The Distribution Toolbox
of GAMLSS

Bob Rigby, Mikis Stasinopoulos, Gillian Heller and Vlasios Voudouris

September 10, 2014
Preface

This book is designed as an practical introduction to univariate statistical parametric distributions. In particular it provides a guide to different types of distributions existing within the framework of Generalised Additive Models for Location, Scale and Shape (GAMLSS), see Rigby and Stasinopoulos [2005].

The book is designed to help:

• practitioners to choose an appropriate parametric distributions for their data.
• students as an introduction to different parametric distributions
• researchers as a quick reference book on distributions and their properties

This book is not a general reference book on univariate distributions. This is almost an impossible task since there are hundreds of different parametric distributions and new distributions are created daily. It does cover only parametric univariate distributions which we will refer to as the gamlss.family because they have been implemented in the gamlss.dist package in the R environment, see R Development Core Team [2011]. The gamlss.dist package is part of the generalised additive models for location scale and shape (GAMLSS) implementation in R, Stasinopoulos and Rigby [2007].
## Contents

1 The GAMLSS Family

1 Types of univariate parametric distributions
   1.1 Introduction .......................................................... 19
   1.1.1 Probability (density) functions ............................... 20
   1.1.2 Cumulative distribution function and its inverse ............ 23
   1.1.3 Survival and hazard functions ............................... 24
   1.2 Population distributions in R: the d, p, q and r functions ....... 24
   1.3 Properties of the distributions .................................. 28
      1.3.1 Mean, variance and moment based measures of skewness and kurtosis . 28
      1.3.2 Centiles and centile based measures of spread, skewness and kurtosis .... 32
      1.3.3 Mode ......................................................... 35
   1.4 Types of distribution within the GAMLSS family ............... 35
   1.5 Displaying GAMLSS family distributions ......................... 38

2 Continuous distributions

   2.1 Introduction .......................................................... 43
   2.2 Continuous distributions on support $(-\infty, \infty)$ .................. 44
      2.2.1 Two parameter distributions on $(-\infty, \infty)$ .................. 44
      2.2.2 Three parameter distributions on $(-\infty, \infty)$ ............... 45
      2.2.3 Four parameter distributions on $(-\infty, \infty)$ ................ 46
   2.3 Continuous distributions on support $(0, \infty)$ .................... 51
      2.3.1 One parameter distributions on $(0, \infty)$ .................... 51
      2.3.2 Two parameter distributions on $(0, \infty)$ .................... 51
      2.3.3 Three parameter distributions on $(0, \infty)$ ................ 54
      2.3.4 Four parameter distributions on $(0, \infty)$ .................... 56
   2.4 Continuous distributions on support $(0, 1)$ ..................... 57
   2.5 Comparison of properties of continuous distributions ............ 57

3 Discrete distributions for count data

   3.1 Introduction .......................................................... 59
   3.1.1 Poisson distribution ............................................. 59
   3.1.2 Overdispersion .................................................. 60
   3.1.3 Excess or shortage of zero values ............................ 61
   3.1.4 Modelling count data in GAMLSS ............................. 61
   3.2 Explicit continuous mixtures of Poisson distributions ........... 62
CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2.1</td>
<td>Negative binomial distribution</td>
<td>63</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Poisson-inverse Gaussian</td>
<td>65</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Sichel distribution</td>
<td>65</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Delaporte distribution</td>
<td>67</td>
</tr>
<tr>
<td>3.3</td>
<td>Zero inflated discrete distributions</td>
<td></td>
</tr>
<tr>
<td>3.3.1</td>
<td>Zero inflated Poisson distribution</td>
<td>68</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Zero inflated Poisson type 2 parameterization</td>
<td>68</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Zero inflated negative binomial type 1 distribution</td>
<td>70</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Zero inflated Poisson inverse Gaussian distribution</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>Zero altered discrete distributions</td>
<td></td>
</tr>
<tr>
<td>3.4.1</td>
<td>Zero altered Poisson distribution</td>
<td>70</td>
</tr>
<tr>
<td>3.5</td>
<td>Zero altered logarithmic distribution, ZALG</td>
<td>72</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Zero altered negative binomial type 1 distribution, ZANBI</td>
<td>72</td>
</tr>
<tr>
<td>3.6</td>
<td>Comparison of the count distributions</td>
<td></td>
</tr>
<tr>
<td>3.7</td>
<td>Discretised continuous distributions method</td>
<td></td>
</tr>
<tr>
<td>3.8</td>
<td>Ad-hoc methods</td>
<td>80</td>
</tr>
<tr>
<td>3.9</td>
<td>Examples: fitting a distribution</td>
<td></td>
</tr>
<tr>
<td>3.10</td>
<td>Bibliography</td>
<td>80</td>
</tr>
</tbody>
</table>

4 Binomial data distributions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Available distributions</td>
<td>83</td>
</tr>
</tbody>
</table>

5 Mixed distributions

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Zero adjusted distributions on zero and the positive real line $[0, \infty)$</td>
<td>85</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Zero adjusted gamma distribution, $ZAGA(\mu, \sigma, \nu)$</td>
<td>85</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Fitting zero adjusted distributions on zero and the positive real line</td>
<td>86</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Example of fitting a response variable on zero and the positive real line</td>
<td>87</td>
</tr>
<tr>
<td>5.2</td>
<td>Distributions on the unit interval $(0,1)$ inflated at 0 and 1</td>
<td>87</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Beta inflated distribution, $BEINF(\mu, \sigma, \nu, \tau)$</td>
<td>87</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Fitting distributions on the unit interval $(0,1)$ inflated at 0 and 1</td>
<td>88</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Beta inflated at 0 distribution, $BEINF0(\mu, \sigma, \nu)$</td>
<td>88</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Beta inflated at 1 distribution, $BEINF1(\mu, \sigma, \nu)$</td>
<td>89</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Example of fitting a response variable on the interval $(0,1]$, the unit interval including value 1</td>
<td>90</td>
</tr>
</tbody>
</table>

6 Finite mixtures

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Introduction to finite mixtures</td>
<td>93</td>
</tr>
<tr>
<td>6.2</td>
<td>Finite mixtures with no parameters in common</td>
<td>94</td>
</tr>
<tr>
<td>6.2.1</td>
<td>The likelihood function</td>
<td>94</td>
</tr>
<tr>
<td>6.2.2</td>
<td>Maximizing the likelihood function using the EM algorithm</td>
<td>94</td>
</tr>
<tr>
<td>6.2.3</td>
<td>Modelling the mixing probabilities</td>
<td>96</td>
</tr>
<tr>
<td>6.2.4</td>
<td>Zero components</td>
<td>97</td>
</tr>
<tr>
<td>6.3</td>
<td>The gamlssMX() function</td>
<td>98</td>
</tr>
<tr>
<td>6.4</td>
<td>Examples using the gamlssMX() function</td>
<td>99</td>
</tr>
<tr>
<td>6.5</td>
<td>Finite mixtures with parameters in common</td>
<td>99</td>
</tr>
<tr>
<td>6.5.1</td>
<td>Maximizing the likelihood using the EM algorithm</td>
<td>100</td>
</tr>
<tr>
<td>6.6</td>
<td>The gamlssNP() function</td>
<td>100</td>
</tr>
<tr>
<td>6.7</td>
<td>Examples using the gamlssNP() function</td>
<td>101</td>
</tr>
<tr>
<td>6.8</td>
<td>Bibliographic notes</td>
<td>101</td>
</tr>
</tbody>
</table>
8.2.2 The function flexDist() .......................... 149
8.3 Examples of fitting continuous distributions to data .......................... 149
  8.3.1 Farzen79 .......................................... 149
  8.3.2 The strengths of glass fibres data .................................. 156
8.3.3 The tensile strength data: response on (0, 1) .......................... 159
8.4 Exercises for Chapter 3 ...................................... 161
  8.4.1 The Turkish stock exchange index .................................. 161
9 Fitting Count Data .............................................. 163
  9.1 Examples: fitting a distribution ..................................... 163
    9.1.1 The computer failure data ....................................... 163
    9.1.2 The lice data ............................................... 165
    9.1.3 A stylometric application ...................................... 167
10 Fitting Binomial data .......................................... 173
  10.1 Examples of fitting binomial data .................................. 173
    10.1.1 The alveolar-bronchiolar adenomas data .......................... 173
    10.1.2 The first year student examination results data ................. 175
11 Methods of generating Distributions .......................... 179
  11.1 Methods of generating continuous distributions ....................... 179
  11.2 Distributions generated by Azzalini’s method ......................... 180
    11.2.1 Azzalini (1985) method ...................................... 180
    11.2.2 Azzalini and Capitano (2003) method ........................... 182
  11.3 Distributions generated by splicing .................................. 183
    11.3.1 Splicing using two components ................................ 183
    11.3.2 Splicing using two components with different scale parameters .. 183
    11.3.3 Splicing using two components with different shape parameters .... 185
    11.3.4 Splicing using three components ................................ 185
  11.4 Distributions generated by a mixture of distributions ............... 186
    11.4.1 Explicitly defined continuous mixture distributions ........... 186
  11.5 Distributions generated by univariate transformation .................. 187
  11.6 Distributions generated by transformation from two or more random variables .. 192
  11.7 Truncation distributions ...................................... 192
    11.7.1 Left truncation .............................................. 193
    11.7.2 Right truncation ............................................. 193
    11.7.3 Both sizes truncation ........................................ 194
  11.8 Systems of distributions .................................... 195
    11.8.1 Pearson system ............................................. 195
    11.8.2 Stable distribution system .................................... 195
    11.8.3 Exponential Family ......................................... 195
    11.8.4 generalised inverse Gaussian family ........................... 196
12 Heaviness of tails of continuous distributions .................. 197
  12.1 Introduction .............................................. 197
  12.2 Types of tails for continuous distributions .......................... 198
  12.3 Classification Tables ...................................... 200
  12.4 Methods for choosing the appropriate tail .......................... 205
    12.4.1 Exploratory Method 1: log survival function plot ................ 205
CONTENTS

12.4.2 Exploratory Method 2: log-log-Survival ............................ 206
12.4.3 Exploratory Method 3: truncated distribution fitting ............. 209
12.5 .......................................................... 209
12.5.1 Lemma B1 .................................................. 209
12.5.2 Lemma B2 .................................................. 210
12.5.3 Lemma B3 .................................................. 212
12.5.4 Corrolary C1 ............................................... 212
12.5.5 Corrolary C2 ............................................... 212

13 Centile based comparisons of continuous distributions ............... 215
13.1 Introduction ..................................................... 215
13.2 Transformed (centile) kurtosis against (centile) central skewness ............................................. 217
13.3 Transformed (centile) kurtosis against (centile) tail skewness ........ 218
13.4 Conclusions ....................................................... 221

II A reference guide .................................................. 223

14 Continuous distributions on \( \mathbb{R} \) ................................. 225
14.1 Continuous two parameter distributions on \( \mathbb{R} \) ................. 225
14.1.1 Normal (or Gaussian) distribution (NO, NO2, NOF) ............. 225
14.1.2 Logistic distribution (LO) .................................... 226
14.1.3 Gumbel distribution (GU) ..................................... 226
14.1.4 Reverse Gumbel distribution (RG) ................................ 226
14.2 Continuous three parameter distributions on \( \mathbb{R} \) ............... 226
14.2.1 Exponential Gaussian distribution (exGAUS) .................... 226
14.2.2 Power Exponential distribution (PE, PE2) ........................ 227
14.2.3 \( t \) family distribution (TF) ................................ 227
14.3 Continuous four parameter distributions on \( \mathbb{R} \) ................. 228
14.3.1 Exponential Generalized Beta type 2 distribution (EGB2) ....... 228
14.3.2 Generalized \( t \) distribution (GT) .................................. 228
14.3.3 Johnson SU distribution (JSUo, JSU) ........................... 228
14.3.4 Normal-Exponential-\( t \) distribution (NET) ..................... 229
14.3.5 Sinh-Arcs sinh (SHASH) ........................................ 230
14.3.6 Sinh-Arcsinh (SHASHo) ......................................... 230
14.3.7 Sinh-Arcsinh (SHASHo2) ...................................... 230
14.3.8 Skew Exponential Power type 1 distribution (SEP1) ............. 230
14.3.9 Skew Exponential Power type 2 distribution (SEP2) ............. 230
14.3.10 Skew Exponential Power type 3 distribution (SEP3) .......... 231
14.3.11 Skew Exponential Power type 4 distribution (SEP4) ........... 231
14.3.12 Skew \( t \) type 1 distribution (ST1) ............................ 232
14.3.13 Skew \( t \) type 2 distribution (ST2) ............................. 232
14.3.14 Skew \( t \) type 3 distribution (ST3) ............................. 232
14.3.15 Skew \( t \) type 4 distribution (ST4) ............................. 232
14.3.16 Skew \( t \) type 5 distribution (ST5) ............................. 233
CONTENTS

15 Continuous distributions in $\mathbb{R}^+$ 235
  15.1 Continuous one parameter distribution in $\mathbb{R}^+$ 235
  15.1.1 Exponential distribution (EXP) 235
  15.2 Continuous two parameter distribution in $\mathbb{R}^+$ 235
  15.2.1 Gamma distribution (GA) 235
  15.2.2 Log Normal distribution (LOGNO, LNO) 235
  15.2.3 Inverse Gaussian distribution (IG) 236
  15.2.4 (WEI, WEI2, WEI3) 236
  15.3 Continuous three parameter distribution in $\mathbb{R}^+$ 237
  15.3.1 Box-Cox Cole and Green distribution (BCCG) 237
  15.3.2 Generalized gamma distribution (GG, GG2) 238
  15.3.3 Generalised inverse Gaussian distribution (GIG) 238
  15.3.4 Zero adjusted Gamma distribution (ZAGA) 238
  15.3.5 Zero adjusted Inverse Gaussian distribution (ZAIG) 239
  15.4 Continuous four parameter distribution in $\mathbb{R}^+$ 239
  15.4.1 Box-Cox $t$ distribution (BCT) 239
  15.4.2 Box-Cox power exponential distribution (BCPE) 239
  15.4.3 Generalized Beta type 2 distribution (GB2) 239

16 Continuous distribution in $\mathbb{R}[0, 1]$ 241
  16.1 Continuous two parameter distribution in $\mathbb{R}[0, 1]$ 241
  16.1.1 Beta distribution (BE, BEo) 241
  16.1.2 Beta inflated distribution (BEINF, BEINF0, BEINF1) 241
  16.1.3 Generalized Beta type 1 distribution (GB1) 242

17 Count data distributions 243
  17.1 Count data one parameter distributions 243
    17.1.1 Poisson distribution (PO) 243
    17.1.2 Logarithmic distribution (LG) 243
  17.2 Count data two parameters distributions 244
    17.2.1 Negative Binomial distribution (NBI, NBII) 244
    17.2.2 Poisson-inverse Gaussian distribution (PIG) 244
    17.2.3 Zero inflated poisson (ZIP, ZIP2) 245
    17.2.4 Zero altered (or adjusted) poisson (ZAP) 246
    17.2.5 Zero altered (or adjusted) logarithmic (ZALG) 246
  17.3 Count data three parameters distributions 247
    17.3.1 Delaporte distribution (DEL) 247
    17.3.2 Sichel distribution (SI, SICHEL) 247
    17.3.3 Zero inflated negative binomial distribution (ZINBI) 248
    17.3.4 Zero altered (or adjusted) negative binomial distribution (ZANBI) 248
    17.3.5 Zero inflated Poisson inverse Gaussian distribution (ZIPIG) 249

18 Binomial type distributions 251
  18.1 Binomial type data one parameter distributions 251
    18.1.1 The Binomial distribution (BI) 251
  18.2 Binomial type data two parameters distributions 251
    18.2.1 Beta Binomial distribution (BB) 251
    18.2.2 Zero altered (or adjusted) binomial (ZABI) 251
    18.2.3 Zero inflated binomial (ZIBI) 252
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.3</td>
<td>Binomial type data three parameters distributions</td>
<td>252</td>
</tr>
<tr>
<td>18.3.1</td>
<td>Zero altered (or adjusted) beta binomial (ZABB)</td>
<td>252</td>
</tr>
<tr>
<td>18.4</td>
<td>Binomial type data three parameters distributions</td>
<td>253</td>
</tr>
<tr>
<td>18.4.1</td>
<td>Zero inflated beta binomial (ZIBB)</td>
<td>253</td>
</tr>
</tbody>
</table>
List of Figures

1.1 Showing how the shape of the exponential distribution changes with different values of $\mu = 0.25, 0.5, 1, 2$ ................................................. 21
1.2 Showing how the shape of the Poisson distribution changes with different values of $\mu = 1, 3, 5, 15$ .................................................. 22
1.3 Showing relationship between the pdf and cdf functions on a continuous distribution. The area under the curve of the pdf until the point $y$ is the value of the cdf at this point $F_Y(y)$ .................................................. 23
1.4 Showing relationship between the pdf and cdf functions on a discrete distribution. The sum of probabilities until $y$ in the probability function is the value of the cdf at this point $F_Y(y)$ ................................................. 24
1.5 Showing the probability function, commutative distribution function, inverse commutative distribution function and a random sample of 1000 observations from a gamma distribution ........................................ 26
1.6 Showing the probability function, commutative distribution function, inverse commutative distribution function and a random sample of 1000 observations from a negative binomial distribution .................................. 27
1.7 Showing different types of continuous distributions .................................. 29
1.8 Showing the difference between the mean median and mode of a positive skew distribution ................................................................. 30
1.9 Showing the $Q_1$, the median and $Q_3$. ................................................. 33
1.10 Showing how $Q_1$, the median and $Q_3$ are derived in a continuous distribution. 34
1.11 Showing how $Q_1$, the median and $Q_3$ are derived in a discrete distribution. ... 34
1.12 Showing the normal distributions, an example of a continuous distribution .... 36
1.13 Showing the Poisson distributions, an example of a discrete distribution ...... 36
1.14 Showing the zero adjusted gamma distributions, an example of a mixed distribution ................................................................. 37
2.1 Two parameter distributions, NO(0,1), GU(0,1) and RG(0,1) .................. 45
2.2 The skew normal type 1 distribution, SN1(0,1, $\nu$) for $\nu=0, 1, 3$ and 100. Changing the sign of $\nu$ reflects the distribution about $y = 0$ ......................... 47
2.3 The power exponential distribution, PE(0, 1, $\nu$), for $\nu=1, 2, 10$ and 1000 .... 47
2.4 The NET distribution NET($\mu = 0$, $\sigma = 1.5$, $\nu = 2$, $\tau = 3$) ............. 49
2.5 The skew exponential power type 1 distribution, SEP1(0.1, $\nu$, $\tau$) ............... 50
2.6 The pdf and hazard functions of the Weibull type 2 distribution, WEI2, for $\mu = 1$ and $\sigma = 0.5, 1, 1.2$ and the gamma, GA, for $\mu = 1$ and $\sigma = 0.5, 1, 2$. ......................... 52
2.7 The pdf and hazard functions of the inverse gamma distribution, IGAMMA, and the inverse Gaussian, IG, for different values of $\sigma = 0.5, 1, 2$ ..................... 54
LIST OF FIGURES

2.8 The log normal family $\mathcal{LN}(0,1,\nu)$ for $\nu = -1, 1, 2$ .......................... 55
2.9 The Probability Density Function $f(y)$ of the Box-Cox $t$ Distribution $BCT(\mu, \nu, \rho, \tau)$ Against $y$. Parameter values (a) $\mu = (8, 10, 12), \sigma = 0.15, \nu = 1, \tau = 10$ (b) $\mu = 10, \sigma = (0.10, 0.15, 0.20), \nu = 1, \tau = 10$ (c) $\mu = 10, \sigma = 0.15, \nu = (-5, 1, 5)$, $\tau = 10$ (d) $\mu = 10, \sigma = 0.15, \nu = 1, \tau = (1, 10, 40)$. The three curves for the parameter varying are (---), (...) and (--) respectively .......................... 56
2.10 The beta distribution family $\mathcal{BE}$: In panel (a) the parameters are set to $\mu = 0.5$ and $\sigma = 0.3, 0.5, 0.7$, while in panel (b) to $\sigma = 0.5$ and $\mu = 0.1, 0.5, 0.9$ ............ 57
3.1 NBI distribution for $\mu = 5$ and $\sigma = (0.01, 0.5, 1, 2)$ ................................. 64
3.2 PIG distribution for $\mu = 5$ and $\sigma = (0.01, 0.5, 1, 2)$ ................................. 66
3.3 ZIP distribution for $\mu = 5$ and $\sigma = (0.01, 0.1, 0.4, 0.7)$ ............................. 69
3.4 Skewness-kurtosis combinations for different distributions for $Y$ (with fixed mean 5 and variance 30) .................................................. 74
3.5 ZIP, NBI, NBItr and PIG distributions for $Y$ (with fixed mean 5 and variance 30) 75
3.6 ZINBI distributions for $Y$ corresponding to four points along the ZINBI skewness-kurtosis curve from ZIP to NBI (with fixed mean 5 and variance 30) ............. 76
3.7 ZANBI distributions for $Y$ corresponding to four points along part of the ZANBI skewness-kurtosis curve from NBI to NBItr (with fixed mean 5 and variance 30) 77
3.8 four SICHEL distributions for $Y$ with increasing values of $\nu$, all having mean 5 and variance 30 .......................... 78
3.9 four DEL distributions for $Y$ with increasing values of $\nu$, all having mean 5 and variance 30 .......................... 79
7.1 The empirical distribution and cumulative distribution function of the Parzen’s snow fall data .................................................. 105
7.2 Showing a schematic plot of true population, the empirical distribution function and the model defined by $f(Y|\theta)$ .......................... 106
7.3 Showing a histogram of the air condition data. .......................... 110
7.4 Showing the likelihood and log likelihood function for the air condition data assuming an exponential distribution. .......................... 111
7.5 Showing different views of the log likelihood function for the air conditioning data assuming a gamma distribution: (a) top left corner shows the two dimensional log likelihood for $\mu$ (theta1) and $\sigma$ (theta2) (b) top right corner shows a contour plot of the two dimensional log likelihood. (c) The bottom left the profile log likelihood for $\mu$ (theta1) (d) bottom right the profile log likelihood for $\sigma$ (theta2) (see text for more details). .......................... 113
7.6 Showing for scenario I the likelihood and the log-likelihood function for the air-cond data, the MLE estimator and confidence bounds for the parameters $\mu$ .......................... 116
7.7 Scenario II log-likelihood based confidence bounds for parameter $\mu$ and $\sigma$ at levels 0.5, 0.1 and 0.05 of the standardised likelihood. .......................... 117
7.8 Showing the log likelihood of the aircond data together with the quadratic approximation of the log likelihood .......................... 124
7.9 Profile global deviance plot for $\mu$ for gamma example .......................... 129
7.10 Profile global deviance plot for $\sigma$ for gamma example .......................... 130
7.11 A schematic presentation of the \( \theta_c \), the ‘closest’ or ‘best’ value for \( \theta \) under the model \( f_Y(y|\theta) \), and the \( \hat{\theta} \) the MLE. \( \theta_c \) is the value of \( \theta \) ‘closest’ to the population using the Kullback-Liebler distance (or risk function), while \( \hat{\theta} \) is the value of \( \theta \) closest to the sample using the empirical risk function. The solid (red) line represents confidence bounds for the parameter \( \theta_c \). .......................................................... 134

8.1 Showing the global deviance for the parameter \( \mu \) at the top and the parameter \( \eta_\mu \) at the bottom .......................................................... 147
8.2 Showing the plot produced by the function \texttt{histDist()} on the \texttt{aircond} using an exponential distribution fit .......................................................... 148
8.3 A histogram of the Parzen’s snowfall data together with the fitted Weibull distribution (red) and a kernel density estimate (blue) .......................................................... 151
8.4 The residual plot from the fitted Weibull model to the snowfall data .......................................................... 152
8.5 The worm plot from the fitted Weibull model to the snowfall data .......................................................... 153
8.6 The profile deviance for the \( \sigma \) parameter for fitted Weibull model .......................................................... 155
8.7 The bootstrap for log \( \sigma \) from model modWEI .......................................................... 156
8.8 The strengths of glass fibres data and the fitted SEP4 distribution model .......................................................... 158
8.9 A worm plot of the residuals for the model using the SEP4 distribution fitted to the strengths of glass fibres data .......................................................... 159
8.10 Tensile strength data with fitted (a) logit-normal , (b) beta, (c) logit- \( t \) (d) logit-SEP3 (e) logit-TS3 and (f) generalised beta type distributions .......................................................... 161

9.1 The computer failure data fit with (a) Poisson , (b) negative binomial (c) Poisson inverse gaussian and (d) Sichel distributions respectively .......................................................... 164
9.2 The lice data with the fitted (a) negative binomial and (b) Sichel distributions respectively .......................................................... 168
9.3 The stylometric data: number of time a word appear in a text against the frequencies .......................................................... 169
9.4 The stylometric data fits with (a) Poisson (b) negative binomial type II (c) De- laporte and (d) Sichel distributions respectively .......................................................... 171

10.1 The proportion of alveolar-bronchiolar adenomas .......................................................... 174
10.2 The proportion and fitted distribution to the artificial alveolar-bronchiolar adenomas data .......................................................... 175
10.3 The first year student examination results data: Fitted probabilities (a) binomial (b) beta binomial distributions .......................................................... 177

11.1 Azzalini’s method (a) pdf, \( f_{Z_1}(z) \) of \( Z_1 \sim \text{NO}(0,1) \) (b) \( 2^*\text{cdf} \), \( 2^*F_{Z_2}(\nu z) \) of \( Z_2 \sim \text{NO}(0,1) \) for \( \nu=0, 1, 2, 1000 \) (note that the solid line \( \nu = 1000 \) is a step function) (c) The skew normal type 1 distribution, \( Z \sim \text{SN1}(0,1,\nu) \) for \( \nu=0, 1, 2, 1000 \). .... 181
11.2 Splicing method \( Y \sim \text{SN2}(0,1,\nu) \) for \( \nu=1, 2, 3, 5 \). Switching from \( \nu \) to \( 1/\nu \) reflects \( f_Y(y) \) about \( y = 0 \) .......................................................... 184

12.1 Figure showing the log of the standardised version of the normal, Cauchy and Laplace distributions .......................................................... 198
12.2 Figure showing the log of the standardised version of the normal, Cauchy and Laplace distributions .......................................................... 199
12.3 Figure showing the shape of the tail for different types of distributions for \( k_1, k_3, k_5 = 1, 2, \) and \( k_2, k_4, k_6 = 1, 2, \) Smaller values i the k’s result heavier tails .......................................................... 201
12.4 Exploratory Method 1 applied to the 90’s film revenues data 207
12.5 Exploratory Method 2 applied to the 90’s film revenues data 208
12.6 QQ plot for the truncated Weibull 209
12.7 Sequencial plot of $\hat{\sigma}$ for the truncated Weibull 210
13.1 The upper boundary of centile central skewness against the transformed centile kurtosis for six distributions on the real line 217
13.2 Contours of centile central skewness against the transformed centile kurtosis for constant values of $\nu$ and $\tau$ for the SHASHo and SB distributions 219
13.3 The upper boundary of centile tail skewness against the transformed centile kurtosis for six distributions on the real line 220
13.4 Contours of centile tail skewness against the transformed centile kurtosis for constant values of $\nu$ and $\tau$ for the SHASHo and SB distributions 222
List of Tables

1.1 Continuous distributions implemented within the `gamlss.dist` package (with default link functions) .............................................. 38
1.2 Discrete distributions implemented within the `gamlss` packages (with default link functions) .................................................. 39
1.3 Mixed distributions implemented within the `gamlss` packages (with default link functions) ....................................................... 39

2.1 Continuous GAMLSS family distributions defined on \((−∞, +∞)\) ......................... 44
2.2 Continuous GAMLSS family distributions defined on \((0, +∞)\) ......................... 51
2.3 Continuous GAMLSS family distributions defined on \((0, 1)\) ......................... 57

3.1 Discrete `gamlss.family` distributions for count data (derived from Poisson mixtures) ............................................................... 62
3.2 Discrete `gamlss.family` distributions for count data .................................. 63

6.1 Table showing the expansion of data for fitting the multinomial model at M step (2) of the EM algorithm ........................................... 98
6.2 Table showing the expansion of data use in M-step (1) of the EM algorithm for fitting the common parameter mixture model ..................... 101

10.1 The first year student examination results data where the binomial denominator is constant at 8 ...................................................... 176

11.1 Showing distributions generated by Azzalini type methods using equation (11.4) 182
11.2 Showing distributions generated by splicing .......................................... 185
11.3 Showing distributions generated by continuous mixtures .......................... 187
11.4 Showing distributions generated by univariate transformation ..................... 189

12.1 Left and right tail asymptotic form of the log of the probability density function for continuous distributions on the real line, where \(c_1^2 = \Gamma\left(\frac{1}{\nu}\right) \left[\Gamma\left(\frac{3}{\nu}\right)\right]^{-1}\) ......................... 202
12.2 Right tail asymptotic form of the log of the probability density function for continuous distributions on the positive real line, where \(c_2 = \left[K_{\nu+1}\left(\frac{1}{\nu}\right)\right]\left[K_{\nu}\left(\frac{1}{\nu}\right)\right]^{-1}\)

\[
K_{\lambda}(t) = \frac{1}{2} \int_{0}^{\infty} x^{\lambda-1} \exp\left\{-\frac{1}{2} t (x + x^{-1})\right\} \, dx
\]

where \(K_{\lambda}(t) = \frac{1}{2} \int_{0}^{\infty} x^{\lambda-1} \exp\left\{-\frac{1}{2} t (x + x^{-1})\right\} \, dx
\) ................................................................................................................. 203

12.3 Mandlebrot’s classification of randomness ............................................. 204
12.4 Asymptotic form of \(\log \tilde{F}_Y(y)\) as \(y \to \infty\) .............................................. 204
12.5 Showing possible relationships of the log\([F_Y(y)\)] against \(t = \log y\) for Method 1 . 205
12.6 Showing possible relationships of the log$[\bar{F}_Y(y)]$ against $t = \log y$ for Method 1 . 208
12.7 Estimated coefficients from exploratory method 2. . . . . . . . . . . . . . . . . . 208
12.8 References for continuous distributions . . . . . . . . . . . . . . . . . . . . . . . 211

No further action is required
Part I

The GAMLSS Family
Chapter 1

Types of univariate parametric distributions

This chapter provides:
1. the definition of a statistical distribution
2. an introduction to properties of the distribution functions
3. an introduction to different types of distributions within GAMLSS
4. the implementation of distributions in \( \mathbb{R} \), the \( \text{d,p,q,r} \) representation

This chapter is essential for understanding the different types of distributions in GAMLSS and especially the need for more complex distributions.

