t^{x-1}\,\mathrm{e}^{-t}\,dt \tag{2.30}$$

for $x > 0$ (Johnson et al., 2005).

Beta function ($B(a, b)$)

The Beta function $B(a, b)$ is as follows,

$$\begin{aligned} B(a,b) &= \int_0^1 t^{a-1}\,(1-t)^{b-1}\,dt \\ &= \tfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \end{aligned} \tag{2.31}$$

where $a, b > 0$ (Wimmer and Altmann, 1999, pg. XXII).

Digamma $(\psi(x))$

The derivatives of the logarithm of $\Gamma(x)$ are often required when calculating log-likelihoods of distributions. The digamma function, $\psi(x)$, is given by

$$\psi(x) = \frac{d}{dx}[\log\Gamma(x)] = \frac{\Gamma'(x)}{\Gamma(x)} \tag{2.32}$$

(Johnson et al., 2005).

Trigamma $(\phi'(x))$

The trigamma function gives the second derivative of the logarithm of $\Gamma(x)$,

$$\psi' = \frac{d}{dx}[\log\Gamma(x)] = \frac{d^2}{dx^2}[\log\Gamma(x)] \tag{2.33}$$

(Johnson et al., 2005).

<u>Hypergeometric Functions</u>

Generalized Hypergeometric functions ($_pF_q$)

The Generalized Hypergeometric function, $_pF_q$, has p numerator parameters and q denominator parameters and is defined as,

$$\begin{aligned} {}_pF_q\,[a_1,\ldots a_p; b_1,\ldots,b_q] &=_p F_q \left[\begin{matrix} {\scriptstyle a_1,\ldots a_p;\,x} \\ b_1,\ldots,b_q \end{matrix}\right] \\ &= \textstyle\sum_{j=0}^{\infty} \frac{(a_1)_j\ldots(a_p)_j\,x^j}{(b_1)_j\ldots(b_q)\,j!} \end{aligned} \tag{2.34}$$

where $b_i \neq 0, -1, -2, \ldots, i = 1, \ldots, q$

Gaussian Hypergeometric function ($_2F_1(a, b; c; x)$)

The Gaussian Hypergeometric function, or often more simply known as the Hypergeometric function is denoted by $_2F_1$ is a special case of the Generalized hypergeometric function where $p = 2$ and $q = 1$ and has the form,

$$\begin{aligned} {}_2F_1(a, b; c; x) &= 1 + \tfrac{a\,b}{c\,1!} + \tfrac{a(a+1)b(b+1)}{c(c+1)2!}x^2 + \cdots \\ &= \textstyle\sum_{j=0}^{\infty} \tfrac{(a)_j\,(b)_j\,x^j}{(c)_j\,j!}\,,\ c \neq 0, -1, -2, \ldots, \end{aligned} \tag{2.35}$$

where $(a)_j$ is Pochammer's symbol.

Confluent Hypergeometric function of the first kind ($_1F_1(a; b; x)$)

The confluent hypergeometric function of the first kind, denoted by $_1F_1(a; b; x)$ is a special case of the Generalized Hypergeometric function where $p = 1$ and $q = 1$. It can be written as a series as follows,

$$\begin{aligned} {}_1F_1(a; b; x) &= 1 + \tfrac{a}{b\,1!}x + \tfrac{a(a+1)}{b(b+1)\,2!}x^2 + \ldots \\ &= \textstyle\sum_{j=0}^{\infty} \tfrac{(a)_j x^j}{(b)_j j!}\,,\ c \neq 0, -1, -2, \ldots, \end{aligned} \tag{2.36}$$

where $(a)_j$ is Pochammer's symbol (Johnson et al., 2005; Wimmer and Altmann, 1999).

Confluent hypergeometric function of the second kind ($U(a, b, x)$)

The confluent hypergeometric function of the second kind, $U(a, b, x)$ is given as,

$$U(a, b, x) = \frac{1}{\Gamma(a)} \int_0^{\infty} \mathrm{e}^{-xt} t^{a-1} (1 + t)^{b-a-1} dt \tag{2.37}$$

for $a > 0$ and $x > 0$ (Johnson et al., 2005; Wimmer and Altmann, 1999).

Bessel functions

Bessel function of the first kind ($J_\nu(x)$)

The Bessel function of the first kind, $J_\nu(x)$ is,

$$J_\nu(x) = \left(\frac{x}{2}\right)^\nu \sum_{j=0}^{\infty} \frac{\left(-x^2/4\right)^j}{j!\,\Gamma(\nu+j+1)} \tag{2.38}$$

where ν is the order of the function (Johnson et al., 2005).

Modified Bessel function of the first kind ($I_\nu(x)$)

The modified Bessel function of the first kind is given by,

$$I_\nu(x) = (-i)^\nu J_\nu(ix) = \sum_{j=0}^{\infty} \frac{\frac{x^2}{4}^{j}}{j!\Gamma(\nu+j+1)} \ , \tag{2.39}$$

where $i^2 = -1$ (Johnson et al., 2005).

Bessel function third kind ($K_\nu(x)$)

The modified Bessel function of the third kind, $K_\nu(\cdot)$, is defined as,

$$K_\nu(x) = \frac{\pi}{2} \frac{I_{-\nu}(x) - I_\nu(x)}{\sin(\nu\pi)} \ , \tag{2.40}$$

when ν is not an integer or zero (Johnson et al., 2005; Wimmer and Altmann, 1999).

Lerch functions

Lerch function ($\Phi(p, a, c)$)

The Lerch function also known as the Hurwitz Zeta function is,

$$\Phi(p, a, c) = \sum_{y=1}^{\infty} \frac{p^y}{(a+y)^c} \ , \quad p > 0, \ a > 0 \ . \tag{2.41}$$

Riemann zeta function, ($\zeta(x)$) and ($\zeta(x, a)$)

The Riemann zeta function, $\zeta(x)$, is as follows,

$$\zeta(x) = \sum_{j=1}^{\infty} j^{-x} \ , \tag{2.42}$$

for $x > 1$ (Johnson et al., 2005). A generalized form of the Riemann zeta function, $\zeta(x, a)$, is defined as,

$$\zeta(x, a) = \sum_{j=1}^{\infty} (j + a)^{-x} \ , \tag{2.43}$$

for $x > 1$ and $a > 0$ (Johnson et al., 2005).

2.2 Basic Distributions

Several discrete distributions which have been well established in the statistical literature are described in this section. Two important classes of *pdf*'s are the exponential family and distributions generated by Urn models. Both are described in the following sections.

Exponential Family

The *exponential family* is a class of probability distributions which includes all distributions (both discrete and continuous) where the *pdf* can be expressed in the form,

$$f(y_i; \theta_i) = \exp\left[d(\theta_i)\, e(y_i) + g(\theta_i) + h(y_i)\right] \ , \tag{2.44}$$

where d, e, g and h are known functions with the same form for all y_i (Dobson, 2002). Alternatively, this can be parametrized to include an additional dispersion parameter, ϕ, that is constant for all y_i and $d(\theta_1) = \theta$ and $e(y_i) = y_i$ are replaced. Called the "natural form" (McCullagh and Nelder, 1983), this can be written as

$$f_Y(y, \theta; a, b, c, d) = \exp\left[\frac{\theta_i\, y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right] \ , \tag{2.45}$$

where a, b, c are all known to have the same form for y_i. The parameter θ_i is known as the canonical parameter, $b(\theta_i)$ is the cumulant, $a(\phi)$ is the scale parameter, equal to 1 in discrete models and $c(y_i, \phi)$ is a normalization term, guaranteeing that the *pdf* sums to one. In the exponential family form of the *pdf* the first and second derivatives of the cumulant with respect to θ give the mean and variance:

$$\mu = b'(\theta_i) \quad \text{and} \quad \sigma^2 = b''(\theta_i) \quad \cdot \qquad (2.46)$$

The exponential family of distributions provides a framework for selecting a parametrization of the distribution via natural parameters and can be used to define sample statistics (McCullagh and Nelder, 1983). Many well-known distributions can be expressed in an *exponential family* form, including the Normal, exponential, gamma, chi-square, beta, Bernoulli, binomial, Poisson, negative binomial and many others.

Urn Models

The concept of urn models have a very long history in probability and has been widely applied in many fields such as genetics, capture-recapture sampling of animal populations, learning processes and filing systems (Johnson et al., 2005) In the basic model, an urn contains n white and m black balls. A ball is drawn randomly from the urn and its colour observed; it is then placed back in the urn, and the selection process is repeated. A variation is that the balls may be drawn without replacement. We are interested in modelling, for example, the distribution of the number of white balls after a fixed number of trials, the outcome of a fixed number of selections or the discrete waiting time until a specified set of conditions are fulfilled. A wide variety of well-known discrete distributions can be obtained in terms of urn models e.g. the Hypergeometric, Binomial, Geometric, negative binomial, beta-binomial and Poisson distributions. *Pdf*'s in the generalized hypergeometric family e.g. the Hermite and Generalized Gegenbauer distributions can be generated in terms of urn outcomes.

This section presents a series of frequently applied discrete distributions that have been well established within the statistical literature: the Bernoulli, Binomial, Geometric, Hypergeometric and Poisson distributions. These distributions can be modified and form the basis of more complex discrete probability models.

2.2.1 Bernoulli (p)

The simplest example of a discrete random variable are *Bernoulli* random variables, named after Jacob Bernoulli (1654-1705) (Hald, 1998)). Bernoulli random variables have outcomes $R_Y = \{0, 1\}$ referred to as successes and failures respectively (Zelterman, 2004). The *Bernoulli* distribution has one parameter p representing the probability of success, where:

$$\begin{aligned} P(Y=1) = &\quad p \\ P(Y=0) = &\quad 1-p \end{aligned} \quad , \tag{2.47}$$

where $0 \leq p \leq 1$ (Zelterman, 2004). The Bernoulli distribution is generated from an urn model where a single ball is sampled from an urn containing black and white balls. For a Bernoulli distributed random variable, Y, the *pdf* is given by,

$$f_Y(y;p) = P(Y=y) = p^y(1-p)^{1-y} \ , \tag{2.48}$$

where $y \in \{0, 1\}$ and $0 < p < 1$ (Wimmer and Altmann, 1999; Rose and Smith, 2002). The Bernoulli distribution plays a key role in many statistical models and is a member of the exponential family. The *pgf* of the Bernoulli distribution is,

$$G(t) = (1-p) + p\,t \ , \tag{2.49}$$

and the distribution has *mgf*,

$$M(t) = 1 + \left(\mathrm{e}^t - 1\right) p \ . \tag{2.50}$$

The mean of the Bernoulli distribution is p and the variance $p(1-p)$. The overdispersion index of the Bernoulli distribution is $(1-p)$ and since the parameter $0 < p < 1$, the

OD index indicates that the Bernoulli distribution will always be underdispersed. The zero-inflation index is given by $1+\frac{\log(1-p)}{1-p}$. For small values of p the ZI index indicates large amounts of zero-inflation in the Bernoulli distribution and as p approaches 1, the ZI index tends to 0. The SI of the Bernoulli distribution is given by,

$$SI_y = (1-p)^{y-1}p^{-y}(1+2(p-1)p) \; . \tag{2.51}$$

Since, p models the probability of success, where p is small the SI is larger where $Y = 1$ than 0 and where p is large this is reversed, i.e. the SI is larger for $Y = 0$.

2.2.2 Binomial (p, n)

A Binomial random variable, represents the number of successes in n trials, where each trial is an independent and identically distributed Bernoulli random variable with two possible outcomes: success with probability p and failure with probability $q = 1 - p$ (Horgan, 2009; Rose and Smith, 2002). The Binomial distribution is an example of an urn model where balls are sampled with replacement from an urn containing p black and $1 - p$ white balls until n balls are drawn. It can also be calculated as the sum of n Bernoulli random variables, with *pdf* modelling the probability that exactly y successes in n trials will occur,

$$f_Y(y; p, n) = P(Y = y) = \binom{n}{y} p^y(1-p)^{n-y} \; , \tag{2.52}$$

for $y = 0, 1, \ldots$ where $\binom{n}{m}$ is the binomial coefficient. The valid parameter values are $0 \leq p \leq 1$ and $n = 1, 2, \ldots$. The Binomial distribution can be derived from the binomial expansion $(p + q)^n$ (Rose and Smith, 2002). Figure 2.1 shows the binomial *pdf* for two samples sizes, n, of 20 and 40 and values of p of 0.2, 0.5 and 0.7. As the parameter n tends to infinity the Binomial distribution is approximated by a Normal distribution with mean $n\,p$ and variance $n\,p\,(1-p)$.

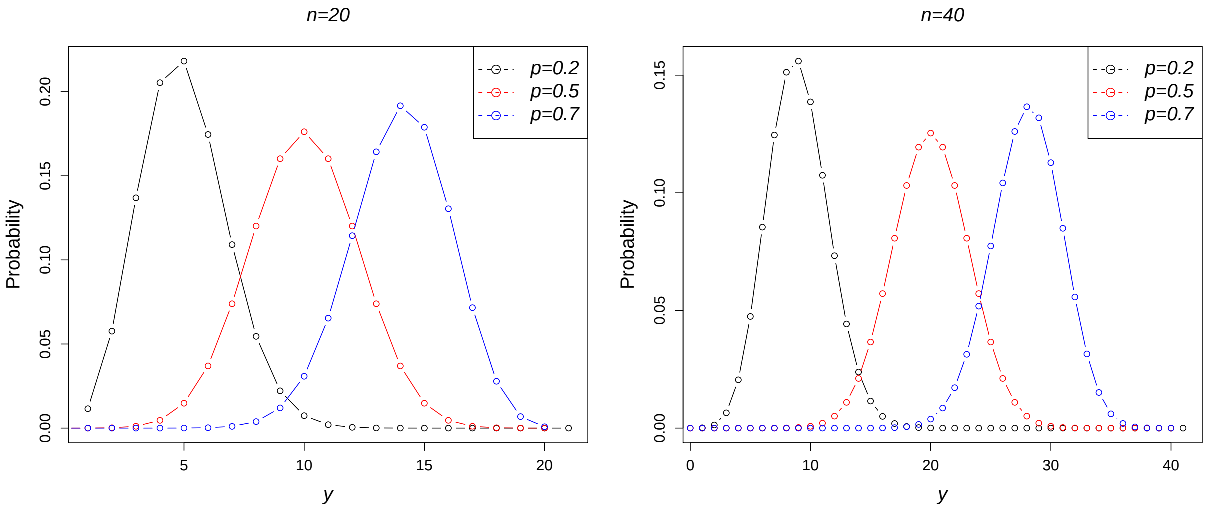


Figure 2.1: Binomial *pdf*

The *pgf* of the Binomial distribution is given by,

$$G(t) = (1 + p(t-1))^n \ , \tag{2.53}$$

and the *mgf*,

$$M(t) = \left(1 + \left(\mathrm{e}^t - 1\right) p\right)^n \ . \tag{2.54}$$

The Binomial distribution has mean $n\,p$ and variance $n\,p\,(1-p)$ and its overdispersion index is $OD = n\,(1-p)$. Where p is small the dispersion in the Binomial distribution is large, as p increases the OD index decreases. For larger values of n the OD also increases. The zero-inflation index is given by $ZI = \dfrac{1 + \log((1-p)^n)}{n\,p}$.

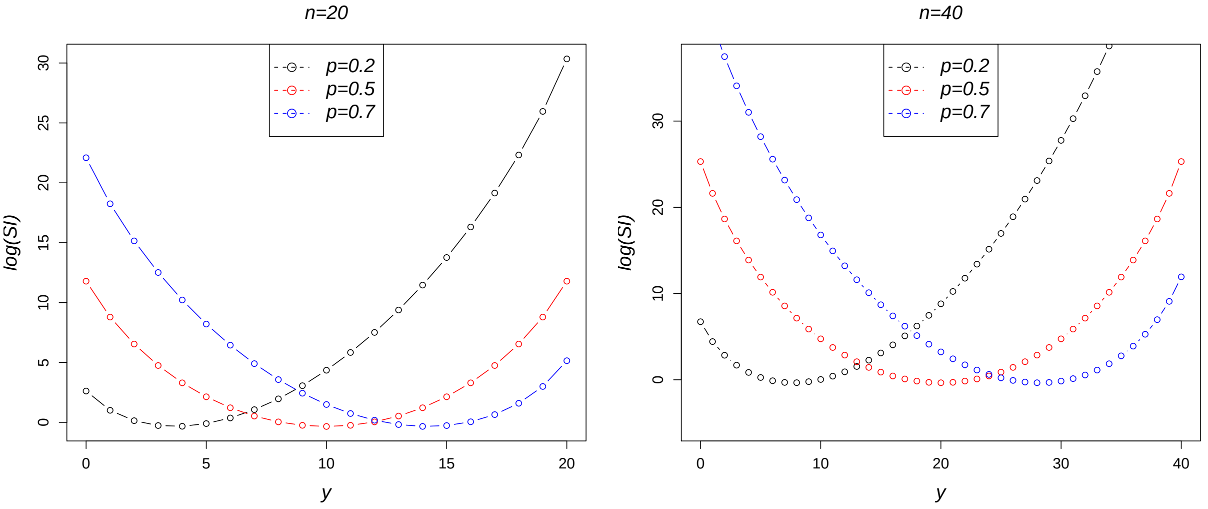


Figure 2.2: log(*SI*)'s for Binomial distributions

The SI of the Binomial distribution is,

$$SI_y = \frac{(1-p)^{y-n}(p-1)^{2n}p^{-y}\,{}_2F_1\left(-n,-n;1;\frac{p^2}{(p-1)^2}\right)}{\binom{n}{y}} . \tag{2.55}$$

A range of SI's are plotted for the Binomial distribution in Figure 2.2. In the first plot n is fixed at 20 with p in the range 0.2, 0.5 and 0.7. For smaller values of p the SI is skewed being more surprising for high Y values but as p increases the skew reverses and low values of Y are more surprising. The second plot plots SI's where $n = 40$ and $p = 0.2, 0.5$ and 0.7 and also illustrates this same pattern.

2.2.3 Geometric (p)

The Geometric distribution also arises from a series of Bernoulli trials. If p denotes the probability of success in repeated independent Bernoulli trials, then we are interested in the probability that the *first* success occurs on the yth trial (Rose and Smith, 2002). This distribution is an urn model where balls are sampled with replacement from an urn containing p white and $1-p$ black balls, until a white ball is drawn. The *pdf* is then given by,

$$f_Y(y;p) = P\left(Y=y\right) = p\left(1-p\right)^y \ , \tag{2.56}$$

where $y = 0, 1, 2, \ldots$ and $0 < p < 1$ (Wimmer and Altmann, 1999). The probability distribution is shown in Figure 2.3 for values of p of 0.2, 0.4, 0.6 and 0.8. As the value of p increases we can see the probability of a low value of y increases.

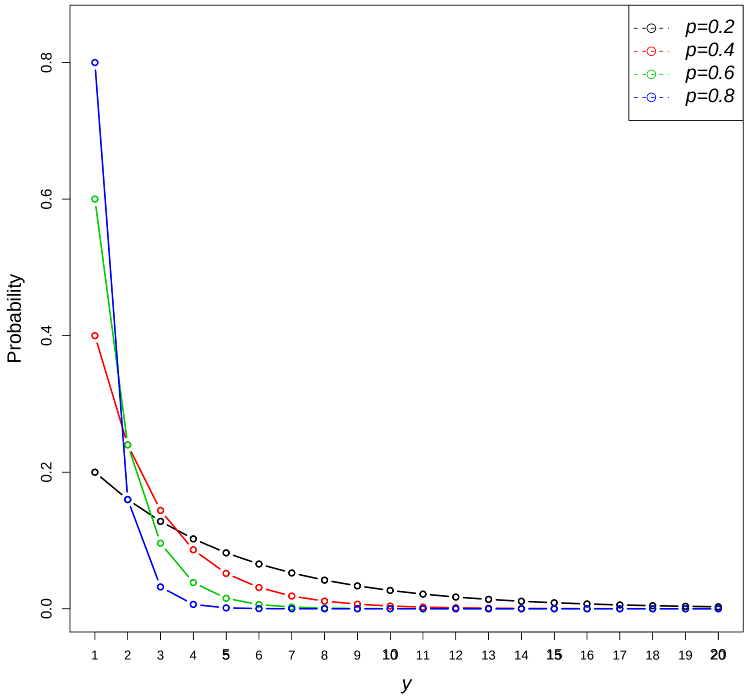


Figure 2.3: Geometric *pdf*

The Geometric distribution has *pgf*,

$$G(t) = \frac{p}{1 + (p-1)t} \, , \tag{2.57}$$

and the *mgf* is given by,

$$M(t) = \frac{p}{1 + \mathrm{e}^t (p-1)} \, , \tag{2.58}$$

The mean and variance of this distribution are given by $\frac{1-p}{p}$ and $\frac{1-p}{p^2}$, respectively and the overdispersion index of the Geometric distribution is $\frac{1}{p}$, with small values of p resulting in an overdispersed distribution, and which as p approaches 1 the dispersion decreases. The zero-inflation index $1 + \frac{p \log(p)}{1-p}$ indicates that as p increases the amount of zero-inflation in the distribution decreases.

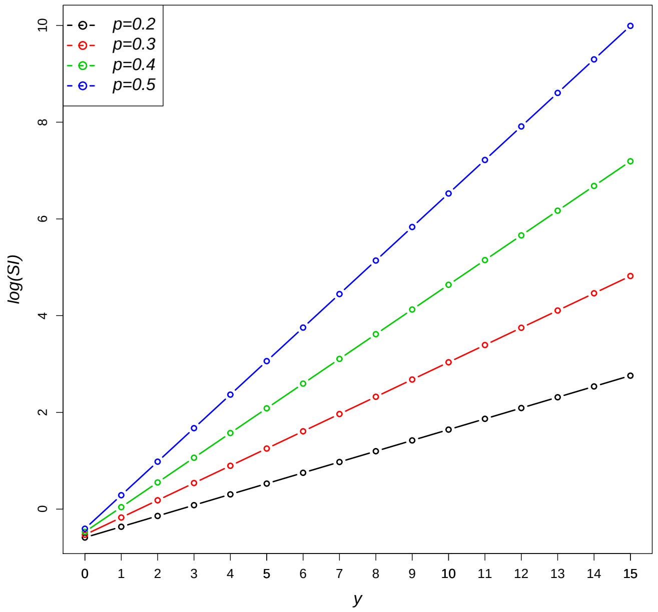


Figure 2.4: log(*SI*) for the Geometric distribution

The SI for the Geometric distribution is,

$$SI_y = \frac{(1-p)^{-y}}{2-p} \; . \tag{2.59}$$

Figure 2.4 plots $\log(SI)$ for the Geometric distribution with values of p of 0.2, 0.3, 0.4 and 0.5. As p increases, the SI increases for large values of y.

2.2.4 Hypergeometric (m, n, k)

Classical urn models in which balls are repeatedly drawn without replacement lead to the Hypergeometric distribution, in contrast to sampling without replacement which produces a Binomial distribution (Rose and Smith, 2002). A hypergeometric random variable Y counts the number of successes in a sample of size k drawn without replacement from a population of size $m + n$ where m is the number of successes in the population and n is the number of failures. The *pdf* gives the probability of getting exactly y

successes when drawing k elements without replacement from $m + n$ and it is:

$$f_Y(y; m, n, k) = P(y = y) = \frac{\binom{m}{y}\binom{n}{k-y}}{\binom{m+n}{k}} \tag{2.60}$$

for $y = 0, 1, \ldots, \min(m, k)$ (Wimmer and Altmann, 1999; Johnson et al., 2005; Horgan, 2009). The hypergeometric distribution can be formulated as an urn model where Y is the number of white balls drawn in a sample of k balls from an urn with m white balls and n black balls.

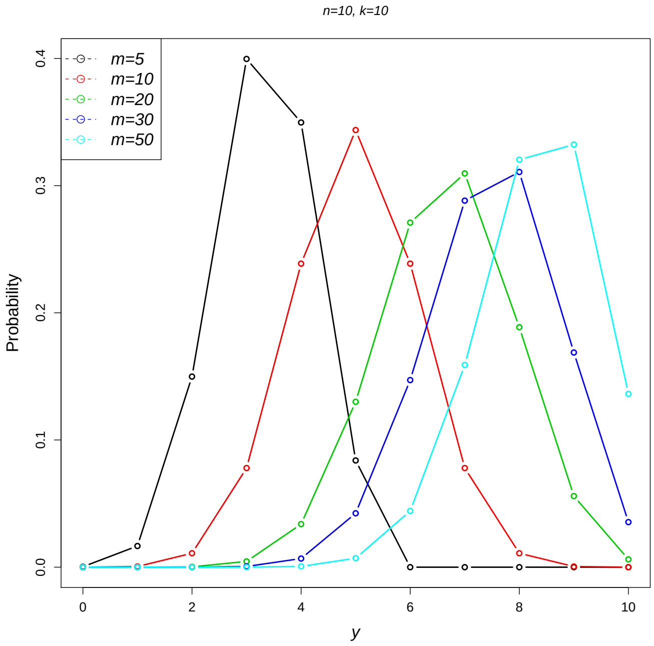


Figure 2.5: Hypergeometric *pdf*

Figure 2.5 plots several hypergeometric *pdf*. The plot shows the distribution for values of m of 5, 10, 20, 30 and 50 where $n = 10$ and $k = 10$. As the value of m increases (i.e. the number of successes in the population) the peak of the distribution i.e. the mean number of successes, increases. Similarly, if the the values of $m = 10$ and $k = 10$ and n (the number of failures) varies by 5, 10, 20, 30 and 50, as the number of failures increases, the mean number of successes decreases, i.e. the distributions peak shifts to the left in a mirror image of Figure 2.5.

The Hypergeometric distribution gets its name from the fact that the Gaussian hypergeometric function features in the *pgf* (Rose and Smith, 2002),

$$G(t) =_2 F_1(-k, -m; -(n+m); 1-t) \ . \tag{2.61}$$

where ${}_2F_1(a, b, c; x)$ is the Gaussian hypergeometric function, with a, b, c and x real numbers. The *mgf* is,

$$M(t) = \frac{\binom{n}{k}}{\binom{m+n}{k}} \, {}_2F_1(-k, -m; n-k+1; \mathrm{e}^t) \ . \tag{2.62}$$

The Hypergeometric distribution has mean and variance,

$$\mu = \frac{k\, m}{m+n} \quad \text{and} \quad \sigma = \frac{k\, m\, n(m+n-k)}{(m+n-1)(m+n)^2} \ , \tag{2.63}$$

The overdispersion index is,

$$OD = \frac{n(m+n-k)}{(m+n-1)(m+n)} \ , \tag{2.64}$$

For larger values of m or k, the OD index increases however when n increases the OD index decreases. The zero-inflation index is,

$$ZI = 1 + \frac{(m+n)\log\left(\frac{\binom{n}{k}}{\binom{m+n}{k}}\right)}{k\, m} \ . \tag{2.65}$$

and the SI is,

$$SI_y = \frac{\binom{n}{k}^2 \, {}_4F_3(-k, -k, -m, -m; 1, 1-k+n, 1-k+n; 1)}{\binom{m}{y}\binom{n}{k-y}\binom{m+n}{k}} \ , \tag{2.66}$$

where ${}_4F_3(a_1, \ldots, a_P; b_1, \ldots, b_Q; x)$ is the generalized hypergeometric function with $P = 4$ and $Q = 3$.

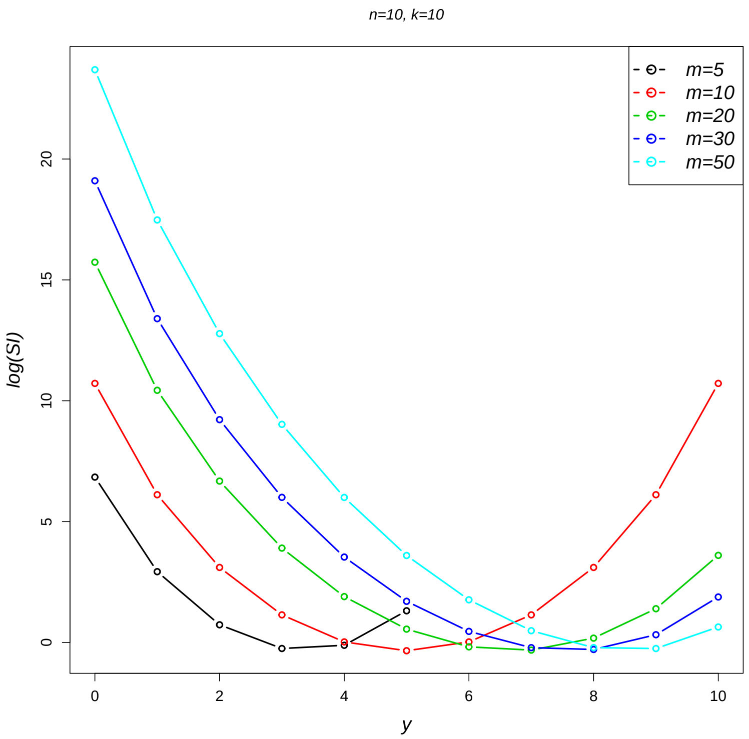


Figure 2.6: log(*SI*)'s for Hypergeometric distributions.

SI's for the Hypergeometric distribution are plotted in Figure 2.6 for values of m of 5, 10, 20, 30 and 50 with $n = 10$ and $k = 10$. As m increases the SI's for low values of Y increases and similarly as n increases the SI's for high values of Y increases.

2.2.5 Poisson (μ)

The most commonly used model for discrete data is the Poisson distribution. It was first discussed by Siméon-Denis Poisson (1781-1840) in 1838 (Hald, 1998). For the random variable Y representing discrete observations the Poisson probability distribution function is,

$$f_Y(y;\mu) = P\left(Y = y\right) = \frac{\mu^y \mathrm{e}^{-\mu}}{y!}\ , \tag{2.67}$$

where $y = 0, 1, 2, \ldots$ are discrete counts and μ is the mean of the Poisson distribution (Johnson et al., 2005; Wimmer and Altmann, 1999). It is considered an urn model where sampling is from an infinite number of urns each with an infinite number of white and black balls, where Y is the number of black balls drawn.

The Poisson probability distribution can also model rates with *pdf*,

$$f_Y(y,t;\mu) = P\left(Y = y\right) = \frac{(\mu t)^y \mathrm{e}^{-(\mu t)}}{y!} \tag{2.68}$$

where t is the length of time during which events occur (Hilbe, 2007). The rate variable, t, can be entered into regression models using its natural logarithm as a known offset in the model, $\ln(\mu) = Y\beta + \ln(t)$ where β is the matrix of covariates and μ the parameter of the Poisson distribution is the mean number of events (Hilbe, 2007). An offset is used to describe the time period in rates, and in this model the number of events y is proportional to the time period t.

Figure 2.7 illustrates the *pdf* of the Poisson distribution for increasing values of the mean, μ of 2, 5, 10 and 20. This graph illustrates the extent of skewness of the Poisson distribution, particularly for small values of μ and shows how the Poisson distribution approaches the Normal distribution as μ tends to infinity.

The Poisson distribution has *pgf*,

$$G(t) = e^{(t-1)\mu}\ , \tag{2.69}$$

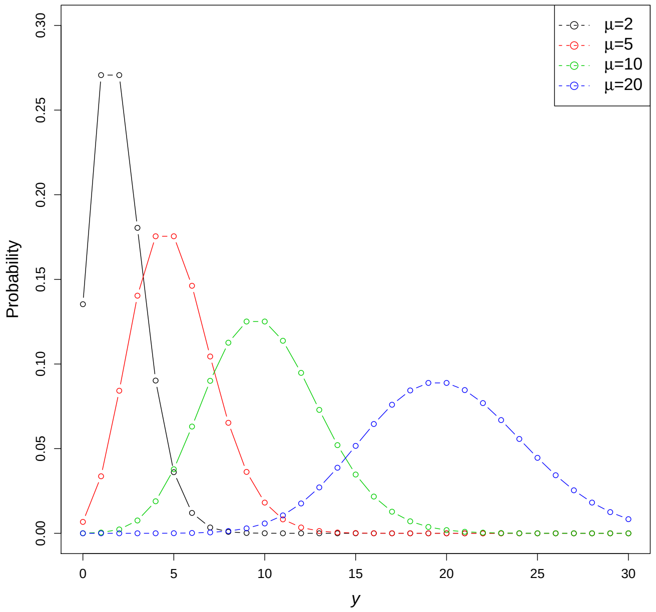


Figure 2.7: Poisson *pdf*

and *mgf*,

$$M(t) = \mathrm{e}^{(\mathrm{e}^{t}-1)\mu} \ . \tag{2.70}$$

A property of the Poisson distribution is that the mean and variance are equal i.e. $\mathrm{E}[Y] = Var[Y] = \mu$ or for rates $\mathrm{E}[Y] = \mathrm{Var}[Y] = \mu t$ and therefore the Poisson distribution cannot model overdispersion (Cox, 1986). The overdispersion index for the Poisson distribution is $OD = \frac{\mu}{\mu} = 1$ indicating no overdispersion is present under a Poisson model and the zero inflation index $ZI = 0$, there is no zero-inflation present in the Poisson distribution. The SI for this distribution is,

$$SI_y = \mathrm{e}^{-\mu}\, \mu^{-y}\, \mathrm{I}_0\,(2\mu)\; y! \ . \tag{2.71}$$

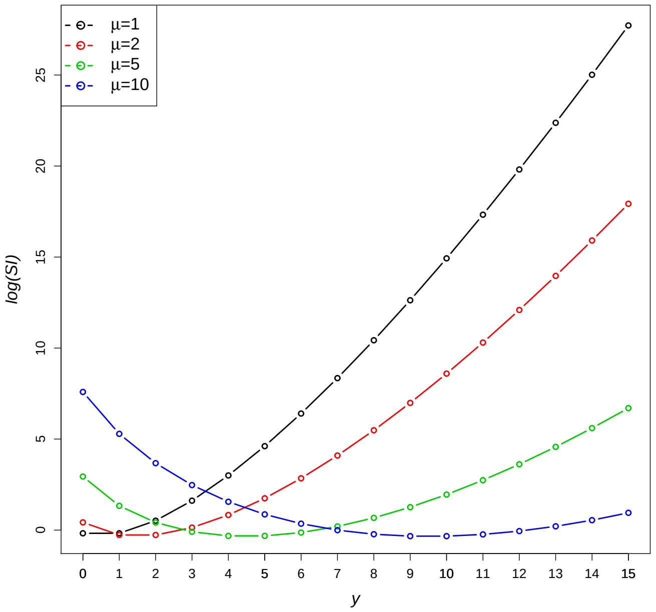


Figure 2.8: log(SI)'s for Poisson distribution

The logarithm of the SI for the Poisson distribution is shown for $\mu = 1, 2, 5, 10$ in Figure 2.8. As the value of μ increases the SI becomes less skew for higher values of y and as μ becomes large the SI is higher for low y values.

In a Poisson model the variance is μ and this means the dispersion in the data is fixed at 1. *Quasi-Poisson models* allow us to deal with overdispersion in a Poisson model by not restricting the the dispersion parameter to 1 but by estimating it from the data (Zeileis et al., 2008). This model has the same parameter estimates as the standard Poisson model but inference is adjusted for over-dispersion. However the Quasi-Poisson model does not have a fully specified likelihood.

2.3 Parameter-Mix Distributions

Distributions with long tails or multi-modality can be formed by the method of *mixing* distributions. Discrete distributions formulated through parameter mixtures of distributions are described in this section. A parameter-mix distribution is defined by the *pdf* of a random variable Y being dependent on the parameters $\theta_1, \theta_2, \ldots, \theta_m$ where some (or all) of those parameters are random variables varying according to other continuous distributions. The new distribution then has the *pdf*,

$$f_Y(y \mid \theta_1, \ldots, \theta_m) \ , \tag{2.72}$$

(Johnson et al., 2005; Rose and Smith, 2002; Willmot, 1986).

If only one parameter θ varies, the following notation can be used to denote a parameter mixture,

$$f_A \bigwedge_{\Theta} f_B \tag{2.73}$$

where f_A represents the original distribution and f_B represents the distribution of the random variable corresponding to the parameter θ known as the mixing distribution with parameter space Θ (Johnson et al., 2005).

Parameter mixtures of Poisson distributions allow for overdispersion by adapting the mean parameter to vary according to another distribution, for example as a frailty model or by incorporating random effects (Johnson et al., 2005). The Negative Binomial, Holla, Sichel and Delaporte distributions are all parameter mixtures of Poisson distributions and are presented in this section. Other distributions included are the Yule, and Waring distributions which are formed as mixtures of Geometric distributions and the Beta-Binomial distribution which is a mixture of a binomial distribution.

2.3.1 Negative Binomial

There are three different ways the negative binomial distribution is commonly parameterized. These are the negative binomial with parameters r and p and the negative binomial types I and II with parameters α and μ.

Negative Binomial (p, r)

The negative binomial distribution can be defined using an expansion of the negative binomial series $(1-p)^{-r} = \sum_{k=0}^{\infty} \binom{k+r-1}{r-1} p^k$. For count data, y, the $(y+1)$th term gives the *pdf*, and produces the probability, p, of observing y failures before the

r^{th} success in a series of Bernoulli trials,

$$f_Y(y;p,r) = P\left(Y = y\right) = \binom{y+r-1}{r-1} p^r(1-p)^y \ , \qquad (2.74)$$

for discrete observations $y = 0, 1, 2, \ldots$, $0 < p < 1$ and $r > 0$ (Johnson et al., 2005). The geometric distribution is also a special case of the negative binomial distribution when $r = 1$ in Equation 2.74.

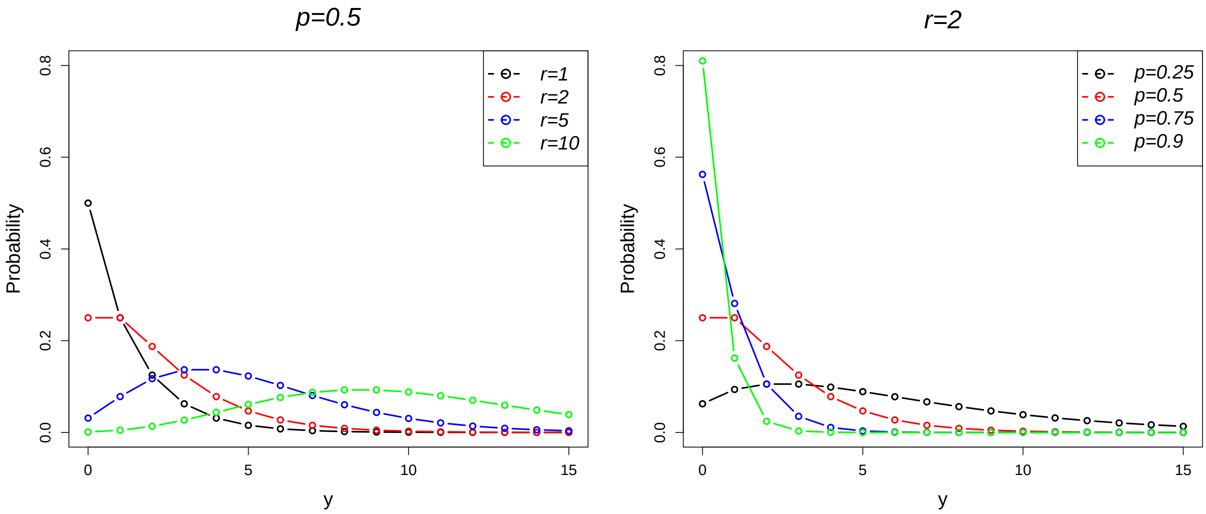


Figure 2.9: Negative Binomial *pdf*

Figure 2.9 shows the negative binomial distribution for increasing values of r of 1, 2, 5 and 10 when p is fixed at 0.5 (first plot) and for values of p of 0.25, 0.5, 0.75 and 0.9 when r is fixed at 2 (second plot). As r increases the distribution tends to a normal distribution. The parameter p adjusts the height of the probability distribution for low counts of y with high values of p closer to 1 resulting in a large skew at low y values and values of p close to 0 result in a low probability at low values of y. The *pgf* for the negative binomial distribution is,

$$G(t) = p^r(1 + (p-1)t)^{-r} \ , \qquad (2.75)$$

with *mgf*,

$$M(t) = \left(1 + \mathrm{e}^t(p-1)\right)^{-r} p^r \ . \qquad (2.76)$$

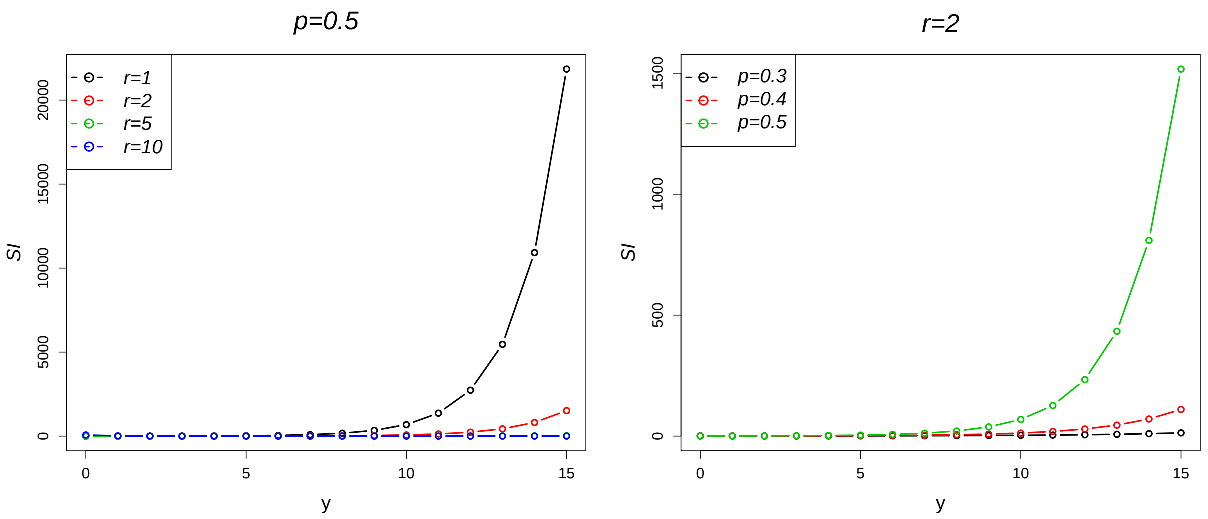


Figure 2.10: log(*SI*)'s for negative binomial distributions

The mean of the NB distribution is $\dfrac{r\,(1-p)}{p}$ and the variance $\dfrac{r\,(1-p)}{p^2}$. The overdispersion index is therefore given by $OD = \dfrac{1}{p}$ and is the same OD as the Geometric distribution. The OD index indicates that as the value of the parameter p approaches 0, the dispersion in the NB distribution increases, i.e. as you have to wait longer to complete r successes. The zero-inflation index is $ZI = 1 + \dfrac{p\log(p)}{(1-p)}$ and for values of p around 0, the ZI index is close to 1, indicating zero-inflation is present. As p increases the ZI index approaches 0.

The SI for the NB distribution is,

$$SI_y = \frac{(1-p)^{-y}\,p^r\,{}_2F_1\left(r,r,1,(p-1)^2\right)}{\binom{y+r-1}{r-1}}\,, \tag{2.77}$$

and is plotted in Figure 2.10. In the first plot, the parameter p is fixed at 0.5, where r is in the range 1, 2, 5 and 10, and demonstrates that when r is small the SI is large for large values of r, but becomes less skew as r increases. In the second plot r is fixed at 2 and the parameter p takes values 0.3, 0.4 and 0.5. The SI is again large for large values of y when p approaches 1.

It is often more convenient to form the negative binomial distribution in terms of the mean and a dispersion parameter, as opposed to the parameters r and p used in

Equation 2.74. Converting Equation 2.74 to the natural form of exponential family (defined in Section 2.2) gives:

$$f\left(y;p,r\right) = P(Y=y) = \exp\left\{ y\ln(p) + r(\ln(p)) + \ln\begin{pmatrix} y+r-1 \\ r-1 \end{pmatrix} \right\} \tag{2.78}$$

Hilbe (2007) and from this the cumulant $b(\theta_i)$ can be recognized as $-r\log(p)$ and thus the mean is $\mu = \dfrac{r(1-p)}{p}$ and the variance $\sigma^2 = \dfrac{r(1-p)}{p^2}$. We can then re-parametrize Equation 2.78 in terms of the mean μ and a dispersion parameter $\alpha = \dfrac{1}{r}$ giving a negative binomial *pdf*,

$$f_Y(y;\mu,\alpha) = P(Y=y) = \begin{pmatrix} y+\alpha-1 \\ \alpha-1 \end{pmatrix} \left(\frac{1}{1+\frac{\mu}{\alpha}}\right)^{\alpha} \left(\frac{\frac{\mu}{\alpha}}{1+\frac{\mu}{\alpha}}\right)^{y}, \tag{2.79}$$

where $y = 0,1,2,\ldots$, the mean, μ, lies in the range $\mu > 0$, and $\alpha > 0$ is the overdispersion parameter. Alternatively the negative binomial distribution can be generated through a Poisson-Gamma parameter-mix distribution, where the Gamma distribution is given by $\Gamma(a,b) = y^{a-1}\dfrac{\mathrm{e}^{-\frac{y}{b}}}{b^a\,\Gamma(a)}$ for $y \geq 0$ and $a,b > 0$ where $\Gamma(x)$ is the Gamma function (Johnson et al., 2005). If the discrete observations, y, follow a Poisson distribution with mean μ, the mean can then be assumed to vary across individuals according to a Gamma distribution with shape and scale parameters a and b. The result is a negative binomial distribution with parameters a and b:

$$f_Y(y;a,b) = P\left(Y=y\right) = \begin{pmatrix} y+a-1 \\ a-1 \end{pmatrix} \left(\frac{b}{b+1}\right)^{y} \left(\frac{1}{b+1}\right)^{a} \tag{2.80}$$

The shape and scale parameters determine either a negative binomial type I or type II distribution as follows (Booth et al., 2003).

Negative Binomial Type I (NBI)

The first form of the negative binomial distribution can be derived directly from the *pdf* or can be derived using a Poisson-Gamma parameter-mix distribution, where observations y are assumed to follow a Poisson distribution with mean μ and μ is assumed to vary according to a Gamma distribution, with shape and scale parameters α and $\frac{\alpha}{\mu}$ i.e. $\text{NB}(Y;\alpha,\frac{\alpha}{\mu}) = Po\,(\mu) \bigwedge\limits_{\mu} \Gamma\left(\alpha,\frac{\alpha}{\mu}\right)$ (Booth et al., 2003). This gives the following *pdf*:

$$f_Y(y;\alpha,\mu) = P\,(Y=y) = \binom{\alpha+y-1}{\alpha-1}\left(\frac{\frac{\alpha}{\mu}}{\frac{\alpha}{\mu}+1}\right)^{y}\left(\frac{1}{\frac{\alpha}{\mu}+1}\right)^{\alpha}, \qquad (2.81)$$

where $\alpha > 1$ is the overdispersion parameter and μ is the mean (Anscombe, 1950; McCullagh and Nelder, 1983, p. 194).

The variance of this form of the negative binomial distribution is $\mu + \frac{\mu^2}{\alpha}$ (Booth et al., 2003). This form uses the canonical link for the negative binomial distribution, $\eta = \log\left(\frac{\mu}{\mu+\alpha}\right)$ (Hilbe, 2007).

Negative Binomial II (NBII)

A second version of the negative binomial distribution can again be formulated from a Poisson-Gamma parameter-mix where $\alpha\,\mu$ and α are the shape and scale parameters respectively, i.e. $\text{NB}(Y;\alpha\,\mu,\mu,\alpha) = Po\,(\mu) \bigwedge\limits_{\mu} \Gamma\,(\alpha\,\mu,\alpha)$ with *pdf*,

$$f_Y(y;\alpha,\mu) = P\,(Y=y) = \binom{\alpha\mu+y-1}{\alpha\mu}\left(\frac{\alpha}{\alpha+1}\right)^{y}\left(\frac{1}{\alpha+1}\right)^{\alpha\mu}, \qquad (2.82)$$

again for $y = 0,1,2,\ldots$, $\alpha > 0$ and $\mu > 0$ (McCullagh and Nelder, 1983, p.132) and (Johnson et al., 2005, p.200). The mean of this distribution is again μ, however the variance is now $\mu + \frac{\mu}{\alpha}$ (Booth et al., 2003). This second type of the negative binomial distribution uses a logarithmic link $\eta = \ln(\mu)$ (Hilbe, 2007). For both versions of the distribution, as α, the overdispersion parameter, tends to infinity the negative binomial distribution becomes the Poisson distribution.

Both the type I and type II forms of the negative binomial distribution can as considered as members of the exponential family of distributions. The type I distribution uses the canonical link and the type II distribution uses a logarithmic link, which allows for comparison of estimates to the Poisson distribution.

2.3.2 Holla (α, θ)

This distribution was initially proposed by Holla (1966) as a parameter mix of a Poisson and an Inverse-Gaussian distribution (IG) (a two parameter continuous probability distribution for $\mu > 0$). The Holla distribution can be written as, Holla(α, θ)=Poisson(μ) $\bigwedge_{\mu}$ IG(θ, α) where the *pdf* of the Inverse-Gaussian distribution with parameters α and θ is as follows,

$$f_M\left(\mu; \alpha, \theta\right) = \frac{(1-\theta)^{-\frac{1}{4}} \left[\frac{2}{(\alpha\theta)}\right]^{-\frac{1}{2}} \mu^{-\frac{3}{2}}}{2K_{\frac{1}{2}}(\alpha\sqrt{1-\theta})} \exp\left[\left(1-\frac{1}{\theta}\right)\mu - \frac{\alpha^2\theta}{4\mu}\right] , \qquad (2.83)$$

for $\mu > 0$, $\alpha > 0$, and $0 < \theta < 1$ (Johnson et al., 2005). This gives rise to a Holla distribution with parameters θ and α,

$$f_Y(y; \theta, \alpha) = P\left(Y = y\right) = \sqrt{\frac{2\alpha}{\pi}} \frac{\exp(\alpha\sqrt{(1-\theta)}(\frac{\alpha\theta}{2}))^y}{y!} K_{y-\frac{1}{2}}(\alpha) , \qquad (2.84)$$

for $y = 0, 1, 2, \ldots$, $\alpha > 0$, and $0 < \theta < 1$ (Johnson et al., 2005).

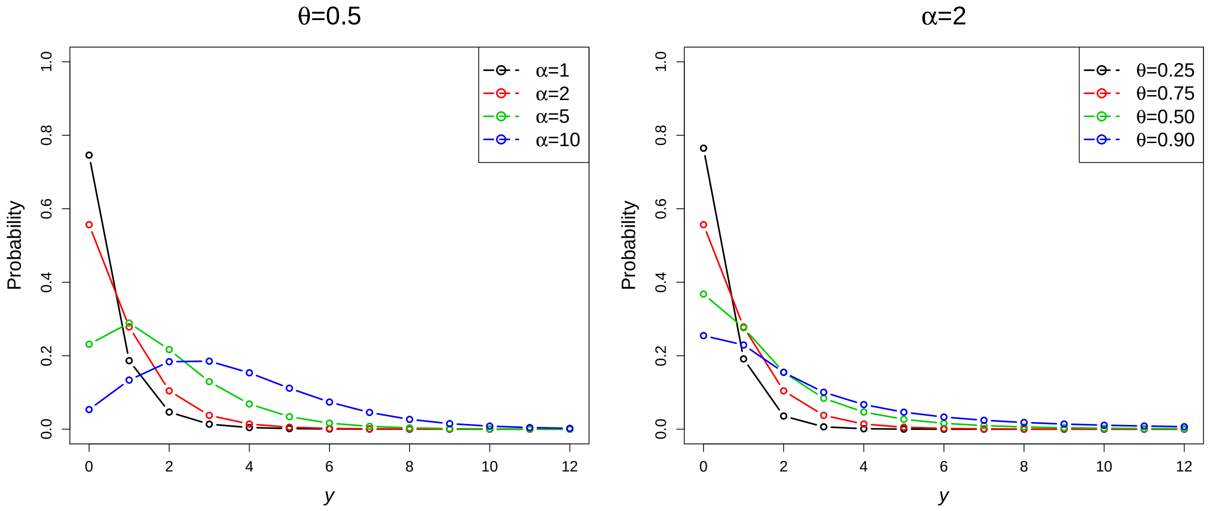


Figure 2.11: Holla *pdf*

Figure 2.11 gives two plots of the Holla density. The first gives a range of values of

$\alpha = 1, 2, 5$ and 10 where θ is fixed at 0.5 and shows that as α increases the distribution becomes less skew and tends to a normal distribution. In the second plot α is fixed at 2 and θ ranges in $0.25, 0.50, 0.75$ and 0.90. As the value of θ decreases the probability of a lower y value increases.

The *pgf* of the Holla distribution is given by,

$$G(t) = \exp\left\{\frac{\alpha}{\theta} - \frac{\alpha}{\theta}\sqrt{1 + \frac{2\theta^2}{\alpha}(1-t)}\right\} \ , \tag{2.85}$$

The *mgf* is,

$$M(t) = \exp\left\{\frac{\alpha}{\theta} - \frac{\alpha}{\theta}\sqrt{1 + \frac{2\theta^2}{\alpha}(1-\mathrm{e}^t)}\right\} \ , \tag{2.86}$$

The mean of the Holla distribution is θ and the variance $\theta + \dfrac{\theta^3}{\alpha}$. The variance of the Holla distribution increases if θ increases or if α decreases. The Holla distribution has overdispersion index $OD = \dfrac{\alpha + \theta^2}{\theta}$. As both α and θ increase the OD index also increases, indicating more dispersion in the distribution and will increase faster with respect to θ. The zero-inflation index is given by $ZI = 1 - \dfrac{\alpha}{\theta}$. As α increases the ZI index approaches $-\infty$ and as θ approaches 0 from 1, the ZI index decreases. The SI is given by,

$$SI_y = \frac{\left(\mathrm{e}^{\frac{1}{2}\alpha^2\sqrt{1-\theta\theta}}\right)^{-y} \sqrt{\frac{\pi}{2}}\, y! \sum_{y=0}^{\infty} \frac{2\left(\mathrm{e}^{\frac{1}{2}\alpha^2\sqrt{1-\theta\theta}}\right)^{2y} \alpha\, K_{y-\frac{1}{2}}(\alpha)^2}{\pi(y!)^2}}{\sqrt{\alpha}\, K_{y-\frac{1}{2}}(\alpha)} \ . \tag{2.87}$$

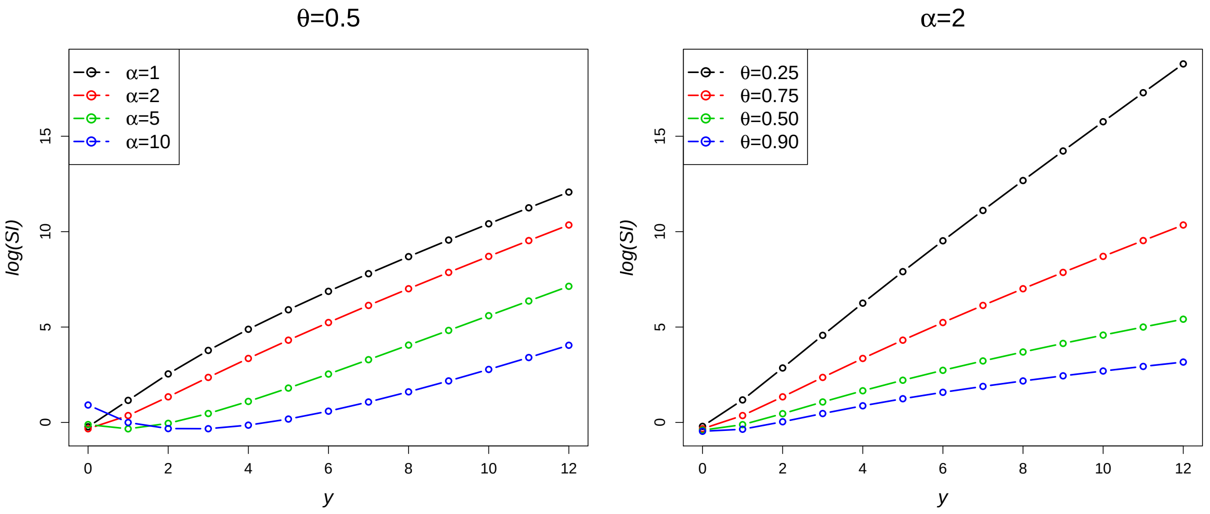


Figure 2.12: log(*SI*)'s for Holla probability distributions

SI's for the Holla distribution are plotted in Figure 2.12 where in the first plot θ is fixed a 0.5 and α has values 1, 2, 5 and 10 and in the second plot α is fixed at 2 and $\theta = 0.25, 0.50, 0.75$ and 0.90. As α increases the SI's become less surprising, whilst as θ decreases from 1 to 0 the SI increases.

2.3.3 Sichel (α, θ, γ)

A more general form of the Holla distribution is the Sichel distribution which was first defined by Sichel (1975) to model word count data. The Sichel distribution can be thought of as a parameter-mixture distribution where $\text{Sichel}(\alpha, \theta, \gamma) = \text{Poisson}(\mu) \bigwedge_{\mu} \text{GIG}(\alpha, \theta, \gamma)$ and is also known as the Poisson-Generalized Inverse Gaussian (GIG) distribution, with parameters *pdf*,

$$f_{\Lambda}(\lambda) = \frac{(1-\theta)^{\frac{\gamma}{2}} \left(\frac{2}{\alpha\,\theta}\right)^{\gamma} \lambda^{\gamma-1}}{2K_{\gamma}(\alpha\sqrt{1-\theta})} \exp\left[\left(1-\frac{1}{\theta}\right)\lambda - \frac{\alpha^2\theta}{a\lambda}\right] , \tag{2.88}$$

for $\lambda > 0$. The Sichel distribution therefore has *pdf*,

$$f_Y\left(y; \alpha, \theta, \gamma\right) = P\left(Y = y\right) = \frac{(1-\theta)^{\frac{\gamma}{2}} \left(\frac{\alpha\theta}{2}\right)^{y}}{y! K_{\gamma}\left(\alpha(1-\theta)^{\frac{1}{2}}\right)} K_{y+\gamma}(\alpha) , \tag{2.89}$$

for $y = 0, 1, 2, \ldots$, $0 < \theta < 0$, $-\infty < \gamma < \infty$, and $\alpha > 0$, where $K_v(\cdot)$ is a modified Bessel function of the third kind (Johnson et al., 2005; Wimmer and Altmann, 1999). When $\gamma = -\dfrac{1}{2}$ the Sichel distribution is equal to the Holla distribution.

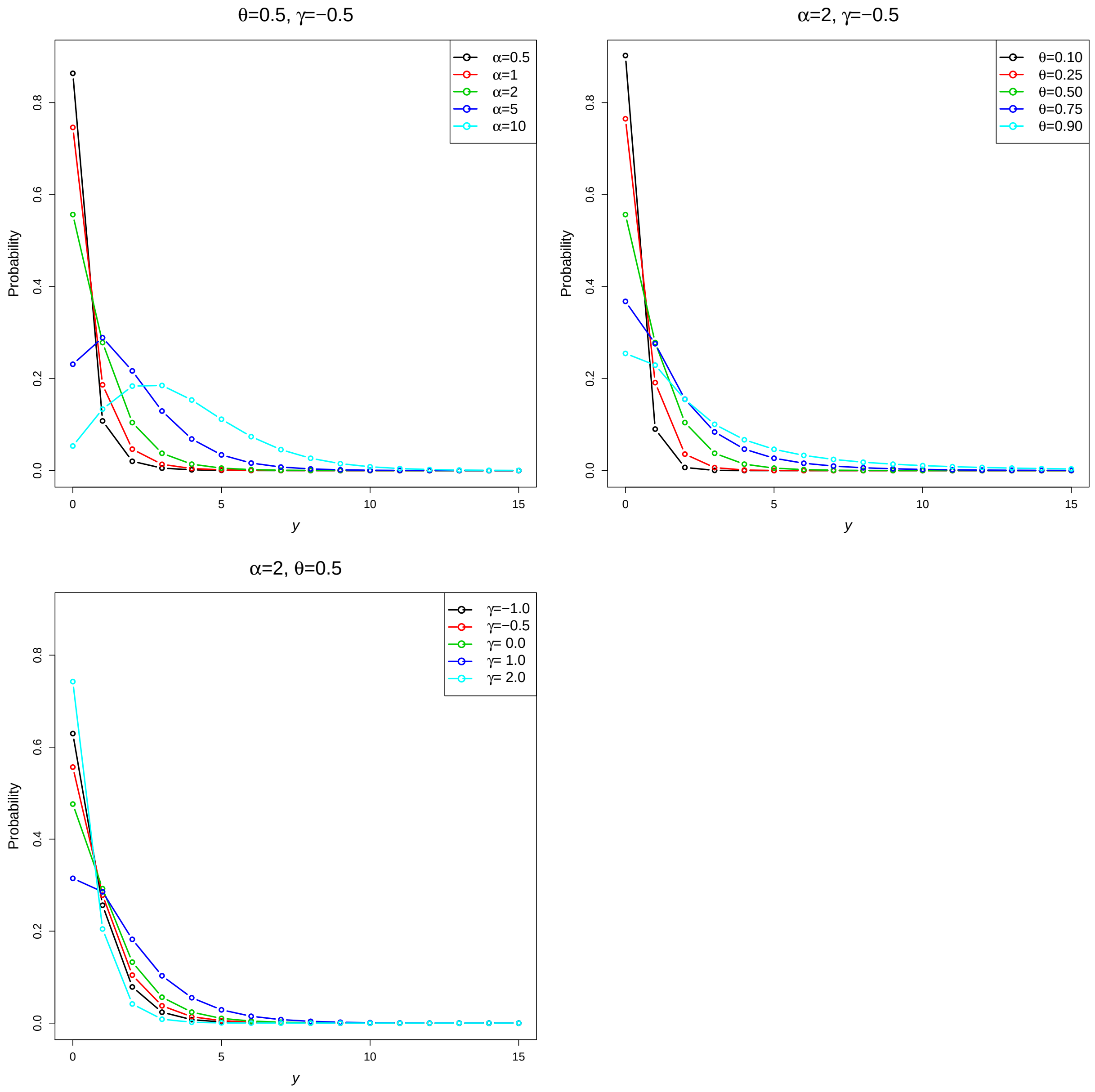


Figure 2.13: Sichel *pdf*

Figure 2.13 shows examples of Sichel *pdf*'s. The first plot shows varying α of 0.5, $1, 2, 5$, and 10 for fixed $\theta = 0.5$ and $\gamma = -0.5$, where α characterises the probability of low values of Y. The second plot shows the distribution for $\theta = 0.10, 0.25, 0.50, 0.75$ and 0.90 where $\alpha = 2$ and $\gamma = -0.5$, illustrating how θ influences the tail of the distribution. In the final plot α is fixed at 2 and $\theta = 0.5$ and $\gamma = -1, -0.5, 0, 1$ and 2 which parametrizes the overall shape of the distribution.

The Sichel distribution has *pgf*,

$$G(t) = \frac{\left(\frac{1-\theta}{1-t\theta}\right)^{\gamma/2} K_\gamma(\alpha\sqrt{1-t\theta})}{K_\gamma(\alpha\sqrt{1-\theta})} , \tag{2.90}$$

and the *mgf* is,

$$M(t) = \frac{\left(\frac{1-\theta}{1-\mathrm{e}^t\theta}\right)^{\gamma/2} K_\gamma(\alpha\sqrt{1-\mathrm{e}^t\theta})}{K_\gamma(\alpha\sqrt{1-\theta})} , \tag{2.91}$$

The mean of the Sichel distribution is,

$$\mu = \frac{\alpha\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})}{2\sqrt{1-\theta}K_\gamma(\alpha\sqrt{1-\theta})} , \tag{2.92}$$

and the variance,

$$\begin{aligned} \sigma^2 = & \tfrac{1}{4(\theta-1)^2}\theta\,(\,4\gamma + 4\gamma^2\theta - \alpha^2(\theta-1)\theta \\ & +(\,\alpha\,(\,2\sqrt{1-\theta}(1+\gamma\theta)K_{\gamma-1}(\alpha\sqrt{1-\theta})K_\gamma(\alpha\sqrt{1-\theta}) \\ & +\alpha(\theta-1)\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})^2\,)\,)\,/K_\gamma(\alpha\sqrt{1-\theta})^2\,) \end{aligned} \;\cdot \tag{2.93}$$

The overdispersion index is

$$OD = \frac{1}{2}\left(-\frac{2(1+\gamma\,\theta)}{\theta-1} + \frac{\alpha\,\theta\left(K_\gamma(\alpha\sqrt{1-\theta})^2 - K_{\gamma+1}(\alpha\sqrt{1-\theta})^2\right)}{\sqrt{1-\theta}K_\gamma(\alpha\sqrt{1-\theta})K_{\gamma+1}(\alpha\sqrt{1-\theta})}\right) , \tag{2.94}$$

If θ is 0, the OD index is equal to one and as θ increases the dispersion in the distribution increases. For large negative values of γ, the OD index approaches 1 and as γ increases the dispersion in the data increases. The zero-inflation index is,

$$ZI = 1 + \frac{2\sqrt{1-\theta}K_{\gamma}(\alpha\sqrt{1-\theta})\log\left(\frac{(1-\theta)^{\frac{\gamma}{2}}K_{\gamma}(\alpha)}{K_{\gamma}(\alpha\sqrt{1-\theta})}\right)}{\alpha\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})}. \qquad (2.95)$$

The ZI is less than 1 across all parameter values, indicating that no zero-inflation is present under this distribution. The SI of the Sichel distribution is,

$$SI_y = \frac{2^y(1-\theta)^{-\gamma/2}(\alpha\theta)^{-y}\mathrm{K}\left(\gamma,\alpha\sqrt{1-\theta}\right)\mathrm{y}!\sum_{\mathrm{y}=0}^{\infty}\frac{4^{-\mathrm{y}}(1-\theta)^{\gamma}(\alpha\theta)^{2\mathrm{y}}\mathrm{K}(\mathrm{y}+\gamma,\alpha)^2}{\mathrm{K}\left(\gamma,\alpha\sqrt{1-\theta}\right)^2(\mathrm{y}!)^2}}{\mathrm{K}(\mathrm{y}+\gamma,\alpha)}, \qquad (2.96)$$

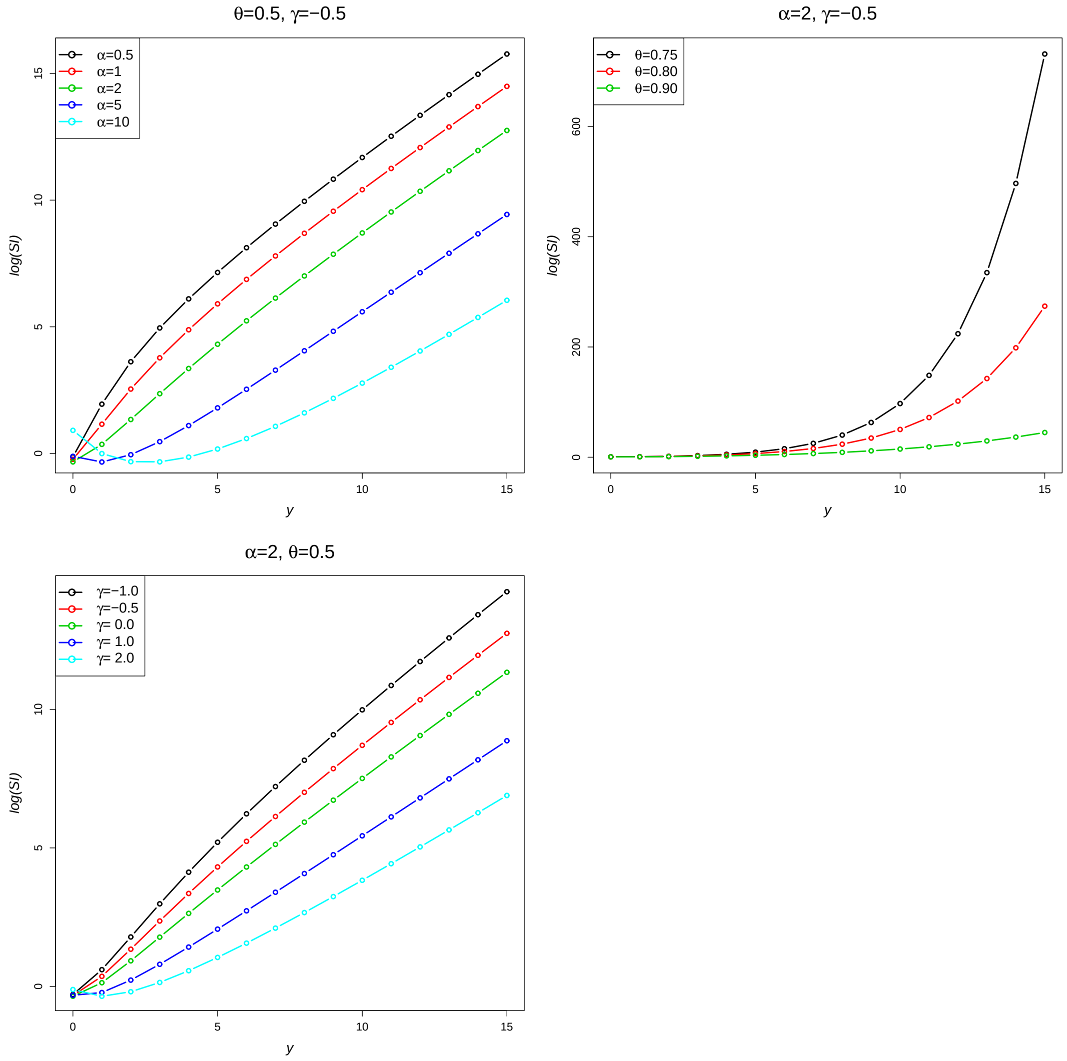


Figure 2.14: log(SI's for Sichel distribution

The logarithm of the SI is plotted for various values of α, θ and γ in Figure 2.14. The first plot has varying values of $\alpha = 0.5, 1, 2, 5$ and 10 where θ and γ are fixed at 0.5 and -0.5 and as α becomes large, the log of the SI increases. In the second plot, $\alpha = 2$ and $\gamma = -0/5$, where θ is in the range $0.75, 0.80$ and 0.90, and indicated that as θ approaches 1 larger values of y become more surprising. Finally, the third plot shows the logarithm of the SI, with α fixed at 2, θ fixed at 0.5 and $\gamma = 1, 0.5, 0, 1$ and 2. As γ increases values of the SI become less surprising.

2.3.4 Delaporte (α, β, γ)

The Delaporte distribution was introduced by Delaporte (1959) for the number of claims in a motor insurance portfolio (Ruohonen, 1988; Willmot, 1989). The number of claims in time Y, can be thought of as the sum of two components, $N_Y = N_{1Y} + N_{2Y}$, where N_{1Y} has a Poisson distribution with expected value γY and N_{2Y} follows a Negative Binomial distribution with parameters r and p. The Delaporte distribution can also be constructed as a parameter mix model of a Poisson and a three-parameter Gamma distribution, with *pdf* given by,

$$f_M(\mu; \alpha, \beta, \gamma) = P(M = \mu) = \frac{\beta^{\alpha}(\mu - \gamma)^{\alpha-1}\mathrm{e}^{-\beta(\mu-\gamma)}}{\Gamma(\alpha)}\ , \tag{2.97}$$

where $\mu > \gamma$, $\alpha \geq 0$ and $\alpha, \beta > 0$ (Wimmer and Altmann, 1999; Ruohonen, 1988). The resulting parameter mix distribution can be written in this notation as

Delaporte(α, β, γ) = Poisson(μ) $\bigwedge_{\mu}$ Gamma(α, β, γ) and has *pdf*,

$$f_Y(y; \alpha, \beta, \gamma) = P(Y = y) = \sum_{j=0}^{n} \frac{\Gamma(j+\alpha)}{\Gamma(\alpha) j!} \left(\frac{\beta}{y+\beta}\right)^{\alpha} \left(\frac{y}{y+\beta}\right)^{j} \frac{(\gamma y)^{n-j}\mathrm{e}^{-\gamma y}}{(n-j)!}\ , \tag{2.98}$$

for $y = 0, 1, 2, \ldots$ (Ruohonen, 1988; Willmot, 1989) and is also known as a Poisson-Negative Binomial convolution distribution as it is also a Poisson distribution generalized by a negative binomial distribution (see Section 2.7) (Wimmer and Altmann, 1999). Stasinopoulos and Rigby (2008) parametrize the Delaporte distribution in terms of the location μ, scale σ and skewness ν parameters,

$$f_Y(y;\mu,\sigma,\nu) = P(Y=y) = \frac{\mathrm{e}^{-\mu\nu}}{\Gamma(\frac{1}{\sigma})}(1+\mu\sigma(1-\nu))^{\frac{-1}{\sigma}} S \ , \tag{2.99}$$

where,

$$S = \sum_{j=0}^{y} \binom{y}{j} \frac{\mu^y \nu^{y-j}}{y!} \left[\mu + \frac{1}{\sigma(1-\nu)}\right]^{-j} \Gamma\left(\frac{1}{\sigma}+j\right) \tag{2.100}$$

for $y = 0, 1, 2 \ldots$ where $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$.

Figure 2.15 shows the Delaporte probability distribution for values of μ=1, 2, 5, 10, σ=1, 5 and ν=0.1, 0.5, 0.9. The mean μ changes the skew of the distribution: as it increases the distribution tends to a normal distribution. The parameter σ characterizes the probabilities of low values of Y with higher values of σ resulting in higher low probabilities and ν affects the overall shape of the distribution: as ν increases the distribution approximates a normal distribution.

This distribution in Equation 2.98 has *pgf*,

$$G(t) = \mathrm{e}^{(t-1)\alpha}(1 - t(1-\gamma))^{-\beta}\gamma^{\beta} \ , \tag{2.101}$$

with *mgf*,

$$M(t) = \mathrm{e}^{\left(\mathrm{e}^{t}-1\right)\alpha}\left(\mathrm{e}^{t}(\gamma-1)+1\right)^{-\beta}\gamma^{\beta} \ . \tag{2.102}$$

The Delaporte distribution has mean,

$$\mu = \frac{\gamma\,(\alpha-\beta)+\beta}{\gamma} \ , \tag{2.103}$$

and variance,

$$\sigma^2 = \frac{\gamma\,(\alpha\,\gamma-\beta)+\beta}{\gamma^2} \ . \tag{2.104}$$

The overdispersion index is,

$$OD = \frac{\alpha\,\gamma - \beta + \frac{\beta}{\gamma}}{\alpha\,\gamma + \beta - \beta\,\gamma} \ , \tag{2.105}$$

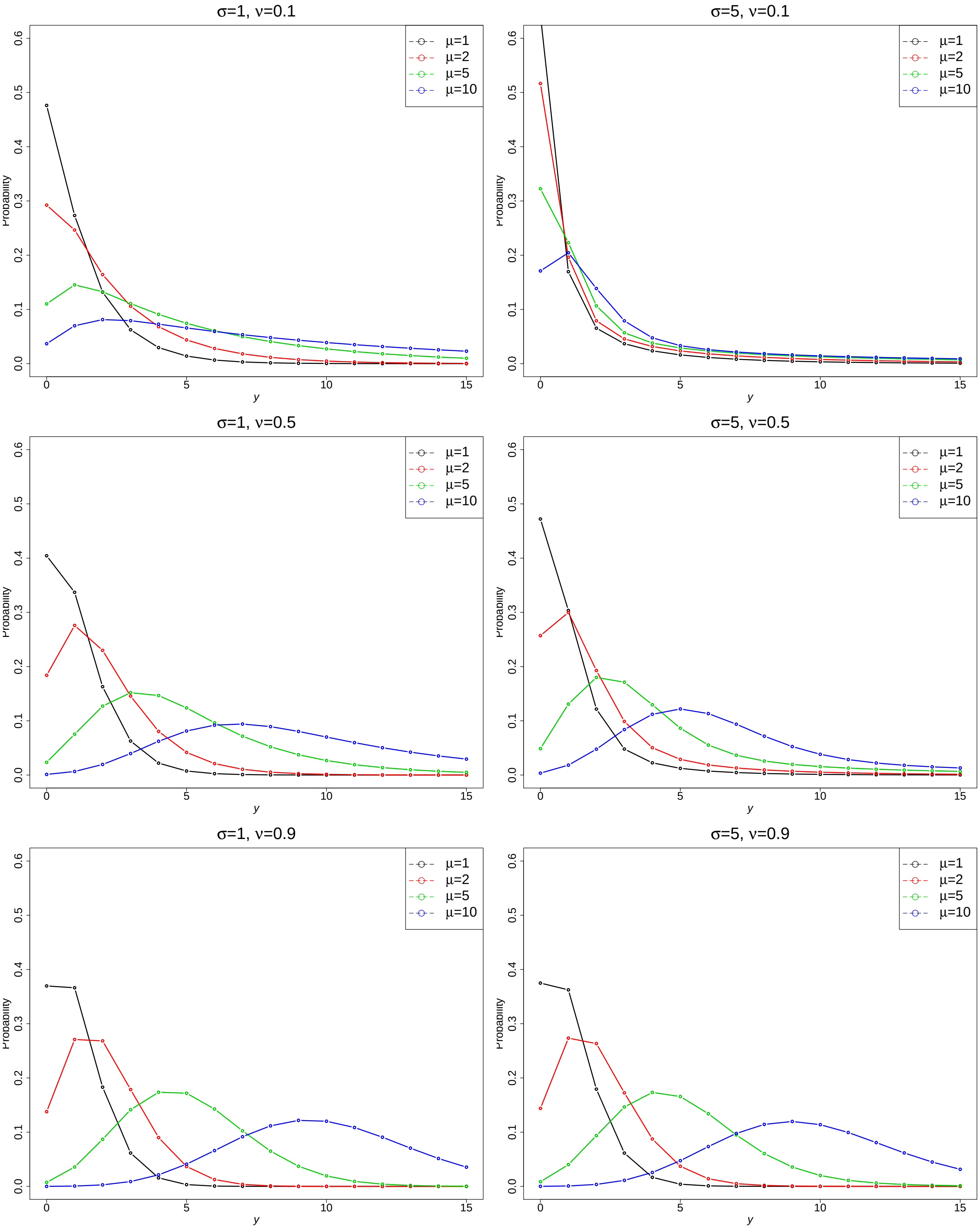


Figure 2.15: Delaporte *pdf*

As α becomes large, the OD index approaches 1 .i.e. the distribution becomes less overdispersed. However, as β and γ increase the OD index also increases. The zero-inflation index is given by

$$ZI = 1 + \frac{\alpha\,\beta\,\gamma\log(\gamma)}{\beta + \alpha\gamma - \beta\gamma}\,. \tag{2.106}$$

where γ is large the ZI index is greater than 1 and for β approaching 0 the ZI index tends to 1. The SI of this distribution is,

$$\begin{aligned} SI_y = \; & \tfrac{1}{U(-n,1-n-\alpha,(y+\beta)\gamma)}\,(\;\mathrm{e}^{y\gamma}y^{-n}\left(\tfrac{\beta}{y+\beta}\right)^{-\alpha}(y+\beta)^n n! \\ & \textstyle\sum_{y=0}^{\infty}\tfrac{1}{n!\Gamma(1+n)}\mathrm{e}^{-2y\gamma}y^n\left(\tfrac{\beta}{y+\beta}\right)^{\alpha}(y+\beta)^{-n}(y\gamma)^n(\beta\gamma)^{\alpha} \\ & U(-n,1-n-\alpha,(y+\beta)\gamma)U(\alpha,1+n+\alpha,(y+\beta)\gamma)\;)\;. \end{aligned} \tag{2.107}$$

where $U(a,b,x)$ is the confluent hypergeometric function.

2.3.5 Yule (λ)

The Yule distribution was originally developed by G.U. Yule (1925) as the limiting case of a distribution in mathematical genetics and was used by Simon (1955) to model word frequencies. This distribution can be constructed as a parameter mix distribution in two ways:

1. Yule(λ) = Geometric$(1-p)\bigwedge\limits_{p}$Beta$(\lambda,1)$
2. Yule(λ) =Geometric$(1-\mathrm{e}^{-a})\bigwedge\limits_{a}$exponential$(\frac{1}{\lambda})$

The Yule distribution is generated as a Geometric-mixture i.e. for each individual the number of failures are counted until the first success with a frailty distribution (Beta or exponential) for p (Wimmer and Altmann, 1999). The *pdf* of the Yule distribution is given by,

$$f_Y(y;\lambda) = P(Y=y) = \frac{\lambda}{y+1}\binom{\lambda+y+1}{\lambda+1}^{-1} = \frac{\lambda\,y!}{(\lambda+1)^{(y+1)}}\,, \tag{2.108}$$

for $\lambda > 0$ and $y = 0, 1, 2, \ldots$ (Wimmer and Altmann, 1999). Alternatively, the *pdf* can be written as,

$$f_Y(y;\lambda) = P(Y = y) = \frac{B(\lambda+1, y+1)}{B(\lambda, 1)} \; , \tag{2.109}$$

where B is the beta function. The Yule *pdf* is highly skewed, as shown in Figure 2.16. As the value of λ increases the probability of a value of zero increases, however this difference decreases as the value of y increases.

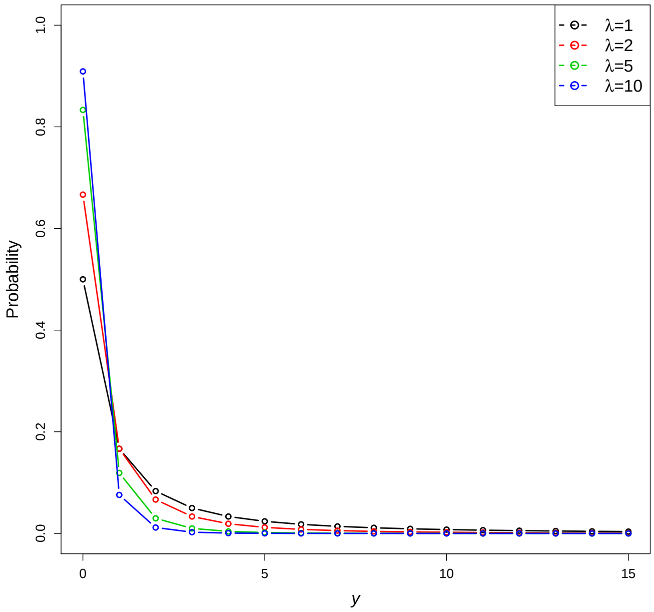


Figure 2.16: Yule *pdf*

The Yule distribution has *pgf*,

$$G(t) = \lambda\,\Gamma(1+\lambda)\,{}_2F_1(1, 1; 2+\lambda, t) \; , \tag{2.110}$$

and *mgf*,

$$M(t) = \lambda\Gamma(1+\lambda)\,{}_2F_1(1, 1; 2+\lambda; \mathrm{e}^t) \; , \tag{2.111}$$

The mean and variance of the Yule distribution are,

$$\mu = \tfrac{1}{\lambda-1} \quad \text{and} \quad \sigma^2 = -\tfrac{1}{(\lambda-1)^2} + \lambda B(1+\lambda, 2)_3F_2(2, 2, 2; 1, 3+\lambda; 1) \; . \tag{2.112}$$

The overdispersion index is given by,

$$OD = \frac{1}{1-\lambda} + (\lambda-1)\lambda\, B(\lambda+1,2)\, {}_3F_2(2,2,2;1,3+\lambda;1) \tag{2.113}$$

For small values of λ, the OD index is large, however as λ increases the dispersion in the distribution is reduced. The zero-inflation index for the Yule distribution is,

$$ZI = 1 + (\lambda-1)\log\left(\frac{\lambda}{\lambda+1}\right) . \tag{2.114}$$

Where $\lambda = 1$ the ZI index is 1, indicating that zero-inflation is present in the distribution and as λ increases the ZI index decreases. The SI of the Yule distribution is,

$$SI_y = \frac{\lambda^4 B(\lambda,1)\Gamma(\lambda)^2\, {}_3F_2(1,1,1;2+\lambda,2+\lambda;1)}{B(1+\lambda,1+y)} . \tag{2.115}$$

The logarithm of the SI for the Yule distribution with values of λ of 1, 2, 5 and 10 is shown in Figure 2.17. As λ increases, values of Y become more surprising.

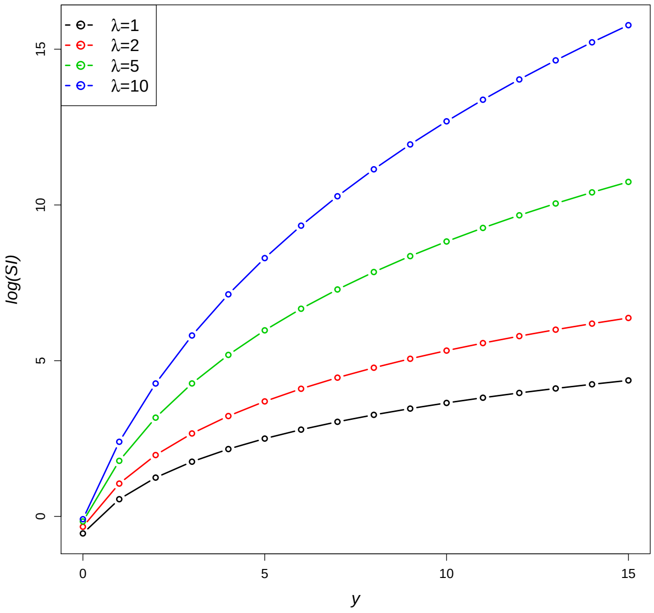


Figure 2.17: SI's for the Yule distribution

2.3.6 Waring (b, n)

The Waring distribution was developed by Irwin (1963) to describe biological distributions with very long tails. The Waring distribution can be generated as a parameter mixture of Geometric or negative binomial distributions with beta or exponential mixing distributions as follows,

1. Waring(b, n) =Geometric$(p) \bigwedge_p$ Beta(b, n)
2. Waring(b, n) =negative binomial $(n, p) \bigwedge_p$ Beta $(b, 1)$
3. Waring(b, n) =negative binomial $(n, e^{-p}) \bigwedge_p$ exponential $(1/b)$

The distribution has *pdf*,

$$f_Y(y; b, n) = P(Y = y) = \frac{B(n + y, b + 1)}{B(n, b)} \ , \tag{2.116}$$

for $y = 0, 1, 2, \ldots$ where $b > 0$ and $n \geq 0$ (Wimmer and Altmann, 1999, P. 643). The Waring distribution is equal to a Yule distribution when $n \to 1$.

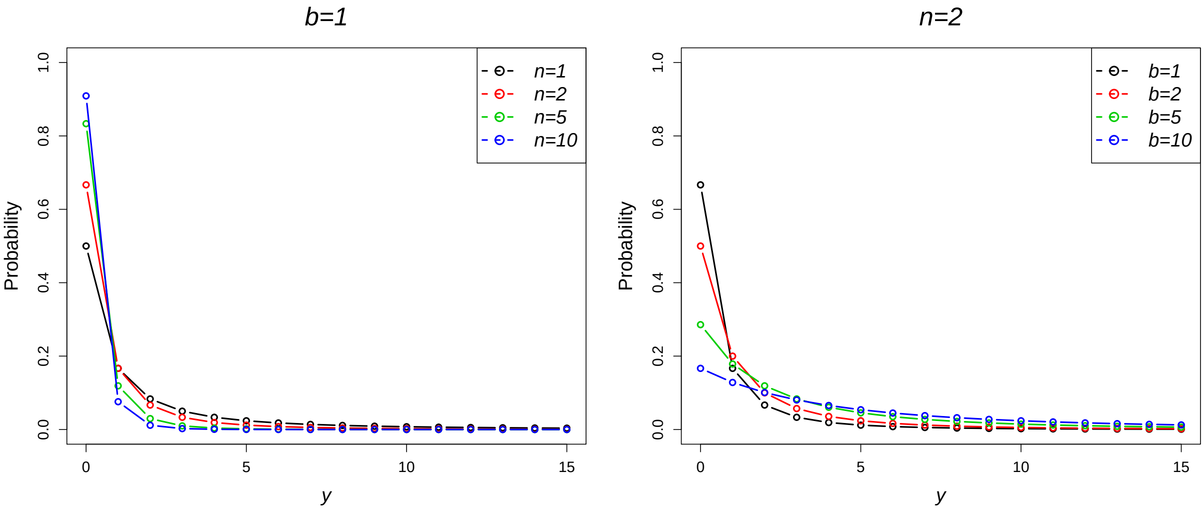


Figure 2.18: Waring *pdf*

Figure 2.18 shows the *pdf* of the Waring distribution for values of n of 1,2,5 and 10 when b is fixed at 1 and also values of b of 1,2,5 and 10 when n is fixed at 2. As

n increases the probability of low values of y decreases and the resulting distribution becomes flatter. As b increases the probability of a low value of y decreases and the distribution becomes les J-shaped.

The *pgf* of the Waring distribution is,

$$G(t) = b\,\Gamma(b+n)_2F_1(1, n; b+n+1; t)\ . \tag{2.117}$$

and the *mgf*,

$$M(t) = b\,\Gamma(b+n)\,_2F_1(1, n; b+n+1; \mathrm{e}^t)\ , \tag{2.118}$$

The Waring distribution has mean,

$$\mu = \frac{n}{b-1}\ , \tag{2.119}$$

and variance,

$$\sigma^2 = \frac{B(1+n, 1+b)_3F_2(2, 2, 1+n; 1, 2+b+n; 1)}{B(n, b)} - \frac{n^2}{(b-1)^2}\ . \tag{2.120}$$

The overdispersion index is given by,

$$OD = \frac{(b-1)\,b\,(b+2n)\,\Gamma(b+n)\,_2F_1(2, 1+n; 2+b+n; 1)}{b-2} - \frac{n}{b-1}\ , \tag{2.121}$$

where b and n are small the OD index is large and as b and n decrease, the dispersion in the distribution decreases. The zero-inflation index is,

$$ZI = 1 + \frac{(b-1)\text{log}\left(\frac{b}{b+n}\right)}{n}\ . \tag{2.122}$$

The ZI index is equal to 1 when $b = 1$ and as b increases the ZI index decreases. For the parameter n, as n increases the ZI index also increases. The SI of this distribution is,

$$SI_y = \frac{b^2\,B(n, b)\,_3F_2(1, n, n; b+n+1, b+n+1; 1)}{(b+n)^2\,B(n+y, b+1)}\ . \tag{2.123}$$

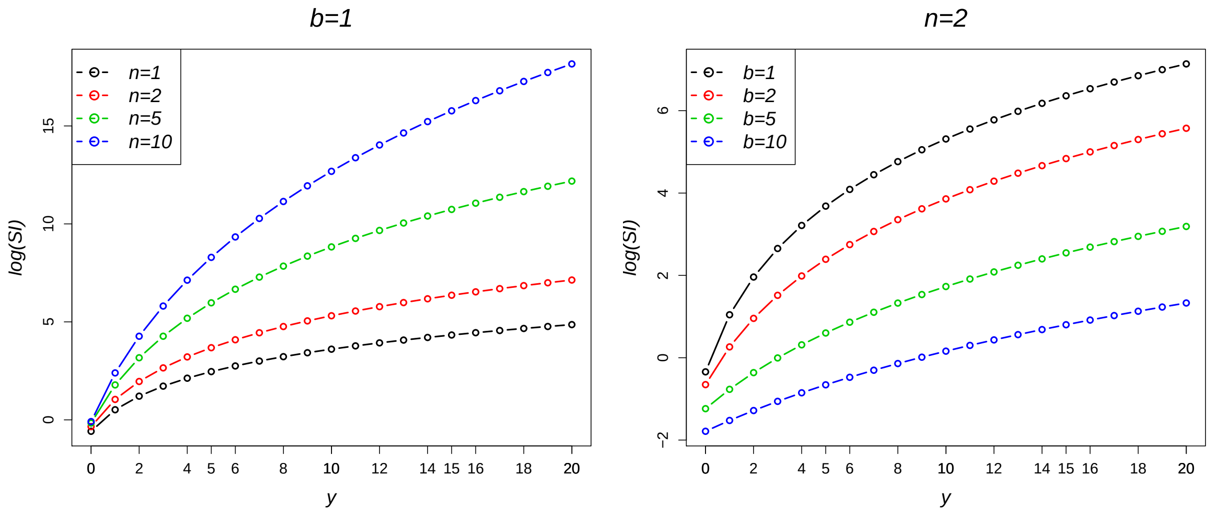


Figure 2.19: log(*SI*)'s for Waring distributions

Figure 2.19 plots the logarithm of SI for the Waring distribution. In the first plot, b is fixed at 1, and n is in the range 1, 2, 5 and 10, with larger values of n resulting in higher SI's for large values of Y. In the second plot, n is fixed at 2, with b values of 1, 2, 5 and 10. For higher values of b the SI is smaller.

2.3.7 Beta-Binomial (a, b, n)

The Beta-Binomial distribution is also known as a contagious binomial, hyperbinomial, hypergeometric waiting time or inverse hypergeometric distribution. The distribution is used to model variation in the number of defective items per lot in inspection sampling (Johnson et al., 2005). Examples of the distribution's application are also found in biology where it is used to estimate population sizes.

There are two different ways of obtaining this distribution. The first is as an urn model and the second is through a parameter mixture of distributions.

The Beta-Binomial distribution can be considered an *Urn Model* arising from random draws from an urn containing a white balls and b black balls. It can be defined by drawing a random ball; if it is a white ball then two white balls are returned to the urn, if a black ball is drawn two black balls are returned to the urn. This is repeated n times and the probability of observing y white balls lies in the range $R_Y = \{0, 1, \ldots, n\}$ and follows a beta-binomial distribution with parameters $a > 0$, $b > 0$ and $n > 0$ (Johnson et al., 2005).

Alternatively, the Beta-Binomial distribution can be constructed as a parameter-mix of a Beta and a Binomial distribution, where in a Binomial distribution with parameters n and p, the latter varies according to a Beta distribution with parameters a and b, i.e. Beta-Binomial(a, b, n) =Binomial$(n, p) \bigwedge_{p}$ Beta(a, b), resulting in the following *pdf*,

$$f_Y(y; a, b, n) = P(Y = y) = \frac{\binom{n}{y} \Gamma(b + n - y)\Gamma(a + y)}{B(a, b)\Gamma(a + b + n)} \tag{2.124}$$

where $a > 0$, $b > 0$ and $n > 0$ (Johnson et al., 2005; Wimmer and Altmann, 1999). When n is 1 the distribution is a Bernoulli distribution and for large values of both a and b tends to a normal distribution. The Beta-Binomial distribution appears often in Bayesian statistics as the predictive distribution of a Binomial with a Beta prior on the success probability.