1.1 Introduction

*Statistical modelling* is the art of building parsimonious statistical models for a better understanding of the phenomena of interest. A statistical model deals with uncertainties and therefore it contains a *stochastic* or *random* part, that is, a component which is trying to describe the uncertain nature of the phenomenon. The main instrument to deal with uncertainty in mathematics is the concept of *probability*. All statistical models incorporate some probabilistic assumptions.

Probabilities, denoted here as \( \text{Pr}() \), are set functions defined on *events*, \( (E) \), which are subsets of the sample space of an experiment, \( S \), and have the properties:

- \( \text{Pr}(E) \geq 0 \)

- \( \text{Pr}(E_1 \cup E_2) = \text{Pr}(E_1) + \text{Pr}(E_2) \) where \( E_1 \) and \( E_2 \) are events which have no common outcomes (i.e. disjoint).

- \( \text{Pr}(S)=1 \)

That is, probabilities take values from zero to one. If an event has a probability of zero then this event will never occur while an event with probability of one is a certain event.
A random variable $Y$ is a function $Y()$ that maps the sample space $S$ of an experiment onto a space of real numbers, i.e. it assigns to each element $s \in S$ one real number $Y = Y(s)$. The reason for doing this is because it is easier to deal with numbers rather than events of a sample space. For example, let us consider the simple experiment of tossing a coin. The sample space is $S = \{T, H\}$ where $T$ represents tail and $H$ represents head. By assigning $Y(T) = 0$ and $Y(H) = 1$ we map the experiment sample space 'head' and 'tail', to the random variable $Y$ with range $R_Y = \{0, 1\}$. The range $R_Y$ of a random variable $Y$ is the set of values that the random variable $Y$ can take. The random variable $Y$ takes only two values 0 and 1 and we have that $Pr(Y = 0) + Pr(Y = 1) = 1$. For a fair coin $Pr(Y = 0) = Pr(Y = 1) = \frac{1}{2}$.

The values that a random variable $Y$ can take, i.e. range $R_Y$, can be:

- a discrete set of real numbers: as in the tossing a coin example above or
- a continuous set of real numbers: for example $Y$ could take any possible values between zero and one.

The most common ranges $R_Y$ for a continuous random variable $Y$ are:

- the real line: $\mathbb{R} = (-\infty, +\infty)$
- the positive real line: $\mathbb{R}^+ = [0, +\infty)$
- values between zero and one: $\mathbb{R}_0^1 = [0, 1]$.

The most common ranges $R_y$ for a discrete random variable $Y$ are:

- the binary values: $\{0, 1\}$
- the binomial values: $\{0, 1, 2, \ldots, n\}$
- the count values: $\{0, 1, 2, \ldots, \infty\}$

Note that the binary values are a special case of the binomial values.

### 1.1.1 Probability (density) functions

A distribution of a random variable $Y$ determines how likely different values of $Y$ are. A discrete distribution is defined by a probability function (pf), while a continuous distribution is defined by a probability density function (pdf).

Let $Y$ be a discrete random variable which takes values within a discrete set of values in $\mathbb{R}$, say $R_Y$ denoting the range of $Y$. Let $f_Y(y) = Pr(Y = y)$, that is $f_Y(y)$ represents the probability that $Y$ is equal to specific value $y$. The function $f_Y(y)$ is said to be a proper probability function of the discrete variable $Y$ if the function $f_Y(y)$ is positive for all values of $y$ within $R_Y$ and if

$$\sum_{R_Y} f_Y(y) = 1. \tag{1.1}$$

Let $Y$ be a continuous random variable which takes values within a subset of $\mathbb{R}$, say $R_Y$. The function $f_Y(y)$ is said to be a proper probability density function of the continuous variable $Y$ if $f_Y(y)$ is positive for all values of $y$ within $R_Y$ and if

$$\int_{R_Y} f_Y(y)dy = 1, \tag{1.2}$$
1.1. INTRODUCTION

i.e. the integral over \( \mathbb{R} \) equals one.

For convenience let \( f_Y(y) = 0 \) for all \( y \) in the real line \( \mathbb{R} \) not in \( \mathbb{R}_Y \). Then \( f_Y(y) \) is defined on \( y \in \mathbb{R} \).

Note here the peculiarity that \( \Pr(Y = y_0) = f_Y(y_0) = 0 \), for any arbitrary value \( y_0 \) of a continuous \( Y \). That is, the probability of a continuous random variable \( Y \) to be equal any specific value is equal to zero. This is because integrals are not defined on a single point but on an interval. In practice this is circumvented by defining the probability on a small interval \( (y_o - \Delta y, y_o + \Delta y) \) around \( y_o \) where \( \Delta y \) has a small value. Then \( \Pr(Y \in (y_o - \Delta y, y_o + \Delta y)) = \int_{y_o-\Delta y}^{y_o+\Delta y} f_Y(y)dy \) is properly defined.

Example 1: the exponential distribution

The function \( f_Y(y) = e^{-y} \), defined on the positive real line \( \{0, \infty\} \) is a proper pdf since \( e^{-y} > 0 \) for \( y \in (0, \infty) \) and \( \int_0^{\infty} e^{-y}dy = 1 \). Its flexibility is enhanced by introducing a single parameter in the definition of the pdf, i.e. \( f_Y(y) = \theta e^{-\theta y} \) where \( \theta > 0 \). It is easy to show that \( \int_0^{\infty} \theta e^{-\theta y}dy = 1 \) so the new function is still a pdf. Because now the pdf \( f_Y(y) \) contains the parameter \( \theta \) it is called a parametric probability density function. The key advantages of introducing the parameter \( \theta \) it that my making the shape of the distribution more flexible it can be used for modelling a set of of historical observations. Finding a suitable parameter value for \( \theta \) for a given set of data is an example of statistical inference.

In order to emphasise the fact that the pdf depends on \( \theta \) we write it as \( f_Y(y|\theta) \) for \( \theta > 0 \). The notation \( f_Y(y|\theta) \) now represents, for different values of \( \theta \), a family of pdf's, see for example Figure 1.1. Note that if instead of parameter \( \theta \) we chose any other one to one function of \( \theta \), e.g. \( \mu = \frac{1}{\theta} \) the pdf family remains the same. For example, equation

\[
f_Y(y|\theta) = \theta e^{-\theta y}
\]  

\( \hat{1.3} \)
and equation

$$f_Y(y|\mu) = \frac{1}{\mu} e^{-\frac{y}{\mu}}$$  \hspace{1cm} (1.4)

define the same pdf family (the exponential distribution). They are different parameterisations of the same family. We shall see later that some parameterisations are preferable to others in practice. The R code for generating Figure 1.1 is given in section 1.2

**Example 2: the Poisson distribution**

The function

$$f(y|\mu) = P(Y = y|\mu) = \frac{e^{-\mu} \mu^y}{y!}$$ \hspace{1cm} (1.5)

where \( y = 0, 1, 2, \ldots, \infty \) defines the probability function of one of the most popular discrete distributions, the Poisson distribution. Figure 1.2 shows how the distribution shape changes with different values of the parameter \( \mu \).

![Figure 1.2: Showing how the shape of the Poisson distribution changes with different values of \( \mu = 1, 3, 5, 15 \).](image)

The R code for generating Figure 1.2 is given in section 1.2

**Summarising**

- Any non-negative function \( f_Y(y) \) which sums or integrates over the sample space to one can be a probability (density) function.
- Probability (density) functions containing say \( k \) parameters \( \theta = (\theta_1, \theta_2, \ldots, k) \) are called parametric and are denoted \( f(y|\theta) \).
1.1. INTRODUCTION

- A one to one re-parameterisation does not change the pdf family.
- Parameters effect the 'shape' of the distribution.

![Diagram showing relationship between pdf and cdf functions on a continuous distribution](image)

Figure 1.3: Showing relationship between the pdf and cdf functions on a continuous distribution. The area under the curve of the pdf until the point $y$ is the value of the cdf at this point $F_Y(y)$.

### 1.1.2 Cumulative distribution function and its inverse

In this section we omit the conditioning on $\theta$ for simplicity of notation.

**cdf function**

The cumulative distribution function (cdf) is defined as the $F_Y(y) = Pr(Y \leq y)$ and is

$$F_Y(y) = \sum_{w \leq y} f_Y(w) \quad (1.6)$$

for a discrete random variable, where the sum is over all $w \in \mathbb{R}_Y$ for which $w \leq y$, and

$$F_Y(y) = \int_{-\infty}^{y} f_Y(w)dw \quad (1.7)$$

for a continuous random variable [since for convenience $f_Y(y) = 0$ for all $y$ in the real line $\mathbb{R}$ not in $\mathbb{R}_Y$].

Figure 1.3 shows the relationship between the pdf and cdf functions for a continuous random variable $Y$. The area under the curve of the pdf up to the point $y$ is the value of the cdf at $y$, i.e. $F_Y(y)$. In fact the pdf is the derivative function of the cdf with respect to $Y$, i.e. $f_Y(y) = dF_Y(y)/dy$. Note that the cdf function domain is $y \in \mathbb{R}_Y$ the same as the pdf. (The domain of a function is the range where it takes it values). The range of values of the cdf function is from zero to one.

For a discrete random variable the cdf is a step function, see for example the lower part of Figure 1.4 where the Poisson distribution pdf and cdf are plotted.
24  \hspace{1cm} \textit{CHAPTER 1. TYPES OF UNIVARIATE PARAMETRIC DISTRIBUTIONS}

Figure 1.4: Showing relationship between the pdf and cdf functions on a discrete distribution. The sum of probabilities until \( y \) in the probability function is the value of the cdf at this point \( F_Y(y) \)

\textbf{Inverse cdf function}

For a continuous random variable \( Y \), the inverse of the cumulative distribution function \( F_Y^{-1}(\cdot) \) given by \( y_p = F_Y^{-1}(p) \) for \( 0 < p < 1 \) is defined by \( y_p \) satisfying \( Pr(Y \leq y_p) = p \), i.e. \( F_Y(y_p) = p \). See the lower left plot of Figure 1.5 for an example of a continuous inverse cdf plot.

For a discrete random variable \( Y \), \( y_p = F_Y^{-1}(p) \) is defined by \( y_p \) satisfying \( Pr(Y < y_p) < p \) and \( Pr(Y \leq y_p) \geq p \), i.e. \( y_p \) is the smallest value of \( y \) for which \( Pr(Y \leq y_p) \geq p \). See the lower left plot of Figure 1.6 for an example of a discrete inverse cdf plot.

\subsection*{1.1.3 Survival and hazard functions}

The \textit{survival function} \( S_Y(y) \) is the probability of \( Y \) ‘surviving’ beyond \( y \), i.e. \( S_Y(y) = Pr(Y > y) = 1 - Pr(Y \leq y) = 1 - F_Y(y) \). Alternative names for the survival function are the \textit{complementary cumulative distribution function} (ccdf), the \textit{tail distribution}, \textit{exceedance}, or the \textit{reliability function} (common in engineering).

The \textit{hazard function} \( h_Y(y) \) is the instantaneous likelihood of ‘dying’ at \( y \) given survival up to \( y \). i.e.

\[ h_Y(y) = \lim_{\delta y \to 0} \frac{Pr(y < Y \leq y + \delta y | Y > y)}{\delta y} = \frac{f_Y(y)}{S_Y(y)} \tag{1.8} \]

\subsection*{1.2 Population distributions in R: the d, p, q and r functions}

All releases of the R statistical environment contain several popular distributions. R uses the \texttt{d}, \texttt{p}, \texttt{q} and \texttt{r} convention where
1.2. POPULATION DISTRIBUTIONS IN R: THE D, P, Q AND R FUNCTIONS

d is the probability density function
p is the cumulative distribution function
q is the inverse of the cumulative distribution function and
r refers to a function which generates random numbers from the distribution

For example \texttt{dnorm}, \texttt{pnorm} \texttt{qnorm} and \texttt{rnorm} are the pdf, cdf, inverse cdf and random generating function of the normal distribution respectively.

The following continuous distributions come as standard in \texttt{R}:

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{beta}</td>
<td>the beta distribution</td>
</tr>
<tr>
<td>\texttt{cauchy}</td>
<td>the Cauchy distribution</td>
</tr>
<tr>
<td>\texttt{chisq}</td>
<td>the chi-squared distribution</td>
</tr>
<tr>
<td>\texttt{exp}</td>
<td>the exponential distribution</td>
</tr>
<tr>
<td>\texttt{f}</td>
<td>the F distribution</td>
</tr>
<tr>
<td>\texttt{gamma}</td>
<td>the gamma distribution</td>
</tr>
<tr>
<td>\texttt{lnorm}</td>
<td>the log-normal distribution</td>
</tr>
<tr>
<td>\texttt{norm}</td>
<td>the normal distribution</td>
</tr>
<tr>
<td>\texttt{t}</td>
<td>the Student’s ( t ) distribution</td>
</tr>
<tr>
<td>\texttt{unif}</td>
<td>the uniform distribution</td>
</tr>
<tr>
<td>\texttt{weibull}</td>
<td>the Weibull distribution</td>
</tr>
</tbody>
</table>

Also the following discrete distributions:

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{binom}</td>
<td>the binomial (including Bernoulli) distribution</td>
</tr>
<tr>
<td>\texttt{geom}</td>
<td>the geometric distribution</td>
</tr>
<tr>
<td>\texttt{hyper}</td>
<td>the hypergeometric distribution</td>
</tr>
<tr>
<td>\texttt{multinom}</td>
<td>the multinomial distribution</td>
</tr>
<tr>
<td>\texttt{nbinom}</td>
<td>the negative binomial distribution</td>
</tr>
<tr>
<td>\texttt{pois}</td>
<td>the Poisson distribution</td>
</tr>
</tbody>
</table>

This book is about distributions in the \texttt{gamlss.dist} package but to see what distributions other packages contain use the following URL link \url{http://cran.r-project.org/web/views/Distributions.html}.

The \texttt{gamlss.dist} package follows the \( d, p, q \) and \( r \) convention but the parametrisation of the distribution that follows can be different from the parametrisation used in the above \texttt{R} distributions. The main reason for that is because the \texttt{gamlss.dist} package is build to support the GAMLSS regression framework. When you build a regression model you need to have a sound interpretation of the parameter you model as a function of explanatory variables.

Graphical representation of the the \( d, p, q \) and \( r \) functions in \texttt{R} can be done easily. Here is an example of how to plot the pdf, cdf, inverse cdf and the random generating sample function.

To create Figure 1.5 an example of a continuous distribution (the gamma) use the following:

```R
PPP <- par(mfrow=c(2,2))
curve(dgamma(x, shape=2), 0.01, 25, ylab="pdf") # pdf
curve(pgamma(x, shape=2), 0.01, 25, ylab="cdf") # cdf
plot(function(x) qgamma(x, , shape=2), 0, 1, , ylab="invcdf") # inverse cdf
y<-rgamma(100, shape=2 ) # randomly generated values
hist(y)
par(PPP)
```
To create figure 1.6 an example of a discrete distribution (the negative binomial) use the following code:

```r
PPP <- par(mfrow=c(2,2))
plot(function(y) dnbinom(y, size=10, p=.5), from=0, to=40, n=40+1, type="h",
     main="pdf", ylab="pdf(x)")
cdf <- stepfun(0:39, pnbinom(0:40, size=10, p=.5), f = 0)
plot(cdf, main="cdf", ylab="cdf(x)", do.points=FALSE)
invcdf <- stepfun(seq(0.01,.99,length=39), qnbinom(seq(0.1,.99,length=40),
                      size=10, p=.5), f = 0)
plot(invcdf, main="inverse cdf", ylab="inv-cdf(x)", do.points=FALSE)
tN <- table(Ni <- rnbinom(1000, size=10, p=.5))
r <- barplot(tN, col='lightblue')
par(PPP)
```

The survival function of any implemented distribution in R can be obtained by setting the `lower.tail` argument of the `p` function to `FALSE`. For example

```r
curve(pgamma(x, shape=2,lower.tail=FALSE ), 0.01, 10, ylab="cdf")
```

will plot the survival function of the gamma distribution above.

Here we present the R code for generating the exponential distribution example in Figure 1.1:

```r
xmax <- 6
x <- seq(0.001, xmax, length=201)
```
1.2. POPULATION DISTRIBUTIONS IN R: THE D, P, Q AND R FUNCTIONS

Figure 1.6: Showing the probability function, commutative distribution function, inverse commutative distribution function and a random sample of 1000 observations from a negative binomial distribution

xx <- rep(0, 201)
yy <- seq(-1,3.5, length=201)
plot(dEXP(x, mu=1 )~x, xlab="Y", ylab=expression(f[Y](y)), type="l",
    ylim=c(0,4), xlim=c(0,xmax), lwd=2, main="The exponential distribution")
lines(dEXP(x, mu=2 )~x, xlab="Y", col="red", lty=2, lwd=2)
lines(dEXP(x, mu=.5)~x, xlab="Y", col="blue", lty=3, lwd=2)
lines(dEXP(x, mu=.25)~x, xlab="Y", col="green4", lty=4, lwd=2)
col <- c("black", "red", "blue", "green4")
ex.leg <- expression(paste( \( \mu = 1 \)), paste( \( \mu = 2 \)), paste( \( \mu = 0.5 \)), paste( \( \mu = 0.25 \)))# 2 ways
abline(h=-1:3, v=1:6, col="gray90")
legend(4, 2.5, ex.leg, lty=c(1,2,3,4), col=col, adj = c(0, .6))

The Poisson example in Figure 1.2 was generated using the following R commands:

op <- par(mfrow=c(2,2))
plot(Y, dP0(x=Y, mu=1), type="h", ylab="Poisson pdf")
text(25,.2, expression(paste( \( \mu = 1 \))))
plot(Y, dP0(x=Y, mu=3), type="h", ylab="Poisson pdf")
text(25,.13, expression(paste( \( \mu = 3 \))))
plot(Y, dP0(x=Y, mu=5), type="h", ylab="Poisson pdf")
text(25,.1, expression(paste( \( \mu = 5 \))))
plot(Y, dP0(x=Y, mu=15), type="h", ylab="Poisson pdf")
text(25,.06, expression(paste( \( \mu = 15 \))))
par(op)
1.3 Properties of the distributions

In this section we are concerned with some general properties of the distributions including measures of location, scale, skewness and kurtosis. For parametric distributions it is common practice that some of the parameters are related to these measures. One of the great advantages of the GAMLSS framework (and its associated distributions in package `gamlss.dist`) is its ability to fit a variety of different distributions to the response variable, so that an appropriate distribution can be chosen among different alternatives. It is assumed that the response variable $Y$ comes from a population distribution which can be modelled by a theoretical probability (density) function $f_Y(y|\theta)$, where the parameter vector $\theta$ can be any number of parameters, but in practice is usually limited to up to four dimensions, i.e. $\theta^T = (\mu, \sigma, \nu, \tau)$. Our task is to find the appropriate distribution $f_Y(y|\theta)$ and estimate the parameters $\theta$. Limiting the parameter vector $\theta$ to four dimensions, i.e. $\theta^T = (\mu, \sigma, \nu, \tau)$ is not usually a serious restriction. The parameters $\mu$, $\sigma$, $\nu$ and $\tau$ usually represent location, scale, skewness and kurtosis parameters respectively, although more generally they can be any parameters of a distribution. A location parameter usually represents the ‘centre’ of the distribution, and is often the mean, median or mode of the distribution. A scale parameter usually represents the ‘spread’ of the distribution and is often the standard deviation or coefficient of variation of the distribution.

A skewness or kurtosis parameter usually strongly affects the skewness or kurtosis of the distribution respectively. The skewness is a measure of asymmetry of the distribution of the variable. In general, a distribution with a longer tail to the right than the left has positive skewness, while one with a longer tail to the left has negative skewness, and a symmetric distribution has zero skewness. Kurtosis is a measure of heavy tails. A distribution with a heavy (i.e. fat) tails (relative to a normal distribution) will have high kurtosis, i.e. leptokurtosis, while a distribution with short (i.e. thin) tails will have low kurtosis, i.e. platykurtosis. Note that leptokurtic and platykurtic distributions are judged by comparison to the normal distribution which is mesokurtic. Skewness and kurtosis measures can be either moment or centile based, see sections 1.3.1 and 1.3.2 respectively.

Distributions can be symmetric, negatively or positively skewed, and also mesokurtic, leptokurtic or platykurtic. Figure 1.7 shows typical i) negatively skew, ii) positively skew, iii) platykurtic and iv) leptokurtic continuous distributions.

1.3.1 Mean, variance and moment based measures of skewness and kurtosis

mean

The population mean (or average or expected) value of $Y$ is denoted by $E(Y)$ is given by expectation operator:

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y | \theta) dy$$

for a continuous variable $Y$ [ since $f_Y(y|\theta) = 0$ for all $y$ in the real line $\mathbb{R}$ not in $R_Y$ ], and

$$E(Y) = \sum_{y \in R_Y} y P(Y = y | \theta)$$

for a discrete variable $Y$. Note that $E(Y)$ is conditional on $\theta$, i.e. $E(Y|\theta)$, but this conditioning is omitted for simplicity of presentation in $E(Y)$ above and in subsequent expected values. Note that the parameter $\mu$ is not always the mean of $Y$, so in general $\mu \neq E(Y)$. 
1.3. PROPERTIES OF THE DISTRIBUTIONS

Figure 1.7: Showing different types of continuous distributions
More generally the population mean (or expected value) of a function $g(Y)$ of $Y$ is given by

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f_Y(y \mid \theta)\,dy$$

for a continuous variable $Y$ and

$$E[g(Y)] = \sum_{y \in RV} g(y)P(Y = y \mid \theta)$$

for a discrete variable $Y$.

![Figure 1.8: Showing the difference between the mean median and mode of a positive skew distribution](image)

properties of expectations

Let, $Y$, $Y_1$ and $Y_2$ be random variables and $a$, $b$ and $c$ constants:

- $E(c) = c$
- $E(cY) = cE(Y)$
- $E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$
- $E(aY_1 + bY_2) = aE(Y_1) + bE(Y_2)$

moments

The $k^{th}$ population central moment (or $k^{th}$ moment about the mean) of $Y$ is given

$$\mu_k = E\left(\left[Y - E(Y)\right]^k\right)$$

for $k = 2, 3, \ldots$.

Note that $\mu_k$ should not be confused with the distribution parameter $\mu$. 
1.3. PROPERTIES OF THE DISTRIBUTIONS

**Variance and standard deviation**

The population variance of $Y$ is given by

$$V(Y) = \mu_2 = \int_{-\infty}^{\infty} (y - E(Y))^2 f_Y(y | \theta) dy$$

for a continuous variable $Y$ and

$$V(Y) = \mu_2 = \sum_{y \in \mathbb{R}^Y} (y - E(Y))^2 P(Y = y | \theta)$$

for a discrete variable $Y$.

Hence the variance of $Y$ is $V(Y) = \mu_2$ and the standard deviation of $Y$ is $\sqrt{V(Y)} = \sqrt{\mu_2}$. The standard deviation gives a measure of the spread of the distribution of the variable.

**Moment based skewness and kurtosis**

The (moment based) population skewness of $Y$ is given by

$$\sqrt{\beta_1} = \mu_3 / (\mu_2)^{1.5},$$

the (moment based) population kurtosis of $Y$ is given by

$$\beta_2 = \mu_4 / (\mu_2)^2,$$

and the (moment based) excess kurtosis of $Y$ is given by $\beta_2 - 3$ and measures the excess kurtosis relative to a normal distribution which has kurtosis $\beta_2 = 3$. Note that $\mu_3$ and $\mu_4$ are defined in a similar fashion to $\mu_2$ above.

One of the problems with moment based summary statistics of the distributions is the fact that in several cases moments do not exist. This is common for fat tail distributions (that is, leptokurtic distributions). The centile based summary statistics described in the next section avoid this problem because they always exist.

**Properties of variances**

Let, $Y$, $Y_1$ and $Y_2$ be random variables and $a$, $b$ and $c$ constants:

- $V(Y) = E\left[ (Y - E(Y))^2 \right] = E(Y^2) - [E(Y)]^2$
- $V(c) = 0$
- $V(cY_1) = c^2V(Y_1)$
- if $Y_1$ and $Y_2$ are independent then $V(Y_1 + Y_2) = V(Y_1) + V(Y_2)$
- if $Y_1$ and $Y_2$ are independent then $V(aY_1 + bY_2) = a^2V(Y_1) + b^2V(Y_2)$
CHAPTER 1. TYPES OF UNIVARIATE PARAMETRIC DISTRIBUTIONS

moment generating function

The moment generating function of $Y$ is given by

$$M_Y(t) = E \left[ e^{tY} \right]$$

provided it exists. It is called the moment generating function because setting $t = 0$ in its $k^{th}$ derivative with respect to $t$ gives $E \left[ Y^k \right]$. i.e.

$$E \left[ Y^k \right] = \left. \frac{d^k M_Y(t)}{dt^k} \right|_{t=0}$$

A specific example? why is it important to know the moment generating function?

probability generating function

The probability generating function of $Y$ for a discrete random variable $Y$ is given by

$$P_Y(s) = E \left[ s^Y \right]$$

provided it exists. It is called the probability generating function because setting $s = 0$ in its $k^{th}$ derivative with respect to $s$ gives

$$P(Y = k) = \left. \frac{1}{k!} \frac{d^k P_Y(s)}{ds^k} \right|_{s=0}$$

A specific example? why is it important to know the probability generating function? how is it related with chapter 4?

1.3.2 Centiles and centile based measures of spread, skewness and kurtosis

Here we omit the conditioning on $\theta$ for simplicity of notation, hence $F_Y(y) = \int_{-\infty}^{y} f_Y(\omega)d\omega$ for a continuous variable, and $F_Y(\omega) = \sum_{\omega \leq y} P(Y = \omega)$ for a discrete variable.

Centile and Quantile

For a continuous random variable $Y$, the $100p^{th}$ centile (or $p^{th}$ quantile) of $Y$ is the value $y_p$ of $Y$ for which the probability of being at or below $y_p$ is $p$, and is defined by

$$P(Y \leq y_p) = p$$

i.e. $F_Y(y_p) = p$ and hence $y_p = F_Y^{-1}(p)$, where $F_Y^{-1}(p)$ is the inverse cumulative distribution function of $Y$ evaluated at $0 < p < 1$. Hence the centile or quantile function is the inverse cumulative distribution function.

For example the $5^{th}$ centile, $y_{0.05}$ is the value of $Y$ for which the probability of being at or below $y_{0.05}$ is 0.05 (i.e. 5%) and is defined by $P(Y \leq y_{0.05}) = 0.05$, i.e. $F_Y(y_{0.05}) = 0.05$, i.e. $y_{0.05} = F_Y^{-1}(0.05)$.

Strictly for a discrete random variable $Y$, $y_p = F_Y^{-1}(p)$ is defined by $P(Y < y_p) < p$ and $P(Y \leq y_p) \geq p$, i.e. $y_p$ is the smallest value of $Y$ for which $P(Y \leq y_p) \geq p$. 

1.3. PROPERTIES OF THE DISTRIBUTIONS

median, $Q_1$ and $Q_3$

For a continuous random variable $Y$, the population median $m$ is the value of $Y$ for which half the population lies at or below $m$, i.e. $P(Y \leq m) = 0.5$. Hence $m = y_{0.5} = F_Y^{-1}(0.5)$.

Similarly the first and third quartiles $Q_1$ and $Q_3$ are given by $P(Y \leq Q_1) = 0.25$ and $P(Y \leq Q_3) = 0.75$ (see Figure 1.9). Hence $Q_1 = y_{0.25} = F_Y^{-1}(0.25)$ and $Q_3 = y_{0.75} = F_Y^{-1}(0.75)$.

Note the corresponding strict definitions of $m = F_Y^{-1}(0.50)$, $Q_1 = F_Y^{-1}(0.25)$ and $Q_3 = F_Y^{-1}(0.75)$ for a discrete random variable $Y$, i.e. $m, Q_1$ and $Q_3$ are the smallest value of $Y$ for which $P(Y \leq m) \geq 0.5$, $P(Y \leq Q_1) \geq 0.25$ and $P(Y \leq Q_3) \geq 0.75$ respectively, see Figure 1.11.

![Figure 1.9: Showing the Q1, the median and Q3.](image)

centile based skewness and kurtosis

Centile based measures of spread, skewness and kurtosis are given by the following.

The interquartile range is $IR = (Q_3 - Q_1)$. The semi-interquartile range $SIR = (Q_3 - Q_1)/2$ is a measure of spread. Galton’s skewness measure $\gamma$ is the mid-quartile minus the median divided by the semi-interquartile range, i.e. $\gamma = [(Q_3 - Q_1)/2 - m]/SIR$.

A centile based kurtosis measure $t_{0.49}$ is given by $t_{0.49} = [y_{0.99} - y_{0.01}]/IR$, Andrews et al. (1972). This measure has been scaled to be one for the normal distribution, giving $st_{0.49} = t_{0.49}/3.449$, Rossenberger and Gasco (1983).

The median and first and third quartiles for any GAMLSS distribution (continuous or discrete) can be found from its $q$ function evaluated at $p=0.5, 0.25$ and $0.75$ respectively for chosen parameter values. For example the median and first and third quartiles of a normal distribution with mean 5 and standard deviation 2 are given by $m = qNO(0.5, 5, 2)$, $Q_1 = qNO(0.25, 5, 2)$ and $Q_3 = qNO(0.75, 5, 2)$. 
Figure 1.10: Showing how $Q_1$, the median and $Q_3$ are derived in a continuous distribution.

Figure 1.11: Showing how $Q_1$, the median and $Q_3$ are derived in a discrete distribution.
1.3.3 Mode

The mode(s) of a random variable $Y$ is the value (or values) of $Y$ which have the highest pdf, $f_Y(y)$, for a continuous variable, and the highest pf, $P(Y = y)$, for a discrete variable.

1.4 Types of distribution within the GAMLSS family

Within the GAMLSS framework the population probability (density) function of the response variable $Y$, $f_Y(y|\theta)$, where $\theta = (\mu, \sigma, \nu, \tau)$, is deliberately left general with no explicit distribution specified. The only restriction that the R implementation of GAMLSS, Stasinopoulos and Rigby (2007), has for specifying the distribution of $Y$ is that the function $f_Y(y|\theta)$ and its first (and optionally expected second and cross) derivatives with respect to each of the parameters of $\theta$ must be computable. Explicit derivatives are preferable, but numerical derivatives can be used (resulting in reduced computational speed). That is, the algorithm used for fitting the regression model needs only this information.

Here we introduce the available distributions within the current implementation of GAMLSS in R. We refer to this set of distributions as the GAMLSS family to be consistent with R where the distributions are defined as `gamlss.family` objects.

Fitting a parametric distribution within the GAMLSS family can be achieved using the command `gamlss(y ~ 1, family="")` where the argument `family` can take any `gamlss.family` distribution, see Tables 1.1, 1.2 and 1.3 for appropriate names. Other ways of fitting a parametric distribution will be discuss in Chapter 8).

The type of distribution to use depends on the type of the response variable. Within the GAMLSS family there are three distinct types of distributions:

1. continuous distributions see Figure 1.12,
2. discrete distributions, see Figure 1.13,
3. mixed distributions see Figure 1.14.

All the continuous distributions in `gamlss.family` are shown in Table 1.1. The columns of the Table shows the names, the `gamlss.family` names and the default link functions of the distribution respectively.

Link functions were introduced by Weddeburn and Nelder (1972) for Generalised Linear Models but are appropriate for all regression models since they guarantee that the distribution parameters remain within the appropriate range. For example take the beta distribution in the first row of Table 1.1. Both $\mu$ and $\sigma$ are defined on $(0, 1)$, see Section 16.1.1. The logit link for $\mu$ uses the predictor $\eta = \log(\frac{\mu}{1-\mu})$ for fitting the $\mu$ parameter in the model. Therefore $\mu = \frac{e^\eta}{1+e^\eta}$ which ensures that $\mu$ is in the right range from 0 to 1.

The continuous distributions, $f_Y(y|\theta)$ in Table 1.1 are defined on $(-\infty, +\infty)$, $(0, +\infty)$ and $(0, 1)$, ranges. Users can restrict those ranges of the response variable $Y$ by defining a truncated `gamlss.family` distribution using the package `gamlss.tr`. Any GAMLSS distribution on $(-\infty, \infty)$ can be transformed to a distribution either on $(0, \infty)$ or $(0, 1)$ see section 11.5.

Discrete distributions $P(Y = y|\theta)$ are usually defined on $y = 0, 1, 2, \ldots, n$, where $n$ is a known finite value or $n$ is infinite such as discrete (count) values. Table 1.2 shows the available discrete `gamlss.family` distributions.

Mixed distributions are a special case of finite mixture distributions described in Chapter 6 and are mixtures of continuous and discrete distributions, i.e. continuous distributions where the range of $Y$ has been expanded to include some discrete values with non-zero probabilities. Those
Figure 1.12: Showing the normal distributions, an example of a continuous distribution

Figure 1.13: Showing the Poisson distributions, an example of a discrete distribution
1.4. TYPES OF DISTRIBUTION WITHIN THE GAMLLS FAMILY

distributions are useful for modelling data like insurance claims where most of the people are not claiming therefore there is a high probability at zero but if they claim then the distribution on the amount of claim is defined in the positive real line. The zero inflated gamma distribution shown in Figure 1.14 is a possible distribution for such data. Table 1.3 shows the available mixed `gamlss.family` distributions.

![Zero adjusted GA pdf](image)

![Zero adjusted Gamma c.d.f.](image)

Figure 1.14: Showing the zero adjusted gamma distributions, an example of a mixed distribution

<table>
<thead>
<tr>
<th>Distributions</th>
<th>R Name</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\nu$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta</td>
<td>BE()</td>
<td>logit</td>
<td>logit</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Box-Cox Cole-Green</td>
<td>BCCG()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>-</td>
</tr>
<tr>
<td>Box-Cox power exp.</td>
<td>BCPE()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>Box-Cox t</td>
<td>BCT()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>exponential</td>
<td>EXP()</td>
<td>log</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>exponential Gaussian</td>
<td>exGAUS()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>-</td>
</tr>
<tr>
<td>exponential gen. beta t2</td>
<td>EGB2()</td>
<td>identity</td>
<td>identity</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>gamma</td>
<td>GA()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>generalised beta type 1</td>
<td>GB1()</td>
<td>logit</td>
<td>logit</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>generalised beta type 2</td>
<td>GB2()</td>
<td>log</td>
<td>identity</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>generalised gamma</td>
<td>GG()</td>
<td>log</td>
<td>log</td>
<td>identity</td>
<td>-</td>
</tr>
<tr>
<td>generalised inv. Gaussian</td>
<td>GIG()</td>
<td>log</td>
<td>log</td>
<td>identity</td>
<td>-</td>
</tr>
<tr>
<td>generalised $t$</td>
<td>GT()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>Gumbel</td>
<td>GU()</td>
<td>identity</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>inverse Gamma</td>
<td>IGAMMA()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>inverse Gaussian</td>
<td>IG()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Johnson’s SU repar.</td>
<td>JSU()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>Johnson’s original SU</td>
<td>JSUo()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>logistic</td>
<td>LO()</td>
<td>identity</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>logistic normal</td>
<td>LOGITNO()</td>
<td>logit</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>log normal</td>
<td>LOGNO()</td>
<td>identity</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>log normal 2</td>
<td>LOGNO2()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>log normal (Box-Cox)</td>
<td>LNO()</td>
<td>identity</td>
<td>log</td>
<td>fixed</td>
<td>-</td>
</tr>
</tbody>
</table>
CHAPTER 1. TYPES OF UNIVARIATE PARAMETRIC DISTRIBUTIONS

<table>
<thead>
<tr>
<th></th>
<th>NET</th>
<th>NET()</th>
<th>identity</th>
<th>log</th>
<th>fixed</th>
<th>fixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>NO</td>
<td>NO()</td>
<td>identity</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>normal family</td>
<td>NOF</td>
<td>NOF()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>-</td>
</tr>
<tr>
<td>Pareto 2 original</td>
<td>PARETO2o</td>
<td>PARETO2o()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pareto 2</td>
<td>PARETO2</td>
<td>PARETO2()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Pareto 2 repar</td>
<td>GP</td>
<td>GP()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>power exponential</td>
<td>PE</td>
<td>PE()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>-</td>
</tr>
<tr>
<td>reverse Gumbel</td>
<td>RG</td>
<td>RG()</td>
<td>identity</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>skew normal type 1</td>
<td>SN1</td>
<td>SN1()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>-</td>
</tr>
<tr>
<td>skew normal type 2</td>
<td>SN2</td>
<td>SN2()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>-</td>
</tr>
<tr>
<td>skew power exp. t1</td>
<td>SEP1</td>
<td>SEP1()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>skew power exp. t2</td>
<td>SEP2</td>
<td>SEP2()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>skew power exp. t3</td>
<td>SEP3</td>
<td>SEP3()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>skew power exp. t4</td>
<td>SEP4</td>
<td>SEP4()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>sinh- arcsinh original</td>
<td>SHASHo</td>
<td>SHASHo()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>sinh-arcsinh original 2</td>
<td>SHASHo2</td>
<td>SHASHo2()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>sinh-arcsinh</td>
<td>SHASH</td>
<td>SHASH()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>skew t type 1</td>
<td>ST1</td>
<td>ST1()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>skew t type 2</td>
<td>ST2</td>
<td>ST2()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>skew t type 3</td>
<td>ST3</td>
<td>ST3()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>skew t type 3 repar</td>
<td>SST</td>
<td>SST()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>logshifto2</td>
</tr>
<tr>
<td>skew t type 4</td>
<td>ST4</td>
<td>ST4()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>log</td>
</tr>
<tr>
<td>skew t type 5</td>
<td>ST5</td>
<td>ST5()</td>
<td>identity</td>
<td>log</td>
<td>identity</td>
<td>log</td>
</tr>
<tr>
<td>t Family</td>
<td>TF</td>
<td>TF()</td>
<td>identity</td>
<td>log</td>
<td>log</td>
<td>-</td>
</tr>
<tr>
<td>t Family repar</td>
<td>TF</td>
<td>TF()</td>
<td>identity</td>
<td>log</td>
<td>logshifto2</td>
<td>-</td>
</tr>
<tr>
<td>Weibull</td>
<td>WEI</td>
<td>WEI()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weibull (PH)</td>
<td>WEI2</td>
<td>WEI2()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Weibull ($\mu$ the mean)</td>
<td>WEI3</td>
<td>WEI3()</td>
<td>log</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

|                          | Table 1.1: Continuous distributions implemented within the gamlss.dist package (with default link functions) |

1.5 Displaying GAMLSS family distributions

Each GAMLSS family distribution has five functions. The 'fitting' function which is used in the argument family of the gamlss() function when fitting a distribution and the usual four R functions, d, p, q and r for the pdf, the cdf, the inverse cdf and the random generating function respectively. The names of the fitting gamlss.family functions are given in column two of Tables 1.1, 1.2 and 1.3 respectively.

For example, the pdf, cdf, inverse cdf and random generating functions of the normal distribution which has the name NO within the gamlss.family are given as dNO, pNO, qNO, rNO respectively.

A gamlss.family population distribution can be displayed graphically in R using the gamlss.demo package. For example, the following commands:

library(gamlss.demo)
### 1.5. Displaying GAMLSS Family Distributions

<table>
<thead>
<tr>
<th>Distributions</th>
<th>R Name</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>$\nu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta binomial</td>
<td>BB()</td>
<td>logit</td>
<td>log</td>
<td>-</td>
</tr>
<tr>
<td>binomial</td>
<td>BI()</td>
<td>logit</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>geometric</td>
<td>GEM()</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>logarithmic</td>
<td>LG()</td>
<td>logit</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Delaporte</td>
<td>DEL()</td>
<td>log</td>
<td>log</td>
<td>logit</td>
</tr>
<tr>
<td>negative binomial type I</td>
<td>NBI()</td>
<td>log</td>
<td>log</td>
<td>-</td>
</tr>
<tr>
<td>negative binomial type II</td>
<td>NBII()</td>
<td>log</td>
<td>log</td>
<td>-</td>
</tr>
<tr>
<td>Poisson</td>
<td>PO()</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Poisson inverse Gaussian</td>
<td>PIG()</td>
<td>log</td>
<td>log</td>
<td>-</td>
</tr>
<tr>
<td>Sichel (the mean)</td>
<td>SICHEL()</td>
<td>log</td>
<td>log</td>
<td>identity</td>
</tr>
<tr>
<td>Yule (the mean)</td>
<td>YULE()</td>
<td>log</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>zero altered beta binomial</td>
<td>ZABB()</td>
<td>logit</td>
<td>log</td>
<td>logit</td>
</tr>
<tr>
<td>zero altered binomial</td>
<td>ZABI()</td>
<td>logit</td>
<td>logit</td>
<td>-</td>
</tr>
<tr>
<td>zero altered logarithmic</td>
<td>ZALG()</td>
<td>logit</td>
<td>logit</td>
<td>-</td>
</tr>
<tr>
<td>zero altered neg. binomial</td>
<td>ZANBI()</td>
<td>log</td>
<td>log</td>
<td>logit</td>
</tr>
<tr>
<td>zero altered poisson</td>
<td>ZAP()</td>
<td>log</td>
<td>logit</td>
<td>-</td>
</tr>
<tr>
<td>zero inflated beta binomial</td>
<td>ZIBB()</td>
<td>logit</td>
<td>log</td>
<td>logit</td>
</tr>
<tr>
<td>zero inflated binomial</td>
<td>ZIBI()</td>
<td>logit</td>
<td>logit</td>
<td>-</td>
</tr>
<tr>
<td>zero inflated neg. binomial</td>
<td>ZINBI()</td>
<td>log</td>
<td>log</td>
<td>logit</td>
</tr>
<tr>
<td>zero inflated poisson</td>
<td>ZIP()</td>
<td>log</td>
<td>logit</td>
<td>-</td>
</tr>
<tr>
<td>zero inflated poisson (the mean)</td>
<td>ZIP2()</td>
<td>log</td>
<td>log</td>
<td>logit</td>
</tr>
<tr>
<td>zero inflated poisson inv. Gaussian</td>
<td>ZIPIG()</td>
<td>log</td>
<td>log</td>
<td>logit</td>
</tr>
</tbody>
</table>

Table 1.2: Discrete distributions implemented within the `gamlss` packages (with default link functions)

| beta inflated (at 0)          | BEIO() | logit   | log | logit | -     |
| beta inflated (at 0)          | BEINFO() | logit | logit | log | -     |
| beta inflated (at 1)          | BEZO() | logit   | log | logit | -     |
| beta inflated (at 1)          | BEFIN() | logit   | log | logit | -     |
| beta inflated (at 0 and 1)    | BEINF() | logit   | log | log | logit |
| zero adjusted GA              | ZAGA() | log     | log | logit | -     |
| zero adjusted IG              | ZAGI() | log     | log | logit | -     |

Table 1.3: Mixed distributions implemented within the `gamlss` packages (with default link functions)
gamlss.demo()

will bring the package *gamlss.demo* in R and display a menu where by choosing the option gamlss family distributions you can proceed to display the different distributions. Alternatively you can just type `demo.NAME()` where NAME is a `gamlss.family` name e.g. `demo.NO()` for normal distribution. This allows any distribution in GAMLSS to be displayed graphically and its parameters adjusted interactively.

An alternative method of graphically displaying the probability (density) functions is to use the pdf.plot function.

```r
pdf.plot(family=PO(),mu=c(1,2,3,4), min=0, max=10,step=1)
```

or if you wish to display plots in one figure:

```r
y<-seq(0,10,by=0.01)
plot(dNO(y, mu=5, sigma=0.5)~y, type="l",col="black",ylab="f(y")
lines(dNO(y, mu=5, sigma=1)~y, col="red")
lines(dNO(y, mu=5, sigma=1.5)~y, col="green")
lines(dNO(y, mu=5, sigma=2)~y, col="blue")
```

**Exercises**

**Practical 1: Introduction to gamlss packages**

- Q1 The *gamlss.dist* package (which is downloaded automatically when *gamlss* is downloaded) contains many distributions. Typing

  ```r
  ?gamlss.family
  ```

  in R will show all the available distributions in the *gamlss* packages.

  You can explore the shape and other properties of the distributions. For example the following R script will produce the probability density function (pdf), cumulative distribution function (cdf), inverse c.d.f., and a histogram of a random sample obtained from a Gamma distribution:

  ```r
  PPP <- par(mfrow=c(2,2))
  plot(function(y) dGA(y, mu=10 ,sigma=0.3),0.1, 25) # pdf
  plot(function(y) pGA(y, mu=10 ,sigma=0.3), 0.1, 25) #cdf
  plot(function(y) qGA(y, mu=10 ,sigma=0.3), 0, 1) # inverse cdf
  hist(rGA(100,mu=10,sigma=.3)) # randomly generated values
  par(PPP)
  ```

  Note that the same type of plots produced say by

  ```r
  plot(function(y) dGA(y, mu=10 ,sigma=0.3), 0, 25) # pdf
  ```

  can also be produced by using the function `curve()` as in

  ```r
  curve(dGA(x=x, mu=10, sigma=.3),0, 25)
  ```
1.5. DISPLAYING GAMLSS FAMILY DISTRIBUTIONS

To explore discrete distributions use:

```r
PPP <- par(mfrow=c(2,2))
plot(function(y) dNBI(y, mu = 10, sigma =0.5 ), from=0, to=40, n=40+1, type="h",
     main="pdf", ylab="pdf(x)")
cdf <- stepfun(0:39, c(0, pNBI(0:39, mu=10, sigma=0.5 ))), f = 0)
plot(cdf,main="cdf", ylab="cdf(x)", do.points=FALSE )
invcdf <-stepfun(seq(0.01,.99,length=39), qNBI(seq(0.01,.99,length=40),
        mu=10, sigma=0.5 ), f = 0)
plot(invcdf, main="inverse cdf",ylab="inv-cdf(x)", do.points=FALSE )
tN <- table(Ni <- rNBI(1000,mu=5, sigma=0.5))
r <- barplot(tN, col='lightblue')
par(PPP)
```

Note that to find moments or to check if a distribution integrates or sums to one, the functions integrate() or sum can be used. For example

```r
integrate(function(y) dGA(y, mu=10, sigma=.1),0, Inf)
```

will check that the distribution integrates to one, and

```r
integrate(function(y) y*dGA(y, mu=10, sigma=.1),0, Inf)
```

will give the mean of the specific gamma distribution.

The density function of a GAMLSS family distribution can be plotted also using the pdf.plot() of the GAMLSS package. Use for example

```r
dpdf.plot(family=GA, mu=10, sigma=c(.1,.5,1,2),
         min=0.01,max=20, step=.5)
```

Try plotting other continuous distributions, e.g. IG (inverse Gaussian), PE (power exponential) and BCT (Box-Cox t) and discrete distributions, e.g. NBI (negative binomial type I), and PIG (Poisson inverse Gaussian).
Chapter 2

Continuous distributions

This chapter provides explanation for:

1. different types of continuous distributions within the GAMLSS family
2. how those distributions model skewness and kurtosis

This chapter is essential for understanding the different types of continuous GAMLSS family distributions.

2.1 Introduction

As we learnt in section 1.3 of the previous Chapter, continuous distributions can be symmetric, negatively or positively skewed, and also mesokurtic, leptokurtic or platykurtic, see Figure 1.7. Note that leptokurtic and platykurtic distributions are judged by comparison to the normal distribution which is mesokurtic. A leptokurtic distribution has thicker (fatter) tails than the normal distribution, while a platykurtic distribution has thinner (or shorter) tails than the normal.

In this chapter we study the GAMLSS family of continuous distributions in more detail. In particular we check the shapes of the distributions and therefore how flexible they are to model skewness and kurtosis in the data.

GAMLSS family distributions have up to four parameters and for consistency they are called \( \mu, \sigma, \nu \) and \( \tau \). Again for consistency we use in general \( \mu \) as a location parameter, \( \sigma \) as a scale parameters, while \( \nu \) and \( \tau \) deal with skewness and/or kurtosis. It is important to emphasise that the distribution parameters, for example, \( \mu \) and \( \sigma \) do not necessarily represent the mean and standard deviation of \( Y \). The relationship between the distribution parameters and moment-based measures (e.g., mean and variance) depends on the properties of the distribution. For example, for the gamma distribution, as parametrised in GAMLSS, the mean is given by the distribution parameter \( \mu \) while the variance is given by \( \mu^2 \sigma^2 \).

Table 2.1 provides a list of the continuous \texttt{gamlss.family} distributions defined on support range \( R_Y = (-\infty, +\infty) \) available in the current version of GAMLSS software, while Tables 2.2 and 2.3 provide a list of distributions defined on \( (0, +\infty) \) and \( (0, 1) \) respectively. Note that the word both in the skewness column of Tables 2.1, 2.2 and 2.3 indicates that the distribution can be negatively or positively skewed, while 'both' in the kurtosis column indicates that the distribution can be platykurtic or leptokurtic. A brackets indicates that the skewness or kurtosis
Table 2.1: Continuous GAMLSS family distributions defined on \((-\infty, +\infty)\)

<table>
<thead>
<tr>
<th>Distributions</th>
<th>family</th>
<th>no parameters</th>
<th>skewness</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential Gaussian</td>
<td>exGAUS</td>
<td>3</td>
<td>positive</td>
<td>-</td>
</tr>
<tr>
<td>Exponential Generalised beta type 2</td>
<td>EGB2</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>Generalised t</td>
<td>GT</td>
<td>4</td>
<td>(symmetric)</td>
<td>lepto</td>
</tr>
<tr>
<td>Gumbel</td>
<td>GU</td>
<td>2</td>
<td>(negative)</td>
<td>-</td>
</tr>
<tr>
<td>Johnson’s SU</td>
<td>JSU</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>Johnson’s original SU</td>
<td>JSUo</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>Logistic</td>
<td>LO</td>
<td>2</td>
<td>(symmetric)</td>
<td>(lepto)</td>
</tr>
<tr>
<td>Normal-Exponential-t</td>
<td>NET</td>
<td>2 + (2 fixed)</td>
<td>(symmetric)</td>
<td>lepto</td>
</tr>
<tr>
<td>Normal</td>
<td>NO-NO2</td>
<td>2</td>
<td>(symmetric)</td>
<td></td>
</tr>
<tr>
<td>Normal Family</td>
<td>NOF</td>
<td>3</td>
<td>(symmetric)</td>
<td>(meso)</td>
</tr>
<tr>
<td>Power Exponential</td>
<td>PE-PE2</td>
<td>3</td>
<td>(symmetric)</td>
<td>both</td>
</tr>
<tr>
<td>Reverse Gumbel</td>
<td>RG</td>
<td>2</td>
<td>positive</td>
<td>-</td>
</tr>
<tr>
<td>Sinh Arcsineh</td>
<td>SHASH-SHASHo,</td>
<td>4</td>
<td>both</td>
<td>both</td>
</tr>
<tr>
<td>Skew Exponential Power (type 1 to 4)</td>
<td>SEP1-SEP4</td>
<td>4</td>
<td>both</td>
<td>both</td>
</tr>
<tr>
<td>Skew normal (type 1 to 2)</td>
<td>SN1-SN2</td>
<td>3</td>
<td>both</td>
<td>-</td>
</tr>
<tr>
<td>Skew t (type 1 to 5)</td>
<td>ST1-ST5</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>t Family</td>
<td>TF</td>
<td>3</td>
<td>(symmetric)</td>
<td>lepto</td>
</tr>
</tbody>
</table>

cannot be modelled independently of the location and scale parameters.

Many of the distributions of Tables 2.1, 2.2 and 2.3 can be generated by one (or more) of the methods described in Chapter ???. The probability density function and history of the distributions can be found in Part Two of this book.

All the distributions presented here can be plotted and investigated interactively using the R package `gamlss.demo`.

### 2.2 Continuous distributions on support \((-\infty, \infty)\)

Table 2.1 gives the continuous distributions on support \((-\infty, \infty)\) in GAMLSS. All the distributions in Table 2.1 are location-scale distributions (with the exception of the exGAUS distribution). For example, for a four parameter distribution \(D\), and random variable \(Y \sim D(\mu, \sigma, \nu, \tau)\) to be location-scale we need \(\varepsilon = (Y - \mu)/\sigma \sim D(0, 1, \nu, \tau)\), i.e. \(Y = \mu + \sigma \varepsilon\), so the random variable \(Y\) is a scaled and shifted function of the random variable \(\varepsilon\). why is the importance of location-scale distributions?

#### 2.2.1 Two parameter distributions on \((-\infty, \infty)\).

The two parameter continuous distributions with range \((-\infty, \infty)\) are:

**GU**: the Gumbel (or extreme value or Gompertz) distribution, see Crowder et al. (1991) p 17 and section 14.1.3.

**LO**: the logistic distribution, see Johnson et al. (1995) p 116, and section 14.1.2.

**NO**: the normal distribution where \(\sigma\) is the standard deviation, see section 14.1.1.
2.2. CONTINUOUS DISTRIBUTIONS ON SUPPORT \((-\infty, \infty)\)

\textbf{NO}: the normal or Gaussian distribution where \(\sigma\) is the variance, see section 14.1.1. This distribution is included in the list for completeness.

\textbf{RG}: the reverse Gumbel distribution or type I extreme value distribution, see Johnson \textit{et al.} (1995) p 2 and p 75 and section 14.1.4

Two parameter distributions are only able to model independently the location \(\mu\) and scale \(\sigma\) of the distribution while skewness and/or kurtosis are defined implicitly from those two parameters. For example, the normal distribution \(NO(\mu, \sigma)\) has location parameter \(\mu\) which is the mean of this distribution and scale parameter \(\sigma\) which is the standard deviation of this distribution. The (moment based) skewness and excess kurtosis of the normal distribution are fixed at zero (indicating a symmetric mesokurtic distribution) and therefore cannot be modelled using this distribution. The logistic distribution \((L0)\) distribution is symmetric but has a higher kurtosis than the normal. For the \(GU\) and \(RG\) distributions the skewness changes with the location and/or scale parameters but cannot be modelled independently of them. The \(GU\) distribution is negatively skew, while \(RG\) is positively skew. Figure 2.1 compares the \(NO(0,1)\), \(GU(0,1)\) and \(RG(0,1)\) distributions.

It is very important to reiterate that the relationship between the distribution parameters \((\mu, \sigma, \nu \text{ and } \tau)\) and moment-based measurements (e.g., mean, variance, skewness and kurtosis) is a function of the properties of the specific distribution. As discussed above, for the \(NO\) distribution in GAMLSS, the mean is defined by the distribution parameter \(\mu\) while the standard deviation is defined by the distribution parameter \(\sigma\). By changing the distribution, for example, the \(GA\) distribution discussed in section 2.3.2, the standard deviation is defined by \(\sqrt{\mu^2 + \sigma^2}\)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.1.png}
\caption{Two parameter distributions, NO(0,1), GU(0,1) and RG(0,1)}
\end{figure}

2.2.2 Three parameter distributions on \((-\infty, \infty)\).

The three parameter distributions with range \((-\infty, \infty)\) are:
exGAUS: the exponential Gaussian distribution or lagged normal distribution, see Johnson et al., (1994), p172 and section 14.2.1.

NOF: the normal family which is a normal distribution with an extra parameter to account for a mean variance relationship, appropriate only for regression models, see 14.1.1.

PE: the power exponential distribution, see Nelson (1991), and section 14.2.2.

PE2: the power exponential distribution (different parametrisation to above), see section 14.2.2.

TF: the t family, see section 14.2.3.

SN1: the skew normal distribution type 1, Azzalini (1985) see section 11.2.

SN2: the skew normal distribution type 2 (different parametrisation to above), see Johnson et al. (1994) p173 and section 11.3.

Three parameter distributions are able to model either skewness or kurtosis in addition to the location and scale parameters. The skew normal type 1 and type 2, SN1 and SN2, distributions and the exponential-Gaussian distribution, exGAUS are able to model skewness.

Figure 2.2 plots the $SN_1(\mu, \sigma, \nu)$ distribution for $\mu = 0$, $\sigma = 1$ and $\nu = 0, 1, 3, 100$. Corresponding negatively skewed $SN_1$ distributions are given by $\nu = -1, -3, -100$ (which reflect the distributions for $\nu = 1, 3, 100$ about the vertical origin $y = 0$). The R code to create the reflected image of the distribution is given below:

```R
y<-seq(-4,4,by=0.01)
plot(dSN1(y, mu=0, sigma=1, nu=-100)~y, type="l", col="green", ylab="f(y)", lt=1)
lines(dSN1(y, mu=0, sigma=1, nu=-3)~y, col="red",lt=2,lw=2)
lines(dSN1(y, mu=0, sigma=1, nu=-1)~y, col="blue",lt=3,lw=2)
lines(dSN1(y, mu=0, sigma=1, nu=0)~y, col="black",lt=4,lw=2)
ex.leg <- expression(paste( \nu, " = -100"), paste( \nu," = -3"), paste( \nu, " = -1"), paste( \nu, " = 0"))
legend("topright", ex.leg, lty=c(1,2,3,4), col=c("green", "red", "blue", "black"), adj = c(0, .6))
```

The t family distribution, TF, is symmetric but able to model leptokurtosis, while the power exponential distribution PE is symmetric but able to model both leptokurtosis and platykurtosis.

Figure 2.3 plots the power exponential distribution $PE(\mu, \sigma, \nu)$ for $\mu = 0$, $\sigma = 1$ and $\nu = 1, 2, 10, 1000$. The power exponential includes the Laplace (i.e. a two-sided exponential) and the normal as special cases $\nu = 1$ and $\nu = 2$ respectively, and the uniform distribution as a limiting case as $\nu \to \infty$. Note that $\nu > 0$ so PE can be more extremely kurtotic than the Laplace when $0 < \nu < 1$.

### 2.2.3 Four parameter distributions on $(-\infty, \infty)$.

The four parameter distributions with range $(-\infty, \infty)$ are:

EGB2: the exponential generalised beta type 2 distribution, see McDonald and Xu [1995] and 14.3.1.
2.2. CONTINUOUS DISTRIBUTIONS ON SUPPORT $(-\infty, \infty)$

Figure 2.2: The skew normal type 1 distribution, SN1(0, 1, $\nu$) for $\nu=0, 1, 3$ and 100. Changing the sign of $\nu$ reflects the distribution about $y = 0$.

Figure 2.3: The power exponential distribution, PE(0, 1, $\nu$), for $\nu=1, 2, 10$ and 1000.
GT: the generalised $t$ distribution, see McDonald [1991], McDonald and Newey [1988] and section 14.3.2.

JSU: the Johnson’s SU distribution, see Johnson [1949] and 14.3.3.

JSU0: the Johnson’s original (parametrisation) SU, see Johnson [1949] and 14.3.3.

NET: the Normal-Exponential-$t$ distribution, introduced by Rigby and Stasinopoulos [1994] as a robust method of fitting the mean and the scale parameters of a symmetric distribution. This is a piecewise distribution which is normal in the middle, continues as a double exponential (Laplace) and finally ends up with a $t$ distribution as a tail, see for example Figure 2.4 and also section 14.3.4.

SEP1: the skew exponential power distribution type 1, see Azzalini (1986) and section 14.3.8.