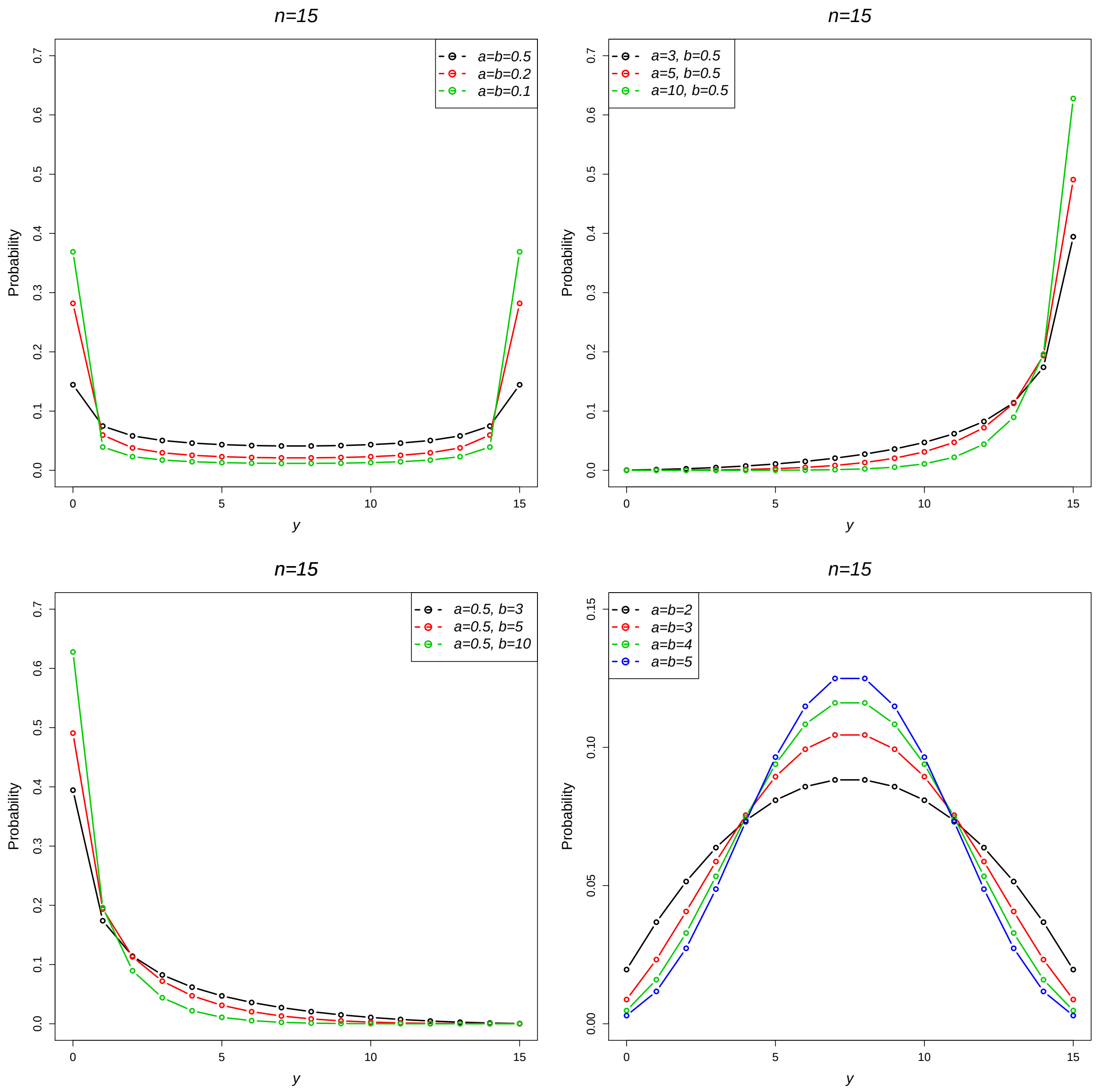


Figure 2.20: Beta-Binomial *pdf*

Figure 2.20 plots the probability density of the Beta binomial distribution where n is fixed at 15. In the first plot the distribution is 'U'-shaped where a and b are both equal have small values i.e. 0.1, 0.2 and 0.5. In the second plot, b is fixed at 0.5 and a has values 3, 5 and 10. These densities are highly right skewed and as a increases the probabilities of y values near $n = 15$ increase. Similarly, the third plot shows the densities for three values of b of 3, 5 and 10 when a is fixed at 0.5. These densities are left skewed and as b increases the probability of a low y value increases. When a and

b are both equal and have large values e.g. 3,4 or 5 the distribution tends to a normal distribution.

The *pgf* for the Beta-Binomial distribution is,

$$G(t) =_2 F_1(-n, a; a+b; t) \ , \tag{2.125}$$

and the *mgf*,

$$M(t) =_2 F_1(-n, a; a+b; \mathrm{e}^t) \ . \tag{2.126}$$

The mean and variance of the distribution are therefore given by

$$\mu = \frac{a\,n}{a+b} \ , \tag{2.127}$$

and

$$\sigma^2 = \frac{a\,b\,n(a+b+n)}{(a+b)^2(1+a+b)} \ , \tag{2.128}$$

with overdispersion index

$$OD = \frac{b(a+b+n)}{(a+b)(a+b+1)} \ . \tag{2.129}$$

For small values of a or b the OD index is 0, but as either of these parameters become large the OD index increases. As the parameter n tends to infinity, the overdispersion index also approaches infinity, as $\sigma^2 \to \infty$ faster than $\mu \to \infty$. The zero-inflation index is given by,

$$ZI = 1 + \frac{(a+b)\log\left(\frac{\Gamma(a)\Gamma(b+n)}{B(a,b)\Gamma(a+b+n)}\right)}{a\,n} \ . \tag{2.130}$$

The ZI index is close to 1 for small values of both a and b, but decreases as either parameter a or b become larger. Again, as n increases the ZI also increases but is always lower than 1 indicating no zero-inflation is present in the distribution. The SI of the Beta-Binomial distribution is,

$$SI_y = \frac{\Gamma(a)^2\Gamma(b+n)^2\,{}_4F_3(a, a, -n, -n; 1, 1-b-n, 1-b-n; 1)}{B(a,b)\binom{n}{y}\Gamma(a+b+n)\Gamma(b+n-y)\Gamma(a+y)} \ . \tag{2.131}$$

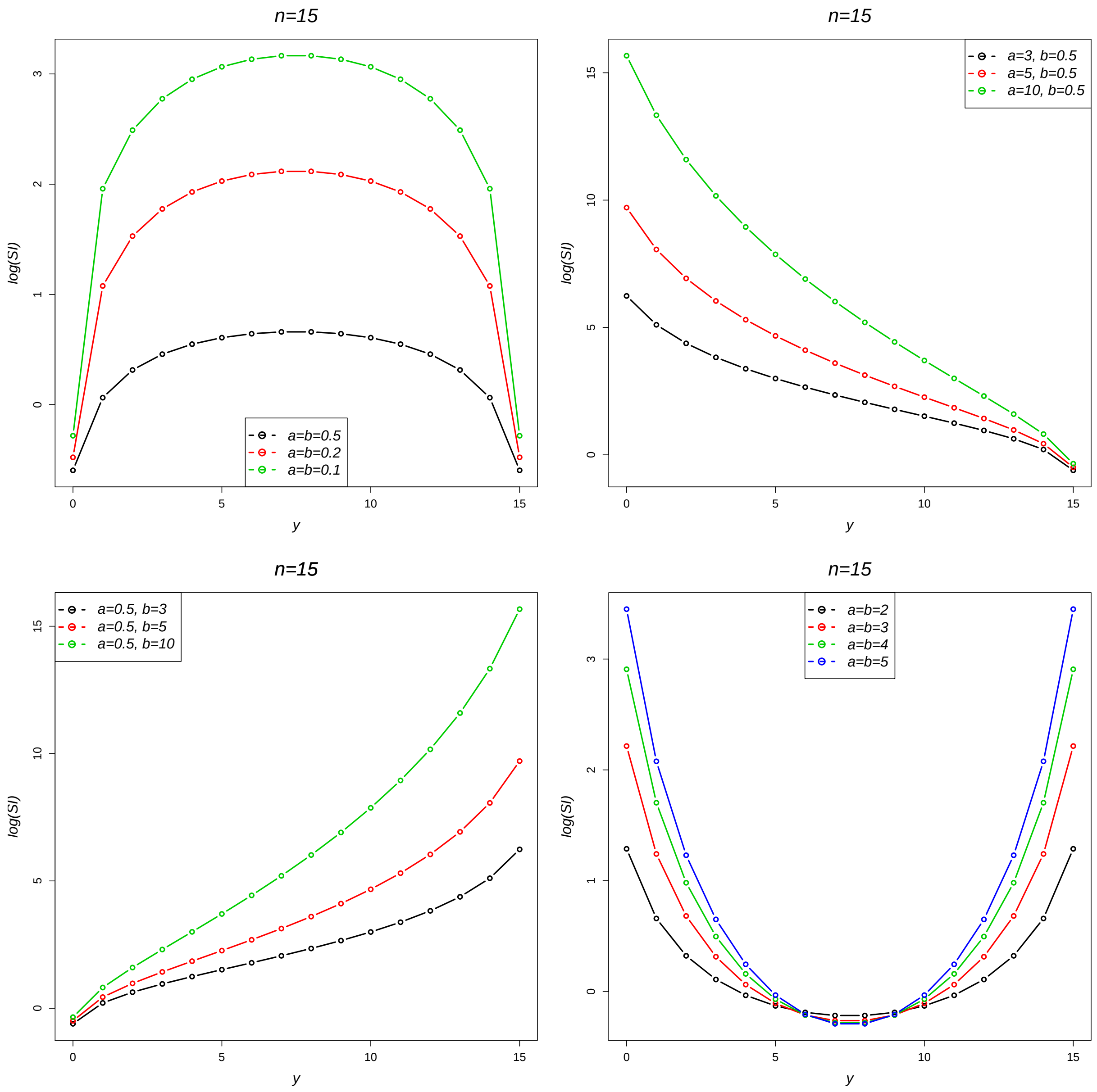


Figure 2.21: log(*SI*)'s for the Beta-Binomial distributions

The $\log(SI)$ is plotted in Figure 2.21 for various values of a and b. Where a and b are both small (first plot) for smaller parameter values very low and very high values of y are less surprising. In the second and third plots as a and b increase, high and low values of y become more surprising, respectively. Finally, the last plot indicates where the parameter values are both equally large, low or high values of y become surprising.

2.4 Component-Mix Distributions

The concept of a component mix of distributions has a long history (Pearson, 1915). This method forms distributions from linear combinations of other distributions (Rose and Smith, 2002; Johnson et al., 2005). For k different component distributions with *pdf*'s $f_1(y), f_2(y), \ldots, f_k(y)$ and mixing weights $\omega_1, \omega_2, \ldots, \omega_k$ where $\omega_j > 0$ and $\sum_{j=1}^{k} \omega_j = 1$, a k-component mixing distribution is defined by taking the weighted average of the f_j's,

$$f_Y(y) = \sum_{j=1}^{k} \omega_j f_j(y) \ , \tag{2.132}$$

(Johnson et al., 2005; Rose and Smith, 2002). This relationship can be written symbolically as $f_A * f_B$ for a component mixture between two distributions f_A and f_B.

Zero-inflated distributions can be formed from a component mix of two distributions and are a special case of Equation 2.132. They allow for zero-inflated data and involve a mix of two distributions where the zeros are modelled separately from the counts,

$$f_Y(y; \theta, \omega) = \begin{cases} \mathrm{P}\,(Y = 0) = \omega + (1 - \omega)\, p_0 & \\ \mathrm{P}\,(Y = j) = (1 - \omega)\, p_j & \text{for}\, j > 0 \end{cases} , \tag{2.133}$$

where ω represents the mixing probability, θ is the vector of parameters of the mixing distribution, p_0 is the probability distribution for the zero counts, and p_j is the probability distribution for the non-zero counts of observations where $j \geq 1$ (Johnson et al., 2005). This type of component mixture is also known as a zero-modified distribution or a distribution with an excess of zeros.

A Poisson component mixture arises through the weighted average of k Poisson distributions, resulting in the *pdf*,

$$f_Y(y; \mu_1, \ldots, \mu_j, \omega_1, \ldots, \omega_j) = P(Y = y) = \sum_{j=1}^{k} \omega_j \frac{\mathrm{e}^{-\mu_j} (\mu_j)^y}{y!} \ , \tag{2.134}$$

where $y = 0, 1, 2, \ldots$ for k components, $\omega_j \neq 0$, $\sum \omega_j = 1$ and $j = 1, 2, \ldots, k$ (Johnson et al., 2005; Karlis and Xekalaki, 1999). This mixture of distributions was studied initially by Feller (1943).

Bimodal distributions can be created through mixtures of two distributions. This section presents three zero-inflated distributions: zero-inflated Poisson, zero-inflated Negative Binomial and zero-inflated Sichel distributions and two bimodal distributions: a two component Poisson-mix and a Poisson-Negative Binomial mix.

2.4.1 Zero-inflated Poisson (ω, μ)

The simplest component mixture distribution is that of a two-component Binomial-Poisson mixture, where the probability of an observation with value zero follows a Binomial distribution and counts of observations greater than or equal to zero follow a Poisson distribution i.e. ZIP(ω, μ) =Bernoulli(ω) * Poisson(μ). The zero-inflated Poisson distribution is the most common zero-inflated distribution within the statistical literature and has the following *pdf*,

$$f_Y(y;\omega,\mu) = P(Y=y) = \begin{cases} P\left(Y=0\right) = \omega + (1-\omega)e^{-\mu} & \\ P\left(Y=j\right) = (1-\omega)\dfrac{e^{-\mu}\mu^{y}}{y!} & \text{for } j > 0 \end{cases}, \tag{2.135}$$

for $y = 0, 1, 2, \ldots$, where $j > 1$ are the non-zero counts, μ is the mean of the Poisson distribution and $0 \leq \omega \leq 1$ is the mixing probability (Ridout et al., 2001; Rose and Smith, 2002; Wimmer and Altmann, 1999; Morgan et al., 2007). This distribution is also sometimes known as the Poisson-with-zeroes or zero-modified Poisson distribution.

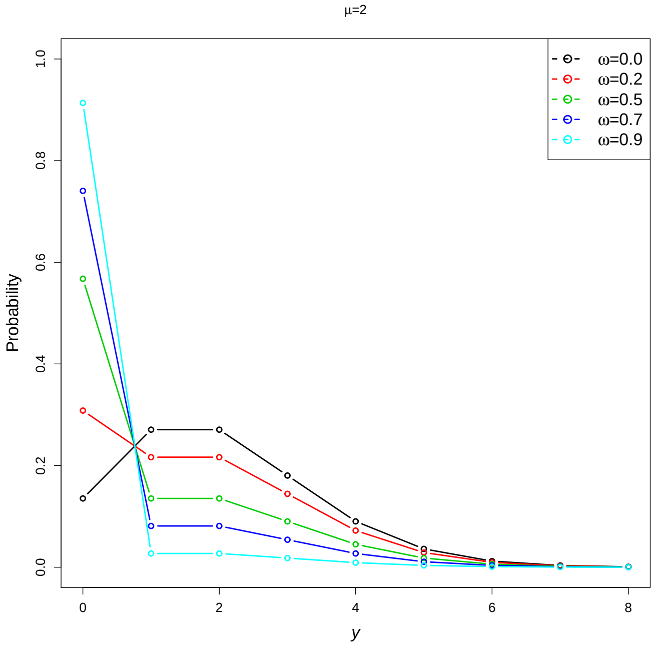


Figure 2.22: Zero-inflated Poisson *pdf*

Figure 2.22 shows how the mixing probability ω affects the probability of y for a zero-inflation Poisson distribution with a mean of 2. As ω increases the probability of a y value of zero increases.

The ZIP distribution has *pgf*,

$$G(t) = \mathrm{e}^{(t-1)\mu}(1-\omega) + \omega \ , \tag{2.136}$$

and *mgf*,

$$M(t) = \mathrm{e}^{(\mathrm{e}^{t}-1)\mu}(1-\omega) + \omega \ . \tag{2.137}$$

The mean of the ZIP is,

$$\mu = \mu(1-\omega) \ , \tag{2.138}$$

and the variance,

$$\sigma^2 = \mu(1-\omega)(1+\mu\,\omega) \ . \tag{2.139}$$

The overdispersion index is given by $OD = 1 + \mu\,\omega$. As ω tends to 0 the OD index becomes close to 1 and as μ increases the OD index also increases. The zero-inflation index for the zero-inflated Poisson distribution is $ZI = 1$ and is therefore always zero-inflated. The ZIP distribution has SI,

$$SI_y = \frac{\mathrm{e}^{-2\mu}\left(\mathrm{e}^{\mu}\omega\left(2+\left(-2+\mathrm{e}^{\mu}\right)\omega\right)+(\omega-1)^2\mathrm{I}_0(2\mu)\right)}{\frac{\mathrm{e}^{-\mu}\mu^{y}(1-\omega)}{y!}+\omega\mathrm{U}_{-y}} \tag{2.140}$$

where U_y is the unitstep function and I_0 is the Bessel function of the first kind.

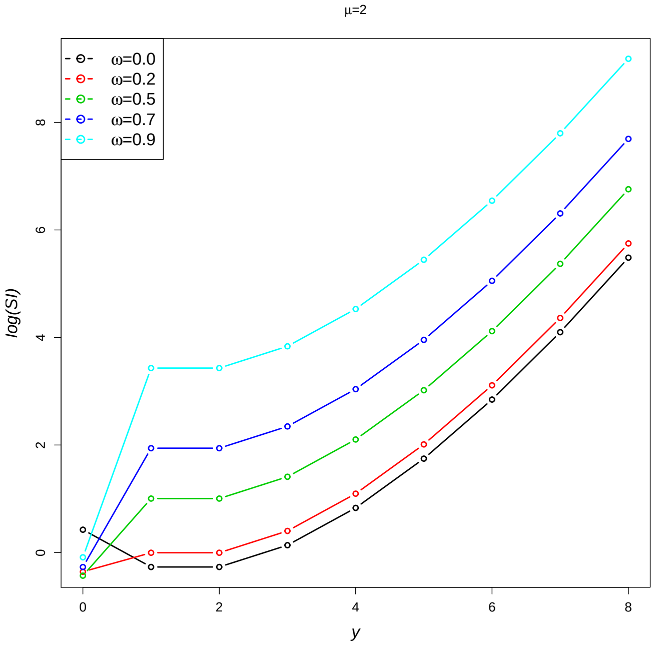


Figure 2.23: log(*SI*)'s for Zero-inflated Poisson distributions

Figure 2.23 plots the SI for the ZIP distribution where $\mu = 2$ and ω is in the range 0, 0.2, 0.5, 0.7 and 0.9. As ω increases to 1, large values of Y become increasingly more surprising, however zero values have low SI values.

2.4.2 Zero-inflated Negative Binomial (ω, p, r)

This distribution is generated as a two-component mixture of a Binomial and Negative Binomial distribution, i.e. ZINB(ω, p, r) =Bernoulli(ω) * NB(p, r). The zero-inflated negative binomial distribution (ZINB) has *pdf*

$$f_Y(y;p,r,\omega) = P(Y=y) = \begin{cases} P\left(Y=0\right) = \omega + (1-\omega)p^r \\ P\left(Y=j\right) = (1-\omega)\binom{y+r-1,y}{p}{}^{r}(1-p)^y \end{cases}, \tag{2.141}$$

where $y = 0, 1, 2, \ldots$, $j > 1$, $r > 0$, $0 < p < 1$ and $0 \leq \omega \leq 1$ (Johnson et al., 2005; Wimmer and Altmann, 1999; Yau et al., 2003).

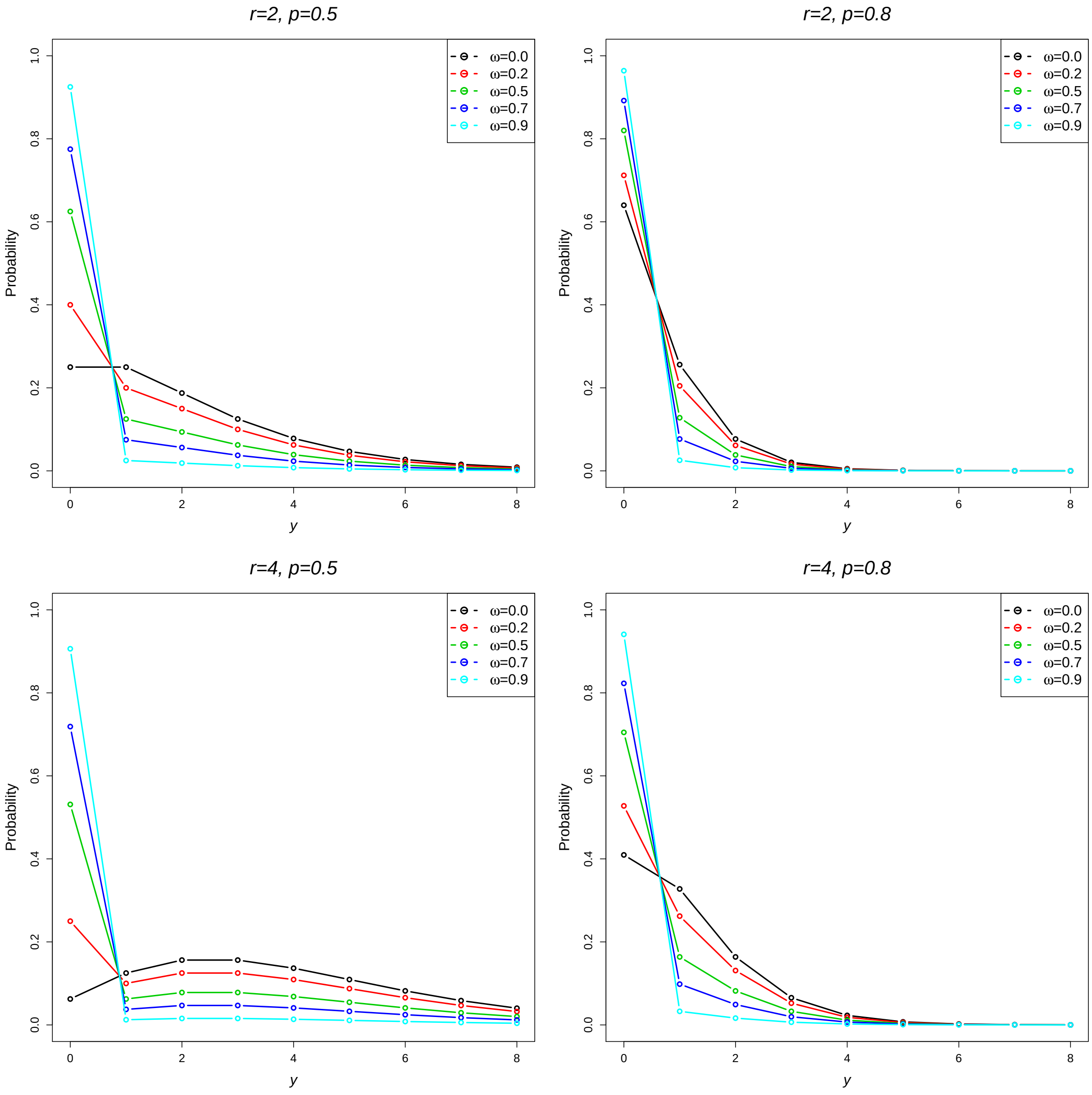


Figure 2.24: Zero-inflated Negative Binomial *pdf*

Figure 2.24 show the *pgf* for values of $\omega = 0, 0.2, 0.5, 0.7, 0.9$ with fixed values of r of 2 and 5 (rows) and p of 0.5 and 0.8 (columns). Again, as ω the zero-inflation parameter index increases, $P(Y = 0)$ increases.

The *pgf* of the ZINB distribution is,

$$G(t) = -p^r(1 + (p-1)t)^{-r}(\omega - 1) + \omega \ . \tag{2.142}$$

The *mgf* is given by,

$$M(t) = -\left(1 + \mathrm{e}^t(p-1)\right)^{-r} p^r(\omega - 1) + \omega \ , \tag{2.143}$$

The mean and variance of the ZINB distribution are given by,

$$\mu = \frac{r}{p}(1-p)(1-\omega) \quad \text{and} \quad \sigma^2 = \frac{r}{p^2}(1-p)(1-\omega)(1-(1-p)r\omega) \ \ . \tag{2.144}$$

The overdispersion index is,

$$OD = \frac{1 + r\,\omega - p\,r\,\omega)}{p} \ , \tag{2.145}$$

As p increases from 0 to 1 the OD approaches 1, indicating overdispersion is present and as either r or ω increase the OD index increases. Under a ZINB distribution the zero-inflation index is again always greater than or equal to 1 demonstrating the zero-inflation present under this distribution. This distribution has SI,

$$SI_y = \frac{\omega\left(-2p^r(\omega-1)+\omega\right) + p^{2r}(\omega-1)^2\,{}_2F_1(r,r;1;(p-1)^2)}{(1-p)^y p^r(1-\omega)\binom{y+r-1}{r-1} \ + \omega \mathrm{U}_{-y}} \ . \tag{2.146}$$

An alternative parameterization of the ZINB distribution is given by Ridout et al. (2001) for both types of the negative binomial distribution, with parameters μ the mean and α the dispersion parameter, with the *pdf* as follows,

$$f_Y(y;\mu,\alpha,c,\omega) = P(Y=y) = \begin{cases} P\left(Y=0\right) = & \omega + (1-\omega)(1+\alpha\mu^c)^{-\mu^{\frac{1-c}{\alpha}}} \\ \\ P\left(Y=j\right) = & (1-\omega)\frac{\Gamma\left(\frac{y+\mu^{1-c}}{\alpha}\right)}{y!\Gamma\left(\frac{\mu^{1-c}}{\alpha}\right)} \\ & (1+\alpha\mu^c)^{-\mu^{\frac{1-c}{\alpha}}}\left(\frac{1+\mu^{-c}}{\alpha}\right)^y \end{cases}, \tag{2.147}$$

for $\mu > 0$, $\alpha \geq 0$ and $0 \leq \omega \leq 1$ for $y = 0, 1, 2, \ldots$ (Ridout et al., 2001). The index c denotes the particular form of the negative binomial distribution: when $c = 1$ a NB I distribution is derived and when $c = 0$ a NB II distribution is formed.

2.4.3 Zero-inflated Sichel $(\omega, \alpha, \theta, \gamma)$

A two-component mix of a Binomial and Sichel distribution results in a zero-inflated Sichel (ZISI) distribution and can be written as ZISI$(\omega, \alpha, \theta, \gamma)$ =Binomial(n, p) * Sichel(α, θ, γ). This distribution has *pdf*,

$$f_Y\left(y;\omega,\alpha,\theta,\gamma\right) = P\left(Y=y\right) = \begin{cases} P\left(Y=0\right) = \omega + (1-\omega)\frac{(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)}{K_\gamma(\alpha\sqrt{1-\theta})} \\ \\ P\left(Y=j\right) = (1-\omega)\frac{(1-\theta)^{\frac{\gamma}{2}}\left(\frac{\alpha\theta}{2}\right)^y}{y!K_\gamma(\alpha(1-\theta)^{\frac{1}{2}})}K_{y+\gamma}(\alpha) \end{cases}, \tag{2.148}$$

for $y = 0, 1, 2, \ldots$, $0 < \theta < 1$, $-\infty < \gamma < \infty$, $\alpha > 0$, and $j > 0$ are the non-zero counts.

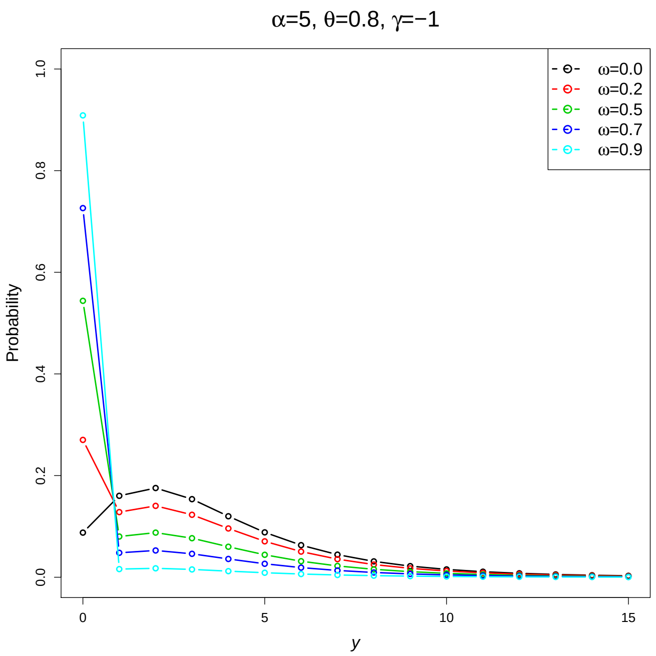


Figure 2.25: Zero-inflated Sichel *pdf*

Figure 2.25 shows the ZISI distribution where the parameters of the Sichel distribution are fixed at $\alpha = 5$, $\theta = 0.8$ and $\gamma = -1$ and ω the zero-inflation parameter ranges in 0.2, 0.5, 0.7 and 0.9. As ω increases the probability of a y value of zero increases.

The *pgf* is given by,

$$G(t) = \omega - \frac{\left(\frac{\theta-1}{t\theta-1}\right)^{\frac{\gamma}{2}}(\omega-1)K_{\gamma}(\alpha\sqrt{1-t\theta})}{K_{\gamma}(\alpha\sqrt{1-\theta})}\ , \tag{2.149}$$

The *mgf* is given by,

$$M(t) = \omega - \frac{\left(\frac{\theta-1}{\mathrm{e}^{t}\theta-1}\right)^{\frac{\gamma}{2}}(\omega-1)K_{\gamma}(\alpha\sqrt{1-\mathrm{e}^{t}\theta})}{K_{\gamma}(\alpha\sqrt{1-\theta})}\ . \tag{2.150}$$

The ZI Sichel distribution has mean,

$$\mu = \frac{\alpha\theta(\omega-1)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{2\sqrt{1-\theta}K_{\gamma}(\alpha\sqrt{1-\theta})}\ , \tag{2.151}$$

and variance,

$$\begin{aligned}\sigma^2 = \quad & \tfrac{1}{4(\theta-1)^2}\theta(\omega-1)\,(\,-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta \\ & +\tfrac{1}{K_\gamma(\alpha\sqrt{1-\theta})^2}\alpha\left(-2\sqrt{1-\theta}(1+\gamma\theta)K_{\gamma-1}(\alpha\sqrt{1-\theta})K_\gamma(\alpha\sqrt{1-\theta})\right)\,)\ .\end{aligned} \tag{2.152}$$

The overdispersion index is,

$$OD = \frac{1}{2}\left(\frac{\alpha\theta\left(K_\gamma(\alpha\sqrt{1-\theta})^2+(\omega-1)K_{\gamma+1}(\alpha\sqrt{1-\theta})^2\right)}{\sqrt{1-\theta}K_\gamma(\alpha\sqrt{1-\theta})K_{\gamma+1}(\alpha\sqrt{1-\theta})}-\frac{2(\gamma\theta+1)}{\theta-1}\right)\ , \tag{2.153}$$

As any of the parameters α, θ, γ or ω increase from 0 the OD index becomes large. The zero-inflation index is given by,

$$ZI = 1-\frac{2\sqrt{1-\theta}K_\gamma(\alpha\sqrt{1-\theta})\log\left(\omega+\frac{(1-\theta)^{\gamma/2}(1-\omega)K_\gamma(\alpha)}{K_\gamma(\alpha\sqrt{1-\theta})}\right)}{\alpha\theta(\omega-1)K_{\gamma+1}(\alpha\sqrt{1-\theta})}\ . \tag{2.154}$$

The ZI Sichel distribution has SI,

$$SI_y = \frac{\sum_{y=0}^{\infty}\frac{4^{-y}\left(\alpha\theta^y(1-\theta)^{\frac{\gamma}{2}}(\omega-1)K_{y+\gamma}(\alpha)-2^y\omega K_\gamma(\alpha\sqrt{1-\theta})y!U_{-y}\right)^2}{K_\gamma(\alpha\sqrt{1-\theta})^2(y!)^2}}{\frac{2^{-y}\alpha\theta^y(1-\theta)^{\frac{\gamma}{2}}(1-\omega)K_{y+\gamma}(\alpha)}{K_\gamma(\alpha\sqrt{1-\theta})y!}+\omega U_{-y}}\ . \tag{2.155}$$

2.4.4 2-component Poisson Mixture (ω, μ, λ)

A two-component Poisson mixture of two Poisson distributions, ie. 2PO(ω, μ, λ) =Poisson(μ) * Poisson(λ), has *pdf*,

$$f_Y(y;\mu,\lambda,\omega) = P(Y=y) = \omega\frac{\mathrm{e}^{-\mu}\mu^y}{y!}+(1-\omega)\frac{\mathrm{e}^{-\lambda}\lambda^y}{y!}\ , \tag{2.156}$$

for $y = 0, 1, 2, \ldots$, where $\mu, \lambda > 0$ are the means of the two Poisson distributions and $0 < \omega < 1$ is the weighting parameter (Rose and Smith, 2002).

Figure 2.26 shows the Poisson-Poisson mix distribution for values of μ of 1 and 2 (rows) and values of λ of 5 and 10 (columns), whilst in each plot the weighting parameter ω varies by 0.2, 0.5 and 0.8. These graphs indicates that for some parameter values the density is bi-modal i.e. where one value of μ or λ is small and the other

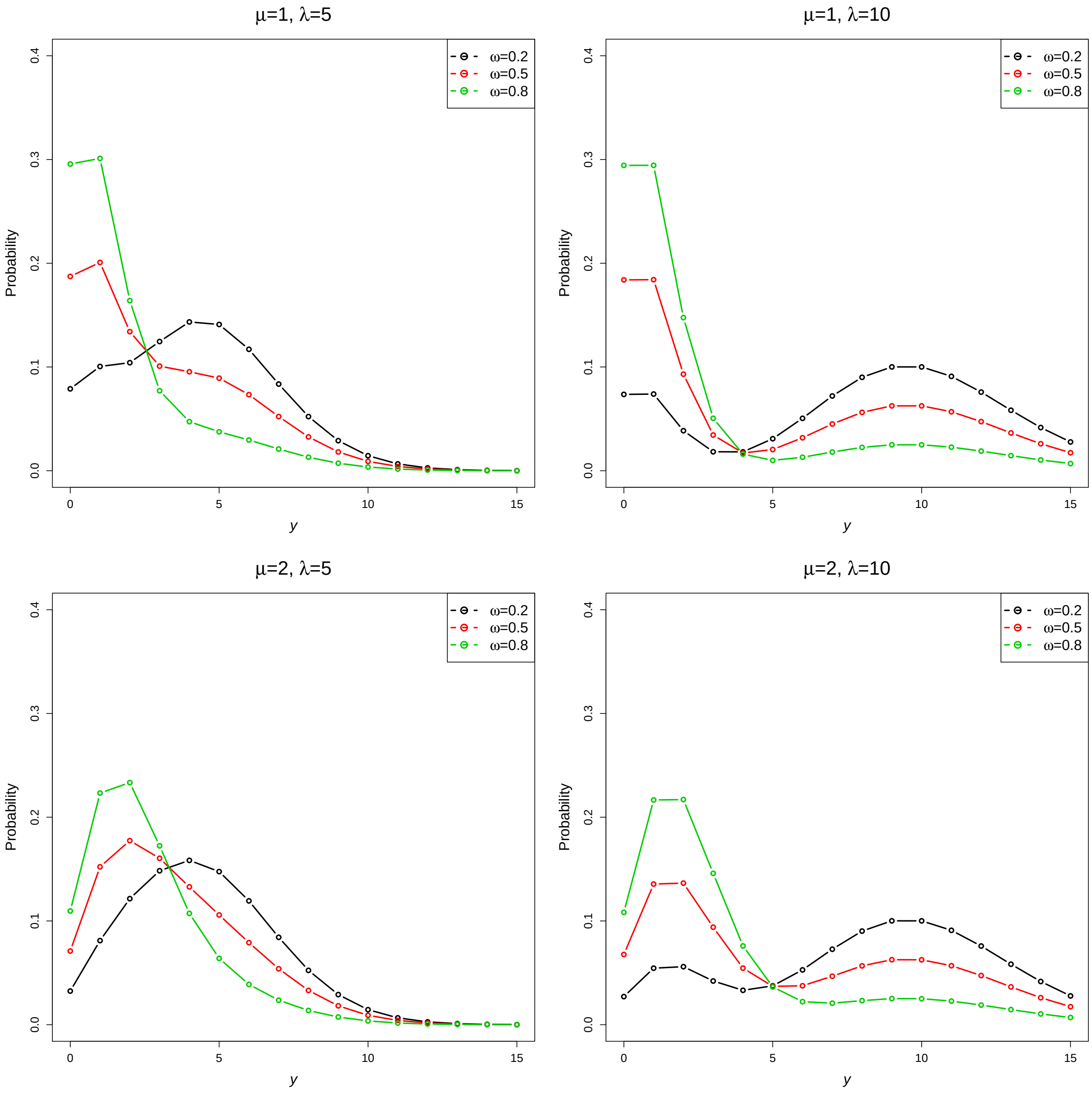


Figure 2.26: Poisson-Poisson mix *pdf*

large. The parameter ω adjusts the weighting between the two Poisson distributions.

The *pgf* of this distribution is,

$$G(t) = -\mathrm{e}^{(t-1)\lambda}(\omega - 1) + e^{(t-1)\mu}\omega \ , \tag{2.157}$$

and the *mgf* is

$$M(t) = -\mathrm{e}^{\left(\mathrm{e}^{t}-1\right)\lambda}(\omega - 1) + \mathrm{e}^{\left(\mathrm{e}^{t}-1\right)\mu}\omega \ . \tag{2.158}$$

This distribution has mean and variance

$$\mu = \lambda - \lambda\omega + \mu\omega \quad \text{and} \quad \sigma^2 = \lambda + (\lambda - \mu - 1)(\lambda - \mu)\omega - (\lambda - \mu)^2\omega^2 \ , \tag{2.159}$$

The overdispersion index is

$$OD = 1 + \mu + \lambda\omega - \mu\omega - \frac{\lambda\mu}{\lambda - \lambda\omega + \mu\omega} \ . \tag{2.160}$$

Where μ and λ are equal the OD index is equal to 1. As either μ or λ increases from 0 the OD index also increases, but is always greater than 1 indicating overdispersion is present in the distribution. The OD index is close to 1 where ω is 0, increases until $\omega = 0.5$ and then decreases to 1 as ω approaches 1. The zero-inflation index is,

$$ZI = 1 + \frac{\log(\mathrm{e}^{-\lambda}(\omega - 1) + \mathrm{e}^{-\mu}\omega)}{\lambda - \lambda\omega + \mu\omega} \ . \tag{2.161}$$

Again, where μ and λ are equal the ZI index is equal to 0. For values of μ is less than λ the ZI index decreases, and where λ is greater than μ the ZI index increases. The SI for this distribution is,

$$SI_y = -\frac{\mathrm{e}^{-\lambda-\mu} y! \left(\mathrm{e}^{2\mu}(\omega - 1)^2 \, \mathrm{I}_0(2\lambda) + \mathrm{e}^{\lambda}\omega \left(\mathrm{e}^{\lambda} \, \omega \mathrm{I}_0(2\mu) - 2\mathrm{e}^{\mu}(\omega - 1) \, {}_0F_1(;1;\lambda\mu)\right)\right)}{\mathrm{e}^{\mu}\lambda^y(\omega - 1) - \mathrm{e}^{\lambda}\mu^y\omega} \ . \tag{2.162}$$

Figure 2.27 plots SI's of the two-component Poisson mixture for values of $\mu = 1, 2$, $\lambda = 1, 10$ and $\omega = 0.2, 0.5$ and 0.8. All the plots show that for larger values of ω the SI is larger for high values of y. Where λ is greater than μ, higher values of y are more surprising.

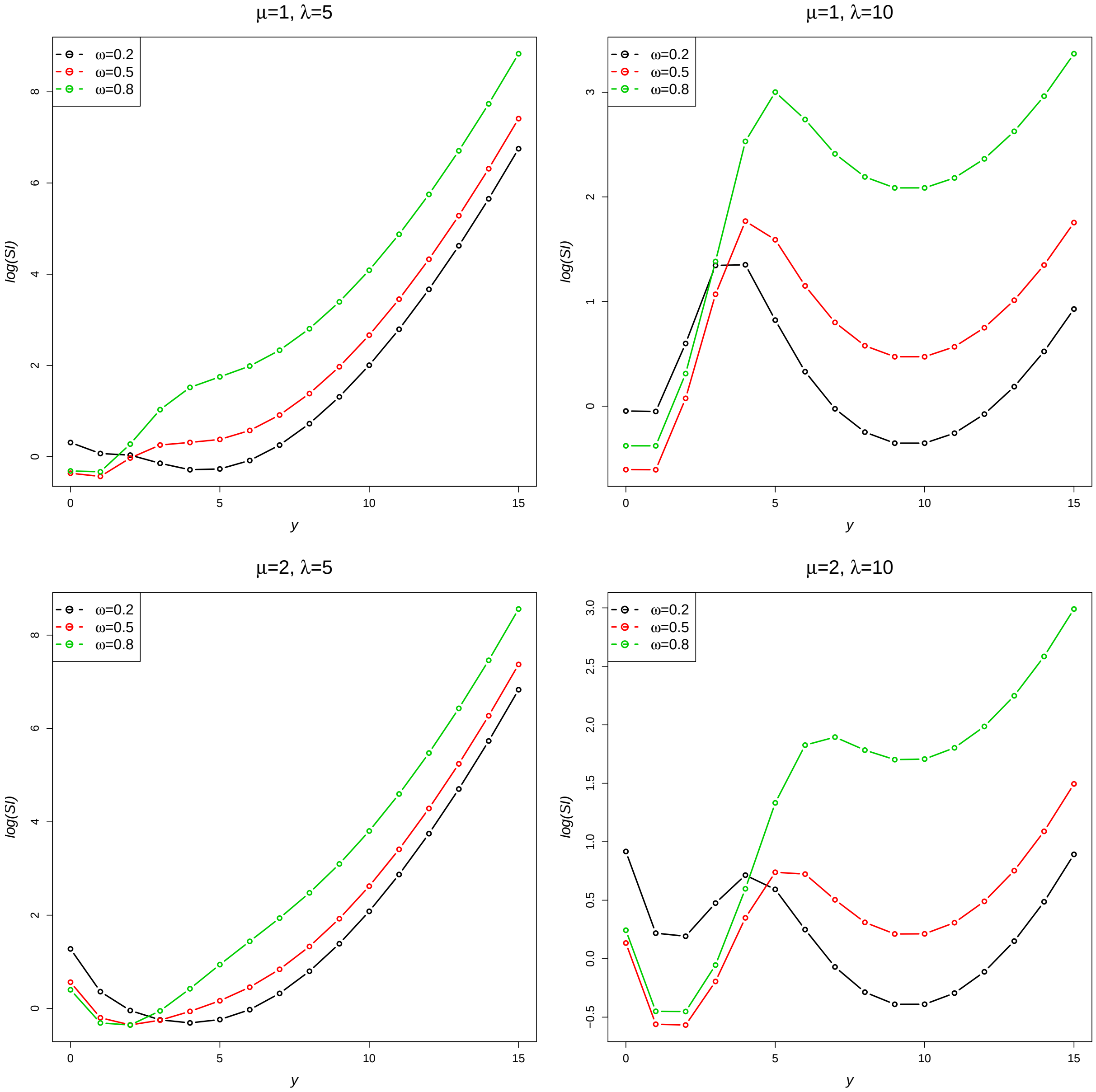


Figure 2.27: log(*SI*)'s for 2-component Poisson mixture distributions

2.4.5 2-component Poisson-Negative Binomial Mixture (ω, μ, r, p)

A component mix of Poisson and Negative Binomial distribution, i.e. 2PNB(ω, μ, r, p) =Poisson(μ) * NB(r, p) results in the following *pdf*,

$$f_Y(y; \mu, r, p, \omega) = P(Y = y) = \omega \frac{e^{-\mu}\mu^y}{y!} + (1-\omega)\binom{y+r-1}{r-1} p^r (1-p)^y \quad (2.163)$$

for $y = 0, 1, 2, \ldots$, for the parameters $\mu > 0$, $r > 0$, $0 < p < 1$ and $0 < \omega < 1$.

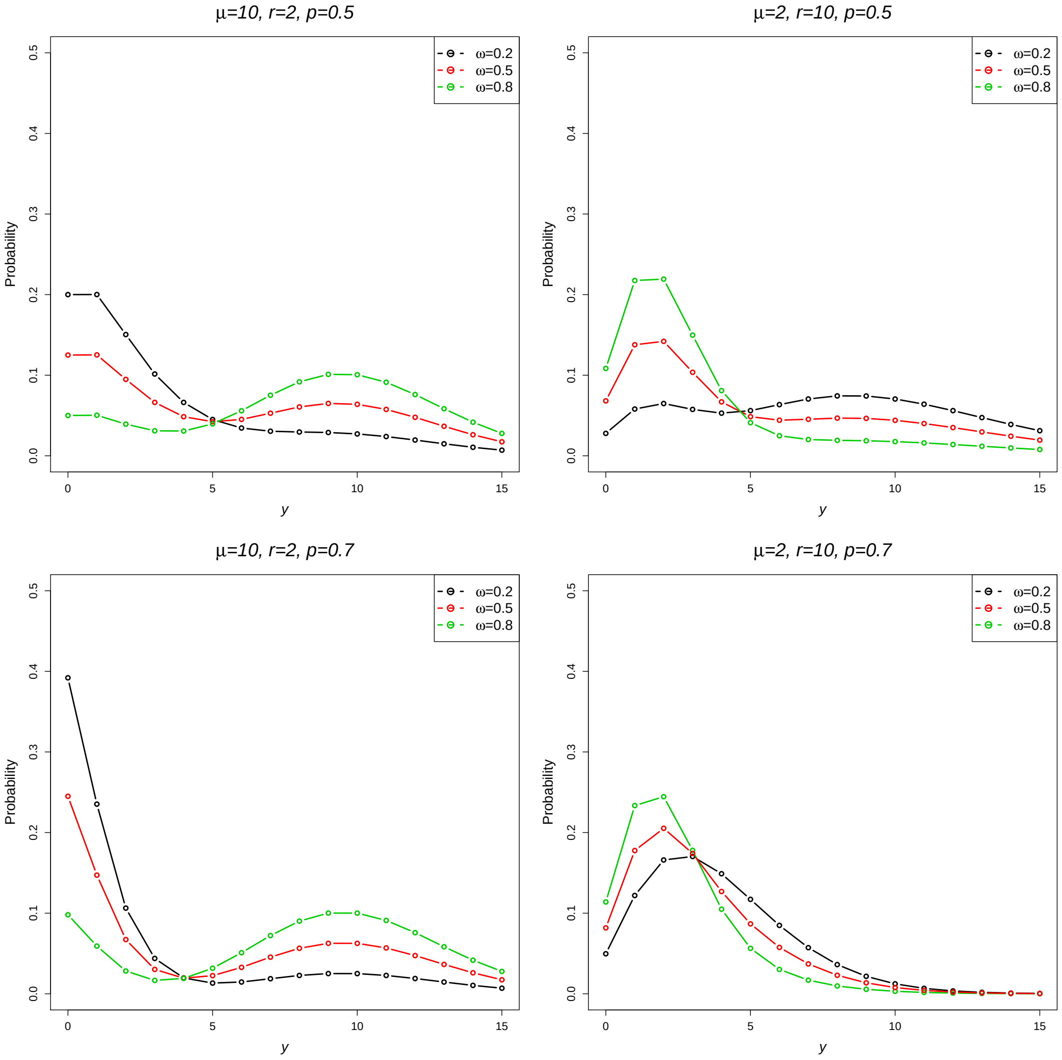


Figure 2.28: Poisson-Negative Binomial mix *pdf*

Figure 2.28 shows the density of the Poisson-Negative Binomial distribution for $\mu = 10$ and 2 (columns) , $r = 2$ and 10 (columns) and $p = 0.5$ and 0.7 (rows) for varying $w = 0.2$, 0.5 and 0.8. As in the Poisson-Poisson mixture distribution, this mixture results in a bi-modal distribution where ω controls the weighting between the two distributions.

The *pgf* of this distribution is given by,

$$G(t) = -p^r(1+(p-1)t)^{-r}(\omega-1) + \mathrm{e}^{(t-1)\mu}\omega \ , \tag{2.164}$$

with *mgf*,

$$M(t) = -\left(1+\mathrm{e}^t(p-1)\right)^{-r} p^r(\omega-1) + \mathrm{e}^{\left(\mathrm{e}^t-1\right)\mu}\omega \ . \tag{2.165}$$

The mean of this distribution is,

$$\mu = \frac{(1-p)r(1-\omega)+p\mu\omega}{p} \ , \tag{2.166}$$

and the variance,

$$\sigma^2 = \frac{(p-1)^2r^2(1-\omega)\omega + p^2\mu\,\omega(\mu-\mu\,\omega+1) + (1-p)r(1-\omega)(1+2p\,\mu\,\omega)}{p^2} \ . \tag{2.167}$$

The overdispersion index is,

$$OD = \frac{(1-p)^2r^2(1-\omega)\omega + p^2\mu\omega(1+\mu-\mu\omega) + (1-p)r(1-\omega)(2p\mu\omega-1)}{p((p-1)r(\omega-1)+p\mu\omega)} \ . \tag{2.168}$$

When ω equals 0, $OD = \frac{1}{p}$ which is the OD index for the negative binomial distribution. As ω equals 1, $OD = 1$ i.e. the value of the OD index for the Poisson distribution. The zero-inflation index is,

$$ZI = 1 + \frac{p\log\left(-p^r(\omega-1)+e^{-\mu}\omega\right)}{(p-1)r(\omega-1)+p\,\mu\,\omega} \ . \tag{2.169}$$

The ZI index is equal to the ZI of the negative binomial distribution $1 - \dfrac{p\log(p^r)}{(p-1)r}$ when ω is 0 and where ω is 1 the OD is equal to 0, as for the Poisson distribution. The two-component Poisson-negative binomial distribution has SI,

$$\begin{aligned} SI_y = \;& \big(\, r\mathrm{e}^{-\mu} y! \,\big(-\omega^2 \mathrm{I}_0(2\mu) + \mathrm{e}^{\mu} p^r (\omega - 1) \,\big(2\omega \,{}_1F_1(r;1;\mu - p\mu) - \\ & \mathrm{e}^{\mu} p^r (\omega - 1) \,{}_2F_1(r,r;1;(p-1)^2) \,\big)\,\big)\,\big) \,/\, \left(\frac{\mathrm{e}^{\mu}(1-p)^y p^r (\omega-1)\Gamma(r+y)}{\Gamma(r)} - \mu^y \omega \right) \end{aligned} \tag{2.170}$$

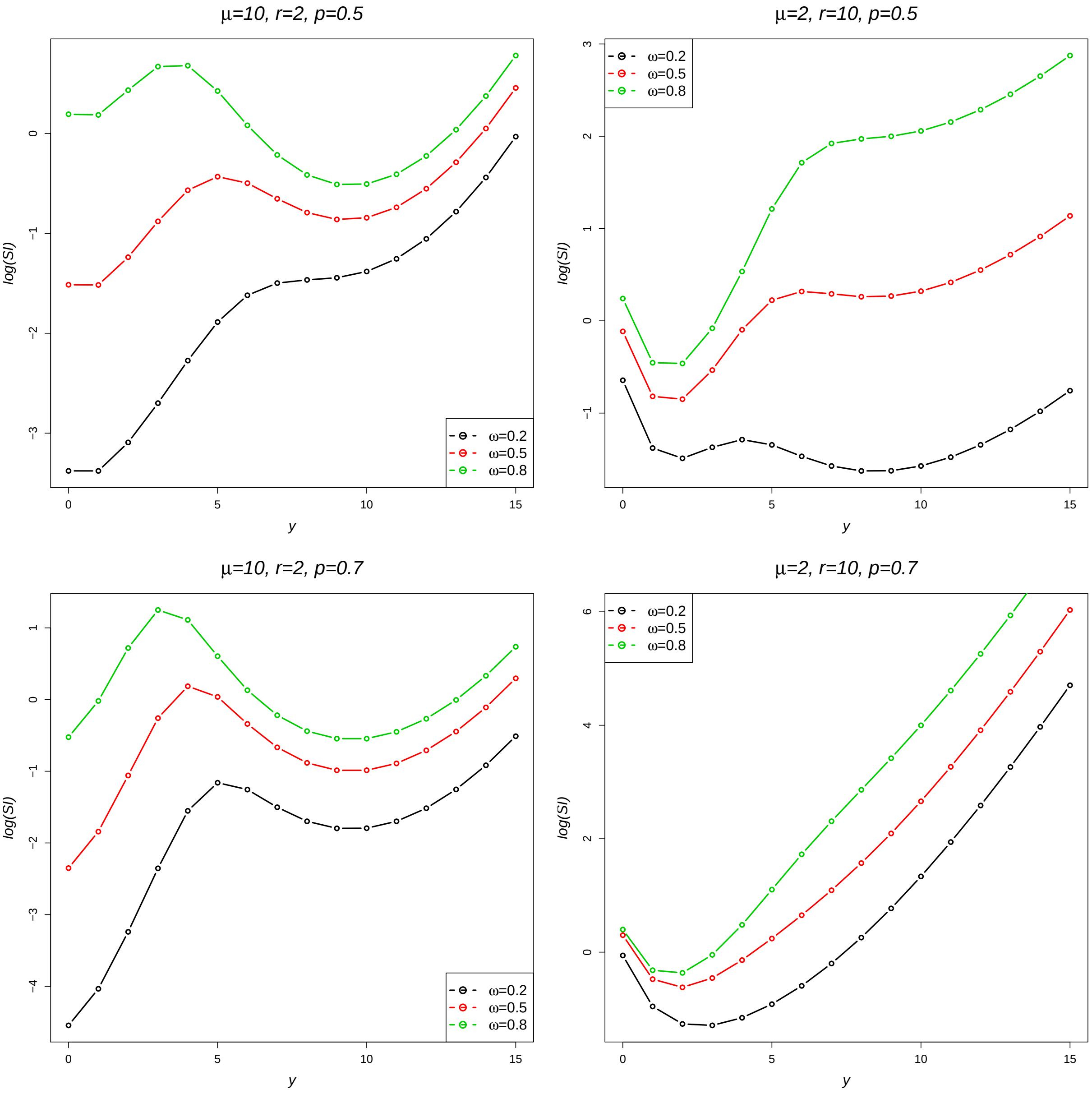


Figure 2.29: log(*SI*)'s for 2-component Poisson-Negative Binomial distributions

The four plots in Figure 2.29 show SI's for the 2-component Poisson-Negative Binomial distributions for values of $\mu = 2, 20$, $r = 2, 10$, $p = 0.5, 0.7$ and $\omega = 0.2, 0.5, 0.8$. Increasing the parameter ω results in a larger surprise index.

2.5 Truncated Distributions

Truncated distributions can be created through the conditional modification of parent distributions (Rose and Smith, 2002; Johnson et al., 2005). Let a single random variable, Y, have *pdf* $f(y)$ and *cdf* $F_Y(y) = P(Y \leq y)$. Further, there is a finite interval T with truncation points a and b inside the range of values taken by Y. If T consists of all values greater than a, then this results in a distribution that is *truncated below* or *left truncated*:

$$f_Y(y \,|\, Y > a) = \frac{f_Y(y)}{1 - F_Y(a)} \tag{2.171}$$

Similarly, if T consists of values less than b the distribution is said to be *truncated above* or *right truncated*:

$$f(y \,|\, Y \leq b) = \frac{f_Y(y)}{F_Y(b)} \tag{2.172}$$

A distribution can also be *doubly truncated*, that is truncated from both below (left) and above (right) where values of T are restricted within the truncation points a and b:

$$f_Y(y \,|\, a < Y \leq b) = \frac{f_Y(y)}{F_Y(b) - F_Y(a)} \tag{2.173}$$

In each case the conditional density is expressed in terms of the parent *pdf* which is scaled by a constant in the denominator to ensure the density still integrates to one (Rose and Smith, 2002).

The commonest form of truncated distribution is the omission of the zero class resulting in zero-truncated also called *positive distributions*. All these distributions have *pdf*'s of the form,

$$f_Y(y \,|\, Y > 0) = \frac{f_Y(y)}{1 - F_Y(0)}\ . \tag{2.174}$$

This section presents the positive forms of the Poisson, Geometric, Negative Binomial, Holla, Sichel and Yule distributions.

2.5.1 Positive Poisson (μ)

A common zero-truncated distribution is the zero-truncated or positive Poisson distribution with *pdf*,

$$f_Y(y;\mu) = P(Y=y) = \frac{\mathrm{e}^{-\mu}\mu^y}{y!(1-\mathrm{e}^{-\mu})}\ , \tag{2.175}$$

for $\mu > 0$ and $y = 1, 2, 3, \ldots$ (Johnson et al., 2005; Wimmer and Altmann, 1999, P.544). This distribution is also known as the conditional Poisson distribution (Cohen, 1960). The Truncated Poisson probability distribution is illustrated in Figure 2.30 for values of μ of 2, 5, 10 and 20. As the parameter μ becomes large the distribution tends to a normal distribution. For small values of μ i.e. a rare event with small mean, the distribution is skew.

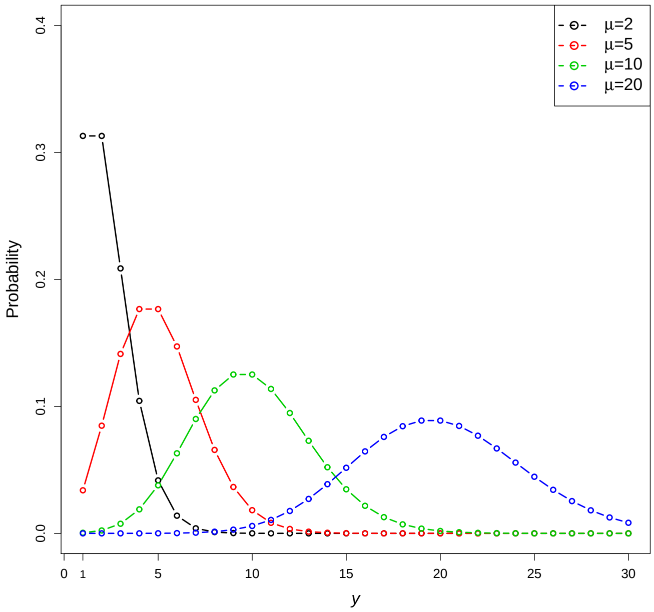


Figure 2.30: Positive Poisson probability *pdf*

The *pgf* for the Positive Poisson distribution is,

$$G(t) = \frac{\mathrm{e}^{t\mu}-1}{\mathrm{e}^{\mu}-1}\ , \tag{2.176}$$

and the *mgf*,

$$M(t) = \frac{\mathrm{e}^{\mathrm{e}^t\mu}-1}{\mathrm{e}^{\mu}-1}\ . \tag{2.177}$$

The mean and variance of the distribution are,

$$\mu = \left(\frac{\mathrm{e}^{\mu}}{\mathrm{e}^{\mu}-1}\right)\mu \quad \text{and} \quad \sigma^2 = \frac{e^{\mu}\left(e^{\mu}-\mu-1\right)\mu}{\left(e^{\mu}-1\right)^2}\ , \tag{2.178}$$

respectively. The overdispersion index is,

$$OD = 1 - \frac{\mu}{\mathrm{e}^{\mu}-1}\ . \tag{2.179}$$

When μ is small, the OD index is 0 and as μ increases the OD approaches 1. The SI of the positive Poisson distribution is,

$$SI_y = \frac{\mathrm{e}^{\mu}\left(1-\mathrm{e}^{-\mu}\right)\mu^{-y}(I_0(2\mu)-1)y!}{\left(\mathrm{e}^{\mu}-1\right)^2}\ . \tag{2.180}$$

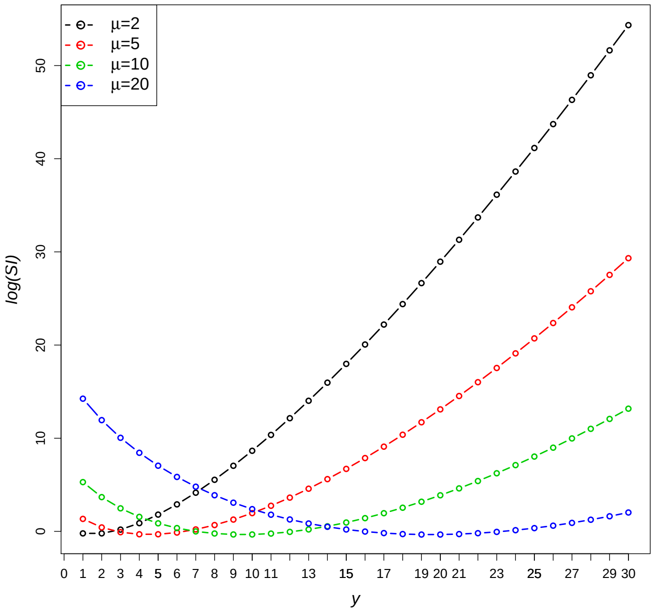


Figure 2.31: log(*SI*)'s for the Positive Poisson distribution

Figure 2.31 plots log(SI)'s for various Positive Poisson distributions where $\mu = 2, 5, 10$ and 20. For small values of μ the SI is large for high values of Y and as μ increases the SI becomes increasingly large for low values of Y.

2.5.2 Positive Geometric (p)

The zero-truncated Geometric distribution has *pdf*,

$$f_Y(y;r,p) = P(Y=y) = (1-p)^{y-1}p \; , \tag{2.181}$$

for $y = 1,2,3\ldots$ and $0 < p < 1$ (Johnson et al., 2005). The Positive Geometric density function is plotted in Figure 2.32 for values of p of 0.2, 0.4, 0.6 and 0.8. For increasing values of p which tend to one, the distribution becomes more skew with a shorter tail and a higher probability of lower values of y.

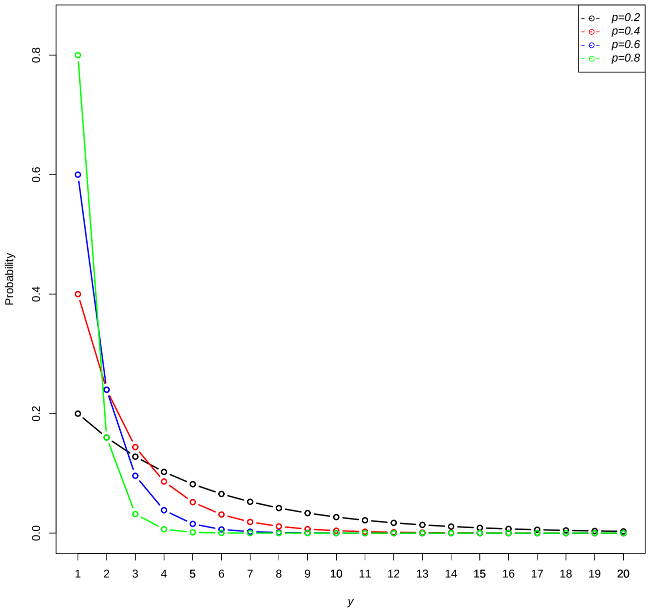


Figure 2.32: Positive Geometric *pdf*

The *pgf* is,

$$G(t) = \frac{pt}{1+(p-1)t} \; , \tag{2.182}$$

and the *mgf*,

$$M(t) = \frac{\mathrm{e}^t p}{1+\mathrm{e}^t(p-1)} \; . \tag{2.183}$$

The mean and variance of the positive geometric distribution are,

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma^2 = \frac{1-p}{p^2} \quad , \tag{2.184}$$

The overdispersion index is,

$$OD = \frac{1-p}{p} \ . \tag{2.185}$$

When $p \leq 0.5$, the OD index is greater than 1 indicating overdispersion is present in the distribution and as $p \to 1$ the OD approaches 0. The SI is given by,

$$SI_y = \frac{(1-p)^{1-y}}{2-p} \ . \tag{2.186}$$

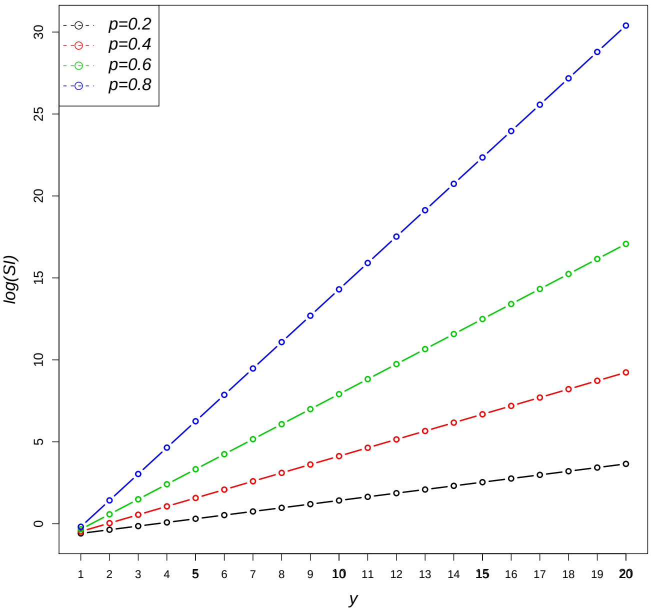


Figure 2.33: log(*SI*)'s for Positive Geometric distributions

As p increases, the SI increases for large Y values, shown in Figure 2.33 which plots SI's for the Positive Geometric distribuion for values of p of 0.2, 0.4, 0.6 and 0.8.

2.5.3 Positive Negative Binomial (r, p)

Also known as the zero-truncated negative binomial, the positive negative binomial distribution has *pdf*,

$$f_Y(y; r, p) = P(Y = y) = \frac{\binom{y+r-1}{y} p^r (1-p)^y}{1 - p^r} , \tag{2.187}$$

for $y = 1, 2, 3 \ldots r \geq 0$ and $0 < p < 1$ (Wimmer and Altmann, 1999, P.540). This distribution is equivalent to a positive geometric distribution when $r = 1$. The first plot in Figure 2.34 shows the effect of varying r at 1,2,5 and 10 when p is fixed at 0.5. For small values of r the distribution is skew and as r increases the distribution becomes flat. In the second plot r is fixed at 2 and p ranges between 0.25, 0.50, 0.75 and 0.9. For values of p near 1 (as p increases) the probability of low values increases.

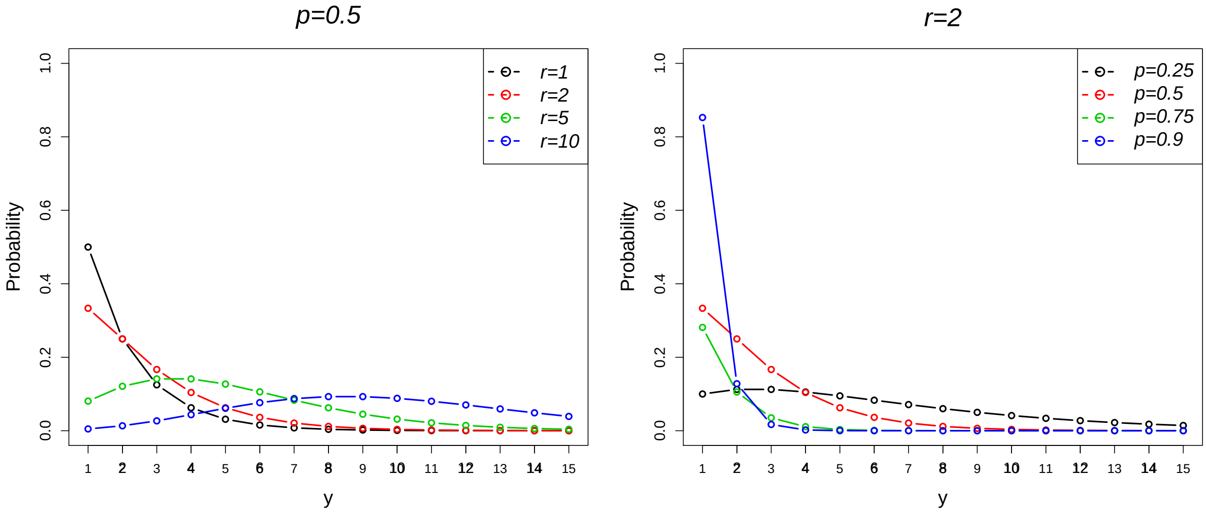


Figure 2.34: Positive Negative Binomial *pdf*

This distribution has *pgf*

$$G(t) = -\frac{p^r \left(1 - (1 + (p-1)t)^{-r}\right)}{1 - p^r} , \tag{2.188}$$

and *mgf*

$$M(t) = \frac{\left(1 - (1 + \mathrm{e}^t (p-1))^{-r}\right) p^r}{p^r - 1} . \tag{2.189}$$

The mean and variance of the distribution are,

$$\mu = \frac{(p-1)r}{p\,(p^r-1)} \quad \text{and} \quad \sigma^2 = \frac{(1-p)r\,(1+p^r((p-1)r-1))}{p^2\,(p^r-1)^2}\,, \tag{2.190}$$

respectively. The overdispersion index is given by,

$$OD = \frac{p^r(1+r-pr)-1}{p\,(p^r-1)} \tag{2.191}$$

For large values of r, the OD index increases and as p tends to 1 the OD index decreases. The SI is given by,

$$SI_y = \frac{(1-p)^{-y}p^r\,(1-p^r)\,(\,{}_2F_1(r,r;1;(p-1)^2)-1)}{(p^r-1)^2\,\binom{y+r-1}{r-1}}\,. \tag{2.192}$$

Figure 2.35 plots log(SI) for the Positive Negative Binomial distribution with parameters of $r = 1, 2, 5, 10$ where $p = 0.5$ and in the second plot $p = 0.25, 0.5, 0.75, 0.9$ with $r = 2$. As r increases the SI becomes smaller and less skew towards high Y values and as p approaches 1, the SI also decreases for high values of Y.

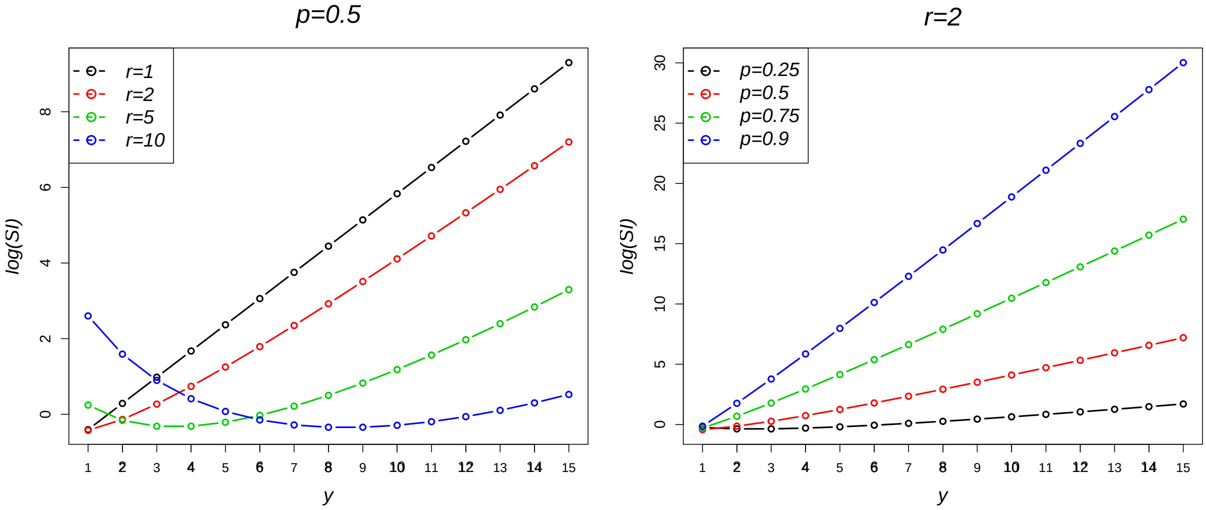


Figure 2.35: log(*SI*)'s for Positive Negative Binomial distributions

2.5.4 Positive Holla (α, θ)

The Holla and Sichel distributions have been widely used to analyse word frequency and species abundance frequency data in the fields of linguistics or ecology, where distributions for counts of species or lengths of words take values in the range $R_y = \{1, 2, \ldots\}$ (Sichel, 1975; Puig et al., 2009; Ginebra and Puig, 2010). The zero-truncated or positive Holla or Poisson-Inverse Gaussian distribution has *pdf*

$$f_Y(y; \alpha, \theta) = P(Y = y) = \frac{\left(\frac{2\alpha}{\pi}\right)^{\frac{1}{2}} \mathrm{e}^{\alpha}}{\left(\mathrm{e}^{\alpha\left(1-(1-\theta)^{\frac{1}{2}}\right)} - 1\right)} \frac{\left(\frac{\alpha\theta}{2}\right)^{y}}{y!} K_{y-\frac{1}{2}}(\alpha) \ , \qquad (2.193)$$

for $y = 1, 2, 3, \ldots$ where $\alpha \geq 0$, $0 < \theta \leq 1$ and $0 \leq \frac{\alpha}{\theta} < 1$ (Wimmer and Altmann, 1999, P.547). The Positive Holla distribution is plotted in Figure 2.36 where θ is first fixed at 0.5 and α is in the range 1,2,5 and 10, (first plot) and then α is fixed at 2 and θ has values 0.25, 0.50, 0.75 and 0.90. As α increases the distribution becomes less skewed and as θ decreases the probability of a low y value increases.

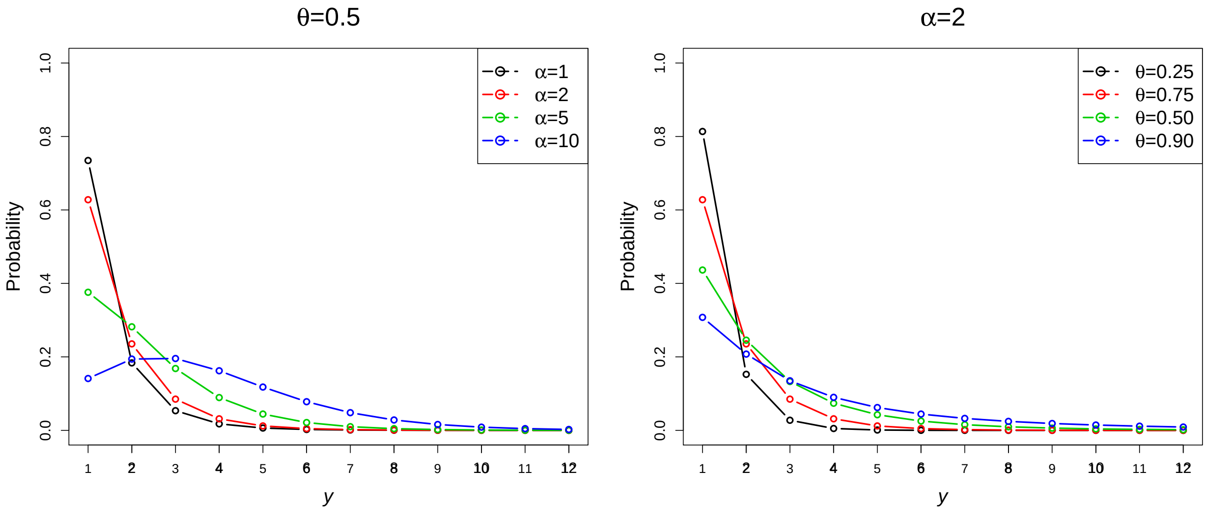


Figure 2.36: Positive Holla *pdf*

The *pgf* is given by,

$$G(t) = \frac{\sqrt{1-\theta} - \sqrt{1-\theta t}}{1 - \sqrt{1-\theta}} \frac{{}_1F_1(1; 2; \alpha\sqrt{1-\theta} - \alpha\sqrt{1-\theta t})}{{}_1F_1(1; 2; \alpha\sqrt{1-\theta} - \alpha)} \ , \qquad (2.194)$$

and the *mgf*

$$M(t) = \frac{\mathrm{e}^{\alpha}\left(\mathrm{e}^{\alpha\left(\sqrt{1-\theta}-\sqrt{1-\mathrm{e}^{t}\theta}\right)}-1\right)}{\mathrm{e}^{\alpha}-\mathrm{e}^{\alpha\sqrt{1-\theta}}}\,. \tag{2.195}$$

The mean is

$$\mu = \frac{e^{\alpha}\alpha\theta}{\left(2e^{\alpha}-2e^{\alpha\sqrt{1-\theta}}\right)\sqrt{1-\theta}}\,, \tag{2.196}$$

and the variance

$$\sigma^2 = \frac{e^{\alpha}\alpha\theta\left(e^{\alpha}(\theta-2)+e^{\alpha\sqrt{1-\theta}}\left(2+\left(\alpha\sqrt{1-\theta}-1\right)\theta\right)\right)}{4\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)^2(1-\theta)^{3/2}}\,. \tag{2.197}$$

and the overdispersion index is therefore,

$$OD = \frac{e^{\alpha}(\theta-2)+e^{\alpha\sqrt{1-\theta}}\left(2+\left(\alpha\sqrt{1-\theta}-1\right)\theta\right)}{2\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)(\theta-1)}\,. \tag{2.198}$$

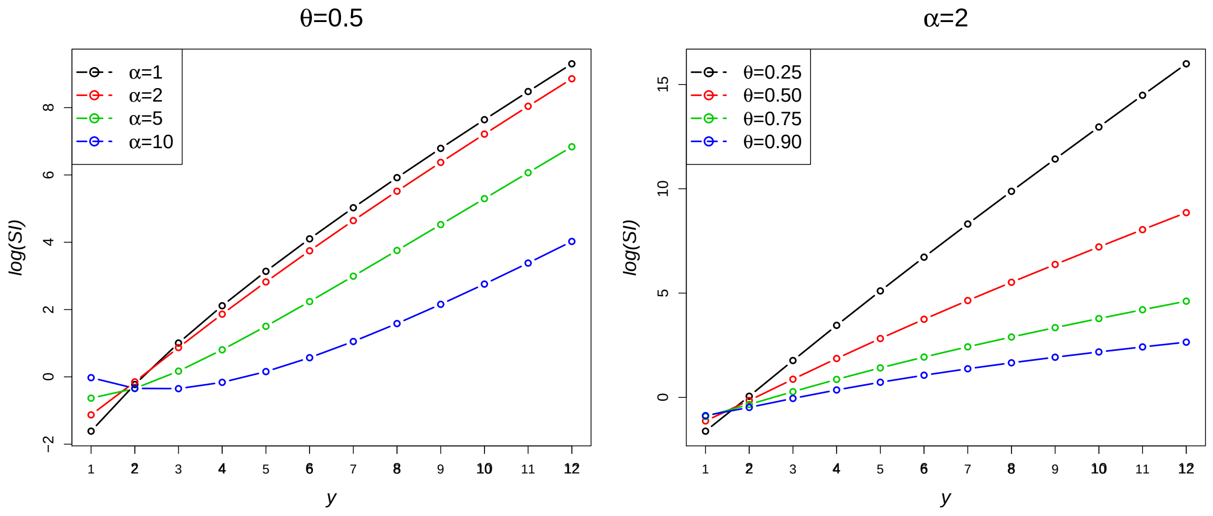


Figure 2.37: log(*SI*)'s for Positive Holla distributions

As either α or θ increase the OD index increases, and for large values of both α and θ the $OD > 1$. The SI of the positive Holla distribution is,

$$SI_y = \frac{\left(e^{\frac{1}{2}\alpha^2\sqrt{1-\theta}\theta}\right)^{-y}\left(1-e^{-\alpha}\right)\sqrt{\frac{\pi}{2}}y!\sum_{y=1}^{\infty}\frac{2e^{2\alpha}\left(e^{\frac{1}{2}\alpha^2\sqrt{1-\theta}\theta}\right)^{2y}\alpha K_{y-\frac{1}{2}}(\alpha)^2}{\left(-1+e^{\alpha}\right)^2\pi(y!)^2}}{\sqrt{\alpha}\,K_{y-\frac{1}{2}}(\alpha)}\,, \tag{2.199}$$

The first plot in Figure 2.37 illustrate the SI's for $\alpha = 1, 2, 5$ and 10, where $\theta = 0.5$ and as α increases values of Y become less surprising. In the second plot $\alpha = 2$ and θ is in the range $0.25, 0.50, 0.75$ and 0.90, with larger values of θ resulting in lower SI's.

2.5.5 Positive Sichel (α, θ, γ)

The *pdf* of the zero-truncated Sichel distribution is given by:

$$f_Y(y; \alpha, \theta, \gamma) = P(Y = y) = \frac{1}{(1-\theta)^{-\frac{\gamma}{2}} K_\gamma(\alpha(1-\theta)^{\frac{1}{2}}) - K_\gamma(\alpha)} \frac{\left(\frac{\alpha\theta}{2}\right)^y}{y!} K_{y+\gamma}(\alpha) \ , \tag{2.200}$$

for $y = 1, 2, 3, \ldots$ where $\alpha > 0$, $0 < \theta < 1$, $\gamma \in \mathbb{R}$ (Puig et al., 2009; Ginebra and Puig, 2010; Wimmer and Altmann, 1999, P.548). This distribution is also sometimes known as a truncated Generalized Inverse Gaussian-Poisson or positive Sichel distribution. It is also equal to a truncated Holla distribution when $\gamma = -\dfrac{1}{2}$ (Wimmer and Altmann, 1999). Figure 2.38 shows the Positive Sichel distribution for values of $\alpha = 0.5, 1, 2, 5, 10$ (first plot), $\theta = 0.10, 0.25, 0.50, 0.75, 0.90$ (second plot) and $\gamma = 1.0, -0.5, 0, 1, 2$ (third plot) where the remaining two parameters are fixed at $\alpha = 2$, $\theta = 0.5$ and $\gamma = -0.5$. As in the non-truncated version of the Sichel distribution the parameter α characterizes the low counts of y, θ influences the tail of distribution and γ parametrizes the overall shape of the distribution.

The *pgf* for the Positive Sichel distribution is

$$G(t) = \left(\frac{1-\theta}{1-\theta t}\right)^{\frac{\gamma}{2}} \frac{K_\gamma(\alpha\sqrt{1-\theta t}) - (1-\theta t)^{\frac{\gamma}{2}} K_\gamma(\alpha)}{K_\gamma(\alpha\sqrt{1-\theta}) - (1-\theta)^{\frac{\gamma}{2}} K_\gamma(\alpha)} \ , \tag{2.201}$$

with *mgf*

$$M(t) = \left(\frac{1-\theta}{1-\theta \mathrm{e}^t}\right)^{\frac{\gamma}{2}} \frac{K_\gamma(\alpha\sqrt{1-\theta \mathrm{e}^t}) - (1-\theta \mathrm{e}^t)^{\frac{\gamma}{2}} K_\gamma(\alpha)}{K_\gamma(\alpha\sqrt{1-\theta}) - (1-\theta)^{\frac{\gamma}{2}} K_\gamma(\alpha)} \ . \tag{2.202}$$

The mean is given by,

$$\mu = \frac{\alpha\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})}{\sqrt{1-\theta}\left(2(1-\theta)^{\frac{\gamma}{2}} K_\gamma(\alpha) - 2K_\gamma(\alpha\sqrt{1-\theta})\right)} \ , \tag{2.203}$$

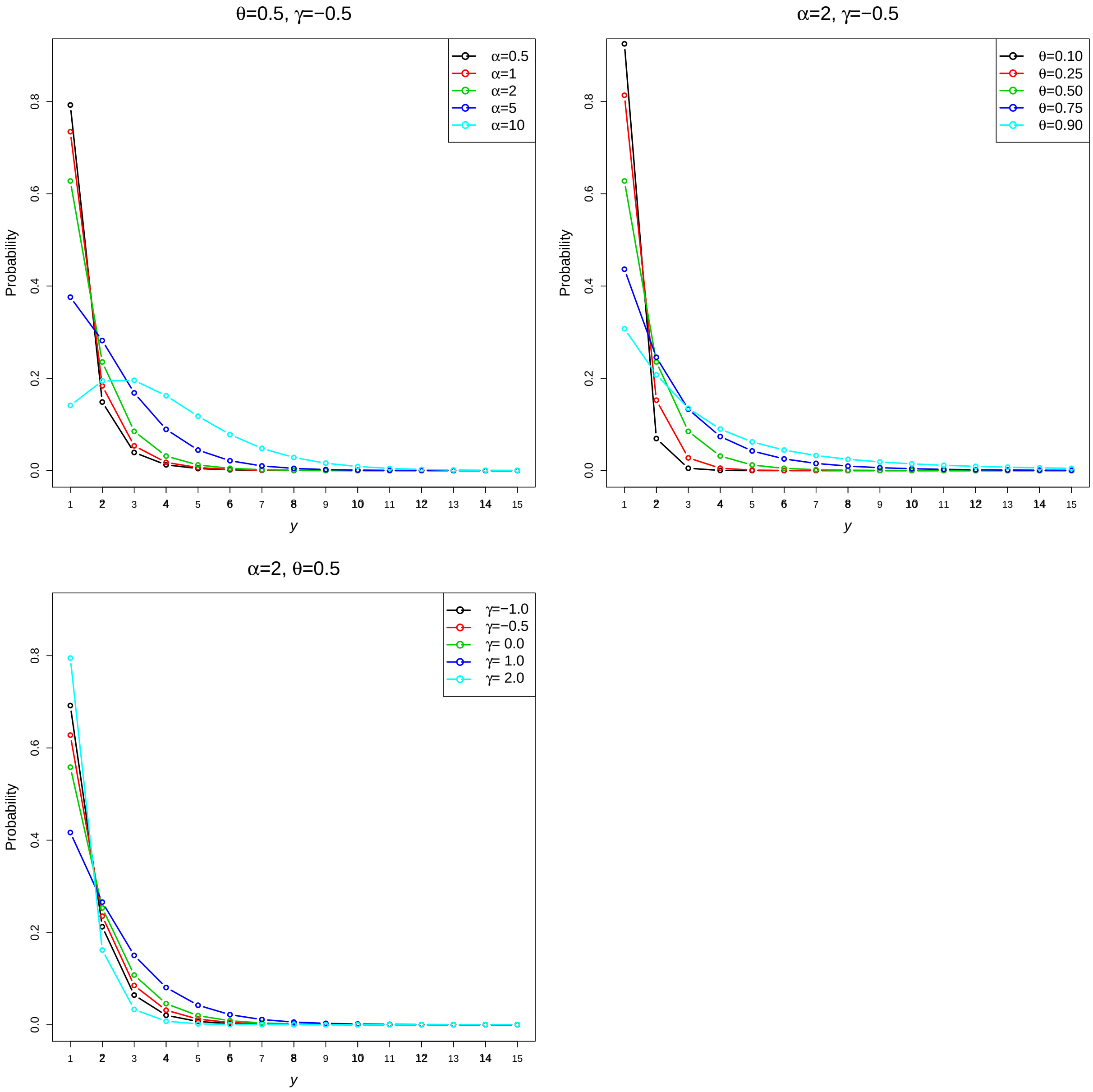


Figure 2.38: Positive Sichel *pdf*

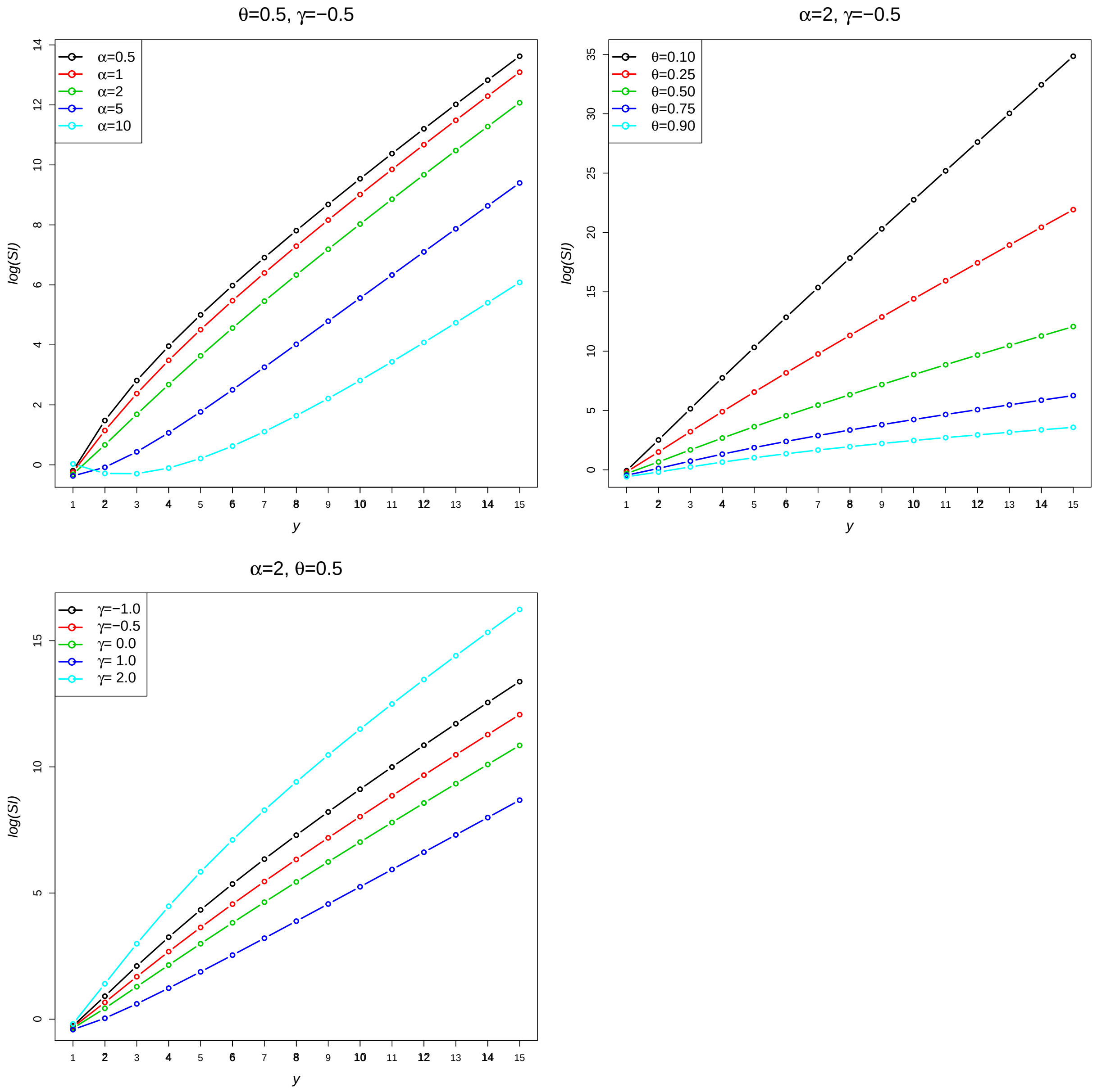


Figure 2.39: log(*SI*)'s for Positive Sichel distributions

and the variance,

$$
\begin{aligned}
\sigma^2 = \; & \tfrac{1}{4}\theta\,(\,(\,\frac{\left(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\right)K_{\gamma-2}(\alpha\sqrt{1-\theta})}{(\theta-1)^2}+ \\
& \frac{2\left(-4(\gamma-1)\gamma(\gamma\theta+1)+\alpha^2(\theta-1)(1+(2\gamma-1)\theta)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{\alpha(1-\theta)^{5/2}}\;)\,/ \\
& \left((1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)-K_\gamma(\alpha\sqrt{1-\theta})\right) \qquad . \\
& +\frac{\alpha^2\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})^2}{(\theta-1)\left(-(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right)^2}\;) \\
. &
\end{aligned}
\tag{2.204}
$$

and the overdispersion index is,

$$
\begin{aligned}
OD = \; & -\frac{1}{4\alpha K_{\gamma+1}(\alpha\sqrt{1-\theta})}\sqrt{1-\theta}\left(2(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)-2K_\gamma(\alpha\sqrt{1-\theta})\right) \\
& (\,(\,\frac{\left(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\right)K_{\gamma-2}(\alpha\sqrt{1-\theta})}{(\theta-1)^2}+ \\
& \frac{2\left(-4(\gamma-1)\gamma(\gamma\theta+1)+\alpha^2(\theta-1)(1+(2\gamma-1)\theta)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{\alpha(1-\theta)^{\frac{5}{2}}}\;)\,/ \quad . \\
& \left((1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)-K_\gamma(\alpha\sqrt{1-\theta})\right)\;) \\
& +\frac{\alpha^2\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})^2}{(\theta-1)\left(-(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right)^2}
\end{aligned}
\tag{2.205}
$$

As the values of α, θ and γ increase the dispersion in the distribution increases. The SI is,

$$
\begin{aligned}
SI_y = \; & \frac{1}{K_{y+\gamma}(\alpha)}2^y(\alpha\theta)^{-y}\left(-K_\gamma(\alpha)+(1-\theta)^{-\gamma/2}K_\gamma(\alpha\sqrt{1-\theta})\right)y!\times \\
. \quad & \sum_{y=1}^{\infty}\frac{4^{-y}(\alpha\theta)^{2y}\mathrm{K}_{\mathrm{y}+\gamma}(\alpha)^2}{\left(\mathrm{K}_\gamma(\alpha)-(1-\theta)^{-\gamma/2}\mathrm{K}_\gamma(\alpha\sqrt{1-\theta})\right)^2(y!)^2} \qquad ,
\end{aligned}
\tag{2.206}
$$

Plots of the SI's (Figure 2.39) indicate that larger values of Y are more surprising. For larger values of α or θ the SI decreases but as γ increases the SI decreases.

2.5.6 Positive Yule (λ)

The zero-truncated Yule distribution has been used to model word frequencies in texts by (Simon, 1955) and also the distribution of under-reporting in incomes by Krishnaji (1970). This distribution has *pdf*

$$f_Y(y;\lambda) = P(Y=y) = \frac{\lambda\Gamma(y)\Gamma(\lambda+1)}{\Gamma(\lambda+y+1)} \;, \tag{2.207}$$

for $y = 1, 2, 3\ldots$ and $\lambda > 0$ (Wimmer and Altmann, 1999; Rose and Smith, 2002, P.107). This distribution can also be generated from a parameter mixture where Yule(λ)=Geometric$(\mathrm{e}^{-W}) \bigwedge_{W}$ Exponential$(\frac{1}{\lambda})$ (Wimmer and Altmann, 1999, P.549). The Yule probability density is plotted for values of λ of 1, 2, 5 and 10 in Figure 2.40. As λ increases the distribution becomes more skew with a higher probability of low counts of y and a reduction in the tail of the distribution.

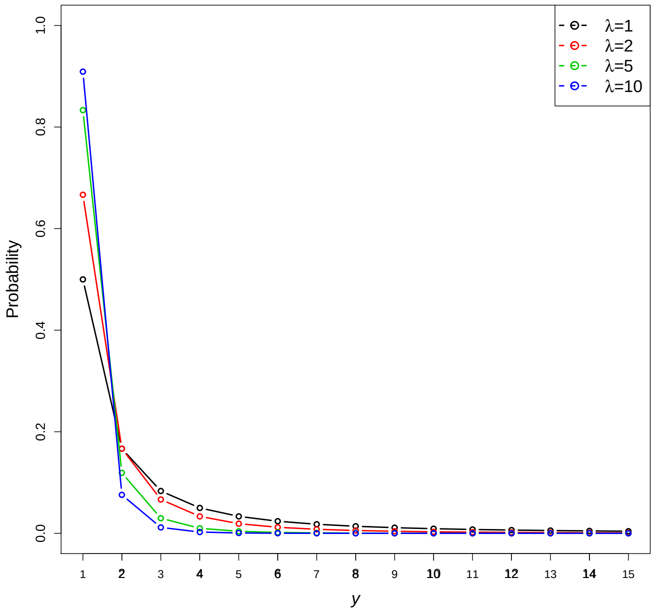


Figure 2.40: Positive Yule *pdf*

The distribution has *pgf*

$$G(t) = t\,\lambda\Gamma(\lambda+1)\,\frac{{}_2F_1(1,2;3+\lambda;t)}{\Gamma(\lambda+2)} \;, \tag{2.208}$$

and *mgf*

$$M(t) = \mathrm{e}^t \lambda \Gamma(\lambda+1)\, \frac{{}_2F_1(1,2;3+\lambda;\mathrm{e}^t)}{\Gamma(\lambda+2)} \,, \tag{2.209}$$

The positive Yule distribution therefore has mean

$$\mu = \frac{\lambda}{\lambda - 1} \,, \tag{2.210}$$

and variance,

$$\sigma^2 = \frac{\lambda^2}{(\lambda-2)(\lambda-1)^2} \,. \tag{2.211}$$

The overdispersion index is,

$$OD = \frac{\lambda}{\lambda^2 - 3\lambda + 2} \,. \tag{2.212}$$

For small values of λ the OD index is large and as λ increases the OD index decreases.