SEP2: the skew exponential power distribution type 2 see Azzalini (1986) and Diciccio and Monti (2004) and section 14.3.9

SEP3: the skew exponential power distribution type 3 see Fernandez et. al. (1995) and section 14.3.10

SEP4: the skew exponential power distribution type 4, see Jones (2005) and section 14.3.11.

SHASH: the sinh-arcsinh distribution (the first parametrisation), see Jones(2005) and section 14.3.5.

SHASHo: the sinh-arcsinh distribution (original parametrisation) see Jones and Pewsey [2009] and section 14.3.6.

SHASHo2: the sinh-arcsinh distribution (amended original parametrisation) type 2, see Jones and Pewsey [2009] and section 14.3.7

ST1: the skew $t$ distribution type 1, see Azzalini (1986) and section 14.3.8

ST2: the skew $t$ distribution type 2, see Azzalini and Capitanio (2003), p382. and section 14.3.13

ST3: the skew $t$ distribution type 3, see Fernandez and Steel (1998), p360, eqn. (5) and section 14.3.14

ST4: the skew $t$ distribution type 4, see section 14.3.15

ST5: the skew $t$ distribution type 5, see Jones and Faddy (2003), p162 and section 14.3.16

Four parameter distributions are able to model both skewness and kurtosis in addition to the location and scale parameters. The only exceptions are the NET and the GT distributions both symmetric distributions where both the two extra parameters model kurtosis.

The EGB2, JSU, JSU0 and ST1–ST5 four parameter distributions are able to model skewness and leptokurtosis. The SHASH and SEP1–SEP4 distributions are able to model skewness and both leptokurtosis and platykurtosis.

Figure 2.5 shows skew SEP1($\mu, \sigma, \nu, \tau$) distributions where $\mu = 0$ and $\sigma = 1$. Figure 2.5(a) and (b) plot the SEP1 distribution for $\nu = 0, 1, 3, 100$ and $\tau = .5$ and $\tau = 2$ respectively. Note the when $\tau = 2$ the SEP1 becomes the skew normal type 1 distribution. Figure 2.5(c)
2.2. CONTINUOUS DISTRIBUTIONS ON SUPPORT $(-\infty, \infty)$

The NET distribution

Figure 2.4: The NET distribution $NET(\mu = 0, \sigma = 1.5, \nu = 2, \tau = 3)$

and (d) plot the SEP1 distribution for $\nu = 0, -2, -4, -6, -8, -1$ and $\tau = 10$ and $\tau = 1000$ respectively.

Changing the sign of $\nu$ in $SEP1(\mu, \sigma, \nu, \tau)$ in Figure 2.5 reflects the distributions about the origin, changing the skewness from positive to negative. Setting $\nu = 0$ in $SEP1(\mu, \sigma, \nu, \tau)$ gives a symmetric distribution which is a reparametrisated power exponential distribution.

The R code to reproduce the NET distribution in Figure 2.4 is given below:

```R
x=seq(-5, 5, length=201)
zero<-rep(0,201)
plot(dNET(x, mu=0, sigma=1.5, nu=2, tau=3)~x, xlab="Y", ylab="f[Y](y)", type="l")
title("The NET distribution")
polygon(c(-2,seq(-2,2, length=100),2), c(0, dNET(seq(-2,2, length=100), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey90", lty=0)
polygon(c(-3,seq(-3,-2, length=50),-2), c(0, dNET(seq(-3,-2, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey60", lty=0)
polygon(c(2,seq(2,3, length=50),3), c(0, dNET(seq(2,3, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey60", lty=0)
polygon(c(-5,seq(-5,-3, length=50),-3), c(0, dNET(seq(-5,-3, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey30", lty=0)
polygon(c(3,seq(3,5, length=50),5), c(0, dNET(seq(3,5, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey30", lty=0)
lines(zero~x)
text(1.3,0.24, "normal")
text(3.5, 0.06, "exponential")
text(4.4, 0.026, "t-distribution")
```

The R code to reproduce the upper left plot in Figure 2.5 is given below:

```R
y<-seq(-4,4,by=0.01)
x=seq(-5, 5, length=201)
zero<-rep(0,201)
plot(dNET(x, mu=0, sigma=1.5, nu=2, tau=3)~x, xlab="Y", ylab="f[Y](y)", type="l")
title("The NET distribution")
polygon(c(-2,seq(-2,2, length=100),2), c(0, dNET(seq(-2,2, length=100), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey90", lty=0)
polygon(c(-3,seq(-3,-2, length=50),-2), c(0, dNET(seq(-3,-2, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey60", lty=0)
polygon(c(2,seq(2,3, length=50),3), c(0, dNET(seq(2,3, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey60", lty=0)
polygon(c(-5,seq(-5,-3, length=50),-3), c(0, dNET(seq(-5,-3, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey30", lty=0)
polygon(c(3,seq(3,5, length=50),5), c(0, dNET(seq(3,5, length=50), mu=0, sigma=1.5, nu=2, tau=3),0), col="grey30", lty=0)
lines(zero~x)
text(1.3,0.24, "normal")
text(3.5, 0.06, "exponential")
text(4.4, 0.026, "t-distribution")
```
Figure 2.5: The skew exponential power type 1 distribution, SEP1(0,1,\nu,\tau) for \tau (a) .5, (b) 2, (c) 10, (d) 1000.
2.3 Continuous distributions on support \((0, \infty)\)

Table 2.2: Continuous GAMLSS family distributions defined on \((0, +\infty)\)

<table>
<thead>
<tr>
<th>Distributions</th>
<th>family</th>
<th>no parameters</th>
<th>skewness</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box-Cox Cole and Green</td>
<td>BCCG</td>
<td>3</td>
<td>both</td>
<td>-</td>
</tr>
<tr>
<td>Box-Cox Power Exponential</td>
<td>BCPE</td>
<td>4</td>
<td>both</td>
<td>both</td>
</tr>
<tr>
<td>Box-Cox-t</td>
<td>BCT</td>
<td>4</td>
<td>both</td>
<td>lepto</td>
</tr>
<tr>
<td>Exponential</td>
<td>EXP</td>
<td>1</td>
<td>(positive)</td>
<td>-</td>
</tr>
<tr>
<td>Gamma</td>
<td>GA</td>
<td>2</td>
<td>(positive)</td>
<td>-</td>
</tr>
<tr>
<td>Generalized Beta type 2</td>
<td>GB2</td>
<td>4</td>
<td>both</td>
<td>both</td>
</tr>
<tr>
<td>Generalized Gamma</td>
<td>GG</td>
<td>3</td>
<td>positive</td>
<td>-</td>
</tr>
<tr>
<td>Generalized Inverse Gaussian</td>
<td>GIG</td>
<td>3</td>
<td>positive</td>
<td>-</td>
</tr>
<tr>
<td>Inverse gamma</td>
<td>IGAMMA</td>
<td>2</td>
<td>(positive)</td>
<td>-</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>IG</td>
<td>2</td>
<td>(positive)</td>
<td>-</td>
</tr>
<tr>
<td>Log Normal</td>
<td>LOGNO</td>
<td>2</td>
<td>(positive)</td>
<td>-</td>
</tr>
<tr>
<td>Log Normal family</td>
<td>LNO</td>
<td>2 + (1 fixed)</td>
<td>positive</td>
<td></td>
</tr>
<tr>
<td>Pareto type 2</td>
<td>PARETO2</td>
<td>2</td>
<td>positive</td>
<td>??</td>
</tr>
<tr>
<td>Reverse Generalized Extreme</td>
<td>RGE</td>
<td>3</td>
<td>positive</td>
<td>-</td>
</tr>
<tr>
<td>Weibull</td>
<td>WEI-WEI3</td>
<td>2</td>
<td>(positive)</td>
<td>-</td>
</tr>
</tbody>
</table>

plot(dSEP1(y, mu=0, sigma=1, nu=100, tau=.5)^y, type="l",col="yellow",ylab="f(y)")
lines(dSEP1(y, mu=0, sigma=1, nu=3, tau=.5)^y, col="red")
lines(dSEP1(y, mu=0, sigma=1, nu=1, tau=.5)^y, col="blue")
lines(dSEP1(y, mu=0, sigma=1, nu=0, tau=.5)^y, col="black")

2.3 Continuous distributions on support \((0, \infty)\)

Table 2.2 gives the continuous distributions on support \((0, \infty)\) in GAMLSS. Many of the distributions in Table 2.2 are scale distributions (BCCG, BCPE, BCT, EXP, GA, GB2, GG, GIG, WEI and WEI3). To explain the concept of a scale distribution consider an example, of a four parameter distribution \(D\). If \(Y \sim D(\mu, \sigma, \nu, \tau)\) then \(\varepsilon = (Y/\mu) \sim D(1, \sigma, \nu, \tau)\), i.e. \(Y = \mu \varepsilon\), so the random variable \(Y\) is a scaled version of the random variable \(\varepsilon\), that is, proportional to \(\varepsilon\).

2.3.1 One parameter distributions on \((0, \infty)\).

There is only one single parameter continuous distribution in the gamlss.family, the exponential distribution, EXP. The parametrisation that GAMLSS uses for the exponential distribution is the one in equation (1.4) where \(\mu\) is the mean, see also section 15.1.1.

2.3.2 Two parameter distributions on \((0, \infty)\).

The two parameter distributions on the range \((0, \infty)\) are:

GA: the gamma distribution see section 15.2.1

IGAMMA: the inverse gamma distribution see section ??.

IG: the inverse Gaussian distribution see section 15.2.3.

LOGNO: the log-normal distribution see section 15.2.2
Figure 2.6: The pdf and hazard functions of the Weibull type 2 distribution, WEI2, for $\mu = 1$ and $\sigma = 0.5, 1, 1.2$ and the gamma, GA, for $\mu = 1$ and $\sigma = 0.5, 1, 2$. 
WEI: the Weibull distribution, see Johnson et al. (1994) p632 and section 15.2.4.

WEI2: the Weibull distribution parametrisation 2 see section 15.2.4

WEI3: the Weibull distribution parametrisation 3 with $\mu$ as the mean of the distribution, see section 15.2.4

Two parameter distributions are only able to model the location and scale of the distribution, (since the skewness and kurtosis are determined by the values of location and scale).

The log normal distribution is associated with geometric Brownian motion in finance. The exponential, EXP, Weibull, WEI, WEI2, and WEI3, and gamma GA distributions are widely used in survival and reliability analysis.

By way of an example, Figure 2.6 shows the pdf and hazard for the Weibull type 2, WEI2 and for the gamma distribution. The parameter values are set to $\mu = 1$ and $\sigma = (0.5, 1, 1.2)$ and $\sigma = (0.5, 1, 2)$ respectively. The R code to produce Figure 2.6 is given below:

```r
op<-par(mfrow=c(2,2))
y<-seq(0,10,by=0.01)
plot(dWEI2(y, mu=1, sigma=1)~y, type="l", col="black", ylab="f(y)", lty=1,
     ylim=c(0,1.5))
lines(dWEI2(y, mu=1, sigma=1.2)~y, col="red",lty=2,lw=2)
lines(dWEI2(y, mu=1, sigma=.5)~y, col="blue",lty=3,lw=2)
ex.leg <- expression(paste( sigma, " = 1"), paste( sigma," = 1.2"), 
paste( sigma, " = 0.5"))# 2 ways
legend("topright", ex.leg, lty=c(1,2,3), col=c("black", "red", "blue"),
       adj = c(0, .6))
title("WEI2 pdf")
# figure 12
plot(dGA(y, mu=1, sigma=1)~y, type="l", col="black", ylab="f(y)", lty=1,
     ylim=c(0,1.5))
lines(dGA(y, mu=1, sigma=2)~y, col="red",lty=2,lw=2)
lines(dGA(y, mu=1, sigma=.5)~y, col="blue",lty=3,lw=2)
ex.leg <- expression(paste( sigma, " = 1"), paste( sigma," = 2"), 
paste( sigma, " = 0.5"))# 2 ways
legend("topright", ex.leg, lty=c(1,2,3), col=c("black", "red", "blue"),
       adj = c(0, .6))
title("GA pdf")
# figure 21
gen.hazard("WEI2")
plot(hWEI2(y, mu=1, sigma=1)~y, type="l", col="black", ylab="h(y)",
     ylim=c(0,2.5))
lines(hWEI2(y, mu=1, sigma=1.2)~y, col="red",lty=2,lw=2)
lines(hWEI2(y, mu=1, sigma=.5)~y, col="blue",lty=3,lw=2)
title("WEI2 hazard")
#-------------------------------------------------------------
gen.hazard("GA")
plot(hGA(y, mu=1, sigma=1)~y, type="l", col="black", ylab="h(y)", ylim=c(0,4))
lines(hGA(y, mu=1, sigma=2)~y, col="red",lty=2,lw=2)
lines(hGA(y, mu=1, sigma=.5)~y, col="blue",lty=3,lw=2)
```
Note that hazard functions are not automatically given in the GAMLSS packages but have to be generated using the function `gen.hazard()` which takes as an argument the `gamlss.family` name. Figure 2.7 shows the pdf and hazard for the inverse gamma, IGAMMA, and inverse Gaussian, IG, distributions. The parameters values are set $\mu = 1$ and $\sigma = (0.5, 1, 2)$.

Figure 2.7: The pdf and hazard functions of the inverse gamma distribution, IGAMMA, and the inverse Gaussian, IG, for different values of $\sigma = 0.5, 1, 2$

### 2.3.3 Three parameter distributions on $(0, \infty)$.

The three parameter distributions on $(0, \infty)$ are:

**BCCG:** the Box-Cox Cole and Green distribution introduced by Cole and Green (1992) see section 15.3.1. This is a re-parametrisation of the original Box Cox family (see equation, 2.1 below) in which parameters $\mu$, $\sigma$ and $\nu$ are all fitted.
2.3. CONTINUOUS DISTRIBUTIONS ON SUPPORT $(0, \infty)$

**GG:** the generalised gamma distribution see Lopatatzidis and Green (2000) and section 15.3.2

**GG2:** the generalised gamma distribution parametrisation type 2 see Johnson et. al. (1995) p401 and section 15.3.2

**GIG:** the generalised inverse Gaussian distribution see section 15.3.3

**LNO:** the log normal family (i.e. original Box-Cox family). The third parameter of the family, $\nu$, is fixed in the fitting (not estimated).

Three parameter distributions are able to model skewness in addition to location and scale. The generalised gamma, GG, and the generalised inverse Gaussian, GIG, have been used in survival and reliability analysis.

The LNO distribution is based on the standard Box Cox transformation, Box and Cox (1964), which is defined as:

$$Z = \begin{cases} 
\frac{(Y^\nu - 1)}{\nu}, & \text{if } \nu \neq 0 \\
\log(Y), & \text{if } \nu = 0.
\end{cases}$$

(2.1)

and where the transformed variable $Z$ is then assumed to have a normal NO($\mu, \sigma$) distribution. The parameter $\nu$ (or sometimes known as $\lambda$ in the literature) is the skewness parameter. Note that $\nu = 0$ corresponds to the log-normal distribution while $\nu = 1$ to the normal distribution. Figure 2.8 shows the LNO$(0, 1, \nu)$ distribution for different values of $\nu = -1, 0, 1, 2$. In the fitting GAMLSS procedure the parameter $\nu$ is treated as fixed because of the difficulty in estimating it. The reason is because the estimate of $\nu$ is highly correlated with that of both $\mu$ and $\sigma$. A similar distribution is the BCCG distribution which uses the Cole and Green (1992) transformation to normality see section 15.3.1. The $\nu$ parameter is much easier to estimate in this case while $\mu$ and $\sigma$ are interpreted as the median and (approximate) coefficient of variation. The distribution BCCG is used in the LMS method, Cole and Green (1992), widely used in centile estimation.

![pdf of LNO](image)

Figure 2.8: The log normal family LNO$(0, 1, \nu)$ for $\nu = -1, 1, 1, 2$. 
Figure 2.9: The Probability Density Function $f(y)$ of the Box-Cox t Distribution $BCT(\mu, \sigma, \nu, \tau)$ Against $y$. Parameter values (a) $\mu = (8, 10, 12)$, $\sigma = 0.15$, $\nu = 1$, $\tau = 10$ (b) $\mu = 10$, $\sigma = (0.10, 0.15, 0.20)$, $\nu = 1$, $\tau = 10$ (c) $\mu = 10$, $\sigma = 0.15$, $\nu = (-5, 1, 5)$, $\tau = 10$ (d) $\mu = 10$, $\sigma = 0.15$, $\nu = 1$, $\tau = (1, 10, 40)$. The three curves for the parameter varying are ( ), ( ) and ( ) respectively.

2.3.4 Four parameter distributions on $(0, \infty)$.

The four parameter distributions on $(0, \infty)$ are:

**BCPE**: the Box-Cox power exponential distribution 15.4.2

**BCT**: the Box-Cox power exponential distribution 15.4.1

**GB2**: the generalised beta distribution 15.4.3

Four parameter distributions can model both skewness and kurtosis. The BCT distribution can model skewness and leptokurtosis, while the BCPE and GB2 can model skewness and both leptokurtosis and platykurtosis. Figure 2.9 shows the BCT distribution for different values of each of the four parameters.
2.4 Continuous distributions on support \((0, 1)\)

Table 2.3 gives the Continuous distributions on support \((0, 1)\) in GAMLSS. There are two parameterizations of the beta distribution, the original parameterization \(BE_0\) and a reparameterization \(BE\) which has mean \(\mu\) and variance \(Var(Y) = \sigma^2 \mu (1 - \mu)\). The beta distribution has two parameters and so is only able to model location and scale. The four parameter generalized beta type 1, \(GB1\), is able to model skewness and kurtosis in addition to location and scale. Note also that any GAMLSS distribution on \(-\infty, \infty\) can be transformed to a distribution on \((0, 1)\) see section 11.5 and the example in section 8.3.3.

![Beta Distribution Family](image)

2.5 Comparison of properties of continuous distributions

The choice of model distribution for a particular response variable \(Y\) is usually based on how well it fits the data as judged by the fitted global deviance \(GDEV = -2 \log \hat{l}\), i.e. minus twice the fitted log likelihood function, and tests and information criteria (e.g. AIC or SBC) based on \(GDEV\). Residual plots also help to assess the adequacy of a chosen distribution.
CHAPTER 2. CONTINUOUS DISTRIBUTIONS

Where more than one distribution fits the data adequately the choice of distribution may be made on other criteria, e.g. properties of the particular distribution. For example an explicit formula for the mean, median or mode of $Y$ may be desirable in a particular application.

The term explicit indicates that the particular function or measure can be obtained using standard functions (available in R), i.e. not requiring numerical integration or numerical solution.

In addition to the range of the distribution, the following are properties of the distribution that may be relevant in choosing the model distribution:

1. Explicit probability density function, cumulative distribution function and inverse cumulative distribution function

2. Explicit moment based measures of location, scale, skewness and kurtosis (e.g. mean, standard deviation, skewness $\sqrt{\beta_1}$ and kurtosis $\beta_2$), see section 1.3.1

3. Explicit centiles and centile based measures of location, scale, skewness and kurtosis (e.g. median, semi-interquartile range, skewness $\gamma$ and kurtosis $st_{0.49}$), see section 1.3.2

4. Explicit mode(s)

5. Continuity of the probability density function $f_Y(y|\mu, \sigma, \nu, \tau)$ and its derivatives with respect to $y$

6. Continuity of the derivatives of the probability density function $f_Y(y|\mu, \sigma, \nu, \tau)$ with respect to $\mu$, $\sigma$, $\nu$ and $\tau$

7. Flexibility in modelling skewness and kurtosis

Many of the distributions of Tables 2.1, 2.2 and 2.3 can be generated by one (or more) of the methods described in Chapter 11, including distributions generated by univariate transformation, Azzalini type methods and splicing distributions.

Distributions generated by univariate transformation often satisfy all the desirable properties above, except perhaps the flexibility in modelling skewness and kurtosis.

An important disadvantage of distributions generated by Azzalini type methods are that their cumulative distribution function (cdf) is not explicitly available, but requires numerical integration. Their inverse cdf requires a numerical search and many integrations. Consequently both functions can be slow, particularly for large data sets. Centiles and centile based measures (e.g. the median) are not explicitly available. Moment based measures are usually complicated, if available. However they can be flexible in modelling skewness and kurtosis.

An important disadvantage of distributions generated by splicing is often a lack of continuity of the second derivatives of the probability density function with respect to $y$ and $\mu$ at the splicing point. However their mode and moment based measures are explicit and they can be flexible in modelling skewness and kurtosis.

Exercises

Q1: Reproduce Figure 2.5.
Chapter 3

Discrete distributions for count data

3.1 Introduction

GAMLSS discrete distributions for unlimited count variables have range $R_Y = (0, 1, 2, 3, \ldots, \infty)$, [except for the logarithmic distribution (LG) which has range $R_Y = (1, 2, 3, \ldots, \infty)$].

3.1.1 Poisson distribution

For count data with range $R_Y = (0, 1, 2, 3, \ldots, \infty)$, the classical approach to model this type of data is using the Poisson distribution.

The probability function of the Poisson distribution, denoted $PO(\mu)$, is given by

$$P(Y = y | \mu) = \frac{e^{-\mu} \mu^y}{y!}$$  \hspace{1cm} (3.1)

where $y = 0, 1, 2, \ldots$, where $\mu > 0$.

However the Poisson distribution, $Y \sim PO(\mu)$ has only one parameter $\mu$, which equals the mean or expected value, $E(Y)$, of $Y$. Hence the moment based variance $V(Y) = \mu$, skewness $\sqrt{\beta_1} = \mu^{-\frac{1}{2}}$ and kurtosis $\beta_2 = 3 + \mu^{-1}$ of the Poisson distribution all depend on $\mu$. The variance of the Poisson distribution equals the mean, i.e. $V(Y) = E(Y)$, while the skewness ($\sqrt{\beta_1}$) and excess kurtosis ($\beta_2 - 3$) both tend to zero as $\mu$ increases.

There are three major problems often encountered when modeling count data using the Poisson distribution.

- overdispersion
- excess (or shortage) of zero values
- long right tail (i.e. high positive skewness).

Overdispersion is defined as the extra variation in a count response variable which is not explained by the Poisson distribution alone. The problem occurs because the population mean of a Poisson distributed random variable is equal to its variance. That is, $E(Y) = V(Y) = \mu$. 

59
Unfortunately very often for a count response variable $V(Y) > E(Y)$, i.e. the distribution is overdispersed Poisson.

The problem of excess (or shortage) of zero values occurs when the response variable has more (or less) zero values than expected from a Poisson distribution. This again is a phenomenon that occurs often in practice.

The third problem is with the right tail of the count data distribution. There are many situations where events appearing in the right tail of the distribution happen more often than the Poisson distribution would suggest. Because the Poisson distribution is a one parameter distribution its skewness and kurtosis are fixed given the mean $E(Y) = \mu$ and consequently events happening in the right tail of the distribution may not be modelled properly.

The one parameter discrete count distributions in GAMLSS, the Geometric, Logarithmic, Poisson, Waring and Yule distributions, (i.e. GEOM, LG, PO, WARING, and YULE respectively) are only able to model the location of the distribution. The details of these distributions are given in Chapter ???

The two parameter discrete count distributions in GAMLSS are able to model the location and scale (or dispersion) of the distribution (the negative binomial type I and type II and the Poisson-inverse Gaussian, i.e. NBI, NBII and PIG respectively), or to model the location and the zero probability of the distribution (the zero altered logarithmic, zero altered Poisson and zero inflated Poisson type 1 and type 2, i.e. ZALG, ZAP, ZIP and ZIP2 respectively).

The three parameter discrete count distributions in GAMLSS are able to model the location, scale (or dispersion) and positive skewness (i.e. long right tail) of the distribution (the Delaporte and two parameterizations of the Sichel, i.e. DEL, SI and SICHEL respectively), or to model the location, scale (or dispersion) and the zero probability of the distribution (the zero altered negative binomial type I, zero inflated negative binomial type I, and zero inflated Poisson inverse Gaussian, i.e. ZANBI, ZINBI and ZIPIG respectively). See Table 1 for a summary.

### 3.1.2 Overdispersion

Overdispersion has been recognised for a long time as a potential problem within the literature of generalised linear models, Nelder and Wedderburn [1972], which originally modelled only the mean of the distribution of the response variable. Over the years several solutions to the problem of overdispersion have been suggested, see e.g. Consul [1989] and Dossou-Gbéto and Mizère (2006). Here we consider three major categories:

(a) random effect at the observation level, dealt with below,

(b) discretised continuous distributions, dealt with in section 3.7,

(c) ad-hoc solutions, dealt with in section 3.8.

It should be noted that method (a) can only deal with overdispersion (and not underdispersion), while methods (b) and (c) can potentially deal with both overdispersion and underdispersion.

#### Random effect at the observation level

The random effect at the observation level accounts for the overdispersion by including an extra random effect variable. This can also solve the problems of excess of zero values and long tails in the data. The random effect methodology at the observation level works like this.

Assume that the conditional distribution of the response variable $Y$ is a discrete probability function $P(Y = y|\Gamma = \gamma)$, given $\Gamma = \gamma$, a value of a continuous random effect variable $\Gamma$ which
has a probability density function $f_{\Gamma}(\gamma)$. Then the marginal probability function of $Y$ is given by

$$P(Y = y) = \int P(Y = y|\Gamma = \gamma)f_{\Gamma}(\gamma)d\gamma.$$  \hspace{1cm} (3.2)

The resulting distribution of $Y$ is called a continuous mixture of discrete distributions.

When the random effect variable $\Gamma$ has a discrete probability function $P(\Gamma = \gamma)$ then

$$P(Y = y) = \sum_{\Gamma} P(Y = y|\Gamma = \gamma)P(\Gamma = \gamma).$$  \hspace{1cm} (3.3)

The resulting distribution of $Y$ is called a discrete mixture of discrete distributions. When $\Gamma$ takes only a finite number of possible values, then the resulting distribution is called a finite mixture of discrete distributions.

Within the random effect at the observation level models we distinguish three different types:

(a)(i) when $P(Y = y)$, the probability function of a continuous mixture of discrete distributions given by (3.2), exists explicitly. This is dealt with in section 3.2,

(a)(ii) when $P(Y = y)$, the probability function of a continuous mixture of discrete distributions given by (3.2), is not explicit but is approximated by integrating out the random effect using an approximation, e.g. Gaussian quadrature. This is dealt with in Chapter 11???,

(a)(iii) when $P(Y = y)$ is the probability function of a finite mixture of discrete distributions given by (3.3) where $\Gamma$ takes only a finite number of possible values. This is dealt with in Chapter 11???.

### 3.1.3 Excess or shortage of zero values

A solution to excess zero values in a particular discrete distribution is a zero inflated discrete distribution, dealt with in section 3.3. A solution to excess and/or shortage of zero values in a particular discrete distribution is a zero altered discrete distribution, dealt with in section 3.4.

### 3.1.4 Modelling count data in GAMLSS

Many random effect at the observation level models in category (a)(i) above can be fitted using \texttt{gamlss} packages. The package \texttt{gamlss.dist} contains a variety of (unlimited) count distributions, which can be fitted using the \texttt{gamlss} package.

Models in category (a)(ii) in which the continuous mixture of discrete distributions is approximated by integrating out a normally distributed random effect using Gaussian quadrature, can be fitted using the package \texttt{gamlss.mx}, as can a finite mixture of discrete distributions in category (a)(iii). In fact package \texttt{gamlss.mx} fits type (a)(ii) and (a)(iii) models above allowing a more general conditional distribution $P(Y = y|\Gamma = \gamma)$ to be used than the Poisson, e.g. a negative binomial distribution (resulting in a negative binomial-normal mixture model and a negative binomial non-parametric mixture model for $Y$ respectively). See Chapter 11???.

Discretised continuous distributions in category (b) can be fitted in GAMLSS using the package \texttt{gamlss.cens}. An example of this is given in section ???
Table 3.1: Discrete \textit{gamlss.family} distributions for count data (derived from Poisson mixtures)

<table>
<thead>
<tr>
<th>Marginal distribution for $Y$</th>
<th>GAMLSS name for $Y$</th>
<th>mixing distribution for $\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>PO$(\mu)$</td>
<td>-</td>
</tr>
<tr>
<td>Negative binomial type I</td>
<td>NBI($\mu, \sigma$)</td>
<td>GA($1, \sigma^2$)</td>
</tr>
<tr>
<td>Negative binomial type II</td>
<td>NBII($\mu, \sigma$)</td>
<td>GA($1, \sigma^2/\mu$)</td>
</tr>
<tr>
<td>Poisson-inverse Gaussian</td>
<td>PIG($\mu, \sigma$)</td>
<td>IG($1, \sigma^2$)</td>
</tr>
<tr>
<td>Sichel</td>
<td>SICHEL($\mu, \sigma, \nu$)</td>
<td>GIG($1, \sigma^2, \nu$)</td>
</tr>
<tr>
<td>Delaporte</td>
<td>DEL($\mu, \sigma, \nu$)</td>
<td>SG($1, \sigma^2, \nu$)</td>
</tr>
</tbody>
</table>

### 3.2 Explicit continuous mixtures of Poisson distributions

These are discrete distributions in category (a)(i). Suppose, given a random variable $\Gamma$ takes value $\gamma$, that $Y$ has a Poisson distribution with mean $\mu\gamma$, i.e. $Y|\Gamma = \gamma \sim PO(\mu\gamma)$, where $\mu > 0$, and suppose that $\Gamma$ has probability density function $f_\Gamma(\gamma)$ defined on $\mathbb{R}^+$, then the (marginal) distribution of $Y$ is a continuously mixed Poisson distribution. Provided $\Gamma$ has mean 1, then $Y$ has mean $\mu$. The model can be considered as a multiplicative Poisson random effect model, provided the distribution of $\Gamma$ does not depend on $\mu$.

For example $Y|\Gamma = \gamma \sim PO(\mu\gamma)$ where $\Gamma \sim GA(1, \sigma^1)$ gives marginal distribution $Y \sim NBI(\mu, \sigma)$. For simplicity of notation we use notation of the form $Y|\gamma \sim PO(\mu\gamma)$ where $\gamma \sim GA(1, \sigma^1)$ throughout this section. Table 3.1 shows some of those (marginal) overdispersed Poisson count data distributions for $Y$ used in applied statistics and their corresponding mixing distribution for $\gamma$, where $Y|\gamma \sim PO(\mu\gamma)$.

Many parameterizations of continuously mixed Poisson distributions [e.g., the Sichel and Delaporte distributions, see Johnson, Kotz and Kemp (2005) and Wimmer and Altmann [1999]] for a discrete count random variable $Y$ have been defined such that none of the parameters of the distribution is the mean of $Y$, and indeed the mean of $Y$ is often a complex function of the distribution parameters, making the distribution difficult to interpret for regression models.

Here we consider several mixed Poisson distribution defined so that the mean $\mu$ is a parameter of the distribution. This allows easier interpretation of models for $\mu$ and generally provides a more orthogonal parameterisation.

Specifically the following distributions with mean exactly equal to $\mu$ are considered: Poisson, negative binomial type I and type II, Poisson-inverse Gaussian, Sichel and Delaporte distributions. The distributions are continuous mixtures of Poisson distributions.

Table 3.2 shows the distributions for count data currently available in \textit{gamlss} packages, together with their R name within \textit{gamlss}, their number of parameters, mean, variance. Table 3.1 shows the mixing distribution for $\gamma$. The probability functions for all the distributions in Tables 3.1 and 3.2 are given in the Appendix ??, (except for SG distribution defined in section 3.6).

In Table 3.2 $\mu > 0$ and $\sigma > 0$ for all distributions, while $-\infty < \nu < \infty$ for the Sichel distribution and $0 < \nu < 1$ for the Delaporte distribution.

The Poisson-inverse Gaussian (PIG) is a special case of the Sichel where $\nu = -0.5$. The Poisson is a limiting case of the other distributions as $\sigma \to 0$ (with $\nu = -0.5$ for the Sichel and $\nu \to 0$ for the Delaporte).
3.2. EXPLICIT CONTINUOUS MIXTURES OF POISSON DISTRIBUTIONS

### Table 3.2: Discrete gamlss family distributions for count data

<table>
<thead>
<tr>
<th>Distributions</th>
<th>R Name</th>
<th>params</th>
<th>mean</th>
<th>variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>PO(µ)</td>
<td>1</td>
<td>µ</td>
<td>µ</td>
</tr>
<tr>
<td>Negative binomial type I</td>
<td>NBI(µ, σ)</td>
<td>2</td>
<td>µ</td>
<td>µ + σµ²</td>
</tr>
<tr>
<td>Negative binomial type II</td>
<td>NBII(µ, σ)</td>
<td>2</td>
<td>µ</td>
<td>µ + σµ²</td>
</tr>
<tr>
<td>Poisson-inverse Gaussian</td>
<td>PIG(µ, σ)</td>
<td>2</td>
<td>µ</td>
<td>µ + σµ²</td>
</tr>
<tr>
<td>Sichel</td>
<td>SICHEL(µ, σ, ν)</td>
<td>3</td>
<td>µ</td>
<td>µ + h(σ, ν)µ²</td>
</tr>
<tr>
<td>Delaporte</td>
<td>DEL(µ, σ, ν)</td>
<td>3</td>
<td>µ</td>
<td>µ + σ(1 - ν)µ²</td>
</tr>
<tr>
<td>Zero inflated Poisson</td>
<td>ZIP(µ, σ)</td>
<td>2</td>
<td>(1 - σ)µ</td>
<td>(1 - σ)µ + σ(1 - σ)µ²</td>
</tr>
<tr>
<td>Zero inflated Poisson type 2</td>
<td>ZIP2(µ, σ)</td>
<td>2</td>
<td>µ</td>
<td>µ + σ(1 - σ)µ²</td>
</tr>
</tbody>
</table>

#### 3.2.1 Negative binomial distribution

The negative binomial type I distribution, denoted NBI(µ, σ), is a continuously mixed Poisson distribution obtained as the marginal distribution of Y when Y|γ ~ PO(µγ) and γ ~ GA(1, σ²), i.e., γ has a gamma distribution with mean 1 and scale parameter σ².

The probability function of the negative binomial distribution type I, denoted NBI(µ, σ), is given by

\[ P(Y = y | µ, σ) = \frac{\Gamma(y + \frac{1}{σ})}{\Gamma(\frac{1}{σ})\Gamma(y + 1)} \left( \frac{σµ}{1 + σµ} \right)^y \left( \frac{1}{1 + σµ} \right)^{1/σ} \]

for y = 0, 1, 2, ..., where µ > 0 and σ > 0. The mean and variance of Y are given by E(Y) = µ and Var(Y) = µ + σµ².

Figure 3.1 plots the negative binomial type I distribution, NBI(µ, σ), for µ = 5 and σ = (0.01, 0.5, 1, 2). The plot was created using the command

`pdf.plot(family="NBI", mu=5, sigma=c(0.01, 0.5, 1, 2), min=0, max=20, step=1)`.

Note that plot for σ = 0.01 is close to a Poisson, PO(5), distribution which corresponds to µ = 5 and σ → 0 in the NBI(µ, σ) distribution.

The negative binomial type II distribution, denoted NBII(µ, σ), is a mixed Poisson distribution obtained as the marginal distribution of Y when Y|γ ~ PO(µγ) and γ ~ GA(1, (σ/µ)²). This is a reparameterization of the NBI distribution obtained by replacing σ by σ/µ.

The probability function of the negative binomial distribution type II, denoted here as NBII(µ, σ), is given by

\[ P(Y = y | µ, σ) = \frac{\Gamma(y + \mu/σ)σ^y}{\Gamma(µ/σ)\Gamma(y + 1)(1 + σ)^{µ+µ/σ}} \]

for y = 0, 1, 2, ..., where µ > 0 and σ > 0. The mean and variance of Y are given by E(Y) = µ and Var(Y) = (1 + σ)µ.
Figure 3.1: NBI distribution for $\mu = 5$ and $\sigma = (0.01, 0.5, 1, 2)$
3.2. EXPLICIT CONTINUOUS MIXTURES OF POISSON DISTRIBUTIONS

The NBI and NBII models differ when there are explanatory variables for \( \mu \) and/or \( \sigma \). The extra \( \sigma \) parameter allows the variance to change for a fixed mean, unlike the Poisson distribution for which the variance is fixed equal to the mean. Hence the negative binomial allows modelling of the variance as well as of the mean. The negative binomial distribution can be highly positively skewed, unlike the Poisson distribution, which is close to symmetric for moderate \( \mu \) and even closer as \( \mu \) increases.

The negative binomial family distribution, denoted \( \text{NBF}(\mu, \sigma, \nu) \) is obtained by replacing \( \sigma \) by \( \sigma \mu^{\nu-2} \) in the NBI distribution. This distribution has mean \( \mu \) and variance \( \sigma \mu^{\nu} \).

3.2.2 Poisson-inverse Gaussian

The Poisson-inverse Gaussian distribution, denoted \( \text{PIG}(\mu, \sigma) \), is a continuously mixed Poisson distribution obtained as the marginal distribution of \( Y \) when \( Y|\gamma \sim PO(\mu \gamma) \) and \( \gamma \sim IG(1, \sigma^{\frac{1}{2}}) \), an inverse Gaussian mixing distribution. This allows for even higher skewness, i.e. longer upper tail, than the negative binomial distribution.

The probability function of the Poisson-inverse Gaussian distribution, denoted \( \text{PIG}(\mu, \sigma) \), is given by

\[
P(Y = y|\mu, \sigma) = \left( \frac{2\alpha}{\pi} \right)^{\frac{1}{2}} \frac{\mu^{\nu e^{\frac{1}{\sigma}} K_{\nu-\frac{1}{2}}(\alpha)}}{(\alpha \sigma)^{\nu y!}}
\]

where \( \alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma^2} \), for \( y = 0, 1, 2, ..., \infty \) where \( \mu > 0 \) and \( \sigma > 0 \) and \( K_{\nu}(t) = \frac{1}{2} \int_{0}^{\infty} x^{\nu-1} \exp\left\{ -\frac{1}{2} t(x + x^{-1}) \right\} dx \) is the modified Bessel function of the third kind. The mean and variance of \( Y \) are given by \( E(Y) = \mu \) and \( \text{Var}(Y) = \mu + \sigma \mu^{2} \).

Figure 3.2 plots the Poisson-inverse Gaussian distribution, \( \text{PIG}(\mu, \sigma) \), for \( \mu = 5 \) and \( \sigma = (0.01, 0.5, 1, 2) \). Note that plot for \( \sigma = 0.01 \) is close to a Poisson, \( PO(5) \), distribution. The plot was created using the command

\[
\text{pdf.plot(family="NBI", mu=5, sigma=c(0.01, 0.5, 1, 2), min=0, max=20, step=1)}
\]

3.2.3 Sichel distribution

The Sichel distribution has been found to provide a useful three parameter model for over-dispersed Poisson count data exhibiting high positive skewness, e.g. Sichel (1992). In the parametrization below \( \mu \) is the mean of the Sichel distribution, while the two remaining parameters \( \sigma \) and \( \nu \) jointly define the scale and shape of the Sichel distribution. In particular the three parameters of the Sichel allow different shapes (in particular the level of positive skewness) of the distribution for a fixed mean and variance, unlike the negative binomial and Poisson-inverse Gaussian distributions. The Sichel distribution therefore allows modelling of the mean, variance and skewness.

The Sichel distribution, denoted \( \text{SICHEL}(\mu, \sigma, \nu) \), is a continuously mixed Poisson distribution obtained as the marginal distribution of \( Y \) when \( Y|\gamma \sim PO(\mu \gamma) \) and \( \gamma \sim GIG(1, \sigma^{\frac{1}{2}}, \nu) \), a generalized inverse Gaussian mixing distribution with probability density function given by

\[
f_{\nu}(\gamma) = \frac{\nu^{\nu-1} \exp\left\{ -\frac{1}{2\sigma} (c\gamma + \frac{1}{c\gamma}) \right\}}{2K_{\nu}(\frac{c}{2})}
\]

for \( \gamma > 0 \), where \( \sigma > 0 \) and \( -\infty < \nu < \infty \). The parameterization (3.4) of the GIG ensures that \( E[\gamma] = 1 \).
Figure 3.2: PIG distribution for $\mu = 5$ and $\sigma = (0.01, 0.5, 1, 2)$
3.3. ZERO INFLATED DISCRETE DISTRIBUTIONS

The probability function of the Sichel distribution, Rigby, Stasinopoulos and Akantziliotou (2008), denoted by SICHEL ($\mu, \sigma, \nu$), is given by

$$P(Y = y|\mu, \sigma, \nu) = \frac{\mu^yK_{y+\nu}(\alpha)}{y!(\alpha\sigma)^y\Gamma(\nu)\Gamma(1/\sigma)}$$ (3.5)

for $y = 0, 1, 2, \ldots, \infty$, where $\alpha^2 = \sigma^{-2} + 2(\alpha \sigma)^{-1}$, $c = R_\nu(1/\sigma)$, $R_\nu(t) = K_{\nu+1}(t)/K_\nu(t)$ and $K_\nu(t) = \frac{1}{\pi} \int_0^\infty \frac{t^{\nu-1}}{x^{\nu+1}} \exp[-\frac{1}{2}t(x+x^{-1})] dx$ is the modified Bessel function of the third kind. The mean and variance of $Y$ are given by $E(Y) = \mu$ and $Var(Y) = \mu + \mu^2 \left[2\sigma(\nu+1)/c + 1/c^2 - 1\right]$ respectively. [Note that for the Sichel distribution $h(\sigma, \nu) = \left[2\sigma(\nu+1)/c + 1/c^2 - 1\right]$ in Table 3.2.]

3.2.4 Delaporte distribution

The Delaporte distribution, denoted DEL ($\mu, \sigma, \nu$), is a mixed Poisson distribution obtained as the marginal distribution of $Y$ when $Y|\gamma \sim PO(\mu \gamma)$ and $\gamma \sim SG(1, \sigma^2, \nu)$, a shifted gamma mixing distribution with probability density function given by

$$f_\gamma(\gamma) = \frac{(\gamma - \nu)^{\frac{1}{2} - 1}}{\sigma^{1/\sigma}(1 - \nu)^{1/\sigma}\Gamma(1/\sigma)} \exp \left[-\frac{(\gamma - \nu)}{\sigma(1 - \nu)}\right]$$ (3.6)

for $\gamma > \nu$, where $\sigma > 0$ and $0 \leq \nu < 1$. This parameterization ensures that $E[\gamma] = 1$.

The probability function of the Delaporte distribution, denoted DEL ($\mu, \sigma, \nu$), is given by

$$P(Y = y|\mu, \sigma, \nu) = \frac{e^{-\mu\nu}}{\Gamma(1/\sigma)} \left[1 + \mu(1 - \nu)/\nu \right]^{-1/\sigma} S$$ (3.7)

where

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

for $y = 0, 1, 2, \ldots, \infty$ where $\mu > 0$, $\sigma > 0$ and $0 \leq \nu < 1$. The mean of $Y$ is given by $E(Y) = \mu$ and the variance by $Var(Y) = \mu + \mu^2\sigma(1-\nu)^2$.

3.3 Zero inflated discrete distributions

A discrete response variable $Y$ can exhibit a greater probability of value zero than that of a particular discrete distribution, denoted here by $D$. This can be modeled by a 'zero inflated' distribution, denoted here by ZID, i.e. $Y \sim ZID$, which is a discrete mixture of two components: value 0 with probability $p$ and a distribution $D$ with probability $1-p$.

Hence $Y = 0$ with probability $p$ and $Y = Y_1$ with probability $(1-p)$, where $Y_1 \sim D$ and $0 < p < 1$.

Hence

$$P(Y = 0) = p + (1-p)P(Y_1 = 0)$$

$$P(Y = y) = (1-p)P(Y_1 = y)$$
if $y \neq 0$ and $y \epsilon R_y$.

Note that the probability that $Y = 0$ has two components, first $p$ and second $(1 - p)P(Y_1 = 0)$ from $Y_1 \sim D$. Note also that $P(Y = 0) > P(Y_1 = 0)$ and hence the distribution is called a ‘zero inflated’ distribution.


### 3.3.1 Zero inflated Poisson distribution

The zero inflated Poisson distribution, denoted ZIP, i.e. $Y \sim ZIP(\mu, \sigma)$, is a discrete mixture of two components: value 0 with probability $\sigma$ and a Poisson distribution with mean $\mu$ with probability $1 - \sigma$.

Hence $Y = 0$ with probability $\sigma$ and $Y = Y_1$ with probability $(1 - \sigma)$, where $Y_1$ has a Poisson distribution with mean $\mu$, i.e. $Y_1 \sim PO(\mu)$, and $0 < \sigma < 1$.

Hence the probability function (pf) of $Y \sim ZIP(\mu, \sigma)$ is given by

$$P(Y = 0|\mu, \sigma) = \sigma + (1 - \sigma)e^{-\mu}$$

$$P(Y = y|\mu, \sigma) = (1 - \sigma)\frac{\mu^y}{y!}e^{-\mu}$$

if $y = 1, 2, 3, ..., \text{ where } \mu > 0 \text{ and } 0 < \sigma < 1$.

Note that the probability that $Y = 0$ has two components, first $\sigma$ and second $(1 - \sigma)e^{-\mu}$ from the Poisson distribution $Y_1 \sim PO(\mu)$. Note also that $P(Y = 0) > P(Y_1 = 0)$ and hence the distribution is called a zero inflated Poisson distribution.

Figure 3.3 plots the zero inflated Poisson distribution, ZIP$(\mu, \sigma)$, for $\mu = 5$ and $\sigma = (0.01, 0.1, 0.4, 0.7)$. Note that plot for $\sigma = 0.01$ is close to a Poisson, PO(5), distribution. The plot was created using the command

```
pdf.plot(family="ZIP", mu=5, sigma=c(0.01, 0.1, 0.4, 0.7), min=0, max=12, step=1).
```

The mean and variance of $Y$ are given by $E[Y] = (1 - \sigma)\mu$ and $V[Y] = (1 - \sigma)\mu + \sigma(1 - \sigma)\mu^2$.

The ZIP distribution can be viewed as a discrete mixed Poisson distribution defined by the marginal distribution of $Y$ where $Y|\gamma \sim PO(\mu\gamma)$ and $\gamma \sim BI(1, 1 - \sigma)$, i.e. $\gamma = 0$ with probability $\sigma$ and $\gamma = 1$ with probability $1 - \sigma$. Note however that $\gamma$ has mean $1 - \sigma$ in this formulation and $Y$ has mean $(1 - \sigma)\mu$.

### 3.3.2 Zero inflated Poisson type 2 parameterization

A different parameterization of the zero inflated poisson distribution, denoted ZIP2$(\mu, \sigma)$, has pf given by

$$P(Y = 0|\mu, \sigma) = \sigma + (1 - \sigma)e^{-\frac{\mu}{1 - \sigma}}$$

$$P(Y = y|\mu, \sigma) = (1 - \sigma)\frac{\mu^y}{y!(1 - \sigma)^y}e^{-\frac{\mu}{1 - \sigma}}$$

if $y = 1, 2, 3, ..., \text{ where } \mu > 0 \text{ and } 0 < \sigma < 1$.

The mean of $Y$ in (17.5) is given by $E(Y) = \mu$ and the variance by $Var(Y) = \mu + \mu^2\frac{\sigma}{(1 - \sigma)}$.

The zero inflated Poisson type 2 distribution, denoted ZIP2, i.e. $Y \sim ZIP2(\mu, \sigma)$, is the marginal distribution for $Y$ where $Y|\gamma \sim PO(\mu\gamma)$ and $\gamma \sim (1 - \sigma)^{-1}BI(1, 1 - \sigma)$. Hence $\gamma$ has mean 1 and $Y$ has mean $\mu$. 

3.3. ZERO INFLATED DISCRETE DISTRIBUTIONS

Figure 3.3: ZIP distribution for $\mu = 5$ and $\sigma = (0.01, 0.1, 0.4, 0.7)$
3.3.3 Zero inflated negative binomial type 1 distribution

The zero inflated negative binomial type 1 distribution, denoted ZINBI, i.e. \( Y \sim ZINBI(\mu, \sigma, \nu) \), is a discrete mixture of two components: value 0 with probability \( \nu \) and a negative binomial type 1 distribution, \( NBI(\mu, \sigma) \), with probability \( 1 - \nu \).

Hence \( Y = 0 \) with probability \( \nu \) and \( Y = Y_1 \) with probability \( (1 - \nu) \), where \( Y_1 \) has a negative binomial type I distribution, i.e. \( Y_1 \sim NBI(\mu, \sigma) \), and \( 0 < \nu < 1 \).

Hence the pdf of \( Y \sim ZINBI(\mu, \sigma, \nu) \) is given by

\[
P(Y = 0|\mu, \sigma, \nu) = \nu + (1 - \nu)P(Y_1 = 0|\mu, \sigma)
\]

\[
P(Y = y|\mu, \sigma, \nu) = (1 - \nu)P(Y_1 = y|\mu, \sigma)
\]

if \( y = 1, 2, 3, \ldots \) where \( Y_1 \sim NBI(\mu, \sigma) \) so

\[
P(Y_1 = 0|\mu, \sigma) = (1 + \sigma \mu)^{-\frac{1}{\sigma}}
\]

and

\[
P(Y_1 = y|\mu, \sigma) = \frac{\Gamma\left(y + \frac{1}{\sigma}\right)}{\Gamma\left(\frac{1}{\sigma}\right)\Gamma(y + 1)} \left(\frac{\sigma \mu}{1 + \sigma \mu}\right)^y \left(\frac{1}{1 + \sigma \mu}\right)^{1/\sigma}
\]

for \( y = 0, 1, 2, 3, \ldots \), where \( \mu > 0, \sigma > 0 \) and \( 0 < \nu < 1 \).

The mean and variance of \( Y \) are given by \( E[Y] = (1-\nu)\mu \) and \( V[Y] = (1-\nu)\mu[1+(\sigma+\nu)\mu] \).

3.3.4 Zero inflated Poisson inverse Gaussian distribution

The zero inflated Poisson inverse Gaussian distribution, denoted by ZIPIG, i.e. \( Y \sim ZIPIG(\mu, \sigma, \nu) \), is a discrete mixture of two components: value 0 with probability \( \nu \) and a Poisson inverse Gaussian distribution \( PIG(\mu, \sigma) \) with probability \( 1 - \nu \).

Hence \( Y = 0 \) with probability \( \nu \) and \( Y = Y_1 \) with probability \( (1 - \nu) \), where \( Y_1 \) has a Poisson inverse Gaussian distribution, i.e. \( Y_1 \sim PIG(\mu, \sigma) \), and \( 0 < \nu < 1 \).

Hence the pdf of \( Y \sim ZIPIG(\mu, \sigma, \nu) \) is given by

\[
P(Y = 0|\mu, \sigma, \nu) = \nu + (1 - \nu)P(Y_1 = 0|\mu, \sigma)
\]

\[
P(Y = y|\mu, \sigma, \nu) = (1 - \nu)P(Y_1 = y|\mu, \sigma)
\]

if \( y = 1, 2, 3, \ldots \) where \( Y_1 \sim PIG(\mu, \sigma) \), where \( \mu > 0, \sigma > 0 \) and \( 0 < \nu < 1 \).

The mean and variance of \( Y \) are given by \( E[Y] = (1-\nu)\mu \) and \( V[Y] = (1-\nu)\mu[1+(\sigma+\nu)\mu] \).

3.4 Zero altered discrete distributions

A discrete response variable \( Y \) can exhibit either a greater or less probability of value zero than that of a particular discrete distribution, denoted here by \( D \).

This can be modeled by a 'zero altered' distribution, denoted here by \( ZAD \), i.e. \( Y \sim ZAD \), which is a discrete mixture of two components: value 0, with probability \( p \), and a distribution \( D_{tr} \), (the distribution \( D \) truncated at zero), with probability \( 1 - p \).

Hence \( Y = 0 \) with probability \( p \) and \( Y = Y_1 \) with probability \( (1 - p) \), where \( Y_1 \sim D_{tr} \) and \( 0 < p < 1 \).
3.4. ZERO ALTERED DISCRETE DISTRIBUTIONS

Hence

\[ P(Y = 0) = p \]

\[ P(Y = y) = (1 - p)P(Y_1 = y) \]

if \( y \neq 0 \) and \( y \in Y_1 \), where \( Y_1 \sim Dtr. \)

Hence

\[ P(Y = 0) = p \]

\[ P(Y = y) = \frac{(1 - p)P(Y_2 = y)}{1 - P(Y_2 = 0)} \] (3.8)

if \( y \neq 0 \) and \( y \in Y_2 \), where \( Y_2 \sim D. \)

Note that the probability that \( Y = 0 \) is exactly \( p \). Note also that \( P(Y = 0) = p \) where \( 0 < p < 1 \), so \( P(Y = 0) \) can be greater than or less than \( P(Y_2 = 0) \) and hence the distribution is called a 'zero altered' distribution. If \( p = P(Y_2 = 0) \) in (5.1) then \( Y = Y_2 \sim D \) and ZAD becomes distribution \( D \). If \( p > P(Y_2 = 0) \) in (5.1) then the zero altered distribution ZAD is a reparameterization of the zero inflated distribution ZID in section 3.3 (but the models are different when the parameters are modelled by explanatory variables). However if \( p < P(Y_2 = 0) \) in (5.1) then the zero altered distribution ZAD is different from (i.e. not a reparameterization of) the zero inflated distribution ZID. Hence ZID is a reparameterized submodel of ZAD.

The mean and variance of \( Y \) are given by

\[ E(Y) = \frac{(1 - p)E(Y_2)}{1 - P(Y_2 = 0)} \]

and

\[ Var(Y) = E(Y) \left[ \frac{V(Y_2)}{E(Y_2)} + E(Y_2) \right] - [E(Y)]^2. \]

### 3.4.1 Zero altered Poisson distribution

A zero altered Poisson distribution, denoted ZAP, i.e. \( Y \sim ZAP(\mu, \sigma) \), is a discrete mixture of two components: value 0, with probability \( \sigma \), and a distribution \( P0tr(\mu) \), (the Poisson distribution \( PO(\mu) \) truncated at zero), with probability \( 1 - \sigma \).

Hence \( Y = 0 \) with probability \( \sigma \) and \( Y = Y_1 \) with probability \( (1 - \sigma) \), where \( Y_1 \sim P0tr(\mu) \) and \( 0 < \sigma < 1 \).

Hence

\[ P(Y = 0|\mu, \sigma) = \sigma \]

\[ P(Y = y|\mu, \sigma) = (1 - \sigma)P(Y_1 = y|\mu) = \frac{(1 - \sigma)P(Y_2 = y|\mu)}{1 - P(Y_2 = 0|\mu)} \]

if \( y \neq 0 \) and \( y = 1, 2, 3, ..., \) where \( Y_1 \sim P0tr(\mu) \) and \( Y_2 \sim PO(\mu) \).

Hence the pf of \( Y \sim ZAP(\mu, \sigma) \) is given by

\[ P(Y = 0|\mu, \sigma) = \sigma \]
\[
P(Y = y|\mu, \sigma) = \frac{(1 - \sigma)e^{-\mu}\mu^y}{y!(1 - e^{-\mu})}
\]
if \(y = 1, 2, 3, \ldots \), where \(\mu > 0\) and \(0 < \sigma < 1\).

Note also that \(P(Y = 0) = \sigma\) and \(0 < \sigma < 1\), so \(P(Y = 0)\) can be greater than or less than the probability of 0 for a Poisson, \(PO(\mu)\) distribution and hence the distribution is called a ‘zero altered’ distribution.

The mean and variance of \(Y\) are given by \(E(Y) = (1 - \sigma)\mu/(1 - e^{-\mu})\) and \(V(Y) = (1 + \mu)E(Y) - [E(Y)]^2\).

### 3.5 Zero altered logarithmic distribution, ZALG

Let \(Y = 0\) with probability \(\sigma\) and \(Y = Y_1\), where \(Y_1 \sim LG(\mu)\) a logarithmic distribution, with probability \((1 - \sigma)\), then \(Y\) has a zero altered logarithmic distribution, denoted by \(ZALG(\mu, \sigma)\), with probability function given by

\[
p_Y(y|\mu, \sigma) = \begin{cases} 
\sigma, & \text{if } y = 0 \\
(1 - \sigma)\alpha^y, & \text{if } y = 1, 2, 3, \ldots 
\end{cases}
\]

where \(\alpha = \frac{\log(1 - \mu)}{\log(1 - e^{-\mu})}\) for \(0 < \mu < 1\) and \(0 < \sigma < 1\). The mean and variance of \(Y\) are given by \(E(Y) = \frac{(1 - \alpha)\mu}{(1 - \mu)}\) and its variance by \(Var(Y) = \frac{(1 - \alpha)\mu(1 - (1 - \sigma)\alpha)\mu}{(1 - \mu)^2}\).

#### 3.5.1 Zero altered negative binomial type 1 distribution, ZANBI

A zero altered negative binomial type 1 distribution, denoted \(ZANBI\), i.e. \(Y \sim ZANBI(\mu, \sigma, \nu)\), is a discrete mixture of two components: value 0, with probability \(\nu\), and a distribution \(NB1tr(\mu, \sigma)\), [the negative binomial type 1 distribution \(NB1(\mu, \sigma)\) truncated at zero], with probability \(1 - \nu\).

Hence \(Y = 0\) with probability \(\nu\) and \(Y = Y_1\) with probability \((1 - \nu)\), where \(Y_1 \sim NB1tr(\mu, \sigma)\) and \(0 < \nu < 1\).

Hence

\[P(Y = 0|\mu, \sigma, \nu) = \nu\]

\[P(Y = y|\mu, \sigma, \nu) = (1 - \nu)P(Y_1 = y|\mu, \sigma)\]

if \(y \neq 0\) and \(y = 1, 2, 3, \ldots\), where \(Y_1 \sim NB1tr(\mu, \sigma)\).

Hence the pf of \(Y \sim ZANBI(\mu, \sigma, \nu)\) is given by

\[P(Y = 0|\mu, \sigma, \nu) = \nu\]

\[P(Y = y|\mu, \sigma, \nu) = \frac{(1 - \nu)P(Y_1 = y|\mu, \sigma)}{1 - P(Y_1 = 0|\mu, \sigma)}\]

if \(y = 1, 2, 3, \ldots\), for \(\mu > 0\), \(\sigma > 0\) and \(0 < \nu < 1\) where \(Y_2 \sim NB1(\mu, \sigma)\) so

\[P(Y_2 = 0|\mu, \sigma) = (1 + \sigma\mu)^{-\frac{1}{\sigma}}\]
and

\[ P(Y_2 = y | \mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma}) \left( \frac{\sigma \mu}{1 + \sigma \mu} \right)^y \left( \frac{1}{1 + \sigma \mu} \right)^{1/\sigma}}{\Gamma(\frac{1}{\sigma}) \Gamma(y + 1)} \]

for \( y = 0, 1, 2, 3, \ldots \). The mean of \( Y \) is given by

\[ E(Y) = \frac{(1 - \nu) \mu}{1 - (1 + \sigma \mu)^{-\frac{1}{\sigma}}} \]

and the variance by

\[ V(Y) = [1 + (\sigma + 1)\mu] E(Y) - [E(Y)]^2 \]

### 3.6 Comparison of the count distributions

Count distributions for \( Y \) can be compared using a diagram of their kurtosis against their skewness. Figure 3.4 displays the skewness-kurtosis combinations for different distributions of \( Y \), where \( Y \) has fixed mean 5 and fixed variance 30. Similar figures were obtained for other combinations of fixed mean and variance of \( Y \).

The zero-inflated Poisson (ZIP), negative binomial (NBI), negative binomial truncated at zero (NBItr) and Poisson-inverse Gaussian (PIG) distributions each have two parameters, so fixing the mean and variance of \( Y \) results in a single combination of skewness-kurtosis, displayed as circles in Figure 3.4.