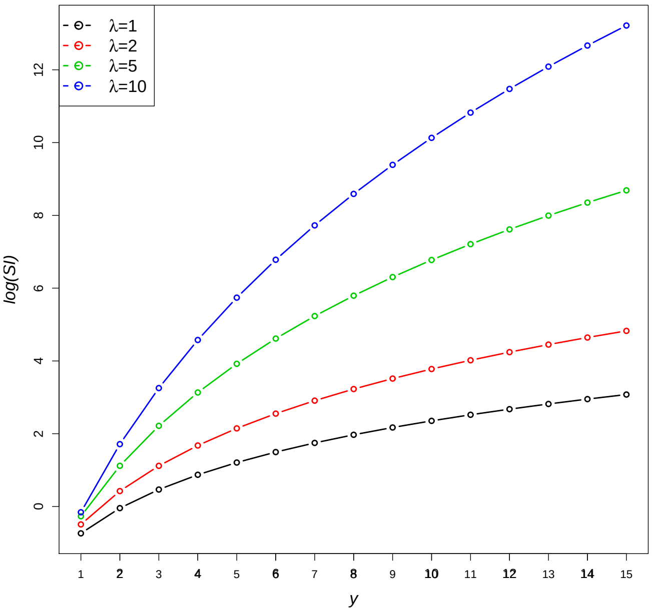


Figure 2.41: log(*SI*)'s for Positive Yule distributions

The SI for the positive Yule distribution is,

$$SI_y = \frac{1}{\Gamma(y)\Gamma(\lambda+1)}\, \lambda^3\, \Gamma(\lambda)^2\, \Gamma(y+\lambda+1)\, {}_3F_2(1,1,1;\lambda+2,\lambda+2;1) \,. \tag{2.213}$$

The SI is larger for higher values of λ, shown in Figure 2.41 which plots the SI's for the Positive Yule distribution with values of $\lambda = 1, 2, 5$ and 10.

2.6 Lerch Family Distributions

The Lerch family of distributions (Kulasekera and Tonkyn, 1992; Kemp, 1995; Zörnig and Altmann, 1995; Doray and Luong, 1997) is formed of distributions based on the Lerch Zeta function (Wimmer and Altmann, 1999, pg. XXIV) defined as,

$$\Phi(p, a, c) = \sum_{y=1}^{\infty} \frac{p^y}{(a+y)^c} , \tag{2.214}$$

where $p > 0$ and $a > 0$. The special case where $p = 1$, $a = 1$ and $c > 1$ is the Reimann Zeta function $\zeta(c)$ and where $p = 1$, $a \neq 0, -1, -2, \ldots$ and $c > 1$ is the Hurwitz Zeta function $\zeta(c, a)$ (Johnson et al., 2005, pg. 527). The general form of the Lerch distribution utilizes the Lerch Zeta function and distributions within the Lerch family have *pgf*'s of the form,

$$G(t) = \frac{\Phi\left(p\,t, a, c\right)}{\Phi\left(p, a, c\right)} \tag{2.215}$$

where $p > 0$ and $a > 0$ for $G(t)$ to be a valid *pgf* with non-negative probabilities and range $0, 1, 2, \ldots$ (Johnson et al., 2005).

The Lerch family of distributions have applications in many fields for example, modelling word frequencies in linguistics (Zipf, 1949), surname distributions (Fox and Lasker, 1983), counts of insurance policies (Seal, 1947), species distributions (Yule, 1925) and ranking size of cities (Brakman et al., 1999). The Estoup, Lotka, Zeta, Zipf and Good distributions can be considered as special cases of the more general Lerch distribution (Zörnig and Altmann, 1995) and are presented in this section.

2.6.1 Lerch (p, a, c)

The *pdf* of the general form of the Lerch distribution is given by,

$$f_Y(y; p, a, c) = P(Y = y) = \frac{p^y}{T \times, (a + y)^c} \ , \tag{2.216}$$

for $y = 1, 2, 3, \ldots$ where $a > 0$, $c \geq 0$ and $0 \leq p < 1$ (Zörnig and Altmann, 1995; Wimmer and Altmann, 1999) where $T = \Phi(p, a, c)$ is the Lerch Zeta function (see equation 2.214 in Section 2.6).

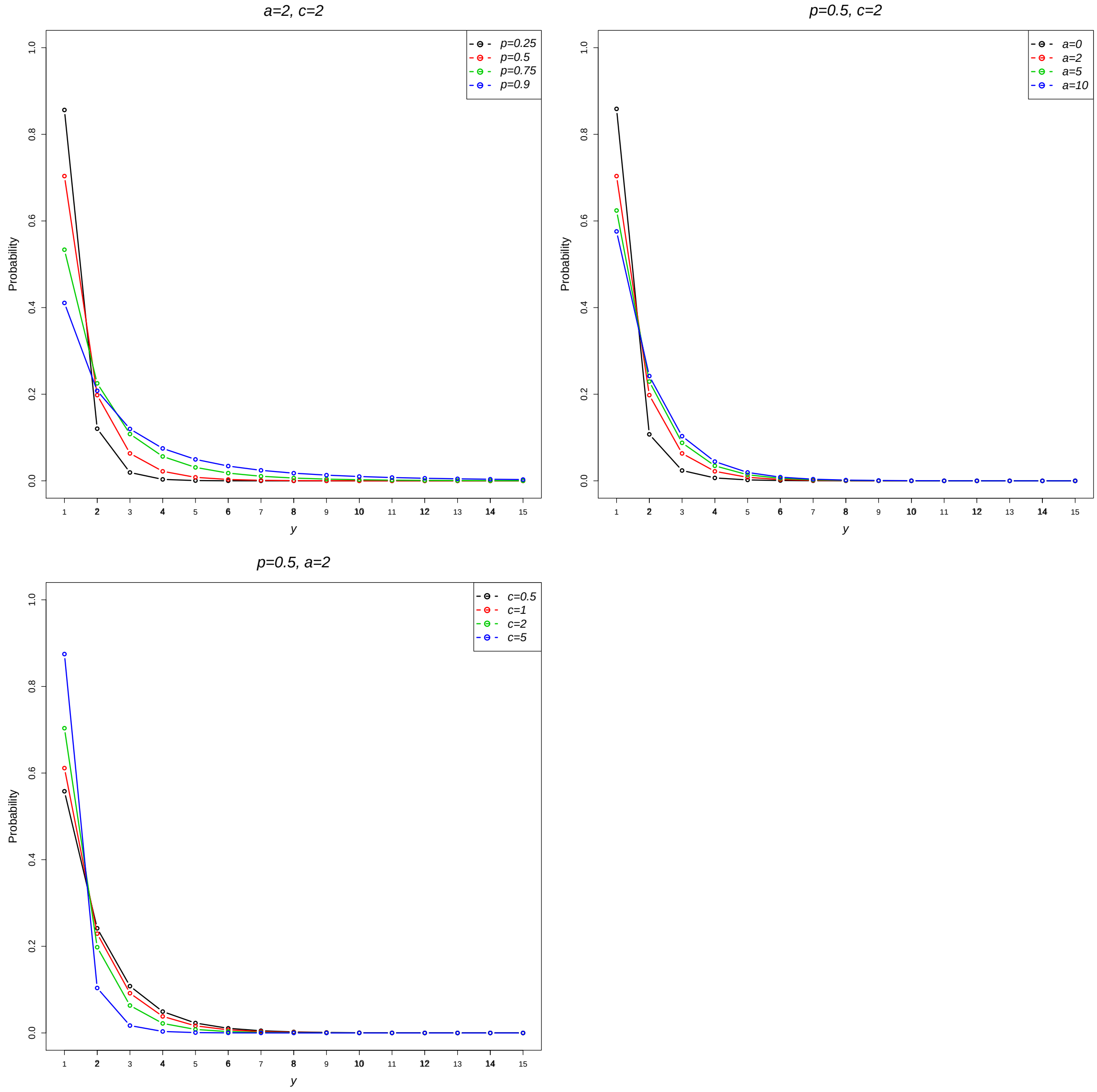


Figure 2.42: Lerch *pdf*

Figure 2.42 shows the Lerch probability density for $p = 0.25, 0.5, 0.75, 0.9$ where

$a = 2$ and $c = 2$ (first plot), $a = 0, 2, 5, 10$ where $p = 0.5$ and $c = 2$ (second plot) and $c = 0.5, 1, 2, 5$ where is $p = 0.5$ and $a = 2$. The parameter p controls the low counts of y and as p decreases the probability of a low y count increases. The parameter, a controls the overall skew of the distribution and as a becomes larger the skew increases in the distribution. The tail of the distribution is characterized by c and as c increases the tail becomes larger.

The *pgf* of the Lerch distribution is given by

$$G(t) = \frac{t\,\Phi(pt, c, a+1)}{\Phi(p, c, a+1)}\ , \tag{2.217}$$

with *mgf*

$$M(t) = \frac{\mathrm{e}^t\Phi(\mathrm{e}^t p, c, a+1)}{\Phi(p, c, a+1)}\ , \tag{2.218}$$

The mean of the Lerch distribution is

$$\mu = \frac{\Phi(p, c-1, a+1) - a\Phi(p, c, a+1)}{\Phi(p, c, a+1)}\ , \tag{2.219}$$

and the variance,

$$\sigma^2 = \frac{\Phi(p, c-2, a+1) - (\Phi(p, c-1, a+1) - a\Phi(p, c, a+1))^2}{\Phi(p, c, a+1)^2}\ . \tag{2.220}$$

with overdispersion index,

$$\begin{aligned} OD = \ & \frac{\Phi(p, c-2, a+1) - a\Phi(p, c-1, a+1)}{\Phi(p, c-1, a+1) - a\Phi(p, c, a+1)} + \\ & \frac{a\Phi(p, c, a+1) - \Phi(p, c-1, a+1)}{\Phi(p, c, a+1)} - a\ . \end{aligned} \tag{2.221}$$

As either p or a increase the OD index becomes large. When $c = 0$ the index is equal to 1 and decreases as c becomes large. The SI of the Lerch distribution is,

$$SI_y = \frac{p^{1-y}(a+y)^c\,\Phi(p, c, a+1)\,\Phi(p^2, 2c, a+1)}{\Phi(p, c, a+1)^2}\ , \tag{2.222}$$

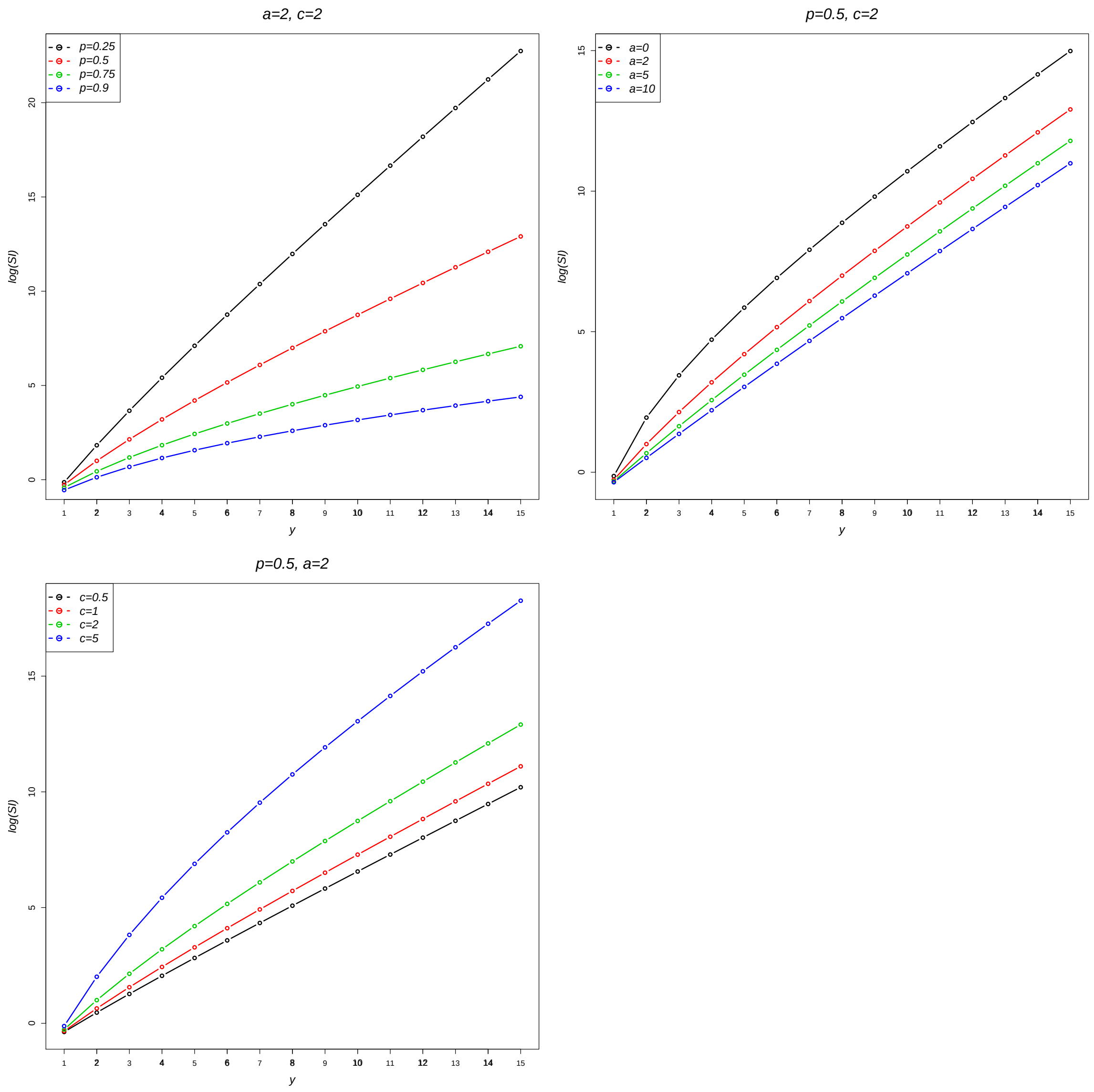


Figure 2.43: log(*SI*)'s for Lerch distributions

Figure 2.43 shows three plots of the logarithm of SI's for the Lerch distribution where $p = 0.25, 0.5, 0.75, 0.9$ with $a = 2$ and $c = 2$, $a = 0, 2, 5, 10$, when $p = 0.5$ and $c = 2$ and finally $c = 0.5, 1, 2, 5$ where $p = 0.5$ and $a = 2$. As p approaches 0 the SI increases for high Y values, and decreases for low Y values. Lower values of a result in an increase in the SI, whilst higher values of c increase the SI.

Several special cases of the Lerch distribution can be found by fixing the parameters of the Lerch distribution. Two examples where all three parameters in the Lerch distribution are fixed are the Estoup and Lotka distributions, presented in the following sections.

Estoup

This distribution was established by Estoup (1916) and is a special case of the Lerch distribution where $p = 1$, $a = 0$ and $c = 1$. It is sometimes known as the Estoup-Zipf law within the linguistics literature (Wimmer and Altmann, 1995, 1999). For the Estoup distribution, the Lerch distribution *pdf* reduces to

$$f_Y(y) = P(Y = y) = \frac{1}{S \times y} , \tag{2.223}$$

where $S = \sum_{y=1}^{n} \frac{1}{y}$, for $y = 1, 2, \ldots, n$. (Zörnig and Altmann, 1995; Wimmer and Altmann, 1999, P.145).

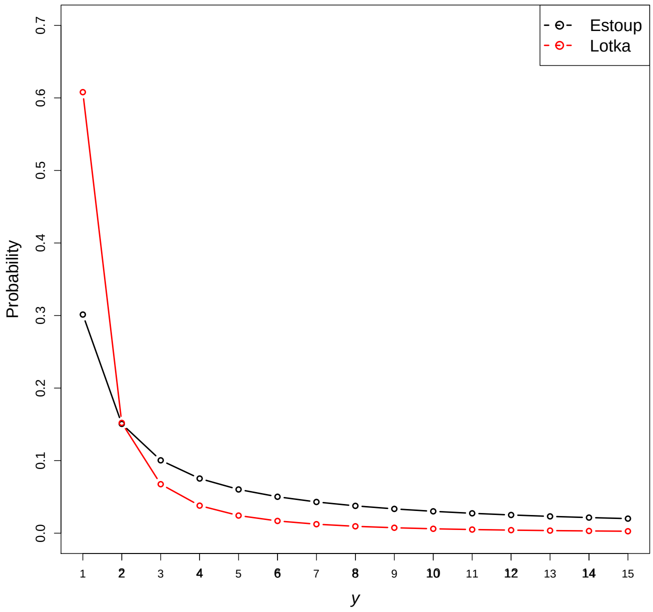


Figure 2.44: Estoup and Lotka *pdf*'s

Lotka

The Lotka distribution is another special case of the Lerch distribution where $p = 1$, $a = 0$ and $c = 2$ (Johnson et al., 2005). Also known as Lotka's Law after Lotka (1926) published his distribution for the frequency of scientific production based on the inverse square law. The *pdf* of this distribution is

$$f_Y(y) = P(Y = y) = \frac{1}{T \times y^2} , \tag{2.224}$$

for $y = 1, 2, \ldots, n.$ and where the corresponding Zeta function is $T = \Phi(1, 0, 2) = \frac{1}{6}\pi^2$ (Zörnig and Altmann, 1995; Wimmer and Altmann, 1999, P. 394).

Figure 2.44 plots the Estoup (shown in black) and Lotka (shown in red) densities for values of y of 1 to 15. Since these distributions have no parameters the densities are fixed. The Lotka distribution is more skewed than the Estoup, with a higher proportion of values of y of one. The Lotka distribution also has a smaller tail compared to the Estoup density.

2.6.2 Zipf (a, c)

This distribution is also often known as the Zipf-Mandelbrot distribution or, less frequently, as the Hurwitz distribution (Wimmer and Altmann, 1999). It has been applied to ranking problems in linguistics and in the analysis of publications citation frequencies (Zipf, 1949; Mandlebrot, 1959).This is a special case of the Lerch distribution, where p is a constant at one and $a > 0$ and $c > 1$,

$$f_Y(y; a, c) = P(Y = y) = \frac{1}{\Phi(1, a, c)\,(a + y)^c} , \tag{2.225}$$

for observations in the range $y = 1, 2, \ldots$, where $\Phi(p, a, c)$ is the the Zeta function (Wimmer and Altmann, 1999; Zörnig and Altmann, 1995, P.666).

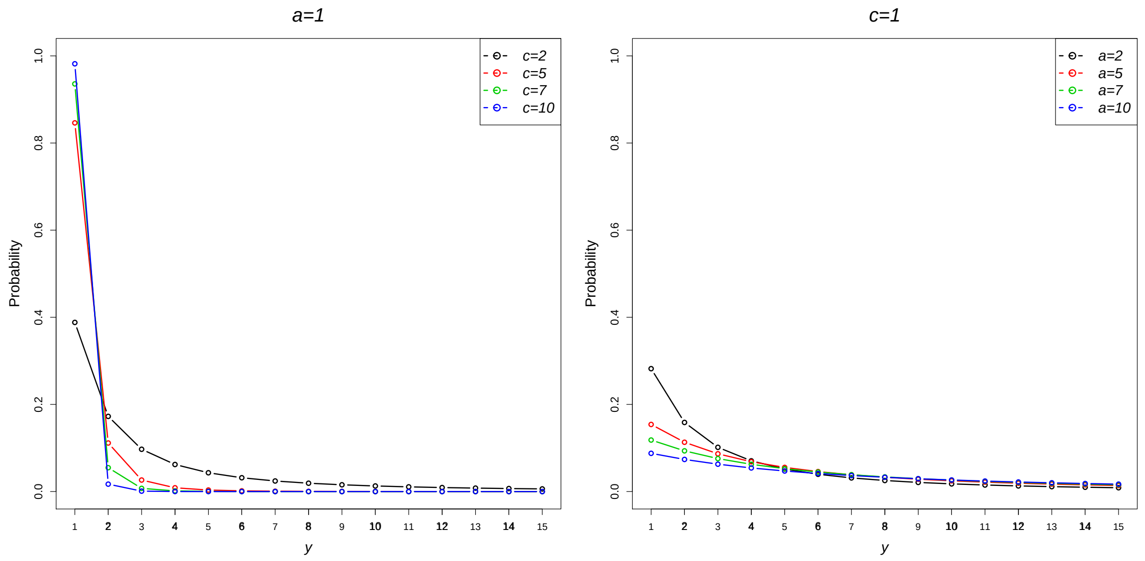


Figure 2.45: Zipf *pdf*

In Figure 2.45 the Zipf probability distribution is shown for values of c of 2,5,7,and 10 with a fixed at one and and a of 2,5,7,and 10 with $c = 1$. The parameter c controls the probability of the distribution where y equals one, whilst a controls the skewness of the distribution.

The *pgf* of this distribution is

$$G(t) = \frac{t\,\Phi(t, c, a + 1)}{\zeta(c, a + 1)} \;, \tag{2.226}$$

and the *mgf* is

$$M(t) = \frac{\mathrm{e}^t\,\Phi(\mathrm{e}^t, c, a + 1)}{\zeta(c, a + 1)} \;, \tag{2.227}$$

The mean is

$$\mu = \frac{\Phi(1, c - 1, a + 1) - a\Phi(1, c, a + 1)}{\zeta(c, a + 1)} \;, \tag{2.228}$$

and variance is given by

$$\begin{aligned} \sigma^2 = \; & \tfrac{1}{\zeta(c,a+1)^2} \big(-(\Phi(1, c - 1, a + 1) - a\Phi(1, c, a + 1))^2 + (\Phi(1, c - 2, a + 1) \\ & +a(-2\Phi(1, c - 1, a + 1) + a\Phi(1, c, a + 1)))\zeta(c, a + 1) \;\big) \end{aligned} . \tag{2.229}$$

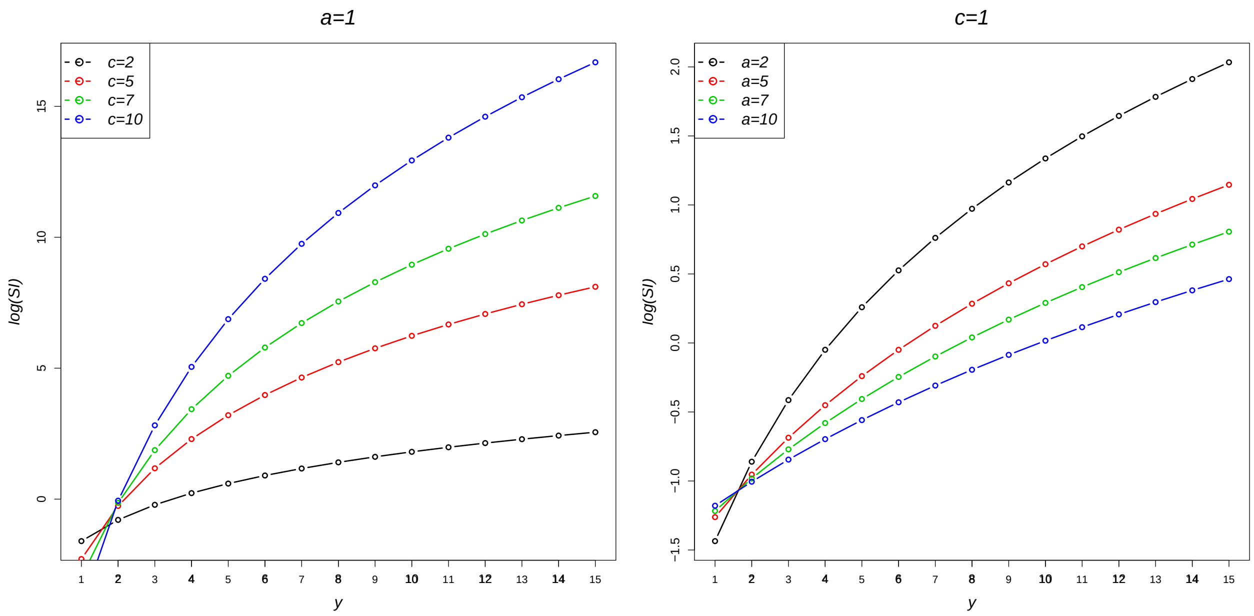


Figure 2.46: log(*SI*)’s for Zipf distributions

The overdispersion index is

$$OD = \frac{\Phi(1, c-2, a+1) - a\Phi(1, c-1, a+1)}{\Phi(1, c-1, a+1) - a\Phi(1, c, a+1)} + \frac{-\Phi(1, c-1, a+1) + a\Phi(1, c, a+1)}{\zeta(c, a+1)} - a\ . \tag{2.230}$$

Increasing values of a result in an increase in the OD index, however as c increases the OD index decreases. The SI of the Zipf distribution is,

$$SI_y = \frac{(a+y)^c \zeta(c, a+1)\zeta(2c, a+1)}{\zeta(c, a+1)^2}\ . \tag{2.231}$$

2.6.3 Good (p, c)

This distribution has been used in linguistics to model the distribution of word frequencies (Good, 1953), the size of business farms (Ijiri and Simon, 1977) and numbers of species per genus (Yule, 1925). The Good distribution arises where $0 < p < 1$, $a = 0$ and $c \in \mathbb{R}$ in the Lerch distribution with the resulting *pdf*

$$f_Y(y; p, c) = P(Y = y) = \frac{p^y}{\Phi(p, 0, c)\, y^c}\ , \tag{2.232}$$

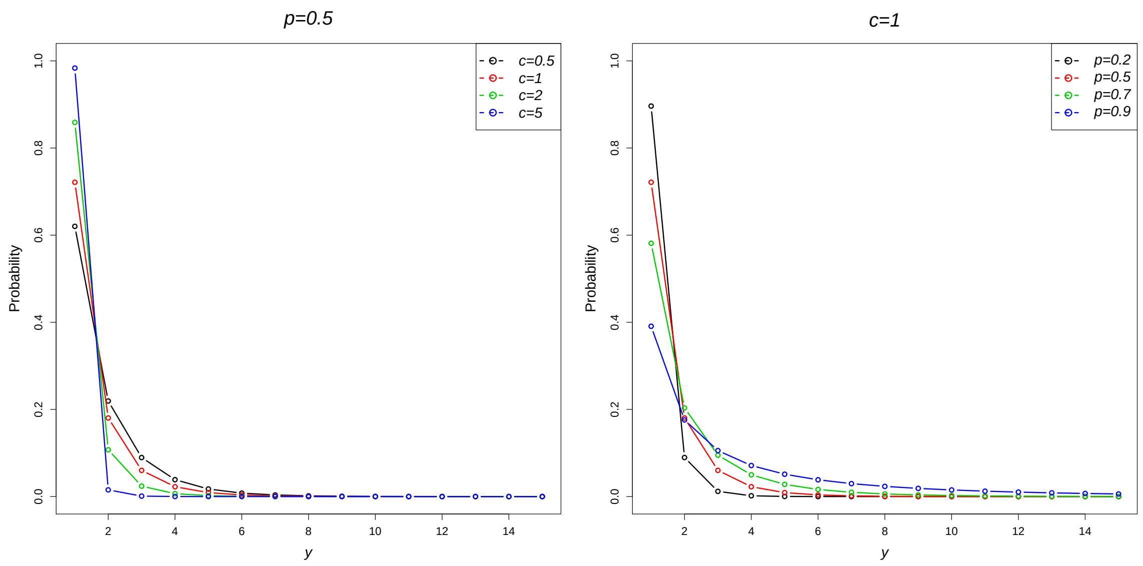


Figure 2.47: Good *pdf*

for $y = 1, 2, \ldots$ (Zörnig and Altmann, 1995; Wimmer and Altmann, 1999, pg. 219). The Good distribution is shown in Figure 2.47 for $c = 0.5, 1, 2, 5$ where $p = 0.5$ in the first plot, and $p = 0.2, 0.5, 0.7, 0.9$ where $c = 1$ in the second plot. As c increases the probability of a low value of y increases. The parameter p controls the tail of the distribution, with values of p closer to one having longer tails.

The Good distribution has *pgf*

$$G(t) = \frac{p^{-c}\mathrm{Li}_c(pt)}{\zeta(c)} \,, \tag{2.233}$$

where $Li_s(z)$ is the polylogarithm given by $Li_s(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^s}$ (Wimmer and Altmann, 1999). The *mgf* is,

$$M(t) = \frac{p^{-c}\mathrm{Li}_c(\mathrm{e}^t p)}{\zeta(c)} \,. \tag{2.234}$$

The mean of the Good distribution is

$$\mu = \frac{p^{-c}\mathrm{Li}_{c-1}(p)}{\zeta(c)} \,, \tag{2.235}$$

with variance

$$\sigma^2 = \frac{p^{-2c}\left(-\mathrm{Li}_{c-1}(p)^2 + p^c\mathrm{Li}_{c-2}(p)\zeta(c)\right)}{\zeta(c)^2} \,. \tag{2.236}$$

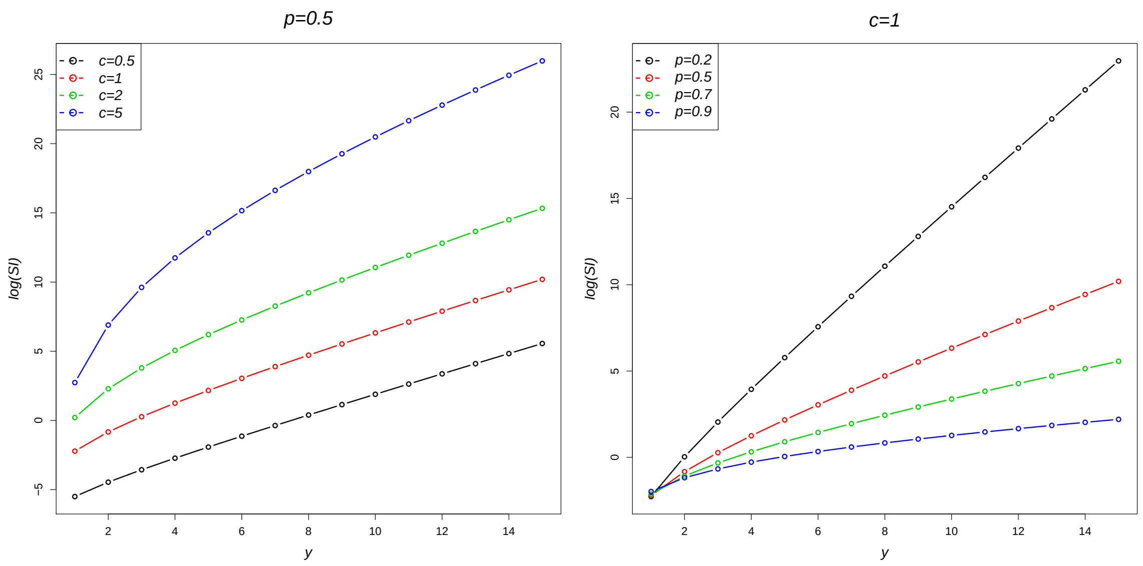


Figure 2.48: log(*SI*)'s for Good distributions

The overdispersion index is given by

$$OD = \frac{\mathrm{Li}_{c-2}(p)}{\mathrm{Li}_{c-1}(p)} - \frac{p^{-c}\mathrm{Li}_{c-1}(p)}{\zeta(c)} \; . \tag{2.237}$$

The OD increases as p approaches 1, but decreases as the value of c increases. The SI is

$$SI_y = p^{c-y}y^c \left(\sum_{i=1}^{\infty} \frac{p^{2i-c}y^{-c}(pi)^{-c}}{\zeta(c)^2} \right) \zeta(c) \; . \tag{2.238}$$

Figure 2.48 plots the log of the SI's for Good distributions with $c = 0.5, 1, 2, 5$ where $p = 0.5$ (first plot) and $p = 0.2, 0.5, 0.7, 0.9$ where $c = 1$ (second plot). As c increases the SI becomes more surprising across all values of Y plotted. Decreasing p reduces the SI for high values of Y.

2.6.4 Zeta (c)

The Zeta distribution is also known as the Reimann Zeta distribution or the discrete Pareto distribution and has been applied to the number of insurance policies by Seal

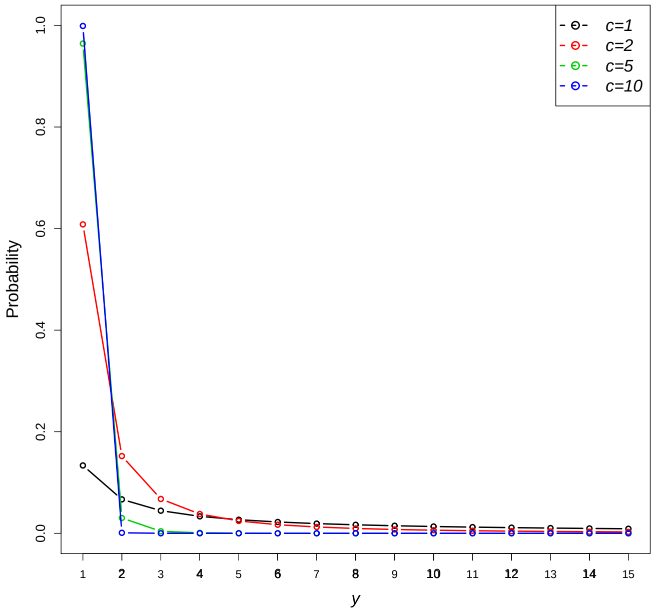


Figure 2.49: Zeta *pdf*

(1947). The *pdf* of the Zeta distribution is given by

$$f_Y(y;c) = P(Y = y) = \frac{1}{T\,y^c}\ , \tag{2.239}$$

for $y = 1, 2, \ldots$, where $T = \Phi(1, 0, c)$ (Zörnig and Altmann, 1995; Wimmer and Altmann, 1999, P. 664) and again, this distribution is a special case of the Lerch distribution where $p = 1$, $a = 0$ and $c > 1$. The Zeta distribution is also sometimes known as the discrete Pareto distribution, the Joos model, or the Riemann zeta distribution.

The zeta distribution is shown in Figure 2.49 for values of c of 1,2,5 and 10. As c increases the Zeta distribution becomes more J-shaped and the probability of a y value of one increases.

The *pgf* is given by

$$G(t) = \frac{\mathrm{Li}_c(t)}{\zeta(c)}\ , \tag{2.240}$$

and the *mgf* is

$$M(t) = \frac{\mathrm{Li}_c(\mathrm{e}^t)}{\zeta(c)}\ . \tag{2.241}$$

The mean and variance of the Zeta distribution are therefore

$$\mu = \frac{\zeta(c-1)}{\zeta(c)} \quad \text{and} \quad \sigma = \frac{\zeta(c-2)\zeta(c) - \zeta(c-1)^2}{\zeta(c)^2} \ . \tag{2.242}$$

The overdispersion index is

$$OD = \frac{\zeta(c-2)}{\zeta(c-1)} - \frac{\zeta(c-1)}{\zeta(c)} \ . \tag{2.243}$$

The SI of the Zeta distribution is

$$SI_y = \frac{y^c \zeta(2c)}{\zeta(c)} \ , \tag{2.244}$$

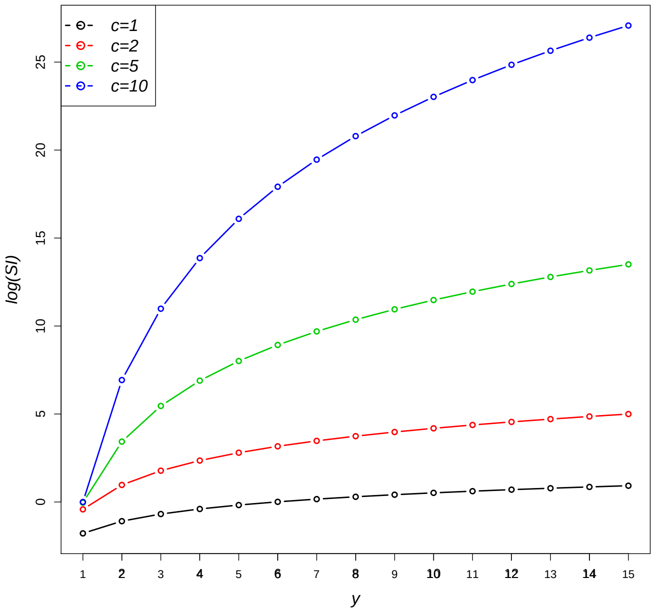


Figure 2.50: log(*SI*) for the Zeta distribution

Logarithms of the SI for the Zeta distribution with c values of 1, 2, 5 and 10, are plotted in Figure 2.50. Larger values of c result in larger SI's for high values of Y.

2.7 Generalized Poisson Distributions

The term Generalized distribution was coined by Feller (1943) to describe the combination of two independently distributed variables. Consider a random variable S that can be represented as a sum,

$$S = Y_1 + Y_2 + \ldots + Y_N \ , \tag{2.245}$$

where N and $Y_1, Y_2, \ldots$ are random variables, and the distribution of N has *pdf* f_N with *pgf* $G_N(t)$ and $\sum_i Y_i$ with *pdf* f_Y and *pgf* $G_Y(t)$. The variable S then has *pgf* $G_N(G_Y(t))$ and is called an f_N distribution generalized by f_Y (Gupta and Jain, 1974; Karlis and Xekalaki, 2005). This is represented using the symbolic notation developed by Feller (1943):

$$f_S = f_N \bigvee f_Y \ . \tag{2.246}$$

The model for this process can be interpreted as the sum of observations from f_Y, where the number of observations to be added is determined by an observation from the distribution f_N i.e. the sum from f_Y observations is stopped by the value of the f_N observation (Johnson et al., 2005, P.381). These distributions are also known by many other names in the statistical literature: compound, composed, stuttering, power series and stopped-sum distributions (Wimmer and Altmann, 1996).

Generalized Poisson distributions are a special case of generalized distributions and have been defined by Gupta and Jain (1974) and more recently by Wimmer and Altmann (1996) as a family of distributions with *pgf*'s of the form,

$$\begin{aligned} G(t) &= \exp\{\mu[G(t) - 1]\} \\ &= \exp\{a_1(t-1) + a_2(t^2-1) + \ldots + a_m(s^m - 1) + \ldots\} \end{aligned} \ , \tag{2.247}$$

where $G(t)$ is also a *pgf* and $\sum a_i = \mu$ (Gupta and Jain, 1974). This is a called the Generalized Poisson family of distributions due to the occurrence of the Poisson *pgf* $G(t) = \mathrm{e}^{\mu(t-1)}$ which is generalized as a compound distribution. For all distributions belonging to the Generalized Poisson family the *pgf* tends to the Poisson distribution as m becomes large (Gupta and Jain, 1974). The negative binomial distribution belongs to this family, where $\mu = -\ln(p)$ and $G(t) = \dfrac{\ln(1-(1-p)t)}{\ln(1-(1-p))}$ (Wimmer and Altmann, 1999).

This section presents five distributions from the Generalized Poisson family including the Neyman Type A, Hermite, Generalized Hermite, Gegenbauer and Generalized Gegenbauer distributions.

2.7.1 Neyman Type A (μ, ϕ)

First established by Jerzgi Neyman (1939) to describe numbers of larvae in a unit of a given area, the use of this model originates in the description of plant and insect distributions, especially when reproduction of the species produces clusters of offspring e.g. by seeds falling near the parent plant (David and Moore, 1954). The Neyman type A distribution can be generated as a generalized distribution, where a Poisson distribution is generalized by another Poisson distribution, i.e. Neyman type A(ϕ, μ) = Poisson(μ) $\bigvee$ Poisson(ϕ). For example, the number of plants follows a Po(μ) distribution and the number of offspring from each plant has a Po(ϕ) distribution. The Neyman Type A distribution is a member of the generalized Poisson family and its *pgf* has parameter μ. This distribution can also be constructed as a parameter mixture of a Poisson distribution with mean ϕ_j, where j varies across individuals according to a Poisson distribution with mean μ, i.e. Neyman type A(ϕ, μ) = Poisson(ϕj) $\bigwedge_j$ Poisson(μ). The *pdf* of this distribution has no closed form and can be written as:

$$f_Y(y; \mu, \phi) = P(Y = y) = \frac{\mathrm{e}^{-\mu}\phi^y}{y!} \sum_{j=0}^{\infty} \frac{(\mu \mathrm{e}^{-\phi})^j j^y}{j!} \ , \tag{2.248}$$

for $y = 0, 1, 2, \ldots$, where $\mu \geq 0$ and $\phi \geq 0$ (Johnson et al., 2005; Wimmer and Altmann, 1999, P. 468). The *pdf* of the Neyman Type A distribution can be seen in Figure 2.51. In the first plot ϕ is fixed at 2 and μ is in the range 1, 2, 5 and 10. As μ increases the distribution becomes almost flat and for small values of μ the distribution is highly skew. The second plot shows ϕ in 1, 2, 5 and 10, where μ is fixed at 2. The parameter ϕ adjusts the shape of the distribution with lower values of ϕ having higher

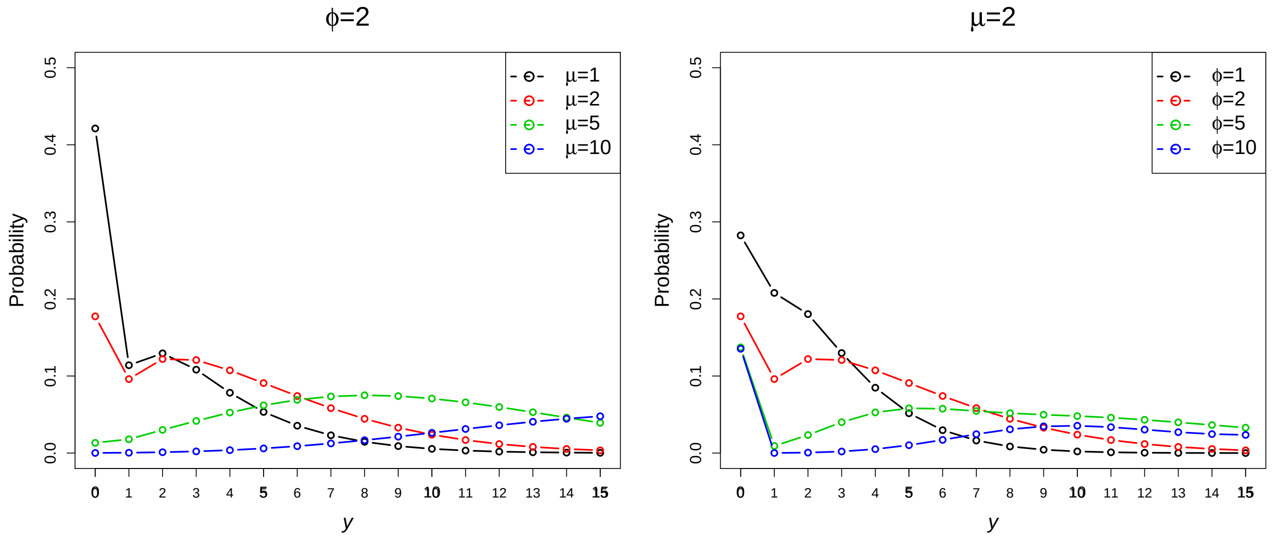


Figure 2.51: Neyman Type A *pdf*

probabilities of low y values.

The *pgf* is,

$$G(t) = \mathrm{e}^{\mu(\mathrm{e}^{(\phi(t-1))}-1)}, \tag{2.249}$$

and *mgf*,

$$M(t) = \mathrm{e}^{\mathrm{e}^{(\phi(t-1)}-1)\mu} \;. \tag{2.250}$$

The Neyman Type A distribution therefore has mean

$$\mu = \mu\phi \;, \tag{2.251}$$

and variance

$$\sigma^2 = \mu\,\phi(1+\phi) \;. \tag{2.252}$$

The overdispersion index is therefore

$$OD = 1 + \phi \;, \tag{2.253}$$

and is independent of μ, taking values greater than 1 and indicating that overdispersion can be accounted for by a Neyman Type A distribution. The zero-inflation index is,

$$ZI = 1 - \frac{\mathrm{e}^{-\phi} - 1}{\phi} \, . \tag{2.254}$$

Again, the ZI index is independent of μ and as ϕ increases the ZI index approaches 1. The SI for the Neyman Type A distribution is,

$$SI_y = \frac{\mathrm{e}^{\mu - \mathrm{e}^{-\phi}\mu} \phi^{-y} y! \sum_{y=0}^{\infty} \frac{\mathrm{e}^{2\left(\mathrm{e}^{-\phi}-1\right)\mu} \phi^{2y} Bl_y(\mathrm{e}^{-\phi}\mu)^2}{(y!)^2}}{Bl_y(\mathrm{e}^{-\phi}\mu)} \tag{2.255}$$

where Bl_y is the Bell polynomial.

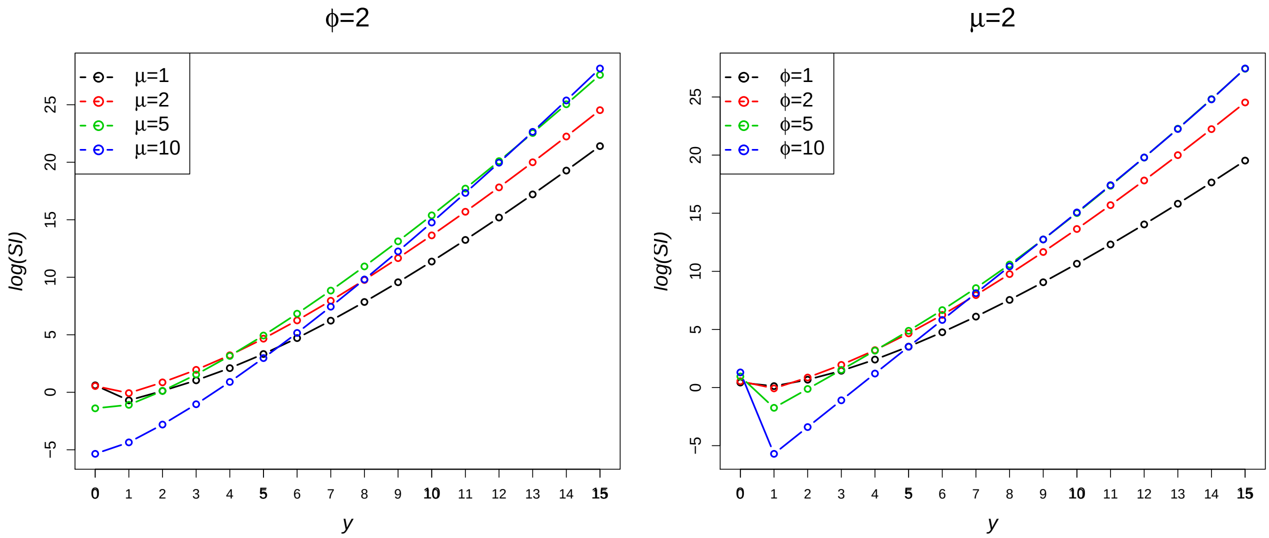


Figure 2.52: log(*SI*)'s for Neyman Type A distributions

The logarithm of the SI is plotted in Figure 2.52 for $\mu = 1, 2, 5$ and 10 where ϕ is fixed at 2 (first plot) and $\phi = 1, 2, 5$ and 10 with $\mu = 2$ (second plot). For increasingly large values of ϕ or μ the SI is larger for high values of Y.

2.7.2 Hermite (a, b)

The Hermite distribution was first derived by McKendrick (1926) as the sum of two correlated Poisson random variables and applied to counts of bacteria in leucocytes. Let the bivariate Poisson distribution equal $(Y_1, Y_2) = (U+V, U+W)$ where U, V and

W are three independent Poisson variables with parameters b, a_1 and a_2 respectively (Ahmed, 1961). Taking the sum $Y_1 + Y_2$ results in a Hermite *pdf* with parameters $a = a_1 + a_2$ and b,

$$f_Y(y; a, b) = P(Y = y) = \mathrm{e}^{-(a+b)} \sum_{j=0}^{\left\lceil \frac{y}{2} \right\rceil} \frac{a^{(y-2j)}\, b^j}{(y-2j)!j!} \ , \tag{2.256}$$

where $\lceil x \rceil$ is a Ceiling function giving the smallest integer greater than or equal to x, valid for $y = 0, 1, 2, \ldots$ $a \geq 0$ and $b \geq 0$ (Johnson et al., 2005; Wimmer and Altmann, 1999, P.254).

This distribution gets its name from the appearance of the Hermite polynomial in the *pdf*, setting $a = \alpha\beta$ and $b = \frac{\alpha^2}{2}$ in Equation 2.256 gives,

$$\begin{aligned} P(Y = 0) &= \mathrm{e}^{-\alpha\beta - \frac{\alpha^2}{2}} \\ P(Y = y) &= \frac{\alpha^y H_y(\beta)}{y!} P(Y = 0) \ , \quad y = 1, 2, \ldots \end{aligned} \tag{2.257}$$

where $H_y(\beta)$ is the Hermite polynomial (Johnson et al., 2005).

This is a generalized Poisson distribution where a Poisson distribution with mean $a + b$ is generalized by a zero-truncated Bernoulli distribution with probability $\frac{b}{a+b}$, i.e Hermite(a, b) = Poisson$(a + b)$ $\bigvee$ Zero-truncated Bernoulli $\left(\frac{b}{a+b}\right)$ and is a member of the generalized Poisson family where the parameter of the generalized Poisson distribution family *pgf* is $\mu = a + b$ (Wimmer and Altmann, 1999, P.254). This distribution is also known as a two-parameter Poisson distribution. The Hermite distribution can also be generated as a component mix of a Poisson distribution and Poisson doublet, where in a Poisson doublet distribution pairs (rather than individuals) follow a Poisson distribution with sample space $0, 2, 4, \ldots$, i.e. Hermite(a, b) =Poisson(a)*Poisson doublet(b) (Johnson et al., 2005). It is also a Binomial-Poisson parameter mix where Hermite(a, b) =Binomial$\left(2j, \frac{2b}{(a+2b)}\right) \bigwedge_j$ Poisson$\left(\frac{(a+2b)^2}{4b}\right)$ (Wimmer and Altmann, 1999).

The Hermite probability density function is plotted in Figure 2.53 firstly for values of a of 1, 2, 5 and 10 where $b = 2$ and in the second plot b of 1, 2, 5 and 10 where

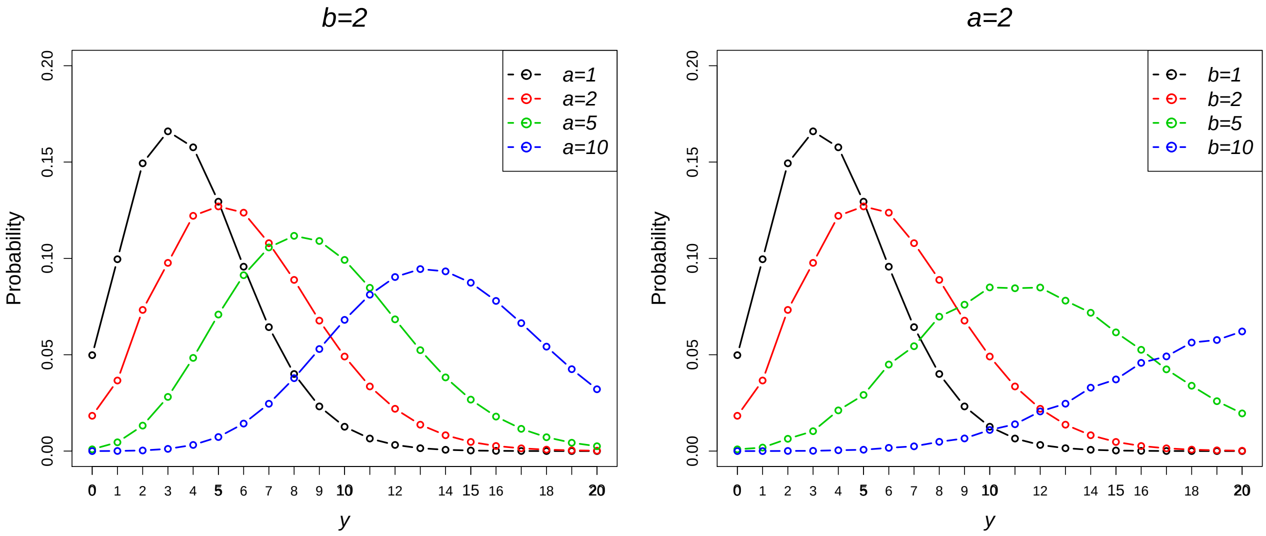


Figure 2.53: Hermite *pdf*

$a = 2$. As the values of a and b become larger the distribution tends to a normal curve.

The *pgf* of this distribution is

$$G(t) = e^{a(t-1)+b(t^2-1)} \ , \tag{2.258}$$

and the *mgf*,

$$M(t) = \mathrm{e}^{(\mathrm{e}^t-1)(a+b+b\mathrm{e}^t)} \ . \tag{2.259}$$

The mean and variance are given by

$$\mu = a + 2b \quad \text{and} \quad \sigma^2 = a + 4b \quad . \tag{2.260}$$

The overdispersion index for the Hermite distribution is

$$OD = \frac{a + 4b}{a + 2b} \ , \tag{2.261}$$

and where $b = 0$ and $a > 0$ the OD is equal to 1. As a increases the OD index increases but when b increases the index slowly decreases. The zero-inflation index is

$$ZI = 1 - \frac{(a + b)}{a + 2b} \ . \tag{2.262}$$

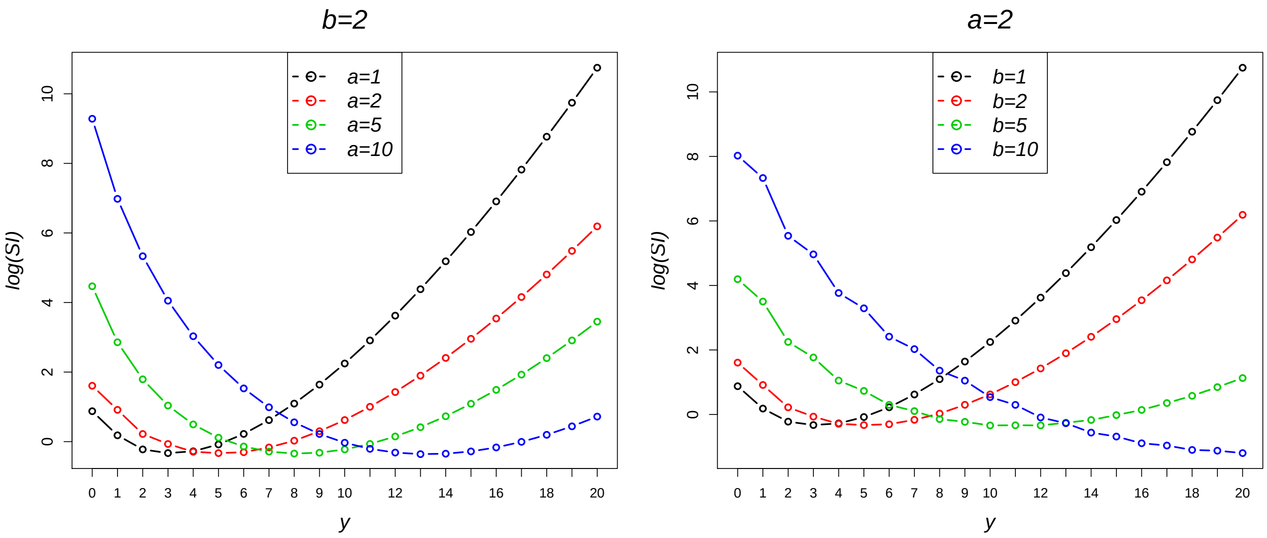


Figure 2.54: log(*SI*)'s for Hermite distributions

Large values of a result in smaller values of the ZI index, i.e. less zero-inflation in the dataset, whilst large values of b increase the ZI index. The SI of the Hermite distribution is,

$$SI_y = \frac{1}{U\left(\frac{1-y}{2},\frac{3}{2},-\frac{a^2}{4b}\right)} \left(2^{1-y}a^{-y}\left(-\frac{a^2}{b}\right)^{-\frac{1}{2}+\frac{y}{2}} \mathrm{e}^{a+b}y!\sum_{i=0}^{\infty}\left(\frac{1}{(i!)^2}\right)2^{2i-y}a^y\left(-\frac{a^2}{b}\right)^{\frac{1}{2}-\frac{y}{2}} b^{\frac{y}{2}}\mathrm{e}^{-2(a+b)}U\left(\frac{1-y}{2},\frac{3}{2},-\frac{a^2}{4b}\right)U\left(-\frac{y}{2},\frac{1}{2},-\frac{a^2}{4b}\right) \right), \tag{2.263}$$

where $U(a,b,x)$ is the confluent hypergeometric function of the second kind. The SI is plotted in Figure 2.54. The first plot illustrates the SI's of the Hermite distrbution where $a = 1, 2, 5, 10$ and $b = 2$. Where a is small the log(SI) is skew with high values of y having large SI's and as a increases there is a reduction in log(SI), with low y values eventually having the highest SI values. In the second plot, $a = 2$ and b is in the range 1, 2, 5 and 10. Again, where b is small, the SI is skew, with high y values having large SI's and as b increases the SI is less skew with low values becoming more surprising.

2.7.3 Generalized Hermite (a, b, m)

Gupta and Jain (1974) extended the Hermite distribution to form the Generalized

Hermite (GH) distribution with $Y = Y_1 + mY_2$, where $Y_1 = U + V$ and $Y_2 = U + W$ and U, V and W are independent Poisson random variables (Johnson et al., 2005, P.399). This distribution has been applied to the frequency of bacteria in leucoytes and frequency of larvae in corn plants by Cortina-Borja (2006). The *pdf* is,

$$f_Y(y; a, b, m) = P(Y = y) = \begin{cases} \mathrm{e}^{-(a+b)} & y = 0 \\ \mathrm{e}^{-(a+b)} \displaystyle\sum_{j=0}^{\left\lceil \frac{y}{m} \right\rceil} \frac{b^j}{j!} \frac{a^{y-mj}}{(y-mj)!} & y = 1, 2, 3, \dots \end{cases} \tag{2.264}$$

for $a \geq 0$, $b \geq 0$ and $m \in \mathbb{N}$ (Wimmer and Altmann, 1999, P.229). The distribution is also known as the Gupta-Jain-Hermite distribution after Gupta and Jain (1974).

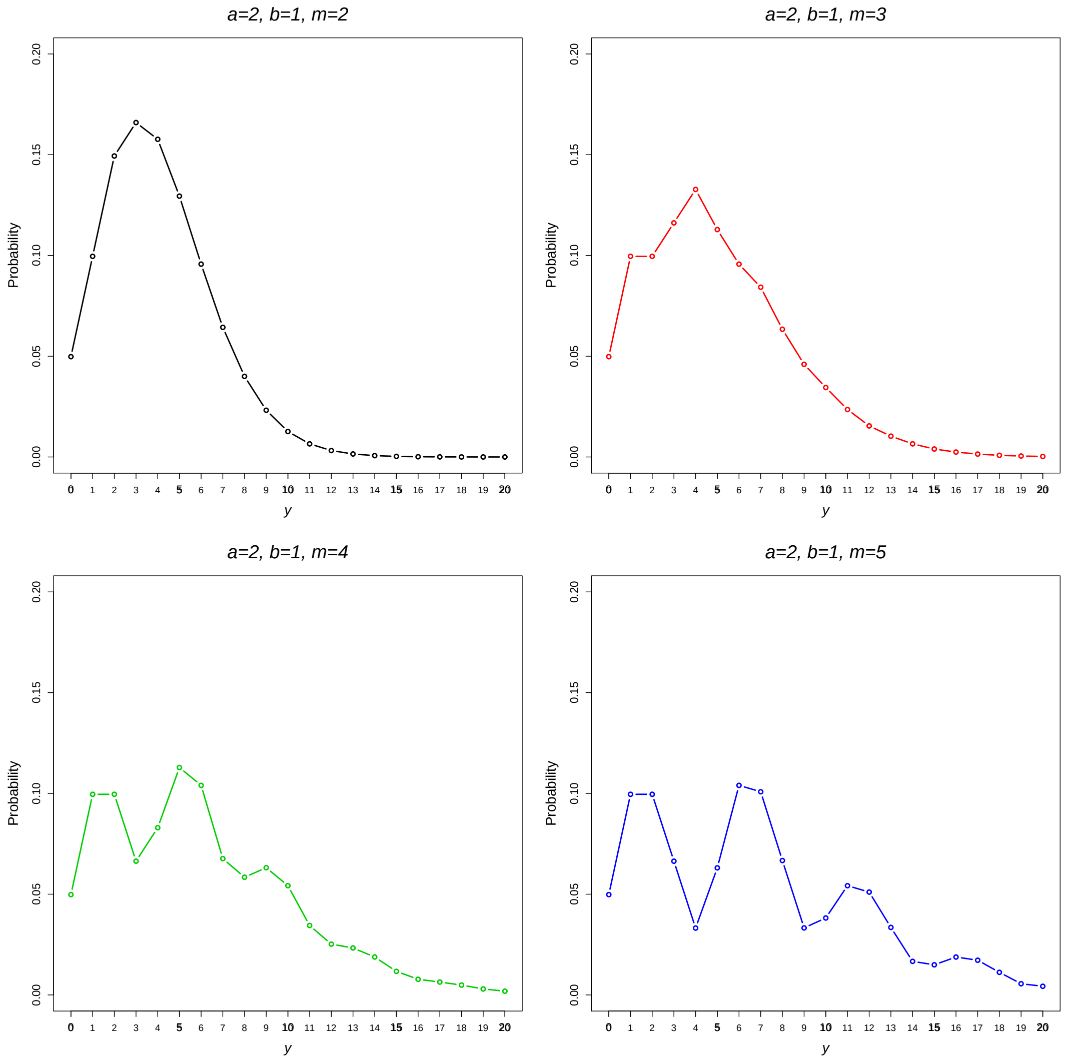


Figure 2.55: Generalized Hermite *pdf*

In the Generalized Hermite distribution m controls the number of modes in the density. Figure 2.55 plots the *pdf* of the Generalized Hermite distribution for values of m of 2, 3, 4 and 5, where a is fixed at 2 and b at 1. In the first plot $m = 2$ results in a uni-modal density, whilst a value of $m = 3$ results in the bimodal density displayed in the second plot. Examples of densities with 3 ($m = 4$) and 4 ($m = 4$) modes can be seen in the third and fourth plots.

This distribution is again a member of the generalized Poisson family where $\mu = a + b$ in the *pgf*

$$G(t) = \mathrm{e}^{-(a+b)}\mathrm{e}^{-at+bt^m} \ , \tag{2.265}$$

and the *mgf* is

$$M(t) = \mathrm{e}^{-a(1+t)-b(1-t^m)} \ . \tag{2.266}$$

The mean and variance are

$$\mu = a + m\,b \quad \text{and} \quad \sigma^2 = a + m^2\,b \ \ , \tag{2.267}$$

respectively. The overdispersion index can be calculated as

$$OD = \frac{a + b\,m^2}{a + m\,b} \ , \tag{2.268}$$

when $b = 0$, the index $OD = 0$ and where $a = 0$ the OD index is equal to m. The OD index is greater than 1 for all parameter values of a, b and m. As b increases the OD increases however when a increases the OD index decreases. The zero-inflation index is given by

$$ZI = \frac{b(m-1)}{a + mb} \ . \tag{2.269}$$

Increasing a results in a decrease in the ZI index, where as increasing b increases the amount of zero-inflation in the distribution. Larger values of m also result in a higher

ZI index. The SI is given by

$$SI_y = \frac{\sum_{y=0}^{\infty} \mathrm{e}^{-2(a+b)} \left(\mathrm{e}^{a+b} \left(\sum_{j=0}^{\lfloor \frac{y}{m} \rfloor} \frac{a^{y-jm} b^j}{j!(y-jm)!} \right) (U_{-y} - 1) - 1 \right)^2}{\mathrm{e}^{-(a+b)} + \left(\sum_{j=0}^{\lfloor \frac{y}{m} \rfloor} \frac{a^{y-jm} b^j}{j!(y-jm)!} \right) (1 - U_{-y})} . \tag{2.270}$$

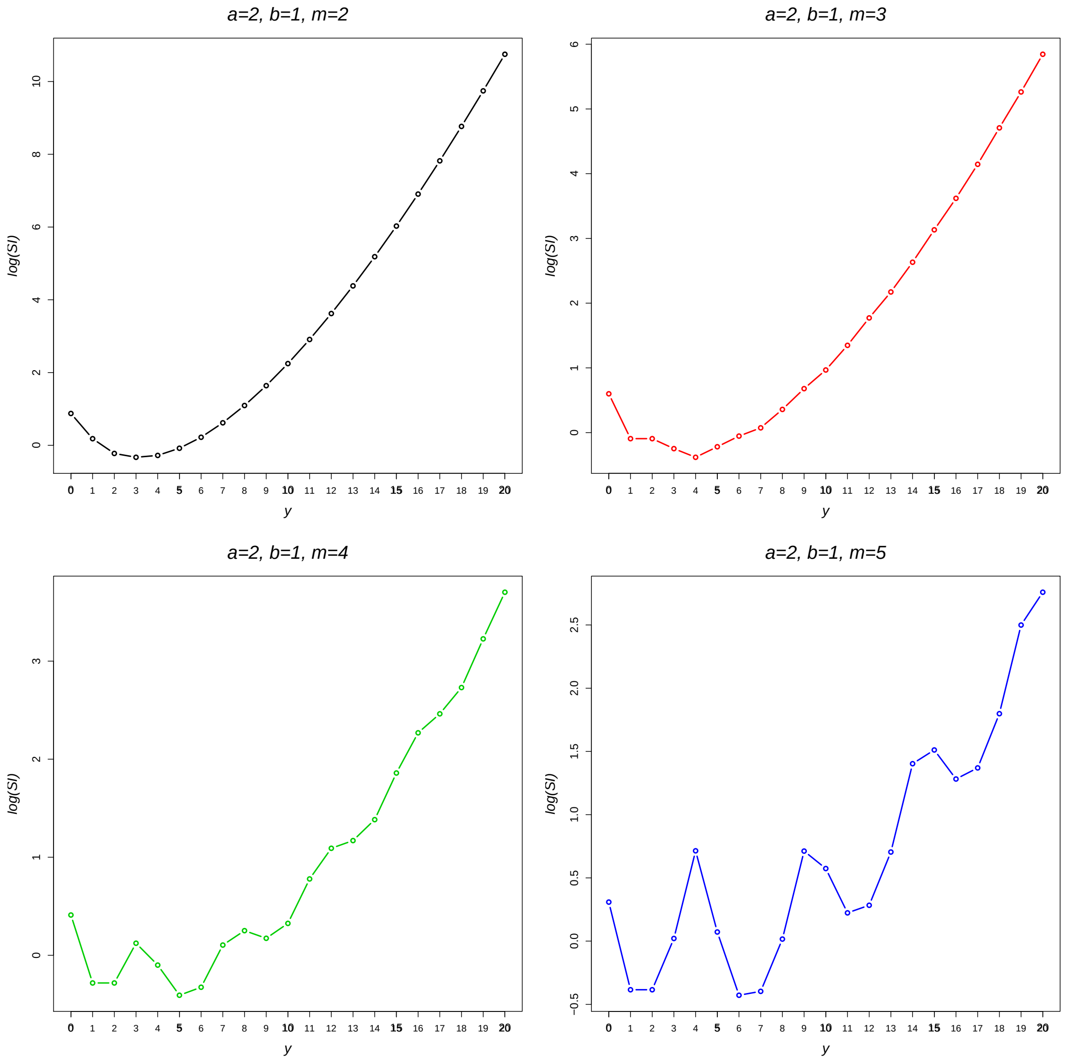


Figure 2.56: log(SI)'s for Generalized Hermite distributions

Figure 2.56 plots log(SI)'s for the Generalized Hermite distribution where $a = 2$ and $b = 1$ for values of m of 2, 3, 4 and 5. The SI's indicate that higher values of y are more surprising. As m increases the SI decreases with heavier tails and becomes more variable due to the multi-modal nature of the distribution.

2.7.4 Gegenbauer (a, b, k)

A parameter mixture of a Hermite and Gamma distributions results in a Gegenbauer distribution where Gegenbauer(a, b, k) = Hermite$(\theta, \frac{\theta a}{b}) \bigwedge_{\theta}$ Gamma$(\frac{a}{(1-a-b)}, k)$ (Wimmer and Altmann, 1999, P.176). The *pdf* of the Gegenbauer distribution is,

$$f_Y(y; a, b, k) = P(Y = y) = \begin{cases} (1-a-b)^k & y = 0 \\ (1-a-b)^k \sum_{j=0}^{[\frac{x}{2}]} \dfrac{b_j k^{(y-j)} a^{y-2j}}{j!\Gamma(y-2j+1)} & y = 1, 2, \ldots \end{cases}, \tag{2.271}$$

for $a \geq 0$, $b \geq 0$, $0 \leq a + b < 1$ and $k \geq 0$ (Plunkett and Jain, 1975; Johnson et al., 2005, P.500). This distribution is a member of the generalized Poisson family with $\mu = -k \ln(1 - a - b)$.

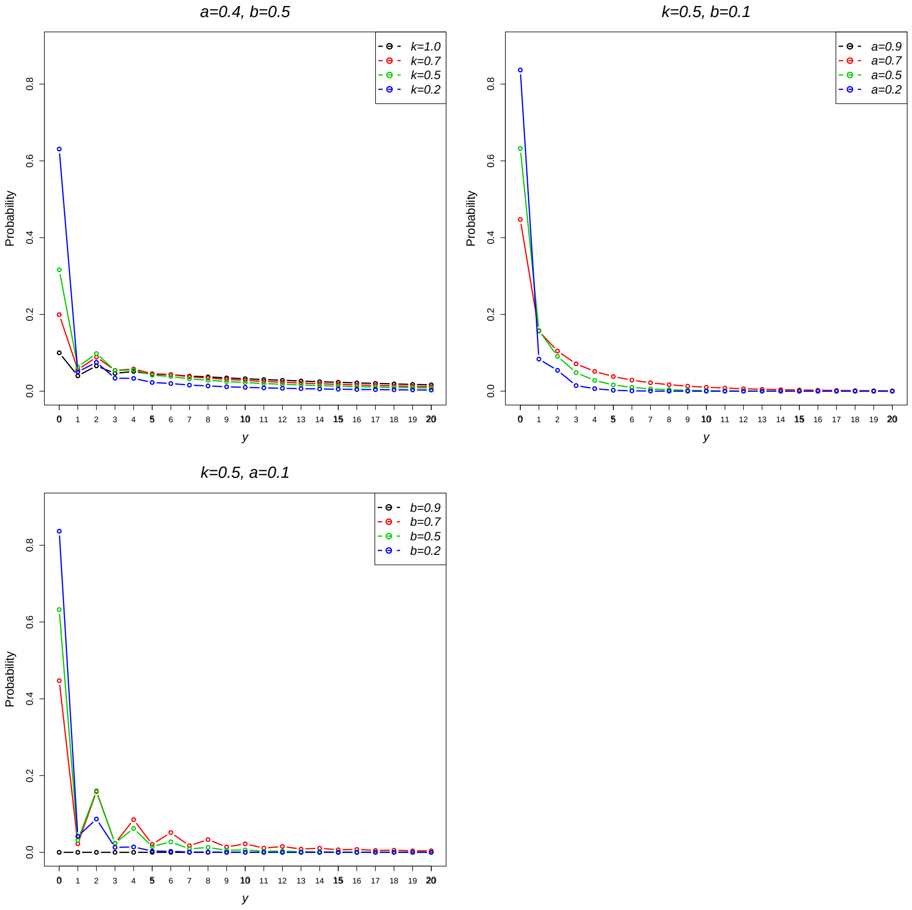


Figure 2.57: Gegenbauer *pdf*

The *pdf* of the Gegenbauer distribution is shown in Figure 2.57. In the first plot k

has values in the range 0.2, 0.5, 0.7 and 1 where $a = 0.4$ and $b = 0.5$ and it illustrates that for smaller values of k the probability of a y value of zero increases. In the second plot $k = 0.5$ and b is fixed at 0.1 whilst a ranges in 0.2, 0.5, 0.7 and 0.9 and in the third plot a is fixed at 0.1 whilst b ranges in 0.2, 0.5, 0.7 and 0.9. In each plot, as a or b decreases the distribution becomes more skew.

The *pgf* of this distribution is

$$G(t) = (1 - a - b)^k \, (1 - at - bt^2)^{(-k)} \, , \tag{2.272}$$

and *mgf*

$$M(t) = (1 - a - b)^k \, (1 - at - bt^{\mathrm{e}^t})^{-k} \, . \tag{2.273}$$

The mean of this distribution is

$$\mu = -\frac{k(a + 2b)}{a + b - 1} \, , \tag{2.274}$$

and the variance

$$\sigma^2 = \frac{k(a - (a - 4)b)}{(a + b - 1)^2} \, . \tag{2.275}$$

The overdispersion index is

$$OD = \frac{a - 2}{a + b - 1} - \frac{a}{a + 2b} \, . \tag{2.276}$$

where either a and b are large the OD index is also large. The zero-inflation index is

$$ZI = 1 - \frac{(a + b - 1)\log((1 - a - b)^k)}{(a + 2b)k} \, . \tag{2.277}$$

The parameters a, b and k all increase the ZI index which approaches a value of 1 as these values are large. The SI of this distribution is

$$
\begin{aligned}
SI_y = \; & -(\,(1-\alpha-\beta)^{-a}\sum_{y=0}^{\infty}(1-\alpha-\beta)^{2a}\left(-\sum_{j=0}^{\left\lfloor \frac{y}{2}\right\rfloor}\left[\frac{a^{y-2}\alpha^{y-2j}\beta^{j}(U_{-y}-1)}{j!\Gamma(1-2j+y)}\right]+U_{-y}\right) \\
& \left(-\sum_{j=0}^{\left\lfloor \frac{y}{2}\right\rfloor}\left[\frac{(a\alpha)^{y-2j}(a\beta)^{j}(U_{-y}-1)}{\Gamma(j+1)\Gamma(1-2j+y)}\right]+U_{-y}\right)\,)\,/ \\
& \left(\sum_{j=0}^{\left\lfloor \frac{y}{2}\right\rfloor}\left[\frac{a^{y-j}\alpha^{y-2j}\beta^{j}}{j!\Gamma(1-2j+y)}\right](U_{-y}-1)-U_{-y}\right)\;,
\end{aligned}
\tag{2.278}
$$

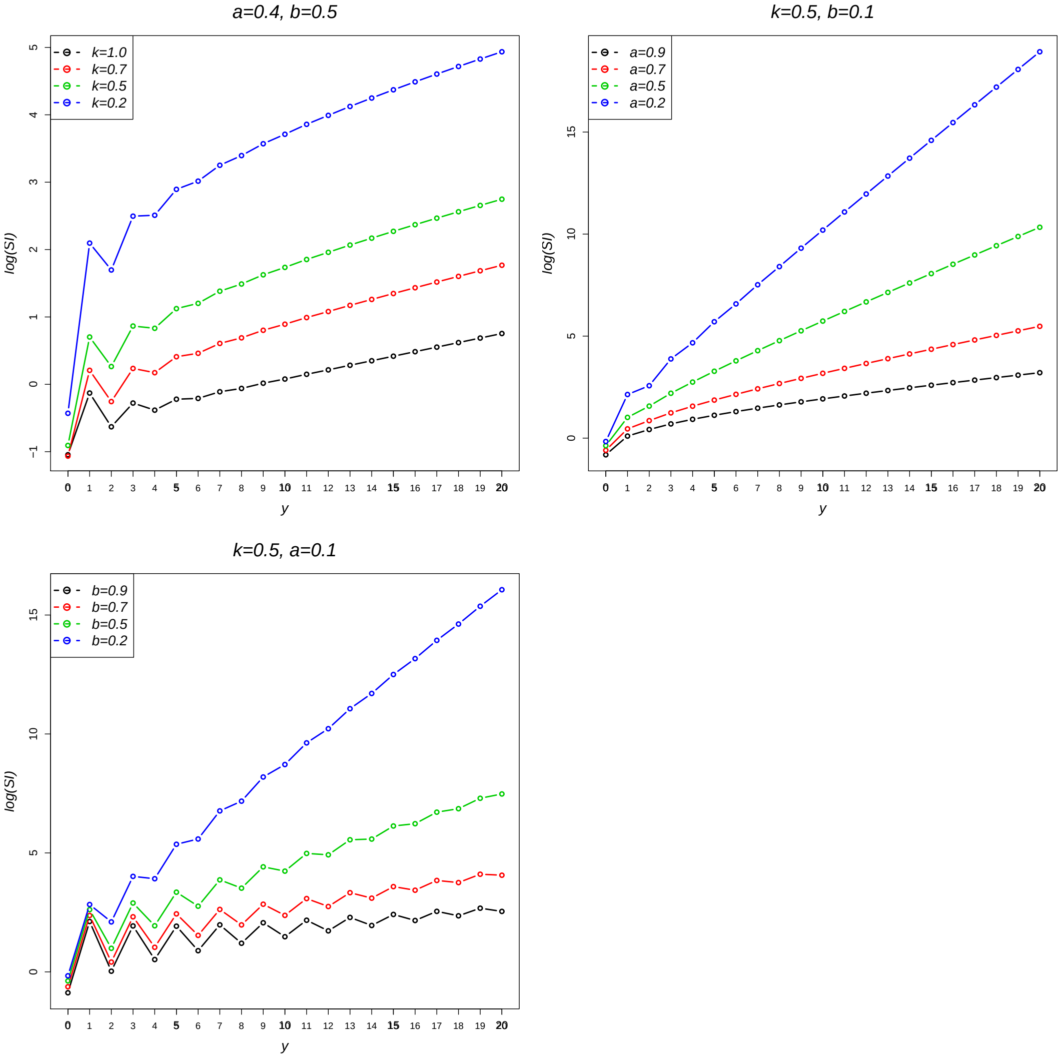


Figure 2.58: log(*SI*)'s for Gegenbauer distributions

Logarithm of SI's are plotted for the Gegenbauer distribuion in Figure 2.58. In the first plot $a = 0.4$, $b = 0.5$ and k is in the range 0.2, 0.5, 0.7 and 1.0, with larger values of k resulting in a higher SI. In the second plot $a = 0.9, 0.7, 0.6, 0.2$ where $b = 0.1$ and in the final plot $b = 0.9, 0.7, 0.6, 0.2$ where $a = 0.1$, with k fixed at 0.5. For smaller values of a or b in these plots the SI is higher for low values of Y.

2.7.5 Generalized Gegenbauer (a, m, α, β)

A generalization of the Gegenbauer distribution by Medhi and Borah (1984) has four parameters with *pdf*,

$$f_Y(y; a, m, \alpha, \beta) = P(Y = y) = \begin{cases} (1-\alpha-\beta)^a & y = 0 \\ \\ (1-\alpha-\beta)^a \displaystyle\sum_{j=0}^{[\frac{y}{m}]} \frac{a^{(y-(m-1)j)}\beta^j \alpha^{y-mj}}{j!\Gamma(y-mj+1)} & y = 1, 2, 3, \ldots \end{cases}, \tag{2.279}$$

for $a > 0$, $\alpha \geq 0$, $\beta \geq 0$, $0 \leq \alpha + \beta < 1$ and $m \in \mathbb{N}$ (Wimmer and Altmann, 1995, 1999, P.407). The density of the Generalized Gegenbauer distribution is plotted in Figure 2.59 for values $\alpha = 0.4$, $\beta = 0.5$ and $a = 0.5$, for four different values of m of 2, 3, 4 and 5. As m increases the number of modes in the distribution also increases and they become more pronounced for higher values of m.

This distribution is also known as the Medhi-Borah distribution and can be obtained by mixing a generalized Hermite distribution with a Gamma distribution (Wimmer and Altmann, 1999) and is a member of the generalized Poisson family where the parameter $\mu = -a\ln(1-\alpha-\beta)$ in the *pgf*:

$$G(t) = (1-\alpha-\beta)^a(1-\alpha t-\beta t^m)^{(-a)} \tag{2.280}$$

mgf:

$$M(t) = (1-\alpha-\beta)^a(1-\alpha \mathrm{e}^t-\beta \mathrm{e}^{tm})^{(-a)} \tag{2.281}$$

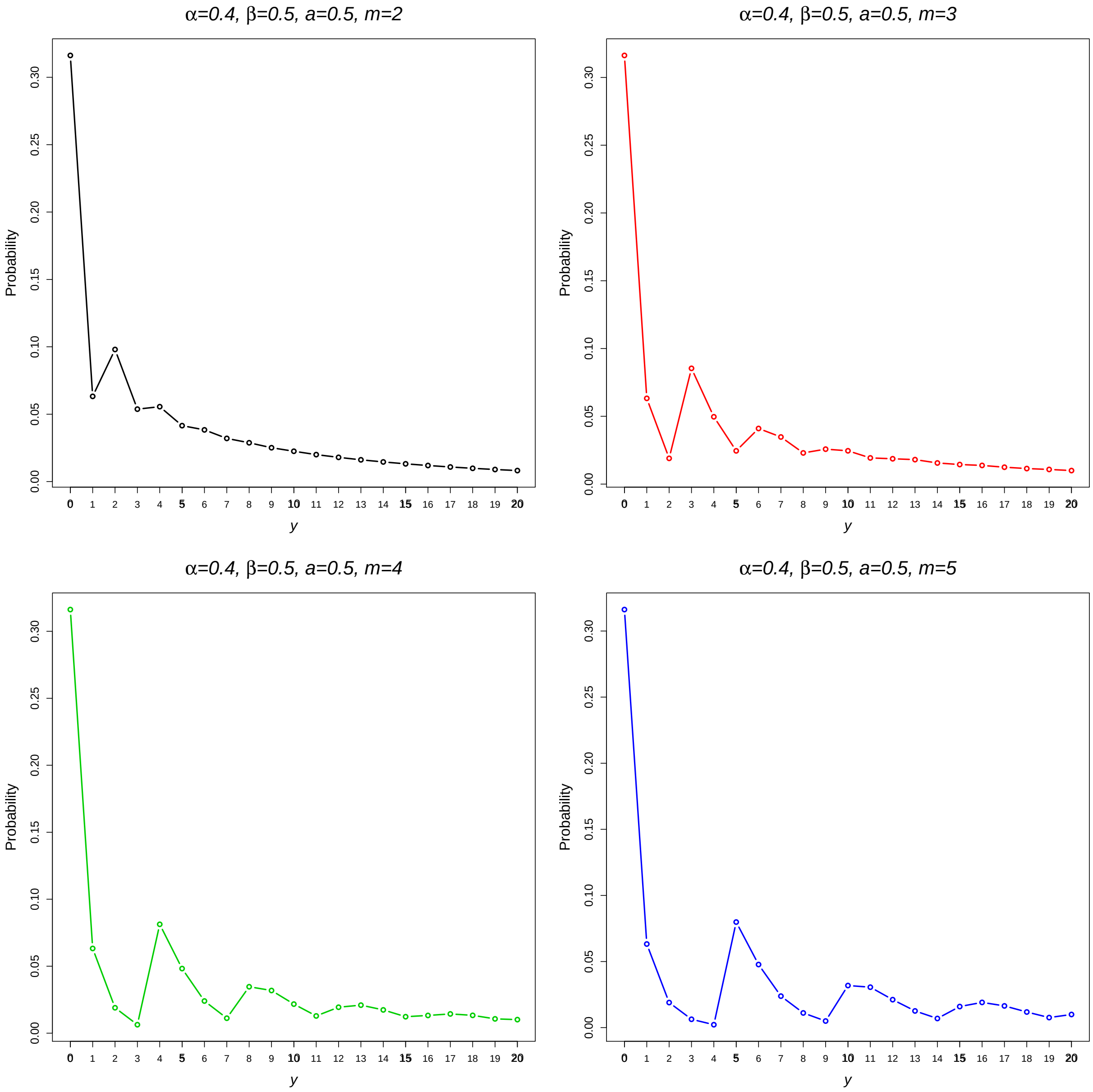


Figure 2.59: Generalized Gegenbauer *pdf*

and the mean and variance are,

$$\mu = -\frac{a(\alpha + m\beta)}{\alpha + \beta - 1} \quad \text{and} \quad \sigma^2 = \frac{a\left(\alpha + \beta\left(-\alpha(m-1)^2 + m^2\right)\right)}{(\alpha + \beta - 1)^2} \quad . \tag{2.282}$$

the overdispersion index is,

$$OD = -\frac{a(\alpha + (m^2 - (m-1)^2\alpha)\beta)}{(\alpha + \beta - 1)^2} \ , \tag{2.283}$$

Where α and β are close to 0 the OD index is near 1. As the values of α, β and m increase the OD index increases, indicating that the distribution becomes more dispersed. The zero-inflation index is

$$ZI = \frac{a(\alpha + m\beta) - (\alpha + \beta - 1)\log((1 - \alpha - \beta)^a}{a(\alpha + m\beta)} \ . \tag{2.284}$$

As m and α increase the ZI index increases, whilst larger values of a and β decreases the ZI index.

The SI is given by

$$\begin{aligned} SI_y = \ & \big(\textstyle\sum_{y=0}^{\infty}(1-\alpha-\beta)^{2a} \left(-\sum_{j=0}^{\lfloor \frac{y}{m} \rfloor} \frac{a^{-j(m-1)+y}\alpha^{y-jm}\beta^j}{j!\Gamma(1-jm+y)}(U_{-y}-1) + U_{-y} \right) \\ & \left(-\textstyle\sum_{j=0}^{\lfloor \frac{y}{m} \rfloor} \frac{(a\alpha)^{y-jm}(a\beta)^j}{\Gamma(j+1)\Gamma(1-jm+y)}(U_{-y}-1) + U_{-y} \right) \big) / \\ & \left((1-\alpha-\beta)^a \left(\textstyle\sum_{j=0}^{\lfloor \frac{y}{m} \rfloor} \frac{a^{-j(m-1)+y}\alpha^{y-jm}\beta^j}{j!\Gamma(1-jm+y)} \right) (1-U_{-y}) + (1-\alpha-\beta)^a U_{-y} \right) \ , \end{aligned} \tag{2.285}$$

Figure 2.60 plots four SI's for the Generalized Gegenbauer distribution with parameter values fixed at $\alpha = 0.4$, $\beta = 0.5$, $a = 0.5$ and where m is in the range 2, 3, 4 and 5. The number of modes in the distribution also is determined by m resulting in a variating SI and as m increases the size of the SI also increases. The Generalized Gegenbauer also has heavy tails in contrast with other *pdf*'s illustrated by the values of SI's in Figure 2.60 which are all less than 4 i.e. not surprising.

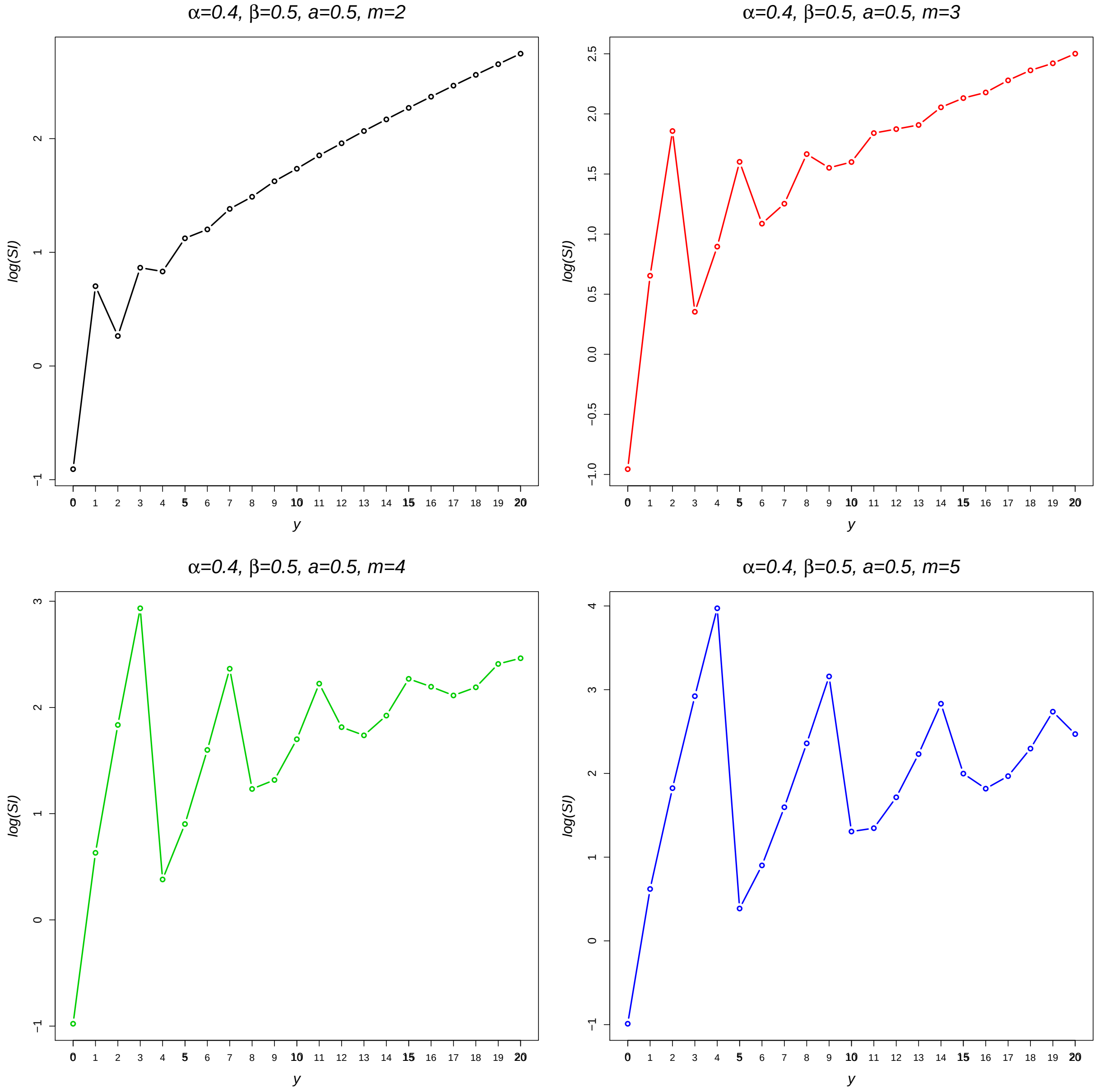


Figure 2.60: log(*SI*)'s for Generalized Gegenbauer distributions

Summary

The purpose of this chapter is to provide a basis for model fitting and introduces the distributions which will be referred to in forthcoming chapters. Common distributions for discrete data presented include the binomial, geometric, hypergeometric, Poisson, and negative binomial distributions, followed by alternatives to these distributions such as parameter-mixtures, component-mixtures and truncation to model highly skew, zero-inflated and/or long-tailed distributions. The Lerch family is a special class of distributions useful for modelling populations and word frequencies. The generalized Poisson family includes the generalized Hermite and generalized Gegenbauer distributions which allow fitting of multi-modal models.

Although many of these distributions have been previously covered in the statistical literature, by bringing this information together we hope to gain an overall understanding of discrete distributions, provide comparisons between distributions and identify suitable instances for their implementation in practice. An outline of each discrete distribution featured in this thesis has been given, including the *pgf*, *mgf*, mean and variance. Application of the surprise, overdispersion and zero-inflation provide new insights into the characteristics of these distributions.

Chapter 3

Fitting the models

This chapter describes estimation methods for fitting the discrete models detailed in the previous chapter. This is followed by a section on model diagnostics.

3.1 Estimation methods

The parameters of a discrete distribution can be estimated in a variety of ways. The methods of rapid estimation, maximum likelihood and the Estimation-Maximization (EM) algorithm are presented in this section. These methods are illustrated using the example of counts of cysts in steroid treated embryonic mouse kidneys presented in Section 1.2.2.

3.1.1 Rapid Estimation

Many methods of model fitting, for example maximum likelihood estimation (see Section 3.1), can be made easier if good initial estimators are obtained. Rapid estimation techniques, presented for discrete distributions by Kemp and Kemp (1988) provide an estimation method which can be used as initial estimates for iterative procedures (e.g. the Newton-Raphson method). These methods do not require iteration and are suitable where quick estimates of a model's parameters are needed.

Let Y be a discrete random variable with *pdf* f_Y and parameters $\underline{\theta} = \{\theta_1, \theta_2, \ldots, \theta_n\}$ and denote the mean $\mu \equiv \mu(\underline{\theta})$, variance $\sigma^2 \equiv \sigma^2(\underline{\theta})$, skewness $\gamma_1 \equiv \gamma_1(\underline{\theta})$ and kurtosis

$\gamma_2 \equiv \gamma_2(\underline{\theta})$. The process of rapid estimation works by equating functions of the sample observations to their expectations, producing equations that can solved simultaneously for the estimators $\underline{\theta}^*$. Three methods of rapid estimation are presented in the following sections; they use different estimating equations to generate parameter estimates.

Method of moments

The simplest example of this technique is the *method of moments* where the sample moments are equated to expressions of the moments for the distribution, giving the equations

$$\begin{aligned} \bar{y} &= \mu(\underline{\theta}^*) \\ s^2 &= \sigma^2(\underline{\theta}^*) \\ \gamma_1 &= \gamma_1(\underline{\theta}^*) \\ &\vdots \end{aligned} , \tag{3.1}$$

where $\bar{y}$, s^2, and γ_1, are the sample mean, variance and skewness of the observed count data y and $\mu(\underline{\theta}^*)$, $\sigma^2(\underline{\theta}^*)$ and $\gamma_1(\underline{\theta}^*)$ give expressions for the moments of the discrete distribution. Solving these equations results in expressions for the moment estimators $\underline{\theta}^*$.

The Poisson distribution presented in Section 2.2.5 illustrates this method using a simple one-parameter distribution. For the method of moments the mean of the Poisson distribution is simply equated to the sample mean of the observations. In the case of the Poisson distribution, the mean is equal to μ. The moment estimator μ^* is therefore given by the sample mean $\bar{y}$ of the data.

The parameters of a zero-inflated Poisson distribution ω and μ (see Section 2.4.1) can also be estimated using the method of moments. The sample mean $\bar{y}$ and variance s^2 of the discrete data y is set equal to the expressions for the mean and variance of the distribution given in Equations 2.138 and 2.139 of Section 2.4.1 as follows,

$$\begin{aligned} \bar{y} &= \mu - \mu\,\omega \\ s^2 &= \mu(1-\omega)(1+\mu\,\omega) \end{aligned} . \tag{3.2}$$

These are solved simultaneously to give moment estimators ω^* and μ^* for the parameters ω and μ, as follows

$$\begin{aligned} \mu^* &= \frac{\bar{y}^2 - \bar{y} + s^2}{\bar{y}} \\ \omega^* &= \frac{s^2 - \bar{y}}{\bar{y}^2 - \bar{y} + s^2} \end{aligned} \; . \tag{3.3}$$

where $\bar{y}$ and s^2 are the sample mean and variance of the y data, respectively.

Method of mean and zero frequency

The *method of mean and zero frequency* is another simple procedure, where the first estimating equation is the sample probability for the distribution at $y = 0$ denoted by f_0 which is equated to the *pdf* of the distribution at $P(Y = 0)$. Estimating equations for the remainder of the parameters in $\underline{\theta}^*$ are estimated using the moment equations,

$$\begin{aligned} f_0 &= P_0(\underline{\theta}^*) \\ \bar{y} &= \mu(\underline{\theta}^*) \\ \sigma^2 &= \sigma^2(\underline{\theta}^*) \\ &\vdots \end{aligned} \; . \tag{3.4}$$

For the Poisson distribution, the *pdf* at $P(Y = 0)$ is $e^{-\mu}$. The estimating equation can be constructed by equalling this to the probability of zero in the data as follows,

$$f_0 = \mathrm{e}^{-\mu} \; , \tag{3.5}$$

where f_0 is the probability of a zero value in the data y. This can be solved to give an estimate μ^*,

$$\mu^* = -\log(f_0) \tag{3.6}$$

as the rapid estimate for the parameter μ .

The zero-inflated Poisson distribution has *pdf* at $P(Y = 0)$ given by $\mathrm{e}^{-\mu}(1-\omega)+\omega$. Estimating equations for the method of mean and zero frequency for the zero-inflated

Poisson distribution are therefore

$$\begin{aligned} f_0 &= \mathrm{e}^{-\mu}(1-\omega)+\omega \\ \bar{y} &= \mu-\mu\,\omega \end{aligned} \quad . \tag{3.7}$$

However there is not a closed form from which to estimate μ^* and ω^* due to the inability to invert the $\mathrm{e}^{-\mu}$ term.