Figure 3.5 shows the corresponding ZIP(10,0.5), NBI(5,1), NBItr(2.77,2.61) and PIG(5,1) distributions each having mean 5 and variance 30.

The zero inflated negative binomial distribution (ZINBI) and the zero adjusted negative binomial distribution (ZANBI) each have three parameters, so their possible skewness-kurtosis combinations are represented by curves. The zero inflated negative binomial distribution (ZINBI) skewness-kurtosis curve is the curve (shown in Figure 3.4 but not labeled) from the skewness-kurtosis of the ZIP to that of the NBI. Figure 3.6 shows ZINBI distributions corresponding to four points along the ZINBI skewness-kurtosis curve from ZIP to NBI, where the extra probability \( \nu \) at \( Y = 0 \) decreases from effectively 0.5 to 0.09091. All four ZINBI distributions have mean 5 and variance 30.

The zero adjusted negative binomial distribution (ZANBI) skewness-kurtosis curve is the curve (shown in Figure 3.4 but not labeled) from the skewness-kurtosis of the ZIP to that of the NBItr. In general the ZANBI distribution can lead to a zero inflated or zero deflated probability relative to the NBI distribution. In the zero inflated case, the ZANBI distribution is a re-parameterization of the ZINBI distribution, and hence has the same skewness-kurtosis curve (between ZIP and NBI in Figure 3.4 for fixed mean 5 and fixed variance 30). In the zero deflated case the ZANBI distribution is different from (i.e. not a reparameterization of) the ZINBI distribution, and its skewness-kurtosis curve lies between NBI and NBItr in Figure 3.4 for fixed mean 5 and fixed variance 30). Figure 3.7 shows ZANBI distributions corresponding to four points along part of the ZANBI skewness-kurtosis curve which lies between NBI and NBItr, where the exact probability \( \nu \) at \( Y = 0 \) decreases from 0.1667 [equal to the NBI(5,1) distribution] to close to zero.

The Sichel, Poisson-Tweedie and Delaporte distributions each have three parameters, so their possible skewness-kurtosis combinations are represented by curves. The three curves meet
at the skewness-kurtosis point of the negative binomial which is a limiting case of the Sichel, an internal special case of the the Poisson-Tweedie and a boundary special case of the Delaporte. The Poisson-Tweedie curve alone continues (as its power parameter decreases from two to one) and stops at the circle market PT between ZIP and NBI. [Note also that the PIG is a special case of both the Sichel and the Poisson-Tweedie distributions.] The Poisson-Tweedie distribution is not explicitly implemented yet in the *gamlss* packages. The probability function for the Poisson-Tweedie is given by Hougaard *et al.* (1997). It is a mixture of Poisson $PO(\mu \gamma)$ distributions, with a Tweedie mixing distribution for $\gamma$. The zero-inflated Poisson reciprocal Gamma (ZIPRG) curve has the highest kurtosis for a given skewness.

Figure 3.8 shows four SICHEL distributions with increasing values of $\nu$, all having mean 5 and variance 30, moving away from the NBI(5,1) distribution.

Figure 3.9 shows four DEL distributions with increasing values of $\nu$, all having mean 5 and variance 30, moving away from the NBI(5,1) distribution.

The Poisson-shifted generalized inverse Gaussian (PSGIG), Rigby, Stasinopoulos and Akantziliotou (2008), is a four parameter distribution and has skewness-kurtosis combinations covering the region between the Sichel and Delaporte curves in Figure 3.4, while the zero-inflated Sichel (ZISichel) covers the region between the ZIPRG and Sichel curves in Figure 3.4.

![Figure 3.4: Skewness-kurtosis combinations for different distributions for Y (with fixed mean 5 and variance 30)](image-url)
3.7 Discretised continuous distributions method

By discretised continuous distributions, category (b) solutions in section 3.1.2, we refer to methods which use continuous distributions to create a discrete one. For example, let $F_W(w)$ to be the cumulative distribution function of a continuous random variable $W$ defined in $\mathbb{R}^+$ then $P(Y = y) = F_W(y + 1) - F_W(y)$ is a discrete distribution defined on $y = 0, 1, 2, \ldots, \infty$. Alternatively let $P(Y = 0) = F_W(0.5)$ and $P(Y = y) = F_W(y + 0.5) - F_W(y - 0.5)$ for $y = 1, 2, \ldots, \infty$. Distributions of this kind can be fitted easily using the `gamlss.cens` package. One potential criticism of the above methods of generating discrete distributions is the fact that if the parameter $\mu_W$ is the mean of the continuous random variable $W$, then the mean of the discrete random variable $Y$ will not in general be exactly $\mu_W$.

Note that the discretised method above can cope with underdispersion, $V(Y) < E(Y)$, as well as overdispersion, $V(Y) > E(Y)$, in count data.
Figure 3.6: ZINBI distributions for $Y$ corresponding to four points along the ZINBI skewness-kurtosis curve from ZIP to NBI (with fixed mean 5 and variance 30)
Figure 3.7: ZANBI distributions for $Y$ corresponding to four points along part of the ZANBI skewness-kurtosis curve from NBI to NBItr (with fixed mean 5 and variance 30)
Figure 3.8: four SICHEL distributions for $Y$ with increasing values of $\nu$, all having mean 5 and variance 30
Figure 3.9: four DEL distributions for $Y$ with increasing values of $\nu$, all having mean 5 and variance 30
3.8 Ad-hoc methods

We refer to ad-hoc solutions, i.e. category (c) in section 3.1.2, as those that have been implemented in the past, mainly for their computational convenience (and some also for good asymptotic properties for the estimation of the mean regression function), but which do not assume an explicit proper distribution for the response variable. The quasi-likelihood function approach proposed by Wedderburn [1974], for example, requires assumptions on the first two moments of the response variable. The quasi-likelihood approach is incapable of modelling the second moment parameter, the dispersion, as a function of explanatory variables, therefore the extended quasi-likelihood (EQL) was proposed by Nelder and Pregibon [1987]. Alternatively approaches are the pseudo-likelihood (PL) method introduced by Carroll and Ruppert [1982] and Efron’s double exponential (EDE) family, Efron [1986]. The PL method effectively approximates the probability function by a normal distribution with a chosen variance-mean relationship, but does not properly maximise the resulting likelihood. See Davidian and Carroll [1988] and Nelder and Lee [1992] for a comparison of the EQL and the PL. The problem with all these methods is that, while they work well with moderate underdispersion or overdispersion, they have difficulty modelling long tails in the distribution of the response variable. They also suffer from the fact, that, for a given set of data, the adequacy of the fit of those methods cannot be compared using a properly maximised log likelihood function $\hat{\ell}$ and criteria based on $\hat{\ell}$, e.g. the (generalised) Akaike information criterion $AIC = -2\hat{\ell} + k.df$, where $k$ is the penalty and $df$ denotes the total (effective) degrees of freedom used in the model. The problem is that they do not properly fit a discrete distribution. For the EQL and EDE methods the distribution probabilities do not add up to one, see for example Stasinopoulos [2006]. Note that with increasing computer power the constant of summation, missing from the EQL and EDE methods, can be calculated so that they represent proper distributions resulting in a true likelihood function that can be maximised. However these models are still computational slow to fit to large data sets, the true probability function cannot be expressed explicitly (except by including an infinite sum for the constant of summation) and their flexibility is limited by usually having at most two parameters. See Lindsey [1999b] for a similar criticism of the ad-hoc methods.

3.9 Examples: fitting a distribution

3.10 Bibliography

Appendix of Chapter ??: Skewness and kurtosis for a mixture of Poisson distributions.

Let $Y|\gamma \sim \text{PO}(\mu \gamma)$ and $\gamma$ have a distribution with cumulative generating function $K_\gamma(t)$, then the cumulative generating function of the marginal distribution of $Y$, $K_Y(t)$, is given by

$$K_Y(t) = K_\gamma \left[ \mu \left( e^t - 1 \right) \right]$$

and hence, assuming that $\gamma$ has mean 1, the cumulants of $Y$ and $\gamma$ are related by $E(Y) = \mu$, $V(Y) = \mu + \mu^2V(\gamma)$, $\kappa_3Y = \mu + 3\mu^2V(\gamma) + \mu^3\kappa_3\gamma$, $\kappa_4Y = \mu + 6\mu^2V(\gamma) + 4\mu^3\kappa_3\gamma + \mu^4\kappa_4\gamma$. 

where $\kappa_{3Y}$ and $\kappa_{4Y}$ are the third and fourth cumulants of $Y$.

The skewness and kurtosis of $Y$ are $\sqrt{\beta_1} = \kappa_{3Y}/[V(Y)]^{1.5}$ and $\beta_2 = 3 + \{\kappa_{4Y}/[V(Y)]^2\}$ respectively. An example of the Sichel distribution for $Y$ is given below.

**Sichel distribution**

If $Y$ has a Sichel($\mu, \sigma, \nu$) distribution then the mean, variance, skewness and kurtosis of $Y$ are obtained using (3.10) from the cumulants of the mixing distribution $\gamma \sim \mathrm{GIG}(1, \sigma^{1/2}, \nu)$, with pdf defined (3.4), are given by

$$E(\gamma) = 1, \quad V(\gamma) = g_1, \quad \kappa_3(\gamma) = g_2 - 3g_1, \quad \kappa_4(\gamma) = g_3 - 4g_2 + 6g_1 - 3g_1^2,$$

where $g_1 = [1/c^2 + 2\sigma(\nu + 1)/c - 1],
g_2 = 2\sigma(\nu + 2)/c^3 + \left[4\sigma^2(\nu + 1)(\nu + 2) + 1\right]/c^2 - 1,
g_3 = \left[1 + 4\sigma^2(\nu + 2)(\nu + 3)\right]/c^4 + \left[8\sigma^3(\nu + 1)(\nu + 2)(\nu + 3) + 4\sigma(\nu + 2)\right]/c^3 - 1.$

The cumulants for the Sichel distribution for $Y$ are given by (3.10), from which the mean, variance, skewness and kurtosis of $Y$ are obtained.

**Exercises for Chapter ??**

- Q1 Gupta et al. (1996) present the following data giving the number of Lamb foetal movements $y$ observed with frequency $f$ recorded by ultrasound over 240 consecutive five second intervals:

<table>
<thead>
<tr>
<th>$y$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>182</td>
<td>41</td>
<td>12</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) Fit each of the following distributions for $y$ to the data (using different model names e.g. mPO etc. for later comparison): PO($\mu$), NBI($\mu, \sigma$), NBII($\mu, \sigma$), PIG($\mu, \sigma$), Sichel($\mu, \sigma, \nu$), DEL($\mu, \sigma, \nu$) and ZIP($\mu, \sigma$). [Note that the default fitting method RS may be slow for the Sichel distribution, so try using e.g. method=mixed(2,100), which performs 2 iterations of the RS algorithm, followed by (up to) 100 iterations of the CG algorithm.]

(b) Use the AIC command with each of the penalties $k = 2, 3.8$ and $5.48 = \log(240)$, [corresponding to criteria AIC, $\chi^2_{1,0.05}$ and SBC respectively], in order to select a distribution model. Output the parameter estimates for your chosen model. [Note that the residuals for frequency data are not currently implemented.]


- Q2 The USA National AIDS Behavioural Study recorded $y$, the number of times individuals engaged in risky sexual behaviour during the previous six months, together with two explanatory factors sex of individual (male or female) and whether they has a risky partner risky (no or yes), giving the following frequency distributions:
The data were previously analysed by Heilbron (1994).

(a) Read the above frequencies (corresponding to the male yes, male no, female yes, female no rows of the above table) into a variable f. Read the corresponding count values into y. Use the following code:

```r
y <- rep(c(0:7), 10, 12, 15, 20, 30, 37, 50, 4)
```

Generate a single factor type for type of individual with four levels (corresponding to male yes, male no, female yes, female no) by

```r
type <- gl(4, 15)
```

(b) Fit each of the following distributions for y to the data (using different model names for later comparison): PO \((\mu)\), NBI\((\mu, \sigma)\), NBII\((\mu, \sigma)\), PIG\((\mu, \sigma)\), SICHEL\((\mu, \sigma, \nu)\), DEL\((\mu, \sigma, \nu)\) and ZIP\((\mu, \sigma)\), using factor type for the mean model and a constant scale (and shape).

(c) Use the AIC command with each of the penalties \(k = 2, 3\) and \(4\), in order to select a distribution model.

(d) Check whether your chosen distribution model needs the factor type in its scale (and shape) models. Check whether the factor type is needed in the mean model.

(e) Output the parameter estimates for your chosen model.

Chapter 4

Binomial data distributions

4.1 Available distributions

The binomial distribution is denoted BI($n, \mu$) in gamlss for $y = 0, 1, \ldots, n$, where $0 < \mu < 1$ and $n$ is a known positive integer called the binomial denominator, (bd in the R code). The binomial distribution has mean $n\mu$ and variance $n\mu(1 - \mu)$.

The beta binomial distribution, is denoted BB($n, \mu, \sigma$) in gamlss for $y = 0, 1, \ldots, n$, where $0 < \mu < 1, \sigma > 0$ and $n$ is a known positive integer, has mean $n\mu$ and variance $n\mu(1 - \mu)[1 + \sigma(n - 1)/(1 + \sigma)]$ and hence provides a model for overdispersed binomial data.
Chapter 5

Mixed distributions

A mixed distribution is a special case of a finite mixture distribution. A mixed distribution is a mixture of two components: a continuous distribution and a discrete distribution, i.e. it is a continuous distribution where the range of $Y$ also includes discrete values with non-zero probabilities.

5.1 Zero adjusted distributions on zero and the positive real line $[0, \infty)$

Zero adjusted distributions on zero and the positive real line are a special case of mixed distributions. These distributions are appropriate when the response variable $Y$ takes values from zero to infinity including zero, i.e. $[0, \infty)$. They are a mixture of a discrete value 0 with probability $p$, and a continuous distribution on the positive real line $(0, \infty)$ with probability $(1 - p)$. The probability (density) function of $Y$ is $f_Y(y)$ given by

$$f_Y(y) = \begin{cases} p & \text{if } y = 0 \\ (1 - p)f_W(y) & \text{if } y > 0 \end{cases} \quad (5.1)$$

for $0 \leq y < \infty$, where $0 < p < 1$ and $f_W(y)$ is a probability density function defined on $(0, \infty)$, i.e. for $0 < y < \infty$.

Zero adjusted distributions on zero and the positive real line are appropriate when the probability of $Y = 0$ is non-zero and otherwise $Y > 0$. For example when $Y$ measures the amount of rainfall in a day (where some days have zero rainfall), or the river flow at a specific time each day (where some days the river flow is zero), or the total amount of insurance claims in a year for individuals (where some people do not claim at all and therefore their total claim is zero).

5.1.1 Zero adjusted gamma distribution, ZAGA($\mu, \sigma, \nu$)

The zero adjusted gamma distribution is a special case of a zero adjusted distribution on zero and the positive real line, i.e. $[0, \infty)$.

The zero adjusted gamma distribution is a mixture of a discrete value 0 with probability $\nu$, and a gamma $GA(\mu, \sigma)$ distribution on the positive real line $(0, \infty)$ with probability $(1 - \nu)$. 

85
CHAPTER 5. MIXED DISTRIBUTIONS

The probability (density) function of the zero adjusted gamma distribution, denoted by $\text{ZAGA} (\mu, \sigma, \nu)$, is given by

$$f_Y(y|\mu, \sigma, \nu) = \begin{cases} \nu & \text{if } y = 0 \\ (1 - \nu) f_W(y|\mu, \sigma) & \text{if } y > 0 \end{cases}$$

(5.2)

for $0 \leq y < \infty$, where $\mu > 0$ and $\sigma > 0$ and $0 < \nu < 1$, and $W \sim \text{GA}(\mu, \sigma)$ has a gamma distribution.

The zero inflated gamma distribution is shown in Figure 1.14.

The default link functions relating the parameters $(\mu, \sigma, \nu)$ to the predictors $(\eta_1, \eta_2, \eta_3)$, which may depend on explanatory variables, are

$$\log \mu = \eta_1$$

$$\log \sigma = \eta_2$$

$$\log \left( \frac{\nu}{1 - \nu} \right) = \eta_3.$$

Model (5.2) is equivalent to a gamma distribution $\text{GA}(\mu, \sigma)$ model for $Y > 0$, together with a binary model for recoded variable $Y_1$ given by

$$Y_1 = \begin{cases} 0 & \text{if } Y > 0 \\ 1 & \text{if } Y = 0 \end{cases}$$

(5.3)

i.e.

$$p(Y_1 = y_1) = \begin{cases} (1 - \nu) & \text{if } y_1 = 0 \\ \nu & \text{if } y_1 = 1 \end{cases}$$

(5.4)

The log likelihood function for the ZAGA model (5.2) is equal to the sum of the log likelihood functions of the gamma GA model and the binary BI model (5.4).

The ZAGA model can be fitted explicitly in GAMLSS.

Alternatively the gamma GA model can be fitted after deleting all cases with $y = 0$, and the binary BI model can be fitted to the recoded variable $Y_1$. This method has the advantage that any GAMLSS distribution on $Y > 0$ can replace the gamma distribution and be fitted to $Y > 0$.

5.1.2 Fitting zero adjusted distributions on zero and the positive real line

The zero adjusted gamma (ZAGA) distribution and the zero adjusted inverse Gaussian (ZAIG) can be fitted explicitly in GAMLSS. However other zero adjusted distributions on zero and the positive real line (5.1) cannot currently be fitted explicitly in GAMLSS. However any GAMLSS distribution on $Y > 0$ which is zero adjusted can be fitted by fitting two models as described at the end of the previous subsection.
5.1.3 Example of fitting a response variable on zero and the positive real line

5.2 Distributions on the unit interval (0,1) inflated at 0 and 1

Distributions on the unit interval (0,1) inflated at 0 and 1 are a special case of mixed distributions. These distributions are appropriate when the response variable \( Y \) takes values from 0 to 1 including 0 and 1, i.e. range \([0,1]\). They are a mixture of three components: a discrete value 0 with probability \( p_0 \), a discrete value 1 with probability \( p_1 \), and a continuous distribution on the unit interval (0,1) with probability \((1 - p_0 - p_1)\). The probability (density) function of \( Y \) is \( f_Y(y) \) given by

\[
f_Y(y) = \begin{cases} 
  p_0 & \text{if } y = 0 \\
  (1 - p_0 - p_1) f_W(y) & \text{if } 0 < y < 1 \\
  p_1 & \text{if } y = 1
\end{cases}
\]

for \( 0 \leq y \leq 1 \), where \( 0 < p_0 < 1 \), \( 0 < p_1 < 1 \) and \( 0 < p_0 + p_1 < 1 \) and \( f_W(y) \) is a probability density function defined on \((0,1)\), i.e. for \( 0 < y < 1 \).

Distributions on the unit interval (0,1) inflated at 0 and 1 are appropriate when the probabilities of \( Y = 0 \) and \( Y = 1 \) are non-zero and otherwise \( 0 < Y < 1 \). For example when \( Y \) measures a proportion, e.g. the proportion of loss given default (LGD), or a score between 0 and 1 which can include 0 and 1, e.g. level of pain score.

5.2.1 Beta inflated distribution, \( \text{BEINF}(\mu,\sigma,\nu,\tau) \)

The beta inflated distribution is a special case of a distribution on the unit interval (0,1) inflated at 0 and 1.

The beta inflated distribution is a mixture of three components: a discrete value 0 with probability \( p_0 \), a discrete value 1 with probability \( p_1 \), and a beta \( \text{BE}(\mu,\sigma) \) distribution on the unit interval (0,1) with probability \((1 - p_0 - p_1)\).

The probability (density) function of the beta inflated distribution, denoted by \( \text{BEINF}(\mu,\sigma,\nu,\tau) \), is defined by

\[
f_Y(y|\mu,\sigma,\nu,\tau) = \begin{cases} 
  p_0 & \text{if } y = 0 \\
  (1 - p_0 - p_1) f_W(y) & \text{if } 0 < y < 1 \\
  p_1 & \text{if } y = 1
\end{cases}
\]

for \( 0 \leq y \leq 1 \), where \( W \sim \text{BE}(\mu,\sigma) \) has a beta distribution with \( 0 < \mu < 1 \) and \( 0 < \sigma < 1 \) and \( p_0 = \nu/(1+\nu+\tau) \) and \( p_1 = \tau/(1+\nu+\tau) \). Hence \( \nu = p_0/p_2 \) and \( \tau = p_1/p_2 \) where \( p_2 = 1 - p_0 - p_1 \). Since \( 0 < p_0 < 1 \), \( 0 < p_1 < 1 \) and \( 0 < p_0 + p_1 < 1 \), hence \( \nu > 0 \) and \( \tau > 0 \).

The beta inflated distribution is shown in Figure ??.

The default link functions relating the parameters \((\mu,\sigma,\nu,\tau)\) to the predictors \((\eta_1,\eta_2,\eta_3,\eta_4)\), which may depend on explanatory variables, are

\[
\log \left( \frac{\mu}{1 - \mu} \right) = \eta_1
\]
\[
\log \left( \frac{\sigma}{1-\sigma} \right) = \eta_2
\]

\[
\log \nu = \log \left( \frac{p_0}{p_2} \right) = \eta_3
\]

\[
\log \tau = \log \left( \frac{p_1}{p_2} \right) = \eta_4
\]

Model (5.6) is equivalent to a beta distribution \(BE(\mu, \sigma)\) model for \(0 < Y < 1\), together with a multinomial model \(MN3(\nu, \tau)\) with three levels for recoded variable \(Y_1\) given by

\[
Y_1 = \begin{cases} 
0 & \text{if } Y = 0 \\
1 & \text{if } Y = 1 \\
2 & \text{if } 0 < Y < 1
\end{cases}
\]

i.e.

\[
p(Y_1 = y_1) = \begin{cases} 
p_0 & \text{if } y_1 = 0 \\
p_1 & \text{if } y_1 = 1 \\
1 - p_0 - p_1 & \text{if } y_1 = 2
\end{cases}
\]

The log likelihood function for the BEINF model (5.6) is equal to the sum of the log likelihood functions of the beta BE model and the multinomial MN3 model (5.8).

The BEINF model can be fitted explicitly in GAMLSS. Alternatively the beta BE model can be fitted after deleting all cases with \(y = 0\) and \(y = 1\), and the multinomial MN3 model can be fitted to the recoded variable \(Y_1\). This method has the advantage that any GAMLSS distribution on \(0 < Y < 1\) can replace the beta distribution and be fitted to \(0 < Y < 1\). Note that any GAMLSS distribution on \(-\infty < Y < \infty\) can be transformed (using the inverse logit transform) to a GAMLSS distribution on \(0 < Y < 1\).

### 5.2.2 Fitting distributions on the unit interval \((0,1)\) inflated at 0 and 1

The BEINF model can be fitted explicitly in GAMLSS. However other distributions on the unit interval \((0,1)\) inflated at 0 and 1 (5.5) cannot currently be fitted explicitly in GAMLSS. However any GAMLSS distribution on \(0 < Y < 1\) which is inflated at 0 and 1 can be fitted by two models as described at the end of the previous subsection.

### 5.2.3 Beta inflated at 0 distribution, \(BEINF0(\mu, \sigma, \nu)\)

The beta inflated at 0 distribution is a mixture of two components: a discrete value 0 with probability \(p_0\), and a beta \(BE(\mu, \sigma)\) distribution on the unit interval \((0,1)\) with probability \((1 - p_0)\).

The probability (density) function of the beta inflated at 0 distribution, denoted by \(BEINF0(\mu, \sigma, \nu)\), is given by

\[
f_Y(y|\mu, \sigma, \nu) = \begin{cases} 
p_0 & \text{if } y = 0 \\
(1 - p_0) f_W(y|\mu, \sigma) & \text{if } 0 < y < 1
\end{cases}
\]
for $0 \leq y < 1$, where $W \sim BE(\mu, \sigma)$ has a beta distribution with $0 < \mu < 1$ and $0 < \sigma < 1$ and $p_0 = \nu/(1 + \nu)$. Hence $\nu = p_0/(1 - p_0)$. Since $0 < p_0 < 1$, hence $\nu > 0$.

The default link functions relating the parameters $(\mu, \sigma, \nu)$ to the predictors $(\eta_1, \eta_2, \eta_3)$, which may depend on explanatory variables, are

$$\log \left( \frac{\mu}{1 - \mu} \right) = \eta_1$$
$$\log \left( \frac{\sigma}{1 - \sigma} \right) = \eta_2$$
$$\log \nu = \log \left( \frac{p_0}{1 - p_0} \right) = \eta_3$$

Model (5.9) is equivalent to a beta distribution $BE(\mu, \sigma)$ model for $0 < Y < 1$, together with a binary model for recoded variable $Y_1$ given by

$$Y_1 = \begin{cases} 
0 & \text{if } 0 < Y < 1 \\
1 & \text{if } Y = 0
\end{cases}$$

(5.10)

i.e.

$$p(Y_1 = y_1) = \begin{cases} 
(1 - \nu) & \text{if } y_1 = 0 \\
\nu & \text{if } y_1 = 1
\end{cases}$$

(5.11)

The log likelihood function for the BEINF0 model (5.9) is equal to the sum of the log likelihood functions of the beta BE model and the binary BI model (5.14).

The BEINF0 model can be fitted explicitly in GAMLSS.

Alternatively the BE model can be fitted after deleting all cases with $y = 0$, and the binary BI model can be fitted to the recoded variable $Y_1$. This method has the advantage that any GAMLSS distribution on $0 < Y < 1$ can replace the beta distribution and be fitted to $0 < Y < 1$.

5.2.4 Beta inflated at 1 distribution, $BEINF1(\mu, \sigma, \nu)$

The beta inflated at 0 distribution is a mixture of two components: a discrete value 1 with probability $p_1$, and a beta $BE(\mu, \sigma)$ distribution on the unit interval $(0, 1)$ with probability $(1 - p_1)$.

The probability (density) function of the beta inflated at 1 distribution, denoted by $BEINF1(\mu, \sigma, \nu)$, is given by

$$f_Y(y|\mu, \sigma, \nu) = \begin{cases} 
p_1 & \text{if } y = 1 \\
(1 - p_1)f_W(y|\mu, \sigma) & \text{if } 0 < y < 1
\end{cases}$$

(5.12)

for $0 < y \leq 1$, where $W \sim BE(\mu, \sigma)$ has a beta distribution with $0 < \mu < 1$ and $0 < \sigma < 1$ and $p_1 = \nu/(1 + \nu)$. Hence $\nu = p_1/(1 - p_1)$. Since $0 < p_1 < 1$, hence $\nu > 0$.

The default link functions relating the parameters $(\mu, \sigma, \nu)$ to the predictors $(\eta_1, \eta_2, \eta_3)$, which may depend on explanatory variables, are

$$\log \left( \frac{\mu}{1 - \mu} \right) = \eta_1$$
\[ \log \left( \frac{\sigma}{1-\sigma} \right) = \eta_2 \]

\[ \log \nu = \log \left( \frac{p_1}{1-p_1} \right) = \eta_3 \]

Model (5.12) is equivalent to a beta distribution \( BE(\mu, \sigma) \) model for \( 0 < Y < 1 \), together with a binary model for recoded variable \( Y_1 \) given by

\[
Y_1 = \begin{cases} 
0 & \text{if } 0 < Y < 1 \\
1 & \text{if } Y = 1 
\end{cases}
\]

i.e.

\[ p(Y_1 = y_1) = \begin{cases} 
(1 - \nu) & \text{if } y_1 = 0 \\
\nu & \text{if } y_1 = 1 
\end{cases} \]

The log likelihood function for the BEINF1 model (5.12) is equal to the sum of the log likelihood functions of the gamma BE model and the binary BI model (5.14).

The BEINF1 model can be fitted explicitly in GAMLSS.

Alternatively the BE model can be fitted after deleting all cases with \( y = 1 \), and the binary BI model can be fitted to the recoded variable \( Y_1 \). This method has the advantage that any GAMLSS distribution on \( 0 < Y < 1 \) can replace the beta distribution and be fitted to \( 0 < Y < 1 \).

### 5.2.5 Example of fitting a response variable on the interval \((0,1]\), the unit interval including value 1

**Lung function data**

Lung function data was recorded on 7209 Caucasian subjects aged between 3 and 80 years, Stanojovic, Wade, Cole et al. (2009). There were 3164 males with no missing values. Here the ratio of forced expiratory volume in one second (FEV1) to forced vital capacity (FVC), i.e. \( Y = \text{FEV1}/\text{FVC} \), is modelled. This ratio is the established index for diagnosing airway obstruction, Quanjer, Stanojovic, Stocks et al. (2010). The range of the ratio \( Y \) is \((0,1]\) including the value 1 (but not value 0). An appropriate distribution for \( Y \) is a distribution on \((0,1]\), the unit interval including value 1. The explanatory variable used here is height incm. A log transform of height was applied to give variable lht.

A BEINF1 distribution was initially fitted to the data, but did not provide an adequate model, in particular the beta model for \( 0 < Y < 1 \) was inadequate.

Consequently a logitSST distribution inflated at 1 was used. The probability (density) function of \( Y \) is \( f_Y(y) \) given by

\[ f_Y(y|\mu, \sigma, \nu, \tau, p) = \begin{cases} 
p & \text{if } y = 1 \\
(1-p)f_W(y|\mu, \sigma, \nu, \tau) & \text{if } 0 < y < 1 
\end{cases} \]

for \( 0 < y \leq 1 \), where \( W \sim \text{logitSST}(\mu, \sigma, \nu, \tau) \) has a logitSST distribution with \(-\infty < \mu < \infty\) and \( \sigma > 0, \nu > 0, \tau > 0 \) and \( 0 < p < 1 \).

The default link functions relate the parameters \((\mu, \sigma, \nu, \tau, p)\) to the predictors \((\eta_1, \eta_2, \eta_3, \eta_4, \eta_5)\), which are modelled as smooth functions of \( \text{lht} = \log(\text{height}) \), i.e.
\[ \mu = \eta_1 = s(lht) \]
\[ \log \sigma = \eta_2 = s(lht) \]
\[ \log \nu = \eta_3 = s(lht) \]
\[ \log \tau = \eta_4 = s(lht) \]
\[ \log \left( \frac{p}{1-p} \right) = \eta_5 = s(lht) \]

Model (5.15) was fitted by fitting two models: a logitSST distribution \( BE(\mu, \sigma) \) model for \( 0 < Y < 1 \), together with a binary model for recoded variable \( Y_1 \) given by (5.13).

The R code for fitting the two models is given below.
Chapter 6

Finite mixtures

This Chapter covers finite mixtures within GAMLSS. In general finite mixture distributions are fitted within GAMLSS using the EM algorithm. Certain specific mixtures distributions are explicitly available in `gamlss` packages. The zero inflated Poisson (ZIP and ZIP2) the zero adjusted (inflated) inverse Gaussian (ZAIG), and the four parameter beta inflated at zero and one (BEINF).

6.1 Introduction to finite mixtures

Suppose that the random variable $Y$ comes from component $k$, having probability (density) function $f_k(y)$, with probability $\pi_k$ for $k = 1, 2, \ldots, K$, then the (marginal) density of $Y$ is given by

$$f_Y(y) = \sum_{k=1}^{K} \pi_k f_k(y) \quad (6.1)$$

where $0 \leq \pi_k \leq 1$ is the prior (or mixing) probability of component $k$, for $k = 1, 2, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$.

More generally the probability (density) function $f_k(y)$ for component $k$ may depend on parameters $\theta_k$ and explanatory variables $x_k$, i.e. $f_k(y) = f_k(y|\theta_k, x_k)$.

Similarly $f_Y(y)$ depends on parameters $\psi = (\theta, \pi)$ where $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$ and $\pi^T = (\pi_1, \pi_2, \ldots, \pi_K)$ and explanatory variables $x = (x_1, x_2, \ldots, x_K)$, i.e. $f_Y(y) = f_Y(y|\psi, x)$, and

$$f_Y(y|\psi, x) = \sum_{k=1}^{K} \psi_k f_k(y|\theta_k, x_k) \quad (6.2)$$

Subsequently we omit the conditioning on $\theta_k, x_k$ and $\psi$ to simplify the presentation. In Sections 6.2, 6.3 and 6.7 we consider respectively maximum likelihood estimation, the corresponding fitting function `gamlssMX` and examples for finite mixtures models with no parameters in common, while in Sections 6.5, 6.6 and 6.7 we consider respectively maximum likelihood estimation, the corresponding fitting function `gamlssNP` and examples for finite mixture models with parameters in common. Throughout this chapter we will assume that all $K$ components of the mixture can be represented by GAMLSS models.
6.2 Finite mixtures with no parameters in common

Here the parameter sets \( (\theta_1, \theta_2, \ldots, \theta_k) \) are distinct, i.e. no parameter is common to two or more parameters sets. Note that what this means in practice within GAMLSS is that the conditional distribution components in (6.1), \( f_k(y) \), can have different \texttt{gamlss.family} distributions, e.g. one can be \texttt{GA} and the other \texttt{IG}.

6.2.1 The likelihood function

Given \( n \) independent observations \( y_i \) for \( i = 1, 2, \ldots, n \), from finite mixture model (6.2), the likelihood function is given by

\[
L = L(\psi, y) = \prod_{i=1}^{n} f_Y(y_i) = \prod_{i=1}^{n} \left[ \sum_{k=1}^{K} \pi_k f_k(y_i) \right]
\]

where \( y = (y_1, y_2, \ldots, y_n) \), \( f_k(y_i) = f_k(y_i|\theta_k, x_{ki}) \), with log likelihood function given by

\[
\ell = \ell(\psi, y) = \sum_{i=1}^{n} \log \left[ \sum_{k=1}^{K} \pi_k f_k(y_i) \right]
\]

We wish to maximize \( \ell \) with respect to \( \psi \), i.e. with respect to \( \theta \) and \( \pi \). The problem is that the log function between the two summations in (6.4) makes it difficult. One solution, especially for simple mixtures where no explanatory variables are involved, is to use a numerical maximization technique, e.g. function \texttt{optim} in R, to maximize the log likelihood in (6.4) numerically, see for example Venables and Ripley (2002) Chapter 16.

6.2.2 Maximizing the likelihood function using the EM algorithm

Here we will use the EM algorithm, (Dempser, Laird and Rubin, 1977) to maximize (6.4) with respect to \( \psi \), treating all the component indicator variables (i.e \( \delta \), defined below) as missing variables.

Let

\[
\delta_{ik} = \begin{cases} 
1, & \text{if observation } i \text{ comes from component } k \\
0, & \text{otherwise} 
\end{cases}
\]

for \( k = 1, 2, \ldots, K \) and \( i = 1, 2, \ldots, n \). Let \( \delta^T = (\delta_{i1}, \delta_{i2}, \ldots, \delta_{ik}) \) be the indicator vector for observation \( i \). If observation \( i \) comes from component \( k \) then \( \delta_i \) is a vector of zeros, except for the \( k^{th} \) value which is \( \delta_{ik} = 1 \). Let \( \delta^T = (\delta_1^T, \delta_2^T, \ldots, \delta_n^T) \) combine all the indicator variable vectors. Then the complete data, i.e. observed \( y \) and unobserved \( \delta \), has complete likelihood function given by

\[
L_c = L_c(\psi, y, \delta) = f(y, \delta) = \prod_{i=1}^{n} f(Y_i, \delta_i) = \prod_{i=1}^{n} \left[ \prod_{k=1}^{K} f_k(y_i)^{\delta_{ik}} \pi_k^{\delta_{ik}} \right],
\]

(6.6)
6.2. FINITE MIXTURES WITH NO PARAMETERS IN COMMON

since if δ_{ik} = 1 and δ_{ik'} = 0 for k' ≠ k, then

\[ f(y_i|\delta_i)f(\delta_i) = f_k(y_i)\pi_k \]

\[ = f_k(y_i)^{\delta_{ik}}\pi_k^{\delta_{ik}} \]

\[ = \prod_{k=1}^{K} f_k(y_i)^{\delta_{ik}}\pi_k^{\delta_{ik}} \]

and hence \( f(y_i|\delta_i)f(\delta_i) = \prod_{k=1}^{K} f_k(y_i)^{\delta_{ik}}\pi_k^{\delta_{ik}} \) for all \( \delta_i \).

From (6.6) the complete log likelihood is given by

\[ \ell_c = \ell_c(\psi, y, \delta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \log f_k(y_i) + \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \log \pi_k \] (6.7)

If \( \delta \) were known then, since \( \theta_1, \theta_2, \ldots, \theta_K \) have no parameter in common, \( \ell_c \) could be maximized over each \( \theta_k \) separately, since the likelihood separates.

The EM algorithm alternates between the E-step and the M-step until convergence. Iteration \((r+1)\) of the EM algorithm comprises an E-step followed by an M-step.

E-step

At the \((r+1)^{th}\) iteration, the E-step finds the conditional expectation of the complete data log likelihood (6.7), over the missing \( \delta \), given \( y \) and the current parameter estimates \( \hat{\psi}^{(r)} \) from iteration \( r \), i.e.

\[ Q = E_{\delta} \left[ \ell_c | y, \hat{\psi}^{(r)} \right] \]

\[ = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log f_k(y_i) + \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log \pi_k \] (6.8)

where

\[ \hat{w}_{ik}^{(r+1)} = E \left[ \delta_{ik} | y, \hat{\psi}^{(r)} \right] \]

\[ = \left[ 1 * p(\delta_{ik} = 1|y, \hat{\psi}^{(r)}) \right] + \left[ 0 * p(\delta_{ik} = 1|y, \hat{\psi}^{(r)}) \right] \]

\[ = p(\delta_{ik} = 1|y, \hat{\psi}^{(r)}) \]

\[ = p(\delta_{ik} = 1, y_i|\hat{\psi}^{(r)}) \]

\[ = \frac{f(y_i|\hat{\psi}^{(r)})}{f(y_i|\hat{\psi}^{(r)})} \]

\[ = \frac{\hat{\pi}_k^{(r)} f_k(y_i|\theta_k^{(r)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(r)} f_k(y_i|\theta_k^{(r)})} \] (6.9)

Note that \( \hat{w}_{ik}^{(r+1)} = p(\delta_{ik} = 1|y_i, \hat{\psi}^{(r)}) \) is the posterior probability that observation \( y_i \) comes from component \( k \), given \( y_i \) and given \( \psi = \hat{\psi}^{(r)} \), while \( \hat{\pi}_k^{(r)} = p(\delta_{ik} = 1|\hat{\psi}^{(r)}) \) is the prior (or
mixing) probability that observation \( y_i \) comes from component \( k \), given \( \psi = \hat{\psi}^{(r)} \) only. On convergence, i.e. \( r = \infty \), \( \hat{\psi}^{(\infty)} \) and \( \hat{\pi}^{(\infty)} \) are the estimated posterior and prior probabilities that the observation \( y_i \) comes from component \( k \), respectively, since \( \psi \) is estimated by \( \hat{\psi}^{(\infty)} \).

**M-step**

At the \((r + 1)^{th}\) iteration, the M step maximizes \( Q \) with respect to \( \psi \). Since the parameters \( \theta_k \) in \( f_k(y_i) \) for \( k = 1, 2, \ldots, K \) are distinct, (i.e. there are no parameters in common to two or more \( \theta_k \)'s), \( Q \) can be maximized with respect to each \( \theta_k \) by maximizing separately the \( k^{th} \) part of the first term in (6.8), i.e. maximize \( \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log f_k(y_i) \) with respect to \( \theta_k \), for \( k = 1, 2, \ldots, K \). Assuming, for \( k = 1, 2, \ldots, K \), that the \( k \) component follows a GAMLSS model, this is just a weighted log likelihood for a GAMLSS model with weights \( \hat{w}_{ik}^{(r+1)} \) for \( i = 1, 2, \ldots, n \). Also the parameter \( \pi \) only occurs in the second term in (6.8) and so can be estimated by maximizing the second term, subject to \( \sum_{k=1}^{K} \pi_k = 1 \), i.e. where \( \pi_k = 1 - \sum_{k=1}^{K-1} \pi_k \), leading to \( \hat{\pi}_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \) for \( k = 1, 2, \ldots, K \).

**Summary of the \((r + 1)^{th}\) iteration of the EM algorithm**

**E-step** Replace \( \delta_{ik} \) in (6.7) by \( \hat{w}_{ik}^{(r+1)} \) [its conditional expectation given \( y \) and given the current estimate \( \hat{\psi}^{(r)} \) from iteration \( r \)] obtained from (6.9) for \( k = 1, 2, \ldots, K \) and \( i = 1, 2, \ldots, n \) to give (6.8).

**M-step**

1. for each \( k = 1, 2, \ldots, K \), obtain \( \hat{\theta}_k^{(r+1)} \) by fitting the GAMLSS model for the \( k^{th} \) component to dependent variable \( y \) with explanatory variables \( x_k \) using prior weights \( \hat{w}_k^{(r+1)} \), where \( w_k = (w_{1k}, w_{2k}, \ldots, w_{nk}) \).
2. \( \hat{\pi}_k^{(r+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \) for \( k = 1, 2, \ldots, K \),
3. \( \hat{\psi}^{(r+1)} = \left[ \hat{\theta}^{(r+1)}, \hat{\pi}^{(r+1)} \right] \) where \( \theta = (\theta_1, \theta_2, \ldots, \theta_K) \).

**6.2.3 Modelling the mixing probabilities**

Here we extend the finite mixture model by assuming that the mixing probabilities \( \pi_k \) for \( k = 1, 2, \ldots, K \) for observations \( i = 1, 2, \ldots, n \) are not fixed constants but depend on explanatory variables \( x_0 \) and parameters \( \alpha_k \), and hence depend on \( i \), so \( f_{Y_i}(y_i) = \sum_{k=1}^{K} \pi_{ik} f_k(y_i) \). We model the mixing probabilities \( \pi_{ik} \) using a multinomial logistic model where \( \delta_i \) is a single draw from a multinomial distribution with probability vector \( \pi \), i.e. \( \delta_i \sim M(1, \pi) \) and

\[
\log \left( \frac{\pi_{ik}}{\pi_{iK}} \right) = \alpha_k^T x_{0i} \tag{6.10}
\]

for \( k = 1, 2, \ldots, K \) and \( i = 1, 2, \ldots, n \). Hence

\[
\pi_{ik} = \frac{\exp \left\{ \alpha_k^T x_{0i} \right\}}{\sum_{k=1}^{K} \exp \left\{ \alpha_k^T x_{0i} \right\}} \tag{6.11}
\]
for \( k = 1, 2, \ldots, K \) and \( i = 1, 2, \ldots, n \) where \( \alpha_K = 0 \). Consequently the complete log likelihood is given by replacing \( \pi_k \) by \( \pi_{ik} \) in equation (6.7) to give

\[
\ell_c = \ell_c(\psi, y, \delta) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \log f_k(y_i) + \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \log \pi_{ik}
\]

(6.12)

This results in replacing \( \pi_k \) with \( \pi_{ik} \) in equations (6.8) and (6.9) of the EM algorithm, i.e.

\[
Q = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log f_k(y_i) + \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log \pi_{ik}
\]

(6.13)

where

\[
\hat{w}_{ik}^{(r+1)} = \frac{\hat{w}_{ik}^{(r)} f_k(y_i | \hat{\theta}_k^{(r)})}{\sum_{k=1}^{K} \hat{w}_{ik}^{(r)} f_k(y_i | \hat{\theta}_k^{(r)})}
\]

(6.14)

**Summary of the \((r + 1)^{th}\) iteration of the EM algorithm**

**E-step** Replace \( \delta_{ik} \) in (6.12) by \( \hat{w}_{ik}^{(r+1)} \), obtained from (6.14), for \( k = 1, 2, \ldots, K \) and \( i = 1, 2, \ldots, n \), to give (6.13).

**M-step**

1. For each \( k = 1, 2, \ldots, K \), obtain \( \hat{\theta}_k^{(r+1)} \), by fitting the GAMLSS model for the \( k^{th} \) component to response variable \( y \) with explanatory variables \( x_k \) using weights \( \hat{w}_k^{(r+1)} \), where \( \hat{w}_k^{(r+1)} = (w_{1k}, w_{2k}, \ldots, w_{nk}) \).
2. Obtain \( \hat{\alpha}^{(r+1)} \) by fitting multinomial logistic model (6.10) to pseudo multinomial response variable \( y_p \) with expanded explanatory variables \( x_{0e} \) using prior weights \( \hat{\omega}^{(r+1)} \),
3. \( \hat{\psi}^{(r+1)} = \left[ (\hat{\theta}^{(r+1)}, \hat{\alpha}^{(r+1)}) \right] \) where \( \theta = (\theta_1, \theta_2, \ldots, \theta_K) \).

Note that M-step (2) is achieved by expanding the data set \( K \) times as shown in Table 6.1. That is, by setting up the pseudo multinomial response variable \( y_p^T \), taking data values \( y_{pik}^T \), where subscript \( p \) stands for pseudo and where \( y_{pik}^T = (0, 0, \ldots, 0, 1, \ldots, 0) \) is a vector of zeros except for one in the \( k^{th} \) cell, for \( k = 1, 2, \ldots, K \) and \( i = 1, 2, \ldots, n \), prior weight variable \( \hat{w}_{ik}^{(r+1)} \) [where \( \hat{w}_k^{T} = (\hat{w}_{1k}^{T}, \hat{w}_{2k}^{T}, \ldots, \hat{w}_{nk}^{T}) \) and \( \hat{w}_k^{T} = (\hat{w}_1^{T}, \hat{w}_2^{T}, \ldots, \hat{w}_n^{T}) \) for \( k = 1, 2, \ldots, K \)] and fitting a multinomial model to \( y_p \) based on expanded explanatory variable \( x_{0e} \) using weights \( \hat{w}_k^{(r+1)} \).

**6.2.4 Zero components**

Special cases of the models described above are distributions which we described earlier as type mixed. For example, the zero adjusted inverse Gaussian distribution (ZAIG) described in Appendix 15.3.5 can be thought of as a finite mixture where the first component is identically zero, i.e. \( y = 0 \), with probability 1. Hence

\[
f_1(y) = \begin{cases} 1, & \text{if } y = 0 \\ 0, & \text{otherwise.} \end{cases}
\]

(6.15)
Distributions of this type can be also fitted with the EM algorithm described in the previous section. The EM algorithm only changes in M-step (1) where the fitting of the first component is omitted (since it has no parameters). The rest of the algorithm is unchanged.

### 6.3 The `gamlssMX()` function

The function to fit finite mixtures with no parameters in common is `gamlssMX()`. In this section we describe how it works. Examples of using the function are given in the next section. The function `gamlssMX()` has the following arguments:

**formula**  This argument should be a single formula (or a list of formulae of length K the number of components in the mixture) for modelling the predictor for the \(\mu\) parameter of the model. If a single formula is used then the \(K\) mixture components have the same predictor for \(\mu\), but different parameters in their predictors (since there are no parameters in common to two or more of the \(K\) components). Note that modelling the rest of the distributional parameters can be done by using the usual `gamlss()` formula arguments, e.g. `sigma.fo∼x`, which passes the arguments to `gamlss()`. Again either a single common formula or a list of formula of length \(K\) is used.

**pi.formula**  This should be a formula for modelling the predictor for prior (or mixing) probabilities as a function of explanatory variables in the multinomial model (6.10). The default model is constants for the prior (or mixing) probabilities. Note that no smoothing or other additive terms are allowed here, only the usual linear terms. The modelling here is done using the `multinom()` function from package `nnet`.

**family**  This should be a `gamlss.family` distribution (or a list of \(K\) distributions). Note that if different distributions are used here, it is preferable (but not essential) that their parameters are comparable for ease of interpretation.

<table>
<thead>
<tr>
<th>i</th>
<th>k</th>
<th>(y_i)</th>
<th>(x_{0i})</th>
<th>(x_{1i})</th>
<th>(x_{Ki})</th>
<th>multinomial response (y_p)</th>
<th>weights (\hat{w}^{(r+1)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>(y)</td>
<td>(X_0)</td>
<td>(X_1)</td>
<td>(X_K)</td>
<td>1 0 0 ... 0</td>
<td>(\hat{w}_1^{(r+1)})</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>(y)</td>
<td>(X_0)</td>
<td>(X_1)</td>
<td>(X_K)</td>
<td>0 1 0 ... 0</td>
<td>(\hat{w}_2^{(r+1)})</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(K)</td>
<td>(y)</td>
<td>(X_0)</td>
<td>(X_1)</td>
<td>(X_K)</td>
<td>0 0 0 ... 1</td>
<td>(\hat{w}_K^{(r+1)})</td>
</tr>
<tr>
<td>2</td>
<td>(K)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>(K)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Table showing the expansion of data for fitting the multinomial model at M step (2) of the EM algorithm
weights   For declaring prior weights if needed.

K      For declaring the number of components in the finite mixture with default K=2

prob   For setting starting values for the prior probabilities.

data   The data frame containing the variables in the fit. Note that this is compulsory if
pi.formula is used for modelling the prior (or mixing) probabilities.

control This argument sets the control parameters for the EM iterations algorithm. The
default setting are given in the MX.control function

g.control This argument can be used to pass to gamlss() control parameters, as in gamlss.control.

zero.component This argument declares whether or not there is a zero component, i.e. y
identically equal to zero, y = 0, in the finite mixture.

... For extra arguments to be passed to gamlss().

6.4 Examples using the gamlssMX() function

6.5 Finite mixtures with parameters in common

Here the K components of the mixture may have parameters in common, i.e. the parameter sets
(\theta_1, \theta_2, \ldots, \theta_k) are not disjoint. The prior (or mixing) probabilities are either assumed to be
constant (as in function gamlssMP()) or may depend on explanatory variables x_0 and parameters
\alpha through a multinomial logistic model as in Section 6.2.3. We assume that the K components
f_k(y) = f_k(y; \theta_k, x_k) for k = 1, 2, \ldots, K can be represented by GAMLSS models. Note that since
some of the parameters may be common to the K components, the distribution used must be the
same for all K components. Similarly the link functions of the distribution parameters must be the
same for all K components. GAMLSS models have up to four distributional parameters \mu, \sigma, \nu and \tau. In our notation in this Chapter, the parameter vector \theta_k contains all the parameters
in the (linear) predictor models for \mu, \sigma, \nu and \tau for component k, for k = 1, 2, \ldots, K. Here
are some examples to clarify this.

Example 1, Mixture of K Poisson regression models: \( f(y) = \sum_{k=1}^{K} \pi_k f(y) \) where \( f_k(y) \)
is PO(\mu_k) for k = 1, 2, \ldots, K, and where log \( \mu_k = \beta_{ok} + \beta_{1} x \). Here the slope parameter \( \beta_{1} \)
a predictor parameter for the distribution parameter \( \mu_k \), is the same for all K components,
but the intercept parameter \( \beta_{ok} \) depends on k, for k = 1, 2, \ldots, K.

Example 2, Mixture of K negative binomials regression models: Let \( f_k(y) \) be NBI(\mu_k, \sigma_k)
for k = 1, 2, \ldots, K, where log \( \mu_k = \beta_{10k} + \beta_{11} x \) and log \( \sigma_k = \log \sigma = \beta_{20} + \beta_{21} x \). Here the
predictor slope parameter \( \beta_{11} \) for \( \mu_k \) and all predictor parameters for \( \sigma \) are the same for
all K components, but the predictor intercept parameter \( \beta_{10k} \) for \( \mu_k \) depends on k, for
k = 1, 2, \ldots, K.

Example 3, Mixture of K BCT models: Let \( f_k(y) = BCT(\mu_k, \sigma_k, \nu_k, \tau_k) \) for k = 1, 2, \ldots, K,
where log \( \mu_k = \beta_{10k} + \beta_{11k} x \), log \( \sigma_k = \beta_{20k} + \beta_{21k} x \), \( \nu_k = \nu = \beta_{30} \) and log \( \tau = \log \tau = \beta_{40} \).
Here predictor parameters \( \beta_{10k} \) and \( \beta_{11k} \) for \( \mu \) and \( \beta_{20k} \) and \( \beta_{21k} \) for \( \sigma \) depend on k for
k = 1, 2, \ldots, K, but parameters \( \beta_{30} \) for \( \nu \) and \( \beta_{40} \) for \( \tau \) are the same for all k components.
6.5.1 Maximizing the likelihood using the EM algorithm

As in Section 6.2.3 the complete log likelihood is given by (6.12). The following is a summary of the EM algorithm suitable for dealing with GAMLSS models with common parameters in the mixture.

Summary of the \((r + 1)^{th}\) iteration of the EM algorithm

**E-step** Replace \( \delta_{ik} \) in (6.12) by \( \hat{w}_{ik}^{(r+1)} \), obtained from (6.14) for \( k = 1, 2, \ldots, K \) and \( i = 1, 2, \ldots, n \) to give (6.13), i.e.

\[
Q = \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log f_k(y_i) + \sum_{k=1}^{K} \sum_{i=1}^{n} \hat{w}_{ik}^{(r+1)} \log \pi_{ik}.
\]

**M-step**

1. Since components \( f_k(y) \) for \( k = 1, 2, \ldots, K \) have parameters in common, \( Q \) cannot be maximized separately with respect to each \( \theta_k \). Obtain \( \hat{\theta}^{(r+1)} \) by fitting a single GAMLSS model to an expanded response variable \( y_e \), with expanded explanatory variable design matrix \( X_e \), using weights \( \hat{w}^{(r+1)} \) (see Table 6.2).

2. Obtain \( \hat{\alpha}^{(r+1)} \) by fitting a multinomial logistic model as in Section 6.2.3.

3. \( \hat{\psi}^{(r+1)} = [\hat{\theta}^{(r+1)}, \hat{\alpha}^{(r+1)}] \)

Note that the M step (1) is achieved by expanding the data set \( K \) times as in Table 6.2. This method is identical to the method used in Aitkin et al. (2006) but here we are not restricting ourselves to the exponential family. The column headed as MASS identifies the \( K \) mixture components. This column is declared as a factor in the R implementation of the EM algorithm. If this factor MASS is included in the predictor for a distribution parameter \( \mu, \sigma, \nu, \) or \( \tau \), then the predictor intercepts differs between the \( K \) components. If an interaction between this factor MASS and an explanatory variable \( x \) is included in the predictor model for a distribution parameter, then the coefficient of \( x \) differ between the \( K \) components. Note however that the syntax used in `gamlssNP()` for the interaction between MASS and \( x \) in the predictor for \( \mu \) is achieved using the `random=\sim x` argument (see Section 6.7 for an example).

6.6 The `gamlssNP()` function

The function to fit finite mixtures with parameters in common is `gamlssMX`. In this section we describe how it works. Examples of using the function are given in the next section. The function `gamlssMX` has the following arguments:

- **formula** This argument should be a formula defining the response variable and explanatory fixed effects terms for the \( \mu \) parameter of the model. Note that modelling the rest of the distribution parameters can be done by using the usual formulae, e.g. `sigma.fo= x`, which passes the arguments to `gamlss()`

- **random** This should be a formula defining the random part of the model (for random effect models). This formula is also used for fixed effect mixture models to define interactions of the factor MASS with explanatory variables \( x \) in the predictor for \( \mu \) (needed to request different coefficients in \( x \) in the predictor of \( \mu \) for the \( K \) components).

- **family** A gamlss family distribution.
### 6.7 Examples using the gamlssNP() function

#### 6.8 Bibliographic notes

There is an extensive literature on mixture distributions and their used in modelling data. Everitt and Hand (1981), Titterington et al. (1985), Lindsay (1995), Böhning (1999) and

---

<table>
<thead>
<tr>
<th>i</th>
<th>MASS</th>
<th>$y_i$</th>
<th>$X_i$</th>
<th>$w_i^{(r+1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$y$</td>
<td>$X$</td>
<td>$w_1^{(r+1)}$</td>
</tr>
<tr>
<td>n</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
<td>$y$</td>
<td>$w_2^{(r+1)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$X$</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$K$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>$K$</td>
<td></td>
<td>$y$</td>
<td>$w_K^{(r+1)}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$X$</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.2: Table showing the expansion of data use in M-step (1) of the EM algorithm for fitting the common parameter mixture model

**data** This should be a data frame. Note that this argument is mandatory for this function even if the data are attached. This is because the data frame is used to expand the data as in Table 6.2.

**K** Declaring the number of mixture components (in fixed effects finite mixture models), or the number of mass points or integration quadrature points (for random effects models)

**mixture** Defining the mixing distribution, "np" for non-parametric finite mixtures or "gq" for Gaussian quadrature.

**tol** This defines the tolerance scalar usually between zero and one, used for changing the starting values.

**weights** For prior weights

**control** This sets the control parameters for the EM iterations algorithm. The default setting is the `NP.control` function.

**g.control** This is for controlling the gamlss control function, `gamlss.control`, passed to the gamlss fit

... For extra arguments
McLachlan and Peel (2000) are some of the books dedicated exclusively on mixture distributions. Aitkin, Francis and Hinde (2005) include useful chapters related to mixture distributions.

6.9 Exercises

6.9.1 Exercise 1

Here we analyse the acidity data, which records the acidity index for 155 lakes in the Northeastern United States [previously analysed as a mixture of Gaussian distributions on the log scale by Crawford et al. (1992, 1994)]. These 155 observations are the log acidity indices for the lakes.

- a) Load the acidity data, print the variable name and obtain a histogram of the acidity index values:

```r
data(acidity)
names(acidity)
hist(y)
```

- b) Fit a mixture of two normal distributions (with different means and variances) [note gamlssMXfits fits the mixture from n=20 different starting values and chooses the best mixture fit]:

```r
mm<-gamlssMXfits(n=20,y~1, family=NO, K=2)
mm
```

- c) Calculate the probability density function for the fitted mixture and plot it with the histogram of the acidity index values

```r
fy<-dMX(y=seq(1,8,.1), mu=list(6.249, 4.33),
        sigma=list(exp(-.6535), exp(-.988)),
        p=list( 0.4039988, 0.5960012) )
hist(y,freq=FALSE)
lines(seq(1,8,.1),fy, col="red")
```
Chapter 7

Fitting distributions to data: theory

This chapter provides theoretical background for fitting a distribution to data. In particular it explains:

1. basic concepts of statistical inference
2. maximum likelihood estimation

This chapter is essential for understanding the theory on how to fit a distribution to data.

7.1 Introduction to parametric statistical inference

The stochastic part of a statistical model usually arises from the assumption that the data we observe are a sample from a larger (unknown) population whose properties we are trying to study. The statistical model will be a simplification of the population and its behaviour. Essential to understanding the basic concepts of statistical inference are the ideas of the population, the sample and the model which we will consider next.

7.1.1 The population

The population is the set of the particular subjects we would like to study. The interest lies usually in some characteristics of the population which manifest themselves as a set of variable(s) say $Y$.

Populations can be real, e.g. the height of adults living in the UK, or can be conceptual, e.g. in a clinical trial we have only a limited amount of people taking the actual treatment, but our interest lies in all possible people who could have taken the same treatment. The number of elements in the population is finite, say $N$, even though the number can be very large.

The range of possible values that $Y$, the population characteristic, can take depends on the type of variable we measure i.e. continuous, discrete or categorical. In practice (even for continuous variables) observed population measures can only take a finite number of values, since $N$ is finite. So the range of all possible observed values for $Y$ can be considered to be discrete (even if $Y$ is a continuous variable in the conventional mathematical sense) but possibly very very large. For a discrete variable $Y$, some of the values of $Y$ in the population may be identical.
Let $D$ be the number of distinct values of $Y$ in the population, and $N_I$ the number of times that the distinct value $Y_I$ occurs in the population. Then the total population size would be $N = \sum_{I=1}^{D} N_I$. Let $P_I = N_I / N$, then the population distribution is defined by a probability function (pf) given by

$$f_P(y) = P(Y = y) = \begin{cases} P_I & \text{if } y = Y_I \text{ for } I = 1, 2, \ldots, D \\ 0 & \text{otherwise} \end{cases}$$ (7.1)

That is, the probability of observing $y$ in the population is the number of times $y$ occurs in the population divided by the total population size. In the special (but common case) where $Y$ is a continuous variable, all its values are distinct, and the probability function is just needle points at the distinct values of $Y$ with equal height probability $1/N$.