Empirical Probability Generating Function (EPGF) method

The previous two pairs of estimating equations are both special cases of a rapid estimation approach based upon the EPGF. The EPGF for a set of discrete data $\{Y_1, Y_2, \ldots, Y_n\}$ is,

$$G_n\left(t\right)=\frac{1}{n}\sum_{i=1}^{n}t^{Y_i}\ , \tag{3.8}$$

for $-1 \le t \le 1$. The method of EPGF estimation equates the EPGF to the *pgf* at selected values of t, resulting in a set of simultaneous equations for $\underline{\theta}^*$,

$$\mathrm{G_n}(t_i)=G(t_i)\ ,\quad i=1,2,\ldots,p \tag{3.9}$$

where p is the number of parameters and the choice of t_i is restricted to $-1 \le t_i \le 1$ (Kemp and Kemp, 1988). As $t_1 \to 1\,\forall\,p$ and $t_2 \to 1\,\forall\,p$ the equations are equal to the estimating equations for the method of moments. Similarly, for $p=2$, as $t_1 \to 1$ and $t_2 = 0$, the equations become equivalent to those of the *mean-and-zero-frequency* method.

Placing the *pgf* of the Poisson distribution (Equation 2.69 of Section 2.2.5) with $t = 0$ equal to the EPGF gives the following equation,

$$G_n(0)=\mathrm{e}^{-2\mu} \tag{3.10}$$

This equation can be solved to give the rapid estimate μ^*,

$$\mu^*=-\log\left(G_n\left(0\right)\right) \tag{3.11}$$

However, when $t = -1$ the estimating equation becomes

$$G_n(-1) = \mathrm{e}^{-2\mu} \tag{3.12}$$

which is solved to estimate μ^* as

$$\mu^* = -\frac{1}{2}\log(G_n(-1)) \tag{3.13}$$

For a zero-inflated Poisson distribution the *pgf* (Equation 2.136 of Section 2.4.1)is given by,

$$G(t) = -\mathrm{e}^{(t-1)\mu}(\omega - 1) + \omega \tag{3.14}$$

when t is set to 0, 1 and -1 these give three possible estimating equations,

$$\begin{aligned} G_n(0) &= -\mathrm{e}^{-\mu}(\omega - 1) + \omega \\ G_n(1) &= 1 \\ G_n(-1) &= -\mathrm{e}^{-2\mu}(\omega - 1) + \omega \end{aligned} \quad . \tag{3.15}$$

The first two equations in 3.15 cannot be used to estimate the parameters μ^* and ω^*, however using the first and third equations gives solutions,

$$\begin{aligned} \mu^* &= \log\left(\frac{1 - G_n(0)}{G_n(0) - G_n(-1)}\right) \\ \omega^* &= \frac{G_n(0)^2 - G_n(-1)}{2G_n(0) - G_n(-1) - 1} \end{aligned} \quad . \tag{3.16}$$

We can apply the example of the estimation of the parameter μ of the Poisson distribution using the example of counts of cysts in steroid treated embryonic mouse kidneys in Section 1.2. For the method of moments, the estimate is simply the mean, therefore $\mu^* = 1.55$. Using the method of mean and zero frequency, the probability of a zero count is $f_0 = 0.59$ giving an estimate of $\mu^* = 0.54$ using Equation 3.6. Finally, the EPGF at $t = 0$ is $G_n(0) = 0.59$ for this dataset and the formula in 3.13 results in an estimate of $\mu^* = 0.54$.

Under a zero-inflated Poisson distribution, moment estimators can be generated for counts of cysts in steroid treated embryonic mouse kidneys using Equation 3.3 where the mean number of cysts is 1.55 and the variance is 8.85, resulting in estimates $\mu^* = 4.66$ and $\omega^* = 0.67$. The EPGF method requires the EPGF where $t = 0$ and $t = -1$, i.e. $G_n(0) = 0.59$ and $G_n(-1) = 0.50$ and the estimates calculated using Equation 3.16 to give parameter estimates $\mu^* = 1.53$ and $\omega^* = 0.47$.

The advantage of this method is that it often provides quick estimates of a model's parameters. However, Kemp and Kemp (1988) provide examples of distributions where rapid estimation methods do not always have explicit solutions (particular cases are the negative binomial, Hermite, zero-inflated Poisson and zero-truncated Poisson distributions) and clearly illustrate that no single method of rapid estimation can be applied to all distributions. The example of a Poisson distribution fitted to the number of cysts in steroid treated kidneys shows how different methods of rapid estimation result in varying estimates. Standard errors of parameter estimates also cannot be calculated using rapid estimation methods.

3.1.2 Maximum Likelihood

The method of *maximum likelihood* is commonly used for estimating a model's parameters. If the observed values of the random variables $Y_1, Y_2, \ldots, Y_n$ are $y_1, y_2, \ldots, y_n$, their likelihood is given by

$$\mathscr{L}(\underline{\theta}|y_1, y_2, \ldots, y_n) = P\left[\bigcap_{j=1}^{n} Y_j = y_j|\theta_1, \theta_2, \ldots, \theta_n\right] , \tag{3.17}$$

for discrete distributions, where $\theta_1, \theta_2, \ldots, \theta_n$ are the model's parameters (Rose and Smith, 2002; Johnson et al., 2005, P. 68). If $Y_1, Y_2, \ldots, Y_n$ are mutually independent and have identical distributions, then the joint *pdf* is,

$$\mathscr{L}(\underline{\theta}|y_1, \ldots, y_n) = f_{1,\ldots,n}(y_1, \ldots, y_n; \underline{\theta}) = \prod_{j=1}^{n} f(y_i; \underline{\theta}) . \tag{3.18}$$

In practice it is often more convenient to work with the logarithm of the likelihood, the *log-likelihood*,

$$\ell = \log \mathscr{L}(\underline{\theta}|y_1, \ldots, y_n) = \sum_{j=1}^{n} \log f(y_1|\underline{\theta}) \ . \tag{3.19}$$

The method of maximum likelihood estimates $\hat{\theta}$ by finding a value of θ that maximizes $\hat{\ell}(\theta|y)$. The values $\hat{\theta_1}, \hat{\theta_2}, \ldots, \hat{\theta_n}$ that maximize the likelihood are called *maximum likelihood estimators* (MLE's). Maximizing the likelihood can be achieved by solving the equations,

$$\frac{\partial \mathscr{L}(y_1, y_2, \ldots, y_n|\theta_1, \theta_2, \ldots, \theta_p)}{\partial \theta_p} = 0 \ , \tag{3.20}$$

called *maximum likelihood equations*. In practice, maximizing the likelihood is equivalent to minimizing the negative likelihood. For many models, a maximum likelihood estimator can be found as an explicit function of the observed data $y_1, \ldots, y_n$. However, often the solutions to these equations are intractable and require iterative procedures (e.g. the Newton-Raphson algorithm) to reach a solution.

Again, the Poisson distribution can be used to illustrate this method of parameter estimation. The log-likelihood of the Poisson distribution is calculated using Equation 3.19, resulting in,

$$\ell(\mu|y_1, \ldots, y_n) = -n\,\mu - \sum_{i=1}^{n} \log\left(y_i!\right) + \log(\mu) \sum_{i=1}^{n} y_i \ . \tag{3.21}$$

Differentiating the log-likelihood with respect to the parameter μ gives,

$$\frac{\partial \ell(\mu|y_1, \ldots, y_n)}{\partial \mu} = -n + \frac{\sum_{i=1}^{n} y_i}{\mu} \ , \tag{3.22}$$

and setting this derivative equal to zero and solving for μ results in the MLE,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.23}$$

which is equal to the mean and is the same as the estimate from the method of moments in Section 3.1.1.

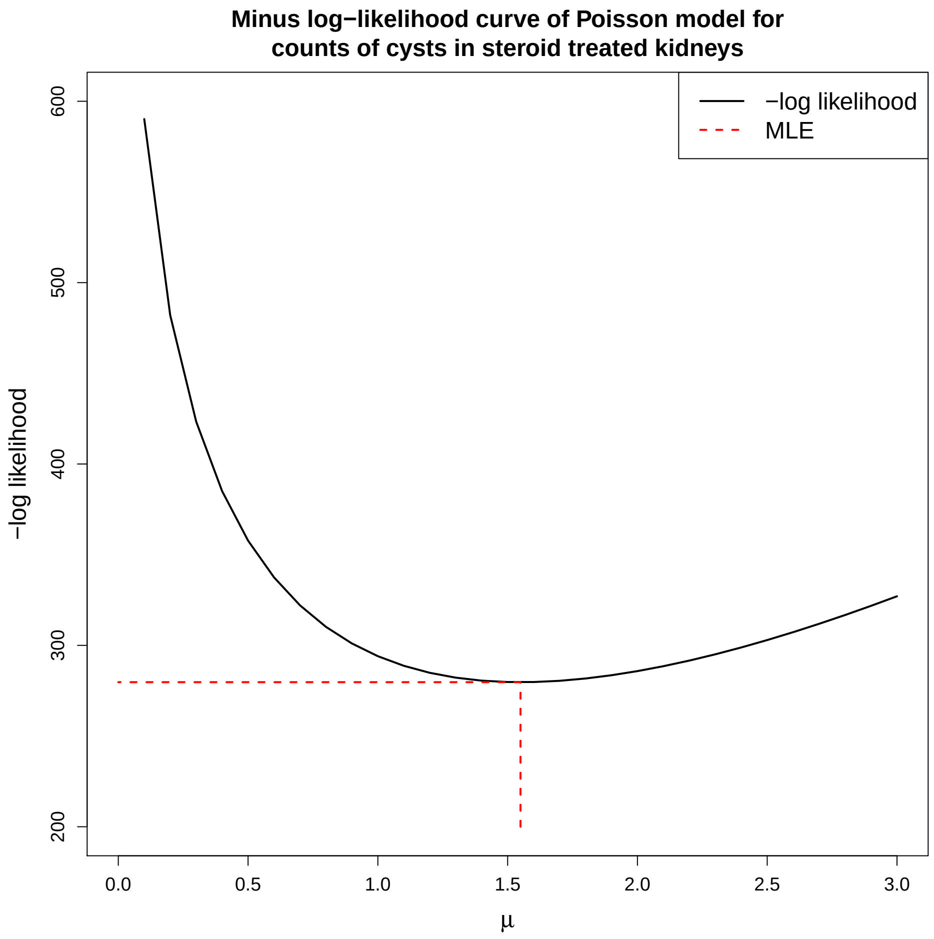


Figure 3.1: Minus log-likelihood curve of Poisson model for counts of cysts in steroid treated kidneys

For the example of counts of cysts in steroid treated embryonic mouse kidneys, Figure 3.1 plots the negative observed log-likelihood (shown in black) of the Poisson distribution for μ in the range 0 to 3. The log-likelihood is minimized at a value of -279.70 (shown by the red line), which corresponds to an estimated $\hat{\mu}$ of 1.55. The curvature of $\ell(\theta)$ gives an indication of $\hat{\theta}$'s precision. If $\ell(\hat{\theta})$ is flat then a lot of possible values are feasible, however if the curve is concentrated around $\ell(\hat{\theta})$ then $\hat{\theta}$ is well defined i.e. a precise estimate. The negative log-likelihood curve for the Poisson model gives similar values of $\ell(\mu)$ around $\hat{\mu}$.

A negative binomial distribution can also be fitted to counts of cysts in steroid treated embryonic mouse kidneys. The negative observed log-likelihood for the negative binomial distribution is plotted as a contour in Figure 3.2 for values of both r and p

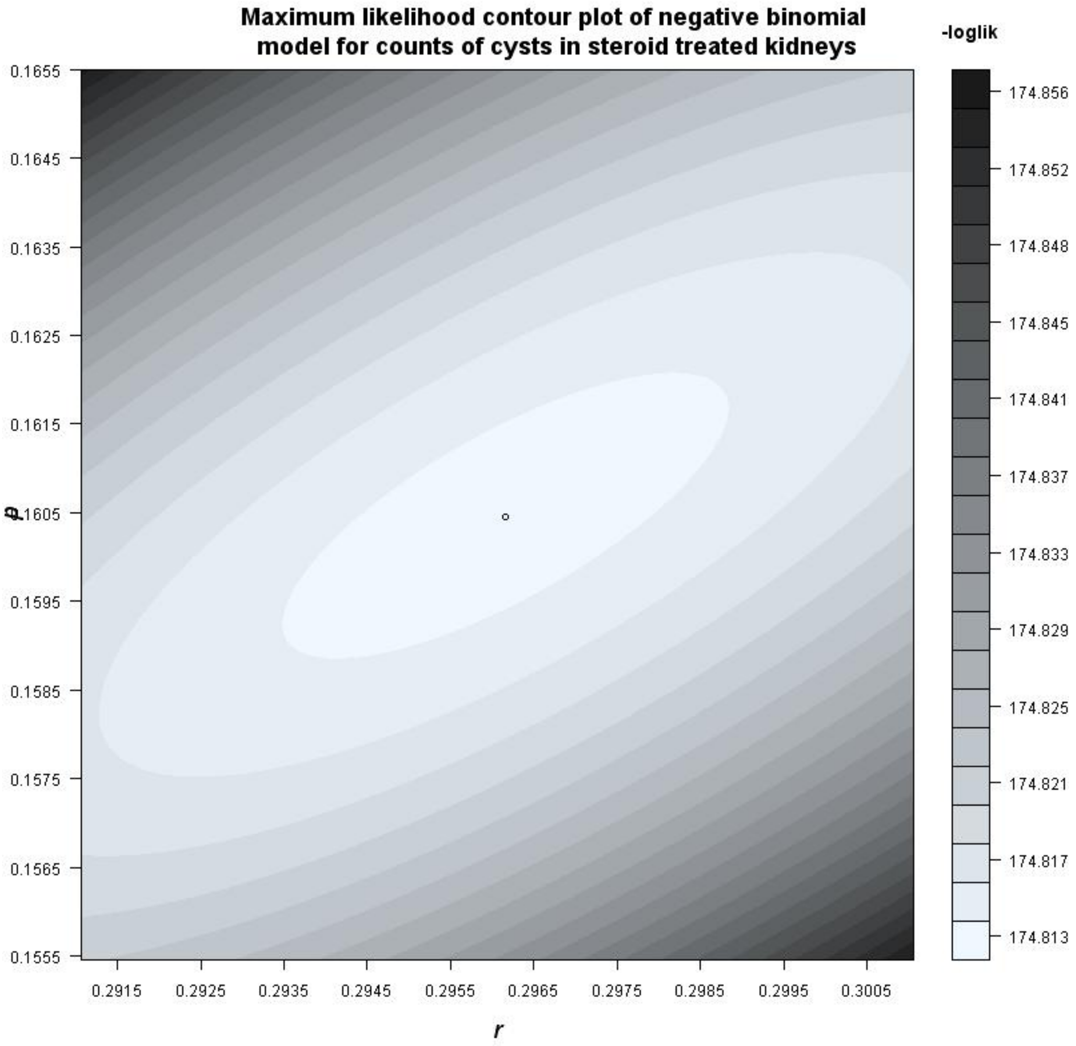


Figure 3.2: Maximum likelihood curve of negative binomial model for counts of cysts in steroid treated kidneys

between 0 and 1. The minimum of this function is indicated on the plot, having a negative log-likelihood of 174.81 resulting in values of $\hat{r} = 0.30$ and $\hat{p} = 0.16$. The negative binomial contour plot of $\ell(\hat{\theta})$ is flat for values around $\hat{r}$ and $\hat{p}$ indicating that a range of possible values are feasible.

The advantage of the maximum likelihood method of estimation is that it provides a consistent approach to parameter estimation and therefore MLE's can be developed for a variety of models. Asymptotically, maximum likelihood estimates are unbiased have approximate normal distributions and their approximate sample variance can be used to generate confidence intervals and hypotheses tests (Johnson et al., 2005). Approximate standard errors for $\hat{\theta}$ can be calculated using the inverse of Fisher's Information matrix, $V(\hat{\theta}) = [I(\theta)]^{-1}$, where the information matrix is the negative of the expected value of the Hessian $I(\theta) = -E\left[\dfrac{\partial \ell(\theta)}{\partial\theta\partial\theta'}\right]$ (Rose and Smith, 2002). These methods are also widely available across statistical software packages (Johnson et al., 2005).

A disadvantage of maximum likelihood estimation is that likelihood equations often need to be numerically optimized, for example using the Newton-Raphson algorithm, where analytically expressions for estimates are not available and this may be difficult (Rose and Smith, 2002). Maximum likelihood estimates may be sensitive to the starting values used in the numerical optimization; poor starting values may result in non-convergence or incorrectly optimising to a local mimimum/maximum instead of the global minimum/maximum.

3.1.3 Expectation-Maximization (EM) algorithm

The Expectation-Maximization (EM) algorithm provides a method for finding maximum likelihood estimates in models which depend on unobserved latent variables i.e. variables that are inferred from other observed variables (Karlis, 2001). The term EM Algorithm was first coined by Dempster et al. (1977) since each iteration of the algorithm requires an Expectation step followed by a Maximization step.

Let the observed data be denoted by y realized from the *pdf* $g(y|\theta)$ with corresponding log likelihood $\ell(\theta) = \log g(y|\theta)$. The aim is to estimate the vector parameter θ by the maximum likelihood estimate (MLE) $\hat{\theta}$ i.e. that value maximizing $\ell(\theta)$. The complete data representation of the problem involves regarding $y = y(x)$ as a statistic calculated from a hypothetical data vector x drawn from a density $f(x|\theta)$, where

$$g(y|\theta) = \int_{x|y(x)=y} f(x|\theta)\,\mathrm{d}x\ , \tag{3.24}$$

The general form of the EM algorithm involves maximizing $f(x|\theta)$ over values of θ, the M-step. Since x is unobservable we replace $\log f(x|\theta)$ by its conditional expectation given y and the current fit, θ, known as the E-step. This is then continued until convergence is achieved.

The two steps of an iteration of the algorithm (Wu, 1983; Green, 1992) can be written as follows:

Let $\theta^{(j)}$ denote the current value of θ after j cycles of the algorithm.

E-step: Using the current estimates $\theta^{(j)}$ taken from the j^{th} iteration, estimate the complete-data sufficient statistics $\log f(x|\theta)$ using,

$$Q(\theta|\theta^{(j)}) = \mathrm{E}(\log f(x|\theta)|y, \theta^{(j)}) \ . \tag{3.25}$$

M-step: Determine $\theta^{(j+1)}$ as the value of θ which maximizes the likelihood equations,

$$\mathrm{E}(\log f(x|\theta)|\theta) = \log f(x|\theta^{(j)}) \ . \tag{3.26}$$

The EM algorithm is a powerful tool for maximum likelihood estimation for data which contain missing values or can be considered as containing missing values e.g. with latent information (Dempster et al., 1977). This formulation is particularly applicable to discrete models which are generated as a mixture of distributions, where the mixing operation can be considered as producing missing data (Karlis, 2001). In this case, the missing data are realizations θ_i of the unobserved mixing parameter for each data point y_i.

The negative binomial distribution can be used as an example of the use of the EM algorithm for maximum likelihood estimation. Suppose y is a vector of observed values from a Poisson distribution with parameter μ, where μ follows a Gamma distribution, denoted here by $h(\mu|r, p)$ with parameters r and p, called the hyperparameters. The parameters of the resultant negative binomial model can be estimated using the EM algorithm with an incomplete data formation for the mixing density.

The MLE of the negative binomial distribution can be estimated through an EM algorithm by computing the maximum likelihood estimates of r and p from the marginal density of the data, $g(y|r,p)$,

$$g(y|r,p) = \int_{\Theta_r \times \Theta_p} \ell(y|\mu)\, h(\mu|r,p)\, d\mu \tag{3.27}$$

where Θ_r and Θ_p are the parameter spaces for r and p, respectively. To implement the EM algorithm we need to obtain $\mathrm{E}(\mu|y)$ and $\mathrm{E}(\log\mu|y)$. For the current estimates, $r^{(j)}$ and $p^{(j)}$, the EM scheme is as follows

E-step: Calculate the pseudo-values t_i and s_i,

$$t_i = \mathrm{E}(\mu_i|y_i) = \frac{y_i + r^{(j)}}{1 + p^{(j)}} \quad \text{and} \quad s_i = \Psi(r^{(j)} + y_i) - \log(p^{(j)} + 1)\ , \tag{3.28}$$

for $i = 1, \ldots, n$ where $\Psi(\cdot)$ is the digamma function (See Section 2.1.6 of Chapter 2).

M-step: Maximize the likelihood of the posterior distribution using t_i and s_i. Using the Expectation/Conditional Maximimization (ECM) algorithm (Meng and Rubin, 1993), update

$$p^{(j+1)} = \frac{r^{(j)}}{\bar{t}}\ , \tag{3.29}$$

and,

$$r^{(j+1)} = r^{(j)} - \frac{\Psi(r^{(j)}) + \log(p^{(j+1)}) - \bar{s}}{\Psi_3(r^{(j)})}\ , \tag{3.30}$$

until convergence is achieved, where $\bar{t}$ and $\bar{s}$ are the expected values of t_i and s_i, respectively and $\Psi_3(\cdot)$ denotes the trigamma function (see 2.1.6 of Chapter 2) .

An advantage of the EM algorithm is that it allows fitting complex models by including both observed data and unobserved or missing data and parameter constraints are often dealt with implicitly within the model. When using an EM algorithm the likelihood is guaranteed to increase at each iteration and does not require derivatives for the estimation. The algorithm is also fast where analytical expressions for the M-step are available.

However, the EM algorithm can be computationally intensive and convergence may be slow, due to the dependence on the unobserved information that needs to be estimated at the E-step (Karlis, 2001). Convergence may also be slow where analytical expressions for the M-step are not available since numerical optimization must be applied.

3.2 Frameworks for model fitting

This section introduces three frameworks for model fitting. The Generalized Linear Models (GLM), Generalized Additive Models (GAM) and Generalized Additive Models for Location, Scale and Shape (GAMLSS) classes all provide frameworks for regression models and incorporate discrete distributions as special cases.

3.2.1 Generalized Linear Models (GLM)

Generalized Linear Models (GLM) are an extension of classical linear models and were first formulated by Nelder and Wedderburn (1972). First consider a linear model, for a set of observations $y_1, y_2, \ldots, y_n$ assumed to be realizations of random variables $Y_1, Y_2, \ldots, Y_n$. Let $X_1, \ldots, X_n$ be a set of *d*-dimensional covariates and $\mu\left(\mathbf{X_i}\right)$ indicate the mean of Y_i. Allowing the mean response to depend on covariates $\mathbf{X}$, a linear model is then given by

$$\mu(\mathbf{X_i}) = \alpha + \sum_{j=1}^{p} \beta_j \mathbf{X}_{ij} \ , \tag{3.31}$$

where β_j is a vector of unknown parameters to be estimated from the data, p is the number of covariates and random variables are assumed to be independently distributed with constant variance of errors. GLM's require that the probability distribution f_y is a member of the exponential class of families (see Section 2.2).

This linear model (3.31) can be extended to a GLM by using a linear predictor η, which is a function of the mean μ_i

$$\eta_i = g\{\mu(X_i)\} = \alpha + \sum_{j=1}^{p} \beta_j X_{ij} \ , \tag{3.32}$$

where $g(\cdot)$ is a link function (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983). The classical linear model in Equation 3.31 has a normal distribution and an identity link function, $\eta = \mu$, whereas GLM's allow for the distribution of the Y_i's to follow an exponential family distribution (see Equation 2.45 in Section 2.2)(McCullagh and Nelder, 1983). In practice, GLM's are written as consisting of three elements:

1. A probability distribution from the exponential family
2. A linear predictor η
3. A link function g

and are defined in terms of μ and $\eta = g(\mu)$, where exponential family distributions can be written in the form $f_Y(y, \mu, \phi)$. In a GLM the link function may be any monotonic differentiable function for a given *pdf* (McCullagh and Nelder, 1983). The canonical link function is the function that expresses θ_i in terms of μ i.e. $\theta_i = b(\mu)$ (Hilbe, 2007). A commonly used link function is the identity link, for which $\eta = \theta$, where θ is a parameter of the exponential family (Hastie and Tibshirani, 1986). Other link functions include the log link, $\eta = \log(\mu)$, the logit link $\eta = \log\left(\frac{\mu}{1-\mu}\right)$ and the inverse link $\eta = \frac{1}{\mu}$.

The maximum likelihood estimate of the parameters $\hat{\beta} = \{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d\}$ for a given GLM with link function g and n observations can be found using a Fisher scoring procedure (Hastie and Tibshirani, 1986). Given a current estimate of the linear predictor $\hat{\eta}$ with corresponding fitted value $\hat{\mu}$, the adjusted dependent variable is given by:

$$Z = \hat{\eta} + (y - \hat{\mu})\left(\frac{d\eta}{d\mu}\right) , \tag{3.33}$$

(Hastie and Tibshirani, 1986). A new estimate of $\hat{\beta}$ can be obtained by regressing Z on $X_1, \ldots, X_d$, with weights W, given by

$$(W)^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 V , \tag{3.34}$$

where V is the variance of Y at $\mu = \hat{\mu}$. Using this estimate of $\hat{\beta}$ a new $\hat{\mu}$ and $\hat{\eta}$ can be computed. A new value of Z can then be calculated with these estimates and the algorithm continues until the change in the deviance,

$$\mathrm{dev}(y, \hat{\mu}) = 2[\ell(y) - \ell(\hat{\mu})] \;, \tag{3.35}$$

is sufficiently small, where $\ell(\mu)$ is the log-likelihood $\sum_{i=1}^{n} \log f_Y(y_i, \mu_i, \phi)$ (Hastie and Tibshirani, 1986).

Alternatively, an iteratively re-weighted least squares algorithm can be used to estimate $\hat{\beta}$ by solving the quasi-likelihood equations,

$$\frac{\partial Q}{\partial \beta} = \frac{\partial \eta}{\partial \beta}\frac{\partial Q}{\partial \eta} = 0 \;, \tag{3.36}$$

where Q is the log quasi-likelihood defined as any function of η satisfying,

$$\frac{\partial Q}{\partial \eta} = V^{-}(\eta)(y - \eta) \;, \tag{3.37}$$

where $V^{-}(\eta)$ is a generalized inverse of η (Green, 1984).

The GLM framework can be illustrated for a Poisson model with *pdf*,

$$f_Y(y; \mu) = \frac{\mathrm{e}^{-\mu}\mu^{y}}{y!} \;, \tag{3.38}$$

giving a log likelihood of,

$$\ell(\mu; y) = \sum \{y\mathrm{log}(\mu) - \mu - \log(y!)\} \;, \tag{3.39}$$

and link, $\log(\mu)$, resulting in the inverse link, $\mu = \exp(X'\beta)$ where $X'\beta$ is a linear predictor, with X being a matrix with length equal to the number of observations in the dataset and columns equal to the number of covariates plus a column of value ones if a constant is specified in the model and β is a vector of coefficients for each of the covariates specified for each column of X (Hilbe, 2007). Substituting the inverse link into (3.39) gives,

$$\ell(\beta; y) = \sum \left\{ y(X'\beta) - \exp(X'\beta) - \log(y!) \right\} , \tag{3.40}$$

which can alternatively be written as,

$$\ell(\beta; y) = \sum \left\{ y(x\beta) - \exp(x\beta) - \log(y!) \right\} , \tag{3.41}$$

The first derivative with respect to β of the Poisson log-likelihood is,

$$\frac{\partial \ell}{\partial \beta} = \sum \left\{ yx - x \exp(x\beta) \right\} . \tag{3.42}$$

The parameter estimates, $\hat{\beta}$, can be obtained by setting (3.42) equal to 0 and solving using one of the Fisher Scoring Procedure or the iteratively re-weighted least squares algorithm.

The main advantage of the GLM framework is that it provides a consistent way of linking together systematic and random elements in a model (Nelder and Wedderburn, 1972). A single algorithm can be used to fit any of the models in a GLM framework and the calculation of the Hessian within the algorithm allows standard errors also to be estimated. However, distributions used for modelling are restricted to only those within the exponential family, which for discrete models is limited to the Bernoulli, Binomial, Poisson, Geometric and NB distributions. The GLM framework is therefore not suitable for models which can account for overdispersion, value-inflation and truncation.

3.2.2 Generalized Additive Models (GAM)

Proposed by Hastie and Tibshirani (1986), the Generalized Additive Models (GAM) class provides a flexible framework for modelling. It is a regression technique that combines the properties of GLM's with an additive component i.e. where the linear function, $\sum_{j=1}^{d} \beta_j X_{ij}$, is replaced by an additive function, $\sum_{j=1}^{d} s_j(X_{ij})$, and hence each covariate is modelled as an unspecified smooth function rather than as a parametric function (Thurston et al., 2000).

The generalized additive class of models extends the GLM class seen in Equation 3.32 by allowing non-linearity between the link η and the covariates X_{ij}. A GAM model is then given by,

$$\eta_i = g\{\mu(X_i)\} = \alpha + \sum_{j=1}^{d} s_j(X_{ij}) \,, \tag{3.43}$$

where each s_j is a smooth function standardized so that $\mathrm{E}_{s_j}(X_j) = 0$ (Hastie and Tibshirani, 1986; Thurston et al., 2000).

Hastie and Tibshirani (1986) present two algorithms known as backfitting and local-scoring to fit GAM's. The estimating procedure for fitting GAM's consists of two loops. Inside each step of the local scoring algorithm (outer loop), a weighted backfitting algorithm (inner loop) is used until convergence. Then, based on the estimates from this weighted backfitting algorithm, a new set of weights is calculated and the next iteration of the scoring algorithm starts. The local scoring and backfitting algorithms are as follows:

Local Scoring Algorithm:

For starting values $s_j = g(\mathrm{E}(y))$ and $s_1^0 = s_2^0 = \ldots = s_p^0 = 0$, given a current estimate of the linear predictor, $\hat{\eta}$, with corresponding fitted value $\hat{\mu}$, the adjusted dependent variable is given by,

$$Z = \hat{\eta} + (Y - \hat{\mu})\frac{\partial \eta}{\partial \mu} \,.$$

The weights W are then formed as,

$$(W)^{-1} = \left(\frac{\partial \mu}{\partial \eta}\right)^2 V$$

where V is the variance of Y at $\mu = \hat{\mu}$. An additive model is fitted to Z using the backfitting algorithm (below) with weights W to obtain estimates of the functions $s_j^m(\cdot)$. The scoring algorithm stops when the deviance of the estimates ceases to decrease.

Backfitting algorithm:

For initial estimates $\alpha = \mathrm{E}(Y)$, $s_1 = s_2 = \ldots = s_p = 0$ and $m = 0$. Calculate at each iteration $m = m + 1$ the j^{th} set of the partial residuals,

$$R_j = Y - \alpha - \sum_{k=1}^{j-1} s_k^{(m)}(X_k) - \sum_{k=j+1}^{p} s_k^{(m+1)}(X_k) \ ,$$

where $s_j^{(m)} = \mathrm{E}(R_j|X_j)$. The iterations continue until,

$$RSS = \mathrm{E}\left[Y - \alpha - \sum_{j=1}^{p} s_j^{(m)}(X_j)\right]^2 \ ,$$

fails to decrease or satisfies the convergence criterion.

Thurston et al. (2000) presents an algorithm, called the alternating profile likelihood algorithm, to fit a negative binomial additive model using the local scoring and backfitting algorithms. The alternating profile likelihood algorithm fits the two parameters of the negative binomial distribution by iterating between the two algorithms. For a negative binomial distribution with parameters μ the mean and α the dispersion parameter specified in Section 2.3.1 of Chapter 2, the structure of the alternating profile likelihood algorithm is as follows:

1. **Iterate the alternating profile likelihood algorithm** Each iteration requires implementation of the local scoring algorithm.

2. **Iterate the local scoring algorithm** Each iteration requires implementation of the backfitting algorithm for a weighted additive model. For this the link function, $\eta = \log\left(\dfrac{\mu}{\mu + \alpha}\right)$ and the inverse link $\mu = \dfrac{\alpha}{\mathrm{e}^{-\eta} - 1}$ are needed. The weights are given by,

$$W = \mu + \frac{\mu^2}{\alpha} = \mu\left(\frac{\mu + \alpha}{\alpha}\right) = \frac{\alpha\, \mathrm{e}^{\eta}}{(\mathrm{e}^{\eta} - 1)^2} \ .$$

3. **Iterate the backfitting algorithm** Each iteration involves a weighted local polynomial smooth, for each predictor X_d.

The GAM framework is a very flexible method for fitting models in the exponential family and other likelihood-based regression models. However, the disadvantage of an increase in flexibility is the potential to over-fit the data by applying overly complex models. Currently GAM models only allow for exponential family likelihoods, which is limited where overdispersion and/or value-inflation is present (Thurston et al., 2000). One disadvantage of GAM's is that they are not as easy to interpret in comparison to GLM's, in particular when they involve complex additive effects.

3.2.3 Generalized Additive Models for Location, Scale and Shape (GAMLSS)

The class of Generalized Additive Models for Location, Scale and Shape (GAMLSS) was developed by Rigby and Stasinopoulos (2005). It allows fitting more complex models in which both the systematic and the random parts of the model are highly flexible. Both the GLM and GAM classes (see Sections 3.2.1 and 3.2.2) assume that the response variable follows an exponential family distribution, in which the models variance, skewness and kurtosis are modelled through their dependence on μ, as opposed to being modelled explicitly in terms of the explanatory variables. In the GAMLSS class the exponential family assumption is relaxed and replaced by a more general family of distributions. This new class allows all the parameters of the distribution of Y to be modelled as parametric and/or additive non-parametric functions of the explanatory variables and/or random effect terms (Rigby and Stasinopoulos, 2005).

A model in this class assumes independent observations, y_i for $i = 1, 2, \ldots, n$ with *pdf* $f(y_i|\boldsymbol{\theta}^{(i)})$ conditional on a vector of four distribution parameters $\boldsymbol{\theta}^{(i)} = (\theta_{1i}, \theta_{2i}, \theta_{3i}, \theta_{4i}) = (\mu_i, \sigma_i, \nu_i, \tau_i)$, which can be functions of the explanatory variables. The first two parameters μ_i and σ_i characterize location and scale, whilst the remaining parameters (if any) characterize shape, often (but not always) skewness and kurtosis.

Let $\mathbf{y}^T = y_1, y_2, \dots, y_n$ denote the vector of response observations. Also, for $k = 1, 2, \dots$ let $g_k(\cdot)$ be a known monotonic link function relating θ_k to explanatory variables and random effects through an additive model given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\boldsymbol{\gamma}_{jk} \ , \tag{3.44}$$

i.e.

$$\begin{aligned} g_1(\boldsymbol{\mu}) &= \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\boldsymbol{\gamma}_{j1} \\ g_2(\boldsymbol{\sigma}) &= \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} \mathbf{Z}_{j2}\boldsymbol{\gamma}_{j2} \\ g_3(\boldsymbol{\nu}) &= \boldsymbol{\eta}_3 = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=1}^{J_3} \mathbf{Z}_{j3}\boldsymbol{\gamma}_{j3} \\ g_4(\boldsymbol{\tau}) &= \boldsymbol{\eta}_4 = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=1}^{J_4} \mathbf{Z}_{j4}\boldsymbol{\gamma}_{j4} \end{aligned} \tag{3.45}$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are vectors of length n, $\boldsymbol{\beta}_k^T = (\beta_{1k}, \beta_{2k}, \dots, \beta_{J'_k k})$ is a parameter vector of length J'_k, $\mathbf{X}_k$ is a design matrix of order $n \times J'_k$, $\mathbf{Z}_{jk}$ is a design matrix $n \times q_{jk}$ and $\boldsymbol{\gamma}_{jk}$ is a q_{jk}-dimensional random variable (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007).

The parameter vectors $\boldsymbol{\beta}_k$ and the random effect parameters $\boldsymbol{\gamma}_{jk}$, for $j = 1, 2, \dots, J_k$ and $k = 1, 2, 3, 4$ can be estimated by maximizing a penalized likelihood function given by,

$$\ell_p = \ell - \frac{1}{2}\sum_{k=1}^{p}\sum_{j=1}^{J_k} \lambda_{jk}\gamma'_{jk}\mathbf{G}_{jk}\gamma_{jk} \ , \tag{3.46}$$

for fixed values of the smoothing hyper-parameters λ_{jk}'s, where $\ell = \sum_{i=1}^{n} \log f(y_i|\boldsymbol{\theta}^{(i)})$ is the log likelihood function (Stasinopoulos and Rigby, 2007). The penalized likelihood given in Equation 3.46 can be maximized using either the Cole-Green (CG) algorithm (Cole and Green, 1992) which uses the first and second and cross derivatives of the likelihood function with respect to the distribution parameters $\boldsymbol{\theta} = (\mu, \sigma, \nu\, \tau)$ (Stasinopoulos and Rigby, 2007). Or alternatively, the Rigby-Stasinopoulos (RS) algorithm, a simpler algorithm used for fitting mean and dispersion additive models (MADAM) which does not use cross derivatives (Rigby and Stasinopoulos, 1996).

The negative binomial type I distribution can be re parametrized in the GAMLSS framework with μ (the mean) a location parameter and σ the scale parameter, where $\alpha = \dfrac{1}{\sigma}$ in the *pdf* of the distribution given in 2.3.1 of Chapter 2. The GAMLSS NBI distribution has *pdf*,

$$f_Y(y;\mu,\sigma) = \frac{\Gamma\left(y+\frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma\left(y+1\right)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^{y}\left(\frac{1}{1+\sigma\mu}\right)^{\frac{1}{\sigma}} \tag{3.47}$$

for $y = 0, 1, 2, \ldots$, where $\mu > 0$ and $\sigma > 0$ with $\mathrm{E}(Y) = \mu$ and $\mathrm{Var}(Y) = \mu + \sigma\mu^2$. The log-likelihood of this distribution is,

$$\begin{aligned} \ell = \; & \log\left(\Gamma\left(y+\tfrac{1}{\sigma}\right)\right) - \log\left(\Gamma\left(\tfrac{1}{\sigma}\right)\right) - \log\left(\Gamma\left(y+1\right)\right) \\ & + y\log\left(\sigma\mu\right) - y\log\left(1+\sigma\mu\right) - \tfrac{1}{\sigma}\log\left(1+\sigma\mu\right) \end{aligned} \tag{3.48}$$

For the CG algorithm the first and expected second and cross derivatives of the likelihood function with respect to the distribution parameters μ and σ are required. The first derivatives of the likelihood with respect to μ and σ are given as

$$\frac{\partial\ell}{\partial\mu} = \frac{y-\mu}{\mu(1+\mu\sigma)}$$

$$\frac{\partial\ell}{\partial\sigma} = -\left(\frac{1}{\sigma}\right)^2\left(\psi\left(y+\frac{1}{\sigma}\right) - \psi\left(\frac{1}{\sigma}\right) - \log\left(1+\mu\,\sigma\right) - \frac{(y-\mu)\sigma}{(1+\mu\sigma)}\right) \; ,$$

The expected second derivatives are given as

$$E\left[\frac{\partial^2\ell}{\partial\mu^2}\right] = -\frac{1}{\mu(1+\mu\,\sigma)} \; ,$$

$$\begin{aligned} E\left[\frac{\partial^2\ell}{\partial\sigma^2}\right] = \; & -\textstyle\sum_{y=0}^{\infty}\left(\left(\frac{1}{1+\mu\sigma}\right)^{\frac{1}{\sigma}}\left(\frac{\mu\sigma}{1+\mu\sigma}\right)^{y}\Gamma\left(y+\frac{1}{\sigma}\right)\left((y-\mu)\sigma + (1+\mu\,\sigma)\log(1+\mu\,\sigma)\right.\right. \\ & \left.\left. -\;(1+\mu\,\sigma)\,\psi\left(y+\tfrac{1}{\sigma}\right) + (1+\mu\,\sigma)\,\psi\left(\tfrac{1}{\sigma}\right)\right)^2\right) / \\ & \left(\sigma^4(1+\mu\sigma)^2\,\Gamma(y+1)\,\Gamma\left(\tfrac{1}{\sigma}\right)\right) \; , \end{aligned} \tag{3.49}$$

and

$$E\left[\frac{\partial^2 \ell}{\partial \mu \sigma}\right] = 0 \ ,$$

where $\psi(x)$ is the digamma function (see Section 2.1.6 in Chapter 2). The expected second derivatives can be replaced in some cases by the negative squared first derivatives, where the expected second derivatives are not analytically tractable (Stasinopoulos and Rigby, 2008).

The main advantage of the GAMLSS framework in comparison to the the GLM and GAM frameworks of models is that distributions do not need to belong to the Exponential family for this class of models. A large number of GAMLSS distributions are available which can account for location, scale, skewness and kurtosis parameters. The GAMLSS framework has the potential to allow for (almost) any probability density to be used when modelling. A benefit of GAMLSS models is that all parameters of the conditional distribution of y can be modelled as parametric and/or additive non-parametric (smooth) functions of explanatory variables and/or random effects terms. The fitting algorithm is also fast enough to fit very large and complex data sets. Software for implementing GAMLSS models is freely available via the `R` language for statistical computing (R Development Core Team, 2009) in the `gamlss` libraries (Stasinopoulos and Rigby, 2008).

Whilst the GAMLSS framework allows for more realistic assumptions when modelling datasets, model selection is more difficult due to the increase in available models to select from. A disadvantage of the framework is that estimation is based upon the first and expected second derivatives of the likelihood with respect to the parameters, which for some distributions can be complex. A numerical algorithm is however available within the `gamlss` libraries which approximates the derivatives.

3.3 Diagnostics

There are three aspects of methods for diagnostic analysis of models: goodness-of-fit methods, model comparison and outlier detection. The Chi-squared goodness-of-fit test and residual analysis are two methods for assessing a model's fit, which assess whether a particular model provides a good fit to a dataset. The fit of a range of distributions to a dataset can also be compared using the Akaike or Bayesian Information Criteria or using a graphical method which plots the EPGF of a dataset. Finally, potential outliers in discrete distributions can be investigated using two methods for outlier detection: the EPGF plot and the surprise index.

3.3.1 Goodness-of-fit

The Chi-squared goodness-of-fit test assesses whether a dataset follows a specified distribution. Residual analysis uses graphical plots of the residuals of a model to determine the quality of fit and detect possible problems with the fit of a model to the dataset.

Chi-squared Goodness-of-fit Test

The success of the fit of a model to a dataset can be determined using a Chi-Squared test of goodness-of-fit (Chernoff and Lehmann, 1954) by comparing the fitted (or expected) data, e, and the observed data, o with the χ^2 statistic is as follows,

$$\chi^2 = \sum \frac{(o - e)^2}{e} \tag{3.50}$$

This can be compared to the χ^2 distribution with $(n - p - 1)$ degrees of freedom, where n is the number of independent observations and p the number of parameters fitted (McCullagh and Nelder, 1983). For the Chi-square approximation to be valid the expected frequencies should all be at least 5. In the case of discrete datasets, expected frequencies are often 0 or very small values and several frequencies may be required to be pooled to ensure the expected frequencies are greater than 5.

For the Poisson model fitted in Section 3.1.2, the χ^2 goodness-of-fit test can be performed using the observed counts of cysts in steroid treated embryonic mouse kidneys (o) and the expected values (e) calculated be substituting the maximum likelihood estimate for μ in the Poisson distribution and scaling by the sample size ($n = 111$): values are presented in Table 3.3.1.

	0	**1**	**2**	**3**	**4 or more**
Observed (o)	65	14	10	6	15
Expected (e)	24	37	28	15	8

Table 3.1: Observed (o) and Expected (e) frequencies of cysts in steroid treated mouse kidneys for a Poisson model.

The null, H_0, and alternative, H_a, hypotheses of the χ^2 goodness-of-fit test are,

H_0 : The data follow a specified distribution

H_a : The data do not follow the specified distribution.

The test statistic for the Poisson model with observed and expected frequencies given in Table 3.3.1 is $\chi^2 = 107.44$. Comparing this to a χ^2_3 distribution, where the degrees of freedom are $df = 5 - 1 - 1 = 3$, gives a $p < 0.005$, indicating that a Poisson distribution is not a suitable model for this dataset. Alternatively, we can fit a negative binomial model to this dataset using maximum likelihood as described in section 3.1.2. This model has parameter estimates $\hat{r} = 0.30$ and $\hat{p} = 0.16$ with log-likelihood $\ell = -174.81$. The pooled observed and expected frequencies for the number of cysts in kidneys is given in Table 3.2.1 for a negative binomial model.

	0	**1**	**2**	**3 or more**
Observed (o)	65	14	10	6
Expected (e)	65	16	9	10

Table 3.2: Observed (o) and Expected (e) frequencies of cysts in steroid treated mouse kidneys for a negative binomial model.

The χ^2 goodness-of-fit test statistic for the negative binomial model is $\chi^2 = 1.96$. The degrees of freedom for this test are $df = 4 - 2 - 1 = 1$, giving a p-value of 0.375 (3sf) when the χ^2 test statistic is compared to a χ^2_1 distribution. This p-value is not significant therefore the null hypothesis H_0 cannot be rejected suggesting that the negative binomial distribution is suitable for this dataset.

Residuals

Residuals are widely used to assess the fit of models (Cox and Snell, 1968). Regression models such as the GLM's presented in Equation 3.31 assume that the response variables, Y_i, are independent and normally distributed having equal variance σ^2 and are linear i.e. the relationship between $\mathrm{E}(Y)$ and explanatory variables X_{ij} is a straight line. Rather than checking these assumptions on the response variables directly, it is convenient to re-express the assumptions in terms of the random errors.

The random errors or raw residuals R are the difference between the observed responses y and the predicted or fitted responses $\hat{y}$ and are given by,

$$R = y - \hat{y}\ , \quad R = y - \mu \quad \text{or} \quad R = y - \mathrm{E}\,(y) \tag{3.51}$$

The following four assumptions of the residuals are equivalent to the assumptions on the response variable,

i. The residuals R are independent.

ii. The residuals R are normally distributed.

iii. The residuals R have constant variance σ_R^2.

iv. The residuals R have zero mean.

A benefit of the raw residuals is they are relatively easy to calculate, however they do not have a constant variance and are therefore not suitable to test the assumption that the underlying errors have a constant variance. The raw residuals can be standardized by subtracting the mean and dividing by the standard deviation to overcome the problem of non-constant variance. Since the mean of the Residuals R is 0, this gives the standardized residuals, R_S

$$R_S = \frac{R}{s\,\sqrt{1 - h_{ii}}} \tag{3.52}$$

where s is an appropriate estimate of the standard deviation σ and h_{ii} is the i^{th} diagonal element of the hat-matrix, $\mathbf{H}$, given by

$$\begin{aligned}\mathbf{H} &= \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{pmatrix}, \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \end{aligned} \tag{3.53}$$

where $\mathbf{X}$ is the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{pmatrix}. \tag{3.54}$$

The standardized residuals will follow a standard normal distribution i.e. they will be normally distributed with mean zero and variance one.

Other forms of standardized residuals are used in the analysis of count response models. The next two sections present Pearson residuals, which are commonly used in residual analysis for GLM's and Anscombe residuals, another standardized residual for discrete models.

Pearson residuals

The Pearson residual is the raw residual, R, scaled by the estimated standard deviation of y (McCullagh and Nelder, 1983):

$$R_p = \frac{y-\mu}{\sqrt{V(\mu)}} \tag{3.55}$$

which have mean 0 and variance ϕ the dispersion parameter of the exponential family of distributions, which is equal to 1 for count response models.

Anscombe residuals

Anscombe residuals attempt to normalize the residuals so that heterogeneity and outliers in the data can be easily identified (Hilbe, 2007). Anscombe (1953) defines a function $A(\cdot)$ which is chosen to ensure the distribution of $A(y)$ is as normal as possible. This is done by utilizing the model variance functions and replacing y with $A(y)$. The function $A(y)$ is given by

$$A\left(\cdot\right) = \int_{-\infty}^{\mu} V(\mu)^{-\frac{1}{3}}\, d\mu \tag{3.56}$$

where $V(\mu)$ is the variance function (Hilbe, 2007). The general formula for Anscombe residuals is

$$R_A = \frac{A\left(y\right) - A\left(\mu\right)}{A'\left(\mu\right)\sqrt{V\left(\mu\right)}}\ , \tag{3.57}$$

where $A'(\mu)$ is the derivative of $A(\mu)$.

Hilbe (2007) gives three special cases of Anscombe residuals for the Poisson, geometric and negative binomial type II distributions. For the Poisson distribution the Anscombe residuals are given by

$$R_A = \frac{3\left(y^{\frac{2}{3}} - \mu^{\frac{2}{3}}\right)}{2\mu^{\frac{1}{6}}}\ , \tag{3.58}$$

where $V = \mu$ and for the Geometric distribution

$$R_A = \frac{\left(3\left(1+y\right)^{\frac{2}{3}} - \left(1-\mu\right)^{\frac{2}{3}}\right) + 3\left(y^{\frac{2}{3}} - \mu^{\frac{2}{3}}\right)}{2\left(\mu^2+\mu\right)^{\frac{1}{6}}} \tag{3.59}$$

where $V = \mu\left(1+\mu\right)$. Finally, the Anscombe residuals for the Negative Binomial Type II distribution are as follows,

$$R_A = \frac{\left(\frac{3}{\alpha}\left(\left(1+\alpha y\right)^{\frac{2}{3}} - \left(1+\alpha\mu\right)^{\frac{2}{3}}\right) + 3\left(y^{\frac{2}{3}} - \mu^{\frac{2}{3}}\right)\right)}{2\left(\alpha\mu^2+\mu\right)^{\frac{1}{6}}}\ , \tag{3.60}$$

where $V = \mu + \alpha\mu^2$ or $V = \mu\left(1+\alpha\mu\right)$. Anscombe residuals for other discrete distribution have not been established.

Several reasons for departures from the fitted model can be investigated using residuals, such as: outliers, further covariates omitted from the model, correlation between residuals, non-constant variance and non-normality (Cox, 1986). The underlying statistical assumptions about the residuals (i-iv) can be assessed using different types of residual plots to check the validity of these assumptions and provide information on how to improve the model.

Residuals vs. fitted values The assumptions that the residuals have constant variation (iii) and zero mean (iv) can be checked by plotting the Residuals against the fitted values. If assumptions (iii) and (iv) are satisfied the residuals are expected to vary randomly around zero and the spread of the residuals to be constant throughout the plot.

Residuals against Index The residuals vs. the index of the data can be used to check the assumption that the errors are independent (i). If the residuals are randomly distributed around zero there will be no drift or patterns in the process.

Normality The assumption that the residuals are normally distributed (ii) is important in the context of discrete data where residuals also take integer values and can be tested in two ways. Firstly, a histogram or plot of the density estimate shows the distribution of the residuals. A symmetric bell-shaped histogram, evenly distributed around 0 indicates the normality assumption is valid. Alternatively, a normal Q-Q plot of the residuals indicates whether the normality assumption of the residuals is appropriate.

A Quantile-Quantile (Q-Q) plot is a scatter plot comparing the fitted quantiles and empirical quantiles of a dataset (McCullagh and Nelder, 1983). It is a graphical technique for determining if a data set come from a distribution. An advantage of Q-Q plots is that they allow for shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can to be detected. If the two sets of quantiles being compared are similar, the points in the Q-Q plot will approximately lie on the line

$y = x$. A normal Q-Q plot can be used in residual analysis to compare the quantiles of the residuals (vertical axis) to a standard normal population (horizontal axis).

The particular problem with discrete datasets is that the response variable takes a small number of distinct values (Dunn and Smyth, 1996). Residuals from discrete responses result in parallel curves corresponding to distinct response values and distract from any information that may be contained in the residual plot. Dunn and Smyth (1996) present randomized quantile residuals which produces continuous residuals for discrete response variables by inverting the fitted distribution function at each response value and finding the equivalent standard normile quantile. This approach includes randomization to achieve continuous residuals for discrete response variables which allows for easier interpretation of the residuals.

Randomized Quantile residuals

Randomized Quantile residuals (Dunn and Smyth, 1996) are given by

$$R_Q = \Phi^{-1}(u_i) \tag{3.61}$$

where Φ^{-1} is the inverse *cdf* of a standard normal distribution with mean 0 and standard deviation 1 and u_i is a random value from the uniform distribution in the interval $\left[F\left((y_i - 1)\,|\hat{\theta^i}\right), F\left(y_i|\hat{\theta^i}\right)\right]$ when y_i is discrete, where $F(y|\theta)$ is the *cdf.*

Randomized quantile residuals retain the useful diagnostic properties of ordinary residuals, but lack their detracting features. The true residuals R_Q follow a standard normal distribution if the model is correct.

The randomization strategy employed prevents masses of overlapping points which occur in plots due to repeated integer values of discrete responses. Dunn and Smyth (1996) implement a process simlar to the strategy of 'jittering' by applying a uniform random component on the cumulative probability scale for each point. It has also been noted that when randomization is used, several randomized sets of residuals should be studied before a deciding upon the the adequacy of a model (Rigby and Stasinopoulos, 2005).

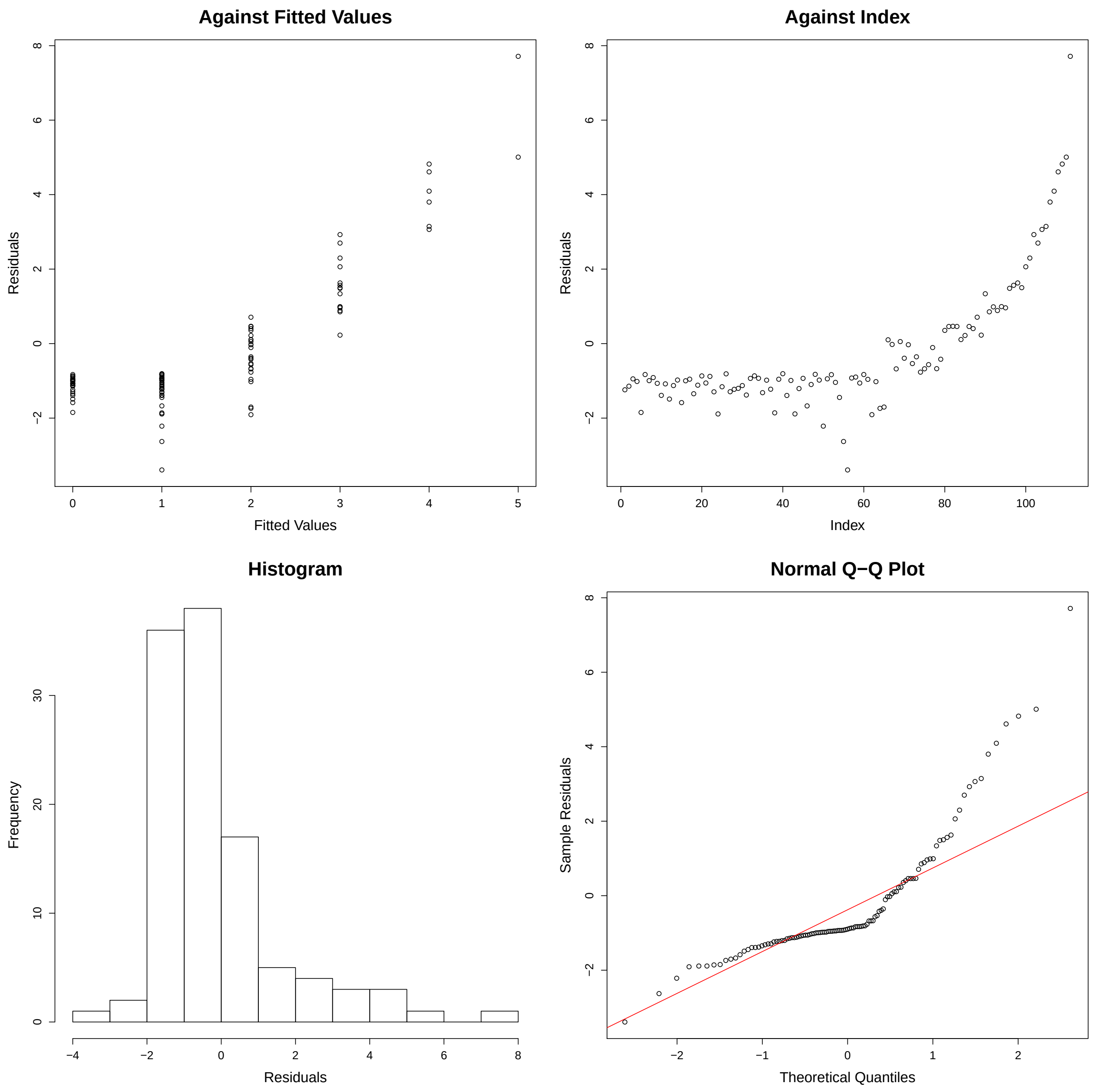


Figure 3.3: Residual analysis using Randomized Quantile Residuals for a Poisson model for counts of cysts in steroid treated kidneys

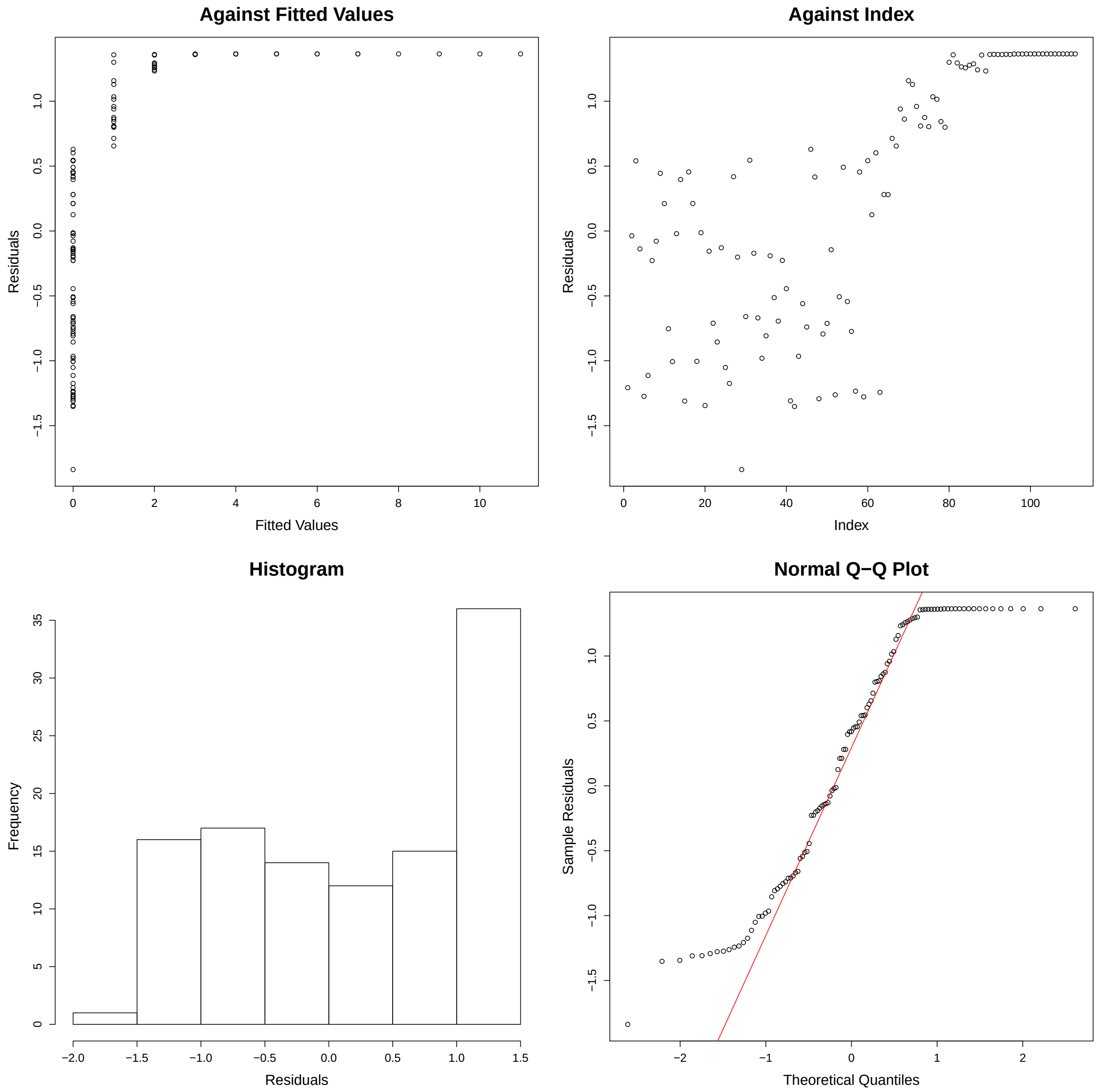


Figure 3.4: Residual analysis using Randomized Quantile residuals for a negative binomial model for counts of seizures in steroid treated kidneys

The fit of the Poisson maximum likelihood model to counts of cysts in steroid treated embryonic mouse kidneys in Section 3.1.2 can be checked by examining various plots of the residuals. The Randomized Quantile Residuals have been calculated for a Poisson model fitted to counts of cysts in steroid treated embryonic mouse kidneys and residual analysis plots are shown in Figure 3.3. The plot of the residuals against the fitted values clearly still indicate the variance is not constant and increases for large fitted values. The histogram of the residuals is also highly skew, with a large positive residual of approximately 8, which can also be seen in the Normal Q-Q plot. Since the assumptions of normality and constant variance of the residuals is shown to be violated, this suggests that this models does not provide a good fit to the data.

A negative binomial model can also be fitted to the counts of cysts using maximum likelihood, as in Section 3.1.2. The Randomized Quantile Residuals for this model are plotted in the usual residual plots in Figure 3.4. The residuals for this model show less variation and the histogram indicates that the distribution is not skew - all of the residuals lie in the range ± 3. The normal QQ plot indicates that the residuals of this model better approximate a normal distribution than those of the Poisson model, suggesting a better fit, although there are still some important deviations from the normal assumption.

3.3.2 Model Comparisons

Choosing the correct model is an important aspect of data analysis. The model makes assumptions about the implicit data generating mechanism present in the dataset and the correct distribution must be chosen to ensure the maximum amount of information is extracted from the data. It is therefore helpful to fit and compare a range of models to a dataset. The Akaike's information criterion (AIC) and Bayesian information criterion (BIC) and EGPF plots provide methods for comparing the fit of multiple discrete distributions to a dataset. AIC and BIC are types of penalized selection criteria which are based upon the deviance of a model and can compare the fit of distributions to a dataset numerically. The EPGF plots assess the fit of several distributions graphically by plotting the EPGF of a dataset and comparing to a range of *pgf*'s for fitted distributions.

Deviance

Measures of discrepancy between data values and a fitted model may be formed in many ways – one such way can be formed from the logarithm of a ratio of likelihoods, known as the *deviance*. Given a sample of n observations, the simplest model that can fitted to the data, known as the *null model*, has one parameter representing a common mean, μ, for all observations y. At the other end of the spectrum, the *full model* has n parameters (one for each observation) and fits the data exactly, providing a baseline for measuring the discrepancy of a model with p parameters.

Let $l\,(\hat{\theta_p})$ be the log likelihood maximised for the model with p parameters and $l\,(\hat{\theta_n})$ be the maximum log likelihood in the full model with n parameters. The deviance (McCullagh and Nelder, 1983) is then given by twice the difference between the two maximum likelihoods:

$$D\,(y) = 2[l\,(\hat{\theta_p}) - l\,(\hat{\theta_n})] \tag{3.62}$$

Here the *full* model is a model with a parameter for every observation so that the data are fitted exactly. In general, the deviance can be expressed for any two nested models, M_1 and M_2, where M_1 contains the parameters in M_2, and k additional parameters, with log-likelihoods L_1 and L_2, respectively. This results in the following deviance,

$$D\,(y) = 2\,(L_1 - L_2) \tag{3.63}$$

The benefit of the deviance is that it is additive for nested sets of models and can be used to compare two models in the likelihood ratio test.

Likelihood ratio Test

The likelihood ratio test is used to compare the fit of two competing models where one model (often called the *alternative* model) is a special case of other (the *full* model). The likelihood of the data under the alternative model is compared to the likelihood of the model under the full model, under the following hypotheses:

H_0 : The null model provides the best fit to the data

H_a : The alternative model provides the best fit to the data.

The test statistic for this test is based on the likelihood ratio of the null model, M_1 with n_1 parameters and the alternative model L_2 with n_2 parameters. Denoted by D, the deviance, the test statistic is written as:

$$D = -2[L_1 - L_2] \tag{3.64}$$

where L_1 and L_2 are the log-likelihoods for the models, M_1 and M_2, respectively. Under the assumption that the null hypothesis H_0 is true, this test statistic will follow a Chi-squared distribution on $n_1 - n_2$ degrees of freedom, where n_1 is the number of parameters in the null model and n_2 is the number of parameters in the alternative model (McCullagh and Nelder, 1983).

When the test statistic, D, is large M_2 the alternative model fits poorly compared with M_1. Large tests statistics and small *p*-values suggest the model M_2 fits more poorly than M_1.

AIC and BIC

Penalized model selection criteria provide a class of goodness-of-fit statistics which allow for comparisons of non-nested models i.e. models for which one model is not a sub model of the other. For example, a model with a covariate X_1 is nested within a model with covariates X_1 and X_2. However, a model with covariates X_1 and X_3 is *not* nested within the model with covariates X_1 and X_2, as the third covariate X_3 does not appear in the first model. Comparisons are made between pairs of candidate models,

M_1 and M_2, with parameter vectors θ_1 and θ_2, respectively, and are of the form:

$$IC = 2[\ell(\hat{\theta_2}) - \ell(\hat{\theta_1})] - a(p_2 - p_1) \tag{3.65}$$

where $\ell(\hat{\theta_2})$ and $\ell(\hat{\theta_1})$ are the log likelihood for models M_1 and M_2 respectively, p_1 and p_2 are their degrees of freedom and a is a positive quantity. It is not necessary for the two models M_1 and M_2 to be nested. For the special case of nested models, where M_1 is nested within M_2, the first term becomes equal to the likelihood ratio test statistic (Kuha, 2004).

Statistics of this kind are known as penalized likelihood criteria due to their formation as sums of two terms. The first term in Equation 3.65 is the deviance and reflects the fit of the two models to the observed data. The second term can be regarded as a penalty for the increased complexity of M_2 over M_1 in terms of the numbers of parameters in the model. These two terms express a trade-off between fit and model complexity, favouring a more parsimonious model unless the more complex model provides an improvement in fit.

The advantages of penalized likelihood criteria are that they allow for comparisons of non-nested as well as nested models. The penalty for a large model with many parameters offsets the large-sample behaviour of significance tests where simple models are increasingly likely to be rejected for large datasets. They are also based on explicit theoretical considerations despite their simplicity.

Many versions of penalized criteria have been proposed in the statistical literature using various theoretical starting points. The first was Akaike's information criterion (AIC) (Akaike, 1974), defined as:

$$AIC = 2[\ell(\hat{\theta_2}) - \ell(\hat{\theta_1})] - 2(p_2 - p_1) \tag{3.66}$$

i.e. $a = 2$. Another widely used penalized criterion is the Bayesian information criterion (BIC) also known as Schwarz's information criterion (SIC or SBIC, (Schwarz, 1978)):

$$BIC = 2[\ell(\hat{\theta_2}) - \ell(\hat{\theta_1})] - \log(n)\,(p_2 - p_1) \tag{3.67}$$

where n is the number of independent observations in the dataset. Lower AIC or BIC values indicate a better fitting model and allow us to compare competing models.

The Akaike and Bayesian information criteria are based on two different model selection approaches. The AIC is aimed at finding the best approximating model to the unknown data generating process, whilst BIC is designed to identify the true model (de Graft Acquah, 2010). The AIC does not depend directly on sample size. Although BIC takes a similar form to the AIC, it is derived within a Bayesian framework and reflects the sample size of the model. BIC values are always higher than those of the AIC as the BIC applies a larger penalty than the AIC, thus it tends to select simpler models than the AIC.

For the Poisson model applied to the cysts data with one parameter the AIC and BIC can be calculated from the log-likelihood,

$$\begin{aligned} AIC &= -(2 \times -279.7035) + (2 \times 1) = 561.4071 \\ BIC &= -(2 \times -279.7035) + (1 \times \log(111)) = 564.1166 \end{aligned} \tag{3.68}$$

and for a negative binomial distribution with 2 parameters,

$$\begin{aligned} AIC &= -(2 \times -174.8132) + (2 \times 2) = 353.6263 \\ BIC &= -(2 \times -174.8132) + (2 \times \log(111)) = 359.0454 \end{aligned} \tag{3.69}$$

The negative binomial model provides a better fit to the counts of cysts in steroid treated kidneys, resulting in lower values for both the AIC and BIC when compared to the Poisson distribution.

Increasing the complexity of the model improves the goodness-of-fit but has the added cost of requiring more independent parameters to be correctly estimated. The BIC is more conservative against over-fitting in comparison to the AIC. Whilst the AIC and BIC are the most often used in practice, a variety of other penalized criteria exist based upon modifications or generalizations of the AIC or BIC (Kuha, 2004). The BIC will be used in this thesis as it accounts for the differing number of parameters in model's when making comparisons and is therefore more conservative against over fitting than the AIC.

EPGF plots

Nakamura and Pérez-Abreu (1993b) present a graphical method of comparing the goodness-of-fit of discrete models based on the empirical probability generating function (EPGF) that provides a method of exploratory analysis of distributions for counts. The EPGF for count data $Y_1, Y_2, \ldots, Y_n$ is,

$$G_n\left(t\right) = \frac{1}{n}\sum_{i=1}^{n} t^{Y_i}\ , \tag{3.70}$$

for $-1 \leq t \leq 1$ and provides a statistical transformation to enable inferences about discrete distributions (Nakamura and Pérez-Abreu, 1993b; Rueda and O'Reilly, 1999). The EPGF can be compared to discrete distributions by plotting the log of the theoretical *pgf* of various candidate models and $G_n(t)$. Let $Y_1, \ldots, Y_n$ be a random sample from a discrete distribution, then $Y(t) = \log\left(G(t)\right)$ and $Y_n(t) = \log\left(G_n(t)\right)$. A graphical plot $Y_n(t)$ against t enables exploratory analysis of the fit discrete distributions (Nakamura and Pérez-Abreu, 1993b).

Nakamura and Pérez-Abreu (1993b) plot the log of the *pgf*, $Y(t)$ against values of t between 0 and 1 for the Poisson, Binomial, negative binomial and zero-truncated (Positive) Poisson distributions for fixed parameter values, shown in Figure 3.5. For the Poisson distribution, the log of the *pgf* is given by $Y(t) = \mu(t-1)$ and is a straight line with an intercept at $-\mu$ and is zero at $t = 1$. The log of the *pgf* for a Binomial distribution yields a concave function, whilst for a negative binomial or other mixtures of Poisson distributions the shape of $Y(t)$ is always convex (Nakamura and Pérez-Abreu, 1993b). For truncated distributions $Y(t)$ diverges to $-\infty$ as t converges to 0. For a truncated Poisson distribution, as $t \to 0$ the *pgf* $Y(t) \to -\infty$ and as $t \to 1$ the log of the *pgf* behaves as a straight line.

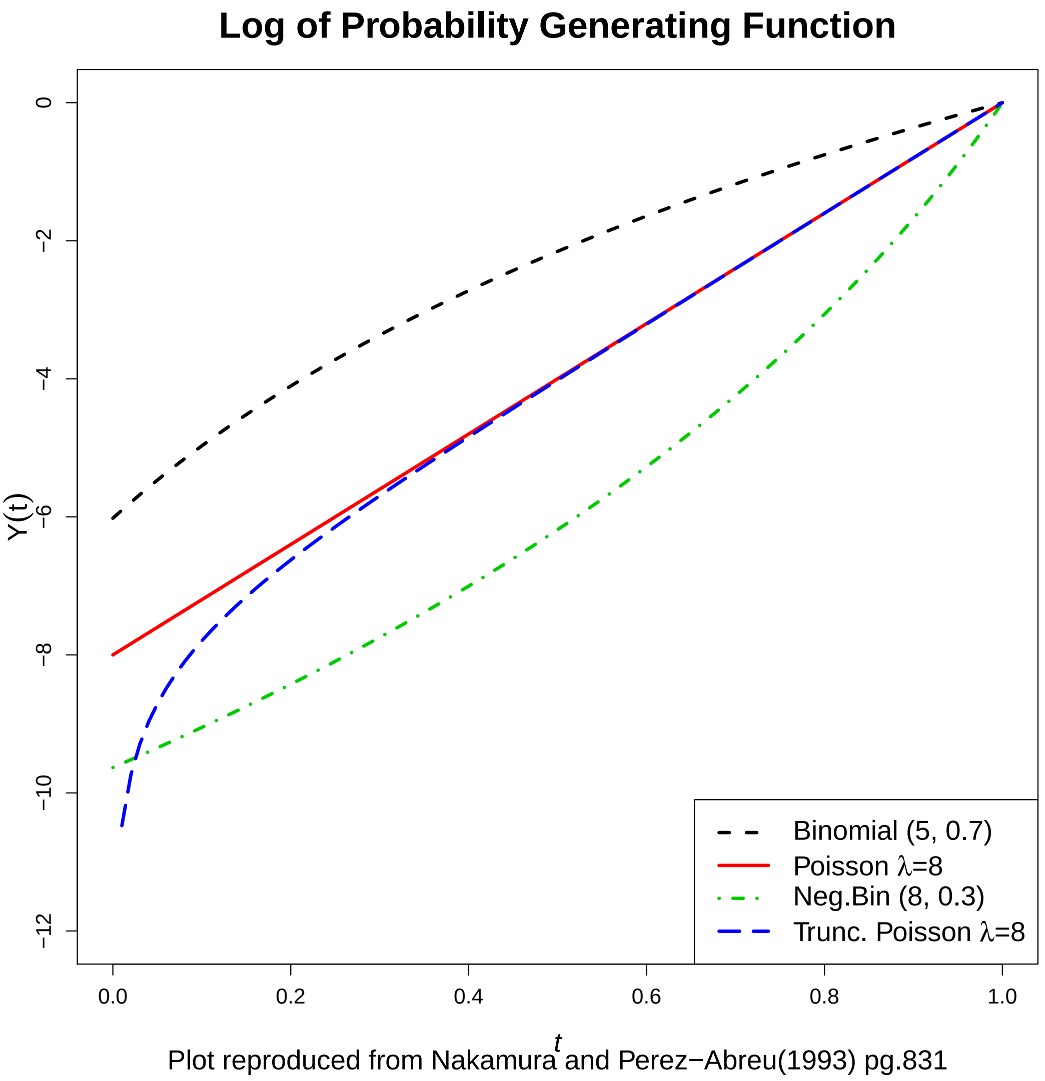
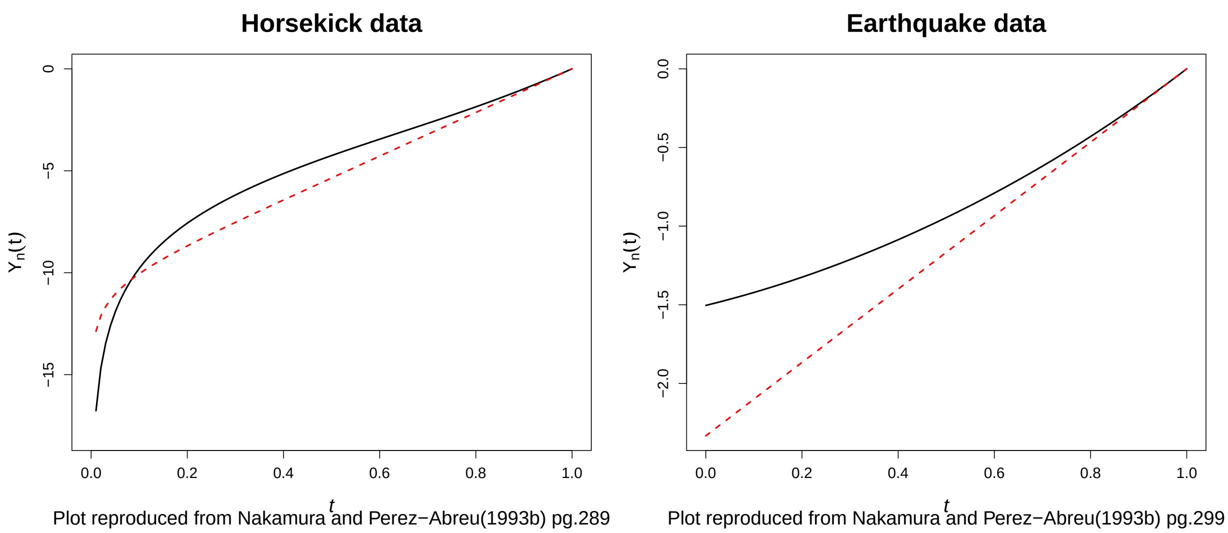


Figure 3.5: Plot of the log of *pgf*'s for a Poisson distribution with $\mu = 8$, Binomial with $n = 5$ and $p = 0.7$, negative binomial with $r = 8$ and $p = 0.3$ and a truncated Poisson distributions with $\mu = 8$.



Plot reproduced from Nakamura and Perez−Abreu(1993b) pg.289

Plot reproduced from Nakamura and Perez−Abreu(1993b) pg.299

Figure 3.6: Plots of the epgf for a) counts of yearly deaths by horse kicks and b) counts of earthquakes in Mexico.

Plots of the log of the EPGF and *pgf*'s provide useful tools in preliminary analysis of count data and allow the comparison of distributions. Nakamura and Pérez-Abreu (1993a) present two examples of the use of the EPGF using previously analysed datasets. The first graph of Figure 3.6 plots the EPGF of the counts of yearly deaths by horse kicks in the Prussian army over a twenty year period between 1875 and 1894 ($n = 20$, min=3, max=18) (Bortkiewicz, 1898). The mean number of deaths by horse kicks is 10.3 (SD=4.51) with median 10.50 (IQR=7.5). The *pgf* for a zero-truncated Poisson distribution is also plotted in the first plot of Figure 3.6 (shown in red) with parameter $\lambda = 10.70$. As t tends to 0, $Y_n(t)$ tends to infinity and suggests the data is from a truncated Poisson distribution. The second graph analyses counts of characteristic subduction earthquakes on Mexico's Pacific coast over periods of ten years between 1806 and 1985 ($n = 18$, min=0, max=7) (Jara and Rosenblueth, 1988). The mean number of earthquakes is 2.33 (SD=2.086) and has median 2 (IQR=2). The second EPGF plot in Figure 3.6 also plots the *pgf* of the Poisson distribution with $\mu = 2.33$. The convex relationship between t and the EPGF $Y_n(t)$ suggests that this dataset is not from a Poisson distribution but displays a mixture of a Poisson or overdispersed behaviour.

This methodology can be extended to compare the EPGF of a dataset with *pgf*'s

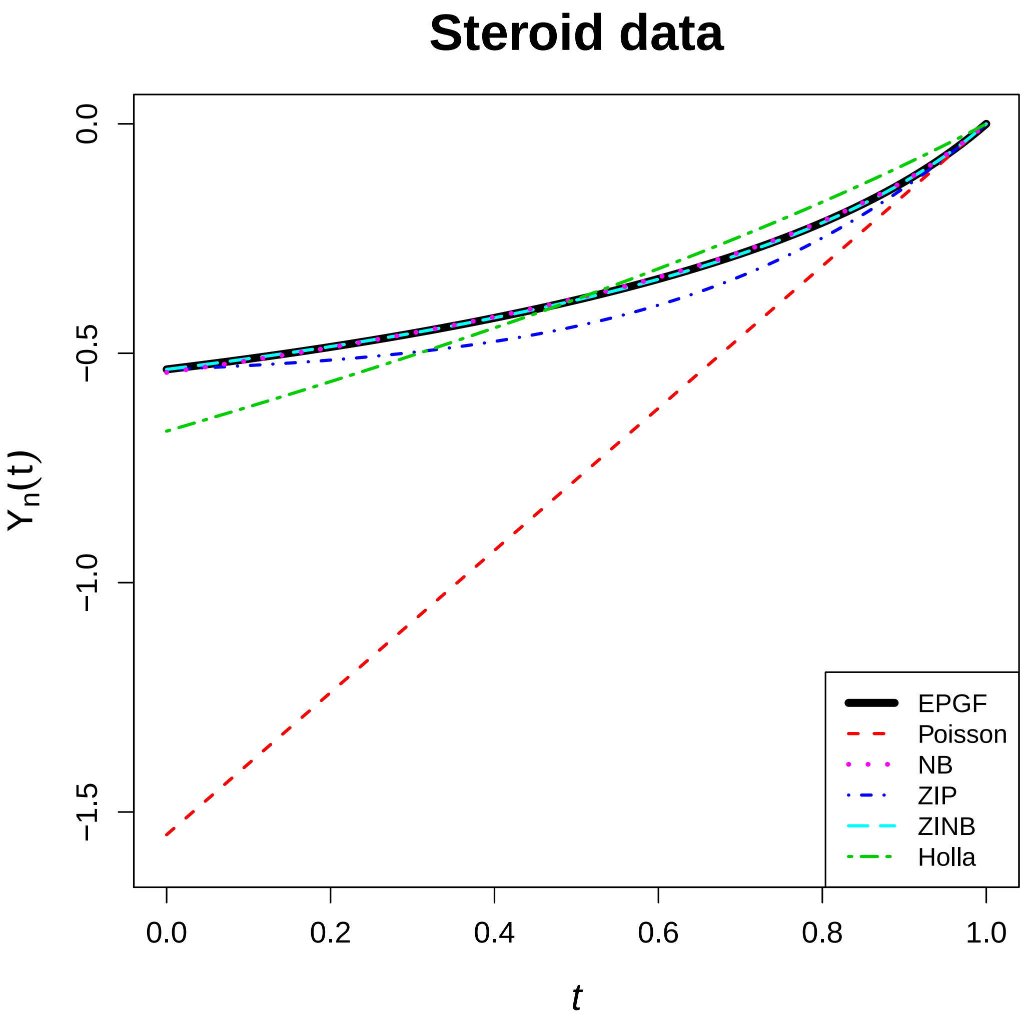


Figure 3.7: EPGF plot of counts of cysts in embryonic mouse kidneys with fitted it pgf's for the Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial and Holla distributions.

calculated using parameter estimates of distributions estimated from the dataset. Plotting both the EPGF and *pgf*('s) on the same graph allows us to compare the fit of a number of discrete distributions using only one plot. The log of the EPGF of the counts of cysts in steroid treated mouse kidneys is plotted in Figure 3.7. The EPGF is shown as a solid black line, whilst a series of distributions fitted to the data are plotted in broken coloured lines. The *pgf* for a Poisson distribution is clearly shown as a straight line (red), as is the Holla distribution (green) which has a smaller gradient. The *pgf*'s for the zero-inflated Poisson (dark blue), zero-inflated negative binomial (light blue) and negative binomial (pink) distributions all are convex and have close fits to the EPGF, with the negative binomial distribution being obscured by the zero-inflated negative binomial which is due to the parameter ω being very close 0. Out of these three distributions, the zero-inflated Poisson does not appear to fit the EPGF as well at the center of the range of t. The EPGF plot therefore indicates that a negative binomial or zero-inflated negative binomial distribution may provide a very good fit to the dataset.

The benefit of the EPGF plot compared with, for example, a histogram of the vector of discrete observations is that it is a continuous plot instead of a function which has jumps at observed data points and is not affected by the occurrence of ties in the data (Nakamura and Pérez-Abreu, 1993b). It also allows the comparison of a range of discrete distributions to an observed dataset.

3.3.3 Outlier Detection

The detection of outliers is especially problematic for discrete distributions where long tails make it difficult to detect outliers. In this section we propose two graphical methods to explore possible outliers in a dataset. The EPGF method graphically detects possible outlying observations by plotting the EPGF using a leave-one-out method and makes no assumptions about the distribution of the dataset. The SI method utilizes the SI by plotting for a fitted distribution over the range of y values to detect outliers.

EPGF Outlier plot

Plotting the EPGF can be useful in detecting possible outlying observations in discrete distributions (Nakamura and Pérez-Abreu, 1993b). If an observation y_i is large in comparison to the other observations, its contribution of t^Y in $G_n(t)$ is small when t is in the range $0 < t < 1$ (Nakamura and Pérez-Abreu, 1993b). Large outlying observations can be examined by evaluating the effect of each observation on $Y_n(t)$ by a 'leave-one-out' procedure, i.e. leave out y_i, use the remaining observations $(n - 1)$ to calculate $Y_{n-1}^{(-i)}(t)$ and plot the resulting n curves on a single plot (Nakamura and Pérez-Abreu, 1993b).

Nakamura and Pérez-Abreu (1993a) utilize a dataset containing frequencies of incidents of international terrorism in the United States between 1968-1974 as an example of the use of the EPGF in detecting outliers (Table 3.3). Figure 3.8 plots $n =$ 76 EPGF curves, where each EPGF is calculated from 75 observations by removing an observation systematically. For most of the curves (full black line) the EPGF increases rapidly for values of t between 1 and 2. However when the observation of 12 incidents is omitted and the EPGF calculated (plotted in a black dashed line) the EPGF curve becomes a straight line. This indicates that this observation has a large effect on the shape of the EPGF when it is removed and hence it can be considered as an outlier with respect to the remainder of the dataset.

Incidents	**0**	**1**	**2**	**3**	**4**	**12**
Frequencies	38	26	8	2	1	1

Table 3.3: Frequencies of incidents of international terrorism in the United States between 1968-1974

Surprise Index

Determining whether or not an observation is an outlier may be problematic for discrete distributions. Weaver (1948)'s SI defined in Section 2.1.5 can be used to assess whether a particular observation can be considered an outlier with respect to the dataset, assuming that the data come from a particular discrete probability model. The SI for

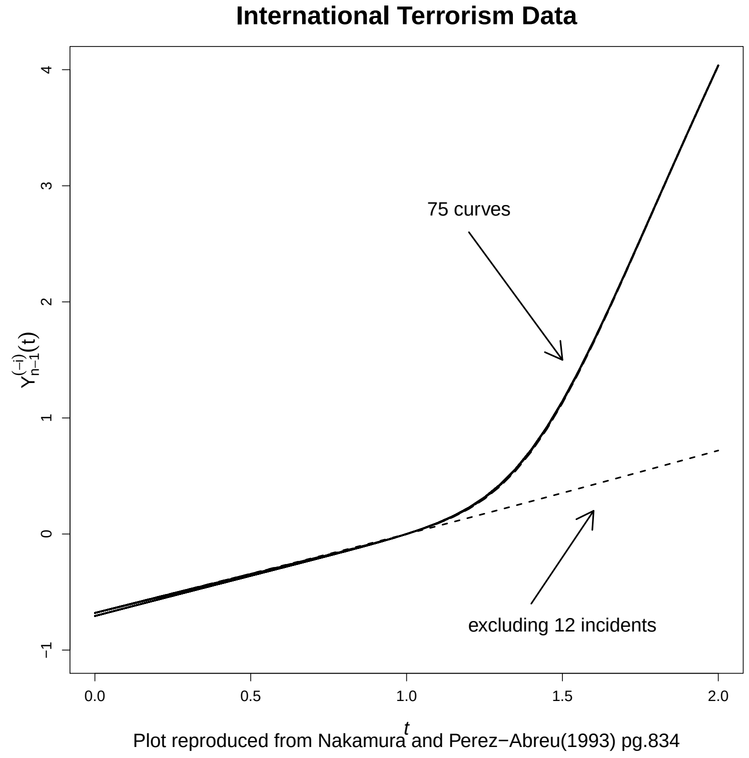


Figure 3.8: EPGF analysis to detect outliers for frequencies of incidents of international terrorism.

the normal case in terms of the standard deviation is presented by (Weaver, 1948). This was extended by Redheffer (1951) to two discrete distributions- the Poisson and binomial cases and are the only examples in the statistical literature. The SI for the distributions presented in Chapter Two of this thesis have all been calculated analytically and the example of counts of cysts in steroid treated embryonic mouse kidneys illustrates the use of the SI in detecting outliers.

The first plot in Figure 3.9 gives the logarithm of the SI for the Poisson distribution calculated using the maximum likelihood estimate of the Poisson distribution for this dataset, $\hat{\mu} = 1.55$. The solid black line plots the logarithm of the SI of the Poisson distribution and the red dashed line indicates values of the SI over 1,000. The plot indicates that counts of cysts greater than 7 are surprising, as their SI is greater than 1,000 and can be considered as outlying observations. A negative binomial distribution is also fitted to the data using maximum likelihood parameter estimates of $\hat{p} = 0.16$ and $\hat{r} = 0.30$ and the SI is plotted in the second graph of Figure 3.9. The dark solid

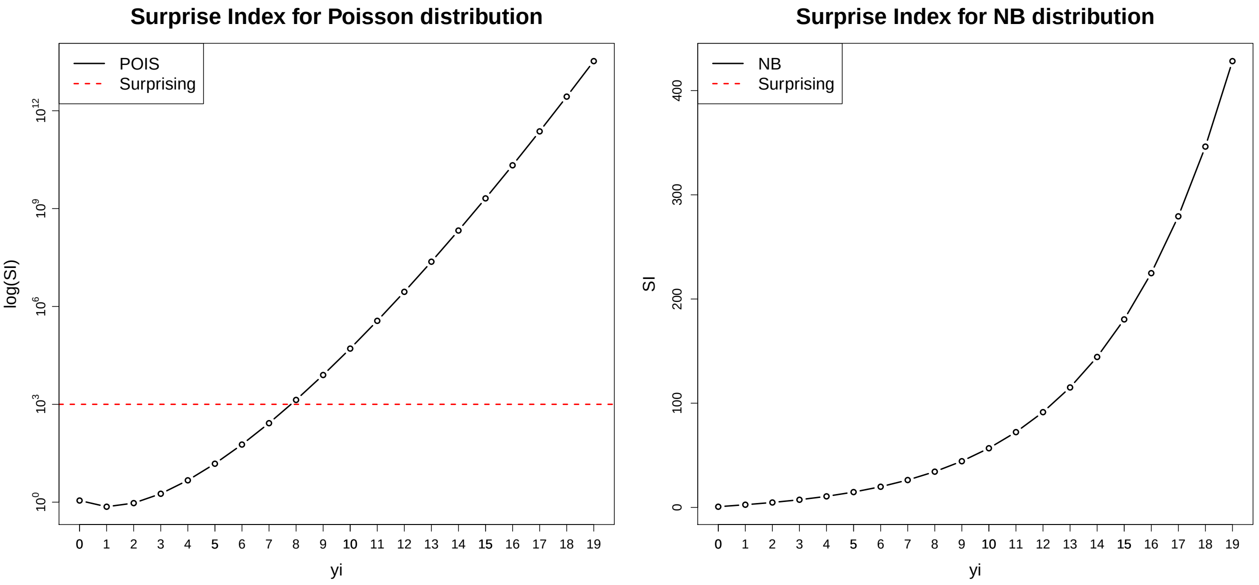


Figure 3.9: Plots of SI for counts of cysts in steroid treated embryonic mouse kidneys for a) a Poisson distribution and b) a negative binomial distribution.

line plots the SI, which is clearly less than 1,000. Therefore, under a negative binomial distribution none of the frequencies of counts of cysts are considered to be surprising.

Summary

This chapter provides a background to estimation methods and frameworks for fitting discrete models described in Chapter 2. Rapid estimation methods give quick estimates of parameters, their main advantage being their use as starting values in maximum likelihood estimation. The method of maximum likelihood provides a consistent approach to parameter estimation, whilst the EM algorithm allows fitting of complex models. The GLM and GAM frameworks offer flexible methods for modelling, however both are limited in their range of distributions which is restricted to those within the exponential family. Distributions in the GAMLSS framework models do not need to belong to the exponential family for this class of models. The GAMLSS framework also allows for location, scale, skewness and kurtosis parameters and has the potential to allow for (almost) any probability density to be used when modelling.

Diagnostic methods for goodness-of-fit, model comparisons and outlier detection for discrete models have also been discussed. The Chi-squared goodness-of-fit test and randomized quantile residuals both provide methods to test the fit of a distribution to a dataset. The AIC and BIC allows comparisons between the fit of distributions and the EPGF can plotted alongside *pgf*'s of distributions to give graphical comparisons. Finally, two different methods for the detection of outliers in discrete distributions have been presented. The EPGF outliers plot is non-parametric, whereas the SI assumes a distribution for the data and requires estimation of the parameters. Software available in statistical environments for the distributions and methods presented in Chapters 2 and 3 will be reviewed in the following chapter.

Chapter 4

Software for fitting discrete probability models

The aim of this thesis is to produce software using the `R` programming language and software environment for statistical computing to analyse models for discrete data. This will provide both statisticians, and clinical and public health scientists with better tools for fitting discrete models. In the next sections we review some software currently available for discrete models and identify potential areas where additional software is required thus highlighting aspects that may benefit from our development.

4.1 Current Software

Statistical software environments differ in their approaches to handling data analysis with some programs allowing command line input, as well as use of graphical user interfaces (GUI). SAS, Stata and PASW (previously known as SPSS) are commonly used examples of statistical packages which provide both command-line and GUI-based analyses. Another approach to statistical software is followed by environments centered on a programming language such as the `R` language and software environment for statistical computing and graphics. The main difference between statistical packages such as SAS, Stata and PASW and languages such as `R` is that this language is object-orientated, meaning that data and methods alike can be stored as 'objects'. We have chosen

R, SAS, Stata and PASW as our primary interest because they are well known and are most frequently used by statisticians or clinicians for scientific research. Two specialized environments, Mathematica (for symbolic computation) and the Altmann fitter (for fitting univariate discrete models), are also chosen as they provide useful computational tools. Many environments have add-on software programs or packages, which provide specialized software routines.

The following sections review software functions available for discrete models within these computational environments. *Functions* refer to routines for statistical analysis or data manipulation. In particular, functions that fit discrete probability densities to single distributions, regression models to identify associations between a discrete outcome and various predictors and goodness–of–fit diagnostics are evaluated.

4.1.1 PASW

Predictive Analytics SoftWare (PASW) previously known as SPSS (Statistical Package for the Social Sciences) (SPSS Inc, 2011) was originally developed in the 1960s as a programming language for conducting statistical analysis and uses both a graphical and a syntactical interface. It provides a range of functions for managing, analysing, and presenting data. The *pdf*'s, (Pdf.-) *cdf*'s, (Cdf.-) and random generating (rv.-) functions for several discrete distributions can be calculated, as shown in Table 4.1.

pdf	*cdf*	**Random generation**
Bernoulli (`Bernoulli`)	Bernoulli (`Bernoulli`)	Bernoulli (`Bernoulli`)
Binomial (`Binom`)	Binomial (`Binom`)	Binomial (`Binom`)
Poisson (`Poisson`)	Geometric (`Geom`)	Geometric (`Geom`)
	Hypergeometric (`Hyper`)	Hypergeometric (`Hyper`)
	Poisson (`Poisson`)	Poisson (`Poisson`)

Table 4.1: Discrete distributions available as *pdf*'s, *cdf*'s and random generations using PASW. Function names are in parenthesis.

Poisson and negative binomial regression models can be fitted within a GLM framework using the GENLIN command. There is a range of optional output statistics for diagnostic analysis: the Chi-squared goodness–of–fit test statistic and *p*-value, log- likelihood, deviance, AIC and BIC. Residuals plots can also be optionally constructed for the fitted model using standardized residuals.

4.1.2 `Stata`

Stata (StataCorp, 2009) is a software package for statistical analysis and provides a wide range of statistical tools and graphical displays. Its `glm` function fits generalized linear models using either maximum likelihood or iteratively re weighted least squares. Models from the exponential family can be fitted, which for discrete response variables are: the Bernoulli or Binomial (`binomial`), Poisson (`poisson`) and negative binomial (`nbinomial`) distributions.

There is also a range of regression models for discrete outcomes: the `poisson` function for fitting for Poisson regression models, `nbreg` for negative binomial regression models or `gnbreg()` for a generalized negative binomial models. Zero-inflated models can be fitted using the `zip` command for zero-inflated Poisson regression and `zinb` for zero-inflated negative binomial regression, where the `inflate()` argument determines the variable list for the zero probability part of the model. Truncated regression models can also be fitted using the commands `tpoisson` and `tnbreg` for zero-truncated Poisson and Negative binomial distributions, respectively. The standard output table for regression models fitted in Stata includes the log-likelihood, deviance and both the AIC and BIC. The `predict` function can calculate raw and standardized Anscombe and Pearson residuals. Stata users can also write their own functions using Stata code.

4.1.3 `SAS`

The Statistical Analysis Systems (SAS) software package has been developed by the SAS Institute since 1976, initially as a project to analyse agricultural research data SAS Institute Inc (2011). It provides a wide range of tools including data management and data mining, report writing and graphics, statistical analysis, alongside many business solution tools such as business planning and forecasting, operations research, project management, data archiving, data storage, web reporting, optimization and quality control.

The PROC GENMOD procedure fits GLM in the SAS software program and can analyse models relating one or several continuous dependent variables to one or several independent variables. This function fits regression models from the exponential family: the binomial, Poisson, geometric and negative binomial distributions. Zero–inflated Poisson regression models can also be fitted using the PROC GENMOD function although it is not strictly a GLM model. The output gives the deviance and Pearson Chi-squared goodness-of-fit tests, log-likelihood, AIC and BIC. Raw Pearson and standardized residuals can also be calculated.

The COUNTREG procedure analyses regression models in which the dependent variable takes count values. The Poisson, negative binomial types I and II , zero-inflated Poisson and zero-inflated negative binomial distributions can be fitted as regression models using maximum likelihood estimation. The output gives the log-likelihood, AIC and BIC, well as parameters estimates and their standard errors.

4.1.4 R

The R language was first developed by Ihaka and Gentleman (1996) as an environment for statistical computing and graphics based on the S-PLUS language (Chambers and Hastie, 1991). A command-driven programming language, R can be used to store and view data, supports many mathematical and statistical functions and provides advanced tools for data analysis and graphical display (Horton et al., 2004). The R project (R Development Core Team, 2009) has been developed since the late 1990s. The software is freely distributed and can be downloaded via the Comprehensive R Archive Network (CRAN) website (CRAN, 2010). The CRAN website features a large amount of background information, documentation and other resources. The R language allows users to write their own functions and one of the main advantage of R is the many packages, also known as libraries, which have been contributed by authors; this allows fitting a wide range of statistical methods beyond the more commonly used functions available in all statistical software packages.

R has two object systems, known informally as S3 and S4 (Chambers, 2008). These systems use object-orientated programming to define the 'class' of an object and then 'method' functions can be associated with a particular type of object. An object in the S3 class is an R object with an additional class attribute, a character vector giving the names of the classes attached to that object. Generic functions can be defined for objects of a certain class. For instance, print is a generic function with alternative definitions for different class types. If a fitted model of class glm is assigned to the object mod, then the command print(mod) will refer to print.glm, the print function for objects of class glm. The S4 class provides an alternative method of attaching classes to objects and can be created using the methods library. A member of the S4 class requires the type of all its components to ensure consistency. In comparison to S3, S4 objects are more rigorous, having a more formal structure.

The R language contains many well-known discrete distributions and provides functions to calculate the *pdf*, *cdf*, quantile and random generating functions for a large range of probability models, with a general notational form adopted to provide a consistent naming scheme. The *pdf* of a distributions name or a shortened version is prefixed by the letter 'd'. Similarly, the *cdf* is prefixed by 'p' and the quantile function 'q'. Functions for random realizations of probability distributions are labelled by prefixing the distribution name with the letter 'r'. Table 4.2 provides a list of discrete distributions available in the base library (part of the core R instillation) of the R language. For example, the first distribution in the table is the Binomial distribution with parameters size and prob which has *pdf* function dbinom, *cdf* function pbinom, quantile function qbinom and random generating function rbinom. Other discrete probability distributions available in add-on libraries are detailed below.

Generalized linear models can be implemented in R using the glm() function, which can fit distributions from binomial, Poisson, geometric, quasi-Binomial and quasi-Poisson distributions using a GLM framework. The glm function returns objects of class glm for which there is a number of generic functions. The summary() function returns summary statistics of a model, including covariate parameter estimates

Distribution	Name	Functions	Parameters
Binomial	`binom()`	(pdqr)	`size`, `prob`
Poisson	`ppois()`	(pdqr)	`lambda`
Geometric	`geom()`	(pdqr)	`prob`
Negative Binomial	`nbinom()`	(pdqr)	`size`, `prob`, `mu`
Hypergeometric	`hyper()`	(pdqr)	`m`, `n`, `k`

Table 4.2: Probability distributions available in the base library of the `R` language

with corresponding Wald tests, the log-likelihood, deviance, AIC and BIC. Raw, Pearson and standardized residuals can be calculated for fitted models using `residual()`. The function `predict` generates predictions from the results of various model fitting functions and `plot` is a generic function for plotting of `R` objects.

A number of models for discrete data are available in `R` through add-on libraries. Libraries in `R` are developed independently by `R` users and therefore there is some overlap in their contents. Several `R` libraries are presented in the following sections which include functions for the analysis of discrete data.

`stats4` **library**

The `stats4` library is available as part of the `R` language environment and provides `S4`-class statistical functions. The `mle` function estimates parameters by maximum likelihood using `R`'s general purpose optimization function, `optim`. This function has usage `mle(minuslogl, ...)`, where `minuslogl` is a function of the negative log-likelihood; the arguments indicated as `'...'` refer to additional ones passed to subsidiary functions in the `mle` call. Objects resulting from this function have class `mle-class` with general methods including: `logLik` which extracts the log-likelihood, `vcoc` which extracts the variance–covariance matrix, `profile` generates the profile likelihoods of the models parameters and `summary` gives a summary of the maximum likelihood estimation including the parameter estimates and model deviance which is minus two times the log-likelihood.

MASS **library**

The MASS library contains functions and datasets supporting the classic text *Modern Applied Statistics with* S-PLUS by Venables and Ripley (2002). Particularly important is that it allows fitting regression models for the negative binomial distribution within the GLM framework using the glm.nb() function. Objects resulting from glm.nb function inherit the glm class.

pscl **library**

The pscl package was developed by Jackman (2010) and contains the zeroinfl() function which can be used for maximum likelihood estimation of zero-inflated models. This function fits regression models using the Poisson, Geometric and Negative Binomial models and allows for zero-inflation to be accounted for in the model either as a constant or including covariates (Zeileis et al., 2008). The function uses maximized likelihood estimation but can also generate parameter estimates using the EM algorithm by setting EM=T. The returned fitted model object is of class zeroinf and is similar to fitted glm objects, the output therefore provides the standard summary and goodness-of-fit statistics.

zicounts **library**

The zicounts package provides an alternative implementation of classical and zero-inflated count data regression models (Mwalili, 2007). The function zicounts() allows for Poisson, zero-inflated Poisson, negative binomial and zero-inflated negative binomial models, with estimates generated using maximum likelihood. There are also functions for regression models for censored count data, zicensor, for the Poisson, zero-inflated Poisson, negative binomial and zero-inflated negative binomial models, where the upper bound response variable is known. This library provides similar models to those in the pscl library, however the interfaces of the zicounts and zicensor functions are less standard, having no class attributed to the output and no generic functions associated with these models (Zeileis et al., 2008).

gamlss family of libraries

The original gamlss package (Stasinopoulos and Rigby, 2008, 2007) was developed to support the generalized additive models for location, scale and shape (GAMLSS) framework of regression models (Rigby and Stasinopoulos, 2005) (see Section 3.2.3); the family of gamlss libraries consists of number of packages related to this framework. The gamlss.dist library contains the p, d, q, r and gamlss family functions for a large range of continuous and discrete probability distributions. The gamlss.cens library provides procedures for fitting censored response variables, whilst gamlss.mx contains algorithms for fitting finite mixture models and the gamlss.tr library can fit truncated models. The gamlss.cens, gamlss.mx and gamlss.tr all fit models using the distributions supplied in gamlss.dist. All available gamlss packages are installed when loading the original gamlss package into an R session.

Distribution	Function	No. of Parameters
Beta Binomial	BB()	2
Binomial	BI()	1
Delaporte	DEL()	3
Negative Binomial I	NBI()	2
Negative Binomial II	NBII()	2
Poisson	PO()	1
Holla (Poisson-Inverse Gaussian)	PIG()	2
Sichel	SI()	3
Sichel (μ the mean)	SICHEL()	3
Zero-altered beta binomial	ZABB()	3
Zero-altered Binomial	ZABI()	1
Zero-altered negative binomial	ZANBI()	2
Zero-inflated beta binomial	ZIBB()	3
Zero-inflated Binomial	ZIBI()	2
Zero-inflated negative binomial	ZINBI()	3
Zero-inflated Poisson	ZIP()	2
Zero-inflated Poisson (μ the mean)	ZIP2()	2
Zero-inflated Holla (ZI Poisson-Inverse Gaussian)	ZIPIG()	3

Table 4.3: Discrete distributions implemented within the gamlss.dist library (Stasinopoulos and Rigby, 2007).

Table 4.3 lists the discrete distributions that can be implemented within the gamlss package. All the distributions in Table 4.3 have p, d, q, and r functions giving the *pdf*, *cdf*, quantiles and random generating functions, respectively. Each distribution also has a GAMLSS family fitting function which provides link functions, first and second derivatives, starting values etc. needed for the fitting procedure in the `gamlss()` function (Stasinopoulos and Rigby, 2007). The arguments of the fitting functions specify the link functions for each of the distribution parameters. The negative binomial type II distribution specified in Section 2.3.1 of Chapter 2 has parameters μ and α, whilst in the `gamlss` library the NBII distribution has parameters μ and $\sigma = \frac{1}{\alpha}$. The fitting function for the negative binomial type II distribution is `NBII()` and has two parameters `mu` and `sigma`, both with default log link functions.

The `gamlss()` function in the `gamlss` library estimates the parameters of regression models using the GAMLSS framework using the methods described in Section 3.2.3. Objects from `gamlss()` fitting have class `gamlss` which have an associated set of generic functions. The `summary()` function returns a standard summary of various statistics of a model, including parameter estimates, log-likelihood, deviance, AIC and BIC (known as the SBC) as part of their output. Randomized quantile residuals can be calculated for fitted models `residual()` and worm plots `wp()` provides a diagnostic tool for checking the residuals within different ranges of the explanatory variables.

`VGAM` **library**

The Vector Generalized Additive Models (VGAM) package implements regression models which use vector generalized linear and additive models (Yee, 2008). The `vglm()` function can be used to fit generalized linear models for the Binomial, `binomialff()`, Poisson `poissonff()` and quasi-Poisson `quasipoissonff()` distributions. Vector generalized additive models can be fitted using the `vgam()` function for distributions by specifying the family argument as a `VGAM` family function. A range of discrete distributions available as family functions are specified in Table 4.4. Distributions with `pdqr` functions available are given in parenthesis. Expressions for Lerch's Phi $\Phi(s, z, v)$, `lerch()`, and Reimann's Zeta function $\zeta(x)$, `zeta()` are also available.

Distribution	Function
Beta-Binomial	`betabinomial()` (dpr)
Beta-Binomial	`betabin.ab()` (dpr)
Generalized Poisson	`genpoisson()`
Geometric	`geometric()`
Negative Binomial	`negbinomial()`
Poisson	`poissonff()`
Poisson-Poisson mix	`mix2poisson()`
Positive negative binomial	`posnegbinomial()` (dpqr)
Postive Poisson	`pospoisson()` (dpqr)
Zeta	`zetaff()`
Zero-altered Negative Binomial	`zanegbinomial()` (dpqr)
Zero-altered Poisson	`zapoisson()` (dpqr)
Zero-inflated Binomial	`zibinomial()` (dpqr)
Zero-inflated Negative Binomial	`zinegbinomial()` (dpqr)
Zero-inflated Poisson	`zipoisson()` (dpqr)
Zero-inflated Poisson	`yip88()`
Zipf	`zipf()` (dp)

Table 4.4: Discrete probability distributions available in `VGAM` library of `R`

The `vgam()` function returns objects with class `vgam`, with generic functions including:`summary()` producing a table of summary statistics including the parameter estimates, log-likelihood and deviance, `residuals()` which gives the residuals.

`zipfR` library

The `zipfR` library provides tools for the analysis of word frequency distributions, including frequency estimation for rare events and functions for plotting word frequency data and vocabulary growth curves (Evert and Baroni, 2008). Zipf models have been applied to many different areas aside from linguistics, e.g. genetics, human geography. More information on the `zipfR` package is available on the `zipfR` website (zipfR, 2010).

Models for word frequency distributions belong to a family of large number of rare events (LNRE) models and can be implemented in `zipfR` using `lnre()`. Currently the Zipf-Mandelbrot (`ZM`), finite Zipf-Mandlebrot (`fZM`) and Generalized Inverse Gauss-Poisson (`GIGP`) (Sichel) models have been programmed. The probability density functions for these distributions can be defined using `dlnre()`, the distribution function `plnre()`, the quantile function qlnre and random sample generation rlnre().

The R language provides software for a good number of discrete distributions across a number of add-on libraries. There are less software routines for the estimation of parameters for discrete regression models. One disadvantage of user-contributed add-on libraries is the resulting overlap of many procedures in the R language. The advantage of the R language environment is that the S3 and S4 classes system provides generic functions which can give summaries and residuals of models.

4.1.5 MATHEMATICA

MATHEMATICA (Inc, 2009) is a command-based application comprised of a symbolic programing language which allows performing complicated algebraic tasks and to create graphics. One package available as an add-on to MATHEMATICA is MATHSTATICA (Rose and Smith, 2002) which uses the MATHEMATICA interface to provide a toolset for mathematical statistics. MATHSTATICA provides the *pdf*'s for the following discrete distributions: Bernoulli, Beta-Binomial, Binomial, Discrete uniform, Geometric, Hypergeometric, Logarithmic, negative binomial, Poisson, Riemann Zeta, Waring, Yule and ZIP. Maximum likelihood estimates can be derived analytically by maximizing the log likelihood of the distributions (Rose and Smith, 2000; Currie, 1995).

4.1.6 Altmann Fitter

The Altmann-fitter (Altmann, 1997) was developed by Gabriel Altmann and fits univariate discrete probability distributions to frequency data. There are 200 discrete distributions currently implemented in this program with applications ranging from the fields of biology and ecology, to economy and linguistics. Wimmer and Altmann (1999) developed a Thesaurus to detail these distributions and many others, their origins and uses.

Distribution parameters are estimated using rapid estimation, often using an iterative procedure. Different methods of rapid estimation are compared to find the best parameter estimates for a particular discrete distribution. Distributions can be fitted singularly, or the range of 200 distributions compared using the chi-squared goodness-of-fit test *p*-value. Predicted values of the range of the observed values can also be calculated and graphics produced plotting observed and fitted values for the models. However, in general the Altmann Fitter does not provide maximum likelihood estimates nor standard errors for the estimates.

4.2 Gaps in methodology

There are several aspects of discrete modelling that have been covered in this thesis, including: discrete probability densities, parameter estimation of univariate models (containing no covariates), regression modelling and model diagnostics. Currently, a large variety of discrete models are available in a number of software packages, although many of the more complex models such as the Hermite or Gegenbauer models, have received less attention in the literature and are not widely available in statistical software packages, if at all. The *pdf*'s or statistical properties of distributions such as the Yule, Waring, and beta-binomial distributions, and many families of distributions such as the generalized Poisson family of distributions which includes the Neyman Type A, Hermite, Generalized Hermite, Gegenbauer and Generalized Gegenbauer or the Lerch family including the Zipf, Zeta and Good distributions, are not currently available in any software packages other than the Altmann Fitter. These models allow for overdispersion, value-inflation and/or long tails which may improve the fit of a dataset and also provide valuable information about the data generating mechanism which yields the data. It is therefore important that a range of distributions are available for modelling to realize the full potential of a dataset. The lack of implementation of such distributions in statistical software packages limits the user in their choice of discrete distribution. The Altman Fitter provides rapid estimates for univariate discrete models, however unlike the object-orientated `R` language there is no flexibility within this program for inferences made with the results. There is therefore a need for more complex discrete models to be made available via open-source software in order that these distributions may be used by statisticians, researchers and clinicians to facilitate interpretation of epidemiological and clinical datasets.

There is even less software available for regression modelling of discrete outcome random variables with non-standard distributions. The standard Poisson and negative binomial distributions are widely available as regression models in many statistical packages. Zero-inflated and censored versions of these distributions are available in Stata , SAS and the `pscl` and `gamlss` libraries in `R`. Routines for fitting GLM's and GAM's can be used to estimate parameters in the `R` language, SAS, Stata and PASW, but these are limited to those of the exponential family. There are a number of discrete distributions available via the `gamlss` library (Table 4.3) which can fit regression models within the GAMLSS framework. These models all include standard goodness-of-fit statistics such as the log-likelihood and deviance alongside parameter estimates and many include some of the Chi-squared goodness-of-fit test, the AIC and BIC. Functions for calculating various types of residuals and plots for residual analysis for models are standard across all of the statistical environments presented in this chapter.

Where a range of models are fitted to a dataset deciding upon the optimum fitting model is an important factor in data analysis, as this ensures that the maximum information is extracted from the data. Although a large number of distributions can be fitted in many of the statistical environments discussed in this chapter, there is no simple way to compare the goodness-of-fit of two or more distributions. For example, when deciding upon the best distribution to fit a particular dataset, there is no convenient way to fit the models and extract only the goodness-of-fit statistics such as the AIC, BIC or Chi-squared test from the output in order to compare the models. Instead, each distribution would need to be fitted to the dataset separately and the relevant goodness-of-fit statistics extracted. The Altmann Fitter has an automatic procedure which estimates the parameters from a large number of distributions and returns the goodness-of-fit statistics in an ordered table. However, the Altmann Fitter only provides rapid estimates for the parameters from probability models and does not extend to a regression setting.

There are therefore three areas of discrete modelling which have been highlighted as benefiting from software development:

1. **Univariate distributions: parameter estimation and model comparisons**

   A set of programs are required to calculate the properties of a range of distributions including the *pdf*, *cdf*, quantile and random generating functions and also the *pgf*'s, moments and SI. Routines for the estimation of parameters in a univariate setting will be performed using maximum likelihood, with rapid estimates providing starting values for the algorithm. The fit of a range of discrete distributions to a dataset will be compared using the Chi-squared goodness-of-fit tests, AIC and BIC values.

2. **Goodness-of-fit tests and model diagnostics**

   There are three issues with assessing the fit of discrete distributions: the goodness-of-fit of a particular distribution to a dataset, model comparisons and outlier detection. The Chi-squared goodness-of-fit test is frequently included in model output tables across all of the software packages. The AIC and BIC are commonly used criteria for model comparisons and will be included in the library. Residual analysis also plays a key role in determining the fit of a model to a dataset and there is a need for residuals for discrete observations. Both of these techniques will be included in the library as methods for the analysis of the fit of distributions to data. Methods for comparing distributions are needed in order to compare the fit of multiple models to a dataset. There is also a need for outlier detection methods particularly suited to discrete data.

3. **Discrete regression models**

   A small range of discrete distribution regression models can be fitted with current statistical software. There is therefore a need for discrete distributions that are not already available for fitting as regression models. The `gamlss` library provides a limited range of discrete distributions, together with procedures for parameter estimation and model diagnostics such as goodness-of-fit statistics and residual analysis. The GAMLSS framework allows users to create their own distributions which can then be fitted using the gamlss function and will be utilized to implement discrete distributions in gamlss.

4.3 Outline of software

The project's main aim is to provide software to fit and analyse discrete data. The R program for statistical computing (R Development Core Team, 2009) can be used to create add-on libraries containing modelling tools for discrete datasets. R has many advantages as a platform to develop new statistical software. The R language is very flexible and lends itself particularly well to the development of new functions, with the S3 and S4 frameworks allowing users to develop generic functions for model classes e.g. summary, residual, plot, that are common functions for most models fitted across R libraries. It provides users with the ability to produce and publish libraries or packages of their own code. These libraries can then be made available to other R users through the Comprehensive R Archive Network (CRAN) (CRAN, 2010) website cran.r-project.org or sourceforge.net to enable a wide range of accessibility to the software. Another benefit of the R program is that it is free to download under the terms of the Free Software Foundation's GNU General Public License. R is increasingly widely used in many fields and has developed a large, worldwide community of users.

The software produced as part of this thesis can be divided into three R libraries, which will include estimation methods, diagnostic and model selection tools for analysing discrete data. They will provide clinicians and researchers within the fields of clinical and population science with tools to fit and interpret complex statistical models with increased ease which may improve the understanding of clinical aspects of disease. The following three sections outline the contents of these libraries.

`Altmann Library` **– Univariate parameter estimation and model comparison**

Univariate parameter estimation for a range of discrete distributions will be the main focus of this library. A number of distributions will be included to allow for more complex analysis of datasets, such as zero-inflation, truncation, long-tailed distributions and other families of distributions. These functions will utilize rapid estimation and maximum likelihood estimation methodologies. Fitted values are obtained for models fitted and functions to plot the fit of distributions will be included. Comparison tools will allow for a large number of distributions to be compared simultaneously. The AIC, BIC and Chi squared goodness-of-fit test p-value will be used to compare the fit of models.

`discrete.diag` **– Model diagnostics**

A Chi squared goodness-of-fit test will be provided to test the fit of a distribution to a dataset. Residuals plots analyses using randomized Quantile residuals will be implemented. Two functions to calculate the AIC or BIC and the plot the EPGF as a graphical tool allowing for model comparisons, are included. For the detection of outliers in discrete data, the SI and the EPGF methods presented in Sections 3.3.3 and 3.3.3 of Chapter 3 provide plots for identifying potentially outlying observations.

`discrete.reg` **– Regression modelling**

The class of generalized additive models for location, scale and shape (GAMLSS) is a useful framework to develop regression models for the variety of distributions described above. The `gamlss` `R` library can be extended to incorporate further distributions. The Geometric, Yule, and Waring distributions have been defined as `gamlss.family` objects to allow for regression modelling within the GAMLSS framework.

Summary

This review of statistical computing environments has demonstrated the current variety of procedures available to analyse discrete data . Areas identified as requiring development in software include the estimation of parameters of discrete distributions using maximum likelihood procedures and also methods for performing comparisons between models. There is also scope to improve the range of discrete distributions available in the `gamlss` library within the `R` language. The next three chapters will present each of the three `R` libraries which provide a toolkit of methods for discrete data: the `Altmann`, `discrete.diag` and `discrete.reg` libraries.

Chapter 5

Altmann Library

This chapter details the R software library developed for univariate parameter estimation of discrete distributions and tools for model comparison, called the 'Altmann library'. This library gets its title from the Altmann Fitter software package (Altmann, 1997). The purpose of this add-on R library is to enable parameter estimation for univariate discrete distributions and facilitates comparisons between the fit of distributions.

The first section of this chapter details several datasets which are included in the Altmann library. Functions to calculate the probability density, cumulative density, quantile and random generations for each distribution will then follow. In the third section, the maximum likelihood estimation functions are explained and in the fourth section plot functions for maximum likelihood models in the Altmann library are presented. This is followed by the altmann.fitter function for comparisons of discrete models. Throughout the first five sections of this chapter the negative binomial distribution is used as an example of the implementation and frameworks of functions in the Altmann library. In the final section in this chapter the implementation of the functions available in the Altmann library is applied to the UK surnames distribution presented in Section 1.2.1 of Chapter one by fitting univariate Zipf distributions to surname frequencies across county districts.

The example of the counts of stillbirths in New Zealand white rabbits (Morgan et al., 2007) illustrates the usage, arguments and outputs of the functions presented in this library in a practical setting. The number of stillbirths in 402 litters of New Zealand white rabbits is shown in Table 5.1. The distribution is seemingly zero-inflated with 78.1% of the litters having no stillbirths and overdispersion is clearly present as the variance (1.51) is much larger than the mean (0.46).

5.1 Datasets

Several discrete datasets are included in the Altmann library and are used as examples in the library help files. These datasets can be loaded into `R` using the `data()` function. The five discrete datasets are as follows:

1. `rabbits`

   The `rabbits` dataset consists of the frequencies of stillbirths in 402 New Zealand white rabbit litters originally discussed in the context of Score Tests by Morgan et al. (2007) (Table 5.1).

No. of Stillbirths	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Frequency	314	48	20	7	5	2	2	1	2	0	0	1

   Table 5.1: Frequency of stillbirths in litters of New Zealand white rabbits

2. `lakota`

   This discourse data come from the Native American language Lakota where the frequency distribution of linguistic items is defined by their length. The variable represents the number of phonemes a linguistic item (word) contains (Pustet and Altmann, 2005) from 1959 words, shown in Table 5.2. Within the grammatical systems of natural languages, zero morphemes are frequently found i.e. morphemes which lack phonetic substance and thus have length 0. The characteristics of Lakota syllable structure automatically lead to a multimodal distribution having modes at even values with blurring at higher values of Y.

No. of phonemes	**0**	**1**	**2**	**3**	**4**	**5**	**6**
Frequency	461	57	524	169	370	106	115
No. of phonemes	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Frequency	41	50	47	12	5	1	1

   Table 5.2: Counts of morpheme length in lakota language

3. `yeast`

   A historic dataset of 400 haemocytometer counts of yeast cells, this has been analyzed by Neyman (1939) and Plunkett and Jain (1975) in the context of Generalized Poisson models. The distribution of counts of yeast cells are shown in Table 5.3 .

Counts of yeast cells	**0**	**1**	**2**	**3**	**4**	**5**
Frequency	213	128	37	18	3	1

Table 5.3: Counts of yeast cells

4. `household`

   Data on household size taken from the Housing Allowance Demand Experiment is presented in Hoaglin and Tukey (1985) and analyzed using EPGF plots by Nakamura and Pérez-Abreu (1993b). Table 5.4 gives the distribution of household size from 1239 households.

Household size	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Frequency	210	315	292	176	125	57	38	18	6	1	0	1

Table 5.4: household size from Housing Allowance Demand Experiment

5. `surnames`

   This dataset presents a table of the frequency of surnames across eight non-overlapping districts, shown in Table 5.5. This data is analyzed by both Zörnig and Altmann (1995) and and Panaretos (1989) to fit truncated discrete models.

Frequency	District							
	1	2	3	4	5	6	7	8
1	832	329	292	243	234	281	349	282
2	151	43	28	17	17	23	30	34
3	39	11	6	4	4	9	73	11
4	20	1	2	2	2	1	1	2
5	11	0	0	0	0	0	0	0
6	2	1	0	0	0	0	0	0
7	4	0	1	0	0	0	0	0
8	5	0	0	0	0	1	0	0
9	0	1	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	2	0	0	0	0	0	0	1

Table 5.5: Frquency of surnames across eight non-overlappping districts

5.2 Summary of discrete datasets

The summary.disc function produces summaries of discrete datasets. This function is based on the summary function in the base library of the R environment and gives descriptive summary statistics, moments and various measures for discrete data including the overdispersion index, zero-inflation index and Gini coefficient. A table of frequencies and probabilities is also given. The function has usage,

```
summary.disc(object, ..., digits = max(options()$digits - 3, 3))
```

where the argument object is a vector of discrete data for which a summary is desired, ... gives additional arguments affecting the summary produced and digits is an integer and specifies the number of digits for statistics produced by the function. This is set to a default value which is greater than or equal to 3. The code for this function is given in Listing 5.1.

```
summary.disc<-
function(object, ..., digits=max(options()$digits-4,2))
{
    require(reldist)
    if(length(levels(object)))
        return(summary.factor(object, ...))
```

```
 7    else
 8    {
 9        nas<- is.na(object)
10        object<- object[!nas]
11        n<- length(object)
12        m1<- mean(object); m2<- var(object)
13        x<- unique(object); fx<- table(object)
14        px<- (fx/n)*100; od<- m2/m1
15        p0<- sum(object==0)
16        m3<- sum( fx*(x-m1)^3)/n; m4<- sum( fx*(x-m1)^4)/n
17        sk<- m3/(m2^1.5); ku<- m4/(m2*m2)
18        kappa3<- m3/m1-1; zi<- 1 + log(p0)/m1
19        gini.coeff<- gini(object)
20        qq <- quantile(object)
21
22        if(any(nas)){
23           qq<-signif(c(qq, n, sum(nas)), digits)
24           names(qq) <- c("Min", "1st␣Q", "Median",
25                          "3rd␣Q", "Max", "n", "NA's")
26        }
27        else {
28           qq<-signif(c(qq, n), digits)
29           names(qq) <- c("Min", "1st␣Q", "Median",
30                          "3rd␣Q", "Max", "n")
31        }
32        moms<- signif(c(m1, m2, sqrt(m2), m3, m4, sk, ku),
33                    digits)
34        names(moms)<- c('mean','var','stddev','m3',
35                      'm4','sk','ku')
36        extras<- signif(c(od, kappa3, zi, gini.coeff),
37                      digits)
38        names(extras)<-c('OD','kappa3','ZI','Gini')
39        tab<- rbind(fx,signif(px,digits))
40        dimnames(tab)<- list(c('freq','\%'),x)
41        value<- list(desc=qq, moms=moms, extras=extras,
42                    tab=tab)
43    }
44    value
45  }
```

Listing 5.1: Summary function for discrete datasets

The first line in summary.dist loads the reldist library if it is not already available in R, which contains functions to calculate the Gini coefficient. In lines 5-6 if the data object contains levels, i.e. categorical data, then a summary of the data is returned using the summary function. Otherwise, the function then procedes to calculate a range of summary staistics and indicies in Lines 9-20. Four tables of summary statistics are provided by this function. Firstly, in lines 22-31 a table of quantile values are calculated for the data vector `object`, which are the minimum, 25% lower quartile, median, 75% upper quartile, maximum and the number of observations in the data vector. A table of moment statistics is calculated in lines 32-35 and includes the mean, variance, standard deviation, third and fourth sample moments, skewness and kurtosis coefficients. The third table (lines 36-38) contains the overdispersion index, kappa3, zero-inflation index and Gini's coefficient which are measures for discrete data. The final table constructed in lines 39 and 40, gives the observed frequencies and probabilities for the range of discrete values of `object`. In lines 41-44 these tables are then returned as the output of this function.

The application and output of this function can be illustrated using the numbers of stillbirths in New Zealand white rabbits presented in Table 5.1; `R` code for this example is shown below.

```
> data(rabbits)
> summary.disc(rabbits)
$desc
   Min  1st Q Median  3rd Q    Max      n
     0      0      0      0     11    402
$moms
   mean     var  stddev      m3      m4      sk      ku
   0.46    1.51    1.23    8.01   61.10    4.31   26.80
$extras
    OD kappa3     ZI   Gini
 3.283 16.410 13.490  0.865
$tab
         0    1     2    3    4     5
freq 314.0 48.0 20.00 7.00 5.00 2.000
prob  78.1 11.9  4.98 1.74 1.24 0.498
         6     7     8    11
freq 2.000 1.000 2.000 1.000
prob 0.498 0.249 0.498 0.249
```

The number of stillbirths is in the range (0,11), with the lower 25% quantile, median and upper 75% quantile all having value 0. This is due to the large amount of zeros present in the data (78.11%) and is supported by a very high ZI index of 13.49 (a ZI value of 0 indicates no ZI is present). The mean is 0.46 with variance

1.51, indicating that the distribution of the number of stillbirths is overdispersed. The high value of 3.28 for the OD index in the thrid table is greater than 1 again indicating overdispserion is present in the data. The skewness coefficient gives a measure of symmetry in the distribution and a positive value of 4.32 indicates a distribution with a long right tail. Similarly, the kurtosis coefficient is also high and gives a measure of peakedness. For this dataset the large positive value shows the distribution is 'leptokurtic' with a peak near the mean and heavy tails. The κ_3 statistic is also high and provides another measure of skewness in the data. Finally, Gini's coefficient measures the size of differences between observations and the small value for this dataset indicates a long tail.

5.3 `pdqr` for distributions

This section presents examples of the probability density function, `d`, the cumulative density function, `p`, the quantile or inverse *cdf* function, `q` and random generating function `r`. The `pdqr` functions have been created for the range of discrete distributions described in the literature review in Chapter 2 and are summarized in Table 5.6. The table shows the shortened name used as a suffix for each distribution, alongside the parameters of the distribution. For example the first entry, the Bernoulli distribution, is shortened to `BER` and hence the *pdf*, *cdf*, inverse *cdf* and random generating functions are labeled `dBER`, `pBER`, `qBER` and `rBER`, respectively, and has parameter `prob`. The negative binomial distribution will be used in this section to provide an example of the implementation of these functions. The next four sections introduce the code for the `d`, `p`, `q` and `r` functions.

5.3.1 Probability density function `d`

`R` code for the *pdf* of the negative binomial distribution is given in Listing 5.2 and has the following usage,

```
dNB(y, r = 2, p = 0.5, log.p = FALSE)
```

Name	Suffix	Parameters
Beroulli	BER	`prob`
Binomial	BIN	`prob, n`
Poisson	POIS	`mu`
Geometric	GEO	`prob`
Negative Binomial	NB	`r, p`
Hypergeometric	HY	`N, M, n`
Holla	HO	`alpha, theta`
Sichel	SICH	`alpha, theta, gama`
Delaporte	DE	`mu, sigma, nu`
Yule	YU	`lambda`
Waring	WA	`b, n`
Beta-binomial	BBI	`a, b, n`
Zero-inflated Poisson	ZIPO	`omega, mu`
Zero-inflated Negative Binomial	ZINB	`omega, r, p`
Zero-inflated Sichel	ZISI	`omega, alpha, theta, gama`
2par Poisson	2PO	`omega, mu, lambda`
Poisson-Negative Binomial mix	2PNB	`omega, mu, r, p`
Zero-truncated Poisson	PPO	`mu`
Zero-truncated Geometric	PGE	`prob`
Zero-truncated Negative Binomial	PNB	`r, p`
Zero-truncated Holla	PHO	`alpha, theta`
Zero-truncated Sichel	PSI	`alpha, theta, gama`
Zero-truncated Yule	PYU	`lambda`
Lerch	LE	`p1, a1, c1`
Zeta	ZE	`c1`
Zipf	ZIPF	`a1, c1`
Good	GO	`p1, c1`
Neyman Type A	NYA	`mu, phi`
Hermite	HE	`a, b`
Generalized Hermite	GHE	`a, b, m`
Gegenbauer	GE	`a, b, k`
Generalized Gegenbauer	GGE	`a, m, alpha, beta`

Table 5.6: Probability distributions available in the Altmann library

with arguments: `y` the range of discrete values on which probabilities are calculated, `r` and `p` are the parameters of the negative binomial distribution set to default values of 2 and 0.5, respectively and finally `log.p` is a logical statement which determines whether the log of the probabilities should be returned as the output.

Lines 2-7 use if statements to determine any specified values of `r`, `p` and `y` which are outside the parameter bounds and a `stop` argument halts the procedure, printing an error message if this occurs. The input parameters of `y`, `r` and `p` can also be given in vector form and results in a matrix of probabilities as output. Line 9 determines the maximum length of these vectors, `ly`, and lines 8-11 replicates the parameter estimates to create vectors. The probability density function is calculated in line 12 and if the `log.p` statement is true the log of the *pdf* is calculated in lines 13-15. In some cases a recurrent form of the probability density function provides a more efficient method of calculating a distributions density probabilities.

```
 1  dNB <-function (y, r = 2, p = 0.5, log.p = FALSE){
 2      if (any(r <= 0))
 3          stop(paste("r␣must␣be␣>␣0)","\n",""))
 4      if (any(p < 0)|any(p > 1))
 5          stop(paste("p␣must␣be␣between␣0␣and␣1)","\n",""))
 6      if (any(y < 0))
 7          stop(paste("y␣must␣be␣>=0", "\n", ""))
 8                  ly <- max(length(y), length(r), length(p))
 9                  y <- rep(y, length = ly)
10                  r <- rep(r, length = ly)
11                  p <- rep(p, length = ly)
12      fy <- (gamma(y+r)/(gamma(r)*gamma(y+1)))*p^r*(1-p)^y
13      if (log.p==TRUE) fy <- lgamma(y+r)-lgamma(y)
14                             -lgamma(y+1)+r*log(p)
15                             +y*log(1-p)
16      fy
17  }
```

Listing 5.2: Probability density function `d`

An example of the application and output of the `dNB` function for a negative binomial distribution with parameters `r`=2 and `p`=0.6 is,

```
> dNB(0:10, r = 2, p = 0.6)
 [1] 0.36000000 0.28800000 0.17280000 0.09216000 0.04608000
 [6] 0.02211840 0.01032192 0.00471859 0.00212337 0.00094372
[11] 0.00041524
```

The negative binomial density function, `dNB` can be used to plot the density. Plot (a) of Figure 5.1 illustrates the *pdf* of negative binomial distribution with parameters of `r`=2 and `p`=0.6.

5.3.2 Cumulative density function `p`

The cumulative density function `pNB` has usage,

```
pNB(q, r = 2, p = 0.5, lower.tail = TRUE, log.p = FALSE)
```

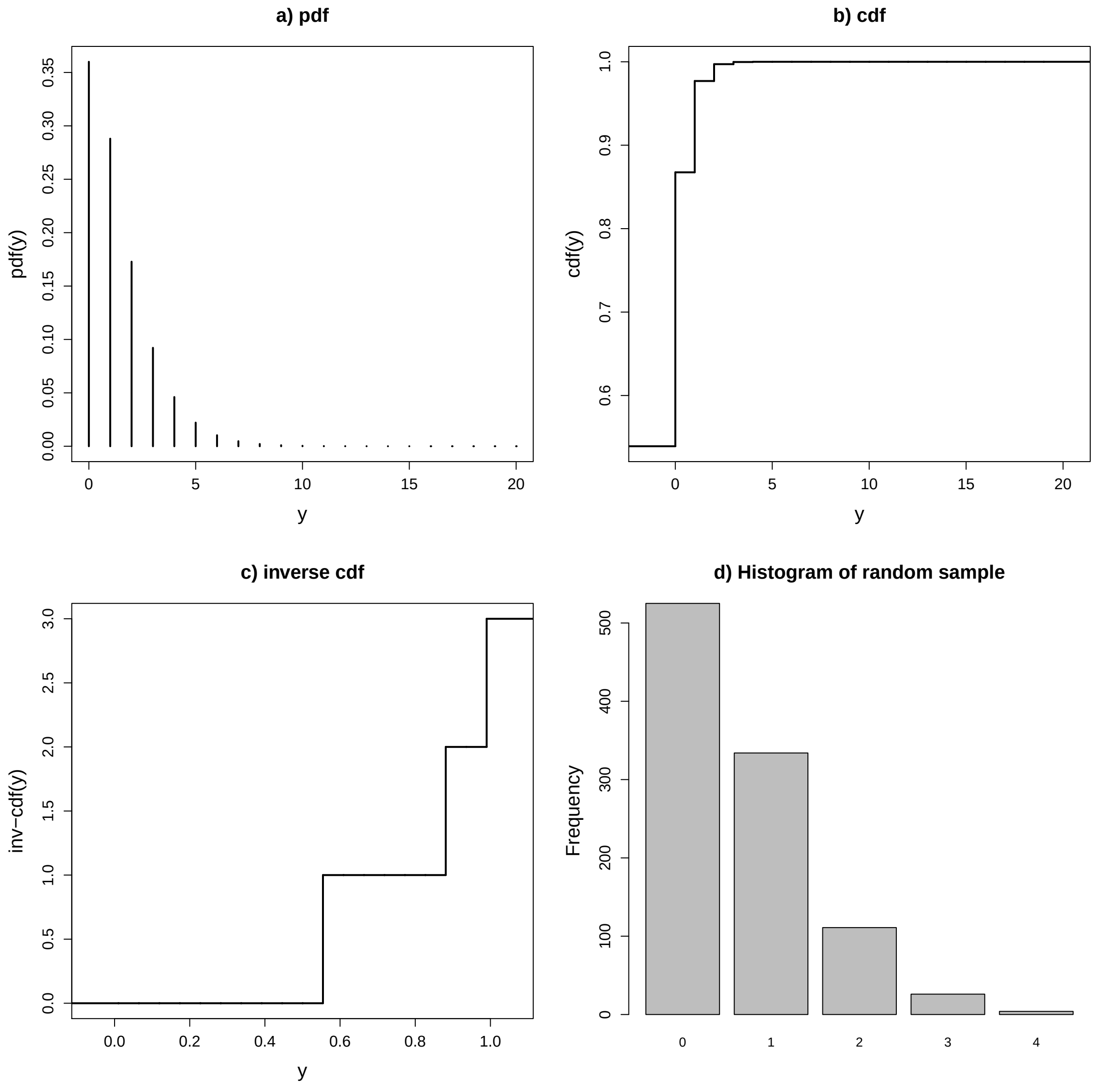


Figure 5.1: Plots of (a) the *pdf*, (b) *cdf*, (c) quantile and (d) a histogram of a random sample of the negative binomial distribution with parameters r$= 2$ and p$= 0.6$, created from the dNB, pNB, qNB and rNB functions.

where the argument q represents the vector of quantiles, whilst r and p are the parameters of the negative binomial distribution, with default values of 2 and 0.5. Two logical statements lower.tail and log.p determine whether the lower tail or log of the distribution probabilities are calculated.

```
1  pNB     <- function (q, r = 2, p = 0.5, lower.tail = TRUE,
2                  log.p = FALSE){
3      if (any(r <= 0))
4          stop(paste("r␣must␣be␣>␣0)","\n",""))
5      if (any(p < 0)|any(p > 1))
6          stop(paste("p␣must␣be␣between␣0␣and␣1)","\n",""))
7      if (any(q < 0))
8          stop(paste("q␣must␣be␣>=0", "\n", ""))
9      ly <- max(length(q), length(r), length(p))
10     q <- rep(q, length = ly)
11     r <- rep(r, length = ly)
12     p <- rep(p, length = ly)
13     cdf <- 1+(1-p)^(floor(q))*(-1+p)*p^r*comb(r+floor(q),
14            -1+r)*hypergeo_2F1(1,1+r+floor(q),2+floor(q),
15            1-p)
16     if (log.p==TRUE) cdf <- log(cdf)
17     cdf
18 }
```

Listing 5.3: Cumulative density function p

Examining the program code for pNB given in Listing 5.3, the function starts with three if statements (lines 3-8) which halt the process if any values of q, r and p supplied to the function are outside of their parameter restrictions. Once again, a vector parameterization is implemented in lines 9-12 for the two model parameters and discrete values of q. The *cdf* is calculated in lines 13-15 using supplementary functions hypergeo_2F1(a,b,c,z), the hypergeometric function ${}_2F_1(a,b;c;z)$ and comb(a,b) gives the binomial combination ${}_aC_b$. The *cdf* or log of the *cdf* is printed in lines 16-17. Alternatively, the cdf can be calculated using a cumulative sum of the pdf by utilizing the dNB function as follows:

```
    s1<- seq(0, max(q))
    cdf<- cumsum(dNB(s1, r=r, p=p))
    s2<-match(q,s1, nomatch=0)
    cdf<- cdf[s2]
```

Implementing this function for a negative binomial distribution with quantiles q between 0 and 20 and parameter values of r=2 and p=0.6 gives the following output:

```
> pNB(0:10, r = 2, p = 0.6)
 [1] 0.53920 0.86752 0.97696 0.99716 0.99973 0.99998 0.99999
 [8] 0.99999 1.00000 1.00000 1.00000
```

This function is plotted graphically in plot (b) of Figure 5.1.

5.3.3 Quantile function q

The arguments of the quantile or inverse *cdf* function for the negative binomial distribution are,

```
qNB(p, r1 = 2, p1 = 0.5, lower.tail = TRUE, log.p = FALSE,
    max.value = 10000)
```

where p is the vector of probabilities, r1 and p1 are the parameters of the negative binomial distribution and the logical statements lower.tail and log.p have the same usage as in the *cdf* function pNB in Section 5.3.2. The max.value argument is used to generate a sequence of values of q for the *cdf* function.

```
qNB <- function (p, r1= 2, p1=0.5, lower.tail = TRUE,
                 log.p = FALSE, max.value = 10000)
{
    if (any(p1 < 0) | any(p1 > 1.0001))
        stop(paste("p1␣must␣be␣in␣[0,1]","\n",""))
    if (any(r1 <= 0))
        stop(paste("r1␣must␣be␣>␣0)","\n",""))
    if (any(p < 0)|any(p > 1))
        stop(paste("p␣must␣be␣between␣0␣and␣1)","\n",""))
          if (lower.tail) p <- p
       else p <- 1 - p
    ly <- max(length(p), length(r1), length(p1))
    p <- rep(p, length = ly)
    QQQ <- rep(0, length = ly)
    r1 <- rep(r1, length = ly)
    p1 <- rep(p1, length = ly)
    for (i in seq(along = p)) {
        cumpro <- 0
        if (p[i] + 1e-09 >= 1)
            QQQ[i] <- Inf
```

```
21          else {
22              for (j in seq(from = 0, to = max.value)) {
23                  cumpro <- pNB(j, r = r1[i], p=p1[i])
24                  QQQ[i] <- j
25                  if (p[i] <= cumpro)
26                    break
27              }
28          }
29      }
30  QQQ
31  }
```

Listing 5.4: Quantile function q

The code shown in Listing 5.4 calculates the quantile distribution by summing the cumulative probabilities. The initial program framework uses stop functions in lines 4-9 to ensure that probabilities and parameters specified in the function are within the appropriate ranges, whilst a vector parameterization for parameters is once again implemented in lines 12-16. The probabilities are processed in lines 17-31 where cumulative probabilities are calculated using the *cdf* function pNB. For a negative binomial distribution with parameters r=2 and p=0.6, the quantile function is,

```
qNB(seq(0,1,length=10), r1=2, p1=0.6)
[1]   0   0   0   0   0   1   1   1   2 Inf
```

Plot (c) in Figure 5.1 gives a step plot for the inverse of the cdf using this function.

5.3.4 Random generating function r

The random generating function for the negative binomial distribution has usage,

```
rNB(n, r1 = 2, p1 = 0.5)
```

with arguments: n the number of random values to be generated from a negative binomial distribution which has parameters r1 and p1, with default values of 2 and 0.5.

```
rNB <- function(n, r1 = 2, p1 = 0.5)
{
    if (any(p1 < 0) | any(p1 > 1.0001))
        stop(paste("p1␣must␣be␣in␣[0,1]","\n",""))
    if (any(r1 <= 0))
```

```
        stop(paste("r1␣must␣be␣>␣0)","\n",""))
    if (any(n <= 0))
        stop(paste("n␣must␣be␣a␣positive␣integer","\n",""))
    n <- ceiling(n)
    p <- runif(n)
    r <- qNB(p, r1=r1, p1=p1)
    r
}
```

Listing 5.5: Random generating function `r`

This function (Listing 5.5) employs the random generating function for the uniform distribution available in `R` to randomly generate probabilities between 0 and 1. The quantile function is then applied to these probabilities to create values of the negative binomial distribution. The final plot (d) in Figure 5.1 shows a histogram of 1,000 observations generated using `rNB`.

Functions to calculate the *pgf*, moments and SI's have also included in the library for each distribution. These functions again use the shortened name for the distribution as suffixes and using the negative binomial distribution example have the following usage,

```
pgfNB(r = 2, p = 0.5, tmin = -1, tmax = 1, log.p = FALSE, n.points = 100)
momentsNB(r = 2, p = 0.5)
siNB(y, r = 2, p = 0.5, log.p = FALSE)
```

where `r` and `p` are the parameters are the negative binomial distribution. In the case of the *pgf* function `tmin` and `tmax` determine the minimum and maximum values of the range of t, whilst the argument`log.p` is a logical function which returns the log of the *pgf* if true. In the SI function `y` is a vector giving the range of discrete values and `log.p` again specifies whether the log of the SI is returned as the functions output.

5.4 Maximum likelihood estimation functions

A series of maximum likelihood estimation functions have been developed as part of the Altmann library to estimate the parameters of discrete distributions. For each discrete distribution there is a separate MLE function and estimates are produced initially by rapid estimation and then using a maximum likelihood procedure. Rapid estimates are calculated using at least one of three methods: moment estimation, method of mean and zero frequency and an EPGF method, described in Section 3.1.1. These rapid estimates then provide starting values for parameters in a maximum likelihood procedure. Maximum likelihood estimation is implemented using the mle function in the stats4 library in R.

Functions are labeled using the shortened name as a suffix in a similar style to the distribution functions in Section 5.3. An example of the programming code of the maximum likelihood function for the negative binomial distribution is given in Listing 5.6. The mle.NB function has the following usage,

```
mle.NB(ydata, method="moments", init.val=NULL, printit=T,
plot.prof=F)
```

where ydata is a vector of the observed frequencies. The method argument refers to which method of rapid estimation used, where 'moments' is the method of moments, 'zerofreq' is the method of mean and zero frequency and 'epgf' is the EPGF method. A vector of inital values for the maximum likelihood procedure can be specified in init.val, which by default is undefined. The argument printit is a logical argument with a default value of TRUE and determines whether a table of results is included in the printed output. The plot.prof argument is also logical and if TRUE profile plots of maximum likelihood estimates are produced. At the initial implementation of any one of these maximum likelihood functions the bbmle library is loaded into R if it has not already been done, seen in line 6 of the code. An example of code for the negative binomial distribution maximum likelihood function is shown in Listing 5.6 . The code for this function can be broken down into three parts: estimation of starting values, maximum likelihood estimation and goodness-of-fit statistics returned as output. The code for these three parts is explained in Sections 5.4.1-5.4.3.

```r
mle.NB <- function(ydata, method="moments", init.val=NULL,
                   printit=TRUE, plot.prof=FALSE){
  #Negative binomial distribution with parameters r and p

  #Load libraries:
  require(bbmle)

  ##Rapid estimation of r and p

if(!is.null(init.val)){
  r0 <- init.val[1]
  p0 <- init.val[2]
} else{

  #Method of moments
  if (method=="moments"){
    ybar <- mean(ydata)
    m2 <- sum((ydata-ybar)^2)/length(ydata)
    r0 <- -(ybar^2/(-m2 + ybar))
    p0 <- ybar/m2
  }

  #Method of mean-and-zero-frequency
  if (method=="zerofreq"){
    f0 <<- sum(ydata==0)/length(ydata)
    ybar<<-mean(ydata)
    zerofreq.fun<-deriv3(
    ~sqrt((ybar-((r*(1-p))/p))^2+(f0-p^r)^2),
    c("r", "p"), c("r", "p"))
    obj.fun<- function(y){
    r<- y[1]; p<- y[2]
    return(zerofreq.fun(r,p))
    }
    fun.sol<-nlminb(objective=obj.fun, start=c(f0, ybar),
                    hessian=TRUE, lower=c(0,0),
                    upper=c(Inf,1))
    r0<-fun.sol$par[1]
    p0<-fun.sol$par[2]
  }

  #EPGF method
  if (method=="epgf"){
    t1 <<- 1/2
    t2 <<- -1
    g1<<-sum(t1^ydata)/length(ydata)
    g2<<-sum(t2^ydata)/length(ydata)
    epgf.fun<-deriv3(
     ~sqrt( (g1 - p^r*(1-(1-p)*t1)^(-r))^2 +
             (g2 - p^r*(1-(1-p)*t2)^(-r))^2 ),
     c("p","r"), c("p","r"))
    obj.fun<- function(y){
    r<- y[1]; p<- y[2]
    return(epgf.fun(r,p))
    }
    fun.sol<- nlminb(objective=obj.fun, start=c(g1, g2),
                     hessian=TRUE, lower=c(0,0),
                     upper=c(Inf,1))
    r0<-fun.sol$par[1]
    p0<-fun.sol$par[2]
  }
}

  ##Maximum Likelihood Estimation
  y <- ydata
  ll.NB<-
    function(r=r, p=p) if(p>1 | p<0 | r<0) NA else
          -sum(lgamma(y+r) - lgamma(y+1) -
               lgamma(r) + y*log(1-p) + r*log(p))
  fit.dist<- try(mle2(ll.NB, start=list(r=r0, p=p0)),
                 silent = TRUE)

  #Plotting profiles
  if(plot.prof==TRUE){
    par(mfrow=c(1,2))
    plot(profile(fit.dist))
  }

  ##Estimates table

  #Parameters (names)
  pars <- c("r", "p")
  #RE Coefficients
  re.coef <- c(r0, p0)
  tab1 <- cbind(re.coef)
  dimnames(tab1) <- list(pars, "re.coef")

  if (class(fit.dist)!="try-error"){

  #MLE Coefficients #MLE coefficients S.E
  mle.coef <- c(coef(fit.dist)[[1]], coef(fit.dist)[[2]])
  mle.se <- c(sqrt(vcov(fit.dist)[1]),
              sqrt(vcov(fit.dist)[3]))
  mle.lci <- confint(profile(fit.dist))[1:2]
  mle.uci <- confint(profile(fit.dist))[3:4]
```

```
tab2<-cbind(signif(mle.coef), signif(mle.se),
            signif(mle.lci), signif(mle.uci))
dimnames(tab2) <- list(pars, c("mle.coef", "mle.se",
                               "mle.LCI", "mle.UCI"))

#Fitted Values
yi <- min(ydata):max(ydata)
observed <- c(sum(ydata==0), tabulate(ydata))
expect <- round(dNB(yi, r=mle.coef[1], p=mle.coef[2])*
                length(ydata), 2)
for (i in 1:length(expect)){
   if(expect[i]=="NaN") expect[i]<-0}
oe.tab <- rbind(observed, expect)
dimnames(oe.tab) <- list(c("obs", "exp"),
                           min(ydata):max(ydata))

#Goodness-of-fit statistics
#Chi sq
exp <- round(dNB(yi, r=mle.coef[1], p=mle.coef[2])*
             length(ydata), 2)
for (i in 1:length(exp)) if(exp[i]==0) exp[i]<-0.1
X2<- chisq.test(observed,p=exp/sum(exp))
chisq<- X2$statistic[[1]]
df <-X2$param[[1]] - length(pars)
p <- 1 - pchisq(chisq, df)
print(warning("expected␣values␣<5␣are␣pooled"))

#-Log Likelihood
logL <- logLik(fit.dist)[1]
#AIC/BIC
aic <- -2*logL+2*length(pars)
bic <- -2*logL+length(pars)*log(length(ydata))
diag.tab <- cbind(chisq, df, p, logL, aic, bic)
dimnames(diag.tab) <- list("model", c("chisq", "df",
                           "p", "logL", "AIC", "BIC"))
}

#Print Output
if (printit==TRUE){

if (class(fit.dist)=="try-error"){
  options(warn=0)
  warning("Maximum likelihood estimates cannot
           be calculated")
  cat("Rapid Estimates", "\n")
  print(tab1)
} else {
  cat("Rapid Estimates", "\n")
143     print(tab1)
144     cat("Maximum Likelihood Estimates", "\n")
145     print(tab2)
146     cat("Fitted Values", "\n")
147     print(oe.tab)
148     cat("Diagnostics", "\n")
149     print(diag.tab)
150   }
151 }
152 
153 #List of output
154  if (class(fit.dist)=="try-error"){
155     out <- list(dataname=deparse(substitute(ydata)),
156                 pars=pars, re.coef=re.coef)
157     out$family <- "NB"
158     out$yrange <- min(ydata):max(ydata)
159     out$npar <- length(pars)
160     class(out) <- "mle"
161     invisible(out)
162 } else {
163     out <- list(dataname=deparse(substitute(ydata)),
164                 pars=pars, re.coef=re.coef,
165                 mle.coef=mle.coef)
166     out$family <- "NB"
167     out$yrange <- min(ydata):max(ydata)
168     out$npar <- length(pars)
169     out$obs <- observed
170     out$exp <- expect
171     out$aic <- aic
172     out$bic <- bic
173     out$chisq <- chisq
174     out$pchi <- p
175     class(out) <- "mle"
176     invisible(out)
177 }
178 }
```

Listing 5.6: Maximum likelihood estimation function for the negative binomial distribution.

5.4.1 Estimation of starting values

Lines 10-61 of the `mle.NB` function code given in Listing 5.6 calculates the starting values of the negative binomial distribution parameters `r` and `p`, denoted by `r0` and `p0`. In lines 10-13 if starting values are specified in the `init.val` argument of the function usage, then `r0` and `p0` are set to these values. Otherwise, the function uses rapid estimation methods to generate starting values. A series of `if` statements are used to determine which method of rapid estimation has been specified in the `method` argument of the function, the default method being the method of moments. For the negative binomial distribution, these can be estimated by the three methods as follows:

Method of Moments

The first two central moments of the negative binomial distribution are given by:

$$\begin{aligned} \mu_1 &= r\left(\frac{1-p}{p}\right) \\ \mu_2 &= r\left(\frac{1-p}{p^2}\right) \end{aligned} \tag{5.1}$$

Equating these expressions to the sample moments of the data- the mean, $\bar{y}$, and the variance, s^2- these can be solved simultaneously for parameter estimates $\hat{p}$ and $\hat{r}$. The following code can be evaluated in Mathematica to calculate these estimates:

```
ln[1]:= Solve[{ybar==r((1 - p)/p), m2==r((1 - p)/p^2)}, {r, p}]
```

which gives solutions,

$$\begin{aligned} \tilde{r} &= -\frac{\bar{y}^2}{\bar{y}-s^2} \\ \tilde{p} &= \frac{\bar{y}}{s^2} \end{aligned} . \tag{5.2}$$

These estimating equations have been implemented in Lines 16-21 of Listing 5.6. In the example of the negative binomial distribution the solution of the moment estimating equations is trivial, which my not always be the case.

Method of Mean and Zero Frequency

Using the method of mean and zero frequency, the two parameters can be estimated by equating the sample mean of the data to the first central moment and the frequency of observations at zero to the probability density at zero, creating the following simultaneous equations,

$$\begin{aligned} f_0 &= p^r \\ \bar{y} &= r(\frac{1-p}{p}) \end{aligned} \tag{5.3}$$

where f_0 is the observed frequency of zeros and $\bar{y}$ is the mean of the data. In the case of the negative binomial distribution these equations cannot be solved analytically to estimate the parameters p and r since the solution involves the intractable inversion of p^r. Alternatively, rapid estimates can be calculated by minimizing a root square error function of the difference between the sample functions of the observed data and their expectations,

$$\sqrt{(f_0 - p^r)^2 + \left(\bar{y} - r\left(\frac{1-p}{p}\right)\right)^2}\,. \tag{5.4}$$

This equation can be minimized in `R` using the `nlminb` function.

This works by formulating the root square error function as a `deriv3` object and passing this as a function of the data into `nlminb`, shown on lines 27-36. In this example the `deriv3` object function is `zerofreq.fun` which comprises of the square root error in Equation 5.4. Initial starting values, alongside lower and upper bounds for the distribution parameters are also required in the minimizing `nlminb` function and the rapid estimates for `r0` and `p0` can then be extracted in lines 37 and 38.

EPGF Method

Simultaneous equations for rapid estimation can be generated by expressing the *pgf* for values of t between $-1 \leq t \leq 1$ for each parameter in the distribution. These are solved for parameter estimates by equating these expressions to the EPGF for corresponding values of t. The *pgf* of the negative binomial distribution (see Equation 2.75 in Chapter 2) is given by,

$$G(t) = p^r(1 + (p-1)\,t)^{-r} \tag{5.5}$$

For values of t of $\frac{1}{2}$ and -1 the negative binomial *pgf* is:

$$\begin{aligned} G_n(\tfrac{1}{2}) \quad &= p^r \left(1+\tfrac{1}{2}(p-1)\right)^{-r} \\ G_n(-1) \quad &= p^r(2-p)^{-r} \end{aligned} \tag{5.6}$$

Once again these simultaneous equations cannot be solved analytically, due to the requirement of the inverse of p^r and we can therefore estimate $\hat{p}$ and $\hat{r}$ by minimizing the root square error function. The root square error function for the two *pgf* equations in Equation 5.6 is,

$$\sqrt{\left(g_1 - p^r\left(1+\frac{1}{2}(-1+p)\right)^{-r}\right)^2 + \left(g_2 - p^r(2-p)^{-r}\right)^2} \tag{5.7}$$

where g_1 and g_2 are the values of the EPGF at $t = \frac{1}{2}$ and $t = -1$, respectively. Lines 43-59 of Listing 5.6 again use the `nlminb` function to perform this minimization, with the only difference being the `deriv3` objective function in lines 47-54 which is this time labeled `epgf.fun` and minimizes Equation 5.7.

5.4.2 Maximum likelihood estimation using `mle`

The maximum likelihood estimation procedure for the negative binomial distribution can be seen in lines 64-70 of Listing 5.6. Maximum likelihood estimation requires a function of the log likelihood, in this example the negative binomial log-likelihood is labeled `ll.NB`, shown in lines 65-68. Tables of the observed frequencies and a sequence of the range of y values are labeled as `tab.y` and `y.tab`, respectively. The log likelihood is then calculate using the *pdf* function `dNB` for parameters $r > 0$ and $0 < p < 1$. The `mle` function from the `stats4` library requires the log likelihood function `ll.NB` and a list of starting values which are provided by the rapid estimates `r0` and `p0`, lines 69-70. The `try()` function is used to evaluate a function and any warnings resulting from non-convergence are suppressed using the argument `silent=TRUE`. If the maximum likelihood does not converge, `fit.dist` will then have class `"try-error"`. Profile plots of the parameter estimates are plotted using lines 73-76 if the function argument `plot.prof` is true.

5.4.3 Goodness-of-fit statistics and Output

In the final part of this function a range of outputs and goodness-of-fit statistics are generated which are returned in tables as output. Firstly, in lines 80 to 85 tables of coefficients for both the rapid estimates are constructed. In line 87 an `if` statement determines whether the model in `fit.dist` has converged. If the model is not of the class `"try-error"` i.e. it has converged, then a table of maximum likelihood estimates and standard errors is constructed in lines 90- 98.

Where the `fit.dist` model converges, fitted values for the distribution are calculated using the maximum likelihood parameter estimates and a table of these values and the observed values is also included in the output, shown in lines 100-109. Goodness-of-fit statistics are calculated in lines 111-130, comprising of a Chi-squared test statistic and *p*-value, the log likelihood, the AIC and BIC values. A warning is given for the chi-squared test, as pooling is needed where expected counts are less than 5. When a large A further table is created for these values to be printed as part of the output.

Finally in lines 132-150, if the logical argument `printit` is true then the output tables are printed. For models where the maximum likelihood estimation procedure does not converge, i.e. the class of `fit.dist` is `"try-error"`, a warning is returned (lines 136-137) and only the table of rapid estimates is returned (lines 138-139). If the model does converge, i.e. the class of `fit.dist` is not equal to `"try-error"`, then tables of the rapid and maximum likelihood estimates, fitted values and model diagnostics are returned (lines 141-148). There is a list of values (lines 153-176) for each of the cases where the model does and does not converge in `mle`, which are attributed to the class `'mle'` which are not printed but are used in plotting and comparing distributions in other functions in the `Altmann` and `discrete.diag` libraries.

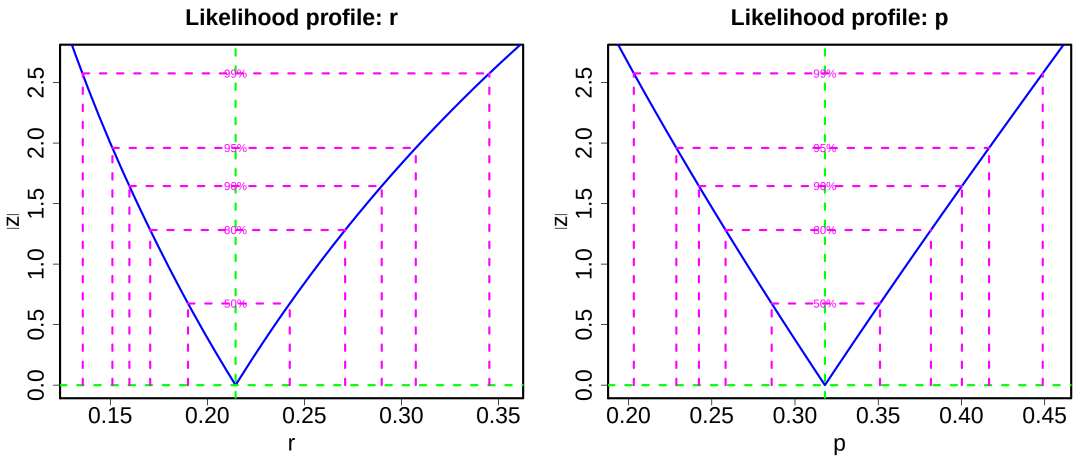


Figure 5.2: Plots likelihood profiles of parameters `r` and `p` for the number of stillbirths in litters of New Zealand white rabbits for the negative binomial distribution using the function `mle.NB`.

As an example of the usage and output of the maximum likelihood estimation functions, a negative binomial distribution can be fitted to counts of stillbirths in New Zealand white rabbits using the `mle.NB` function:

```
> library(Altmann)
> data(rabbits)
> mle.NB(rabbits, plot.prof=T)
Rapid Estimates
    re.coef
r 0.2022910
p 0.3053495
Maximum Likelihood Estimates
  mle.coef    mle.se  mle.LCI  mle.UCI
r 0.214549 0.0387163 0.151199 0.307419
p 0.317970 0.0389246 0.228720 0.416587
Fitted Values
         0  1     2    3    4    5    6    7    8    9   10   11
obs 314.00 48 20.00 7.00 5.00 2.00 2.00 1.00 2.00 0.00 0.00 1.00
exp 314.39 46 19.05 9.59 5.26 3.02 1.79 1.08 0.67 0.42 0.26 0.17
Diagnostics
         chisq df         p      logL      AIC      BIC
model 8.588087  9 0.4761319 -337.1773 678.3545 686.3474
```

Firstly, we require the `Altmann` library and rabbits dataset to be loaded into the `R` console. The first table of output from the `mle.NB` function gives the rapid estimation coefficients, followed by a table of the maximum likelihood coefficients and corresponding standard errors for the parameters `r` and `p`. Observed and fitted values and a table of goodness-of-fit statistics are also given. The default method of rapid estimation is the method of moments and the moment estimates of `r`=0.20 and `p`=0.31 shown in the first table are similar to those achieved using a maximum likelihood estimation procedure of `r`=0.21 and `p`=0.32. Figure 5.2 plots the likelihood profiles of `r` and `p` and shows the square root of the deviance difference, $|z|$, for `r` between 0.15 and 0.35 (first plot) and `p` between 0.20 and 0.45 alongside confidence intervals. The V-shape of the profile likelihoods indicate that the optimization procedure used in the maximum likelihood estimation has worked well. The fitted values provide a close fit to the observed data with the frequency of zero observations of stillbirths being predicted almost exactly. The Chi-squared test statistic given in the goodness-of-fit statistics table is $\chi^2 = 8.59$ with 9 degrees of freedom and *p*-value=0.48 and is not significant at the 5% level. Therefore we can conclude that the negative binomial distribution is a good fit to the data.

5.5 Plotting mle objects

The `S3` framework uses classes to define objects and corresponding generic functions can be created for those objects of a certain class. The `plot.mle` function uses objects of the class `"mle"` which result from the fitting of a maximum likelihood model function described in Section 5.4. The `plot.mle` function, shown in Listing 5.7, produces plots of fitted values for a distribution against the observed values. This function has usage,

```
plot(mleobject, type="bar", ylog=FALSE, xlog=FALSE)
```

where `mleobject` is an object of class `mle`. The `type` argument determines the type of graph plotted, where `bar` produces a bar plot, `l` produces a line plot or `pl` which plots both points and lines. The argument `ylog` is logical and determines whether the frequencies should be plotted on a log scale, whilst `xlog` (also logical) plots the discrete values on a log scale.

```
plot.mle <-
function(mleobject, type="bar", ylog=FALSE, xlog=FALSE){

dname<-mleobject$dataname
fname<-mleobject$family
obs.data<- mleobject$obs
exp.data <- mleobject$exp
range.y<-mleobject$yrange

###Plot different types of graphs
if (type=="l"){
  ##Line Plot
  if(ylog==TRUE){
      limits<-max(log(obs.data), log(exp.data))
      plot(range.y, log(exp.data), type="l", lwd=2,
           col="cornflowerblue", ylim=c(0,limits),
           xlab="Counts", ylab="log(Frequency)")
      title(main=paste("Plot␣of", dname, sep="␣"))
      lines(range.y, log(obs.data), type="l", lwd=2,
            col="midnightblue")
      legend("topright", lty=1, col=c("midnightblue",
             "cornflowerblue"), lwd=2,
             legend=c("Observed", fname))
  }
  if (xlog==TRUE){
      plot(range.y, exp.data, type="l", lwd=2,
           col="cornflowerblue", xlab="Counts",
           ylab="Frequency", log=’x’)
      title(main=paste("Plot␣of", dname, sep="␣"))
      lines(range.y, obs.data, type="l", lwd=2,
            col="midnightblue", log=’x’)
      legend("topright", lty=1, col=c("midnightblue",
             "cornflowerblue"), lwd=2,
             legend=c("Observed", fname))
  }
  if(ylog==FALSE && xlog==FALSE){
      limits<-max(obs.data, exp.data)
      plot(range.y, exp.data, type="l", lwd=2,
           col="cornflowerblue", ylim=c(0,limits),
           xlab="Counts", ylab="Frequency")
      title(main=paste("Plot␣of", dname, sep="␣"))
      lines(range.y, obs.data, type="l", lwd=2,
            col="midnightblue")
      legend("topright", lty=1, col=c("midnightblue",
             "cornflowerblue"), lwd=2,
             legend=c("Observed", fname))
  }
```

```
}

if (type=="pl"){
##Line Plot with points
  if(ylog==TRUE){
      limits<-max(log(obs.data), log(exp.data))
      plot(range.y, log(exp.data), type="b", lwd=2,
           col="cornflowerblue", ylim=c(0,limits),
           xlab="Counts", ylab="log(Frequency)")
      title(main=paste("Plot␣of", dname, sep="␣"))
      lines(range.y, log(obs.data), type="b", lwd=2,
            col="midnightblue")
      legend("topright", lty=1, col=c("midnightblue",
             "cornflowerblue"), lwd=2,
             legend=c("Observed", fname))
  }

  if(xlog==TRUE){
      plot(range.y, exp.data, type="b", lwd=2,
           col="cornflowerblue", xlab="Counts",
           ylab="Frequency", log=’x’)
      title(main=paste("Plot␣of", dname, sep="␣"))
      lines(range.y, obs.data, type="b", lwd=2,
            col="midnightblue", log=’x’)
      legend("topright", lty=1, col=c("midnightblue",
             "cornflowerblue"), lwd=2,
             legend=c("Observed", fname))
  }

  if(ylog==FALSE && xlog==FALSE){
      limits<-max(obs.data, exp.data)
      plot(range.y, exp.data, type="b", lwd=2,
           col="cornflowerblue", ylim=c(0,limits),
           xlab="Counts", ylab="Frequency")
      title(main=paste("Plot␣of", dname, sep="␣"))
      lines(range.y, obs.data, type="b", lwd=2,
            col="midnightblue")
      legend("topright", lty=1, col=c("midnightblue",
             "cornflowerblue"), lwd=2,
             legend=c("Observed", fname))
  }
}

if(type=="bar"){
###Bar Plot
  if(ylog==TRUE){
      limits<-max(log(obs.data), log(exp.data))
      bar.data<-cbind(log(obs.data), log(exp.data))
96        names<-0:(length(obs.data)-1)
97        barplot(t(bar.data), beside=TRUE,
98               col=rep(c("midnightblue","cornflowerblue"),
99               length(obs.data)),
100              names.arg=names, ylim=c(0,limits),
101              xlab="Counts", ylab="log(Frequency)",
102              legend=c("Observed", fname))
103       title(main=paste("Barplot␣of", dname, sep="␣"))
104   }
105   if(xlog==TRUE){
106       bar.data<-cbind(table(log(rep(range.y, obs.data))),
107                       table(log(rep(range.y, obs.data))))
108       names<-dimnames(bar.data)
109       barplot(t(bar.data), beside=TRUE,
110               col=rep(c("midnightblue","cornflowerblue"),
111               dim(bar.data)[1]),
112               names.arg=names[[1]], xlab="Counts",
113               ylab="Frequency",
114               legend=c("Observed", fname))
115       title(main=paste("Barplot␣of", dname, sep="␣"))
116   }
117  if(ylog==FALSE && xlog=FALSE){
118       limits<-max(obs.data, exp.data)
119       bar.data<-cbind(obs.data, exp.data)
120       names<-0:(length(obs.data)-1)
121       barplot(t(bar.data), beside=TRUE,
122               col=rep(c("midnightblue","cornflowerblue"),
123               length(obs.data)), names.arg=names,
124               ylim=c(0,limits), xlab="Counts",
125               ylab="Frequency",
126               legend=c("Observed", fname))
127       title(main=paste("Barplot␣of", dname, sep="␣"))
128   }
129 }
130 }
```

Listing 5.7: Plot function for class ‘mle’

The initial lines of the function (lines 4-8) extract the dataset and family name, observed and fitted values from the mle object and set the range of the observed values, y. Separate plots are specified for each type of graph and combinations of log scales on the x and y axes using if statements. The `plot` and `barplot` functions in `R` are used to create the specified graphs.

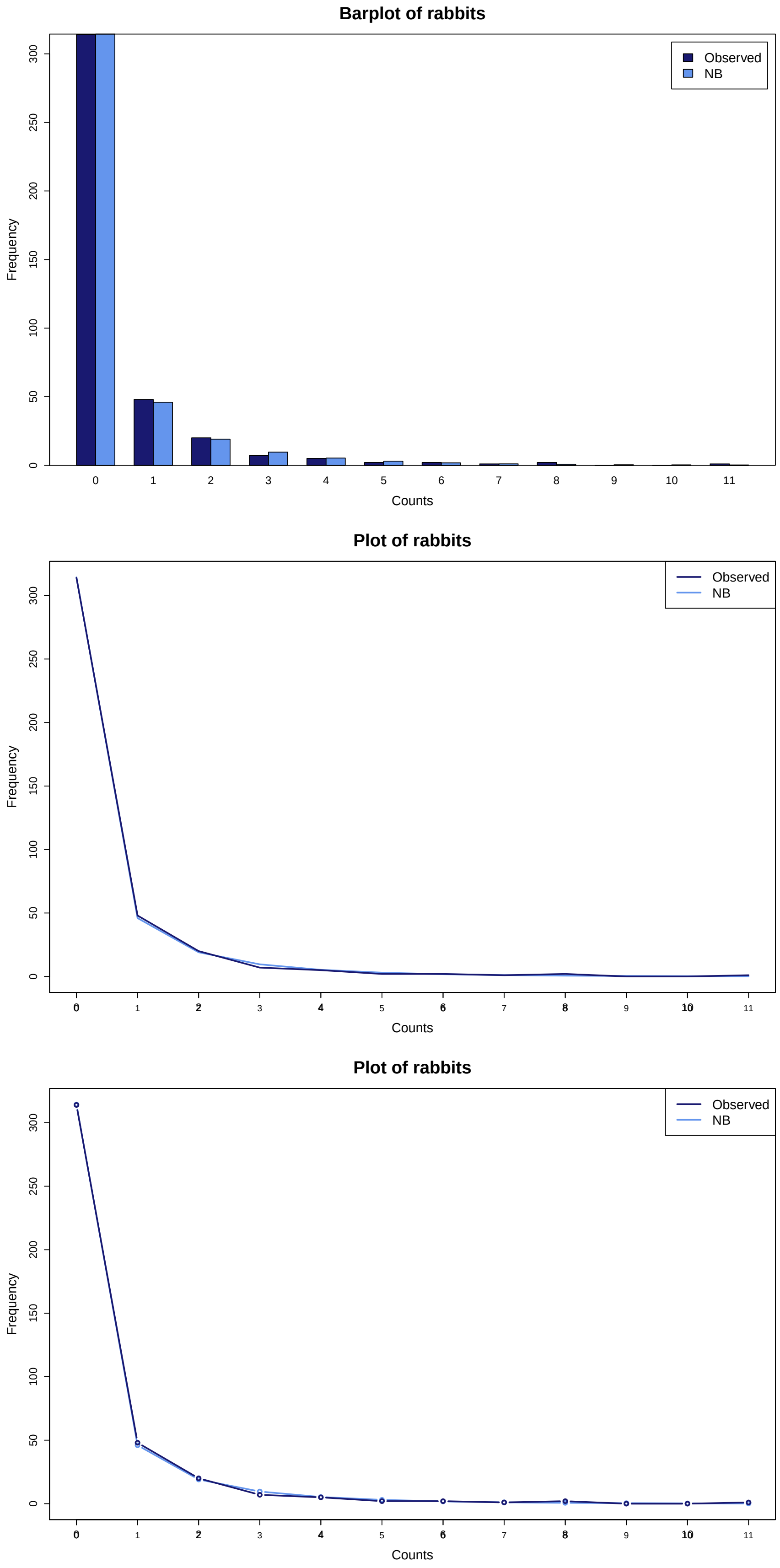


Figure 5.3: Plots of observed and expected frequencies of stillbirths for the negative binomial distribution using the function `plot.mle`.

The observed and expected frequencies for the negative binomial model fitted to frequencies of stillbirths in New Zealand white rabbits can be plotted using the `plot.mle` function, as follows:

```
NB.mod<-mle.NB(rabbits, printit=F)
plot(NB.mod)
plot(NB.mod, type="l")
plot(NB.mod, type="pl")
```

Figure 5.3 shows a barplot, line plot and line and points plot for the negative binomial model using the `plot.mle` function.

5.6 Model comparisons

The `altmann.fitter` function compares the fit of a range of discrete distributions to a dataset. It produces a table of goodness-of-fit statistics including a Chi-squared test statistic and *p*-value, the BIC and the number of parameters in the model. The table values can be ordered to determine which distribution best models the data. This function can fit various groups or families of distributions such as zero-inflated, truncated and Lerch family distributions. The `altmann.fitter` function has usage,

```
altmann.fitter(ydata, family, ord = "BIC", opt.warn = -1)
```

where `ydata` is a vector of discrete observations, the `family` argument determines which group of distributions are fitted and can either be a string of distribution names, i.e. `c("POIS", "NB", "ZIP")` or a group. One family group is `"All"` and fits all distributions allowing for zeros in the vector of observations, otherwise specifying `"Trunc"` will fit distributions for zero-truncated data. Other family groups include `"Lerch"` for the Lerch family, `"ZI"` which fits zero-inflated distributions and `"GPois"` which fits distributions from the generalized Poisson family. The argument `ord` specifies the order in which the results table is sorted, the default being `"BIC"` the Bayesian information criterion. The table may also be ordered according to `"AIC"` the Akaike Information Criterion, `"npar"` the number of parameters or `"pchi"` the Chi-squared goodness-of-fit test *p*-value. The argument `opt.warn` determines what warning messages are displayed. The default setting is negative and therefore all warnings are ignored, however if `opt.warn` is positive they will be printed.

```
altmann.fitter<- function (ydata, family, ord = "BIC",
                           opt.warn=-1){
    options(warn = opt.warn)

    if (family[1] == "All")
        family <- c("POIS", "GEO", "NB", "HY", "HO",
                    "SICH", "DE", "YU", "WA", "ZIPO",
                    "ZINB", "ZISI", "2PO", "2PNB", "NYA")
    if (family[1] == "Trunc")
        family <- c("PPO", "PGE", "PNB", "PHO", "PSI",
                    "PYU")
    if (family[1] == "Lerch")
        family <- c("LE", "ZE", "ZIPF", "GO")
    if (family[1] == "ZI")
        family <- c("ZIPO", "ZINB", "ZISI")
    if (family[1] == "GPois")
        family <- c("NYA", "HE", "GHE", "GE", "GGE")

    c.npar <- rep(NA, length = length(family))
    c.bic <- rep(NA, length = length(family))
    c.aic <- rep(NA, length = length(family))
    c.chisq <- rep(NA, length = length(family))
    c.pchi <- rep(NA, length = length(family))

    for (i in 1:length(family)) {
        if (family[i] == "BER")
            mod <- mle.BER(ydata, printit = FALSE)
        if (family[i] == "BIN")
            mod <- mle.BIN(ydata, printit = FALSE)
                     .
                     .
                     .
        if (family[i] == "GE")
            mod <- mle.GE(ydata, printit = FALSE)
        if (family[i] == "GGE")
            mod <- mle.GGE(ydata, printit = FALSE)

        if (is.null(mod$mle.coef) == FALSE) {
            c.npar[i] <- mod$npar
            c.bic[i] <- mod$bic
            c.aic[i] <- mod$aic
            c.chisq[i] <- mod$chisq
            c.pchi[i] <- round(mod$pchi, 4)
        } else {
```

```
45              c.npar[i] <- NA
46              c.bic[i] <- NA
47              c.aic[i] <- NA
48              c.chisq[i] <- NA
49              c.pchi[i] <- NA
50              cat(paste("warning:  Maximum likelihood
51                         estimates cannot be calculated
52                         for", family[i], "distribution"),
53                         "\n")
54          }
55      }
56
57      options(warn=0)
58      c.df <- length(tabulate(ydata))+any(ydata==0)-c.npar-1
59      data <- data.frame(family, c.npar, c.aic, c.bic,
60                         c.chisq, c.df, c.pchi)
61
62      if (ord == "npar") ord2 <- order(data$c.npar)
63      if (ord == 'BIC')  ord2<- order(data$c.bic)
64      if (ord == 'AIC') ord2<- order(data$c.aic)
65      if (ord == 'pchi') ord2<- order(data$c.pchi,
66                                      decreasing=TRUE)
67      result<- data[ord2, ]
68      names(result)<- c('Distribution','n.par','AIC','BIC',
69                        'chisq','df','chisq.p')
70         print(result)
71      invisible(result)
72  }
```

Listing 5.8: Altmann Fitter Model Comparison Function

The R code for the Altmann.fitter function is given in Listing 5.8. Firstly, the handling of printed warnings is determined in line 3. In lines 5-17 a series of if statements are used to create family groups of distributions. Vectors are constructed for the storage of goodness-of-fit statistics: the numbers of parameters, BIC, AIC, Chi-squared test statistics and p-values for each distribution, in lines 19-23. An iterative sequence locates each distribution specified and fits the model (lines 25-36). In lines 38-43 the goodness-of-fit statistics for each model are extracted and assigned to the storage vectors using an *if* statement to determine if maximum likelihood estimates have been produced. Otherwise, if the maximum likelihood estimation fails and only rapid estimates are returned as output, then the goodness-of-fit statistics are returned as NA in lines 44-54, with a warning that the maximum likelihood estimates cannot be calculated for that distribution. Lines 68-67 create a table of output values, which is ordered according to the criteria specified in `ord`, which is printed in line 70 and stored as an invisible table in line 71.

The fit of several distributions to the number of stillbirths in litters of New Zealand white rabbits can be compared using the `altmann.fitter` function. The following table compares the number of parameters, AIC and BIC's, Chi-squared test statistic and *p*-values for a range of models:

```
> altmann.fitter(rabbits, family="All")
   Distribution n.par      AIC      BIC      chisq df chisq.p
8            YU     1 679.5224 683.5188  11.927863 10  0.2899
3            NB     2 678.3545 686.3474   8.588087  9  0.4761
5            HO     2 678.4542 686.4471   7.333848  9  0.6024
9            WA     2 678.7231 686.7160   7.206004  9  0.6157
6          SICH     3 679.9299 691.9192   6.938156  8  0.5433
7            DE     3 680.0329 692.0223   7.302000  8  0.5044
11         ZINB     3 680.6871 692.6765   9.717613  8  0.2854
14         2PNB     4 681.6373 697.6231   6.163812  7  0.5208
12         ZISI     4 681.8943 697.8801   6.690988  7  0.4617
13          2PO     3 691.2427 703.2320  62.056267  8  0.0000
15          NYA     2 697.0951 705.0880 131.327759  9  0.0000
10         ZIPO     2 718.3784 726.3713 126.254564  9  0.0000
2           GEO     1 733.6000 737.5965 186.426297 10  0.0000
1          POIS     1 883.6870 887.6834 287.300439 10  0.0000
4            HY     3 963.5239 975.5132 434.068701  8  0.0000
```

This table is ordered according to the BIC and determines the Yule distribution with a BIC value of 683.52 to provide the best fit to the data of the models fitted. The model with the second lowest BIC is the negative binomial distribution, which has a BIC of 686.35 followed by the Holla distribution of 686.45. The BIC is more conservative against overfitting than the BIC- if we compare the AIC values for the two models the negative binomial distribution is lower at 678.35 compared to the Yule distribution's AIC value of 679.52. Alternatively, we can compare the fit of the distributions using the Chi-squared goodness-of-fit test statistic as a comparison criteria:

```
> altmann.fitter(rabbits, family="All", ord="pchi")
   Distribution n.par      AIC      BIC      chisq df chisq.p
9            WA     2 678.7231 686.7160   7.206004  9  0.6157
5            HO     2 678.4542 686.4471   7.333848  9  0.6024
```

```
6          SICH     3 679.9299 691.9192   6.938156  8  0.5433
14         2PNB     4 681.6373 697.6231   6.163812  7  0.5208
7            DE     3 680.0329 692.0223   7.302000  8  0.5044
3            NB     2 678.3545 686.3474   8.588087  9  0.4761
12         ZISI     4 681.8943 697.8801   6.690988  7  0.4617
8            YU     1 679.5224 683.5188  11.927863 10  0.2899
11         ZINB     3 680.6871 692.6765   9.717613  8  0.2854
1          POIS     1 883.6870 887.6834 287.300439 10  0.0000
2           GEO     1 733.6000 737.5965 186.426297 10  0.0000
4            HY     3 963.5239 975.5132 434.068701  8  0.0000
10         ZIPO     2 718.3784 726.3713 126.254564  9  0.0000
13          2PO     3 691.2427 703.2320  62.056267  8  0.0000
15          NYA     2 697.0951 705.0880 131.327759  9  0.0000
```

The Waring distribution has the largest Chi-squared test *p*-value for the models fitted of 0.616, followed by a Holla distribution with *p*-value 0.602 and a Sichel distribution with *p-value* 0.543. The Chi-squared goodness-of-fit test statistic *p*-values for the negative binomial distribution is much larger than the Yule distribution of 0.290 We can compare the fit of the Yule and Waring distributions using the maximum likelihood estimation functions:

```
> mle.YU(rabbits)
Rapid Estimates
        re.coef
lambda 3.172973
Maximum Likelihood Estimates
       mle.coef  mle.se mle.LCI mle.UCI
lambda  3.19126 0.29514 2.66699 3.83313
Fitted Values
         0     1     2    3    4    5    6    7    8    9   10
obs 314.00 48.00 20.00 7.00 5.00 2.00 2.00 1.00 2.00 0.00 0.00
exp 306.09 58.96 19.05 7.95 3.88 2.11 1.24 0.78 0.51 0.35 0.25
      11
obs 1.00
exp 0.18
Diagnostics
         chisq df         p      logL      AIC      BIC
model 11.92786 10 0.2899134 -338.7612 679.5224 683.5188

> mle.WA(rabbits)
Rapid Estimates
   re.coef
b 3.609268
n 1.200782
Maximum Likelihood Estimates
```

```
  mle.coef   mle.se  mle.LCI mle.UCI
b  2.46029 0.556232 1.619770 3.98402
n  0.70331 0.197151 0.417895 1.27179
Fitted Values
         0     1     2    3    4    5    6    7    8    9   10
obs 314.00 48.00 20.00 7.00 5.00 2.00 2.00 1.00 2.00 0.00 0.00
exp 312.63 52.81 17.42 7.64 3.95 2.28 1.42 0.93 0.64 0.46 0.34
      11
obs 1.00
exp 0.26
Diagnostics
         chisq df        p      logL      AIC     BIC
model 7.206004  9 0.615681 -337.3616 678.7231 686.716
```

Although the distribution has the lowest of the BIC values in the first model comparison table this is due to its only having one parameter and the predicted frequencies of stillbirths are not as good as those from the Waring or negative binomial models, which is reflected in the lower Chi-squared test *p-value*. Fitted values for the number of stillbirths can be compared for the negative binomial and Waring distributions. The expected frequency of 0 and 1 stillbirths are 314.39 and 46.00 in the negative binomial model (observed frequency is 48) compared to 312.63 and 52.81 in the Waring model. Although the Waring model does not fit the distribution as well as the negative binomial for low numbers of stillbirths, the tail of the distribution is a better fit. We can conclude, therefore that the Waring distribution provides the best fit to the numbers of stillbirth in litters of New Zealand white rabbits out of the range of models fitted. This model has an interesting interpretation as a Geometric-Beta parameter mixture where the number of stillbirth offspring in every litter can be thought of as a Geometric-distributed random variable with the probability of a stillbirth varying according to a Beta distribution perhaps reflecting natural variation in the maternal susceptibility to produce stillbirth rabbits.

5.7 Validation of the functions

There are five functions for each distribution in the `Altmann` library: the *pdf*, `d`, the *cdf*, `p`, the inverse *cdf* function `q`, a random generating function `r` and a maximum likelihood estimation function which begins with the prefix `mle`. A template for each of these types of functions was created, which was then replicated and altered for the correct formulae and parameters for each distribution. This method of creating the functions ensured the consistency of programs in the library. The `d`, `p`, `q` and `r` functions follow the standard format for distribution functions in `R`. The remaining functions in the `Altmann` library (mle fitting, altmann comparison and plot functions) were built using a trial and error process, with the basic functions expanded to achieve the desired function output. Final versions of the code were checked using `Rs check` function when the libraries were created. This runs a series of checks to test if the functions in a library work correctly, including testing the functions and helpfiles for syntax errors and testing the examples in the helpfiles.

The functions in the Altmann library were tested in a variety of ways, using data randomly generated from discrete distributions and also using real datasets, with existing parameter estimates and fitted values found in papers and books to use as comparisons. For each distribution, the `d`, `p`, `q` and `r` functions were tested for a range of parameter values. The `d` functions were tested on a range of y values from 0 to a large number (say around 1,000) to confirm that the *pdf* values sum to one. The `p` and `q` functions were also tested to check that for a set of randomly generated values from the selected distribution, the inverse *cdf* and *cdf* functions also sum to one. The tails of the `p` functions were also tested to ensure that the cumulative probabilities for the lower and upper tails added together summed to one. Finally, to test the *pdf* and *cdf* functions match for a series of randomly generated values from the selected distribution, the sum of probability densities calculated from 0 to the generated value was compared to the *cdf* of the value. Plotting the output of the `d`, `p`, `q` and `r` functions allows a visual inspection of the values generated, to confirm that the restrictions upon the functions hold, i.e. to test the *pdf* sums to one. Any violations of the restrictions placed on the *pdf*, `cdf` and inverse *cdf* found during this testing process prompted an investigation for errors in the function code. This process was continued until the *pdf*, `cdf` and inverse *cdf* functions met the restrictions on the functions.

Maximum likelihood estimation functions were validated in a number of ways. One method of verifying the results from the maximum likelihood estimating functions that was used is to generate a random sample of values from the distribution, with known parameter estimates. This data can then be used to determine that the function can correctly estimate the parameters of the distribution.

An alternative method used to test the maximum likelihood estimation functions was to use published data as a comparison. For a discrete dataset, estimates of the parameters and fitted values of a distribution were compared to the results for the same distribution using maximum likelihood estimation in the `Altmann` library. For example, (Plunkett and Jain, 1975) present parameter estimates and fitted values for a Gegenbauer distribution fitted to 400 haemocytometer counts of yeast cells. Table 5.7 gives the fitted values of the Gegenbauer distribution, with parameter estimates $\hat{a} = 0.198$, $\hat{b} = 0.004$ and $\hat{k} = 2.898$ given by (Plunkett and Jain, 1975). Similar fitted values for the Gegenbauer distribution are obtained from the maximum likelihood estimation in the `Altmann` library to those given by Plunkett and Jain (1975).

No. of yeast cells	0	1	2	3	4	5	6
Observed Freq	213	128	37	18	3	1	0
Gegenbauer (Plunkett and Jain, 1975)	214.15	123.00	44.88	13.36	3.55	0.86	0.20
Altmann library MLE	214.84	121.3	45.72	13.65	3.49	0.79	0.17
Altmann fitter (Altmann, 1997)	213.10	122.68	45.54	13.79	3.70	1.19	-

Table 5.7: Fitted values for a Gegenbauer distribution fitted to 400 haemocytometer counts of yeast cells.

The Altmann fitter program (Altmann, 1997) can also be used to compare the fit of the Gegenbauer distribution to counts of yeast cells, also shown in Table 5.7. This model has parameter estimates $\hat{a} = 0.164$, $\hat{b} = 0.0002$ and $\hat{k} = 3.516$. The rapid estimates calculated by this program do not provide as close a fit to the data as those from the `Altmann` library. A benefit of comparing the results from the Altmann fitter program to those in the `Altmann R` library, is that goodness-of-fit statistics, such as the Chi-square test statistic and *p*-value can also be compared. Counts of yeast cells provide one illustration of the methods for which the maximum likelihood estimation functions have been tested for validity.

These methods of testing the maximum likelihood functions were employed for each distribution using a variety of datasets from published sources. The Poisson, negative binomial, hypergeometric, geometric, parameter-mix, zero-inflated and component mix distributions were tested using counts of stillbirths in New Zealand white rabbits (Morgan et al., 2007), counts of yearly deaths by horse kicks in the Prussian army between 1875-1894 and counts of earthquakes on the coast of Mexico (Nakamura and Pérez-Abreu, 1993a), amongst others. Truncated distributions were compared to results from models fitted to household size data (Nakamura and Pérez-Abreu, 1993a) and numbers of births occurring to HIV-infected women, presented in Section 5.8.2. Lerch family distributions were testing using the frequency of surnames from eight districts analyzed by Zörnig and Altmann (1995) and Panaretos (1989). Distributions from the Generalized Poission family including were checked using the lakota dataset of the number of phonemes of words (Pustet and Altmann, 2005), haemocytometer counts of yeast cells (Plunkett and Jain, 1975) and counts of the number of European red mites on apple leaves (Medhi and Borah, 1984). Many of these datasets were also included in the `Altmann` library.

5.8 Further Examples

This section presents two further examples of the application of functions available in the Altmann library: the number of automobile accidents claims for drivers in Belgium, 1978 and numbers of births occurring in the UK and Ireland to HIV-infected women.

5.8.1 Automobile accidents claims for drivers in Belgium, 1978

Table 5.8 presents the number of automobile accidents claims for drivers in Belgium, 1978 ($n = 9,461$) by Denuit (1997). This dataset was analysed using generalized Poisson, negative binomial and Holla models by Nikoloulopoulos and Karlis (2008a) who concluded that only the Holla model provides an acceptable fit to the number of accident claims. Summary statistics for the number of accident claims can be found using the `summary.disc` function, shown below.

```
> summary.disc(belgiandrivers)
$desc
   Min   1st Q Median   3rd Q    Max      n
     0       0      0       0      7   9461

$moms
   mean      var   stddev      m3      m4      sk      ku
 0.2144   0.2889   0.5375  0.5407  1.8020  3.4810 21.5900

$extras
    OD  kappa3     ZI    Gini
 1.348   1.522 42.830   0.858

$tab
             0         1         2          3        4         5
freq 7840.0000 1317.0000 239.00000 42.000000 14.00000 4.0000000
prob    0.8287    0.1392   0.02526  0.004439  0.00148 0.0004228
             6         7
freq 4.0000000 1.0000000
prob 0.0004228 0.0001057
```

Number of claims	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frequency	7840	1317	239	42	14	4	4	1

Table 5.8: Number of automobile accidents claims for drivers in Belgium, 1978

The mean number of claims is 0.21, with variance 0.29 and the overdispersion index $OD = 1.348$ indicates that some overdispersion is present in the data. The ZI index is very large with a value of 42.83 and reflects the high proportion of zeros present in this dataset. This is also reflected in the values of the median, 25% lower and 75% upper quartiles which are all skewed with values 0. The skewness and kurtosis coefficients are also very large indicating the number of claims has a highly positivly skewed distribution with a peak near the mean and heavy tails.

The fit of a range of distributions to the frequency of accident claims can be compared using the `altmann.fitter`:

```
> altmann.fitter(belgianDrivers1978, family='All')
warning: Maximum likelihood estimates cannot be calculated for
```

```
ZINB distribution
warning: Maximum likelihood estimates cannot be calculated for
ZISI distribution
   Distribution n.par      AIC      BIC       chisq df chisq.p
5            HO     2 10691.02 10705.33   10.404852  5  0.0645
9            WA     2 10693.40 10707.71   14.002380  5  0.0156
6          SICH     3 10690.71 10712.17    5.163696  4  0.2709
7            DE     3 10692.55 10714.02    8.249035  4  0.0829
3            NB     2 10700.08 10714.39   32.236500  5  0.0000
2           GEO     1 10711.36 10718.52   95.562515  6  0.0000
14         2PNB     4 10690.49 10719.11    2.242583  3  0.5236
13          2PO     3 10701.92 10723.39   24.998198  4  0.0001
15          NYA     2 10721.21 10735.52  152.465797  5  0.0000
10         ZIPO     2 10755.23 10769.54  451.595954  5  0.0000
8            YU     1 10792.82 10799.97   98.526822  6  0.0000
1          POIS     1 10983.56 10990.72 1111.074577  6  0.0000
4            HY     3 11187.79 11209.26 1840.257776  4  0.0000
11         ZINB    NA       NA       NA          NA NA      NA
12         ZISI    NA       NA       NA          NA NA      NA
```

This table shows a Holla distribution provides the best fit to the data if we compare the BIC values, as found by Nikoloulopoulos and Karlis (2008a). If we instead compare the distributions according to the Chi-squared goodness-of-fit p-values there are several models which perform better than the Holla, which has $p > 0.05$. The Delaporte ($p = 0.645$), Sichel ($p = 0.271$) and Poisson-negative binomial mixture ($p = 0.524$) models all have p-values suggesting they provide an adequate fit to the data. One parameter models such as the Poisson, Geometric and Yule distributions are unsurprising not a good fit to this dataset and it is interesting to note that two of the three zero-inflated models failed to converge whilst the zero-inflated Poisson model has been placed near the center of the comparison table.

We can compare the fit of the Holla, Sichel and Poisson-negative binomial distributions,

```
> mle.HO(belgiandrivers, plot.prof=T)
Rapid Estimates
        re.coef
alpha 0.8025763
theta 0.4102219
Maximum Likelihood Estimates
      mle.coef    mle.se  mle.LCI  mle.UCI
alpha 0.839766 0.0676676                          NA      0.990529
theta 0.396567                   NaN      NaN 0.446073
```

```
Fitted Values
          0       1      2     3     4    5    6    7
obs 7840.00 1317.00 239.00 42.00 14.00 4.00 4.00 1.00
exp 7844.01 1306.12 238.23 53.27 13.75 3.89 1.17 0.37
Diagnostics
         chisq df          p      logL      AIC      BIC
model 10.40485  5 0.06454375 -5343.511 10691.02 10705.33
```

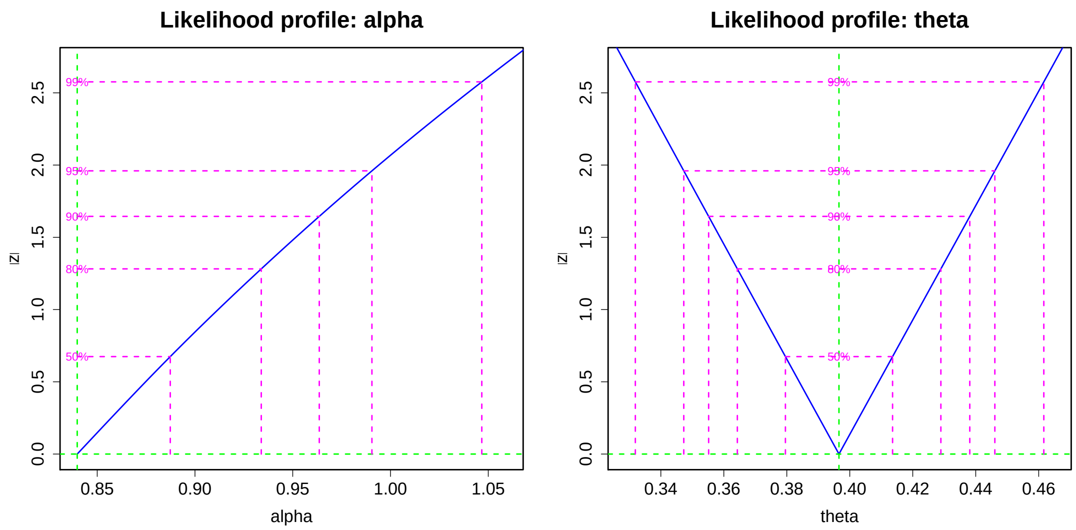


Figure 5.4: Profile likelihood plots for Holla model for number of automobile accidents claims for drivers in Belgium, 1978

For the number of automobile accidents claims for drivers in Belgium, 1978 the Holla model has parameter estimates of $\hat{\alpha} = 0.84$ and $\hat{\theta} = 0.40$. Figure 5.4 plots the profiles for the Holla model. Whilst the profile for theta shows a 'V'-shape indicating the estimate has converged, this is not true for alpha and standard errors for $\hat{\theta}$ have not been unable to be calculated. The fitted values for this model indicate a reasonable fit to the data.

```
> mle.SICH(belgiandrivers, plot.prof=T)
Rapid Estimates
      re.coef
alpha      NA
theta      NA
gama       NA
Maximum Likelihood Estimates
       mle.coef    mle.se   mle.LCI   mle.UCI
alpha  1.011630 0.0863524  0.842381  1.180880
theta  0.597967 0.1534120  0.297280  0.898654
```

```
gama  -1.335610 0.4832330 -2.282750 -0.388476
Fitted Values
              0          1       2       3       4     5     6     7
obs 7840.00 1317.00 239.0 42.00 14.00 4.00 4.00 1.00
exp 7837.82 1325.63 225.5 50.07 14.18 4.75 1.78 0.72
Diagnostics
            chisq df            p        logL       AIC       BIC
model 5.163696  4 0.2709098 -5342.353 10690.71 10712.17
```

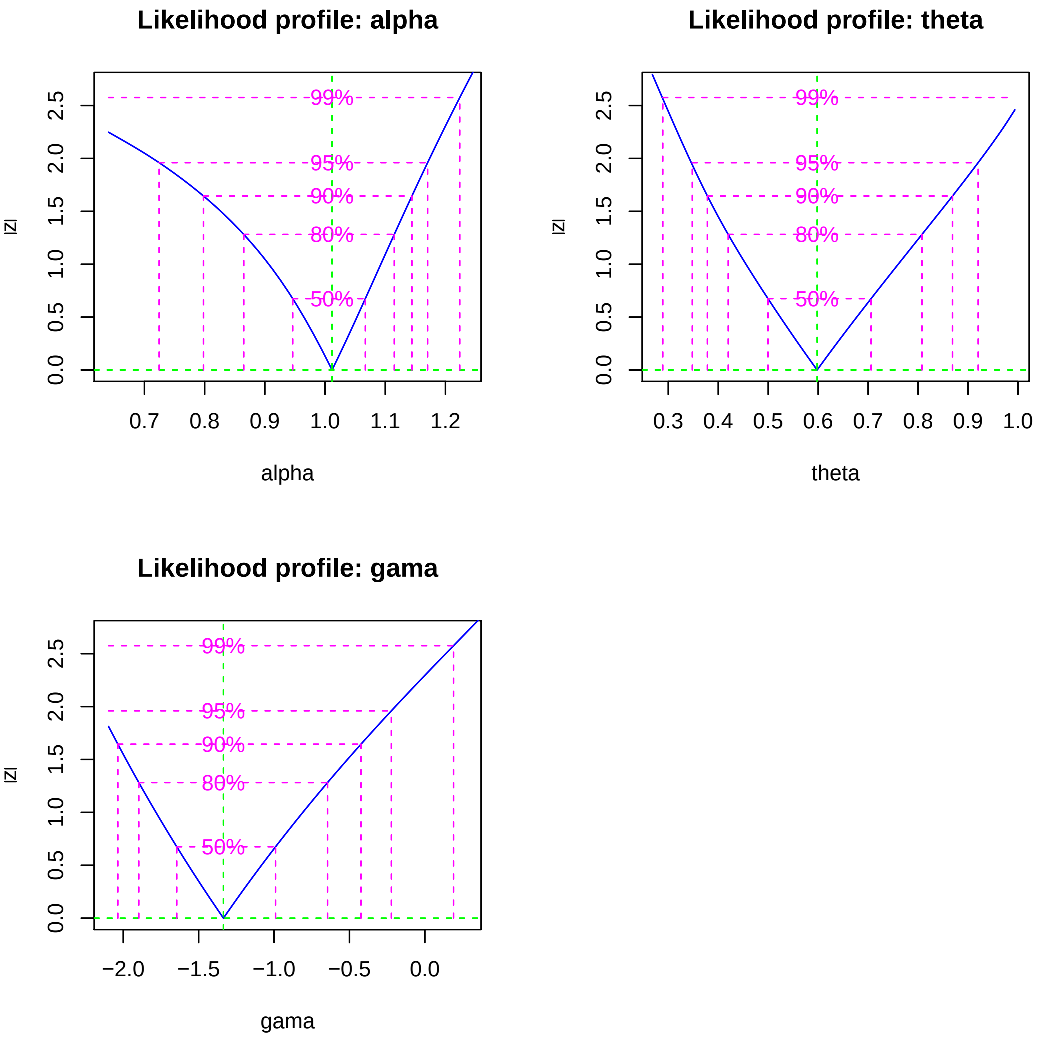


Figure 5.5: Profile likelihood plots for Sichel model for number of automobile accidents claims for drivers in Belgium, 1978

The Sichel model has parameter estimates of $\hat{\alpha} = 1.01$, $\hat{\theta} = 0.60$ and $\hat{\gamma} = -1.34$. Figure 5.5 plots the profiles for the Sichel model, showing 'V' shapes for each parameter indicating that the model has correctly convereged. Fitted values for the Sichel model do not fit the observed counts of automobile accidents as closely as the Holla model, although this model has a large Chi-squared *p*-value indicating it provides a good fit to the data. We can alternatively fit a Poisson-negative binomial mixture to this dataset,

```
> mle.2PNB(belgiandrivers, plot.prof=T)
Rapid Estimates
        re.coef
omega 0.1931172
mu    0.2841505
r     0.5292820
p     0.7305169
Maximum Likelihood Estimates
      mle.coef    mle.se  mle.LCI  mle.UCI
omega 0.489536 0.0590904 0.316242 0.605718
mu    0.297771 0.0467485 0.155788 0.375493
r     0.133017 0.0918020 0.031953 0.543894
p     0.497613 0.0930044 0.313933 0.674960
Fitted Values
          0       1      2     3     4    5   6    7
obs 7840.00 1317.00 239.00 42.00 14.00 4.00 4.0 1.00
exp 7840.08 1318.08 236.16 45.03 12.89 4.95 2.1 0.92
Diagnostics
         chisq df         p      logL      AIC      BIC
model 2.242583  3 0.5236097 -5341.246 10690.49 10719.11
```

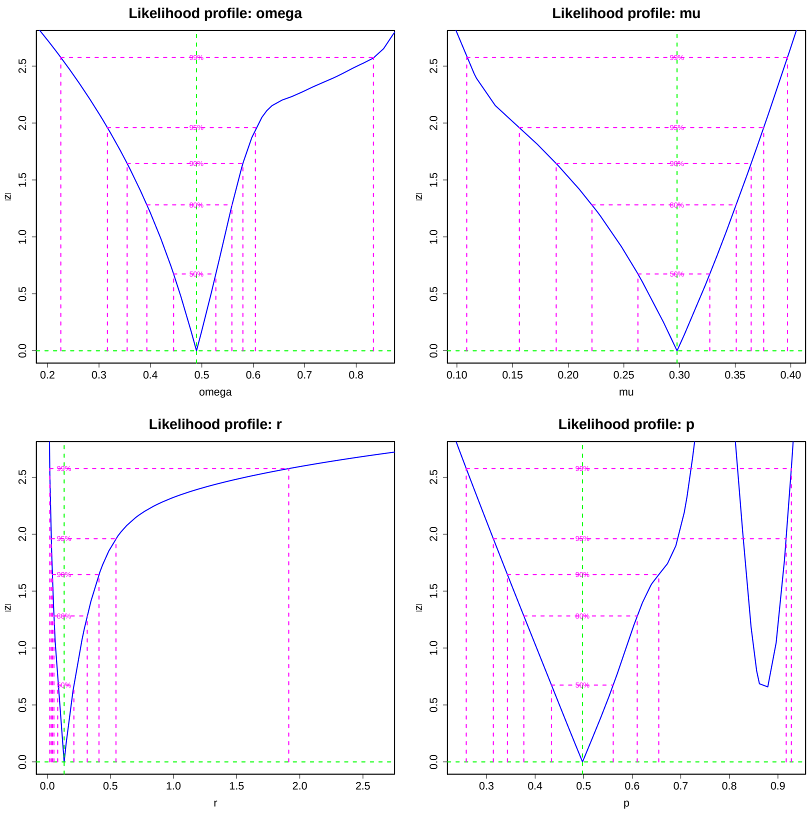


Figure 5.6: Profile likelihood plots for Poisson-negative binomial mixture model for number of automobile accidents claims for drivers in Belgium, 1978

Profile plots shown in Figure 5.6 indicate that all the parameters have converged and we can note that p looks to both a local and global minimum. Although complex (this model has four parameters) this model provides an excellent fit indicated by the expected values. We can conclude that the Poisson-negative binomial mixture provides a good fit to the data, due to the increased complexity in this model. This example illustrates the trade-off between a a model with a higher goodness-of-fit and a large number of parameters (Poisson-negative binomial mixture) and a simpler model with fewer parameters to interpret but a lower goodness-of-fit in comparison.

5.8.2 Numbers of births occurring to HIV-infected women

An example of a truncated dataset is the numbers of births occurring in the UK and Ireland to HIV-infected women reported to the National Study of HIV in Pregnancy and Childhood, between 2000 an 2010 (French, 2011). This dataset is truncated as mothers could have had more than one birth outside of UK which we do not have information on.

Number of births	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Frequency	7655	1895	430	96	19	3	2

A range of positive (also known as zero-truncated) distributions can be fitted to the numbers of births occurring to HIV-infected women by selecting the `"Trunc"` family in the `altmann.fitter` function:

```
> altmann.fitter(HIV.births, family="Trunc")
  Distribution n.par      AIC      BIC      chisq df chisq.p
2          PGE     1 14524.59 14531.81   6.076759  5  0.2988
3          PNB     2 14521.29 14535.73   1.959864  4  0.7431
4          PHO     2 14521.53 14535.97   1.717276  4  0.7876
5          PSI     3 14523.39 14545.05   1.738426  3  0.6284
1          PPO     1 14602.20 14609.42 165.955166  5  0.0000
6          PYU     1 14932.92 14940.14 338.326865  5  0.0000
```

The above table of output shows the positive Geometric distribution has the lowest BIC but the positive negative binomial distribution has lowest AIC. We can again fit both of these models to compare their fit:

```
> mle.PGE(HIV.births)
Rapid Estimates
       re.coef
prob 0.7624943
Maximum Likelihood Estimates
     mle.coef     mle.se  mle.LCI mle.UCI
prob 0.762494 0.00369754 0.755197 0.76969
Fitted Values
          1       2      3      4    5    6    7
obs 7655.00 1895.00 430.00  96.00 19.0 3.00 2.00
exp 7701.19 1829.08 434.42 103.18 24.5 5.82 1.38
Diagnostics
         chisq df         p      logL      AIC      BIC
model 6.076759  5 0.2988215 -7261.294 14524.59 14531.81
```

This model shows a good fit with the parameter estimated at $\hat{p} = 0.76$, which can be interpreted as each woman has a probability of 76% of continuing to have children until she has the number of child she wants. The positive negative binomial model can be fitted using the `mle.PNB` function and gives the following output:

```
> mle.PNB(HIV.births)
Rapid Estimates
    re.coef
p 0.4318472
r 0.6954731
Maximum Likelihood Estimates
  mle.coef    mle.se  mle.LCI  mle.UCI
p  0.81179 0.0216737 0.769235 0.853944
r  1.62971 0.3457400 1.076690 2.497320
Fitted Values
          1      2      3     4     5    6    7
obs 7655.00 1895.0 430.00 96.00 19.00 3.00 2.00
exp 7655.12 1894.4 431.39 93.97 19.91 4.14 0.85
Diagnostics
         chisq df        p      logL      AIC      BIC
model 1.959864  4 0.743141 -7258.646 14521.29 14535.73
```

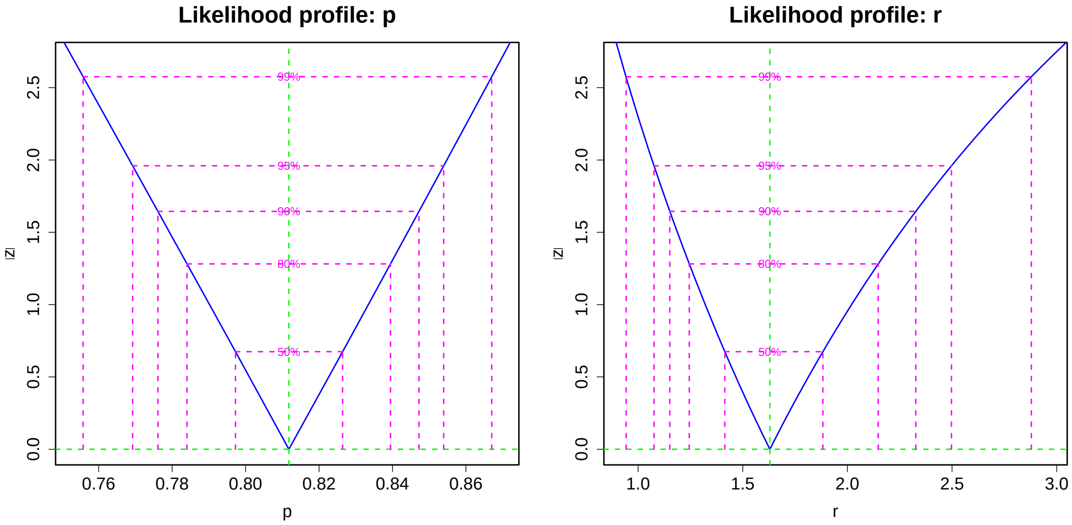


Figure 5.7: Profile likelihood plots for a positive negative binomial model fitted to numbers of births occurring in the UK and Ireland to HIV-infected women

This model is a very good fit, shown by the expected frequencies which closely model frequencies of numbers of births in HIV-infected women. Figure 5.7 gives the profile likelihood plots for $\hat{r}$ and $\hat{p}$, which display V-shaped curves and indicates the maximum likelihood has converged correctly. This example illustrates that the BIC can sometimes over adjust for the number of parameters in a model.

5.9 Application to UK surnames distribution

The Zipf distribution is applied to data that is ranked by size; for example, occurrences of words in a sample of speech or writing, or numbers of species per genus in ecology (Mandlebrot, 1959; Zipf, 1949). Fox and Lasker (1983) applied the Zeta distribution (a variation of the Zipf distribution where $a = 0$) to the frequency distributions of surnames of 4,794 people married within nine districts in England in a 12 month period in 1972/3. For each district, the frequency distribution is defined to be the number of people in that district with a given surname.

In this section, we fit the Zipf distribution to frequencies of surnames in the UK, introduced in Section 1.2.1 of Chapter One. This dataset gives the distribution of surnames across 436 districts (12 regions) of the UK. We are interested in comparing the fit of the Zipf distribution to surname frequencies for each district. Estimates of c and a the parameters of the Zipf distribution are obtained via maximum likelihood estimation for the surname frequencies at each district using the maximum likelihood estimation function, `mle.ZIPF`, in the `Altmann` library.

Figure 5.8 plots the logarithm of the parameters c (x axis) and a (y axis) of the Zipf distribution fitted to surname frequencies for each district. A colour key is used to denote the regions to which each district belongs. The average values of the parameters is $c = 2.24$ and $a = 1.94$, indicated in Figure 5.8 as a white 'x'. There is a clear relationship between the parameters c and a. Districts that have larger values of $log(c)$, have larger values of $log(a)$, whilst for lower values of $log(c)$ there is more variation in the range of $log(a)$.

The plot shows that the majority of districts have values of $log(c)$ and $log(a)$ which are centered around the mean, i.e. $log(a)$ between 0 and 1 and $log(c)$ between 0.6 and 1. There are some cases where districts are clustered by region. For example, London districts (black) have higher values of $log(c)$ than the average district and districts in Northern Ireland (orange) have lower values of $log(a)$ than the average district. Many cities can be seen as outliers on the plot e.g. Edinburgh, Birmingham, Bristol, Liverpool, Leeds - all of which have lower values of $log(a)$ and $log(c)$.

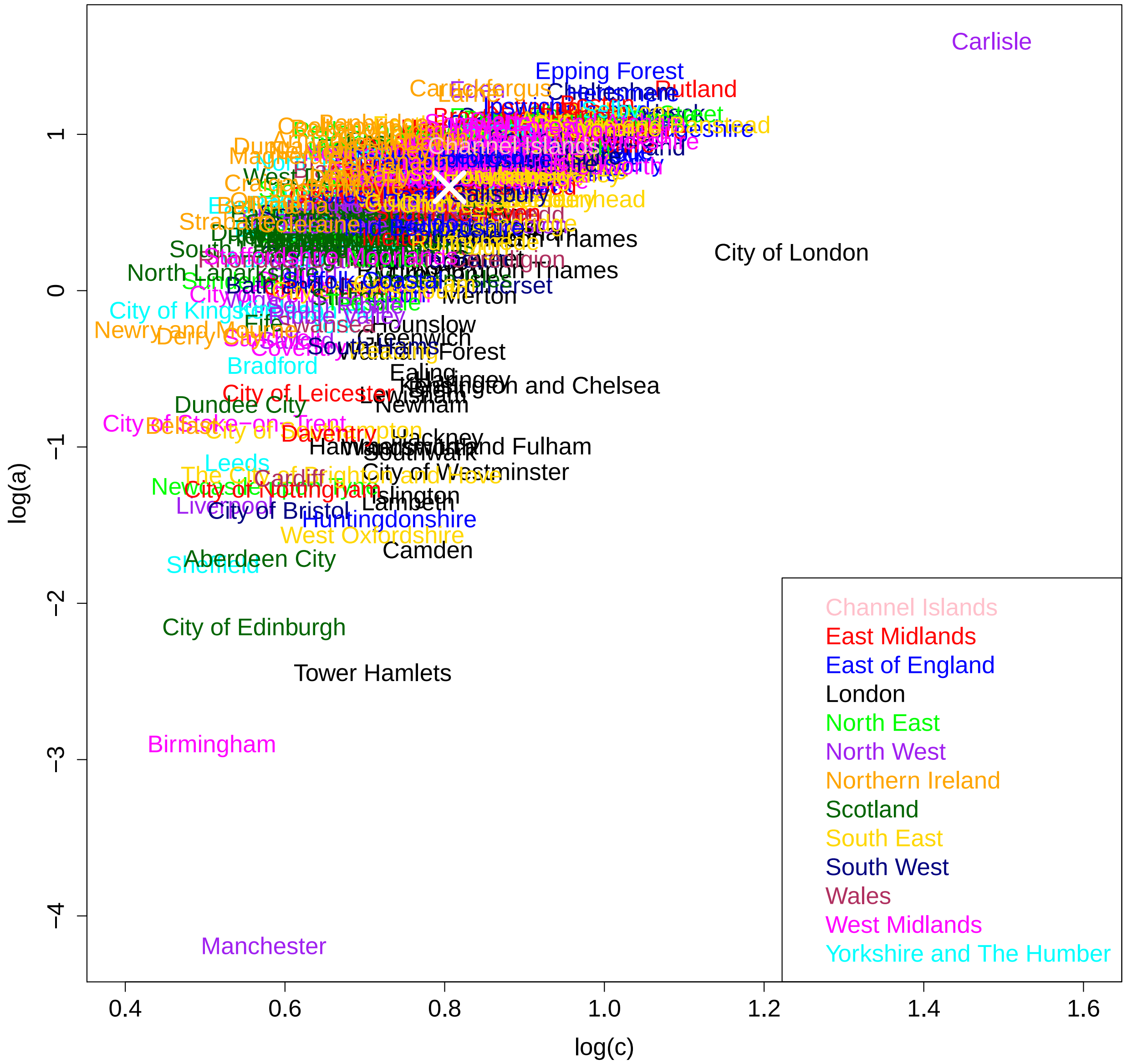


Figure 5.8: UK Surname distribution.

In chapter 2, it was noted that the parameter c controls the the probability of the distribution where the frequency equals one, whilst the parameter a controls the degree of skew of the distribution. This suggests that districts with lower values of c have a lower probability of a frequency of one, whilst higher values of c are more one-inflated. Similarly, for smaller values of a the distribution becomes more skew. Since, many of the cities have lower values of c and a we can infer that the distribution of surnames for these districts have a lower probability of a frequency of one, i.e. fewer unique surnames, and have a lower degree of skew than the average fitted Zipf model. Districts in Northern Ireland, have lower values of c than the average suggesting lower proportions of unique surnames but average skew, whilst districts in London have higher values of $log(c)$ i.e. a higher proportion of unique surnames but with a larger than average amount of skew, suggesting a distribution with high frequency of ones but a long tail. These results reflect the findings from the study of diversity in the UK surnames distribution (McElduff et al., 2010), which found that London, the South East and the East of England have higher surname diversity and Northern Ireland, Scotland, Wales has a less varying surname distribution.

As an example, the fit of Zipf models to the surname distribution in three districts can be examined: Manchester, an outlier with low values of c and a, the London district of Hackney (shown as central of the London district in Figure 5.8) and Carlisle, which has high values of both c and a. There are 313,241 people recorded in Manchester in the enhanced electoral register in 2001, with 6.10% of those people having unique surnames. The most common surname in Manchester is Smith (0.96% of the population), followed by Jones (0.82% of the population) and Taylor (0.62% of the population). The results of the maximum likelihood estimation function for the Zipf distribution (fitted values not shown) for frequencies of surnames in Manchester is:

```
> mle.ZIPF(sfreq.man)
Rapid Estimates
    re.coef
a1 1.147607
c1 1.496989
Maximum Likelihood Estimates
```

```
     mle.coef     mle.se     mle.LCI   mle.UCI
a1 0.0151426 0.01657920 -0.0173525 0.0476378
c1 1.7742500 0.00944124  1.7557400 1.7927500
Diagnostics
         chisq  df p      logL      AIC      BIC
model 12062.97 259 0 -77560.94 155125.9 155142.9
```

The Zipf model for frequencies of surnames in Manchester has parameter estimates $a = 0.015$ and $c = 1.77$. Although the Chi-squared goodness-of-fit p-value is significant, the test statistic is very large with a high degree of freedom due to the large sample size of the dataset. The fitted values for this distribution are plotted as a red solid line in Figure 5.9 showing the distribution of surnames in Manchester, with black points indicating observed values. There are a high number of unique surnames and the distribution is skew with a long tail. The model provides a good fit for unique surnames, but does not predict the tail of the distribution as well.

The number of persons recorded in the 2001 enhanced electoral register in Hackney was 132,771 and the proportion of people having unique surnames is 4.55% The most common surname in Hackney is once again Smith (0.70% of the population), with Williams (0.61% of the population) in second place and then Brown (0.50% of the population). The maximum likelihood estimation function for a Zipf model for the frequencies of surnames in Hackney is:

```
> mle.ZIPF(sfreq.hack)
Rapid Estimates
     re.coef
a1 0.9199753
c1 2.6639170
Maximum Likelihood Estimates
   mle.coef    mle.se  mle.LCI  mle.UCI
a1 0.385751 0.0253160 0.336132 0.435371
c1 2.204230 0.0168417 2.171220 2.237240
Diagnostics
         chisq  df p      logL      AIC      BIC
model 4014.895 119 0 -57988.37 115980.7 115997.6
```

This Zipf model has parameters $a = 0.39$ and $c = 2.20$. Again, we can note the goodness-of-fit test statistics are all large due the large sample size. Figure 5.10 plots the surname distribution in Hackney, with observed values of the frequencies of

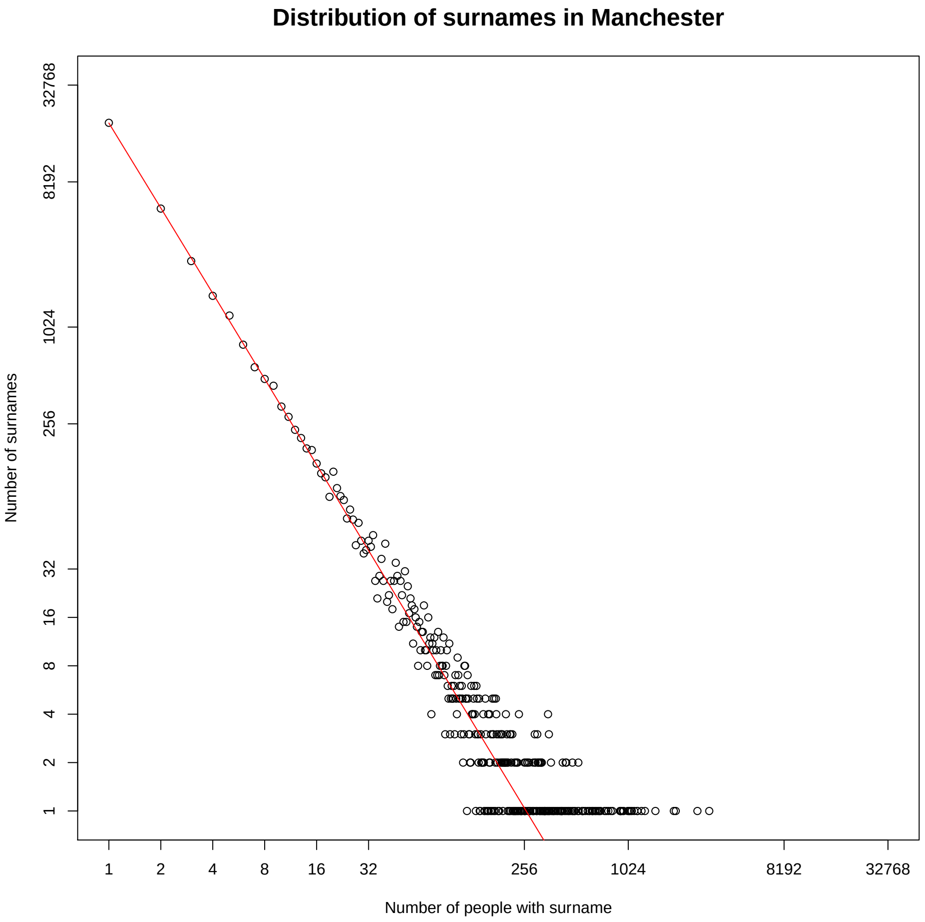


Figure 5.9: Observed and fitted values of Zipf model for surname frequencies of Manchester. Observed values are black points and the fitted model is shown in red.

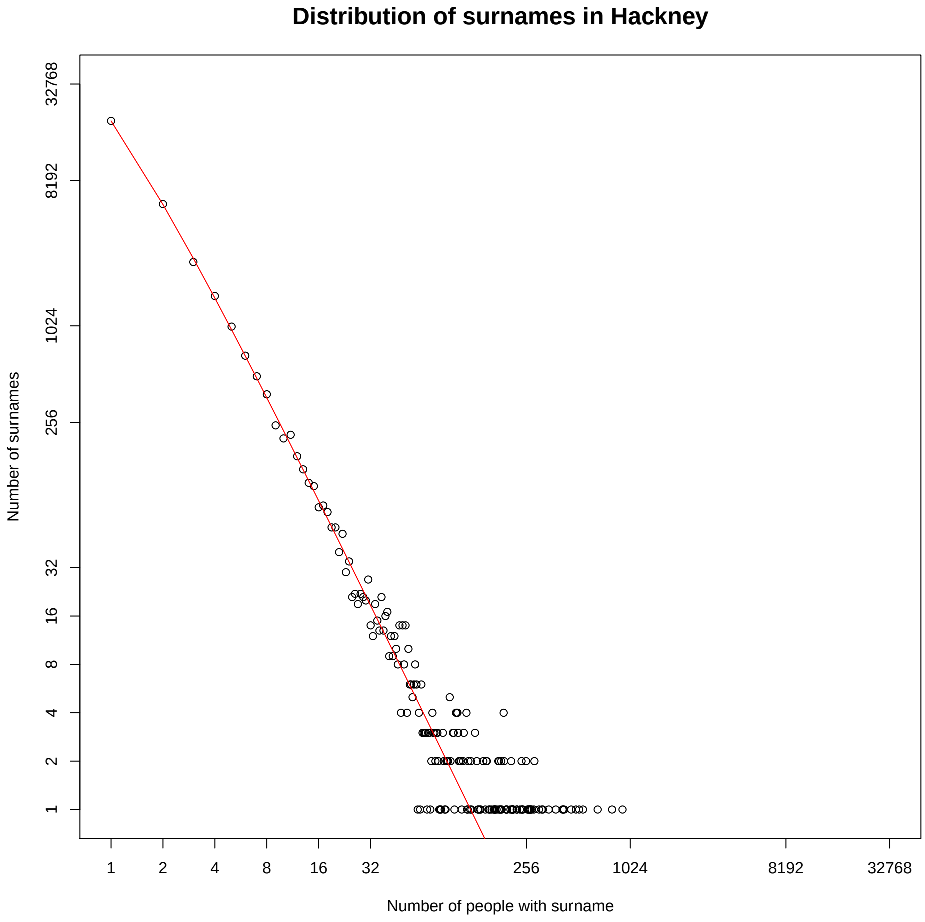


Figure 5.10: Observed and fitted values of Zipf model for surname frequencies of Hackney. Observed values are black points and the fitted model is shown in red.

surnames as black points and fitted values shown as a solid red line. This distribution has a lower value of c than Manchester and therefore has a lower proportion of unique surnames and the distribution of surnames is not as skewed with a shorter tail.

The number of people recorded in the 2001 enhanced electoral register in Carlisle is 81,069 and the percentage of people with unique surnames is 2.90%. The most popular surname in Carlisle is Graham (0.02% of population), followed by Bell (0.01% of population) and then Smith (0.01% of population). The Zipf model for surname frequencies in Carlisle fitted using maximum likelihood estimation is:

```
> mle.ZIPF(sfreq.carl)
Rapid Estimates
     re.coef
a1 -1.595062
```

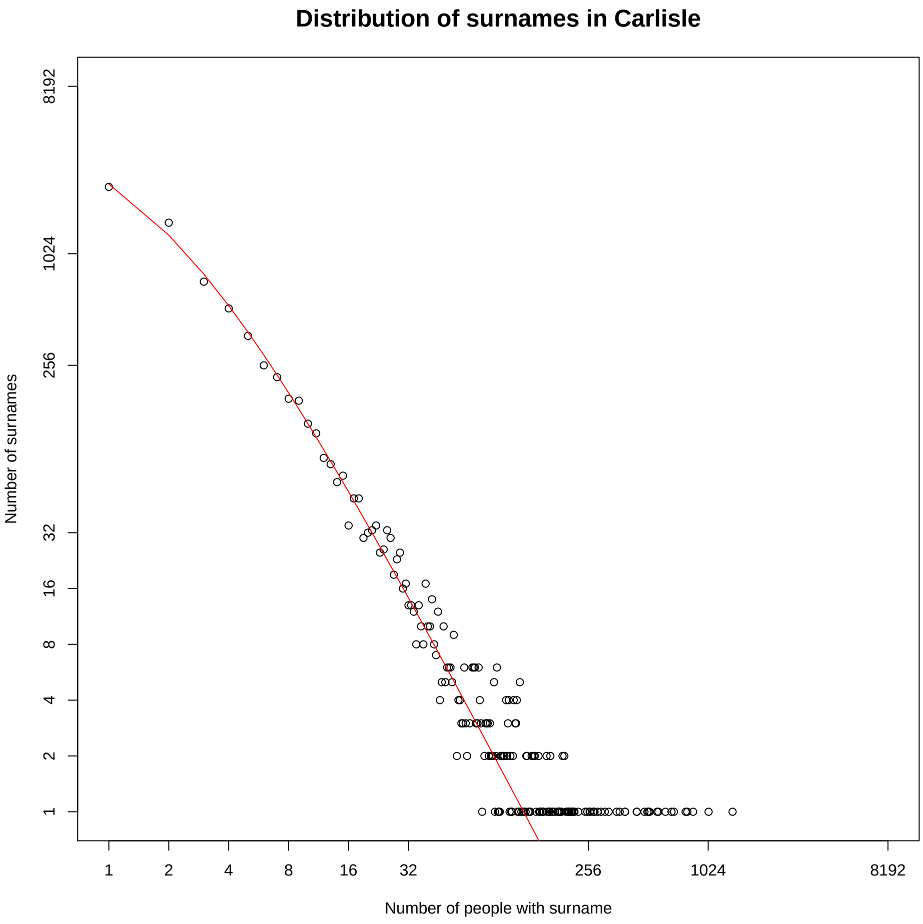


Figure 5.11: Observed and fitted values of Zipf model for surname frequencies of Carlisle. Observed values are black points and the fitted model is shown in red.

```
c1  5.566085
Maximum Likelihood Estimates
   mle.coef    mle.se mle.LCI mle.UCI
a1  1.75857 0.1026070 1.55746 1.95968
c1  2.05386 0.0276999 1.99957 2.10816
Diagnostics
        chisq  df p      logL     AIC      BIC
model 3919.226 118 0 -21012.05 42028.1 42042.01
```

This model has parameters $a = 1.76$ and $c = 2.05$. Goodness-of-fit statistics for this model are also large due to the large sample size. The surname distribution for Carlisle is shown in Figure 5.11 which plots the observed values of surname frequencies as black points and fitted values as a red solid line. The proportion of unique surnames is lower in Carlisle than the average and the distribution of surnames frequencies is less skewed than average with a shorter tail. The Zipf model appears to fit well for low frequencies of people with a given surname but underestimates the large frequencies of people with a given surname.

This example illustrates the application of the `Altmann` library in the modeling of Zipf distribution to frequencies of surnames in the UK. Parameter estimates for Zipf distributions in each district can be used to identify the distribution of surnames and provide an indication of the diversity and proportion of unique surnames in a district. Cities have been shown to have a higher surname diversity in comparison to other areas of the UK.

Summary

The `Altmann` library fits and compares discrete distributions using maximum likelihood estimation. Discrete distributions from a range of families have been implemented and provide increased complexity when modeling, improving the interpretation of discrete data. The UK surname distribution is an example of the application of this library.

Chapter 6

discrete.diag Library

The discrete.diag R library provides diagnostic analysis for univariate discrete models. Functions in this library fall into one of three categories: goodness-of-fit methods, model comparisons and techniques for outlier detection. These functions have been programmed as a supplement to the maximum likelihood estimation functions provided by the Altmann library. When the discrete.diag library is loaded into the R environment the Altmann library is automatically installed, if not already loaded. Functions in the discrete.diag use the Altmann library in one of two ways: objects of class 'mle' resulting from maximum likelihood estimation functions are used with S3 generic functions or alternatively, the maximum likelihood estimation functions are called directly within the diagnostic function.

In the first section, two functions for determining a distribution's goodness-of-fit are presented: the Chi-squared goodness-of-fit test and residual analysis functions. Model comparison functions to calculate a model's AIC and BIC and an EPGF plot function are presented in Section two. The third section, describes methods for the detection of outliers, these are the EPGF outliers plot and the Surprise Index. The usage and output of these functions is illustrated using the number of stillbirths in New Zealand white rabbits, which is used as an example in the previous Chapter. In the final section the outlier detection methods in the discrete.diag library are applied to a dataset featuring counts of cysts in steroid treated kidneys presented in Section 1.2 of Chapter One.

6.1 Goodness-of-fit Methods

These methods assess the fit of one particular distribution to an observed dataset. The first goodness-of-fit method is the Chi-squared goodness-of-fit test and the second calculates and plots randomized quantile residuals of a fitted model.

6.1.1 Chi-squared Goodness-of-fit Test

This function performs a Chi-squared goodness-of-fit test which tests the null hypothesis that the data follows a certain distribution i.e. the distribution provides a good fit to the observed data, against the alternative hypothesis that the distribution is not a good fit to the observed data. This function differs to the `chisq.test` function available in `R` as it adjusts the degrees of freedom for the number of parameters fitted in the model. The `chi.test` function has usage,

```
chi.test(yi, obs, exp, par)
```

where `yi` is the range of values of the discrete variable Y, `obs` and `exp` are the observed and expected frequencies of `yi` under a specified discrete distribution, respectively and `par` is the number of parameters estimated in the discrete distribution. The R code for the `chi.test` program is shown in Listing 6.1.

```
chi.test <-
function(yi, obs, exp, par){
# Chi-Squared Goodness of fit test

for(i in 1:length(exp)) if(exp[i]==0) exp[i]<-0.1
X2 <- chisq.test(obs, p=exp/sum(exp))
chisq <- X2$statistic[[1]]
df <- X2$parameter[[1]]-par
p <- 1-pchisq(chisq, df)

#Output
cat("Chi-square Goodness-of-fit test", "\n")
tab <- cbind(chisq, df, p)
dimnames(tab) <- list("model", c("chisq", "df", "p"))
print(tab)
}
```

Listing 6.1: Chi-squared Goodness-of-fit Test

Line 5 replaces any expected values of 0 by a small value. The Chi-squared test statistic is calculated in line 6, the degrees of freedom is then adjusted in line 8 and the corresponding *p*-value produced in line 9. A table of these values is created in lines 12-15 and is returned as the output of this function.

We continue to use the example of the number of stillbirths in litters of New Zealand White rabbits presented in Chapter 5 to demonstrate the use of the `chi.test` function. This dataset first needs to be loaded into `R` from the `Altmann` library.

```
> library(Altmann)
> data(rabbits)
```

In Chapter 5, the `altmann.fitter` model comparison shows the Waring distribution provides the best fit to the number of stillbirths. A Waring model can be fitted using `mle.WA` and a Chi-squared test performed using the resulting model of class `"mle"`,

```
> mod<-mle.WA(rabbits, printit=FALSE)
> chi.test(0:11, mod$obs, mod$exp, mod$npar)
Chi-square Goodness-of-fit test
         chisq df        p
model 7.206004  9 0.615681
```

The observed values are `mod$obs` the fitted values for the negative binomial distribution are `mod$exp` and `mod$npar` gives the number of parameters in the fitted model. The output shows the *p*-value is not significant at the 5% level and we can conclude that the data does follow a Waring distribution.

As a further example of the application of the `chi.test` function, we can also fit a Poisson distribution to the number of stillbirths,

```
> mod<-mle.POIS(rabbits, printit=FALSE)
> chi.test(0:11, mod$obs, mod$exp, mod$npar)
Chi-square Goodness-of-fit test
         chisq df p
model 287.3004 10 0
```

The Chi-squared goodness-of-fit *p*-value is very significant at the 5% level indicating that a Possion distribution does not provide a good fit to the number of stillbirths in New Zealand white rabbits.

6.1.2 Residuals

S3 objects can be utilized in R to construct a generic residuals function which calculates the randomized Quantile residuals of a model of class "mle" and produces plots for residual analysis. The residuals function has usage,

```
residuals(mleobject, family)
```

where mleobject is a model fitted using the maximum likelihood estimation functions in the Altmann library and family specifies the distribution fitted. The R code for this function is shown in Listing 6.2.

```
residuals.mle <- function (mleobject, family)
{
    y <- rep(mleobject$yrange, mleobject$obs)
    y.hat <- rep(mleobject$yrange, mleobject$exp)
    diff <- length(y.hat)-length(y)
    ifelse(diff>0,
           y.hat <- y.hat[-(1:abs(diff))],
           y.hat <- c(rep(0, abs(diff)), y.hat))
    mle.coef<-mleobject$mle.coef

    pfun <- paste("p",family,sep="")
    a<-rep(NA, length(y))
    for (i in 1:length(y)){
    a[i] <- ifelse((y[i]-1)>=0,
                   eval(call(pfun, y[i]-1, mle.coef)),
                   0)}
    b <- eval(call(pfun, y, mle.coef))
    u <- runif(n = length(y), min = a, max = b)
    R <- qnorm(u)
    par(mfrow = c(2, 2))
    plot(y.hat, R, main = "Against␣Fitted␣Values",
         xlab = "Fitted␣Values", ylab = "Residuals")
    plot(1:length(R), R, main = "Against␣Index",
         xlab = "Index", ylab = "Residuals")
    hist(R, main = "Histogram", xlab = "Residuals")
    qqnorm(R, main = "Normal␣Q-Q␣Plot",
           ylab = "Sample␣Residuals")
    qqline(R, col = "red")
    invisible(R)
}
```

Listing 6.2: Residual Analysis

The residuals function code has two parts: in the first part the randomized quantile residuals are calculated and the second creates plots of the residuals. Lines 3-9 extract vectors of the observed and fitted values and maximum likelihood parameter estimates from the `mleobject`. The cumulative probability function for the distribution in `family` is specified in line 11. To calculate the residuals, the observed values, `y`, are transformed to an interval (`a`,`b`) using the cumulative probability density and values are randomly generated from a uniform distribution between this interval (lines 12-18). In line 19, resulting uniform probabilities are used to produce randomized quantiles by using the inverse cumulative distribution function of a standard normal random variable. The residual R therefore gives the z-score for the specific observation.

The second part of the function uses the calculated residuals to construct a series of plots for residual analysis. Four plots are produced in lines 20-28:

- the residuals against fitted values
- the residuals against the index
- a Kernel density estimate of the residuals
- QQ normal plot of the residuals

The final command of the function (line 29) attaches an invisible copy of the residuals to the function. The residuals are therefore not printed as part of the function unless assigned to an object.

Once again the dataset containing the number of stillbirths in litters of New Zealand white rabbits illustrates the use of the `residuals` function. A negative binomial distribution can be fitted to the data and residual plots created using the following code,

```
> data(rabbits)
> mod1<-mle.NB(rabbits)
> residuals(mod1, family="NB")
```

Figure 6.1 plots the results of the `residuals` command for a negative binomial model. The plots show some residuals with high values which can be seen in the skewed histogram and evidence of non normality in the Q-Q plot.

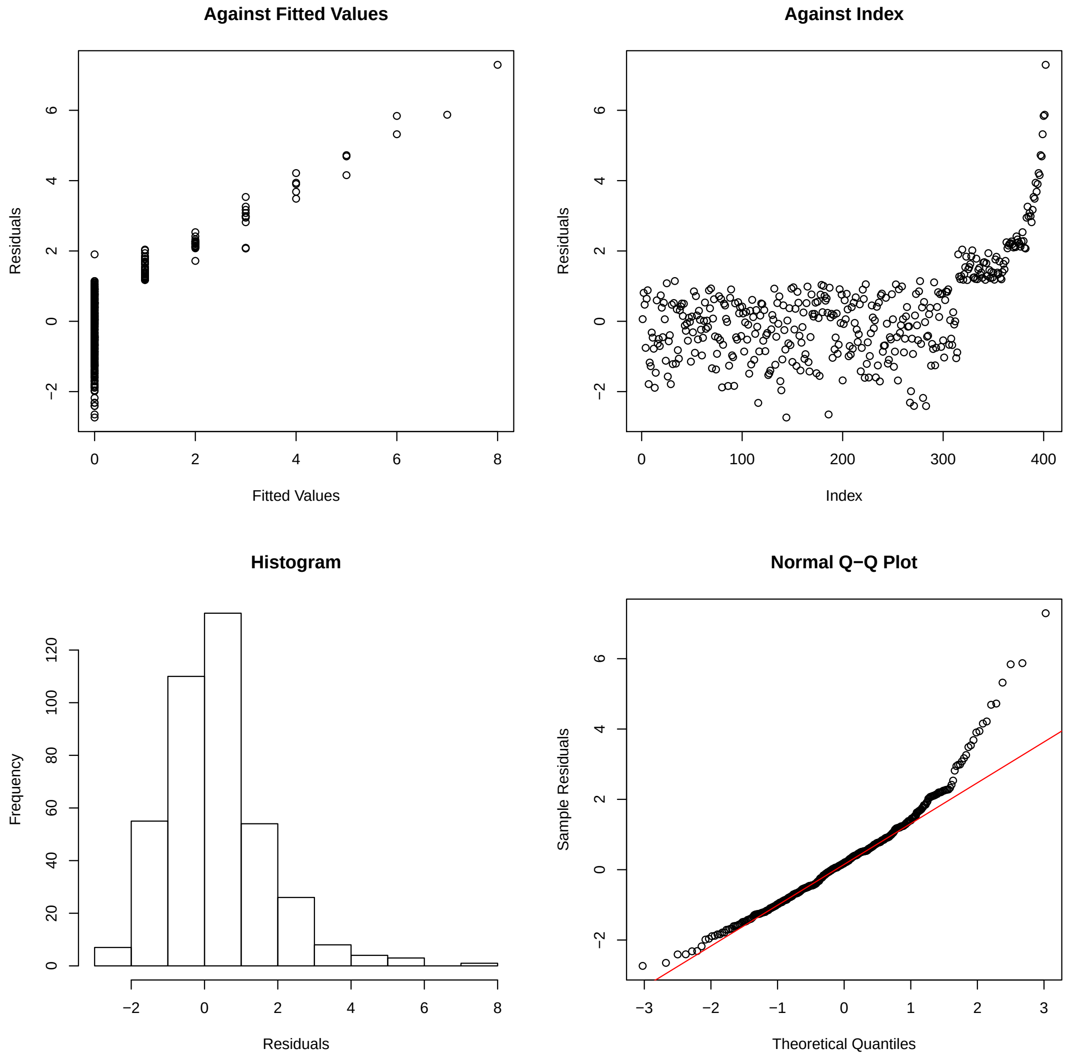


Figure 6.1: Residual plots for the number of stillbirths in litters of New Zealand White Rabbits under a negative binomial model.

6.2 Model Comparison

In this section two functions are presented to perform comparisons between fitted models: a function to calculate the AIC or BIC and the `epgf.plot` function. The `AIC` function produces a statistic that allows for a numerical comparison, whereas the `epgf.plot` provides a graphical representation of the data and fitted distributions.

6.2.1 AIC and BIC

This function makes use of the S3 object system to create a generic function that calculates the AIC or BIC. The AIC function has usage,

```
AIC(mleobject, bic = FALSE)
```

where `mleobject` is a model fitted using the maximum likelihood estimation functions in the `Altmann` library and `bic` is a logical argument specifiying whether the BIC should be returned as output, the default being `FALSE`. The code for this function is given in Listing 6.3 and comprises of if statements to determine whether the AIC or BIC is to be extracted from the mle object in lines 3 or 5. The value of AIC or BIC is then returned as output.

```
1 AIC.mle <-function(mleobject, bic=FALSE){
2 if(bic==FALSE){
3         aic<-mleobject$aic}
4 if(bic==TRUE){
5         aic<-mleobject$bic}
6 aic
7 }
```

Listing 6.3: `AIC` function

This function can be demonstrated for the frequency of stillbirths in litters of New Zealand white rabbits by fitting a negative binomial distribution using the `mle.NB` function. The `AIC` function extracts the AIC and BIC from this model:

```
data(rabbits)
> mod<-mle.NB(rabbits, method="moments", printit=FALSE,
>               plot.prof=FALSE)
> AIC(mod)
[1] 678.3545
```

```
> AIC(mod, bic=TRUE)
[1] 686.3474
```

The higher BIC values compared to the AIC is due to the penalty term of the BIC which adjusts for the number of parameter in the model.

6.2.2 EPGF plots

The `epgf.plot` function plots the EPGF for an observed dataset together with the *pgf*'s for a range of discrete distributions, allowing for comparisons between the fit of distributions. This function has usage,

```
   epgf.plot(ydata, family, tmin=-1, tmax=1, npts=100,
printit=FALSE, plotit=TRUE)
```

where the argument `ydata` is a vector of discrete observations and `family` gives a list of distributions to be fitted. The variables `tmin` and `tmax` give the minimum and maximum values of t for the EPGF and *pgf*'s, with the condition `tmin<tmax`. The argument `npts` is used to calculate values of t within the range `tmin`, `tmax`. The `printit` argument is logical and determines whether a matrix of EPGF and *pgf* values is printed, with the default being `FALSE` and `plotit` is also logical and determines whether a plot is produced, with default `TRUE`.

```
epgf.plot <-
function(ydata, family, tmin=-1, tmax=1, npts=100,
                         printit=FALSE, plotit=TRUE){

#Load packages
require(hypergeo)

###small printing functions
print.dist<-function(name1)
{
   print(paste(rep("=",40),sep=""),quote=F)
   print(paste("Dist", name1, sep="␣=␣"),quote=F)
   invisible(NULL)
}

#Set up the t's and range of yi
t1<- seq( tmin,tmax, length=npts)
yi<-min(ydata):max(ydata)
```

```

#A loop to work out the epgf
phin<-rep(NA,npts)
for(i in 1:npts) phin[i]<- log(mean( t1[i]^ydata))

pgf.current<-matrix(NA, ncol=length(family),
                    nrow=length(t1))

#A loop to work out epgf for different families
for (j in 1:length(family)){

#BERNOULLI
      if (family[j]=="BER") {
      if(printit) print.dist("Bernoulli")
      mod.BER<- mle.BER(ydata, printit=FALSE,
                        plot.prof=FALSE)
      prob<-mod.BER$mle.coef
      pgf<-1+prob*(t1-1)
      pgf.current[,j]<-pgf
      }
             .
             .
             .
             .
             .
#GENERALIZED GEGENBAUER
      if (family[j]=="GGE") {
      if(printit) print.dist("GG")
      mod.GGE<- mle.GGE(ydata, printit=FALSE,
                        plot.prof=FALSE)
      a<-mod.GGE$mle.coef[1]
      m<-mod.GGE$mle.coef[2]
      alpha<-mod.GGE$mle.coef[3]
      beta1<-mod.GGE$mle.coef[4]
      pgf <- (1-alpha-beta1)^a*
             (1-alpha*t1-beta1*t1^m)^(-a)
      pgf.current[,j]<-pgf
      }

}

if (plotit==TRUE){
matplot(t1,cbind(phin, log(pgf.current)),type="l",lwd=2,
        xlab="t", col=c(1, rainbow(length(family))),
        ylab="log␣(PGF)", lty=seq(1,length(family)+1, 1))
title(main=paste("EPGF␣plot␣of",
      deparse(substitute(ydata)), sep="␣"))
leg<-c("epgf", family)
67 legend(x="bottomright", leg, lwd=2,
68        lty=seq(1,length(family)+1, 1),
69        col=c(1, rainbow(length(family))))
70 }
71 else if (plotit==FALSE){
72 invisible(data.frame(family=family))
73 }
74 }
```

Listing 6.4: EPGF Plot function

The code for the `epgf.plot` function is given in Listing 6.4. Line 6 calls in the required `R` library `hypergeo` for the hypergeometric functions. A function is given in lines 9-14 which prints the name of the distribution as part of the output if the `printit` function is specified as `TRUE`. At the beginning of the function the range of t values is calculated in line 17, followed by the range of the y observations in line 18. Following this the EPGF is calculated in lines 20-22 using an iterative function. A storage matrix is provided for the *pgf*'s in line 24-25 with dimensions t number of rows and the number of distributions to be fitted as the number of columns.

This function once again uses the maximum likelihood estimating functions in the `Altmann` library to provide parameter estimates for distributions. For each distribution, the parameter estimates are extracted from the model and used to calculate the *pgf*. The values for each distribution are stored in the matrix. A series of if statements are used to select the appropriate distribution from the list given in the `family` argument and an iterative sequence performs this technique for each element in the vector of distributions specified by the `family` argument.

Following the calculation of the *pgf* matrix, if the `plotit` command is specified as `TRUE` then a plot of the EPGF and *pgf*'s is plotted using the commands in lines 60-70. If the `plotit` argument is `FALSE` an invisible table of the EPGF and *pgf*'s is instead returned.

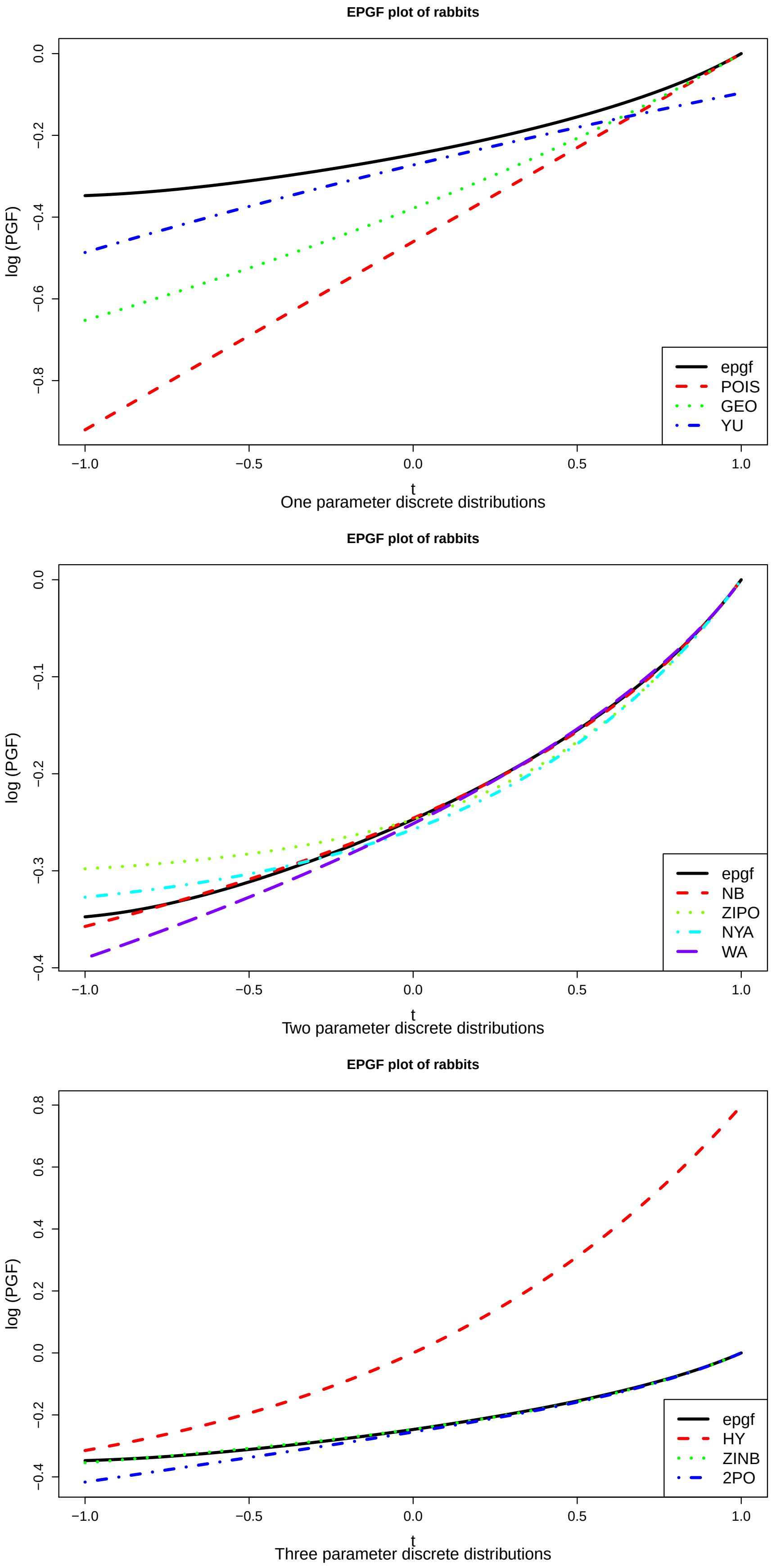


Figure 6.2: EPGF plots for the number of stillbirths in New Zealand white rabbits with a) Poisson, Geometric and Yule distributions, b) negative binomial, zero-inflated Poisson, Neyman type A and Waring distributions and c) hypergeometric, zero-inflated negative binomial and Poisson-Poisson mixture distributions.

The `epgf.plot` function can be applied to the numbers of stillbirths in New Zealand white rabbits as follows,

```
par(mfrow=c(3,1))
epgf.plot(rabbits, family=c("POIS", "GEO", "YU"))
title(sub="One␣parameter␣discrete␣distributions")
epgf.plot(rabbits, family=c("NB", "ZIPO", "NYA", "WA"))
title(sub="Two␣parameter␣discrete␣distributions")
epgf.plot(rabbits, family=c("HY", "ZINB", "2PO"))
title(sub="Three␣parameter␣discrete␣distributions")
```

Figure 6.2 shows 3 EPGF plots with a) one b)two and c) three parameter distributions, respectively. In each plot the EPGF of the number of stillbirths in litters of New Zealand white rabbits is shown by the solid black line for values of t between -1 and 1. The first plot shows that the *pgf* of the Yule distribution provides the closest fit to the EPGF of the three one parameter distributions. The Negative Binomial distribution *pgf*, shown by the red line in the second plot, indicates the negative binomial distribution is a good fit to the data. The zero-inflated Poisson, Neyman type A and Waring distributions do not fit the EPGF closely for values of t around -1. The final plot suggests that the Hypergeometric and Poisson-Poisson mix distributions are not good fits to the data, whilst the *pgf* of the zero-inflated negative binomial distribution follows the EPGF well. We can conclude from these plots that the negative binomial and zero-inflated negative binomial distribution appear to be the best fit to the data.

Whilst the EPGF plot provides a good visual comparison of the fit of several discrete distributions to a dataset, there can be some difficulty in deciding the most appropriate distribution for the data using the plots alone. The use of the AIC and/or BIC alongside can provide further insight into the fit of these distributions to a dataset. EPGF plots can be compared to the output from the `altmann.fitter` function in Section 5.6, which provides a table of goodness-of-fit statistics including the AIC and BIC. For the number of stillbirths in New Zealand white rabbits, it has been identified in Section 5.6 that the Yule distribution, provide the best fit to the data of the one parameter distributions. The Waring distribution has the highest chi-squared test p-value, however the EPGF plot perhaps indicates that this model does not fit the data as well as other models. It is therefore recommended that the EPGF plots are used as an exploratory tool, to identify several possible candidate models which can then be compared using goodness-of-fit statistics, such as those in the `altmann.fitter`.

6.3 Outlier Detection

In this section two functions for outlier detection are presented: the `outliers.plot` a non-parametric graphical method which uses the EPGF and the `surprise.plot` a parametric method which plots the *SI* for a distribution to determine if outliers are present within a dataset.

6.3.1 EPGF Outliers plot

This function plots the EPGF for a dataset using a leave-one-out procedure to determine if any outliers are present. The `outliers.plot` function has usage,

```
   outliers.plot(ydata, tmin = 0, tmax = 2, npts = 100, title0
= NULL)
```

where `ydata` is a vector of discrete observations, `tmin` and `tmax` give the minimum and maximum of the range of t, whilst `npts` calculates values of t within the range (`tmin`, `tmax`). The argument `title0` allows the user to specify a title for the plot produced. `R` code for this function can be found in Listing 6.5.

```
outliers.plot<- function (ydata, tmin = 0, tmax = 2,
                          npts = 100, title0=NULL)
{
    require(TeachingDemos)
    t1 <- seq(tmin, tmax, length = npts)
    epgf <- matrix(NA, nrow = npts, ncol = length(ydata))
    for (i in 1:length(ydata)) {
        r.ydata <- ydata[-i]
        ybar <- mean(r.ydata)
        phin <- rep(NA, npts)
        for (j in 1:100) {
            phin[j] <- log(mean(t1[j]^r.ydata))
        }
        epgf[, i] <- phin
    }
    matplot(t1, epgf, type = "l", lwd = 1, lty = 1,
        col = 1, pch = rep(1, length(ydata)), xlab = "t",
        ylab = "Log␣of␣PGF")
    title0<- ifelse(is.null(title0), paste("EPGF␣plot␣of",
        deparse(substitute(ydata)), sep = "␣"), title0)
    title(main = title0)

    l.epgf <- epgf[npts, ]
    dist.epgf <- rep(NA, length = length(l.epgf))
    for (j in 1:length(ydata)) {
        mean.epgf <- mean(l.epgf[-j])
        dist.epgf[j] <- abs(mean.epgf - l.epgf[j])
    }
    out.box<- boxplot(dist.epgf,plot=FALSE)$out
    n.out1<- length(out.box)
    if(n.out1 > 0)
    {
      subplot(boxplot(dist.epgf,col='lightblue'),
           x=tmin+(tmax-tmin)*0.1, y=max(as.vector(epgf)),
            vadj=1, hadj=0)
       epgf.out <- max(out.box)
       pos <- l.epgf[dist.epgf == epgf.out]
       out.pos<- (1:length(l.epgf))[l.epgf==unique(pos)]
       freq.out<- length(out.pos)
       y.out<- unique(ydata[out.pos])
       n.diff.val.out<- length( y.out)
       tab.out<- table(ydata[out.pos])
       for ( k in 1:n.diff.val.out)
       {
        print(paste(paste("A potential outlier of",
                    y.out[k]),
                    paste("with frequency", tab.out[k]),
                    paste("is detected in positions:",
                    paste(out.pos, sep='', collapse=',␣'))),
              quote=F)
       }
       for ( j in out.pos) lines(t1, epgf[, j],
                             col = "red", lwd = 1)
    }
    else
    {
       print ("there are no outliers", quote=F)
       out.pos<- 0
    }
    invisible(out.pos)
}
```

Listing 6.5: EPGF outliers plot

This function has two parts: the EPGF's are calculated and plotted using a leave-one-out procedure and then possible outlying values are identified. Firstly, the `TeachingDemo` library is loaded in line 4. To calculate the EPGF's the values of t are generated in line 5 for values between `tmin` and `tmax`. A storage matrix is then constructed in line 6 for the EPGF values. An iterative sequence is used to remove each observation systematically and the EPGF for the remainder of the dataset is calculated in lines 7-15. The EPGF curves are plotted in lines 16-21.

In the second part of the function the EPGF curves are used to identify the maximum possible outlying value. In line 23 the EPGF values are extracted at the maximum value of t, `tmax`. Lines 24-28 calculate the absolute difference between each curve at this value of t and the mean of the remaining EPGF values, which are stored in the vector `dist.epgf`. An unplotted boxplot of these absolute differences is used to identify the maximum values of any outliers, i.e. any observations with values, that differ from the average of the remaining differences. If there are outlying observations then a boxplot of the data is plotted in the top left hand corner of the plot (lines 31-54). The function then returns the value of the maximum possible outlying observation and highlights the EPGF curve for which is calculated excluding this observation. If no outliers are detected, in lines 55-59 a statement is printed to this affect. Line 60 attaches an invisible table of EPGF values to the function.

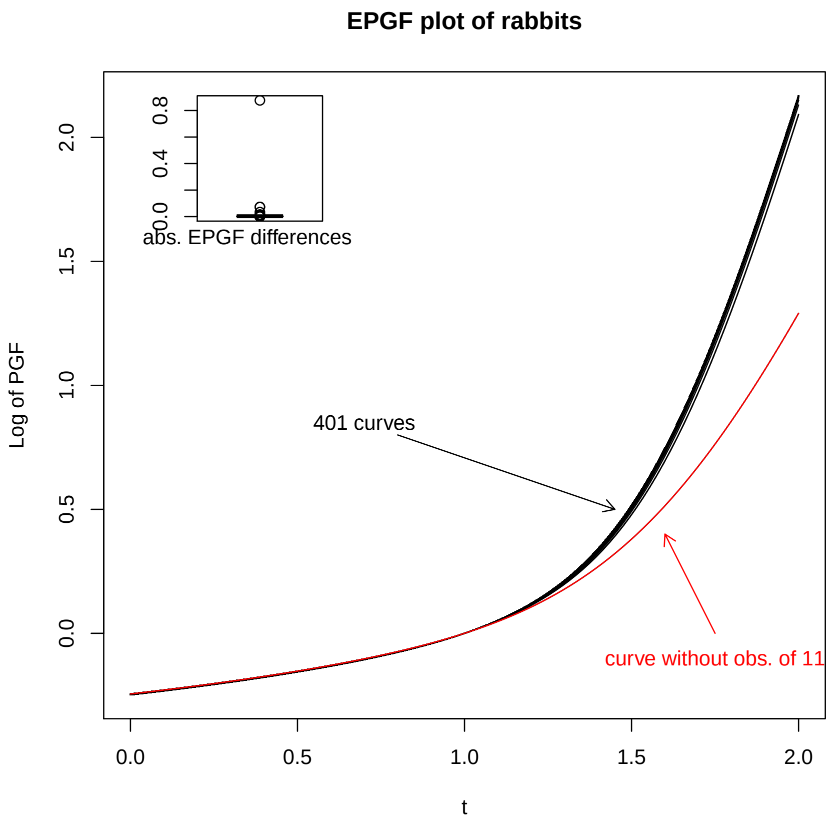


Figure 6.3: EPGF outliers plot of the rabbits data,

The dataset comprising of the number of stillbirths in litters of New Zealand white rabbits contains an observation of 11 stillbirths which may be an outlier with respect to the remainder of the dataset. To determine whether this observation is an outlier we use the `epgf.plot` function,

```
> data(rabbits)
> par(mfrow=c(2,1))
> outliers.plot(rabbits)
[1] A potential outlier of 11 with frequency 1
 is detected in positions:  402
```

The EPGF plot shown in Figure 6.3 shows 402 EPGF curves each calculated with 401 observations. The output of the function indicates that the observation with 11 stillbirths is considered an outlier. The EPGF curve calculated without this observation can clearly be seen (highlighted on the plot in red) and is substantially different to the other curves. Removing this observation we can again produce a further plot of the

EPGF curves,

```
> outliers.plot(rabbits[-402])
[1] A potential outlier of 8 with frequency 2 is
    detected in positions:  400, 401
```

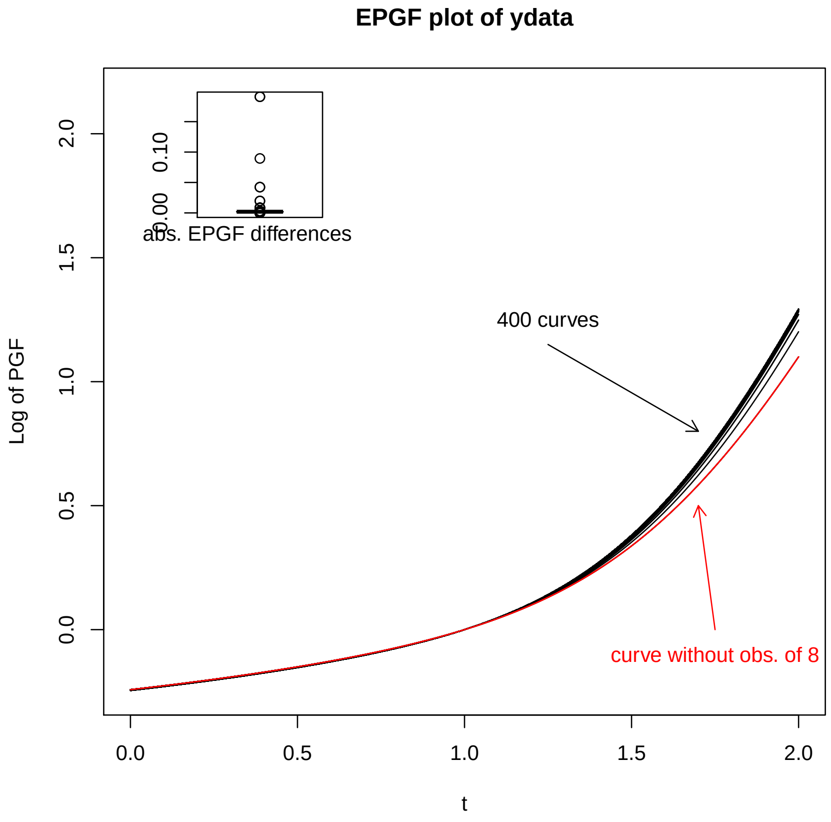


Figure 6.4: EPGF outliers plot of the rabbits data,

The above output and Figure 6.4 shows that the next possible outlying observations would have 8 stillbirths in a litter. The EPGF curve for the dataset without this observation shows no substantial differences to the other curves. We can conclude that the EPGF plots indicate that the observation of 11 stillbirths can be considered as an outlier with respect to the remainder of the dataset.

6.3.2 Surprise Index plot

The `surprise.plot` function calculates and plots the *SI* for a selected distribution where any y values with *SI*'s greater than 1,000 are considered to be outlying observations.

This function has usage,

```
surprise.plot(ydata, family = "POIS", ylim = 100,
              plot.log = TRUE)
```

with arguments `ydata` a vector of discrete observations, the `family` argument specifies the discrete distribution fitted to the observations and used to calculate the *SI*, `y.lim` is a constant giving the limit of y sum in the *SI* and `plot.log` is a logical argument with default `TRUE` determining whether the logarithm of the *SI* should be plotted. `R` code for this function is given in Listing 6.6.

```
surprise.plot <-
function(ydata, family="POIS", ylim=100, plot.log=TRUE){

require(gsl)
require(hypergeo)

yi<-min(ydata):max(ydata)

#BERNOULLI
if (family=="BER") {
  mod.BER<-mle.BER(ydata, printit=FALSE, plot.prof=FALSE)
  prob<-mod.BER$mle.coef
  SI<- -(((1-prob)^(-1+yi)*(-1+prob)^4*prob^(-yi))/
        (-1+2*prob))
}
              .
              .
              .
              .
              .
#GENERALIZED GEGENBAUER
if (family=="GGE") {
  mod.NB<-mle.GGE(ydata, printit=FALSE, plot.prof=FALSE)
  a<-mod.GGE$mle.coef[1]
  m<-mod.GGE$mle.coef[2]
  alpha<-mod.GGE$mle.coef[3]
  beta1<-mod.GGE$mle.coef[4]
  py <- dGGE(yi, a, m, alpha, beta1)
  SI <- sum(py^2)/py
}

#Plot the graph

if(plot.log==TRUE){
```

```
35         plot(yi, SI, type="b", lwd=2, log="y",
36              ylab="log(SI)")
37         title(main=paste("Plot␣of␣Surprise␣Index␣for",
38               deparse(substitute(ydata)), sep="␣"))
39         abline(h=log(1000), lty=2, lwd=2, col="red")
40         leg<-c(family, "Surprising")
41         legend("topleft", leg, lwd=2, lty=1:2,
42                col=c("black", "red"))
43 }
44
45 if(plot.log==FALSE){
46         plot(yi, SI, type="b", lwd=2, ylab="SI")
47         title(main=paste("Plot␣of␣Surprise␣Index␣for",
48               deparse(substitute(ydata)), sep="␣"))
49         abline(h=1000, lty=2, lwd=2, col="red")
50         leg<-c(family, "Surprising")
51         legend("topleft", leg, lwd=2, lty=1:2,
52                col=c("black", "red"))
53 }
54 }
```

Listing 6.6: Surprise Index function

Lines 4 and 5 load in the required libraries into `R`, whilst line 7 calculates the range of `ydata`. For each distribution an if statement locates the family specified in the `family` argument. Parameter estimates are extracted from models fitted using maximum likelihood functions in the `Altmann` library and then the *SI* calculated. The *SI* is then plotted in lines 34-53 using if statements to determine whether the *SI* or log of the *SI* should be plotted as specified in the `plot.log` argument.

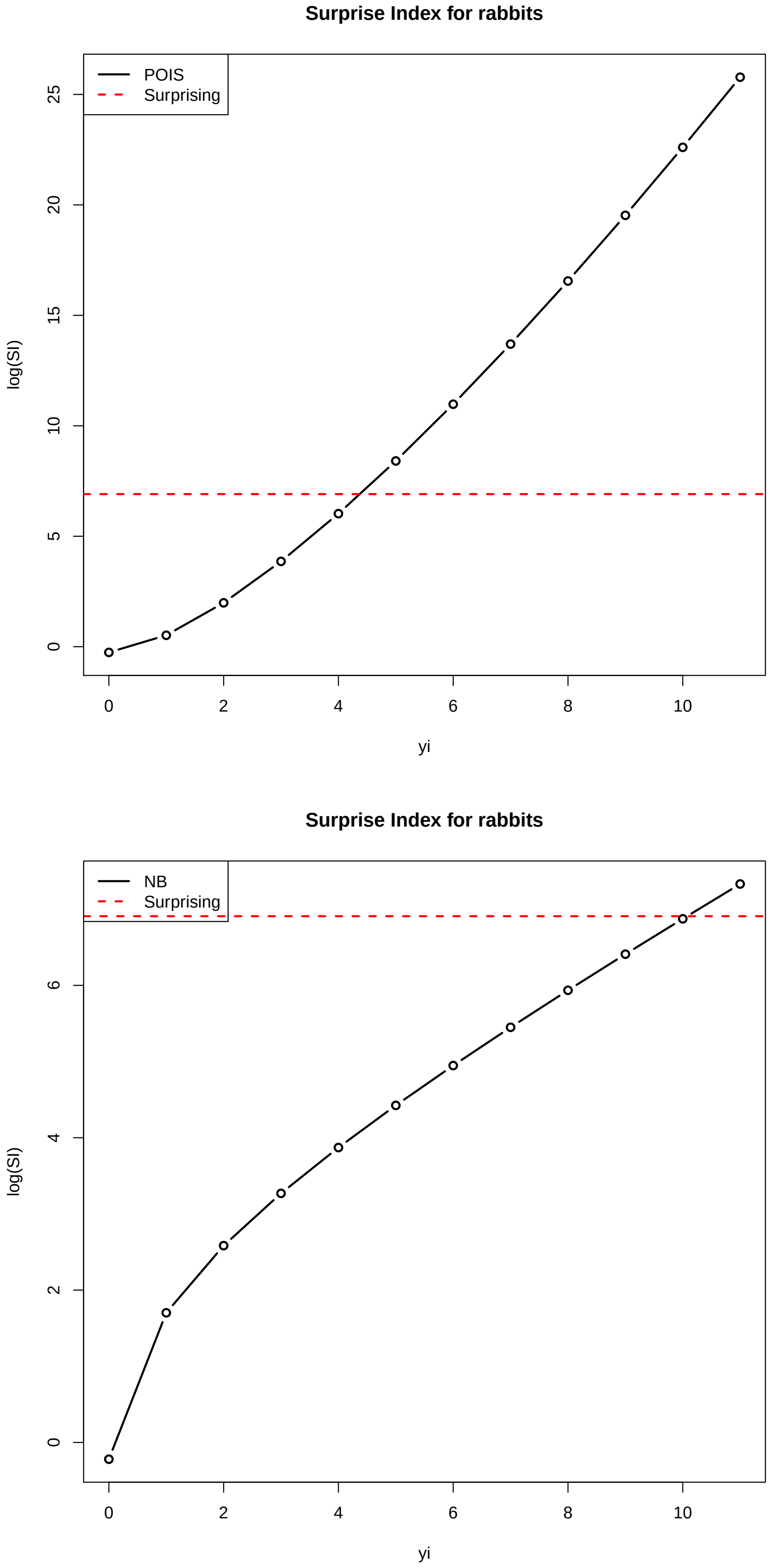


Figure 6.5: *SI*'s for the number of stillbirths in New Zealand White rabbits under Poisson and Negative Binomial distributions.

We can again investigate whether the extreme value of 11 stillbirths in litters of New Zealand white rabbits could be considered an outlier. The surprise.plot function can be used to calculate the *SI* under Poisson and negative binomial models,

```
> data(rabbits)
> par(mfrow=c(2,1))
> surprise.plot(rabbits, family="POIS", plot.log=F)
      0     1     2     3      4        5        6        7
1 0.772 1.678 7.291 47.53 413.13 4488.601 58521.65 890161.8
         8         9         10           11
1 15474380 302628687 6576039582 157185119082
> surprise.plot(rabbits, family="NB", plot.log=F)
      0     1      2      3      4     5       6       7
1 0.803 5.486 13.245 26.308 47.998 83.49 140.853 232.622
        8       9      10       11
1 378.206 607.553 966.732 1526.427
```

The *SI*'s for the Poisson and negative binomial distributions plotted as output from the surprise.plot function are given in Figure 6.5. These graphs indicate that under a Poisson distribution values greater than 4 are considered to be surprising. However, under a negative binomial distribution a value of 11 stillbirths would be considered surprising and thus an outlier.

6.4 Validation of the functions

Functions in the discrete.diag library include the chi.test, residuals.mle, AIC, epgf.plot, outliers.plot and surprise.plot. These were built using a trial and error process, with the basic functions initially programmed and then expanded to incorporate other arguments and produce output tables. The chi.test function was validated using comparisons to the Chi-squared statistics for models estimated using the Altmann fitter program (Altmann, 1997). This was performed alongside testing for the maximum likelihood estimation functions. The outputs from the epgf.plot and outliers.plot methods were also compared to examples given in Nakamura and Pérez-Abreu (1993b) and Nakamura and Pérez-Abreu (1993a) (shown in Section 3.3.2). Functions to calculate surprise indices used in the surprise.plot function were also tested for a range of parameter values for each distribution and plotted using various simulated and real test datasets, to confirm that the functions were performing correctly and resulted in the correct values.

6.5 Application to counts of cysts in steroid treated foetal mouse kidneys

Section 1.2 in the first chapter presents data from a study on the effect of a low protein diet in mice on kidney development in their offspring. Data on counts of cysts in embryonic mouse kidneys which had been subjected to steroids were featured in this study. This dataset was analysed to compare counts of cysts from $n = 111$ steroid treated kidneys and $n = 103$ untreated (control) kidneys using *t*-tests, Wilcoxon-Mann-Whitney tests and discrete regression modelling (McElduff et al., 2010). Cyst counts for the steroid and untreated kidney groups are given in Tables 6.1 and 6.2. The steroid group has one kidney with 19 cysts, which is much higher than the maximum number of cysts found in the control group of kidneys (maximum=3). A high number of cysts indicates abnormal kidney growth and so we investigate whether the kidney with a count of 19 cysts in the steroid treated group is an outlying observation.

Frequency	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**19**
Cyst Counts	65	14	10	6	4	2	2	2	1	1	1	2	1

Table 6.1: Counts of Cysts in steroid treated kidneys

Frequency	**0**	**1**	**2**	**3**
Cyst Counts	94	4	4	1

Table 6.2: Counts of Cysts in control kidneys

The analysis in this section is presented in three parts. Firstly, counts of cysts in steroid treated kidneys are assessed for outliers using the EPGF method. A range of models are fitted in the second section to determine the distribution which best models the counts of cysts. In the final section, the presence of outliers in the dataset is tested for a range of models by plotting the *SI*.

6.5.1 Outlier Detection using the EPGF

Two methods for the detection of outliers in discrete distributions can be found in the `discrete.diag` library. The EPGF method of detecting outliers is non-parametric and does not assume the data follows any type of model. We can analyse counts of cysts in steroid-treated kidneys for outliers using the `outlier.plot` function as follows,

```
> outliers.plot(steroid)
[1] A potential outlier of 19 with frequency 1 is
    detected in positions:  3
```

The result of the above `R` command is shown in Figure 6.6. The graph plots 111 EPGF curves each calculated with 110 observations. One EPGF curve (highlighted in red) differs substantially from the remainder of the curves, with large values of the EPGF for t between 1 and 2. Removing the observation with 19 cysts which is in position two of the steroid data vector, the EPGF outliers plot can be refitted to see the affect that observation has on the output.

```
> outliers.plot(steroid[-3])
[1] A potential outlier of 11 with frequency 2 is
    detected in positions:  12, 31
```

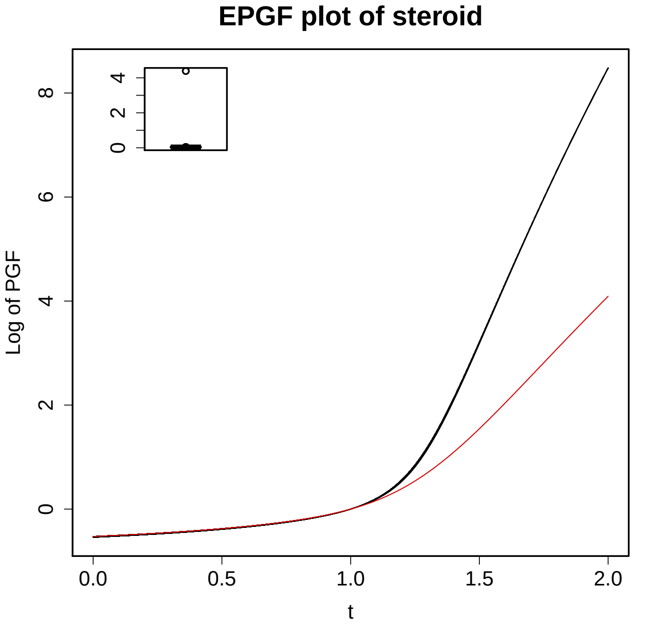


Figure 6.6: EPGF outlier plots of counts of cysts in steroid treated foetal mouse kidneys

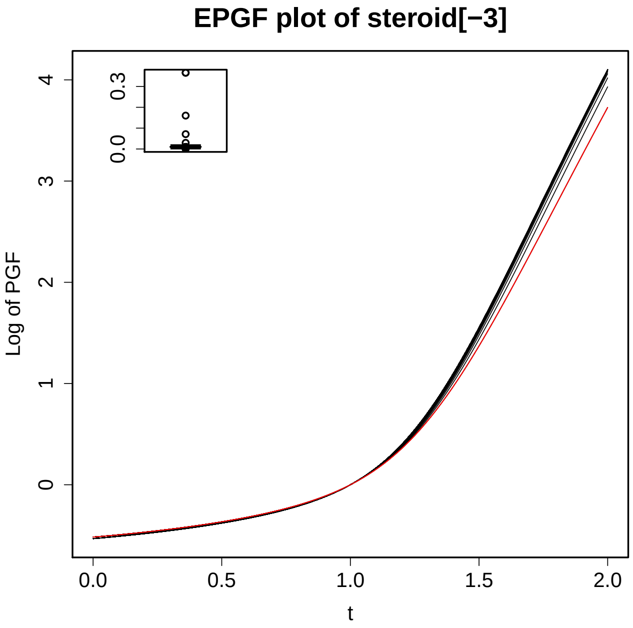


Figure 6.7: EPGF outlier plots of counts of cysts in steroid treated foetal mouse kidneys without observation of 19 cysts.

The resulting output for this command is featured in Figure 6.7 which plots 110 EPGF curves each calculated with 109 observations. The curve highlighted in red for the kidney with 11 cysts does not differ from the other curves. This leads us to conclude that the kidney with 19 cysts would be considered as an outlier with respect to the remainder of the dataset.

6.5.2 Model fitting

The second outlier detection method utilizes the *SI* which is dependent upon the distribution fitted to the data. Using the `altmann.fitter` a range of 12 distributions can be fitted to the counts of cysts in steroid treated kidneys and compared using the goodness-of-fit values,

```
> altmann.fitter(steroid, family=c("POIS", "GEO", "NB", "HY",
+                                  "HO", "YU", "WA", "ZIPO",
+                                  "ZINB", "2PO", "2PNB", "NYA"))
   Distribution n.par      AIC      BIC     chisq df chisq.p
3            NB     2 353.6263 359.0454  14.37728 17  0.6402
6            YU     1 357.2297 359.9393  16.68914 18  0.5446
7            WA     2 356.5583 361.9774  18.88797 17  0.3350
5            HO     2 357.3904 362.8094  20.32699 17  0.2578
```

```
9          ZINB     3 355.3196 363.4482  16.46417 16  0.4211
11         2PNB     4 357.3625 368.2006  17.78310 15  0.2742
12          NYA     2 367.8920 373.3111  29.00046 17  0.0345
2           GEO     1 381.0691 383.7787  50.93147 18  0.0001
10          2PO     3 377.7285 385.8570  28.86159 16  0.0249
8          ZIPO     2 408.6673 414.0863 120.79333 17  0.0000
1          POIS     1 561.4071 564.1166 257.19810 18  0.0000
4            HY     3 659.3490 667.4776 390.84982 16  0.0000
```

The Negative Binomial distribution is the best fit to the data of the 12 models fitted to counts of cysts in embryonic mouse kidneys. This model has the smallest BIC value at 359.05 and the highest χ^2 test statistic p-value of 0.64. These results therefore suggests that the negative binomial distribution provides a good fit to the data.

The data-generating mechanism of the negative binomial model can be used to explain the distribution of counts of cysts. We assume that the data is generated from a Poisson-Gamma parameter-mix, with counts of cysts following a Poisson distribution with one parameter, the mean number of cysts, which varies according to a Gamma distribution. This interpretation of the negative binomial model suggests the underlying capacities of the kidneys may or may not be identical. A Poisson model assumes that they are the same, whereas the negative binomial model allows for variation. Parameter estimates and goodness-of-fit statistics for the negative binomial model can be fitted by maximum likelihood estimation using the `mle.NB` function,

```
> cysts.NB1<-mle.NB(steroid)
Rapid Estimates
    re.coef
r 0.3325390
p 0.1766862
Maximum Likelihood Estimates
  mle.coef    mle.se   mle.LCI  mle.UCI
r 0.296159 0.0650951 0.1918850 0.455564
p 0.160460 0.0439025 0.0932997 0.246146
Fitted Values
        0     1     2    3    4   5    6    7   8    9
obs 65.00 14.00 10.00 6.00 4.00 2.0 2.00 2.00 1.0 1.00
exp 64.56 16.05  8.73 5.61 3.88 2.8 2.08 1.57 1.2 0.93
      10   11   12   13   14   15   16   17   18  19
obs 1.00 2.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.0
exp 0.72 0.57 0.45 0.36 0.28 0.23 0.18 0.15 0.12 0.1
Diagnostics
         chisq df         p      logL      AIC      BIC
model 14.37728 17 0.6402388 -174.8132 353.6263 359.0454
```

This model estimates the parameters of the negative binomial distribution as $r = 0.30$ (95% CI: (0.17, 0.42)) and $p = 0.16$ (95% CI: (0.07, 0.25)). Fitted values estimate the observed values well at lower values of numbers of cysts.

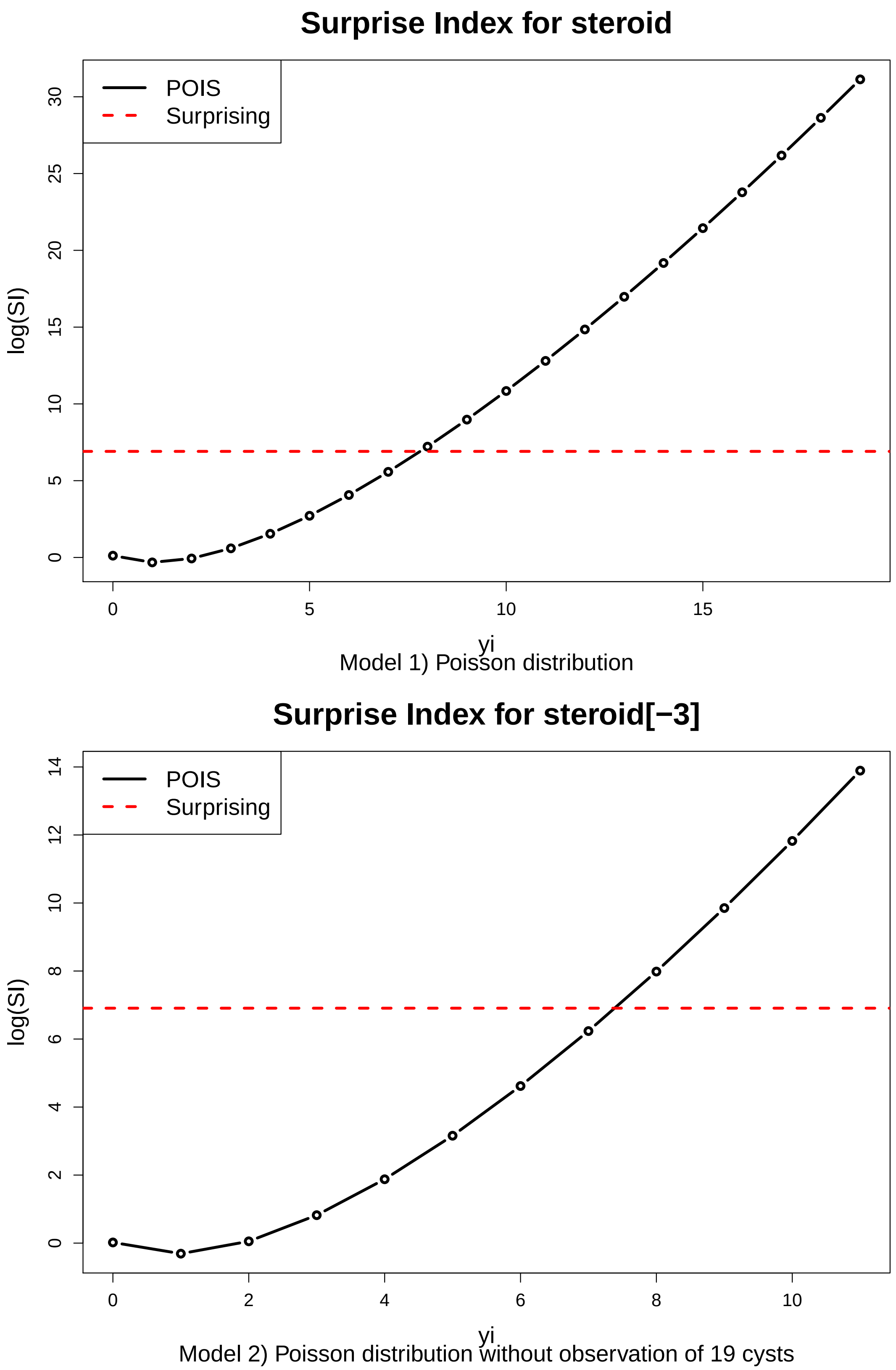
6.5.3 Outlier detection using Surprise Index

We can fit the *SI* for the Poisson and negative binomial distributions using the `surprise.plot` function. Four models are fitted to the number of cysts in steroid treated kidneys: 1) a Poisson model, 2) a Poisson model excluding the value of 19 cysts, 3) a negative binomial model and 4) the negative binomial model excluding the kidney with 19 cysts. Plots of the *SI* for these four models are shown in Figures 6.8 and 6.9 and values of the *SI* for the four models can be found in Table 6.3.

Frequency	**1. Poisson**	**2. Poisson without outlier**	**3. NB**	**4. NB without outlier**
0	1.12	1.02	0.63	0.64
1	0.73	0.73	2.56	2.47
2	0.94	1.05	4.71	4.59
3	1.81	2.27	7.33	7.31
4	4.68	6.54	10.60	10.84
5	15.10	23.51	14.69	15.44
6	58.43	101.40	19.82	21.43
7	263.96	510.30	26.25	29.22
8	1362.76	2935.04	34.29	39.30
9	7915.12	1.90×10^4	44.31	52.33
10	5.11×10^1	1.37×10^5	56.77	69.12
11	3.64×10^5	1.08×10^6	72.24	90.70
12	2.81×10^6	–	91.41	–
13	2.36×10^7	–	115.12	–
14	2.13×10^8	–	144.38	–
15	2.06×10^9	–	180.44	–
16	2.13×10^{10}	–	224.81	–
17	2.33×10^{11}	–	279.35	–
18	2.71×10^{12}	–	346.28	–
19	3.32×10^{13}	–	428.33	–

Table 6.3: Table of *SI*'s for the four models.

SI's for the Poisson distribution in model 1) are given in Table 6.3 and are plotted in the first plot of Figure 6.8. Any observations with values greater than 7 are considered to be outliers. The table of model comparisons in the previous section shows this to be



Model 1) Poisson distribution

Model 2) Poisson distribution without observation of 19 cysts

Figure 6.8: *SI* plots of counts of cysts in steroid treated foetal mouse kidneys for models 1) a Poisson distribution and 2) a Poisson distribution excluding the kidney with 19 cysts.

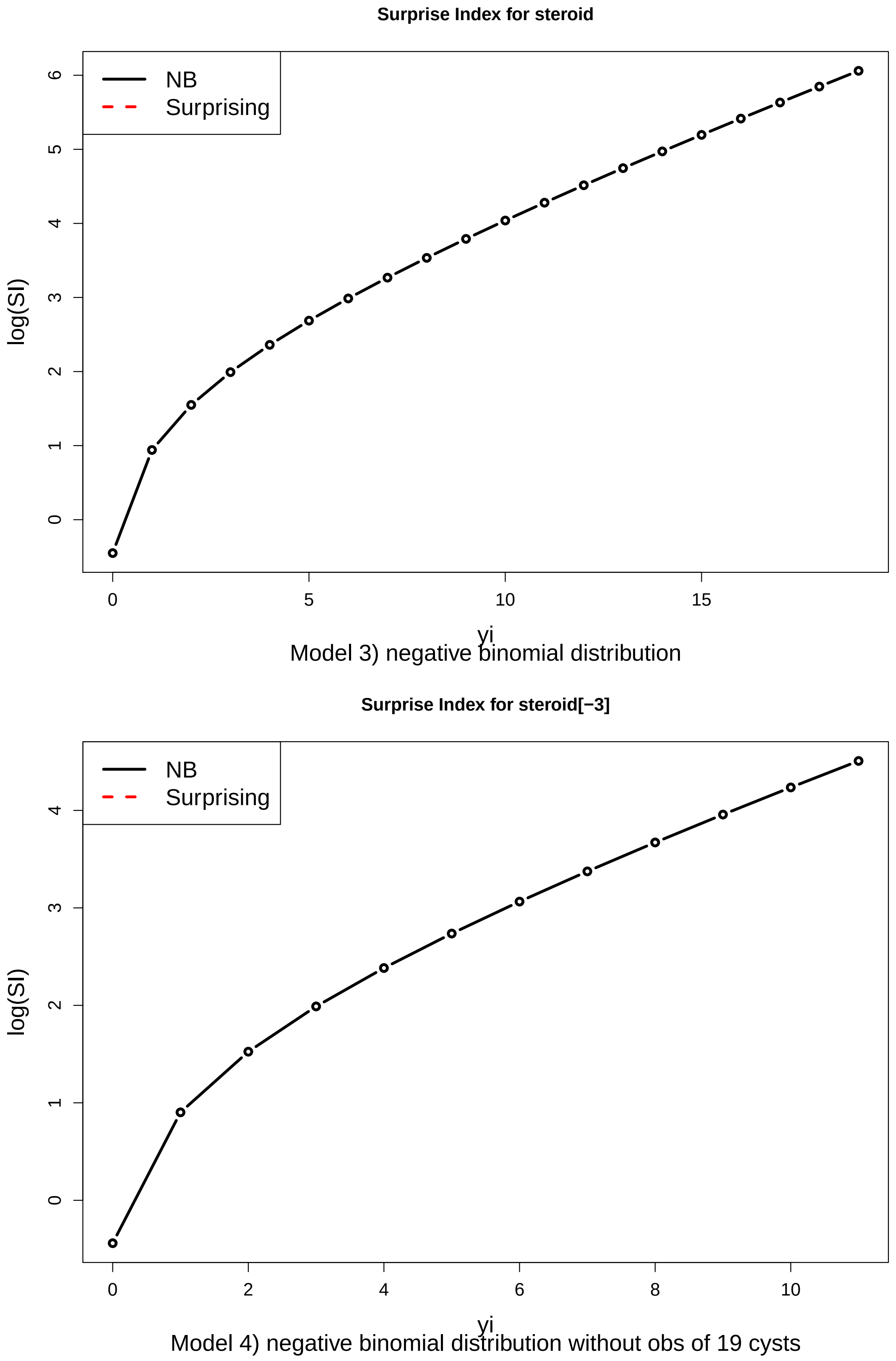


Figure 6.9: *SI* plots of counts of cysts in steroid treated foetal mouse kidneys for models 1) a negative binomial distribution and 2) a negative binomial distribution excluding the kidney with 19 cysts.

the worst of the models fitted to the data. If we fit a Poisson distribution without the observation of 19 cysts in a kidney (model 2) shown in the second plot of Figure 6.8 there are still surprising values detected and anything greater than 7 is still considered an outlier.

SI plots for model 3, the negative binomial model, and model 4, the negative binomial model which excludes the potentially outlying observation of 19, are given in Figure 6.9. In the first plot, under a negative binomial model there are no values that are considered to be surprising. If the kidney with 19 cysts is excluded from the *SI* calculation, then under the negative binomial model (model 4) there are still no surprising values and therefore no outliers detected. These results suggests that the value of 19 cysts is to be considered an outlier under a Poisson model but if we assume a negative binomial model, which can account for overdispersion in the model, then this observation is not an outlier. Model 4 which fits a negative binomial distribution removing the observation of 19 cysts can be fitted using the `mle.NB` function,

```
> cysts.NB2<-mle.NB(steroid[-3])
Rapid Estimates
    re.coef
r 0.4146782
p 0.2296639
Maximum Likelihood Estimates
  mle.coef    mle.se  mle.LCI  mle.UCI
r 0.321606 0.0742157 0.204391 0.507190
p 0.187782 0.0511094 0.109870 0.286349
Fitted Values
        0     1     2    3    4    5    6    7    8    9  10   11
obs 65.00 14.00 10.00 6.00 4.00 2.00 2.00 2.00 1.00 1.00 1.0 2.00
exp 64.24 16.78  9.01 5.66 3.82 2.68 1.93 1.42 1.05 0.79 0.6 0.46
Diagnostics
         chisq df         p      logL      AIC      BIC
model 6.385447  9 0.7008068 -167.5441 339.0881 344.4891
```

The above model's parameters values are similar to those of the negative binomial distribution in model 3 and the BIC is improved only slightly in comparison.

We can conclude that the observation of 19 may be considered an outlier under the Poisson distribution which does not allow for overdispersion in the model. If we account for overdispersion by fitting a negative binomial distribution then we do not consider this value to be an outlier and we can include it in the model - indeed including this values does not have a great effect on the outcome of the models parameters. This example illustrates the importance of the choice of distribution when analysing discrete data.

Summary

This chapter has demonstrated the use of functions for diagnostic analysis featured in the `discrete.diag` `R` library. Goodness-of-fit, model comparison and outlier detection methods help inform the choice of distribution in analyses and improves the understanding of the fitted model. Outlier detection in counts of cysts in embryonic mouse kidneys provides an example of the application of this library in practice. In the next chapter, distributions for fitting GAMLSS models are developed within the `gamlss` library.

Chapter 7

discrete.reg library

The `discrete.reg` library contains functions to fit discrete regression models within a GAMLSS framework and utilizes the `gamlss R` library by Stasinopoulos and Rigby (2008). The GAMLSS framework requires distributions to be parametrized in terms the location, μ, (often the mean) and scale, σ. The first three sections present alternative parametrizations of the Geometric, Yule and Waring distributions. For each distribution, the probability density, cumulative density, quantile and random generating functions have been re parametrized for the location and (where appropriate) scale parameters as required for the GAMLSS framework. The `gamlss.family` object is necessary for the `gamlss()` fitting procedure and is also defined. The use of these distributions for modelling is demonstrated using counts of stillbirths in New Zealand white rabbits, previously used as an example throughout Chapters 5 and 6. In the final section discrete regression models using distributions found in the `discrete.reg` library are applied to a study on the incidence of ES in paediatric coma patients introduced in Section 1.2.3 in Chapter One.

7.1 Geometric Distribution

The Geometric distribution presented in Section 2.3 has one parameter p, with mean $\dfrac{1-p}{p}$. Setting the mean of the Geometric distribution equal to the location parameter

μ, the *pdf* can be rearranged to give,

$$f_Y(y;\mu) = P(Y = y) = \left(\frac{\mu}{\mu + 1}\right)^y \left(\frac{1}{\mu + 1}\right) . \tag{7.1}$$

The probability density, cumulative density, quantile and random generating functions follow the format given in Section 5.3 and each have the following usage,

```
dGEOM(x, mu = 2, log = FALSE)
pGEOM(q, mu = 2, lower.tail = TRUE, log.p = FALSE)
qGEOM(p, mu = 2, lower.tail = TRUE, log.p = FALSE,
     max.value = 10000)
rGEOM(n, mu = 2)
```

where `x` and `q` are vectors of discrete quantiles, `p` is a vector of probabilities and `n` gives the number of random values to return. The argument `mu` is a vector of positive `mu` values, whilst `lower.tail` and `log.p` are both logical arguments. If the `lower.tail` argument is set to the default value `TRUE` probabilities are $P[Y \leq y]$, otherwise $P[Y > y]$. For the argument `log.p` if this is `TRUE` the probabilities, p, are given as $\log(p)$. A constant argument `max.value` generates a sequence of values for the cumulative distribution function.

The probability density, cumulative density, quantile and random generating functions have the same applications and result in the similar output as those previously described in Section 5.3. These functions are required as they are utilized in the `gamlss.family` object. The `GEOM` function provides the information required by `gamlss` for fitting the Geometric distribution. This function has usage,

```
GEOM(mu.link = "log")
```

where `mu.link` defines the link to be used for the `mu` parameter, with `"log"` link as the default. The code for this function is given in Listing 6.1 and uses functions provided by the gamlss library as a template on which to base this function.

```
GEOM<-function (mu.link = "log")
{
mstats <- checklink("mu.link", "Geometric",
                    substitute(mu.link),
                    c("log", "probit", "cloglog",
 6                        "cauchit", "log", "own"))
 7  structure(list(family = c("GEOM", "Geometric"),
 8          parameters = list(mu = TRUE),
 9          nopar = 1,
10          type = "Discrete",
11          mu.link = as.character(substitute(mu.link)),
12          mu.linkfun = mstats$linkfun,
13          mu.linkinv = mstats$linkinv,
14          mu.dr = mstats$mu.eta,
15          dldm = function(y, mu){
16            dldm <- (y - mu)/(mu + (mu^2))
17            dldm
18          },
19          d2ldm2 = function(mu){
20            d2ldm2 <- -1/(mu+(mu^2))
21            d2ldm2
22          },
23          G.dev.incr = function(y, mu, ...) -2 *
24                               dGEOM(y, mu, log = TRUE),
25          rqres = expression(rqres(pfun = "pGEOM",
26                              type = "Discrete",
27                              ymin = 0, y = y, mu = mu)),
28          mu.initial = expression(mu <- rep(mean(y),
29                                        length(y))),
30          mu.valid = function(mu) all(mu > 0) ,
31          y.valid = function(y) all(y >=0)),
32          class = c("gamlss.family", "family"))
33  }
```

Listing 7.1: Geometric GAMLSS family distribution function.

R code for gamlss.family objects follow a template which provide certain information required for fitting in gamlss. The gamlss.family distribution functions have three fields: i) the definition of the link functions, ii) the information needed for fitting the distributions and iii) the class definition (Stasinopoulos and Rigby, 2008). The gamlss.family function code for the Geometric distribution in Listing 7.1 can be explained by these three separate sections.

i) Definition of the link function

The mstats object found in lines 3-6 of the GEOM provides the definition of the link function for the mu parameter. The which.link argument specifies which parameter the link is for and the which.dist argument determines the current distribution, in this case "Geometric". The link is specified in the link argument and link.List gives a list of the possible links for the specific parameter. In the case of the Geometric distribution the parameter mu is limited to values greater than zero and hence a log link is used to restrict the mu parameter values to positive values.

ii) Fitting information

In this section information needed in the fitting procedure is specified, including the family name of the distribution, which parameters will be fitted (in this case only the mu parameter) and the number of parameters. The type argument determines the type of distribution, i.e. discrete. The mu.link, mu.linkfun, mu.linkinv and mu.dr objects give details of the mu link detailed in the mstats object.

The key aspect of this function is the specification of the first and expected second derivatives of the log likelihood function. The log-likelihood, ℓ, of the Geometric distribution is,

$$\ell(\mu) = y \log\left(\frac{\mu}{\mu+1}\right) + \log\left(\frac{1}{\mu+1}\right) \ . \tag{7.2}$$

Expressions for the derivatives can be calculated analytically using Mathematica. The first derivative of the likelihood, ℓ, with respect to the location parameter μ is,

$$\frac{\partial \ell}{\partial \mu} = \frac{y-\mu}{\mu+\mu^2} \ .$$

This derivative is given in lines 15-17 of the code as the object dldm. Also needed is the expected second derivative of the likelihood with respect to μ,

$$\mathrm{E}\left[\frac{\partial^2 \ell}{\partial \mu^2}\right] = -\frac{1}{\mu+\mu^2} \ .$$

This derivative can be found in lines 19-22 as the d2ldm2 object.

Also found in this list is the global deviance G.dev.incr which utilizes the dGEOM function in its calculation. Expressions for the initial starting values of the parameters are given in mu.initial whilst the range of values for the parameters and the response variable are given in mu.valid and y.valid.

iii) Class

In the resulting function each family is defined as a `gamlss.family` object and is used to define the family in the `gamlss()` fit.

GAMLSS regression models can be fitted for distributions using the `gamlss` fitting procedure. The `gamlss` function has usage,

```
gamlss(formula, sigma.formula = ~1, nu.formula = ~1,
       tau.formula = ~1, family = NO(), data, ... )
```

where `formula` is a formula object with the equation for the model, with the response and model terms separated using a '∼'. The arguments `sigma.formula`, `nu.formula` and `tau.formula` can optionally be used to specify models for the `sigma`, `nu` and `tau` parameters. The GAMLSS distribution to be fitted is specified in `family` which must be a `gamlss.family` object. The `data` argument specifies a data frame containing the variables occurring in the model formula. More details on other arguments of the the `gamlss` function and other functions in the `gamlss` libraries can be found in the GAMLSS `R` manual (Stasinopoulos and Rigby, 2008). Fitting of the Geometric distribution in the GAMLSS framework using the `gamlss.family` object, `GEOM`, to numbers of stillbirths in litters of New Zealand white rabbits can be illustrated with the following `R` commands:

```
> mod <- gamlss(rabbits~1, family=GEOM)
GAMLSS-RS iteration 1: Global Deviance = 731.6
> summary(mod)
*******************************************************************
Family:  c("GEOM", "Geometric")

Call:  gamlss(formula = rabbits ~ 1, family = GEOM)

Fitting method: RS()

-------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
-7.761e-01   4.442e-02  -1.747e+01   3.084e-51

-------------------------------------------------------------------
No. of observations in the fit:  402
```

```
Degrees of Freedom for the fit:  1
      Residual Deg. of Freedom:  401
                      at cycle:  1

Global Deviance:     731.6
            AIC:     733.6
            SBC:     737.5965
********************************************************************
> mod$mu.fv[1]
       1
0.460199
> histDist(rabbits, family=GEOM)
```

The generic function `summary` produces a summary of the results of `gamlss` models which have class `"gamlss"`. The fitted Geometric model for the number of stillbirths (Y) is given by $Y \sim \text{Geometric}(\hat{\mu})$. The output estimates the coefficient of the μ function as -0.078 and we can therefore estimate the mean parameter as $\hat{\mu} = \exp(-0.078) = 0.46$. This value can also be extracted from the fitted values of the model, using the command `mod$mu.fv[1]`. Plot a) in Figure 7.1 shows the fitted Geometric distribution created using the `histDist` command, shown above. This distribution is not a good fit to the data as it underestimates the proportion of zeros and overestimates the probability of one or two stillbirths.

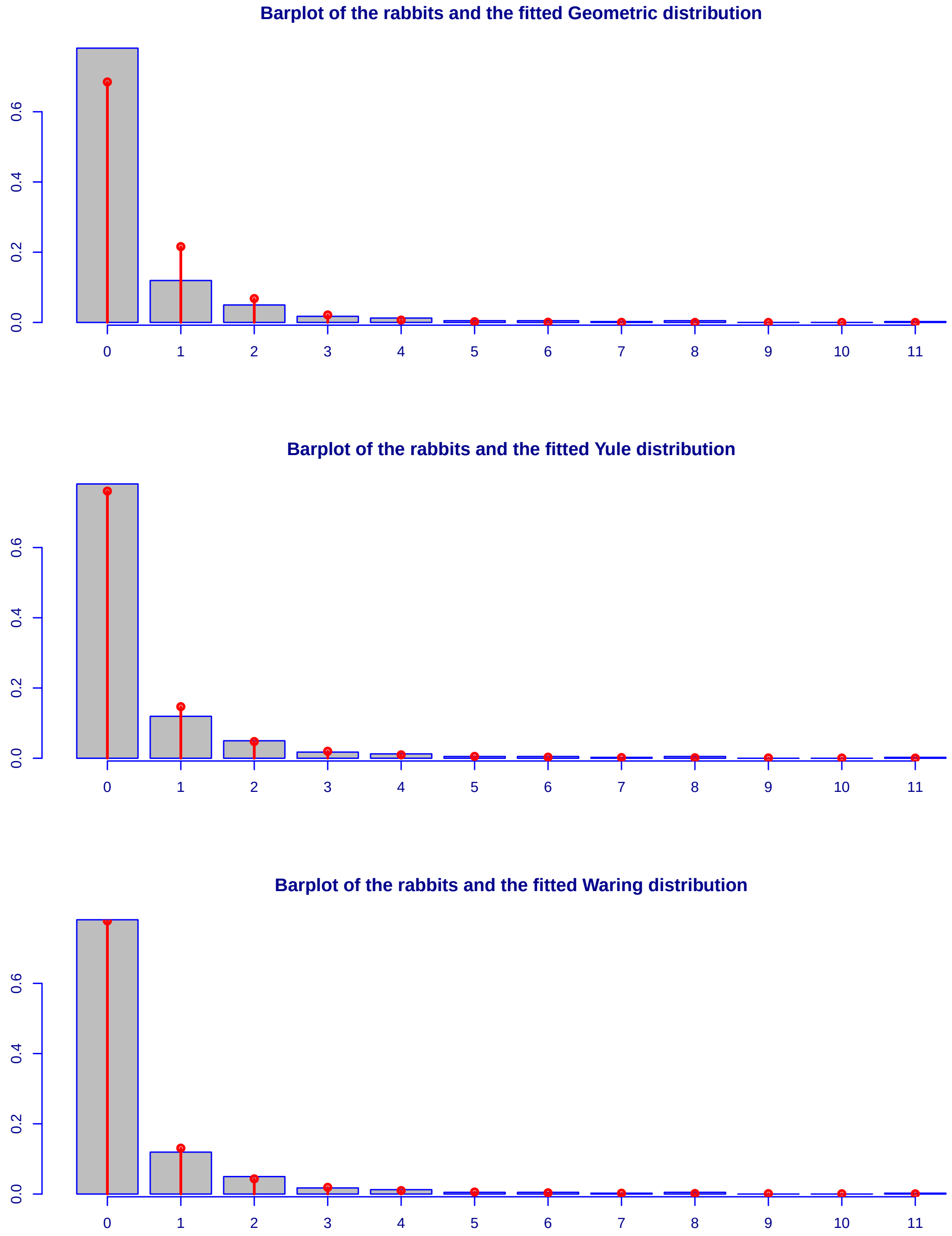
7.2 Yule Distribution

Initially presented in Section 2.16 of Chapter 2, the Yule distribution has one parameter, λ with *pdf*,

$$f_Y(y;\lambda) = \frac{B\left(\lambda+1, y+1\right)}{B\left(\lambda, 1\right)} \ . \tag{7.3}$$

This distribution can be reparameterized in the GAMLSS framework with location parameter μ equal to the mean, given by $\mu = \dfrac{1}{\lambda - 1}$. By substituting $\lambda = \dfrac{\mu+1}{\mu}$, into the *pdf* of the Yule distribution it then becomes,

$$f_Y(y;\mu) = P(Y=y) = \frac{B\left(\frac{2\mu+1}{\mu}, y+1\right)}{B\left(\frac{\mu+1}{\mu}\right)} \ , \tag{7.4}$$



Barplot of the rabbits and the fitted Geometric distribution

Barplot of the rabbits and the fitted Yule distribution

Barplot of the rabbits and the fitted Waring distribution

Figure 7.1: Numbers of stillbirths in New Zealand White rabbits with fitted a) Geometric b) Yule and c) Waring distributions respectively

As for the Geometric distribution there are five functions for the Yule distribution: the probability density, cumulative density, quantile and random generating functions alongside a distribution function in the form of a `gamlss.family` object. The `pdqr` functions for the Yule distribution with parameter `mu` have the following usage,

```
dYUL(x, mu = 2, log.p = FALSE)
pYUL(q, mu = 2, lower.tail = TRUE, log.p = FALSE)
qYUL(p, mu = 2, lower.tail = TRUE, log.p = FALSE,
 max.value = 10000)
rYUL(n, mu = 2)
```

where the arguments of these functions are the same as those given in the previous section for the Geometric distribution. The `YUL` `gamlss.family` function has usage,

```
YUL(mu.link = "log")
```

with argument `"mu.link"` specifying the link of the `mu` parameter.

```
YUL<-function (mu.link = "log")
{
    mstats <- checklink(which.link="mu.link",
    which.dist="Yule", link=substitute(mu.link),
     link.List="log")

    structure(list(family = c("YUL", "Yule"),
        parameters = list(mu = TRUE),
        nopar = 1,
        type = "Discrete",
        mu.link = as.character(substitute(mu.link)),
        mu.linkfun = mstats$linkfun,
        mu.linkinv = mstats$linkinv,
        mu.dr = mstats$mu.eta,
        dldm = function(y, mu){
         lambda <- (mu+1)/mu
         dldm <- (digamma(lambda+1) - digamma(lambda+y+2)
                 +(1/lambda))*(-1/(mu^2))
         dldm
        },
        d2ldm2 = function(y, mu){
         d2ldm2 <- 1/(mu*(mu-1))
         d2ldm2
        },
        G.dev.incr = function(y, mu, ...)
            -2 * dYUL(y, mu = mu, log = TRUE),
27          rqres = expression(rqres(pfun = "pYUL",
28             type = "Discrete", ymin = 0, y = y, mu = mu)),
29          mu.initial = expression(mu <- rep(mean(y),
30                                     length(y))),
31          mu.valid = function(mu) all(mu > 0) ,
32          y.valid = function(y) all(y >=0)),
33          class = c("gamlss.family", "family"))
34 }
```

Listing 7.2: Yule Family distribution function.

R code for the Yule gamlss.family distribution function, YUL is given in Listing 7.2. This function again follows the template provided by the gamlss library for gamlss.family objects and since this function has the same parameter as the GEOM function there are many similarities between these two functions. The YUL function also uses a log link for the parameter mu in the mstats object in lines 3-5. The family argument now specifies that a Yule distribution is to be fitted.

The expressions for the first and expected second derivatives can be found by making use of the log-likelihood for the λ parameterization of the Yule distribution in Equation 7.3 given by,

$$\ell(\lambda) = \log\Gamma(\lambda+1) + \log\Gamma(y+1) - \log\Gamma(\lambda+y+2) + \log\lambda \; . \tag{7.5}$$

The first derivative of the log-likelihood ℓ, with respect to the location parameter μ can be derived by using the chain rule, as follows,

$$\begin{aligned} \frac{\partial\ell}{\partial\mu} &= \frac{\partial\ell}{\partial\lambda} \times \frac{\partial\lambda}{\partial\mu} \\ &= \left(\psi(\lambda+1) - \psi(\lambda+y+2) + \tfrac{1}{\lambda}\right)\left(-\tfrac{1}{\mu^2}\right) \end{aligned} \; , \tag{7.6}$$

where $\psi^{(n)}(z)$ gives the n^{th} derivative of the digamma function. Also needed is the expected second derivative of the log-likelihood with respect to μ,

$$\mathrm{E}\left[\frac{\partial^2\ell}{\partial\mu^2}\right] = \frac{1}{\mu(\mu-1)} \; . \tag{7.7}$$

These derivative can be found as the objects dldm and d2ldm2 in lines 16-25 of the

YUL function code. In, the gamlss.family template the commands to calculate the global deviance for the model G.dev.incr and the quantile residuals rqres use the dYUL and pYUL functions in their computation.

We can also fit the Yule distribution as a GAMLSS model to the numbers of stillbirths in litters of New Zealand white rabbits using the following R code,

```
> mod <- gamlss(rabbits~1, family=YUL)
GAMLSS-RS iteration 1: Global Deviance = 677.5246
GAMLSS-RS iteration 2: Global Deviance = 677.5239
> summary(mod)
GAMLSS-RS iteration 1: Global Deviance = 677.5246
GAMLSS-RS iteration 2: Global Deviance = 677.5239
*******************************************************************
Family:  c("YUL", "Yule")

Call:  gamlss(formula = rabbits ~ 1, family = YUL)

Fitting method: RS()

-------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
-7.898e-01   6.737e-02  -1.172e+01   1.708e-27

-------------------------------------------------------------------
No. of observations in the fit:  402
Degrees of Freedom for the fit:  1
      Residual Deg. of Freedom:  401
                      at cycle:  2

Global Deviance:     677.5239
            AIC:     679.5239
            SBC:     683.5204
*******************************************************************
> mod$mu.fv[1]
        1
0.4539463
> histDist(rabbits, family=YUL)
```

The fitted Yule distribution for the number of stillbirths (Y) in litters of New Zealand white rabbits is given by $Y \sim \text{YUL}(\hat{\mu})$ where $\hat{\mu} = \exp(-0.79) = 0.45$. In Figure 7.1 the second plot shows the fitted Yule distribution to the numbers of stillbirths, again produced using the histDist function. The plot shows that a Yule distribution provides a better fit to the data in comparison to the Geometric distribution and is supported by the lower BIC value of 683.52 in contrast to a value of 737.60 for the Geometric model.

7.3 Waring Distribution

The Waring distribution presented in Section 2.3.6 of Chapter 2 has two parameters n and b with *pdf*,

$$f_Y(y; n, b) = P(Y = y) = \frac{B(n + y, b + 1)}{B(n, b)} \ , \tag{7.8}$$

where $b \geq 0$ and $n \geq 0$ (Wimmer and Altmann, 1999, P. 643). The mean of this distribution is

$$\mu = \frac{n}{b-1} \ , \tag{7.9}$$

with variance

$$\sigma^2 = \frac{\mathrm{B}(1 + n, 1 + b)_P F_Q(\{2, 2, 1 + n\}, \{1, 2 + b + n\}, 1)}{\mathrm{B}(n, b)} - \frac{n^2}{(b-1)^2} \ . \tag{7.10}$$

Since the expression for the variance contains a hypergeometric function, when attempting to solve these as simultaneous equations the solution is intractable for expressions of b and n. If we let the location parameter $\mu = \dfrac{n}{b-1}$ and set $\sigma = \dfrac{1}{b-1}$, we can reparameterize the Waring distribution in Equation 7.8 where $b = 1 + \dfrac{1}{\sigma}$ and $n = \mu\,(b - 1)$, giving the following *pdf*,

$$f_Y(y; \mu, \sigma) = P(Y = y) = \frac{(1-\sigma)\ \Gamma\left(y + \frac{\mu}{\sigma}\right)\ \Gamma\left(\frac{\mu+\sigma+1}{\sigma}\right)}{\sigma\,\Gamma\left(y + \frac{\mu+1}{\sigma} + 2\right)\ \Gamma\left(\frac{\mu}{\sigma}\right)} \ . \tag{7.11}$$

where $\mu > 0$ and $\sigma > 0$. The pdqr functions for the Waring distribution with parameters μ and σ have usage,

```
dWAR(y, mu = 2, sigma = 2, log.p = FALSE)
pWAR(q, mu = 2, sigma = 2, lower.tail = TRUE, log.p = FALSE)
qWAR(p, mu = 2, sigma = 2, lower.tail = TRUE, log.p = FALSE,
     max.value = 10000)
rWAR(n, mu = 2, sigma = 2)
```

where mu and sigma are vectors of positive mu and sigma parameters. The arguments y, q, p, n, lower.tail, log.p and max.value are as described for the Yule pdqr functions in the previous section. The WAR function defines the Waring distribution as a gamlss.family object and has usage,

```
WAR(mu.link = "log", sigma.link = "log")
```

with two arguments mu.link and sigma.link for the links of the parameters mu and sigma. R code for this function is given in Listing 7.3.

```
WAR <- function (mu.link = "log", sigma.link = "log")
{
    mstats <- checklink("mu.link", "WAR",
                        substitute(mu.link), "log")
    dstats <- checklink("sigma.link", "WAR",
                        substitute(sigma.link), "log")
    structure(list(family = c("WAR", "Waring"),
              parameters = list(mu = TRUE, sigma = TRUE),
              nopar = 2, type = "Discrete",
              mu.link = as.character(substitute(mu.link)),
        sigma.link = as.character(substitute(sigma.link)),
        mu.linkfun = mstats$linkfun,
        sigma.linkfun = dstats$linkfun,
        mu.linkinv = mstats$linkinv,
        sigma.linkinv = dstats$linkinv,
        mu.dr = mstats$mu.eta, sigma.dr = dstats$mu.eta,
        dldm = function(y, mu, sigma) {
            dldm <- (1/sigma) * (digamma((mu/sigma) + y)
                    - digamma(y + (mu + 1)/sigma) + 2)
                    - digamma(mu/sigma)
                    + digamma((mu + sigma + 1)/sigma))
            dldm
        }, d2ldm2 = function(y, mu, sigma) {
            dldm <- (1/sigma) * (digamma((mu/sigma) + y)
                    - digamma(y + ((mu + 1)/sigma) + 2)
                    - digamma(mu/sigma)
                    + digamma((mu + sigma + 1)/sigma))
            d2ldm2 <- -dldm * dldm
            d2ldm2
        }, dldd = function(y, mu, sigma) {
            dldd <- (1/sigma^2) * (-1 + (1/(sigma + 1))
                    - mu * harmonic(y + (mu/sigma) - 1)
                    + (mu + 1) * harmonic(y +
                    ((mu + 1)/sigma) + 1)
                    - (mu + 1) * harmonic((mu + 1)/sigma)
                    + mu * (-digamma(1)
                    + digamma(mu/sigma)))
            dldd
        }, d2ldd2 = function(y, mu, sigma) {
            dldd <- (1/sigma^2) * (-1 + (1/(sigma + 1))
                    - mu * harmonic(y + (mu/sigma) - 1)
                    + (mu + 1) * harmonic(y +
                    ((mu + 1)/sigma) + 1)
                    - (1 + mu) * harmonic((mu + 1)/sigma)
                    + mu * (-digamma(1)
                    + digamma(mu/sigma)))
            d2ldd2 <- -dldd * dldd
            d2ldd2
        }, d2ldmdd = function(y, mu, sigma) {
            dldm <- (1/sigma) * (digamma((mu/sigma) + y)
                    - digamma(y + ((mu + 1)/sigma) + 2)
                    - digamma(mu/sigma)
                    + digamma((mu + sigma + 1)/sigma))
            dldd <- (1/sigma^2) * (-1 + (1/(sigma + 1))
                    - mu * harmonic(y + (mu/sigma) - 1)
                    + (mu + 1) * harmonic(y +
                    ((mu + 1)/sigma) + 1)
                    - (1 + mu) * harmonic((mu + 1)/sigma)
                    + mu * (-digamma(1)
                    + digamma(mu/sigma)))
            d2ldmdd <- -dldm * dldd
            d2ldmdd
        }, G.dev.incr = function(y, mu, sigma, ...) -2 *
               dWAR(y, mu, sigma, log = TRUE),
            rqres = expression(rqres(pfun = "pWAR",
               type = "Discrete", ymin = 0, y = y,
               mu = mu, sigma = sigma)),
            mu.initial = expression(mu <-
               (y + mean(y))/2),
            sigma.initial = expression(sigma <-
               rep(2, length(y))),
            mu.valid = function(mu) all(mu > 0),
            sigma.valid = function(sigma) all(sigma > 0),
            y.valid = function(y) all(y >= 0)),
            class = c("gamlss.family", "family"))
}
```

Listing 7.3: Waring family distribution function.

The Waring distribution's two parameters, mu and sigma are reflected in this template of this function. In the first section of the code, there is an additional object dstats which specifies the link of the sigma parameter in the same way that mstats specifies the link of the mu parameter. For this distribution "log" links are once again used as both parameters are restricted to positive values.

Within the second section of the distribution, additional information is needed on the `sigma` parameter which can be extracted from the `dstats` object detailing the link function. The log-likelihood for the Waring distribution given in Equation 7.11 is,

$$\begin{aligned} \ell(\mu,\sigma) = \; & \log\left(1+\frac{1}{\sigma}\right) + \log\Gamma\left(y+\frac{\mu}{\sigma}\right) - \log\Gamma\left(y+\frac{\mu+1}{\sigma}+2\right) \\ & -\log\Gamma\left(\frac{\mu}{\sigma}\right) + \log\Gamma\left(\frac{\mu+\sigma+1}{\sigma}\right) \end{aligned} \quad . \tag{7.12}$$

The first derivative of the log-likelihood of the Waring distribution ℓ, with respect to μ:

$$\frac{\partial\ell}{\partial\mu} = \frac{1}{\sigma}\left(\psi\left(y+\frac{\mu}{\sigma}\right) - \psi\left(y+\frac{\mu+1}{\sigma}+2\right) - \psi\left(\frac{\mu}{\sigma}\right) + \psi\left(\frac{\mu+\sigma+1}{\sigma}\right)\right) \; , \tag{7.13}$$

where H_n gives the n^{th} harmonic number. This derivative is specified in lines 17-23 of the `WAR` function as the object `dldm`. For the second parameter, σ, the first derivative of the log-likelihood of the Waring distribution with respect to σ is:

$$\begin{aligned} \frac{\partial\ell}{\partial\sigma} = \; & \frac{1}{\sigma^2}\,(\,\frac{1}{\sigma+1} - \mu\,H\left(y+\frac{\mu}{\sigma}-1\right) + (1-\mu)\,H\left(y+\frac{\mu+1}{\sigma}+1\right) \\ & -(1+\mu)\,H\left(\frac{\mu+1}{\sigma}\right) + \mu\,\left(\gamma+\psi\left(\frac{\mu}{\sigma}\right)\right) - 1\,)\;, \end{aligned} \tag{7.14}$$

where γ is Euler's constant with numerical value ≈ 0.577216 (Johnson et al., 2005, P.9). The derivative for the `sigma` parameter `dldd` and `d2ldd2` are computed in lines 30-39. Expressions for the expected second derivatives can be replaced for this distribution by the negative squared first derivatives, shown in lines 39-49 for the expected second derivative of the log-likelihood with respect to μ (`d2ldm2`), lines 39-49 for the expected second derivative of the log-likelihood with respect to σ (`d2ldd2`) and lines `49-63` for the expected cross derivative of the log-likelihood with respect to μ and σ (`d2ldmdd`).

In the final section of the template code, the `G.dev.incr` object now utilizes the probability density function specified for the Waring distribution in calculating the global deviance of the model and the cumulative density function is used to calculate the quantile residuals in the object `rqres`. Initial values and valid parameter bounds are given for `sigma` in lines 28 and 40.

The Waring distribution can also be fitted as a GAMLSS model to the numbers of stillbirths in litters of New Zealand white rabbits as follows:

```
> mod <- gamlss(rabbits~1, family=WAR)
GAMLSS-RS iteration 1: Global Deviance = 680.5955
GAMLSS-RS iteration 2: Global Deviance = 679.5804
                        .
                        .
                        .
GAMLSS-RS iteration 12: Global Deviance = 675.889
GAMLSS-RS iteration 13: Global Deviance = 675.8882
> summary(mod)
*******************************************************************
Family:  c("WAR", "Waring")

Call:  gamlss(formula = rabbits ~ 1, family = WAR)

Fitting method: RS()

-------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
-7.179e-01   1.745e-01  -4.113e+00   4.734e-05

-------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
  Estimate  Std. Error     t value    Pr(>|t|)
   -0.3459      0.3803     -0.9096      0.3636

-------------------------------------------------------------------
No. of observations in the fit:  402
Degrees of Freedom for the fit:  2
      Residual Deg. of Freedom:  400
                      at cycle:  13

Global Deviance:     675.8882
            AIC:     679.8882
            SBC:     687.8811
*******************************************************************
> mod\$mu.fv[1]
       1
0.4877834
> mod\$sigma.fv[1]
        1
0.7075593
```

```
> histDist(rabbits, family=WAR)
```

The summary of the output shows that the fitted Waring distribution for the number of stillbirths (Y)in litters of New Zealand white rabbits is given by $Y \sim \mathrm{WAR}(\hat{\mu}, \hat{\sigma})$ where $\hat{\mu} = \exp(-0.72) = 0.49$ and $\hat{\sigma} = \exp(-0.35) = 0.71$. The final plot in Figure 7.1 shows the fitted Waring distribution against the numbers of stillbirths, created using the above `histDist` command. Comparing this plot and the BIC value of 687.88 to the fitted distributions and BIC's of the Geometric and Yule distributions, we can conclude that of the three models fitted the Waring distribution provides the best fit to the data.

7.4 Validation of the functions

In this library, the three distributions each have `d`, `p`, `q` and `r` distribution functions and a GAMLSS family function. The distribution functions were tested for a range of parameter values in the same way as the distribution functions in the `Altmann` library. GAMLSS family functions can were also tested using a randomly generated sample from the selected distribution, with known parameter estimates. A GAMLSS model was then fitted to the sample to confirm that the model produced approximately the same parameter estimates. Correct convergence was checked inside the GAMLSS models using `i.control = glim.control(glm.trace=T)` , where at each iteration it was confirmed that the deviance was reducing. Residual analysis of these GAMLSS models also established that the models were performing adequately.

The results of the GAMLSS models can also be compared to fitted values from published datasets in a similar way to those in the `Altmann` library. The datasets used were the counts of stillbirths in New Zealand white rabbits (Morgan et al., 2007) and haemocytometer counts of yeast cells (Plunkett and Jain, 1975) datasets presented in the `Altmann` library. The results from the GAMLSS models for each distribution were compared to the results using maximum likelihood estimation in the `Altmann` library and also the rapid estimated produced by the Altmann fitter software program

(Altmann, 1997), to check for consistency in the parameter estimates and fitted values.

7.5 Application to Electroencephalographic Seizures in coma patients

Data from a study on the incidence of electroencephalographic seizures (ES) in comatose patients is presented in Section 1.3 of Chapter One. The aim of this study is to use continuous EEG monitoring to document the incidence of ES in children unconscious from a variety of aetiologies. Regression models are used to investigate potential predictors of incidence of ES, exploratory variables are: centre (UK, UK neonate or Kenya), aetiology (with levels: Encephalitis, Head Injury, Hypoxic-ischaemic, Maleria, Meningitis, Reye's and other), EEG classification (with levels: Burst suppression, Diffuse slowing, diffuse slowing with some fast activity, Isoelectric, Low amplitude, Normal), the presence of clinical seizures at any time (yes/no) and the following variables on admission: Pediatric Index of Mortality (PIM) score, Adelaide Coma Scale (ACS) score, temperature, the use of drugs benzodiazepine (yes/no) and phenytoin/phenobarbitone (yes/no). The number of ES is adjusted by the duration of monitoring, which is included in models as an offset. ES may be clinically subtle or only manifest electroencephalographically and differ from clinical seizures which manifest physically.

The `gamlss` library can be used to fit regression models using the GAMLSS framework for a range of discrete distributions. Regression models for discrete outcomes predict the mean number of ES and coefficients for explanatory variables yield rate ratios, which estimate the rate of change in the mean number of ES.

A stepwise model analysis was performed to select a predictor variables by minimizing the BIC. In `gamlss`, stepwise model selection can be performed using the functions `stepGAIC` and `stepGAICAll.B`. The function `step.GAIC` is used to build models for individual parameters of the distribution of the response variable, while the function `stepGAICAll.B` builds a model for all the parameters. For each distribution, a null model was fitted which includes no covariates. This is illustrated for a Geometric distribution, GEOM, as follows,

```
> seiz.mod.geom<-gamlss(NSEIZEEG~1, offset=DURNMON, data=seizures,
GAMLSS-RS iteration 1: Global Deviance = 1412.330
```

The step function stepGAIC and stepGAICAll.B have similar usage,

```
stepGAIC(object, scope, direction = c("both", "backward",
              "forward"), k = 2, ... )
stepGAICAll.B(object, scope, direction = c("both",
              "backward", "forward"), k = 2, ... )
```

where the scope argument defines the range of models examined, with lower detailing terms always included in the model and upper the most complicated model that the procedure would consider. The penalization parameter a can be specified as k=log(n) to give the BIC, where n is the number of independent observations. The argument direction determines the mode of stepwise search, with "both" performing forward stepwise model selection. For the above Geometric model a stepwise model selection can be implemented as follows,

```
> geom.mod <- stepGAICAll.B(seiz.mod.geom, direction="both",
                            k=log(184), scope=list(lower=~1,
                            upper=~as.factor(UKENUNEO)+
                            as.factor(AETIOLOGY)+PIM+ACSOA+TOA+
                            as.factor(EEGOA)+SEIZURE+PXOTHER+DIAZPRE))
Start:  AIC= 1417.55
 NSEIZEEG ~ 1
                       Df    AIC
+ as.factor(AETIOLOGY)  6 1257.3
+ as.factor(EEGOA)      5 1341.9
+ SEIZURE               1 1349.1
+ as.factor(UKENUNEO)   2 1354.8
+ TOA                   1 1377.3
+ PIM                   1 1391.5
<none>                    1417.5
+ DIAZPRE               1 1421.8
+ ACSOA                  1 1422.2
+ PXOTHER               1 1422.8

Step:  AIC= 1257.28
 NSEIZEEG ~ as.factor(AETIOLOGY)

                       Df     AIC
+ SEIZURE               1  884.53
+ DIAZPRE               1 1162.84
+ TOA                   1 1211.65
+ as.factor(EEGOA)      5 1215.80
+ PIM                   1 1250.09
<none>                    1257.28
+ ACSOA                 1 1262.38
+ PXOTHER               1 1262.48
+ as.factor(UKENUNEO)   2 1266.36
- as.factor(AETIOLOGY)  6 1417.55

Step:  AIC= 884.53
 NSEIZEEG ~ as.factor(AETIOLOGY) + SEIZURE

                       Df     AIC
+ TOA                   1  870.88
+ DIAZPRE               1  883.52
<none>                     884.53
+ PIM                   1  886.99
+ as.factor(EEGOA)      5  887.50
+ ACSOA                 1  887.86
+ as.factor(UKENUNEO)   2  888.17
+ PXOTHER               1  889.74
- SEIZURE               1 1257.28
- as.factor(AETIOLOGY)  6 1349.11

Step:  AIC= 870.88
 NSEIZEEG ~ as.factor(AETIOLOGY) + SEIZURE + TOA

                       Df     AIC
+ PIM                   1  866.15
<none>                     870.88
+ ACSOA                 1  870.92
+ DIAZPRE               1  871.07
+ as.factor(EEGOA)      5  873.91
+ PXOTHER               1  875.24
+ as.factor(UKENUNEO)   2  875.61
- TOA                   1  884.53
- SEIZURE               1 1211.65
- as.factor(AETIOLOGY)  6 1229.10

Step:  AIC= 866.15
 NSEIZEEG ~ as.factor(AETIOLOGY) + SEIZURE + TOA + PIM

                       Df     AIC
<none>                     866.15
+ DIAZPRE               1  866.78
+ ACSOA                 1  868.29
- PIM                   1  870.88
+ PXOTHER               1  871.20
+ as.factor(UKENUNEO)   2  872.19
+ as.factor(EEGOA)      5  874.42
```

```
- TOA                    1  886.99
- as.factor(AETIOLOGY)   6 1151.50
- SEIZURE                1 1184.84
```

Beginning with the null model (containing no covariates) at each step of the process each variable is added to the model in turn and the BIC values compared to the current model in a table. Current variables in a model are also systematically removed, the model BIC values calculated and included in the table to determine if removing any of the current models in a backwards procedure improves the fit of the model. In the output given above, in the table showing the first step of the procedure for a Geometric model, each variable is systematically added to the null model. The addition of the variable AETIOLOGY to the null Geometric model (shown in the first line of the table for the first step) produces the lowest BIC value of the models at 1257.3. The variables SEIZURE, TOA, and PIM and DIAZPRE are then systematically added in the next 3 steps to result in a final selection of the Geometric model containing: AETIOLOGY, SEIZURE, TOA and PIM.

Distribution	**Significant Covariates (in order of addition)**	**BIC**
Negative binomial	SEIZURE, CENTRE	743.17
Sichel	SEIZURE, AETIOLOGY	750.15
Waring	SEIZURE, CENTRE, TOA, DIAZPRE	762.56
Delaporte	SEIZURE, DIAZPRE, TOA, ACSOA	773.26
Zero-inflated negative binomial	SEIZURE, TOA	813.15
Yule	TOA, DIAZPRE, PIM	858.29
Geometric	AETIOLOGY, SEIZURE, TOA, PIM	866.14
Zero-inflted Poisson	DIAZPRE, EEGOA, CENTRE TOA, SEIZURE, PXOTHER	4254.40
Poisson	AETIOLOGY, SEIZURE, EEGOA, TOA, CENTRE, PXOTHER, ACSOA, PIM, DIAZPRE	5316.85

Table 7.1: Summary of discrete regression models resulting from stepwise model selection fitted to incidence of ES dataset.

Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, Sichel, Delaporte, Geometric, Yule and Waring models were all fitted and a comparison of BIC values can be found in Table 7.1. The variable SEIZURE is added into all but one model (the Yule distribution) however the addition of other variables varies by distribution. SEIZURE is also added firstly in the top five fitting models. The negative binomial distribution provides the best fit to the data with a BIC of 743.17 and includes the variables SEIZURE the presence of seizures and CENTRE the location of the site. This model is fitted with the R commands below,

```
> nb.mod<-gamlss(NSEIZEEG~as.factor(CENTRE)+SEIZURE,
               sigma.formula=~as.factor(CENTRE)+SEIZURE,
               data=seizures, offset=DURNMON, family=NBII())
GAMLSS-RS iteration 1: Global Deviance = 710.3076
GAMLSS-RS iteration 2: Global Deviance = 708.0219



GAMLSS-RS iteration 17: Global Deviance = 701.4534
GAMLSS-RS iteration 18: Global Deviance = 701.4529
> summary(nb.mod)
*******************************************************************
Family:  c("NBII", "Negative Binomial type II")

Call:
gamlss(formula = NSEIZEEG ~ as.factor(CENTRE) + SEIZURE,
       sigma.formula = ~as.factor(CENTRE) +  SEIZURE,
       family = NBII(), data = seizures, offset = DURNMON)

Fitting method: RS()

-------------------------------------------------------------------
Mu link function:  log
Mu Coefficients:
                      Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)             1.8708      1.1127    1.681  9.447e-02
as.factor(CENTRE)2     -2.4800      0.5192   -4.776  3.746e-06
as.factor(CENTRE)3     -0.8409      0.7311   -1.150  2.516e-01
SEIZURE                 1.9107      1.1434    1.671  9.648e-02

-------------------------------------------------------------------
Sigma link function:  log
Sigma Coefficients:
                      Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)             6.5221      1.2116    5.383  2.314e-07
as.factor(CENTRE)2     -1.6942      0.6451   -2.626  9.395e-03
as.factor(CENTRE)3     -0.6326      0.8999   -0.703  4.830e-01
SEIZURE                -1.5083      1.2550   -1.202  2.310e-02

-------------------------------------------------------------------
No. of observations in the fit:  184
Degrees of Freedom for the fit:  8
```

```
      Residual Deg. of Freedom:  176
                      at cycle:  18

Global Deviance:     701.4529
            AIC:     717.4529
            SBC:     743.1724
******************************************************************
```

The summary command provides tables of parameter estimates and p-values for the negative binomial model. Discrete regression models yield rate ratios (RRs), which estimate the change in the relative (rather than absolute) mean number of events between the groups. RRs can be expressed in different ways. For example, $RR = 1.25$ indicates that the mean in one group is, on average, 1.25 times higher or, alternatively, that there is a 25% increase in one group compared with the other. On the other hand, e.g. RR=0.83 indicates a 17% decrease in one group compared with the other. In this example, the model predicts the mean number of ES and is adjusted by the duration of monitoring (measured in hours) as an offset.

For the parameter μ, the mean number of ES for patients in Kenya (Centre 2) is 91.63% lower (RR:exp(-2.48)=0.084, 95% CI:(0.03,0.23), $p < 0.001$) than those in the UK Intensive Care Unit (ICU). Similarly, for UK Neonate patients (Centre 3) the mean number of ES is 56.87% lower (RR:exp(-0.8409)=0.43, 95% CI:(0.10, 1.81), $p = 0.025$) than those in the UK ICU. In patients where clinical seizures were present, the mean number of ES was around 5.5 times higher (RR:exp(1.9107)=6.76, 95% CI:(0.72, 63.55), $p = 0.096$) than those who did not have clinical seizures.

In the GAMLSS parameterization of the negative binomial type II distribution the σ parameter represents the dispersion of the distribution. Patients in Kenya have 81.63% lower dispersion (RR: exp(-1.6942)=0.18, 95% CI:(0.05, 0.65), $p = 0.009$) than those in the UK ICU. For UK Neonate patients the dispersion is 46.88% lower (RR: exp(-0.6326)=0.53, 95% CI:(0.09, 3.10), $p = 0.48$) compared to those in the UK ICU. Where actual seizures were present in a patient, the dispersion is 77.87% lower (RR:exp(-1.5083)=0.22, 95% CI:(0.02, 2.59), $p = 0.023$) than those who did not have actual seizures.

The goodness-of-fit of this model can be determined through plots of the randomized quantile residuals and can be obtained for models of class `"gamlss"` using the generic function `plot()`,

```
> plot(nb.mod)
*******************************************************************
         Summary of the Randomized Quantile Residuals
                                mean   =  -0.05770572
                            variance   =  1.054727
                    coef. of skewness  =  0.04187008
                    coef. of kurtosis  =  2.702588
Filliben correlation coefficient       =  0.9981934
*******************************************************************
```

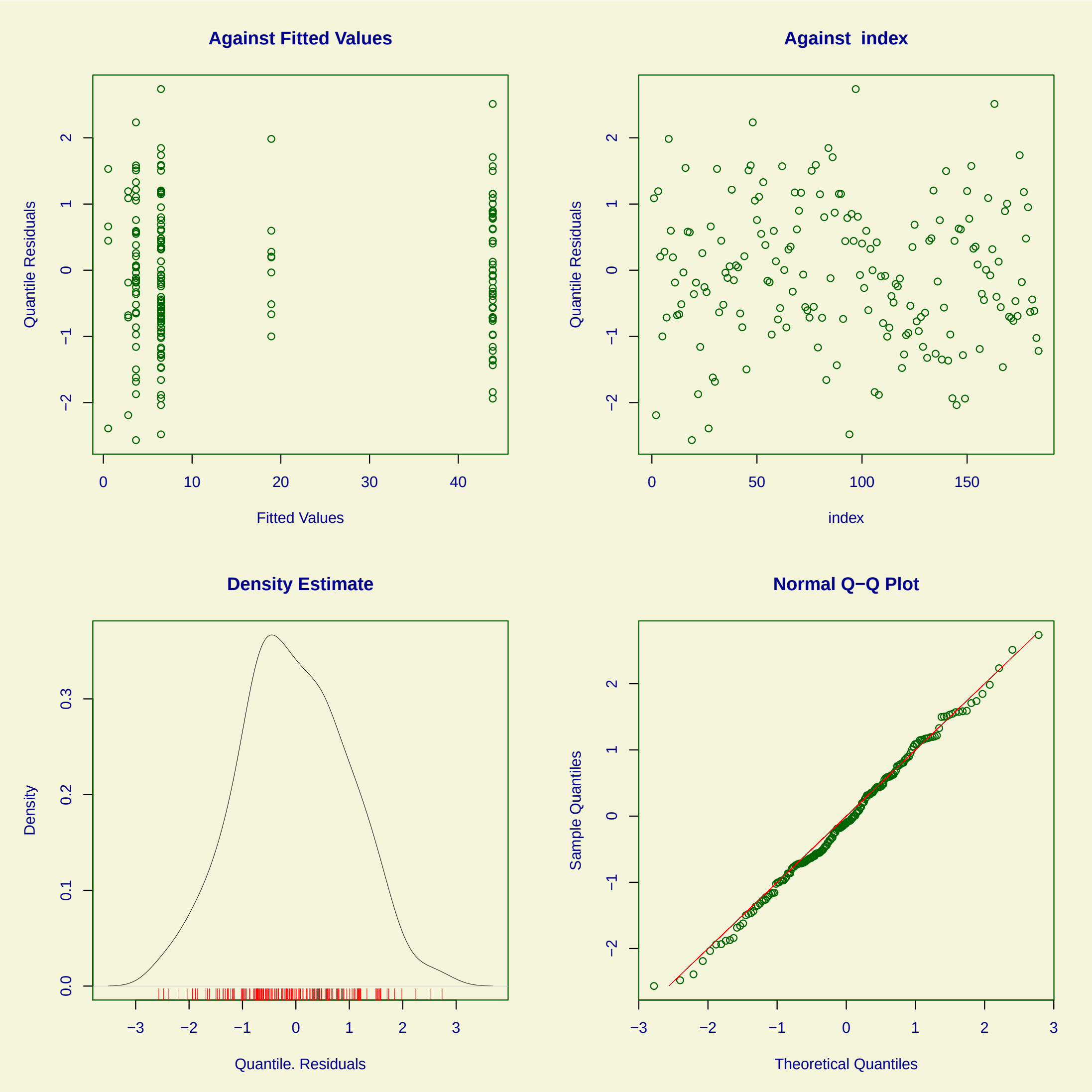


Figure 7.2: Residual plot from the fitted negative binomial model

Figure 7.2 shows plots of the (normalized quantile) residuals: i) against the fitted values ii) against a index iii) a non-parametric kernel density estimate and iv) a normal

Q-Q plot. The residuals follow normal distribution, indicated by the density estimate and normal Q-Q plot and plots of the residuals against the fitted values and index show no signs of non-constant variance or violations of independence. We can therefore determine that the fit of the model is adequate. The `gamlss` library can also implement worm plots of the randomized quantile residuals,

```
> wp(nb.mod)
```

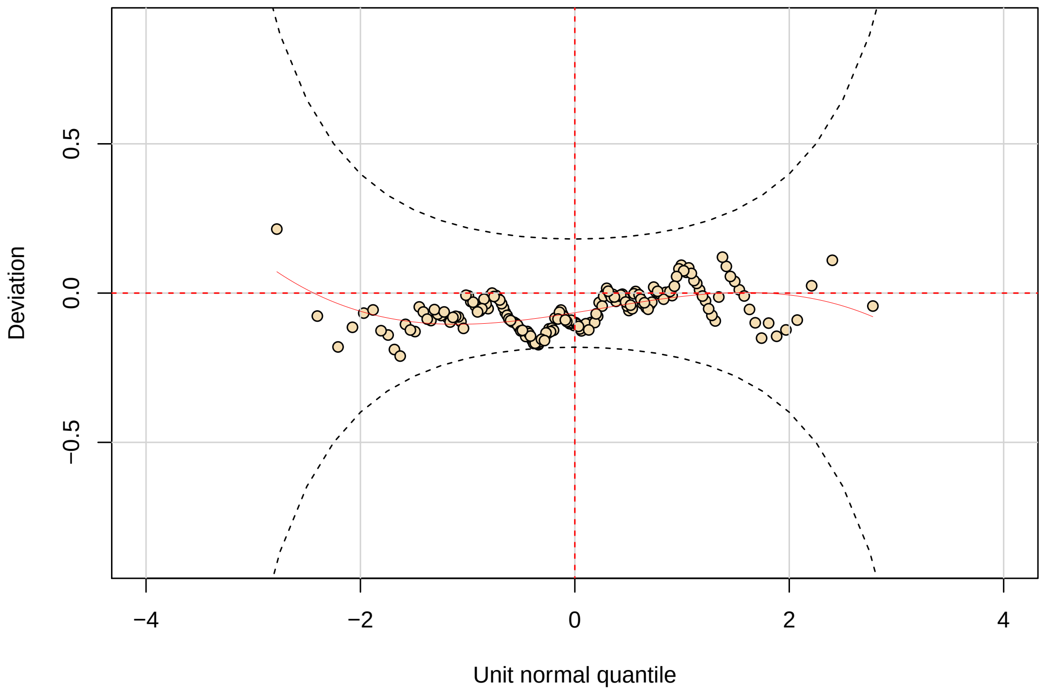


Figure 7.3: Worm plot from the fitted negative binomial model

The worm plot (van Buuren and Frederiks, 2001) is a de-trended normal Q-Q plot of the residuals, and points plotted outside of the (dotted) confidence bands indicate a possible inadequacy in modelling the distribution. The worm plot produced for the negative binomial model in Figure 7.3 supports the conclusion that the model is an adequate fit to the data as there are no points outside of the confidence bands.

Estimated values for the parameters of the negative binomial model can be constructed using the `predict` generic function. This function first requires a data frame containing new values for the explanatory variables used in the model. `R` commands to produce the parameter estimates are as follows,

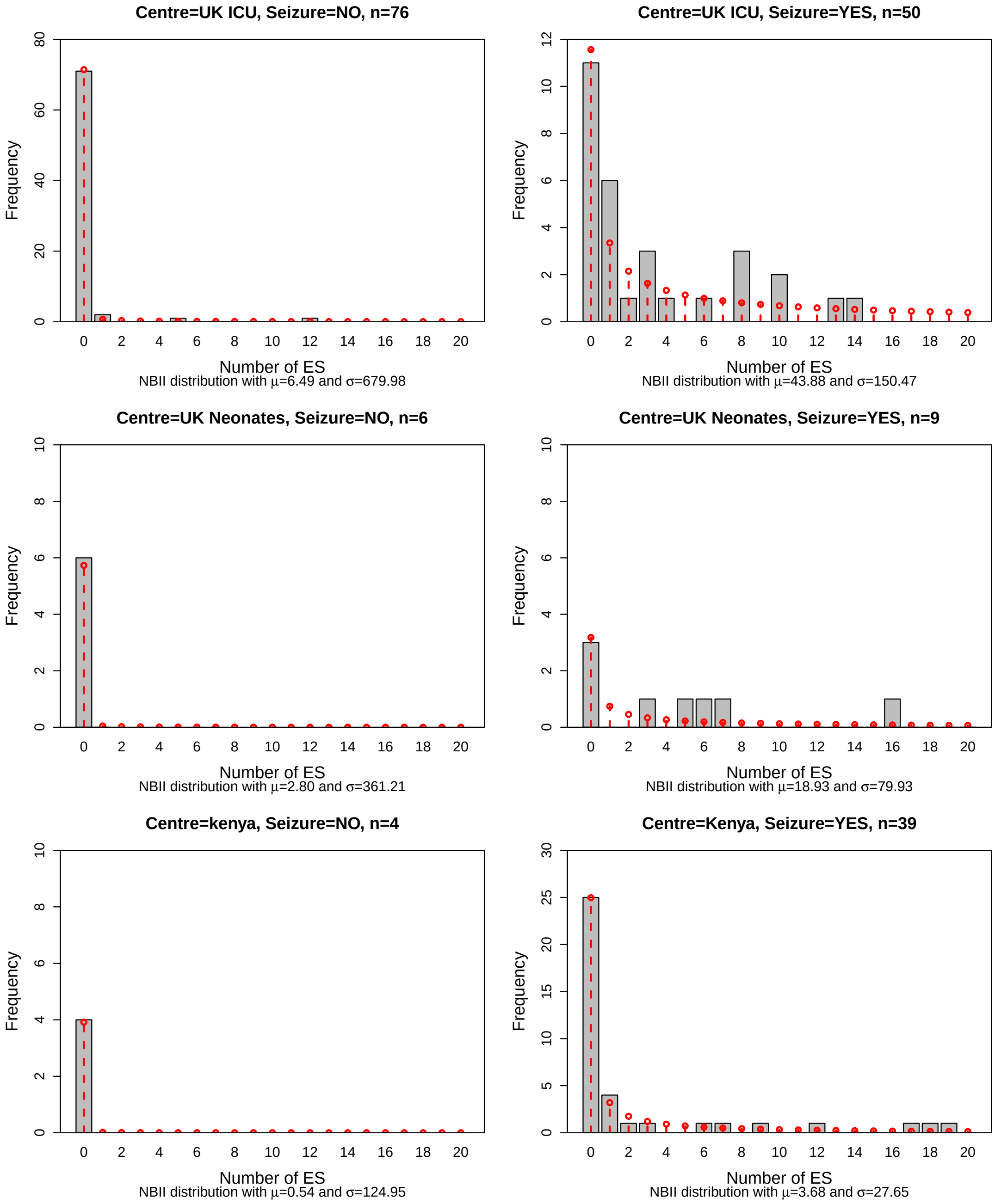


Figure 7.4: Predictions from the fitted negative binomial model across three centres by the presence of seizures

```
> new.seiz<-data.frame(CENTRE=c(1,1,2,2,3,3),
                        SEIZURE=c(0,1,0,1,0,1))
> pred.seiz <- predictAll(nb.mod, newdata=new.seiz)
> pred.seiz
$mu
[1]  6.4935144 43.8831234  0.5438137  3.6750893
[5]  2.8008424 18.9280726

$sigma
[1] 679.97566 150.47312 124.94515  27.64935 361.20718
[6]  79.93223

attr(,"family")
[1] "NBII"
[2] "Negative␣Binomial␣type␣II"
```

The above values of the parameters μ and σ can be used to plot the observed data and fitted negative binomial model for the number of ES. Figure 7.4 plots the predictions from the fitted negative binomial model for the three different site locations in `CENTRE` and where clinical seizures (`SEIZURE`) are present and are not present.

Regression analysis of the incidence of ES in paediatric coma patients has shown a negative binomial model provides the best fit to the data. There are two significant predictors of incidence of ES in this model: centre of study and presence of clinical seizures. This model suggests there is no association between incidence of ES and aetiology, EEG classification, PIM, ACS, temperature, use of drugs benzodiazepine (yes/no) and phenytoin/phenobarbitone on arrival. The centre of study is associated with incidence of ES- patients in UK neonatal units and Kenya have lower mean incidence of ES than those in UK ICU units. The presence of clinical seizures also decreases the incidence of ES in comparison with those who were not affected by clinical seizures. The dispersion parameter of the negative binomial distribution is lower for patients in Kenya and UK Neonatal units than for those in the UK ICU and is also lower for patients with clinical seizures in comparison to those where clinical seizures are not present.

Summary

The aim of the `discrete.reg` library is to extend the range of discrete distributions which can be fitted within the GAMLSS framework. Alongside model fitting procedures, the `gamlss` library includes many useful tools for model selection, predictions and goodness-of-fit assessments. The inclusion of a dispersion parameter when modelling the incidence of ES in paediatric coma patients demonstrates the need for more complex regression models for discrete outcomes. Such models improve the interpretation and understanding of discrete datasets.

Chapter 8

Discussion

This chapter will first discuss the unique contributions of the `Altmann`, `discrete.diag` and `discrete.reg` libraries to current software available for modelling discrete data. The implications this software has in the analysis of discrete data will then be considered in the second section, followed by a discussion of the scope and limitations of the software. In the final section, possible areas of extending this research will be addressed.

8.1 Contributions to software

8.1.1 `Altmann` library

A large variety of models for discrete data have been implemented in the `Altmann` library, which include: parameter-mix distributions such as the Delaporte, Sichel, Yule and Waring; component-mixtures including adjustments for zero-inflation and mixtures of distributions; truncated distributions such as the positive Holla and Sichel; the Lerch family including the Good, Zeta, Zipf and Lerch distribution and finally, distributions in the Generalized Poisson family, which are the Neyman type A, Hermite, generalized Hermite, Gegenbauer and generalized Gegenbauer. Many of these distributions have not previously been implemented in `R`. The benefit of the `Altmann` library is that these models can be found together, allowing the fit of these distributions to a dataset to easily be compared using goodness-of-fit statistics in the altmann.fitter function. A novel aspect of the parameter fitting procedure in this library is the use of rapid estimates as starting values in the maximum likelihood algorithm, which improves the efficiency of the estimation procedure.

8.1.2 discrete.diag library

The randomized quantile method of calculating residuals produces residuals for discrete response variables on a continuous scale. These residuals have a standard normal distribution and are utilized in plots for residual analysis. This has been implemented for the range of distributions which can be fitted using the maximum likelihood estimation functions in the Altmann library. The EPGF plot is a new implementation of the methodology presented by Nakamura and Pérez-Abreu (1993b) and Rueda and O'Reilly (1999) which provides model comparisons through the use of the EPGF and fitted pgf*'s* of a dataset. Previously, there were no appropriate software techniques available for the detection of outliers in discrete data. The discrete.diag library implements the EPGF outliers method for investigating outliers and presents a novel use of the SI as a tool to detect outliers.

8.1.3 discrete.reg library

Three discrete distributions the geometric, Yule and Waring are introduced for the GAMLSS framework. These additional distributions can be fitted as regression models using the gamlss() procedure in the gamlss library.

8.2 Implications for data analysis

The purpose of the R libraries is to facilitate the interpretation of discrete data. The libraries provide a larger variety of distributions including more complex models which enables an appropriate distribution to be chosen to fit the data. Methods for comparing distributions, the epgf.plot and altmann.fitter functions, offer improved ability to compare the fit of distributions, ensuring that the distribution with the optimum fit is chosen to model the data. Diagnostic methods can also be used to test the adequacy of the fit and check for possible outlying observations. The techniques in these libraries ensures the distribution chosen provides the best possible approximation of the data, resulting in the maximum information available to be extracted from the data. This improves interpretation of the data and may enhance the understanding of clinical aspects of disease, offering new strategies for treatment and prevention.

Three examples of discrete data from the fields of child health and epidemiology illustrate the benefits of improved analysis capabilities afforded by the `R` libraries. In Chapter 5, Zipf distributions predict the surname distribution across districts in the UK. The fitting of Zipf distributions to the surnames distribution instead of the usual one-parameter Zeta distribution, which has been previously used to model surname frequencies, enable the interpretation of the parameters of the Zipf distributions to be used as measures for assessing the diversity of surnames in the UK.

Outlier detection methods applied to counts of cysts in steroid treated embryonic mouse kidneys in Chapter 6 indicate the importance of the choice of model fitted to the data. Under a Poisson distribution it would appear that the observation of 19 cysts is an outlier but under a negative binomial distribution, which includes a dispersion parameter, this observations is not considered an outlier. The inclusion or exclusion of the potential outlier has an impact on the interpretation of the model, as a high cyst count indicates an abnormality in kidney growth.

Finally, a series of regression models were used to analyse the incidence of ES in paediatric coma patients in Chapter 7. A negative binomial distribution is the best model of those fitted to the dataset, with the incidence of ES associated with centre of study and presence of clinical seizures. This model includes a dispersion parameter, which allows the dispersion to vary according to the two covariates. This example illustrates the need for complex distributions to model discrete outcome variables.

8.3 Limitations of libraries

The `Altmann` library can estimate 32 distributions. The Altmann Thesaurus (Wimmer and Altmann, 1999) is perhaps the most complete source documenting discrete distributions, containing 100's of distributions and the Altmann fitter software implements approximately 200 of these distributions (Altmann, 1997). There is therefore potential to include more distributions in the `Altmann` `R` library. The maximum likelihood estimation functions provide reliable parameter estimates but the procedure does not always converge. This may be due to the the incorrect specification of the model to the data or unsuitable starting values for parameters given by the rapid estimates. Parameter values resulting from rapid estimation may be outside the parameter bounds. The inclusion of an optional argument in the maximum likelihood estimation functions to specify alternative starting points may help users to ensure convergence. Similarly, alternative methods of minimization selected in the `optim` function used by the maximum likelihood algorithm allow the user to adjust the maximum likelihood procedure to improve convergence.

A non-parametric method, the EPGF technique of outlier detection places no assumptions on the dataset, instead the empirical *pgf* is used to create a smooth transformation of the data from which we take our inferences. The *SI* plot is a parametric method and relies on the underlying assumptions of the model used to generate parameter estimates to calculate *SI*'s. We assume the chosen model is an appropriate and good fit to the data and any parameter estimates are correct. The *SI* cannot be used to compare the fit of distributions but informs us which values are surprising under a specified model. The benefit of this method is that it is not necessary to graphically display the *SI* to detect outliers- if a *SI* value is greater than the threshold of 1,000 then it is considered to be an outlier.

The geometric, Yule and Waring distributions have been programmed as `gamlss.family` objects in the `discrete.reg` library. There is the potential to program more distributions using the GAMLSS framework. For example, the Zipf, Zeta and Good distributions are members of the Lerch family and have not been implemented as regression models in `R`. The generalized Poisson family also has distributions which could be introduced as regression models using the gamlss library. These are the Neyman type A, Hermite, generalized Hermite, Gegenbauer and generalized Gegenbauer. The GAMLSS framework requires that the distribution be parameterized in terms of the location, scale and shape parameters which proves difficult where there are more than two parameter and/or expressions for the mean and variance are complex. There are also limitations due to the derivatives of the likelihood, which for some distributions are complex. Procedures for the numerical estimation of derivatives are available in the gamlss library which utilizes the density function of the distribution and can be used to estimate the derivatives in cases where analytical solutions are unavailable. However, the disadvantage of using numerical derivatives is the resulting estimation procedure is slower.

8.4 Further Work

The R environment is provided with a command line interface (CLI) which requires users to have a good knowledge of the language. CLI's can be intimidating for beginners and therefore graphic user interfaces (GUI) are often preferable. The R libraries in this thesis could be made more user-friendly through the creation of a GUI to perform analyses. There are various types of GUI, such as: menus and dialog boxes (MDB) which are commonly found in statistical environments such PASW (SPSS Inc, 2011), spreadsheets such as Microsoft Excel (Microsoft, 2010b), notebook style GUI's which are an extensions of word processors, for example MATHEMATICA (Inc, 2009) and web-based interfaces in which active web pages with forms trigger analyses on a server. Several projects develop or offer the opportunity to develop alternate GUI (CRAN, 2010). The tcltk library (R Development Core Team, 2009), available as part of the R language when downloaded, provides access to the platform-independent Tcl scripting language and Tk GUI elements and allows building of custom dialog boxes to create GUI. Alternatively, the R-(D)COM server allows access to R using Microsoft COM to build an R GUI client using tools such as Microsoft C++ (Microsoft, 2010a), Microsoft Visual Basic (Microsoft, 2010c) or Microsoft Excel (Microsoft, 2010b). These resources could be utilized to create a GUI for the `Altmann` and `discrete.diag` libraries.

Data on several related discrete outcome measurements can be modelled jointly using a multivariate approach. For example, in studies of birth defects several variables measuring facial growth can be used to characterize a gradient of effect (Sammel et al., 1997). Johnson et al. (1997) present a range of analyses for discrete bivariate and multivariate data, however these distributions have not been addressed as part of this thesis. There is a need for software to analyse bivariate and multivariate discrete response data. The libraries developed in this thesis provide will be extended to incorporate distributions which allow for bivariate and multivariate discrete data.

Longitudinal studies allow investigation of the effect of repeated measurements where observations are grouped into levels. These repeated measurements are correlated and the correct statistical approach requires random effects or a multilevel model. An example of a longitudinal dataset is found in a study of Picture Exchange Communication System (PECS) training in Autistic children in Section 1.2.4 of Chapter 1, which yields repeated outcome measures- the frequency of initiations, frequency of PECS use and the frequency of speech- across three treatments schedules and over three time periods. This study was previously analysed using multilevel logistic regression models (Howlin et al., 2007) and later Poisson multilevel regression models to test for an interaction between treatments and baseline measures (Gordon et al., 2011).

Software for fitting multilevel or random effects models can be found in a range of statistical environments, including `Stata` (StataCorp, 2009) and `R` (R Development Core Team, 2009). Functions for panel models in `Stata` allow for fixed effects, random-effect and population-averaged models for the Poisson (`xtpoisson`) and negative binomial (`xtnbreg`) distributions and the `gllamm` add-on package fits Generalized Linear Latent and Mixed Models (Rabe-Hesketh et al., 2004). Random-effects models can also be fitted in the `gamlss` package in `R` using the `random()` function. The MLwiN software environment (Rasbash, J and Charlton, C and Browne, W J and Healy, M and Cameron, B, 2009; Rasbash et al., 2009) provides the specification and analysis of a wide range of multilevel models, including Binomial and Poisson multilevel regression models for discrete data with repeated measurements or clustered levels. A negative binomial distribution can also be fitted as a regression model as an extra option of the error distribution for Poisson regression models. Software for models incorporating random effects into more complex distribution regression models, such as the zero-inflated Poisson or zero-inflated negative binomial distributions will be included in extensions of the libraries.

8.5 Conclusion

The aim of this thesis has been to develop software to implement models for discrete epidemiological and clinical data. It has been identified that there is a need for software to make more complex methodologies for the analysis of discrete data available to the clinical and scientific community. Three add-on libraries for the R environment for statistical programming provide univariate parameter estimation, model diagnostics and regression modelling within the GAMLSS framework. These libraries provide a toolkit of methods for analysing discrete data, allowing clinical scientist to fit and interpret relatively complex statistical models for a wide range of data with increased ease, thus offering an improved understanding of discrete data.

References

Afifi, A., Kotlerman, J., Ettner, S., and Cowan, M. (2007). Methods for Improving Regression Analysis for Skewed Continuous or Counted Responses. *Annual Review of Public Health.*, 28:95–111.

Ahmed, M. S. (1961). On a locally most powerful boundary randomized similar test for the independence of two Poisson variables. *The Annals of Mathematical Statistics*, 32:809–827.

Akaike, H. (1974). A new look at the statistical model identification. *IEEA Transactions on Automatic Control*, 19:716–722.

Altmann, G. (1997). *Altmann-Fitter: iterative fitting of probability distributions*. Ludenscheid: RAM-Verlag (Software).

Anscombe, F. J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37:358–382.

Anscombe, F. J. (1953). Contribution to the discussion of H.Hotelling's paper. *Journal of the Royal Statistical Society-Series B*, 15:193–232.

Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D., and Charman, T. (2006). Prevalance of disorders of autism spectrum in a population cohort of children in South Thames: The Special Needs and Austism Project. *Journal of the American Academy of Orthopaedic Surgeons*, 368:210–215.

Booth, J. G., Casella, G., Friedl, H., and Hobert, J. P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, 3:179–191.

Bortkiewicz, L. (1898). Das Gesetz der Kleinen Zahlen.

Brakman, S., Garretsen, H., Van Marrewijk, C., and van den Berg, M. (1999). The return of Zipf: Towards a further understanding of the rank-size distribution. *Journal of Regional Science*, 39:183–213.

Brown, K., Ridout, D., Goldman, A., Hoskote, A., and Penny, D. (2003). Risk factors for long intensive care unit stay after cardiopulmonary bypass in children. *Critical Care Medicine.*, 31:28–33.

Chambers, J. (2008). *Software for Data Analyis: Programming with R*. Springer, New York, USA.

Chambers, J. and Hastie, T. (1991). *Statistical Models in S*. Chapman and Hall, London, UK.

Chan, S. K., Riley, P. R., Price, K. L., McElduff, F., Winyard, P. J., Welham, S. J. M., Woolf, A. S., and Long, D. A. (2010). Corticosteroid-induced kidney dysmorphogenesis is associated with deregulated expression of known cystogenic molecules, as well as indian hedgehog. *American Journal of Physiology: Renal Physiology*, 298:F346–F356.

Chernoff, H. and Lehmann, E. L. (1954). The use of maximum likelihood estimates in chi squared tests for goodness-of-fit. *Annals of Mathematical Statistics*, 25:579–586.

Cohen, A. C. (1960). Estimation in the Truncated Poisson Distribution when Zeros and Some Ones are Missing. *Journal of the American Statistical Association*, 55:342–348.

Colantonio, S. E., Lasker, G. W., Kaplan, B. A., and Fuster, V. (2003). Use of surname models in human population biology: a review of recent developments. *Human Biology*, 75:785–807.

Cole, T. J. and Green, P. J. (1992). Smoothing Reference Centile Curves: The LMS Method and Penalized Likelihood. *Statistics in Medicine*, 11:1305–1319.

Cook, D., Hewitt, D., and Milner, J. (1972). Uses of the surname in epidemiological research. *American Journal of Epidemiology*, 96:38–44.

Cortina-Borja, M. (2006). Some remarks on the generalized Hermite and generalized Gegenbauer probability distributions and their applications. In Grzybek, P. and Kohler, R., editors, *Exact methods in the study of language and text*. de Gruyter.

Cox, D. R. (1986). Some remarks on overdispersion. *Biometrika*, 70(1):269–274.

Cox, D. R. and Snell, E. J. (1968). A general definition of Residuals. *Journal of the Royal Statistical Society- Series B (Methodological)*, 30:248–275.

CRAN (2010). Comprehensive R Archive Network (CRAN).

Currie, I. D. (1995). Maximum Likelihood Estimation and Mathematica. *Applied Statistics*, 44:379–394.

Darwin, G. H. (1875). Marriages between first cousins in England and their effects. *Journal of the Statistical Society of London*, 38:153–184.

David, F. N. and Moore, P. G. (1954). Notes on Contagious Distributions in Plant Populations. *Annals of Botany*, 18:47–53.

de Graft Acquah, H. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymetric price relationship. *Journal of Development and Agricultural Economics*, 2:1–6.

Delaporte, P. (1959). Quelques problemes de statistique mathematique poses par lassurance automobile et le bonus non sinistre. *Bulletin Trimestriel de llnstitut des Actuuires FrunCuis*, 227:87–102.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society-Series B (Methodological)*, 39:1–38.

Denuit, M. (1997). A new distribution of poisson-type for the number of claims. *Astin Bulletin*, pages 229–242.

Dobson, A. J. (2002). *An Introduction to Generalized Linear Models*. Chapman and Hall, 2nd edition edition.

Doray, L. G. and Luong, A. (1997). Efficient estimatiors for the Good family. *Communications in Statistics - Simulation and Computation*, 26:1075–1088.

Duncore, J. M., Parikh-Patel, A., and Gold, E. B. (2008). Cancer Occurence in Southeast Asian Children in California. *Journal of Pediatric Hematology/Oncology*, 26:613–618.

Dunn, P. K. and Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, 5:236–244.

Estoup, J. B. (1916). *Les gammes stenographiques*. Institute Stenographique, Paris.

Evert, S. and Baroni, M. (2008). *zipfR: Statistical models for word frequency distributions*. R package version 0.6-5.

Feller, W. (1943). On a General Class of “Contagious”Distributions. *The Annals of Mathematical Statistics*, 14:389–400.

Fox, W. R. and Lasker, G. W. (1983). The Distribution of Surname Frequencies. *International Statistics Review*, 51:81–87.

French, C. (2011). Personal communication.

Ginebra, J. and Puig, X. (2010). On the measure and the estimation of evenness and diversity. *Computational Statistics and Data Analysis*, 54:2187–2201.

Good, I. J. (1953). The population of word frequencies of species and the estimation of population parameters. *Biomatrika*, 40:237–264.

Gordon, K., Pasco, G., McElduff, F., Wade, A., Howlin, P., and Charman, T. (2011). A Communication-Based Intervention for Nonverbal Children With Autism: What

Changes? Who Benefits? *Journal of Consulting and Clinical Psychology*, 79:447–457.

Green, P. J. (1984). Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *Journal of the Royal Statistical Society- Series B*, 46:149–192.

Green, P. J. (1992). On the use of the EM for Penalizing Likelihood Estimation. *Journal of the Royal Statistical Society- Series B (Methodological)*, 52:443–452.

Groeneveld, R. and Meeden, G. (1984). Measuring Skewness and Kurtosis. *The Statistician.*, 33:391–399.

Gupta, R. P. and Jain, G. C. (1974). A generalized Hermite distribution and its properties. *SIAM Journal for Applied Mathematics*, 27:359–363.

Hald, A. (1998). *A History of Mathematical Statistics From 1750 to 1930*. Wiley series in Probability and Statistics.

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1:297–310.

Hilbe, J. M. (2007). *Negative Binomial regression*. Cambridge University Press.

Hoaglin, D. C. and Tukey, J. W. (1985). *Checking the shape of discrete distributions. In D.C. Hoaglin, F. Mosteller and J.W. Tukey (Eds.), Exploring Data Tables, Trends and Shapes. Chapter 9*. Wiley, New York.

Holla, M. S. (1966). On a Poisson-inverse Gaussian distribution. *Metrika*, 11:115–121.

Horgan, J. M. (2009). *Probability with R: An introduction with computer science applications*. Wiley.

Horton, N. J., Brown, E. R., and Quian, L. (2004). Use of R as a Toolbox for Mathematical Statistics Exploration. *The American Statistician*, 58:343–357.

Howlin, P., Gordon, R. K., Pasco, G., Wade, A., and Charman, T. (2007). The effectivness of Picture Exchange Communication System (PECS) training for teachers of children with autism: a pragmatic, group randomised controlled trial. *Journal of the Child Psychology and Psychiatry*, 48:473–481.

Ihaka, R. and Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

Ijiri, Y. and Simon, H. A. (1977). *Skew distributions and the Size of Business firms*. North Holland, Amsterdam.

Inc, W. R. (2009). Mathematica. Version 7.0, Champaign, IL.

Irwin, J. O. (1963). The place of mathematics in medical and biological statistics. *Journal of the Royal Statistical Society- Series A*, 126:1–41.

Jackman, S. (2010). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University. R package version 1.03.3.

Jara, J. and Rosenblueth, E. (1988). Probability distributions of times between characteristic subduction earthquakes. *Eartquakes Spectra*, 4:499–529.

Jobling, M. A. (2001). In the name of the father:surnames and genetics. *Trends in Genetics*, 17:353–357.

Johnson, N., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. Wiley-Interscience, NY, USA.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate discrete distributions*. Wiley.

Karlis, D. (2001). A general EM approach for maximum likelihood estimation in mixed Poisson regression models. *Statistical Modelling*, 1:305–318.

Karlis, D. and Xekalaki, E. (1999). On testing for the number of components in a mixed Poisson model. *Annals of the Institute of Statistical Mathematics*, 51:149–162.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson Distributions. *International Statistical Review*, 73:35–58.

Kemp, A. W. (1995). Splitters, lumpers and species per genus. *Mathematical Scientist*, 20:107–118.

Kemp, C. D. and Kemp, A. W. (1988). Rapid estimation for discrete distributions. *The Statistician*, 37:243–255.

Kirkwood, B. R. and Sterne, J. A. C. (2003). *Essentials in medical statistics*. Wiley-Blackwell, 2nd edition edition.

Krishnaji, N. (1970). A Characteristic Property of the Yule distribution. *Sankya: The Indian Journal of Statistics, Series A*, 32:343–346.

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods Research*, 33:188.

Kulasekera, K. B. and Tonkyn, D. W. (1992). A new distribution, with applications to survival, dispersal and dispersion. *Communications in Statistics - Simulation and Computation*, 21:499–518.

Langley-Evans, S. C., Phillips, G. J., Benediktsson, R., Gardner, D. S., Edwards, Jackson, A. A., and Seckl, J. R. (1996). .. Protein intake in pregnancy, placental glucocorticoid metabolism and the programming of hypertension in the rat. *Placenta*, 17:169–172.

Lasker, G. W. (1985). *Surnames and genetic structure*. Cambridge University Press, Cambridge, UK.

Lord, C., Risi, S., Lambercht, L., Cook, E. H., and Leventhal, B. L. (1999). The Autism Diagnostics Observation Scheduale- Generic: A standard measure of social and communication deficites associated with the spectrum of autism. *Journal of the Autism and Developmental Disorders*, 30:205–223.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16:317–323.

Mandlebrot, B. (1959). A note on a class of skew distribution functions:Analysis and critique of a paper by H. A. Simon. *Information and control*, 2:90–99.

McCullagh, P. and Nelder, J. A. (1983). *Generalized Linear Models*. Chapman and Hall.

McElduff, F., Cortina-Borja, M., Chan, S.-K., and Wade, A. (2010). When *t*-tests or Wilcoxon-Mann-Whitney tests won't do. *Advances in Physiology Education*, 34:128–133.

McElduff, F., Mateos, P., Wade, A., and Cortina-Borja, M. (2008). What's in a name? The frequency and geographic distributions of UK surnames. *Significance*, 5:189–192.

McKendrick, A. G. (1926). Applications of Mathematics to Medical Problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130.

Medhi, J. and Borah, M. (1984). On a generalized Gegenbauer polynomials and associated probabilities. *Sankhya: The Indian Journal of Statistics- Series B*, 46:157–165.

Meng, X. L. and Rubin, D. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, 80:267–278.

Microsoft (2010a). *Microsoft C++*. Redmond, Washington.

Microsoft (2010b). *Microsoft Excel*. Redmond, Washington.

Microsoft (2010c). *Microsoft Visual Basic*. Redmond, Washington.

Monaco, J., Abbott, L., and Kahana, M. (2007). Lexico-semantic structure and the word-frequency effect in recognition memory. *Learning memory.*, 14:204–213.

Morgan, B. J. T., Palmer, K. J., and Ridout, M. S. (2007). Score Test Oddities. *The American Statistician*, 61:285–288.

Mullen, E. (1999). *Mullen Scales of Early Learning*. American Guidence Services, Circle Pines, MN.

Mwalili, S. M. (2007). *zicounts: Counts data models: zero-inflation as well as interval icensored*. R package version 1.1.4.

Nakamura, M. and Pérez-Abreu, V. (1993a). Empirical probability generating function: An overview. *Insurance: Mathematics and Economics*, 12:287–295.

Nakamura, M. and Pérez-Abreu, V. (1993b). Exploratory Data Analysis for Counts Using the Empirical Probability Generating Function. *Communications in Statistics- Theory and Methods*, 22:827–842.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society- Series (A) General*, 135:370–384.

Neyman, J. (1939). On a New Class of 'Contagious' Distributions, Applicable in Entomologyand Bacteriology. *The Annals of Mathematical Statistics*, 17:53–61.

Nikoloulopoulos, A. K. and Karlis, D. (2008a). On modeling count data: a comparison of some well-known discrete distributions. *Journal of Statistical Computation and Simulation*, 78:437–457.

Nikoloulopoulos, A. K. and Karlis, D. (2008b). On modeling count data: a comparison of some well-known discrete distributions. *Journal of Statistical Computation and Simulation*, 78:437–457.

Panaretos, J. (1989). On the evolution of surnames. *International Statistics Review*, 57:161–179.

Pearson, K. (1915). On certain types of compound frequency distributions in which the components can be individually described by binomial series. *Biometrika*, 11:139–144.

Piazza, A., Rendine, S., Zei, G., Moroni, A., and Cavalli-Sforza, L. L. (1987). Migration rates of human populations from surname distribution. *Nature*, 329:714–716.

Plunkett, A. G. and Jain, G. C. (1975). Three generalised negative binomial distributions. *Biometrische Zeitschrift*, 17:286–302.

Puig, P. (2003). Characterizing Additively Closed Discrete Models by a Property of Their Maximum Likelihood Estimators, With an Application to Generalized Hermite Distributions. *Journal of the American Statistical Association*, 98:687–692.

Puig, P. and Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, 101:332–340.

Puig, X., Ginebra, J., and Perez-Casany, M. (2009). Extended truncated Inverse Gaussian-Poisson model. *Statistical Modelling*, 9:151–171.

Pustet, R. and Altmann, G. (2005). Morpheme Length Distribution in Lakota. *Journal of Quantitative Linguistics*, 12:1744–5035.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). GLLAMM Manual. U.C. Berkeley Division of Biostatistics Working Paper Series. Paper 160. Universirt of California, Berkley.

Rankin, J., Silf, K. A., Pearce, M. S., Parker, L., and Ward Platt, M. (2008). Congenital

Anomaly and Childhood Cancer: A Population-Based, Record Linkage Study. *Pediatric Blood Cancer*, 51:608–612.

Rasbash, J., Steele, F., Browne, W. J., and Goldstein, H. (2009). *A Users Guide to MLwiN, v2.10*. Centre for Multilevel Modelling, University of Bristol.

Rasbash, J and Charlton, C and Browne, W J and Healy, M and Cameron, B (2009). *MLwiN Version 2.1*. Centre for Multilevel Modelling, University of Bristol.

Raspall-Chaure, M., Chin, R. F., Neville, B. G., and Scott, R. C. (2006). Outcome of paediatric convulsive status epilepticus: a systematic review. *Lancet*, 5:769–779.

Redheffer, R. M. (1951). A Note on the Surprise Index. *The Annuals of Mathematical Statistics*, 22:128–130.

Ridout, M. S., Hinde, J., and Demetrio, C. G. B. (2001). A score test for testing a Zero-Inflated Poisson regression model against Zero-Inflated Negative binomial alternatives. *Biometrics*, 57:219–223.

Rigby, R. A. and Stasinopoulos, D. M. (1996). A Semi-parametric Additive Model for Variance Heterogeneity. *Statistical Computing*, 6:57–65.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive model for location, scale and shape. *Applied Statistics*, 54:507–554.

Rose, C. and Smith, M. D. (2000). Symbolic maximum likelihood estimation with Mathematica. *The Statistician*, 49:229–240.

Rose, C. and Smith, M. D. (2002). *Mathematical Statistics with Mathematica*. Springer.

Rueda, R. and O'Reilly, F. O. (1999). Tests of fit for discrete distributions based on the Probability Generating Function. *Communications in Statistics- Simulation and Computation*, 28:259–274.

Ruohonen, M. (1988). On a model for the claim number process. *Astin Bulletin*, 18:57–68.

Rutter, M., Bailey, A., and Lord, C. (2003). *Social Communication Questionniare (SCQ)*. Western Psychological Services, Los Angeles.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society-Series B*, 59:667–678.

SAS Institute Inc (2011). *SAS software, Version 9.3 of the SAS System for Windows*. Cary, NC, USA.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annuals of Statistics*, 6:461–464.

Seal, H. L. (1947). A probability distribution of deaths at age x when policies are counted instead of lives. *Skandinavisk Aktuarietidskrift*, 30:18–43.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association*, 70:542–547.

Simon, H. A. (1955). On a Class of Skew Distribution Functions. *Biometrika*, 42:425–440.

SPSS Inc (2011). *SPSS for Windows, Rel. 19*. Chicago.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23:1–46.

Stasinopoulos, M. and Rigby, B with contributions from Akantziliotou, C. (2008). *gamlss: Generalized Additive Models for Location Scale and Shape*. R package version 1.9-4.

StataCorp (2009). *Stata: Release 11. Statistical Software*. StataCorp LP, College Station, TX.

Testa, M. and Simonson, D. (1996). Assessment of quality-of-life outcomes. *The New England Journal of Medicine.*, 334:835–840.

Thurston, S. W., Wand, M. P., and Wiencke, J. K. (2000). Negative Binomial Additive Models. *Biometrics*, 56:139–144.

Valencia, I., Lozano, G., Kothcare, S. V., Melvin, J. J., Khurana, D. S., Hardison, H. H., Yum, S. S., and Legido, A. (2006). Epileptic Seizures in the pediatric intensive care unit setting. *Epileptic Disorders*, 8:227–284.

van Buuren, S. and Frederiks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, 20:1259–1277.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Volkmar, F. R., Lord, C., Bailey, A., Schultz, R. T., and Klin, A. (2004). Autism and pervasive developmental disorders. *Journal of Child Psychology and Pyschiatry*, 45:135–170.

Voracek, M. and Sonneck, G. (2007). Surname study of suicide in Austria: Differences in regional suicide rates correspond to the genetic structure of the population. *Wien Klin Wochenschr*, 119:355–360.

Weaver, W. (1948). Probability, Rarity, Interest, and Surprise. *The Scientific Monthly*, 67:390–392.

Welham, S. J., Riley, P. R., Wade, A., Hubank, M., and Woolf, A. S. (2005). Maternal diet programs embryonic kidney gene expression. *Genomics*, 22:48–56.

Welham, S. J., Wade, A., and Woolf, A. S. (2002). . Protein restriction in pregnancy is associated with increased apoptosis of mesenchymal cells at the start of rat metanephrogenesis. *Kidney*, 61:1231–1242.

Willmot, G. (1986). Mixed compound Poisson distributions. *ASTIN Bulletin*, 16:S59–S79.

Willmot, G. E. (1989). Limiting tail behaviour of some discrete compound distributions. *Insurance: Mathematics and Economics*, 8:175–185.

Wimmer, G. and Altmann, G. (1995). Generalized Gegenbauer Distribution Revised. *Sankhya: The Indian Journal of Statistics- Series B*, 57:450–452.

Wimmer, G. and Altmann, G. (1996). The multiple Poisson distribution, Its Characteristics and a Variety of Forms. *Biometrical Journal*, 38:995–1011.

Wimmer, G. and Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Stamm.

Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11:95–103.

Yau, K. K. W., Wang, K., and Lee, A. H. (2003). Zero-inflated Negative Binomial Mixed Regression Modelling of Over-Dispersed Count Data with Extra Zeros. *Biometrical Journal*, 45:437–452.

Yee, T. W. (2008). The VGAM Package. *R News*, 8:28–39.

Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willia, F.R.S. *Philosophical Transactions of the Royal Society: Series B - Biological Sciences*, 213:21–87.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*, 27:1–25.

Zelterman, D. (2004). *Discrete distributions: applications in the health sciences*. Wiley.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.

zipfR (2010). zipfR:user-friendly LNRE modelling in R.

Zörnig, P. and Altmann, G. (1995). Unified representation of Zipf distributions. *Computational Statistics & Data Analysis*, 19:461–473.

Appendix A

Distribution Moments

The raw and central moments for the distributions presented in Chapter 2 are additionally presented in this Appendix. The first raw moment $\mu_1^{'}$ gives the mean, whilst the second central moment μ_2 gives the variance. The third and fourth central moments are used in calculations for the the skewness and kurtosis coefficients.

Basic Distributions

Bernoulli (p)

The raw moments of the Bernoulli distribution are all equal, where $\mu_n^{'} = p$. Central moments:

$$\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= p(1-p) \\ \mu_3 &= p(1-p)(1-2p) \\ \mu_4 &= p(1-p)(3p^2-3p+1) \end{aligned} \quad . \tag{A.1}$$

Binomial (p, n)

Raw moments:

$$\begin{aligned} \mu'_1 &= np \\ \mu'_2 &= np(1-p+np) \\ \mu'_3 &= np(1-3p+3np+2p^2-3np^2+n^2p^2) \\ \mu'_4 &= np(1-7p+7np+12p^2-18np^2+6n^2p^2-6p^3+11np^3-6n^2p^3+n^3p^3) \end{aligned} \quad , \tag{A.2}$$

Central moments:

$$\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= np(1-p) \\ \mu_3 &= np(1-p)(1-2p) \\ \mu_4 &= np(1-p)\left(3p^2(2-n)+3p(n-2)+1\right) \end{aligned} \quad . \tag{A.3}$$

Geometric (p)

Raw moments:

$$\begin{aligned} \mu_1' &= \tfrac{(1-p)}{p} \\ \mu_2' &= \tfrac{(2-p)(1-p)}{p^2} \\ \mu_3' &= \tfrac{(1-p)(6+(p-6)p)}{p^3} \\ \mu_4' &= \tfrac{(2-p)(1-p)(12+(p-12)p)}{p^4} \end{aligned} \quad , \tag{A.4}$$

Central moments:

$$\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= \tfrac{1-p}{p^2} \\ \mu_3 &= \tfrac{(p-2)(p-1)}{p^3} \\ \mu_4 &= -\tfrac{(p-1)(9+(p-9)p)}{p^4} \end{aligned} \quad . \tag{A.5}$$

Hypergeometric (m, n, k)

Raw moments:

$$\begin{aligned} \mu_1' &= \tfrac{km}{m+n} \\ \mu_2' &= \tfrac{km(k(m-1)+n)}{(m+n-1)(m+n)} \\ \mu_3' &= \tfrac{km(k^2(m-2)(m-1)+3k(m-1)n+n(n-m))}{(m+n-2)(m+n-1)(m+n)} \\ \mu_4' &= \tfrac{km(k^3(m-3)(m-2)(m-1)+6k^2(m-2)(m-1)n-k(m-1)(4m-7n-1)n+n(m+m^2+n-4mn+n^2))}{(m+n-3)(m+n-2)(m+n-1)(m+n)} \end{aligned} \quad , \tag{A.6}$$

Central moments:

$$\begin{aligned}
\mu_1 = {} & 0 \\
\mu_2 = {} & \tfrac{kmn(m+n-k)}{(m+n-1)(m+n)^2} \\
\mu_3 = {} & -\tfrac{km(k-m-n)(2k-m-n)(m-n)n}{(m+n-2)(m+n-1)(m+n)^3} \\
\mu_4 = {} & \tfrac{1}{(m+n-3)(m+n-2)(m+n-1)(m+n)^4}\left(kmn(-k+m+n)\right)(m+n)^2 \\
& (m+m^2+n-4mn+n^2)+3k\left(m^3(n-2)+2m^2n^2-2n^3+mn^3\right) \\
& -3k^2\left(m^2(n-2)-2n^2+mn(n+2)\right))
\end{aligned} \quad . \tag{A.7}$$

Poisson (μ)

Raw moments:

$$\begin{aligned}
\mu_1' &= \mu \\
\mu_2' &= \mu\,(1+\mu) \\
\mu_3' &= \mu\,(1+\mu\,(3+\mu)) \\
\mu_4' &= \mu\,(1+\mu\,(7+\mu\,(6+\mu)))
\end{aligned} \quad , \tag{A.8}$$

Central moments:

$$\begin{aligned}
\mu_1 &= 0 \\
\mu_2 &= \mu \\
\mu_3 &= \mu \\
\mu_4 &= \mu(1+3\mu)
\end{aligned} \quad . \tag{A.9}$$

Parameter Mix Distributions

Negative Binomial (p, r)

Raw moments:

$$\begin{aligned}
\mu_1' &= r(\tfrac{1}{p}-1) \\
\mu_2' &= \tfrac{(p-1)r((p-1)r-1)}{p^2} \\
\mu_3' &= \tfrac{r(2-3p+p^2+3(p-1)^2r-(p-1)^3r^2)}{p^3} \\
\mu_4' &= \tfrac{r(6-(p-4)(p-3)p+11r+p((19-4p)p-26)r-6(p-1)^3r^2+(p-1)^4r^3)}{p^4}
\end{aligned} \quad , \tag{A.10}$$

Central moments:

$$\begin{aligned}\mu_1 &= 0\\ \mu_2 &= \tfrac{r-pr}{p^2}\\ \mu_3 &= \tfrac{((p-2)(p-1)r)}{p^3}\\ \mu_4 &= -\tfrac{(p-1)r\left(p^2+3(2+r)-3p(2+r)\right)}{p^4}\end{aligned} \tag{A.11}$$

Holla (α, θ)

Raw moments:

$$\begin{aligned}\mu_1' &= \theta\\ \mu_2' &= \theta+\theta^2+\tfrac{\theta^3}{\alpha}\\ \mu_3' &= \tfrac{\theta(3\theta^4+3\alpha\theta^2(\theta+1)+\alpha^2(1+\theta(\theta+3)))}{\alpha^2}\\ \mu_4' &= \tfrac{1}{\alpha^3}\theta(15\theta^6+3\alpha\theta^4(6+5\theta)+\alpha^2\theta^2(7+6\theta(3+\theta))+\alpha^3(1+\theta(7+\theta(6+\theta))))\end{aligned}, \tag{A.12}$$

Central moments:

$$\begin{aligned}\mu_1 &= 0\\ \mu_2 &= \theta+\tfrac{\theta^3}{\alpha}\\ \mu_3 &= \theta+\tfrac{3\theta^3}{\alpha}+\tfrac{3\theta^5}{\alpha^2}\\ \mu_4 &= \tfrac{1}{\alpha^3}\left(\theta\left(15\theta^6+3\alpha\theta^4(6+\theta)+\alpha^3(1+3\theta)+\alpha^2\theta^2(7+6\theta)\right)\right)\end{aligned}, \tag{A.13}$$

Sichel (α, θ, γ)

Raw moments:

$$\begin{aligned}
\mu_1' = {} & \tfrac{1}{2\sqrt{1-\theta}K_\gamma(\alpha\sqrt{1-\theta})}\big(\alpha\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})\big) \\
\mu_2' = {} & \tfrac{1}{4(\theta-1)^2}\left(\theta\left(4\gamma+4\gamma^2\theta-\alpha^2(-1+\theta)\theta+\tfrac{2\alpha\sqrt{1-\theta}(1+\gamma\theta)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})}\right)\right) \\
\mu_3' = {} & \tfrac{1}{8(-1+\theta)^3}\left(\theta\left(-24\gamma^2\theta-2\alpha^2(\theta-3)(\theta-1)\theta-8\gamma^3\theta^2+4\gamma\left(-2+\theta\left(-2+\alpha^2(\theta-1)\theta\right)\right)\right)\right) \\
& +\tfrac{\alpha\sqrt{1-\theta}\left(-4+\theta\left(-4-12\gamma-4\gamma^2\theta+\alpha^2(-1+\theta)\theta\right)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})} \\
\mu_4' = {} & \tfrac{1}{16(-1+\theta)^4}\theta\left(96\gamma^3\theta^2+16\gamma^4\theta^3+\alpha^2(\theta-1)\theta\left(-28+\theta\left(8+\left(-4+\alpha^2(\theta-1)\right)\theta\right)\right)\right. \\
& +4\gamma\left(4+\theta\left(16+\left(4+3\alpha^2(-4+\theta)(-1+\theta)\right)\theta\right)\right) \\
& +4\gamma^2\theta\left(28+\theta\left(16-3\alpha^2(-1+\theta)\theta\right)\right) \\
& +\tfrac{1}{K_\gamma(\alpha\sqrt{1-\theta})}4\alpha\sqrt{1-\theta}\,(2+\theta\,(8+12\gamma^2\theta+\left(2+\alpha^2(-3+\theta)(-1+\theta)\right)\theta \\
& +2\gamma^3\theta^2+\gamma\left(14+\theta\left(8-\alpha^2(-1+\theta)\theta\right)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})
\end{aligned} \tag{A.14}$$

Central moments:

$$\begin{aligned}
\mu_1 = {} & 0 \\
\mu_2 = {} & \tfrac{1}{4(\theta-1)^2}\theta\,(4\gamma+4\gamma^2\theta-\alpha^2(\theta-1)\theta \\
& +\big(\alpha\,\big(2\sqrt{1-\theta}(1+\gamma\theta)K_{\gamma-1}(\alpha\sqrt{1-\theta})K_\gamma(\alpha\sqrt{1-\theta}) \\
& +\alpha(\theta-1)\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})^2/K_\gamma(\alpha\sqrt{1-\theta})^2 \\
\mu_3 = {} & \tfrac{1}{4}\theta\,\Bigg(-\tfrac{\alpha^2(\theta-3)(\theta-1)\theta-8\gamma^3\theta^2+\gamma\left(4+\theta\left(4+\alpha^2(\theta-1)\theta\right)\right)}{(\theta-1)^3} \\
& +\big(\alpha\,\big(-3\alpha\sqrt{1-\theta}\theta(1+\gamma\theta)K_{\gamma-1}(\alpha\sqrt{1-\theta})^2K_\gamma(\alpha\sqrt{1-\theta}) \\
& +\left(2+\theta\left(2-6\gamma-10\gamma^2\theta+\alpha^2(\theta-1)\theta\right)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})K_\gamma(\alpha\sqrt{1-\theta})^2 \\
& -\alpha^2(\theta-1)\theta^2K_{\gamma+1}(\alpha\sqrt{1-\theta})^3/\left((1-\theta)^{\frac{5}{2}}K_\gamma(\alpha\sqrt{1-\theta})^3\right) \qquad . \\
\mu_4 = {} & \tfrac{1}{16}\alpha\theta\,\Bigg(\tfrac{\alpha\theta\left(-28+\theta\left(8+4\gamma(-6+\theta)-4\gamma^2\theta+\left(-4+\alpha^2(-1+\theta)\right)\theta\right)\right)}{(-1+\theta)^3} \\
& -\big(K_{\gamma+1}\left(\alpha\sqrt{1-\theta}\right)(-4\left(2+\theta\left(8+12\gamma^2\theta+\left(2-\alpha^2(-3+\theta)(\theta-1)\right)\theta\right.\right. \\
& +2\gamma^3\theta^2+\gamma\left(14+\theta\left(8+\alpha^2(\theta-1)\theta\right)\right)K_\gamma\left(\alpha\sqrt{1-\theta}\right)^3 \\
& +2\alpha\sqrt{1-\theta}\theta\left(8+\theta\left(8+24\gamma+8\gamma^2\theta+\alpha^2(\theta-1)\theta\right)\right)K_\gamma\left(\alpha\sqrt{1-\theta}\right)^2K_{\gamma+1}\left(\alpha\sqrt{1-\theta}\right) \\
& +12\alpha^2(\theta-1)\theta^2(1+\gamma\theta)K_\gamma\left(\alpha\sqrt{1-\theta}\right)K_{\gamma+1}\left(\alpha\sqrt{1-\theta}\right)^2 \\
& +3\alpha^3(1-\theta)^{3/2}\theta^3K_{\gamma+1}\left(\alpha\sqrt{1-\theta}\right)^3/\left((1-\theta)^{7/2}K_\gamma\left(\alpha\sqrt{1-\theta}\right)^4\right)
\end{aligned} \tag{A.15}$$

Delaporte (α, β, γ)

Raw moments:

$$\begin{aligned}
\mu_1' = & \ \alpha + \beta\left(\frac{1}{\gamma} - 1\right) \\
\mu_2' = & \ \frac{1}{\gamma^2}\left(\beta^2(\gamma-1)^2 + \alpha(\alpha+1)\gamma^2 - \beta(\gamma-1)(2\alpha\gamma+1)\right) \\
\mu_3' = & \ \frac{1}{\gamma^3}\left(-\beta^3(\gamma-1)^3 + \alpha(1+\alpha(3+\alpha))\gamma^3 + 3\beta^2(\gamma-1)^2(1+\alpha\gamma)\right. \\
& \left. -\beta(\gamma-1)(2+\gamma(-1+3\alpha(1+\gamma+\alpha\gamma)))\right) \\
\mu_4' = & \ \frac{1}{\gamma^4}\left(\beta^4(\gamma-1)^4 + \alpha(1+\alpha(7+\alpha(6+\alpha)))\gamma^4 - 2\beta^3(\gamma-1)^3(3+2\alpha\gamma)\right. \\
& +\beta^2(\gamma-1)^2(11+2\gamma(-2+3\alpha(2+\gamma+\alpha\gamma))) - \beta(\gamma-1) \\
& \left.(6+\gamma\left(-6+\gamma+2\alpha\left(4+\gamma+3\alpha\gamma+2(1+\alpha(3+\alpha))\gamma^2\right)\right)\right)
\end{aligned} \tag{A.16}$$

Central moments:

$$\begin{aligned}
&\mu_1 = 0 \\
&\mu_2 = \alpha + \frac{\beta-\beta\gamma}{\gamma^2} \\
&\mu_3 = \frac{\beta(\gamma-2)(\gamma-1)+\alpha\gamma^3}{\gamma^3} \\
&\mu_4 = \frac{1}{\gamma^4}\left(3\beta^2(-1+\gamma)^2 + \alpha(1+3\alpha)\gamma^4 - \beta(-1+\gamma)(6+\gamma(-6+\gamma+6\alpha\gamma))\right)
\end{aligned} . \tag{A.17}$$

Yule (λ)

Raw moments:

$$\begin{aligned}
\mu_1' = & \ \lambda\Gamma(1+\lambda)\,{}_2F_1(2,2,3+\lambda,1) \\
\mu_2' = & \ \lambda\Gamma(1+\lambda)\,{}_2F_1(2,2,3+\lambda,1) + 2\,{}_2F_1(3,3,4+\lambda,1) \\
\mu_3' = & \ \frac{\lambda(6+\lambda)\,{}_2F_1(2,2,3+\lambda,1)}{(\lambda-3)(\lambda^2-4)} \\
\mu_4' = & \ \lambda\Gamma(1+\lambda)\left(\,{}_2F_1(2,2,3+\lambda,1) + 28\,{}_2F_1(3,3,4+\lambda,1) + \right. \\
& 216\,{}_2F_1(4,4,5+\lambda,1) + 576\,{}_2F_1(5,5,6+\lambda,1)
\end{aligned} , \tag{A.18}$$

Central moments:

$$\begin{aligned}
\mu_1 = {} & 0 \\
\mu_2 = {} & \lambda\Gamma[1+\lambda]\,(\,{}_2F_1(2,2,3+\lambda,1) - \lambda\Gamma(1+\lambda)\,{}_2F_1(2,2,3+\lambda,1)^2 \\
& +28\,{}_2F_1(3,3,4+\lambda,1) + 216\,{}_2F_1(4,4,5+\lambda,1) \\
& +576\,{}_2F_1(5,5,6+\lambda,1) \\
\mu_3 = {} & \lambda\,{}_2F_1(2,2,3+\lambda,1)\left(\tfrac{(6+\lambda)\Gamma(2+\lambda)}{6-5\lambda+\lambda^2} + \tfrac{3\lambda\Gamma(\lambda+1)\,{}_2F_1(2,2,3+\lambda,1)}{2+\lambda-\lambda^2}\right. \\
& +2\lambda^5\Gamma(\lambda)^3\,{}_2F_1(2,2,3+\lambda,1)^2 \\
\mu_4 = {} & \lambda\Gamma(\lambda+1)\left(\,{}_2F_1(2,2,3+\lambda,1) - \tfrac{4\lambda(6+\lambda)\Gamma(4+\lambda)\,{}_2F_1(2,2,3+\lambda,1)^2}{36-13\lambda^2+\lambda^4}\right. \\
& +\tfrac{6\lambda^4(2+\lambda)\Gamma(\lambda)^2\,{}_2F_1(2,2,3+\lambda,1)^3}{-2+\lambda} - 3\lambda^6\Gamma(\lambda)^3\,{}_2F_1(2,2,3+\lambda,1)^4 \\
& +28\,{}_2F_1(3,3,4+\lambda,1) + 216\,{}_2F_1(4,4,5+\lambda,1) + 576\,{}_2F_1(5,5,6+\lambda,1)
\end{aligned} . \tag{A.19}$$

Waring (b, n)

Raw moments:

$$\begin{aligned}
\mu'_1 = {} & bn\Gamma(b+n)\,{}_2F_1(2,n+1,b+n+2,1) \\
\mu'_2 = {} & \tfrac{bn(b+2n)\Gamma(b+n)\,{}_2F_1(2,n+1,b+n+2,1)}{b-2} \\
\mu'_3 = {} & \tfrac{bn\left(b+b^2+6bn+6n^2\right)\Gamma(b+n)\,{}_2F_1(2,n+1,b+n+2,1)}{6-5b+b^2} \\
\mu'_4 = {} & bn\Gamma(b+n)\,(\,{}_2F_1(2,n+1,b+n+2,1) + 2(1+n)) \\
& \left(\tfrac{(15+7b+18n)\,{}_2F_1(3,n+2,b+n+3,1)}{b-3} + 12(n+2)(n+3)\,{}_2F_1(5,n+4,b+n+5,1)\right)
\end{aligned} . \tag{A.20}$$

Central moments:

$$\begin{aligned}
\mu_1 = \;& 0 \\
\mu_2 = \;& bn\Gamma(b+n)\,{}_2F_1(2,1+n,2+b+n,1)\,\big(\tfrac{b+2n}{b-2} - bn\Gamma(b+n) \\
& {}_2F_1(2,1+n,2+b+n,1) \\
\mu_3 = \;& \Big(\tfrac{b+b^2+6bn+6n^2}{6-5b+b^2} + bn\Gamma(b+n)\,{}_2F_1(2,1+n,2+b+n,1) \\
& \Big(-\tfrac{3(b+2n)}{-2+b} + 2bn\Gamma(b+n)\,{}_2F_1(2,1+n,2+b+n,1)\Big) \qquad . \\
\mu_4 = \;& bn\Gamma[b+n]\,(\,{}_2F_1(2,1+n,2+b+n,1) \\
& -\tfrac{4bn\left(b+b^2+6bn+6n^2\right)\Gamma(b+n)\,{}_2F_1(2,1+n,2+b+n,1)^2}{6-5b+b^2} + \tfrac{6b^2n^2(b+2n)\Gamma(b+n)^2\,{}_2F_1(2,1+n,2+b+n,1)^3}{-2+b} \\
& -3b^3n^3\Gamma(b+n)^3\,{}_2F_1(2,1+n,2+b+n,1)^4 + 2(1+n) \\
& \Big(\tfrac{(15+7b+18n)\,{}_2F_1(3,2+n,3+b+n,1)}{-3+b} + 12(2+n)(3+n)\,{}_2F_1(5,4+n,5+b+n,1)\Big)
\end{aligned} \tag{A.21}$$

Beta-Binomial (a, b, n)

Raw moments:

$$\begin{aligned}
\mu_1' = \;& \frac{a\,n}{a+b} \\
\mu_2' = \;& \frac{a\,n\,(b+n+a\,n)}{(a+b)\,(a+b+1)} \\
\mu_3' = \;& \frac{a\,n\,(b\,(b-a) + 3(a+1)\,b\,n + (a+1)\,(a+2)\,n^2)}{(a+b)\,(a+b+1)\,(a+b+2)} \\
\mu_4' = \;& (a\,n\,(b\,(a^2+(b-1)b - a(4b+1)) + (a+1)b(7b-4a-1)n+ \\
& 6(a+1)(a+2)bn^2 + (a+1)(a+2)(a+3)n^3 \\
& ((a+b)(a+b+1)(a+b+2)(a+b+3))
\end{aligned} \tag{A.22}$$

Central moments:

$$\begin{aligned}
\mu_1 = {} & 0 \\
\mu^2 = {} & \frac{a\,b\,n(a+b+n)}{(a+b)^2(a+b+1)} \\
\mu^3 = {} & \frac{a(a-b)\,b\,n\,(a+b+n)(a+b+2n)}{(a+b)^3}(a+b+1)(a+b+2) \\
\mu^4 = {} & (a\,b\,n\,\left((a+b)^3\left(a^2+(b-1)b-a(1+4b)\right)+\right. \\
& (a+b)^2\left(a^2(7+3b)+b(7b-1)+a(-1+b(3b-10)\right)n+ \\
& 6\left(2a^2b^2+2b^3+ab^3+a^3(2+b)\right)n^2+3\left(a(b-2)b+2b^2+a^2(2+b)\right)n^3/ \\
& ((a+b)^4(1+a+b)(2+a+b)(3+a+b))
\end{aligned} \quad . \tag{A.23}$$

Component Mix Distributions

Zero-inflated Poisson (ω, μ)

Raw moments:

$$\begin{aligned}
\mu'_1 &= \mu(\omega-1) \\
\mu'_2 &= \mu(1+\mu)(\omega-1) \\
\mu'_3 &= \mu(1+\mu(\mu+3))(\omega-1) \\
\mu'_4 &= \mu(1+\mu(7+\mu(\mu+6)))(\omega-1)
\end{aligned} \quad , \tag{A.24}$$

Central moments:

$$\begin{aligned}
\mu_1 = {} & 0 \\
\mu_2 = {} & \mu(-1)(1+\mu\,\omega) \\
\mu_3 = {} & \mu(\omega-1)(1+\mu\omega(3+\mu(-1+2\omega))) \\
\mu_4 = {} & \mu(\omega-1)\left(1+\mu\left(3+\omega\left(4+6\mu\omega+\mu^2(1+3(\omega-1)\omega)\right)\right)\right)
\end{aligned} \quad . \tag{A.25}$$

Zero-inflated Negative Binomial (ω, p, r)

Raw moments:

$$\begin{aligned}
\mu'_1 &= \tfrac{(p-1)r(\omega-1)}{p} \\
\mu'_2 &= -\tfrac{(p-1)r((p-1)r-1)(\omega-1)}{p^2} \\
\mu'_3 &= \tfrac{(p-1)r\left(2-p-3(p-1)r+(p-1)^2r^2\right)(\omega-1)}{p^3} \\
\mu'_4 &= -\tfrac{(p-1)r\left(p^3r^3+3p(r+1)^2(r+2)-(r+1)(r+2)(r+3)-p^2(r+1)(1+3r(r+1))\right)(\omega-1)}{p^4}
\end{aligned} \quad , \tag{A.26}$$

Central moments:

$$
\begin{aligned}
\mu_1 \quad &= 0 \\
\mu_2 \quad &= -\tfrac{(p-1)r(\omega-1)(-1+(p-1)r\omega)}{p^2} \\
\mu_3 \quad &= \tfrac{(p-1)r(\omega-1)\left(2-p-(p-1)r(3+(p-1)r)\omega+2(p)-1^2r^2\omega^2\right)}{p^3} \\
\mu_4 \quad &= \tfrac{1}{p^4}(p-1)r\left(p^3r^3\omega(1+\omega(-4-3(\omega-2)\omega)) + p^2(\omega-1)\right. \quad \cdot \\
&\quad (1+r\omega\left(4+6r\omega+r^2(3+9(\omega-1)\omega)\right)) - 3p(\omega-1) \\
&\quad (2+r\left(1+\omega\left(4+4r\omega+r^2(1+3(\omega-1)\omega)\right)\right)) + (\omega-1) \\
&\quad (6+r\left(3+\omega\left(8+6r\omega+r^2(1+3(\omega-1)\omega)\right)\right))
\end{aligned}
\tag{A.27}
$$

Zero-inflated Sichel $(\omega, \alpha, \theta, \gamma)$

Raw moments:

$$
\begin{aligned}
\mu_1' = \quad & \tfrac{\alpha\theta(\omega-1)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{2\sqrt{1-\theta}K_\gamma(\alpha\sqrt{1-\theta})} \\
\mu_2' = \quad & \tfrac{\theta(\omega-1)\left(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta-\frac{2\alpha\sqrt{1-\theta}(1+\gamma\theta)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{K}{}_{\gamma}(\alpha\sqrt{1-\theta})\right.}{}) 4(\theta-1)^2 \\
\mu_3' = \quad & \tfrac{1}{8(\theta-1)^3}\theta(\omega-1)\left(24\gamma^2\theta+2\alpha^2(\theta-3)(\theta-1)\theta+8\gamma^3\theta^2\right. \\
& +\gamma\left(8+4\theta\left(2-\alpha^2(\theta-1)\theta\right)\right) + \tfrac{\alpha\sqrt{1-\theta}\left(4+\theta\left(4+12\gamma+4\gamma^2\theta-\alpha^2(\theta-1)\theta\right)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})K_\gamma(\alpha\sqrt{1-\theta})}{}) \\
\mu_4' = \quad & \left(\theta(\omega-1)\left(\tfrac{1}{\sqrt{1-\theta}}\alpha\left(\alpha^6(\theta-1)^3\theta^3+16\alpha^2(\gamma-1)(\theta-1)\right.\right.\right. \\
& (2+\theta\left(-6+21\gamma+6(1+\gamma(4\gamma-3))\theta+(-2+\gamma(7+\gamma(5\gamma-9)))\theta^2\right)) \\
& -4\alpha^4(\theta-1)^2\theta(7+\theta(7(\theta-2)+6\gamma(3+(\gamma-2)\theta))) \\
& -64(\gamma-2)(\gamma-1)\gamma(1+\theta(4+\theta+\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))) \\
& K_{\gamma-4}(\alpha\sqrt{1-\theta}) - \tfrac{1}{\theta-1}4\left(\alpha^6(\theta-1)^3\theta^2(3-3\theta+2\gamma\theta)\right. \\
& +8\alpha^2(\gamma-2)(\gamma-1)(\theta-1)\left(3+\theta\left(-9+28\gamma+\left(9-26\gamma+30\gamma^2\right)\theta\right.\right. \\
& +(-3+2\gamma(5+3(\gamma-2)\gamma))\theta^2 - 2\alpha^4(\theta-1)^2 \\
& (1+\theta\left(-24+21\gamma+45\theta+6\gamma(6\gamma-13)\theta+(-22+\gamma(47+2\gamma(-18+5\gamma)))\theta^2\right)) \\
& -32(\gamma-3)(\gamma-2)(\gamma-1)\gamma(1+\theta(4+\theta+\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))) \\
& K_{\gamma-3}(\alpha\sqrt{1-\theta})/\left(16\alpha^3(1-\theta)^{9/2}K_\gamma(\alpha\sqrt{1-\theta})\right)
\end{aligned}
\quad ,
\tag{A.28}
$$

Central moments:

$$
\begin{aligned}
\mu_1 = {}& 0 \\
\mu_2 = {}& \tfrac{1}{4(\theta-1)^2}\theta(\omega-1)\left(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\right. \\
& +\tfrac{1}{K_\gamma(\alpha\sqrt{1-\theta})^2}\alpha\left(-2\sqrt{1-\theta}(1+\gamma\theta)K_{\gamma-1}(\alpha\sqrt{1-\theta})K_\gamma(\alpha\sqrt{1-\theta})\right. \\
& +\alpha(\theta-1)\theta(\omega-1)K_{\gamma+1}(\alpha\sqrt{1-\theta})^2 \\
\mu_3 = {}& \tfrac{1}{8}\theta(\omega-1)\left(\tfrac{24\gamma^2\theta+2\alpha^2(\theta-3)(\theta-1)\theta+8\gamma^3\theta^2+\gamma\left(8+4\theta\left(2-\alpha^2(\theta-1)\theta\right)\right)}{(\theta-1)^3}\right. \\
& +\tfrac{\alpha\left(-4+\theta\left(-4-12\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\right)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{(1-\theta)^{5/2}K_\gamma(\alpha\sqrt{1-\theta})} \\
& +\left(\alpha^2\theta(\omega-1)K_{\gamma+1}(\alpha\sqrt{1-\theta})\right)\left(3\alpha(\theta-1)\theta K_\gamma(\alpha\sqrt{1-\theta})^2\right. \\
& -6\sqrt{1-\theta}(1+\gamma\theta)K_\gamma(\alpha\sqrt{1-\theta})K_{\gamma+1}(\alpha\sqrt{1-\theta}) \\
& +2\alpha(\theta-1)\theta(\omega-1)K_{\gamma+1}(\alpha\sqrt{1-\theta})^2/\left((1-\theta)^{5/2}K_\gamma(\alpha\sqrt{1-\theta})^3\right) \\
\mu_4 = {}& \Big(\theta(\omega-1)\Big(\tfrac{1}{(1-\theta)^{9/2}}\Big(\tfrac{1}{\sqrt{1-\theta}}\alpha\left(\alpha^6(\theta-1)^3\theta^3\right. \\
& +16\alpha^2(\gamma-1)(\theta-1)\left(2+\theta\left(-6+21\gamma+6(1+\gamma(4\gamma-3))\theta\right.\right. \\
& +(-2+\gamma(7+\gamma(5\gamma-9)))\theta^2-4\alpha^4(\theta-1)^2\theta(7+\theta(7(\theta-2) \qquad . \\
& +6\gamma(3+(\gamma-2)\theta)))-64(\gamma-2)(\gamma-1)\gamma(1+\theta(4+\theta \\
& +\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))K_{\gamma-4}(\alpha\sqrt{1-\theta}) \\
& -\tfrac{1}{\theta-1}4\left(\alpha^6(\theta-1)^3\theta^2(3+(2\gamma-3)\theta)+8\alpha^2(\gamma-2)(\gamma-1)(\theta-1)\right. \\
& \left(3+\theta\left(-9+28\gamma+\left(9-26\gamma+30\gamma^2\right)\theta+(-3+2\gamma(5+3(\gamma-2)\gamma))\theta^2\right)\right) \\
& -2\alpha^4(\theta-1)^2\left(1+\theta\left(-24+21\gamma+45\theta+6\gamma(6\gamma-13)\theta\right.\right. \\
& +(-22+\gamma(47+2\gamma(-18+5\gamma)))\theta^2-32(\gamma-3)(\gamma-2)(\gamma-1)\gamma \\
& (1+\theta(4+\theta+\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))K_{\gamma-3}(\alpha\sqrt{1-\theta})K_\gamma(\alpha\sqrt{1-\theta})^3 \\
& -\tfrac{1}{(1-\theta)^{7/2}}4\alpha^4\theta(\omega-1)\left(24\gamma^2\theta+2\alpha^2(\theta-3)(\theta-1)\theta+8\gamma^3\theta^2\right. \\
& +\gamma\left(8+4\theta\left(2-\alpha^2(\theta-1)\theta\right)\right)+\tfrac{\alpha\sqrt{1-\theta}\left(4+\theta\left(4+12\gamma+4\gamma^2\theta-\alpha^2(\theta-1)\theta\right)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{K_\gamma(\alpha\sqrt{1-\theta})} \\
& K_\gamma(\alpha\sqrt{1-\theta})^3K_{\gamma+1}(\alpha\sqrt{1-\theta})-\tfrac{1}{(\theta-1)^3}6\alpha^5\theta^2(\omega-1)^2K_\gamma(\alpha\sqrt{1-\theta}) \\
& \left(-2\alpha\sqrt{1-\theta}(\gamma\theta+1)K_{\gamma-1}(\alpha\sqrt{1-\theta})-\left(4\gamma+4\gamma^2\theta-\alpha^2(\theta-1)\theta\right)K_\gamma(\alpha\sqrt{1-\theta})\right) \\
& K_{\gamma+1}(\alpha\sqrt{1-\theta})^2-\tfrac{3\alpha^7\theta^3(\omega-1)^3K_{\gamma+1}(\alpha\sqrt{1-\theta})^4}{(\theta-1)^2}/\left(16\alpha^3K_\gamma(\alpha\sqrt{1-\theta})^4\right)
\end{aligned}
\tag{A.29}
$$

2-component Poisson Mixture (ω, μ, λ)

Raw moments:

$$\begin{aligned}
\mu_1' = &\quad \lambda - \lambda\omega + \mu\omega \\
\mu_2' = &\quad -\lambda(\lambda+1)(\omega-1) + \mu(\mu+1)\omega \\
\mu_3' = &\quad \lambda(1+\lambda(\lambda+3)) + (-\lambda(1+\lambda(\lambda+3)) + \mu + 3\mu^2 + \mu^3)\omega \quad . \\
\mu_4' = &\quad \lambda(1+\lambda(7+\lambda(6+\lambda))) + (-\lambda(1+\lambda(7+\lambda(6+\lambda))) \\
&\quad +\mu(1+\mu(7+\mu(6+\mu))))\omega
\end{aligned} \tag{A.30}$$

Central moments:

$$\begin{aligned}
\mu_1 &= 0 \\
\mu_2 &= \lambda + (\lambda-\mu-1)(\lambda-\mu)\omega - (\lambda-\mu)^2\omega^2 \\
\mu_3 &= \lambda - (\lambda-\mu)\left(1+\lambda^2+\mu(3+\mu) - \lambda(3+2\mu)\right)\omega \\
&\quad +3(\lambda-\mu-1)(\lambda-\mu)^2\omega^2 - 2(\lambda-\mu)^3\omega^3 \qquad . \\
\mu_4 &= \lambda(3\lambda+1) + (\lambda-\mu)\left(\lambda^3 - 3\lambda^2\mu + \lambda(1+3\mu(\mu+2)-1)\right. \\
&\quad -\mu(7+\mu(6+\mu))\omega - 2(\lambda-\mu)^2\left(2+\lambda(2\lambda-3)\right. \\
&\quad +6\mu - 4\lambda\mu + 2\mu^2\omega^2 + 6(\lambda-\mu-1)(\lambda-\mu)^3\omega^3 - 3(\lambda-\mu)^4\omega^4
\end{aligned} \tag{A.31}$$

2-component Poisson-Negative Binomial Mixture (ω, μ, r, p)

Raw moments:

$$\begin{aligned}
\mu_1' = &\quad \tfrac{(p-1)r(\omega-1)}{p} + \mu\omega \\
\mu_2' = &\quad \tfrac{(p-1)r(\omega-1)-(p-1)^2r^2(\omega-1)+p^2\mu(\mu+1)\omega}{p^2} \\
\mu_3' = &\quad \tfrac{1}{p^3}\left(-(p-2)(p-1)r(\omega-1) - 3(p-1)^2r^2(\omega-1) + (p-1)^3r^3(\omega-1)\right. \\
&\quad +p^3\mu(1+\mu(\mu+3))\omega \qquad , \\
\mu_4' = &\quad \tfrac{1}{p^4}\left((p-1)(6+(p-6)p)r(\omega-1) + (p-1)^2(4p-11)r^2(\omega-1)\right. \\
&\quad +6(p-1)^3r^3(\omega-1) - (p-1)^4r^4(\omega-1) + p^4\mu(1+\mu(7+\mu(6+\mu)))\omega
\end{aligned} \tag{A.32}$$

Central moments:

$$
\begin{aligned}
\mu_1 = {} & 0 \\
\mu_2 = {} & \left(\mu - (r+\mu)^2(\omega-1)\right)\omega - \tfrac{r(\omega-1)(r\omega+1)}{p^2} + \tfrac{r(\omega-1)(1+2(r+\mu)\omega)}{p} \\
\mu_3 = {} & \tfrac{1}{p^3}\left((p-1)^3 r^3(\omega-1)\omega(2\omega-1) + 3(p-1)^2 r^2(-1+\omega)\omega(-1+p\mu(-1+2\omega))\right. \\
& +(-1+p)r(\omega-1)\left(2+p\left(-1-3\mu(1+p+p\mu)\omega+6p\mu^2\omega^2\right)\right) \\
& \left.+p^3\mu\omega(1+\mu(\omega-1)(\mu(2\omega-1)-3)\right) \\
\mu_4 = {} & -\tfrac{1}{p^4}\left((-p)^4 r^4(\omega-1)\omega(1+3(\omega-1)\omega) + 2(p-1)^3 r^3(\omega-1)\omega(-3\omega\right. \\
& +2p\mu(1+3(\omega-1)\omega)) + p^4\mu\omega\left(-1+\mu\left(-7+4\omega+\mu(\omega-1)\right.\right. \\
& (6+\mu-3(\mu+2)\omega+3\mu\omega^2) + (p-1)^2 r^2(\omega-1)(3+2\omega(4+p(-2-6\mu\omega \\
& +3p\mu(1-\omega+\mu(1+3(\omega-1)\omega))))) + (p-1)r(\omega-1)\left(-6+p\left(6+8\mu\omega\right.\right. \\
& \left.\left.+p\left(-1+2\mu\omega\left(-2-3\mu\omega+2p\left(1+\mu\left(3+\mu-3(1+\mu)\omega+3\mu\omega^2\right)\right)\right)\right)\right)\right)
\end{aligned}
\quad . \tag{A.33}
$$

Truncated Distributions

Positive Poisson (μ)

Raw moments:

$$
\begin{aligned}
\mu_1' &= \left(1+\tfrac{1}{\mathrm{e}^\mu-1}\right)\mu \\
\mu_2' &= \tfrac{\mathrm{e}^\mu\mu(\mu+1)}{\mathrm{e}^\mu-1} \\
\mu_3' &= \tfrac{\mathrm{e}^\mu\mu(\mu(\mu+3)+1)}{\mathrm{e}^\mu-1} \\
\mu_4' &= \tfrac{\mathrm{e}^\mu\mu(\mu(7+\mu(6+\mu))+1)}{\mathrm{e}^\mu-1}
\end{aligned}
\quad , \tag{A.34}
$$

Central moments:

$$
\begin{aligned}
\mu_1 &= 0 \\
\mu_2 &= \tfrac{\mathrm{e}^\mu(\mathrm{e}^\mu-\mu-1)\mu}{(\mathrm{e}^\mu-1)^2} \\
\mu_3 &= \tfrac{\mathrm{e}^\mu\mu\left(1+e^{2\mu}+\mu(\mu+3)+\mathrm{e}^\mu((\mu-3)\mu-2)\right)}{(\mathrm{e}^\mu-1)^3} \\
\mu_4 &= \tfrac{\mathrm{e}^\mu\mu\left(\mathrm{e}^{3\mu}(3\mu+1)-1-\mathrm{e}^{2\mu}\left(13\mu+\mu^3+3\right)-\mu(7+\mu(\mu+6))+\mathrm{e}^\mu(3+\mu(17-(\mu-6)\mu))\right)}{(\mathrm{e}^\mu-1)^4}
\end{aligned}
\quad . \tag{A.35}
$$

Positive Geometric (p)

Raw moments:

$$\begin{aligned} \mu_1' &= \tfrac{1}{p} \\ \mu_2' &= \tfrac{2-p}{p^2} \\ \mu_3' &= \tfrac{6+(p-6)p}{p^3} \\ \mu_4' &= -\tfrac{(p-2)(12+(p-12)p)}{p^4} \end{aligned} \quad . \tag{A.36}$$

Central moments:

$$\begin{aligned} \mu_1 &= 0 \\ \mu_2 &= \tfrac{1-p}{p^2} \\ \mu_3 &= \tfrac{(p-2)(p-1)}{p^3} \\ \mu_4 &= -\tfrac{(p-1)(9+(p-9)p)}{p^4} \end{aligned} \quad . \tag{A.37}$$

Positive Negative Binomial (r, p)

Raw moments:

$$\begin{aligned} \mu_1' &= \tfrac{(p-1)r}{p(p^r-1)} \\ \mu_2' &= -\tfrac{(p-1)r((p-1)r-1)}{p^2(p^r-1)} \\ \mu_3' &= \tfrac{r\left(3p-p^2-3(p-1)^2r+(p-1)^3r^2-2\right)}{p^3(p^r-1)} \\ \mu_4' &= \tfrac{1}{p^4(p^r-1)}r\left(12p-7p^2+p^3+(p-1)^2(4p-11)r+6(p-1)^3r^2-(p-1)^4r^3-6\right) \end{aligned} \quad ,$$

Central moments:

$$\begin{aligned} \mu_1 = {} & 0 \\ \mu_2 = {} & -\tfrac{(p-1)r(1+p^r((p-1)r-1))}{p^2(p^r-1)^2} \\ \mu_3 = {} & \tfrac{r\left(2(p-1)^3r^2+3(p-1)^2(p^r-1)r((p-1)r-1)+(p^r-1)^2\left(+3p-p^2-3(-1+p)^2r+(-1+p)^3r^2-2\right)\right)}{p^3(p^r-1)^3} \\ \mu_4 = {} & \tfrac{1}{p^4(p^r-1)^4}r\left(-3(p-1)^4r^3-6(p-1)^3\left(p^r-1\right)r^2((p-1)r-1)\right. \\ & -4(p-1)\left(p^r-1\right)^2r\left(3p-p^2-3(p-1)^2r+(p-1)^3r^2-2\right) \\ & +\left(p^r-1\right)^3\left(-6+12p-7p^2+p^3+(p-1)^2(4p-11)r+6(p-1)^3r^2\right. \\ & -(p-1)^4r^3 \end{aligned} \quad .$$

Positive Holla (α, θ)

Raw moments:

$$\begin{aligned}
\mu_1' &= \frac{\mathrm{e}^{\alpha}\alpha\theta}{\left(2\mathrm{e}\alpha-2\mathrm{e}^{\alpha\sqrt{1-\theta}}\right)\sqrt{1-\theta}}\\
\mu_2' &= \frac{\mathrm{e}^{\alpha}\alpha\theta\left(2+\left(\alpha\sqrt{1-\theta}-1\right)\theta\right)}{4\left(\mathrm{e}^{\alpha}-\mathrm{e}^{\alpha\sqrt{1-\theta}}\right)(1-\theta)^{\frac{3}{2}}}\\
\mu_3' &= -\frac{\mathrm{e}^{\alpha}\alpha\theta\left(-4+\theta\left(2+3\alpha\sqrt{1-\theta}(\theta-2)-\theta+\alpha^2(\theta-1)\theta\right)\right)}{8\left(\mathrm{e}^{\alpha}-\mathrm{e}^{\alpha\sqrt{1-\theta}}\right)(1-\theta)^{\frac{5}{2}}}\\
\mu_4' &= -\frac{\mathrm{e}^{\alpha}\alpha\theta\left(-8+\theta\left(-4+(-4+\theta)\theta-6\alpha^2(-2+\theta)(-1+\theta)\theta-\alpha^3(1-\theta)^{\frac{3}{2}}\theta^2+\alpha\sqrt{1-\theta}(-28+(20-7\theta)\theta)\right)\right)}{16\left(\mathrm{e}^{\alpha}-\mathrm{e}^{\alpha\sqrt{1-\theta}}\right)(1-\theta)^{\frac{7}{2}}}
\end{aligned} \quad , \tag{A.40}$$

Central moments:

$$\begin{aligned}
\mu_1 = {} & 0\\
\mu_2 = {} & -\frac{e^{\alpha}\alpha\theta\left(e^{\alpha}(-2+\theta)+e^{\alpha\sqrt{1-\theta}}\left(2+\left(-1+\alpha\sqrt{1-\theta}\right)\theta\right)\right)}{4\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)^2(1-\theta)^{3/2}}\\
\mu_3 = {} & \frac{1}{8\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)^3(1-\theta)^{5/2}}\left(e^{\alpha}\alpha\theta\left(e^{2\alpha}(4+(-2+\theta)\theta)\right.\right.\\
& +e^{\alpha+\alpha\sqrt{1-\theta}}\left(-8+\theta\left(4+3\alpha\sqrt{1-\theta}(-2+\theta)-2\theta-\alpha^2(-1+\theta)\theta\right)\right)\\
& +e^{2\alpha\sqrt{1-\theta}}\left(4+\theta\left(-2-3\alpha\sqrt{1-\theta}(-2+\theta)+\theta-\alpha^2(-1+\theta)\theta\right)\right)\\
\mu_4 = {} & -\frac{1}{16\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)^4}e^{\alpha}\alpha\theta\Bigg(\frac{3e^{3\alpha}\alpha^3\theta^3}{(-1+\theta)^2}-\frac{6e^{2\alpha}\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)\alpha^2\theta^2\left(2+\left(-1+\alpha\sqrt{1-\theta}\right)\theta\right)}{(1-\theta)^{5/2}}\\
& +\frac{4e^{\alpha}\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)^2\alpha\theta\left(-4+\theta\left(2+3\alpha\sqrt{1-\theta}(-2+\theta)-\theta+\alpha^2(-1+\theta)\theta\right)\right)}{(-1+\theta)^3}\\
& +\frac{\left(e^{\alpha}-e^{\alpha\sqrt{1-\theta}}\right)^3\left(-8+\theta\left(-4+(-4+\theta)\theta-6\alpha^2(-2+\theta)(-1+\theta)\theta-\alpha^3(1-\theta)^{3/2}\theta^2+\alpha\sqrt{1-\theta}(-28+(20-7\theta)\theta)\right)\right)}{(1-\theta)^{7/2}}
\end{aligned} \quad . \tag{A.41}$$

Positive Sichel (α, θ, γ)

Raw moments:

$$
\begin{aligned}
\mu_1' = \;& -\frac{\alpha\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})}{\sqrt{1-\theta}\Big(2(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)-2K_\gamma(\alpha\sqrt{1-\theta})\Big)} \\
\mu_2' = \;& \Bigg(\theta\Bigg(\frac{\Big(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\Big)K_{\gamma-2}(\alpha\sqrt{1-\theta})}{(\theta-1)^2} \\
& +\frac{2\Big(-4(\gamma-1)\gamma(1+\gamma\theta)+\alpha^2(\theta-1)(1+(2\gamma-1)\theta)\Big)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{\alpha(1-\theta)^{\frac{5}{2}}}\Bigg)\Bigg)/ \\
& \Big(4\left((1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)-K_\gamma(\alpha\sqrt{1-\theta})\right)\Big) \\
\mu_3' = \;& (\theta\,(\alpha\,(\alpha^4(\theta-1)^2\theta^2-4\alpha^2(\theta-1)(1+\theta(-2+\theta+3\gamma(2+(\gamma-1)\theta))) \\
& +16(\gamma-1)\gamma(1+\theta(1+\gamma(3+\gamma\theta)))K_{\gamma-3}(\alpha\sqrt{1-\theta}) \\
& +\tfrac{1}{\sqrt{1-\theta}}2\,(3\alpha^4(\theta-1)^2\theta(1+(\gamma-1)\theta)-4\alpha^2(\gamma-1)(\theta-1) \\
& (2+\theta(-4+9\gamma+(2+\gamma(-5+4\gamma))\theta))+16(\gamma-2)(\gamma-1)\gamma(1+\theta(1+\gamma(3+\gamma\theta))) \\
& K_{\gamma-2}(\alpha\sqrt{1-\theta})/\left(8\alpha^2(1-\theta)^{7/2}\left(-(1-\theta)^{\gamma/2}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right)\right), \\
\mu_4' = \;& \Big(\theta\,\Big(\tfrac{1}{\sqrt{1-\theta}}\alpha\,(\alpha^6(\theta-1)^3\theta^3+16\alpha^2(\gamma-1)(\theta-1)\,(2+\theta\,(-6+21\gamma \\
& +6(1+\gamma(4\gamma-3))\theta+(-2+\gamma(7+\gamma(-9+5\gamma)))\theta^2 \\
& -4\alpha^4(\theta-1)^2\theta(7+\theta(7(\theta-2)+6\gamma(3+(\gamma-2)\theta))) \\
& -64(\gamma-2)(\gamma-1)\gamma(1+\theta(4+\theta+\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))) \\
& K_{\gamma-4}(\alpha\sqrt{1-\theta})-\tfrac{1}{\theta-1}4\,(\alpha^6(\theta-1)^3\theta^2(3-3\theta+2\gamma\theta) \\
& +8\alpha^2(\gamma-2)(\gamma-1)(\theta-1)\,(3+\theta\,(-9+28\gamma+(9-26\gamma+30\gamma^2)\,\theta \\
& +(-3+2\gamma(5+3(\gamma-2)\gamma))\theta^2-2\alpha^4(\theta-1)^2\,(1+\theta\,(-24+21\gamma+45\theta \\
& +6\gamma(-13+6\gamma)\theta+(-22+\gamma(47+2\gamma(-18+5\gamma)))\theta^2 \\
& -32(\gamma-3)(\gamma-2)(\gamma-1)\gamma(1+\theta(4+\theta+\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))) \\
& K_{\gamma-3}(\alpha\sqrt{1-\theta})/\left(16\alpha^3(1-\theta)^{9/2}\left((1-\theta)^{\gamma/2}K_\gamma(\alpha)-K_\gamma(\alpha\sqrt{1-\theta})\right)\right)
\end{aligned}
\tag{A.42}
$$

Central moments:

$$
\begin{aligned}
\mu_1 = {} & 0 \\
\mu_2 = {} & \tfrac{1}{4}\theta \Bigg(\frac{\frac{\left(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\right)K_{\gamma-2}(\alpha\sqrt{1-\theta})}{(\theta-1)^2}+\frac{2\left(-4(\gamma-1)\gamma(\gamma\theta+1)+\alpha^2(\theta-1)(1+(2\gamma-1)\theta)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})}{\alpha(1-\theta)^{5/2}}}{(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)-K_\gamma(\alpha\sqrt{1-\theta})} \\
& +\frac{\alpha^2\theta K_{\gamma+1}(\alpha\sqrt{1-\theta})^2}{(\theta-1)\left(-(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right)^2} \\
\mu_3 = {} & \tfrac{1}{8\alpha^2(1-\theta)^{7/2}}\,\theta\left(\left(\alpha\left(\alpha^4(\theta-1)^2\theta^2-4\alpha^2(\theta-1)(1+\theta(-2+\theta+3\gamma(2+(\gamma-1)\theta)))\right.\right.\right. \\
& +16(\gamma-1)\gamma(1+\theta(1+\gamma(3+\gamma\theta)))K_{\gamma-3}(\alpha\sqrt{1-\theta}) \\
& +\tfrac{1}{\sqrt{1-\theta}}2\left(3\alpha^4(\theta-1)^2\theta(1+(\gamma-1)\theta)\right. \qquad , \\
& -4\alpha^2(\gamma-1)(\theta-1)(2+\theta(-4+9\gamma+(2+\gamma(4\gamma-5))\theta)) \\
& +16(\gamma-2)(\gamma-1)\gamma(1+\theta(1+\gamma(3+\gamma\theta)))K_{\gamma-2}(\alpha\sqrt{1-\theta})/ \\
& \left(-(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right) \\
& +\left(3\alpha^2\sqrt{1-\theta}\theta\left(\alpha\sqrt{1-\theta}\left(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\right)K_{\gamma-2}(\alpha\sqrt{1-\theta})\right.\right. \\
& +2\left(-4(\gamma-1)\gamma(1+\gamma\theta)+\alpha^2(\theta-1)(1+(2\gamma-1)\theta)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta}) \\
& K_{\gamma+1}(\alpha\sqrt{1-\theta})/\left(-(1-\theta)^{\frac{\gamma}{2}}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right)^2-\tfrac{2\alpha^5(\theta-1)^2\theta^2K_{\gamma+1}(\alpha\sqrt{1-\theta})^3}{\left((1-\theta)^{\gamma/2}K_\gamma(\alpha)-K_\gamma(\alpha\sqrt{1-\theta})\right)^3}
\end{aligned}
\tag{A.43}
$$

$$
\begin{aligned}
\mu_4 = \ & \tfrac{1}{16\alpha^3}\theta \Big(\Big(\tfrac{1}{\sqrt{1-\theta}}\alpha \left(\alpha^6(\theta-1)^3\theta^3 + 16\alpha^2(\gamma-1)(\theta-1)\right.\left(2+\theta\left(-6+21\gamma\right.\right.\\
& +6(1+\gamma(4\gamma-3))\theta + (-2+\gamma(7+\gamma(-9+5\gamma)))\theta^2 - 4\alpha^4(\theta-1)^2\theta(7+\theta(7(\theta-2)\\
& +6\gamma(3+(\gamma-2)\theta))) - 64(\gamma-2)(\gamma-1)\gamma(1+\theta(4+\theta+\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))\\
& K_{\gamma-4}(\alpha\sqrt{1-\theta}) - \tfrac{1}{\theta-1}4\left(\alpha^6(\theta-1)^3\theta^2(3+(2\gamma-3)\theta) + 8\alpha^2(\gamma-2)(\gamma-1)(\theta-1)\right.\\
& \left(3+\theta\left(-9+28\gamma+\left(9-26\gamma+30\gamma^2\right)\theta + (-3+2\gamma(5+3(\gamma-2)\gamma))\theta^2\right)\right)\\
& -2\alpha^4(\theta-1)^2\left(1+\theta\left(-24+21\gamma+45\theta+6\gamma(6\gamma-13)\theta+(\gamma(47+2\gamma(5\gamma-18))\right.\right.\\
& -22)\theta^2 - 32(\gamma-3)(\gamma-2)(\gamma-1)\gamma(1+\theta(4+\theta+\gamma(7+\theta(4+\gamma(6+\gamma\theta)))))\\
& K_{\gamma-3}(\alpha\sqrt{1-\theta})/\left((1-\theta)^{9/2}\left((1-\theta)^{\gamma/2}K_\gamma(\alpha) - K_\gamma(\alpha\sqrt{1-\theta})\right)\right)\\
& -\left(4\alpha^2\theta\left(\alpha\left(\alpha^4(\theta-1)^2\theta^2 - 4\alpha^2(\theta-1)(1+\theta(-2+\theta+3\gamma(2+(\gamma-1)\theta)))\right.\right.\right.\\
& +16(\gamma-1)\gamma(1+\theta(1+\gamma(3+\gamma\theta)))K_{\gamma-3}(\alpha\sqrt{1-\theta})\\
& +\tfrac{1}{\sqrt{1-\theta}}2\left(3\alpha^4(\theta-1)^2\theta(1+(\gamma-1)\theta) - 4\alpha^2(\gamma-1)(\theta-1)(2+\theta(-4+9\gamma\right.\\
& +(2+\gamma(4\gamma-5))\theta)) + 16(\gamma-2)(\gamma-1)\gamma(1+\theta(1+\gamma(3+\gamma\theta)))K_{\gamma-2}(\alpha\sqrt{1-\theta})\\
& K_{\gamma+1}(\alpha\sqrt{1-\theta})/\Big((\theta-1)^4\left(-(1-\theta)^{\gamma/2}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right)^2\Big)\\
& +\left(6\alpha^4\theta^2\left(\alpha\sqrt{1-\theta}\left(-4\gamma-4\gamma^2\theta+\alpha^2(\theta-1)\theta\right)K_{\gamma-2}(\alpha\sqrt{1-\theta})\right.\right.\\
& +2\left(-4(\gamma-1)\gamma(1+\gamma\theta)+\alpha^2(\theta-1)(1+(2\gamma-1)\theta)\right)K_{\gamma-1}(\alpha\sqrt{1-\theta})\\
& K_{\gamma+1}(\alpha\sqrt{1-\theta})^2/\Big((1-\theta)^{7/2}\left((1-\theta)^{\gamma/2}K_\gamma(\alpha)-K_\gamma(\alpha\sqrt{1-\theta})\right)^3\Big)\\
& -\tfrac{3\alpha^7\theta^3K_{\gamma+1}(\alpha\sqrt{1-\theta})^4}{(\theta-1)^2\left(-(1-\theta)^{\gamma/2}K_\gamma(\alpha)+K_\gamma(\alpha\sqrt{1-\theta})\right)^4}
\end{aligned}
\quad . \tag{A.44}
$$

Positive Yule (λ)

Raw moments:

$$
\begin{aligned}
\mu'_1 = \ & \tfrac{\lambda(\lambda+1)\,{}_2F_1(1,2,3+\lambda,1)}{\lambda^2+\lambda-2}\\
\mu'_2 = \ & \mathrm{B}(\lambda+1,2)\Gamma\lambda+3)\left(\tfrac{\lambda(\lambda+1)(\lambda+5)\,{}_2F_1(1,2,3+\lambda,1)}{\lambda-1} + 12(\lambda-3)\,{}_2F_1(3,4,4+\lambda,1)\right)\\
\mu'_3 = \ & \tfrac{1}{(\lambda-1)\lambda}\mathrm{B}(\lambda+1,2)\Gamma(\lambda+3)\left(\lambda^2(\lambda+1)(\lambda+13)\,{}_2\mathrm{F}_1(1,2,3+\lambda,1)\right.\\
& \left.+72((\lambda-3)(\lambda-1)\lambda\,{}_2F_1(3,4,4+\lambda,1)+2(\lambda-5)(\lambda-4)\,{}_2F_1(4,5,\lambda+4,1)\right)\\
\mu'_4 = \ & \Gamma(\lambda-1)\left(\lambda^2(\lambda+1)(\lambda+29)\,{}_2F_1(1,2,\lambda+3,1)\right.\\
& \left.+60(5(\lambda-3)(\lambda-1)\lambda\,{}_2F_1(3,4,\lambda+4,1)+24(\lambda-5)(\lambda-2)\,{}_2F_1(4,5,\lambda+4,1)\right)
\end{aligned}
\quad , \tag{A.45}
$$

Central moments:

$$\begin{aligned}
\mu_1 = &\quad 0 \\
\mu_2 = &\quad -\lambda^4(\lambda+1)^4\Gamma(\lambda-1)^2\,{}_2F_1(1,2,3+\lambda,1)^2+\mathrm{B}(\lambda+1,2)\Gamma(\lambda+3) \\
&\quad \left(\tfrac{\lambda(\lambda+1)(\lambda+5)\,{}_2F_1(1,2,\lambda+3,1)}{\lambda-1}+12(\lambda-3)\,{}_2F_1(3,4,\lambda+4,1)\right) \\
\mu_3 = &\quad -\tfrac{3\lambda^2(\lambda+1)(\lambda+5)\_2F_1(1,2,\lambda+3,1)^2}{(\lambda+\lambda^2-2)^2}+2\lambda^6(\lambda+1)^6\Gamma(\lambda-1)^3\,{}_2F_1(1,2,\lambda+3,1)^3 \\
&\quad +\tfrac{1}{\lambda-1}\left(-36(\lambda-3)\lambda\Gamma(\lambda+2)^2\,{}_2F_1(1,2,\lambda+3,1)\,{}_2F_1(3,4,\lambda+4,1)\right. \\
&\quad +\Gamma(\lambda)\left(\lambda^2(\lambda+1)(\lambda+13)\,{}_2F_1(1,2,\lambda+3,1)+72((\lambda-3)(\lambda-1)\lambda\right. \\
&\quad {}_2F_1(3,4,\lambda+4,1]+2(\lambda-5)(\lambda-4)\,{}_2F_1(4,5,\lambda+4,1)) \qquad . \\
\mu_4 = &\quad \Gamma(\lambda-1)\left(6\lambda^6(\lambda+1)^5(\lambda+5)\Gamma(\lambda-1)^2\,{}_2F_1(1,2,\lambda+3,1)^3\right. \\
&\quad -3\lambda^8(\lambda+1)^8\Gamma(\lambda-1)^3\,{}_2F_1(1,2,\lambda+3,1)^4+4\lambda^4(\lambda+1)^3\Gamma(\lambda-1) \\
&\quad {}_2F_1(1,2,\lambda+3,1)^2(-13-\lambda+18(\lambda-3)\Gamma(\lambda+2)\,{}_2F_1(3,4,\lambda+4,1)) \\
&\quad +60(5(\lambda-3)(\lambda-1)\lambda\,{}_2F_1(3,4,\lambda+4,1)+24(\lambda-5)(\lambda-2)\,{}_2F_1(4,5,\lambda+4,1)) \\
&\quad +\lambda^2(\lambda+1)\,{}_2F_1(1,2,\lambda+3,1)(29+\lambda+288(\lambda+1)\Gamma(\lambda-1)(-(\lambda-3)(\lambda-1)\lambda \\
&\quad {}_2F_1(3,4,\lambda+4,1)-2(\lambda-5)(\lambda-4)\,{}_2F_1(4,5,\lambda+4,1)))
\end{aligned} \tag{A.46}$$

Lerch Family Distributions

Lerch (p,a,c)

Raw moments:

$$\left.\begin{aligned}
\mu'_1 &= \tfrac{\Phi(p,c-1,a+1)-a\Phi(p,c,a+1)}{\Phi(p,c,a+1)} \\
\mu'_2 &= \tfrac{\Phi(p,c-2,a+1)-2aPhi(p,c-1,a+1)+a^2\Phi(p,c,a+1)}{\Phi(p,c,a+1)} \\
\mu'_3 &= \tfrac{\Phi(p,c-3,a+1)-3a\Phi(p,c-2,a+1)+3a^2\Phi(p,c-1,a+1)-a^3\Phi(p,c,a+1)}{\Phi(p,c,a+1)} \\
\mu'_4 &= \tfrac{\Phi(p,c-4,a+1)-4a\Phi(p,c-1,a+1)+6a^2\Phi(p,c-2,a+1)-4a^3\Phi(p,c-1,a+1)+a^4\Phi(p,c,a+1)}{\Phi(p,c,a+1)}
\end{aligned}\right. \quad , \tag{A.47}$$

Central moments:

$$\begin{aligned}
\mu_1 = \;& 0 \\
\mu_2 = \;& \tfrac{1}{\Phi(p,c,a+1)^2}\,(-(\Phi(p,c-1,a+1)-a\Phi(p,c,a+1))^2+(\Phi(p,c-2,a+1) \\
& +a(-2\Phi(p,c-1,a+1)+a\Phi(p,c,a+1)))\Phi(p,c,a+1) \\
\mu_3 = \;& \tfrac{1}{\Phi(p,c,a+1)^3}\,(2(\Phi(p,c-1,a+1)-a\Phi(p,c,a+1))^3-3(\Phi(p,c-1,a+1) \\
& -a\Phi(p,c,a+1))(\Phi(p,c-2,a+1)+a(-2\Phi(p,c-1,a+1)+a\Phi(p,c,a+1))) \\
& \Phi(p,c,a+1)+(\Phi(p,c-3,a+1)-a(3\Phi(p,c-2,a+1)+a(-3\Phi(p,c-1,a+1) \\
& +a\Phi(p,c,a+1))))\Phi(p,c,a+1)^2 \qquad . \\
\mu_4 = \;& \tfrac{1}{\Phi(p,c,a+1)^4}\,(-3(\Phi(p,c-1,a+1)-a\Phi(p,c,a+1))^4+6(\Phi(p,c-1,a+1) \\
& -a\Phi(p,c,a+1))^2(\Phi(p,c-2,a+1)+a(-2\Phi(p,c-1,a+1)+a(p,c,a+1))) \\
& \Phi(p,c,a+1)-4(\Phi(p,c-1,a+1)-a\Phi(p,c,a+1))(\Phi(p,c-3,a+1) \\
& -a(3\Phi(p,c-2,a+1)+a(-3\Phi(p,c-1,a+1)+a\Phi(p,c,a+1))))\Phi(p,c,a+1)^2 \\
& +(\Phi(p,c-4,a+1)+a\,(-4\Phi(p,c-3,a+1)+a\,(6\Phi(p,c-2,a+1) \\
& -4a\Phi(p,c-1,a+1)+a^2\Phi(p,c,a+1)\Phi(p,c,a+1)^3
\end{aligned} \tag{A.48}$$

Zipf (a,c)

Raw moments:

$$\begin{aligned}
\mu'_1 &= \tfrac{\Phi(1,c-1,a+1)-a\Phi(1,c,a+1)}{\zeta(c,a+1)} \\
\mu'_2 &= \tfrac{\Phi(1,c-2,a+1)-2a\Phi(1,c-1,a+1)+a^2\Phi(1,c,a+1)}{\zeta(c,a+1)} \\
\mu'_3 &= \tfrac{\Phi(1,c-3,a+1)-3a\Phi(1,c-2,a+1)+3a^2\Phi(1,c-1,a+1)-a^3\Phi(1,c,a+1)}{\zeta(c,a+1)} \\
\mu'_4 &= \tfrac{\Phi(1,c-4,a+1)-4a\Phi(1,c-3,a+1)+6a^2\Phi(1,c-2,a+1)-4a^3\Phi(1,c-1,a+1)+a^4\Phi(1,c,a+1)}{\zeta(c,a+1)}
\end{aligned} \quad , \tag{A.49}$$

Central moments:

$$
\begin{aligned}
\mu_1 = \quad & 0 \\
\mu_2 = \quad & \tfrac{1}{\zeta(c,a+1)^2}\left(-(\Phi(1,c-1,a+1)-a\Phi(1,c,a+1))^2+(\Phi(1,c-2,a+1)\right. \\
& +a(-2\Phi(1,c-1,a+1)+a\Phi(1,c,a+1)))\zeta(c,a+1) \\
\mu_3 = \quad & \tfrac{1}{\zeta(c,a+1)^3}\left(2(\Phi(1,c-1,a+1)-a\Phi(1,c,a+1))^3-3(\Phi(1,c-1,a+1)\right. \\
& -a\Phi(1,c,a+1))(\Phi(1,c-2,a+1)+a(-2\Phi(1,c-1,a+1) \\
& +a\Phi(1,c,a+1)))\zeta(c,a+1)+(\Phi(1,c-3,a+1)-a(3\Phi(1,c-2,a+1) \\
& +a(-3\Phi(1,c-1,a+1)+a\Phi(1,c,a+1))))\zeta(c,a+1)^2 \\
\mu_4 = \quad & \tfrac{1}{\zeta(c,a+1)^4}\left(-3(\Phi(1,c-1,a+1)-a\Phi(1,c,a+1))^4+6(\Phi(1,c-1,a+1)\right. \\
& -a\Phi(1,c,a+1))^2(\Phi(1,c-2,a+1)+a(-2\Phi(1,c-1,a+1) \\
& +a\Phi(1,c,a+1)))\zeta(c,a+1)-4(\Phi(1,c-1,a+1)-a\Phi(1,c,a+1)) \\
& (\Phi(1,c-3,a+1)-a(3\Phi(1,c-2,a+1)+a(-3\Phi(1,c-1,a+1) \\
& +a\Phi(1,c,a+1))))\zeta(c,a+1)^2+(\Phi(1,c-4,a+1)+a\left(-4\Phi(1,c-3,a+1)\right. \\
& \left.+a\left(6\Phi(1,c-2,a+1)-4a\Phi(1,c-1,a+1)+a^2\Phi(1,c,a+1)\right)\right)\zeta(c,a+1)^3
\end{aligned}
$$

(A.50)

Good (p,c)

Raw moments:

$$
\begin{aligned}
\mu_1' &= \tfrac{p^{-c}\mathrm{Li})_{c-1}(p)}{\zeta(c)} \\
\mu_2' &= \tfrac{p^{-c}\mathrm{Li})_{c-2}(p)}{\zeta(c)} \\
\mu_3' &= \tfrac{p^{-c}\mathrm{Li})_{c-3}(p)}{\zeta(c)} \\
\mu_4' &= \tfrac{p^{-c}\mathrm{Li})_{c-4}(p)}{\zeta(c)}
\end{aligned}
\quad , \tag{A.51}
$$

Central moments:

$$\begin{aligned}
\mu_1 = {} & 0 \\
\mu_2 = {} & \tfrac{p^{-2c}\left(-\mathrm{Li}_{c-1}(p)^2+p^c\mathrm{Li}_{c-2}(p)\zeta(c)\right)}{\zeta(c)^2} \\
\mu_3 = {} & \tfrac{1}{\zeta(c)^3}p^{-3c}\left(2\mathrm{Li}_{c-1}(p)^3-3p^c\mathrm{Li}_{c-2}(p)\mathrm{Li}_{c-1}(p)\zeta(c)+p^{2c}\mathrm{Li}_{c-3}(p)\zeta(c)^2\right) \quad . \\
\mu_4 = {} & \tfrac{1}{\zeta(c)^4}p^{-4c}\left(-3\mathrm{Li}_{c-1}(p)^4+6p^c\mathrm{Li}_{c-2}(p)\mathrm{Li}_{c-1}(p)^2\zeta(c)\right. \\
& \left.-4p^{2c}\mathrm{Li}_{c-3}(p)\mathrm{Li}_{c-1}(p)\zeta(c)^2+p^{3c}\mathrm{Li}_{c-4}(p)\zeta(c)^3\right.
\end{aligned} \tag{A.52}$$

Zeta (c)

Raw moments:

$$\begin{aligned}
\mu'_1 &= \tfrac{\zeta(c-1)}{\zeta(c)} \\
\mu'_2 &= \tfrac{\zeta(c-2)}{\zeta(c)} \\
\mu'_3 &= \tfrac{\zeta(c-3)}{\zeta(c)} \\
\mu'_4 &= \tfrac{\zeta(c-4)}{\zeta(c)}
\end{aligned} \quad , \tag{A.53}$$

Central moments:

$$\begin{aligned}
\mu_1 &= 0 \\
\mu_2 &= \tfrac{\zeta(c-2)\zeta(c)-\zeta(c-1)^2}{\zeta(c)^2} \\
\mu_3 &= \tfrac{2\zeta(c-1)^3-3\zeta(c-2)\zeta(c-1)\zeta(c)+\zeta(c-3)\zeta(c)^2}{\zeta(c)^3} \\
\mu_4 &= \tfrac{-3\zeta(c-1)^4+6\zeta(c-2)\zeta(c-1)^2\zeta(c)-4\zeta(c-3)\zeta(c-1)\zeta(c)^2+\zeta(c-4)\zeta(c)^3}{\zeta(c)^4}
\end{aligned} \quad . \tag{A.54}$$

Generalized Poisson Distributions

Neyman Type A (μ, ϕ)

Raw moments:

$$\begin{aligned}
\mu'_1 &= \mathrm{e}^{\mu+1}\mu\phi \\
\mu'_2 &= \mathrm{e}^{\mu+1}\mu\phi(1+\phi+\mu\phi) \\
\mu'_3 &= e^{-1+\mu}\mu\phi(1+\phi(3+\phi+\mu(3+(3+\mu)\phi))) \\
\mu'_4 &= e^{-1+\mu}\mu\phi\left(1+\phi\left(7+7\mu+6\phi+6\mu(3+\mu)\phi+(1+\mu(7+\mu(6+\mu)))\phi^2\right)\right)
\end{aligned} \quad , \tag{A.55}$$

Central moments:

$$\begin{aligned}
\mu_1 = &\ 0 \\
\mu_2 = &\ e^{-2+\mu}\mu\phi\left(-e^{\mu}\mu\phi + e(1+\phi+\mu\phi)\right) \\
\mu_3 = &\ e^{-3+\mu}\mu\phi\left(2e^{2\mu}\mu^2\phi^2 - 3e^{1+\mu}\mu\phi(1+\phi+\mu\phi)\right. \\
&\left. + e^2(1+\phi(3+\phi+\mu(3+(3+\mu)\phi)))\right) \\
\mu_4 = &\ e^{-4+\mu}\mu\phi\left(-3e^{3\mu}\mu^3\phi^3 + 6e^{1+2\mu}\mu^2\phi^2(1+\phi+\mu\phi)\right. \\
&+ e^3\left(1+\phi\left(7+7\mu+6\phi+6\mu(3+\mu)\phi+(1+\mu(7+\mu(6+\mu)))\phi^2\right)\right) - 4 \\
&\left. e^{2+\mu}\mu\phi(1+\phi(3+\phi+\mu(3+(3+\mu)\phi)))\right)
\end{aligned} \quad . \tag{A.56}$$

Hermite (a, b)

Raw moments:

$$\begin{aligned}
\mu_1' = &\ a + 2b \\
\mu_2' = &\ a + 2b + (a+2b)^2 \\
\mu_3' = &\ a + 8b + 2(a+2b)(a+4b) + (a+2b)(a+4b+(a+2b)^2) \\
\mu_4' = &\ a + 16b + 3(a+2b)(a+8b) + 3(a+4b)(a+4b+(a+2b)^2) \\
&+ (a+2b)(a+8b+2(a+2b)(a+4b)+(a+2b)(a+4b+(a+2b)^2))
\end{aligned} \quad , \tag{A.57}$$

Central moments:

$$\begin{aligned}
\mu_1 &= 0 \\
\mu_2 &= a + 4b \\
\mu_3 &= a + 8b \\
\mu_4 &= a(1+3a) + 8(2+3a)b + 28b^2
\end{aligned} \quad . \tag{A.58}$$

Generalized Hermite (a, b, m)

Raw moments:

$$\begin{aligned}
\mu_1' = &\quad a + mb \\
\mu_2' = &\quad a + m^2 b + (a+mb)^2 \\
\mu_3' = &\quad a^3 + 2a^2(1+mb) + m^3 b(1+b(3+b)) + a(1+3mb(1+m+mb)) \\
\mu_4' = &\quad a^4 + a^3(6+4mb) + a^2(7+6mb(2+m+mb)) + m^4 b(1+b(7+b(6+b))) \\
&\quad +a(1+2mb(2+m(3+2m)+mb(3+6m+2mb))
\end{aligned} \quad , \tag{A.59}$$

Central moments:

$$\begin{aligned}
\mu_1 &= 0 \\
\mu_2 &= a + m^2 b \\
\mu_3 &= a + m^3 b \\
\mu_4 &= 3a^2 + m^4 b(1+3b) + a(1+6m^2 b)
\end{aligned} \quad . \tag{A.60}$$

Gegenbauer (a, b, k)

Raw moments:

$$\begin{aligned}
\mu_1' = &\quad -\tfrac{k(a+2b)}{a+b-1} \\
\mu_2' = &\quad \tfrac{1}{(a+b-1)^2}\left(k(ka^2 + 4b(kb+1) + a(1+(4k-1)b))\right) \\
\mu_3' = &\quad -\tfrac{1}{(a+b-1)^3}\left(k\left(a^3 k^2 + a^2\left(1+3k+b\left(1-3k+6k^2\right)\right) + 8b\left(1+b\left(1+3k+bk^2\right)\right)\right.\right. \\
&\quad +a\left(1+18bk+b^2(6k(2k-1)-1)\right) \\
\mu_4' = &\quad \tfrac{1}{(a+b-1)^4} k\left(a^4 k^3 + a^2\left(7k+2b+4\left(4+k+24k^2\right) + b^2(4+k(7+24(k-1)k))\right)\right. \\
&\quad +16b\left(1+b\left(4+7k+b\left(1+4k+6k^2+bk^3\right)\right)\right) + a^3\left(1+4k+6k^2\right. \\
&\quad +b(2k(2+k(-3+4k))-1) + a(1+b(13+64k+b(-13+8k(1+15k) \\
&\quad +b(8(-1+k)k(1+4k)-1))))
\end{aligned} \tag{A.61}$$

Central moments:

$$\begin{aligned}
\mu_1 = &\quad 0 \\
\mu_2 = &\quad \tfrac{k(a-(a-4)b)}{(a+b-1)^2} \\
\mu_3 = &\quad -\tfrac{(b+1)(a(a-b+1)+8b)k}{(a+b-1)^3} \\
\mu_4 = &\quad \tfrac{1}{(a+b-1)^4}\left(k\left(-a^3(b-1)+16b+16b^2(b+3k+4)\right.\right. \\
&\quad \left.\left.-a(b-1)(1+b(14+b+24k))+a^2(4+3k+b(8-6k+b(4+3k)))\right)\right)
\end{aligned} \tag{A.62}$$

Generalized Gegenbauer (a, m, α, β)

Raw moments:

$$\begin{aligned}
\mu_1' = &\quad -\tfrac{a(\alpha+m\beta)}{\alpha+\beta-1} \\
\mu_2' = &\quad -\tfrac{a\left(\alpha+\beta m^2-\frac{(1+a)(\alpha+\beta m)^2}{\alpha+\beta-1}\right)}{\alpha+\beta-1} \\
\mu_3' = &\quad -\tfrac{a\left(\alpha+\beta m^3-\frac{2(1+a)(\alpha+\beta m)\left(\alpha+\beta m^2\right)}{\alpha+\beta-1}-\frac{(1+a)(\alpha+\beta m)\left(\alpha+\beta m^2-\frac{(2+a)(\alpha+\beta m)^2}{\alpha+\beta-1}\right)}{\alpha+\beta-1}\right)}{\alpha+\beta-1} \qquad\qquad , \\
\mu_4' = &\quad -\tfrac{1}{\alpha+\beta-1}a\Bigg(\alpha+\beta m^4-\frac{3(1+a)(\alpha+\beta m)\left(\alpha+\beta m^3\right)}{\alpha+\beta-1}-\frac{3(1+a)\left(\alpha+\beta m^2\right)\left(\alpha+\beta m^2-\frac{(2+a)(\alpha+\beta m)^2}{\alpha+\beta-1}\right)}{\alpha+\beta-1} \\
&\quad -\tfrac{(1+a)(\alpha+\beta m)\left(\alpha+\beta m^3-\frac{2(2+a)(\alpha+\beta m)\left(\alpha+\beta m^2\right)}{\alpha+\beta-1}-\frac{(2+a)(\alpha+\beta m)\left(\alpha+\beta m^2-\frac{(3+a)(\alpha+\beta m)^2}{\alpha+\beta-1}\right)}{\alpha+\beta-1}\right)}{\alpha+\beta-1}
\end{aligned} \tag{A.63}$$

Central moments:

$$\begin{aligned}
\mu_1 = &\quad 0 \\
\mu_2 = &\quad \tfrac{a\left(\alpha+\beta\left(-\alpha(m-1)^2+m^2\right)\right)}{(\alpha+\beta-1)^2} \\
\mu_3 = &\quad -\tfrac{a\left(\alpha^2\left(1+\beta(m-1)^3\right)+\beta(1+\beta)m^3+\alpha\left(1-\beta^2(m-1)^3-\beta(m-2)(m+1)(-1+2m)\right)\right)}{(\alpha+\beta-1)^3} \\
\mu_4 = &\quad \tfrac{1}{(\alpha+\beta-1)^4}k\left(\alpha^3\left(1-\beta(m-1)^4\right)+\beta(1+\beta(4+\beta+3a))m^4\right. \qquad\qquad , \\
&\quad +\alpha^2\left(4+3a+\beta\left(-8-6a(m-1)^2+\beta(4+3a)(m-1)^4\right.\right. \\
&\quad +16m-8m^3+3m^4+\alpha\left(1+\beta\left(-3+m\left(4+m\left(6+6a+4m-3m^2\right)\right)\right.\right. \\
&\quad +\beta\left(3-\beta(-1+m)^4-6a(-1+m)^2m^2-8\left(m-2m^3+m^4\right)\right)
\end{aligned} \tag{A.64}$$

Appendix B

Publications and posters arising from this research

B.1 List of publications

McElduff, F, Mateos, P, Wade, A, and Cortina-Borja, M (2008). Whats in a name? The frequency and geographic distributions of UK surnames. *Significance*, 5:189192.

Chan, SK, Riley, PR, Price, KL, McElduff, F, Winyard, PJ, Welham, SJM, Woolf, AS, and Long, DA (2010). Corticosteroid-induced kidney dysmorphogenesis is associated with deregulated expression of known cystogenic molecules, as well as indian hedgehog. *American Journal of Physiology: Renal Physiology*, 298:F346F356.

McElduff, F, Cortina-Borja, M, Chan, S-K, and Wade, A (2010). When t-tests or Wilcoxon-Mann-Whitney tests wont do. *Advances in Physiology Education*, 34:128133.

Gordon, K, Pasco, G, McElduff, F,Wade, A, Howlin, P, and Charman, T (2011). A Communication-Based Intervention for Nonverbal Children With Autism: What Changes? Who Benefits? *Journal of Consulting and Clinical Psychology*, 79:447457.

B.2 List of Posters

McElduff, F, Mateos, P, Wade, A, and Cortina-Borja, M. The UK Surname distribution and potential applications. University of Edinburgh, UK. Royal Statistical Society Conference. September 2009 (Awarded 3rd prize in poster competition). and University of Lancaster, UK. Research Students Conference in Probability and Statistics. April 2009 (Awarded Best Poster).

McElduff, F, Wade, A, Chan, S-K, Woolf, A and Cortina-Borja, M. When are outliers surprising? University of Warwick, UK. Research Students Conference in Probability and Statistics. April 2010.

McElduff, F, Wade, A and Cortina-Borja, M. Outlier detection in discrete distributions. Brighton, UK. Royal Statistical Society Conference. September 2010.

The UK surname distribution and potential applications

Fiona McElduff[1], Pablo Mateos[2], Angie Wade[1] and Mario Cortina-Borja[1]

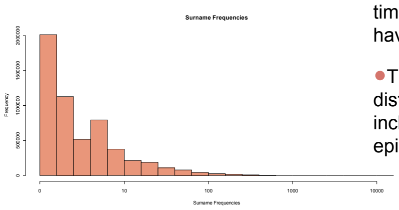
[1] MRC Centre of Epidemiology for Child Health, University College London Institute of Child Health

[2]Department of Geography, University College London

Background

- There are strong relationships between surname frequencies and the ethnic and genetic structures in a population.
- Surnames can be used in the field of child health as indicators of ethnicity in probabilistic record linkage[1].
- Examples of surnames as an indicator of ethnic origin in record linkage can be seen in studies of childhood cancer[2,3].
- Surnames are often patrilinearly inherited so they correlated well with Y-chromosomes[4] and can be used to identify genetic factors in certain diseases/conditions.

Figure 1: Surname frequencies



- Figure 1 illustrates the skew nature of the surname frequency distribution. Most surnames occur relatively few times with some common surnames having very high frequencies.
- This long-tailed, value-inflated distribution makes it an ideal dataset to include in my PhD ('Models for discrete epidemiological and clinical data').

Dataset

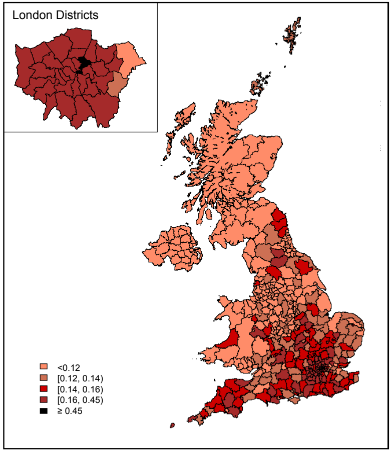
- The 'enhanced electoral register' contains the names and addresses of all adults entitled to vote in the UK, with additional non-registered voters sourced from commercial surveys and credit scoring databases.
- In 2001 the register contained 45,614,126 individuals, with a total of 817,391 surnames; 65.07% were unique but only correspond to 4.41% of the population.
- The dataset contains surname frequencies from 434 districts which can be grouped into 12 regions according to the official Government Office Regions.

Methods

- An established measure of diversity in linguistics used to quantify literary style is Yule's K[5], which is proportional to the probability of two persons randomly selected sharing the same surname. **Larger values of *K* indicate lower diversity in a population and hence greater uniformity of surnames.**
- The number of different surnames divided by the number of people in the population measures the volume of surname diversity.
- Surnames can be categorized by their geographical origin using the National Trust profiler (http://www.nationaltrustnames.org.uk/).

Results

Figure 2: Ratio of the number of different surnames divided by the number of people in each district



- Figure 2 shows that districts in the South of Britain have a higher proportion of surnames per head of population than those in the North.
- Figure 3 demonstrates there is a trend for those districts with a large proportion of people with unique surnames to have low values of *K* (indicating greater diversity).
- Scotland and Northern Ireland tend to have large *K* values and Wales clearly has the largest, signifying a higher rate of uniformity of surnames than in the rest of the UK.
- Districts in London have much higher proportions of people with unique surnames and lower values of *K*.

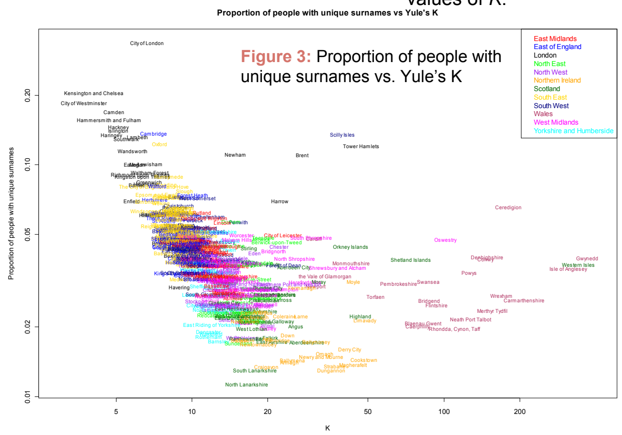


Figure 3: Proportion of people with unique surnames vs. Yule's K

- Oxford and Cambridge are clear outliers. Also the London districts of Tower Hamlets, Brent and Newham have a large number of unique surnames but much lower diversity.
- *Smith* is the most frequent surname in 308 out of the 434 districts- 1.02% of the population are *Smith*'s.
- The percentage of the population with the top 10 ranked surnames for each country, is given in Table 1.
- Wales has the highest cumulative percentage of the population with surnames in the top 10 (24.5%), indicating a lower diversity of surnames than those of the other countries.
- Irish surnames, e.g. *Kelly* and *O'Neill*, and Scottish surnames, e.g. *Campbell,* and *Johnston*, rank highly in the top surnames in Northern Ireland.
- English originating surnames, however, occur in all four UK countries, for example *Brown* arise in the top 10 surnames for all countries.

Rank	England	Scotland	Wales	Northern Ireland
1	Smith (1.26)	Smith (1.28)	Jones (5.75)	Wilson (0.75)
2	Jones (0.75)	Brown (0.94)	Williams (3.72)	Campbell (0.75)
3	Taylor (0.59)	Wilson (0.89)	Davies (3.72)	Kelly (0.74)
4	Brown (0.56)	Robertson (0.78)	Evans (2.47)	Johnston (0.69)
5	Williams (0.39)	Thomson (0.78)	Thomas (2.43)	Moore (0.62)
6	Wilson (0.39)	Campbell (0.77)	Roberts (1.53)	Thompson (0.61)
7	Johnson (0.37)	Stewart (0.73)	Lewis (1.53)	Smyth (0.60)
8	Davies (0.34)	Anderson (0.70)	Hughes (1.23)	Brown (0.59)
9	Robinson (0.32)	Scott (0.55)	Morgan (1.16)	O'Neill (0.57)
10	Wright (0.32)	Murray (0.53)	Griffiths (0.96)	Doherty (0.54)

Table 1: Top Surnames by Country (%)

Conclusion

- In this study we found that geographical regions of the UK have different surname structures. The spatial distribution of surnames reflects the genetic pool of the country's population[6].
- London, the South East and the East of England have higher surname diversity; Wales has a less varying surname distribution.
- A potential application of surnames frequencies is their use in childhood disease epidemiology as an indicator of genetic association.

This study was funded by:
Department of Health's NIHR Biomedical Research Centre's and Research and Development funding, as well as the Medical Research Council.

MRC Centre of Epidemiology for Child Health
30 Guilford Street
London WC1N 1EH
Email: F.McElduff@ich.ucl.ac.uk

References

1. Cook, D, et al. (1972) *American Journal of Epidemiology.* **96**(1):38-44.
2. Rankin, J, et al. (2008) *Paediatric Blood Cancer.* **51**:608-612.
3. Ducore, J, et al. (2008) *Journal of Pediatric Hematology/Oncology.* 26(10):613-618
4. Jobling, M. A. (2001) *Trends in Genetics,* **17,** 353–357.
5. Yule, G. U. (1944) *Cambridge University* Press.
6. McElduff, F, et al. (2008) *Significance.* **5**(4): 189-192.

The
child
first
and
always
Institute of Child Health

MRC Centre of Epidemiology for Child Health

When are outliers surprising?

Fiona McElduff[1], Angie Wade[1], Shun-Kai Chan[2], Adrian Woolf[2] and Mario Cortina-Borja[1]

[1] MRC Centre of Epidemiology for Child Health, University College London Institute of Child Health
[2]Department of Nephrourology, University College London Institute of Child Health

UCL

Background

- An outlier is an observation which appears to be inconsistent with the remainder of the dataset[1].
- Outlying observations can distort any inferences that are drawn from the sample.
- The detection of outliers poses particular problems when the data are discrete and/or the underlying distribution is highly skew with a long tail.
- This problem often arises whilst analysing data from paediatric clinical and epidemiological studies.

Methods

- The Surprise Index[2] (SI) provides an empirical measure of how unexpected an observed value is.
- If a random event has values $V_1, V_2, \dots, V_k$ occurring with probabilities $p_1, p_2, \dots, p_k$ then the SI is defined for each value x with corresponding probability p_x as:

SI_x = expected value of p ($E(p) = \Sigma^k_{i=1} p_i^2$) divided by the probability that the variable takes the value x (p_x)

- **A large SI indicates a more surprising event.**

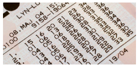
- The following categories can be used as guidelines to quantify how surprising an event is with respect to a chosen probability model[2]:

<5	Not Surprising
10	Begins to be surprising
1,000	Definitely Surprising
1,000,000	Very Surprising
10^{12}	Miracle!

- A rare event is not necessarily surprising but a surprising event is always rare. For example:

(i) Winning the Lottery is rare, but any combination of winning numbers is not in itself surprising since all combinations are equally likely.

(ii) Tossing a Coin The coin could land heads, tails or on its edge. Landing on its edge is a surprising event since the probability of this occur ring is low in relation to the probability of heads or tails, it is also a rare event because it has a very small probability.



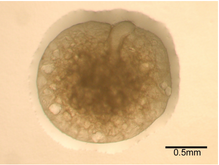
- We calculated analytical expressions for the SI for discrete distributions, estimated their parameters using R[3] and compared models using Bayesian Information Criterion (BIC), where a low BIC value indicates a better fit.

Case Study: Cysts in Embryonic Mice Kidneys

Dataset

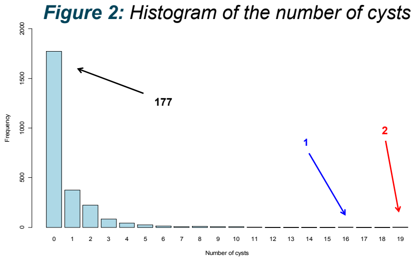
- This study is interested in the effects of mothers diets on the growth of kidney's in their unborn baby.
- Embryonic mice kidney cells (*n*=*2,559*) were examined for cysts (**Figure 1**).
- The distribution of the number of cysts per kidney cell (**Figure 2**) is highly skew with 69.2% of cells having no cysts. There are three possibly outlying cells with16, 19 and19 cysts.

Figure 1: Embryonic Kidney with cysts



Aims: To determine the model that best fits the data and to characterize any outliers.

Figure 2: Histogram of the number of cysts



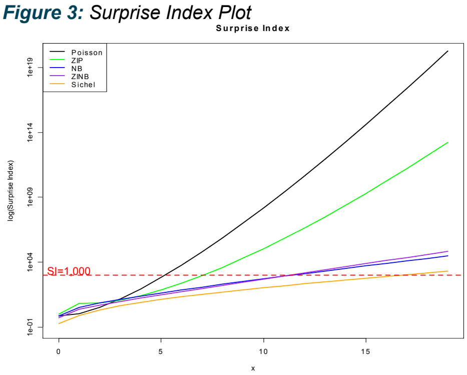
Results

- Probability models are compared in **Table 1** using the BIC and **Figure 3** shows the plotted SI's for the distributions fitted.

Distribution	BIC
Negative Binomial (NB)	5530.2
Zero-inflated negative binomial (ZINB)	5542.6
Sichel	5550.3
Zero-inflated Poisson (ZIP)	5838.8
Poisson	6933.0

Table 1: BIC values for models fitted to cysts.

Figure 3: Surprise Index Plot



- The Poisson distribution has the highest BIC value and observations greater than 5 are considered to be surprising.
- The ZIP distribution is a slight improvement to the fit of the data but observations greater than 8 are still surprising.
- For the NB and ZINB distributions observations greater than 12 (3 observations, 0.1%) are considered to be surprising.
- Surprise indices for the Sichel distribution only regard observations of 19 as outliers.
- The NB has the lowest BIC and hence provides the best overall fit to the data.

Conclusion

- The observations of 16 and 19 are considered surprising and therefore outliers, we should: (a) check for possible observational errors in these values (b) If no errors are found, quantify the sensitivity of any conclusions drawn to the presence of these surprising values.

This study was funded by: Department of Health's NIHR Biomedical Research Centre's and Research and Development funding. FM's work is funded by a capacity building PhD studentship from the Medical Research Council.

MRC Centre of Epidemiology for Child Health
30 Guilford Street
London WC1N 1EH
Email: F.McElduff@ich.ucl.ac.uk

References:
1. Barnett, V. and Lewis, T. (1978) *'Outliers in Statistical Data'* (3rd Edition) New York: John Wiley & sons.
2. Weaver, W (1948) *'Rarity, Probability, Interest and Surprise'* The Scientific Monthly, 67, 390-392.
3. R Development Core Team (2007) *'R: Language and Statistical Computing'* R Foundation for Statistical Computing, Vienna, Austria.

The child first and always
Institute of Child Health

Outlier Detection in Discrete Distributions

Fiona McElduff, Angie Wade and Mario Cortina Borja

MRC Centre of Epidemiology for Child Health,
University College London, Institute of Child Health.

MRC Centre of Epidemiology for Child Health

UCL

Background

- An outlier is an observation which appears to be inconsistent with the remainder of the dataset[1].
- Outlying observations can lead to distorted inferences from the sample.
- The detection of outliers poses particular problems when the data are discrete and/or the underlying distribution is highly skew with a long tail.

Methods

Surprise Index

- The Surprise Index[2] (SI) provides an empirical measure of how unexpected an observed value is.
- If a random event has values V_1, V_2, …, V_k occurring with probabilities p_1, p_2, …, p_k then the SI is defined for each value x with corresponding probability p_x as:

$$SI_x = \frac{E(p)}{p_x} = \frac{\text{average of } p}{\text{probability of event } x}$$

- A large value of SI indicates a more surprising event. The following categories can be used as guidelines to quantify how surprising an event is with respect to a chosen probability model[2]:

<5	Not surprising
10	Begins to be surprising
1,000	Definitely surprising
1,000,000	Very surprising
10^{12}	Miracle!

- We obtained analytical expressions for the SI of several discrete distributions, estimated their parameters using R[3] and compared models using the Bayesian Information Criterion (BIC), where a low BIC value indicates a better fit.
- A rare event is not necessarily surprising but a surprising event is always rare. For example:

(i) Winning the Lottery is rare, but any combination of winning numbers is not in itself surprising since all combinations are equally likely.

(ii)Tossing a coin The coin could land heads, tails or on its edge. Landing on its edge is a surprising event since the probability of this occurring is low in relation to the probability of heads or tails, it is also a rare event because it has a very small probability.

Empirical Probability Generating Function

- The Empirical Probability Generating Function (EPGF) provides a smooth projection of the observed data V_1, V_2, …, V_k:

$$G_k = \frac{1}{k}\sum_{i=1}^{k} t^{V_k}$$

- Where -1 ≤ t ≤ 1.
- If an observation has a large effect on the distribution of the dataset the epgf calculated without the observation will be substantially different, hence

Dataset

- We analyze the frequencies of stillbirths in 402 litters of New Zealand white rabbits[5] (**Table 1**).
- The distribution is zero-inflated with 78.1% of the litters having no stillbirths. Overdispersion is clearly present as the variance (1.51) is much larger than the mean (0.46).
- A possibly outlying observation of 11 stillbirths in one litter can be seen in the tail end of the distribution.

Application

Distribution	**Number of stillbirths**												BIC
	0	1	2	3	4	5	6	7	8	9	10	11	
Observed	**314**	**48**	**20**	**7**	**5**	**2**	**2**	**1**	**2**	**0**	**0**	**1**	---
Poisson	254	117	27	4	·	·	·	·	·	·	·	·	887.7
ZIP	314	33	28	16	7	2	1	·	·	·	·	·	726.4
NB	314	46	19	10	5	3	2	1	1	·	·	·	686.3
ZINB	314	46	19	10	5	3	2	1	1	·	·	·	692.3
Sichel	314	49	18	9	5	3	2	1	1	·	·	·	691.9
ZI Sichel	314	48	18	9	5	3	2	1	1	·	·	·	697.9
SI- NB	0.8	5.5	13.2	26.3	48.0	83.5	140.8	232.5	378.0	607.1	965.9	1524.9	---

Table 1: Frequencies of stillbirths and BIC values.

- Comparing the BIC's (**Table 1**) shows the negative binomial model provides the best fit to the data. A mean of 0.46 and dispersion of 2.15 can be estimated from the fitted model.
- The probability of 11 stillbirths under this model is 0.0004, (SI=1524.9) indicating that this rare event with a low probability can also be regarded as surprising.
- Removing the outlying observation of 11 stillbirths in one litter and fitting a negative binomial model to the remainder of the dataset produces an estimated mean of 0.43 and dispersion of 1.88.

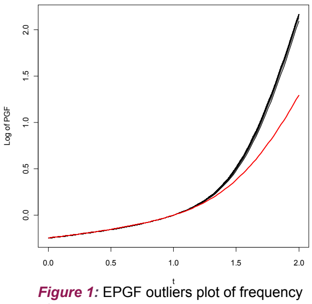


Figure 1: EPGF outliers plot of frequency of stillbirths.

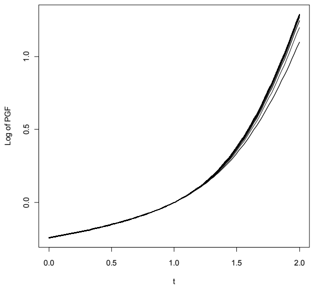


Figure 2: EPGF outliers plot of frequency of stillbirths without the observation with 11 stillbirths.

- 402 EPGF curves each with 401 observations are plotted in **Figure 1**.
- The curve for the EPGF not including the observation with 11 stillbirths (highlighted in red) is different to the remaining 401 curves.
- Removing this observation and reconstructing the EPGF outlier plot (**Figure 2**) the remaining 401 curves do not indicate any outliers.

Summary

- The observation of 11 stillbirths in one litter is considered to be an outlier in this dataset.
- Assuming that the underlying probability model is correct the SI is preferable as it yields a numerical value; the EPGF method is a graphical, non-parametric procedure.
- With any dataset containing potential outliers various methods should be used for formal identification. If any outliers are detected and if no recording errors are found, then sensitivity analyses should be undertaken to assess their influence on the study conclusions.

This study was funded by: Department of Health's NIHR Biomedical Research Centre's and Research and Development funding, as well as the Medical Research Council. FM is funded by an MRC capacity building studentship.

MRC Centre of Epidemiology for Child Health
30 Guilford Street
London WC1N 1EH
Email: F.McElduff@ich.ucl.ac.uk

References:
1. Barnett V, Lewis T (1978) John Wiley & sons.
2. Weaver W (1948) The Scientific Monthly, 67, 390-2.
3. R Development Core Team (2009) Vienna, Austria.
4. Nakamura M, Perez-Abreu V (1993) Commun Stat- Theor M, 22, 827-42.
5. Morgan BJT, Palmer KJ, Ridout MS (2007) Am Stat, 61, 285-7.