The population cumulative distribution function (cdf), $F_P(y) = P(Y \leq y)$, the sum of $P_I$ over all $I$ for which $Y_I \leq y$, is a step function increasing each time a distinct value appears in the population. Figure 7.1 (b) shows a sample cdf for a continuous variable but if you imagine that the number of distinct values in the population is a very large number, then the population cdf may look almost continuous.

### 7.1.2 The sample

The 'true' population probability distribution $f_P(y)$ is usually unknown unless we are prepared to invest time, effort and money to obtain data from the whole population. Suppose we observe a subset of the population which we called the sample.\footnote{Strictly this is a random sample without replacement from a finite population.} The sample is denoted here as a vector $y$ of length $n$. Let $d$ be the number of distinct values of $Y$ in the sample $y$, and $n_I$ the number of times that the distinct value $y_I$ occurred in the sample. The total sample size would be $n = \sum_{I=1}^{d} n_I$. Let $p_I = n_I/n$, then the sample distribution is defined by an empirical probability function (epf) of the sample $y$ given by

$$f_E(y) = P(Y = y) = \begin{cases} p_I & \text{if } y = y_I \text{ for } I = 1, 2, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$ (7.2)

The empirical probability function (epf), $f_E(y)$, can be plotted as a bar (or needle) plot as in Figure 7.1(a). Plotting the cdf for a continuous variable $Y$ is not a very informative graph and statisticians use instead histograms or a smooth version of the plot called a nonparametric density estimator (see Silverman [1988] or Wand and Jones [1999] for the theory or section 8.2 for R functions). The empirical cumulative distribution function, (ecf), $F_E(y) = \hat{P}(Y \leq y)$, the sum of $p_I$ over all $I$ for which $Y_I \leq y$, is a step function increasing each time a distinct value appears in the sample. Figure 7.1 shows typical empirical probability and cumulative distribution functions from a continuous univariate data example. This specific example will be analysed in more detail in Section 8.3.1. The data are the annual snowfall in Buffalo, NY (inches) for 63 years, from 1910 to 1972 inclusive and were obtained from Hand et al. (1994). A non-parametric density estimator function is also superimposed within the epf plot of Figure 7.1 (a) representing approximately the sample distribution of snowfall in a more informative way. The empirical probability and cumulative distribution functions play a very important role in statistical (parametric and non-parametric) inference.

### 7.1.3 The model

Statistical inference is the process in which a sample is used to make inference about the population distribution. A statistical model involves a set of assumptions on how the sample is gener-
Parametric models

Classical parametric statistical inference assumes that the true population distribution \( f_P(y) \) belongs to a particular family of probability (density) functions, given by \( f_Y(y|\theta) \) for a range of values of \( \theta \), and \( f_P(y) = f_Y(y|\theta_T) \) for all \( y \) for a particular unknown value \( \theta_T \) of \( \theta \), where \( \theta \) is a vector of parameters.

The problem of the classical parametric statistical inference is then to determine possible values for the parameter \( \theta \) given the sample \( y \). Note that \( \theta \) is used here as generic parametric notation, since any one to one transformation of \( \theta \) (called a re-parametrization) will also perfectly define the theoretical distribution \( f_Y(y|\theta) \). That is, \( f_Y(y|\theta) \) and \( f_Y(y|\phi) \) are equivalent assumptions if \( \phi = g(\theta) \) and the function \( g() \) is a one to one transformation from the parameter space of \( \theta \) to the parameter space of \( \phi \). Note that in our notation above:

- We use \( Y \) to describe both the population characteristic we are studying and also the random variable involved in the theoretical probability (density) function \( f_Y(y|\theta) \), (which can be continuous).

- By using a theoretical distribution \( f_Y(y|\theta) \), we have replaced the observable characteristic(s) of the population variable \( Y \), (which has a finite range of values), with a theoretical random variable \( Y \).

Figure 7.2 shows a schematic view\(^2\) of the true population distribution, the empirical distribution and the the model \( f_Y(y|\theta) \). Both population (“Population”) and empirical (“Sample”)

\(^2\)In a schematic view we are trying to represent the concepts rather that correct mathematically defined spaces.
distributions are denoted as points while the “Model” \( f_Y(y|\theta) \) is represented as a line. Different values of the parameter \( \theta \) represent different points on the line. The different elliptical regions in Figure 7.2, around the true population distribution, represent loosely the variation of samples around the population distribution, so samples on a contour line have equal probability to be observed.

Note that the model probability (density) function \( f_Y(y|\theta) \), is not real in the sense that it actually exists, but it is treated here as a surrogate of the unknown population distribution \( f_P(Y) \). Given that \( f_Y(y|\theta) \) is ‘close’ to \( f_P(Y) \) and given that we choose \( \theta \) carefully, we should have a very good approximation of ‘true’ population distribution \( f_P(Y) \). Note that within this set up there are two unknown quantities:

(i) the theoretical probability (density) function \( f_Y() \) which can be selected from a large number of appropriate distributions within the statistical literature and

(ii) \( \theta \), the parameter(s).

Classical parametric statistical inference assumes that the true population distribution \( f_P(y) \) belongs to a particular family of probability (density) functions, given by \( f_Y(y|\theta) \) for a range of values of \( \theta \), and \( f_P(y) = f_Y(y|\theta) \) for all \( y \) for a particular value of \( \theta \), where \( \theta \) is a vector of unknown parameters, i.e. the population distribution lies on the model line in Figure 7.2.

If this is true, then the problem of statistical inference becomes one of finding a suitable value for \( \theta \), which leads to point or interval estimation for \( \theta \).

Classical inference has put more emphasis on inference about \( \theta \), assuming a particular theoretical \( f_Y(y|\theta) \), rather than finding an appropriate family of distributions for \( f_Y(y|\theta) \) in the first place.
7.2 Likelihood function

If \( f_Y(y; \theta) \) is not a good approximation to the true population distribution for any \( \theta \), then this part of the classical statistical inference can be irrelevant to the particular problem. The dependence of inferential conclusions on the specification of the family of distributions \( f_Y(y; \theta) \) for a range of values of \( \theta \) has lead statisticians to the development of non-parametric methods of inference.

Non-parametric models

Non-parametric statistical assumptions do not involve unknown parameters. Examples of models where non-parametric assumptions are invoked are the so called non-parametric tests. In these tests it is assumed that the sample is generated from an unknown population in an independent and identical distributed manner (iid), but the population distribution is not explicitly defined. Another example is the density estimator we met earlier in Figure 7.1. Density estimators are part of statistics called non-parametric smoothing techniques. Strictly speaking in those techniques there are parameters to be determined from the data called the smoothing parameters (see also section 8.2).

The empirical probability and cumulative functions are of paramount importance within the non-parametric statistical inference. This is due to the fact that for a random sample of size \( n \) coming from the true population probability (density) function (pdf) \( f_P(Y) \), the ecf \( F_E(y) \) is a consistent estimator as \( n \to \infty \) of the ‘true’ population cdf \( F_P(y) \), (assuming either random sampling from a population of infinite size, \( N = \infty \), or random sampling with replacement from a finite population size \( N \)). The ecf also provides, through the plug-in principle (see Efron and Tibshirani [1993] page 35), a simple method of non-parametric estimation. To demonstrate the plug-in principle, let us assume, that we are interested in a specific characteristic of the population, say \( \vartheta = t(F_P) \). This characteristic, \( \vartheta \), is a function, say \( t() \), of the ‘true’ population cumulative distribution function (cdf) \( F_P(Y) \), e.g., the mean of the population. The plug-in estimate of the parameter \( \vartheta \) is given by replacing the true cdf \( F_P \) in the function \( t() \) by the ecf \( F_E \), i.e., \( \hat{\vartheta} = t(F_E) \). So for example, according to this principal, the sample mean \( \bar{y} = \hat{\mu} = \sum_{i=1}^{d} y_i p_i \) is the plug-in estimate of the population mean, \( \mu_p = \sum_{i=1}^{D} Y_i P_i \).

7.2 Likelihood function

The concept of the likelihood function is of vital importance in parametric statistical inference. It is based on the reasoning that ‘parameter values which make the data appear relatively probable according to the model are more likely to be correct than parameter values which make the data appear relatively improbable according to the model ”, (Wood [2013]) .

The major statistical schools of inference, Bayesian, Classical and pure likelihood, use the likelihood function as their main inferential tool.

Bayesian inference uses the likelihood function as the source of information given by the data, which is combined with a prior distribution for the parameters, to form the posterior distribution of the parameters (see section 7.3.1 and Gelman, A. Carlin, J. B. Stern, H. S. and Rubin [2004]).

Classical inference treats parameters as unknown constants, assumes the data are a realization of potentially repeated sampling from the assumed model, and makes inference about the parameters using the likelihood, (see section 7.3.2 and Cox and Hinkley [1979]).

A third smaller group supporting the pure likelihood approach where the likelihood function is used exclusively for information about the parameters, see section 7.3.3 and Edwards [1972] or Lindsey [1996].
In the data mining community the use of empirical risk function in parameter estimation leads to maximum likelihood estimation, (see appendix 7.9.1 and Hastie, T. J., Tibshirani, R. J. and Friedman [2001]).

7.2.1 Definition of likelihood function

The likelihood function, $L(\theta)$, is the probability of observing the sample, viewed not as a function of the sample $y$ but as a function of the parameter(s) $\theta$.

Let $y = (y_1, y_2, \ldots, y_n)$ be an observed random sample, from an assumed discrete population parametric probability function $f_Y(y|\theta)$ with a known functional form except for unknown parameter(s) $\theta$. A random sample is a sequence of independently and identically distributed (iid) random variables with a particular population distribution. [Assume either random sampling from a population of infinite size, $N = \infty$, or random sampling with replacement from a finite population size $N$.] The probability of observing the sample under the assumed model is:

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i|\theta). \quad (7.3)$$

That is, the joint probability of the sample is the product of the individual probabilities since the observations are assumed to be independent. Note the change of emphasis in the argument of the likelihood function from $y$ to $\theta$. Given the observed values $y$, the likelihood is not a function of the sample $y$, since this has been observed and therefore is fixed, but a function of the parameter(s) $\theta$.

The classical method of fitting a parametric family to an observed random sample of values is the method of maximum likelihood estimation, that is, maximising the likelihood of equation (7.3) with respect to the parameter(s) $\theta$. In practice it is more convenient to work with the logarithm of the likelihood. The log-likelihood is defined as

$$\ell(\theta) = \sum_{i=1}^{n} \log f_Y(y_i|\theta). \quad (7.4)$$

In this book we deal with iid samples which justifies equation (7.3). More generally the definition of the likelihood as the probability of observing the sample still holds but not (7.3). For example, in time series data, where we are assuming that an observation at time $t$ is conditional on the previous history of the $y_t$, the likelihood takes the form:

$$L(\theta) = f_Y(y_1|\theta)f_Y(y_2|(y_1, \theta))f_Y(y_3|(y_1, y_2, \theta)) \cdots f_Y(y_n|y_1, \ldots, y_{n-1}, \theta). \quad (7.5)$$

7.2.2 Clarification of the likelihood function for a continuous variable

Note that equation (7.3) is the exact probability of observing the data $y$ given the parameters $\theta$, provided the distribution of $Y$ is discrete. If however the distribution of $Y$ is continuous, then in practice a specific value $y_i$ is observed to a certain level of accuracy, say $y_i \pm \Delta_i$. [For example, if $y_i$ is rounded to the nearest first decimal place then $\Delta_i = 0.05$ and, for example, an observed value $y_i = 5.7$ corresponds to $5.65 < y < 5.75.$] Hence the true likelihood (i.e. the true probability of observing the data $y$) can be defined as:

$$L(\theta) = \prod_{i=1}^{n} P(y_i - \Delta_i < Y < y_i + \Delta_i|\theta) = \prod_{i=1}^{n} [F_Y(y_i + \Delta_i|\theta) - F_Y(y_i - \Delta_i|\theta)] \quad (7.6)$$
where \( F_Y() \) is the cumulative distribution function of \( Y \). The definition of the likelihood in (7.6) is bounded above by one so cannot go to infinity, something which could happen if the definition (7.3) of the likelihood is used instead. Assume the \( \Delta_i \)'s are sufficiently small then

\[
L(\theta) \approx \prod_{i=1}^{n} f_Y(y_i|\theta)\Delta_i = \left[ \prod_{i=1}^{n} \Delta_i \right] \left[ \prod_{i=1}^{n} f_Y(y_i|\theta) \right].
\]

(7.7)

Hence the log likelihood \( \ell(\theta) \) is given approximately by:

\[
\ell(\theta) = \sum_{i=1}^{n} \log f_Y(y_i|\theta) + \sum_{i=1}^{n} \log \Delta_i.
\]

(7.8)

Clearly the second summation does not depend on \( \theta \) and hence when maximising \( \ell(\theta) \) over \( \theta \) only the first term needs to be maximised. Occasionally this creates problems, (especially in flexible models such as GAMLSS) where the fact that we ignored the accuracy with which the response variable is measured can occasionally lead to the likelihood shooting up to infinity. To demonstrate the point consider a single observation \( y \) from a normal distribution, i.e. \( Y \sim \text{NO}(\mu, \sigma) \). The likelihood is maximised as \( \mu \to y \) and \( \sigma \to 0 \) and the likelihood goes to \( \infty \). The problem is avoided by taking account of the measurement accuracy by using (7.6) instead of (7.7).

Within GAMLSS we have adopted the definition of the likelihood given in (7.3). Models can be maximised using (7.6) with the help of the package \texttt{gamlss.cens} which is designed for censored or interval response variables. In this case one can think of the response variable having the form

- \(( -\infty, y_{i2} ) \) if the response is left censored
- \(( y_{i1}, +\infty ) \) if the response is right censored
- \(( y_{i1}, y_{i2} ) \) if the response lies within an interval

In all three cases the likelihood takes the form

\[
L(\theta) = \prod_{i=1}^{n} \left[ F_Y(y_{i2}|\theta) - F_Y(y_{i1}|\theta) \right].
\]

(7.9)

### 7.2.3 Air conditioning example: likelihood function

The following data reported by Proschan (1963), refer to the intervals, in service-hours, between failures of the air-conditioning equipment in a Boeing 720 aircraft. Proschan reports data on 10 different aircraft but here we are following the \texttt{rpanel} package, Bowman \textit{et al.} (2007), and use only 24 observations from one of the aircraft:

50 44 102 72 22 39 3 15 197 188 79 88 46 5 36 22 139 210 97 30 23 13 14.

A histogram of the data is shown in Figure 7.3. All data points are positive so we require a distribution defined on the positive line. We are going to follow two different scenarios here.
Scenario I: exponential distribution

First assume that the data are independent identically distributed observations coming from an exponential distribution (a one parameter distribution) as in equation (1.4). Under this scenario the likelihood function is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\mu} e^{-y_i/\mu}$$  \hspace{1cm} (7.10)

with log likelihood function given by

$$\ell(\mu) = \sum_{i=1}^{n} \left\{ -\frac{y_i}{\mu} - \log(\mu) \right\}$$

$$= -\frac{1}{\mu} \sum_{i=1}^{n} y_i - n \log(\mu)$$  \hspace{1cm} (7.11)

Note that the likelihood function depends on the the data only through the function $S = \sum_{i=1}^{n} y_i$. Functions of the data only, are generally called statistics. Functions of the data appearing in the likelihood, like $S$, are called sufficient statistics. Sufficient statistics are important within the Classical methodology of inference because they provide a way of finding good estimators. Unfortunately sufficient statistics exist only if the assumed distribution belongs to the Exponential Family of distributions (see section 11.8.3). That itself limits their usefulness.

The likelihood and the log-likelihood function of equations (7.10) and (7.11) respectively for the parameter $\mu$ are shown in Figure 7.4 (a) and (b) respectively.

The following R code was used for the plot:

```r
aircond <- c(50, 44, 102, 72, 22, 39, 3, 15, 197, 188, 79, 88, 46, 5,
```
Scenario II: gamma distribution

Under the second scenario assume that the data are independent identically distributed observations coming from a gamma distribution (a two parameter distribution) as in equation (15.2). Under the gamma distribution scenario the likelihood function is

\[ L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{(\sigma^2 \mu)^{1/\sigma^2}} \left( \frac{1}{\Gamma(1/\sigma^2)} \right) \frac{y_i^{-1/\sigma^2} e^{-y_i/(\sigma^2 \mu)}}{\Gamma(1/\sigma^2)} \]  

(7.12)
with log Likelihood

\[
\ell(\mu, \sigma) = \sum_{i=1}^{n} \left[ -\frac{1}{\sigma^2} \left( \log \sigma^2 + \log \mu \right) + \left( \frac{1}{\sigma^2} - 1 \right) \log y_i - \frac{y_i}{\sigma^2 \mu} - \log \Gamma \left( \frac{1}{\sigma^2} \right) \right]
\]

\[
= -\frac{n}{\sigma^2} (\log \sigma^2 + \log \mu) - n \log \Gamma \left( \frac{1}{\sigma^2} \right) + \left( \frac{1}{\sigma^2} - 1 \right) \sum_{i=1}^{n} \log y_i - \frac{\sum_{i=1}^{n} y_i}{\sigma^2 \mu}
\]

(7.13)

There are two sufficient statistics in the gamma case \( S = \sum_{i=1}^{n} y_i \) and \( T = \sum_{i=1}^{n} \log y_i \) because the gamma distribution does belong to the exponential family. Note also that the gamma distribution used by (15.2) has population mean \( E(Y) = \mu \) and population variance \( V(Y) = \sigma^2 \mu^2 \).

The likelihood function in this case is two dimensional because the gamma distribution has two parameters \( \mu \) and \( \sigma \). Different views of the log-likelihood function of equation (7.13) for the \( \mu \) and \( \sigma \) parameters are shown in Figure 7.5.

The top left corner of Figure 7.5 shows the two dimensional log likelihood function \( \ell(\mu, \sigma) \) against \( \mu \) (\( \theta_1 \)) and \( \sigma \) (\( \theta_2 \)). The top right corner shows a contour plot of the log likelihood function against \( \mu \) and \( \sigma \). The bottom left and right corners of Figure 7.5 show what the two dimensional log likelihood function looks like from the direction of the parameters \( \mu \) and \( \sigma \) respectively. They show the profile log likelihood functions for each of the two parameters \( \mu \) (\( \theta_1 \)) and \( \sigma \) (\( \theta_2 \)) respectively. See section 7.6.1 for the definition of the profile log likelihood function.

The following R code was used for the contour plot of the likelihood.

```r
grid2d <- data.frame(expand.grid(mu=seq(30,120,1),sigma=seq(0.5,1.7,0.05)))
lg <- dim(grid2d)[1]
Lik <- rep(0, lg)
for (i in 1:lg) Lik[i] <-sum(dGA(aircond, mu=grid2d$mu[i],
    sigma=grid2d$sigma[i], log=TRUE))
mu <-seq(30,120,1)
sigma <-seq(0.5,1.7,0.05)
op <- par(mfrow=c(1,1))
contour(mu,sigma, matrix(Lik, nrow=length(mu)), nlevels=30, ylab="sigma",
    xlab="mu")
```

The three dimensional plots are done using the function \texttt{rp.likelihood()} from the package \texttt{rpanel} of Bowman, A. Crawford, E. Alexander, G. and Bowman [2007].

Maybe we should demonstrate the use of the alternative definition of the likelihood, interval censored?

### 7.3 Using the likelihood function for statistical inference

In the previous section we defined the likelihood and the log-likelihood functions. The next question is how those functions can be used for statistical inference about the parameters. The Bayesian and Classical schools of statistics use the likelihood function differently. Bayesian inference uses the likelihood as the only source of information about the parameters coming from the data, which is then combined with a prior distribution for the parameters to give the posterior distribution of the parameters. Classical inference uses the likelihood for inference
Figure 7.5: Showing different views of the log likelihood function for the air conditioning data assuming a gamma distribution: (a) top left corner shows the two dimensional log likelihood for $\mu$ (theta1) and $\sigma$ (theta2) (b) top right corner shows a contour plot of the two dimensional log likelihood. (c) The bottom left the profile log likelihood for $\mu$ (theta1) (d) bottom right the profile log likelihood for $\sigma$ (theta2) (see text for more details).
about the parameters, assuming repeated sampling of the data is allowed. Likelihood inference uses purely the likelihood for inference.

Sections 7.3.1, 7.3.2, and 7.3.3, briefly show how the likelihood function is used by respectively the Bayesian, Classical and pure likelihood schools of inference. Particular attention is given to the following questions:

• how to use the likelihood function for inference about the parameter(s) \( \theta \)?

• how to eliminate nuisance parameters?

• how to choose between different models?

The three different schools of statistical inference answer these questions differently. The three questions are answered in more detail for Classical inference in sections 7.4 to 7.7.

### 7.3.1 Bayesian inference

Bayesian inference uses the posterior distribution, \( f_\theta(\theta|y) \), for \( \theta \), given the observed \( y \), to draw inference about \( \theta \). The posterior distribution is defined as:

\[
f_\theta(\theta|y) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)\,d\theta}
\]

\[
\propto L(\theta)\pi(\theta)
\]

(7.14)

where \( \pi(\theta) \) is a prior distribution for the parameters \( \theta \) and where \( \int L(\theta)\pi(\theta)\,d\theta \) is a constant to ensure that the posterior \( f_\theta(\theta|y) \) integrates to one. It is obvious from equation (7.14) that the only information coming from the data is contained in the likelihood function \( L(\theta) \).

The following comments are related to Bayesian inference:

• By having a posterior distribution for all the parameters in \( \theta \), probabilistic conclusions can be drawn about them. For example the mean, mode (maximum a posteriori or MAP), variance or quantiles of the posterior distribution of \( \theta \) can be obtained. The parameters in \( \theta \) are random variables as far as Bayesian inference is concerned.

• The Bayesian school is fully conditioning its inference on the given data (and the appropriateness of the assumed model). It is not concerned with what could have happened but only on what did happen.

• The derivation of the \( f_\theta(\theta|y) \) involves integration, possibly in high dimensional spaces, something that had held up the spread of Bayesian techniques for a long time. Nowadays the use of computer simulation techniques such as the Monte Carlo Markov chains (MCMC) has changed this, see Gilks, W.R. Richardson, S. and D.J. [1996] and Gill [2006].

• A problem with the Bayesian school of inference has to do with the priors \( \pi(\theta) \). Bayesian theory requires any prior information about \( \theta \) to be expressed as a prior probability (density) function. There can be in informative or non-informative priors. Informative priors are based on prior information or relative beliefs, while non-informative priors are uninformative relative to the information contained in the data and therefore have minimal influence on the inference. Note that the prior \( \pi(\theta) \) does not have to be a proper distribution as long as the posterior \( f_\theta(\theta|y) \) is. For more information about priors see Barnett [1999], Gelman, A. Carlin, J. B. Stern, H. S. and Rubin [2004] or Chapter 2 in Aitkin [2010].
7.3. USING THE LIKELIHOOD FUNCTION FOR STATISTICAL INFERENCE

- If we are interested only in one of the parameters in the vector $\theta$, say $\theta_1$, we can eliminate the rest of the parameters (sometimes called the nuisance parameters) by integrating them out i.e. $f_{\theta_1}(\theta_1) = \int_{\theta_2} \cdots \int_{\theta_k} f_\theta(\theta|y) d\theta_2 \cdots d\theta_k$. That is, Bayesian inference provides a consistent way of eliminating nuisance parameters.

- From the statistical modelling point of view, note that the likelihood $\mathbb{L}(\theta)$ refers to only one possible model. In practice, for a given data set, we are often faced with the situation where several different models perform well. Bayesian factors and Deviance information criterion are used in those circumstances to choose between models, Raftery [1999], Gelman, A. Carlin, J. B. Stern, H. S. and Rubin [2004].

- Bayesian inference for a particular model deals well with parameter uncertainty, but is very vulnerable to model mis-specification.

7.3.2 Classical inference

Here we give an overview of how Classical inference answers the three question given at the start of section 7.3.

In section 7.4 maximum likelihood estimation is considered. An example is given in section 7.4.1. In section 7.5, the statistical properties of the maximum likelihood estimator (MLE) are given assuming the population model is correct, i.e. assuming that the population distribution $f_P(y)$ belongs to a particular family of probability (density) functions $f_Y(y|\theta)$ for a range of values of $\theta$ and $f_P(y) = f_Y(y|\theta_T)$ for all $y$ for a particular value of $\theta = \theta_T$, where $\theta_T$ is the true value of $\theta$. Standard error based approximate confidence intervals and Wald tests for a parameter $\theta$ are given in section 7.5.3.

More accurate profile confidence intervals and generalized likelihood ratio tests for a parameter $\theta$ are given in sections 7.6.1 and 7.7.1 respectively. Section 7.7.3 investigates model selection using the generalized Akaike information criterion.

In section 7.8, the statistical properties of the maximum likelihood estimator (MLE) are given when the population model is mis-specified, i.e. $f_P(y) \neq f_Y(y|\theta)$ for all $y$ for any value of $\theta$. Under model mis-specification, robust standard error based approximated confidence intervals and tests for $\theta_c$, [where $\theta_c$ is the value of $\theta$ which makes $f_Y(y|\theta)$ closest to $f_P(y)$, as measured by the Kullback-Liebler distance] are also given in section 7.8.

7.3.3 Pure likelihood inference

The likelihood function provides a way for ordering possible values of the parameter(s) $\theta$ given the data, but it is not a probability (density) function. For two different values of $\theta$, say $\theta_1$ and $\theta_2$, the likelihood ratio

$$\frac{L(\theta_1)}{L(\theta_2)}$$

provides the ratio of how likely the data is, given the value of $\theta_1$, relative to the value of $\theta_2$. For example, a likelihood ratio of 10 means that the data is ten times more likely given $\theta_1$ than given $\theta_2$. This interpretation of the likelihood function is the basis of the pure likelihood theory approach in statistical inference, Edwards [1972], Lindsey [1996]. This approach is fully conditional on the data, so denies the repeated sampling arguments of the classical school, but also rules out prior distributions. Note that an important point here is that points on the curve of the likelihood function have a meaning but not areas under the curve. The likelihood function
can not be interpreted as a posterior distributions unless it is combined with a prior distribution as in equation (7.14).

It is common to standardize the likelihood function by dividing it by its maximum. The resulting quantity

\[ L_s(\theta) = \frac{L(\theta)}{L(\hat{\theta})} \]

is called the standardized likelihood function. The standardized likelihood can be used to create 'confidence' bounds for the \( \theta \) parameters. One can take the region of all values of \( \theta \) where \( L_s(\theta) > 0.10 \) as a possible 'confidence' bound region. This should not interpreted as a classical confidence interval, but rather as a support region where all the values within are plausible.

Another possibility, by using Classical inferential arguments is to use the region of values of \( \theta \) for which \( L_s(\theta) > 0.1466 \) as an approximate 95% confidence region. The value 0.1466 is calculated by 0.1466 = \( \exp\left(-\chi^2_{10.05}/2\right) \approx \exp(-3.84/2) \) and is justified by section 7.7.1.

Figure 7.6 shows the standardized likelihood and log likelihood functions for \( \mu \) in the aircond data under scenario I. On the top panel is the likelihood while in the bottom the log-likelihood. In addition, MLE of \( \mu \) is shown together with the construction of confidence bounds. In the top panel the horizontal line is at the point of 0.1466. The limits of the confidence bounds for \( \mu \) are where this horizontal lines crosses the likelihood. The actual values for the bounds for \( \mu \) are [44.3, 98.3]. The horizontal line on the log-likelihood is at log(0.1466) = -1.92.

Figure 7.7 considers the two dimensional case of the gamma distribution of scenario II. It shows confidence contours defined by the standardised likelihood values of 0.5, 0.1 and 0.05 which correspond to values of -0.693, -2.302 and -2.995 in the log of the standardized likelihood function. Values of the parameters within say the middle confidence contour, 0.1, are more
likely to be true since they are 'supported' more by the data than the points outside. However the choice of 0.1 or any other values is rather arbitrary and does not correspond to any probabilistic statement. Probabilistic statements about the parameters can only be made by Bayesian inference.

7.4 Maximum likelihood estimation

The maximum likelihood estimate (MLE), is the value of \( \theta \), say \( \hat{\theta} \), which maximizes the likelihood function \( L(\theta) \) or equivalently maximizes the log-likelihood function \( \ell(\theta) \). The maximum likelihood estimate is called a point estimate in Classical statistical inference. Point estimators are the solutions to the following inferential problem: If I have to guess the true value \( \theta_T \) of the parameters \( \theta \) what will this be? Point estimators are rather naive in the sense that even if we believe that there is a "true" parameter \( \theta_T \) our guess would inevitably be wrong. [Standard error based confidence intervals for a parameter \( \theta \) are considered in section 7.5.3 while more reliable profile confidence intervals for \( \theta \) are considered in section 7.6.1]

The maximum likelihood estimate (MLE) of \( \theta_T \) is a function of the observed sample \( y \), while the maximum likelihood estimator (MLE) of \( \theta_T \) is the same function of the random sample (of iid random variables) \( Y \). Hence the MLE of \( \theta_T \) is a vector of constant numerical values calculated from the observed \( y \), while the MLE of \( \theta_T \) is a vector of random variables which are functions of \( Y \).

Maximum likelihood estimates can be derived analytically or numerically. Analytical solutions are rather rare, so numerical computation of MLe's is frequently needed.
7.4.1 Air conditioning example continued: maximum likelihood estimation

Scenario I: exponential distribution

This demonstrates how to find the MLE analytically, using scenario I in section 7.2.3 where an exponential distribution was assumed. Differentiating the log likelihood function, \( \ell(\mu) \), of equation (7.11) with respect to \( \mu \) gives:

\[
\frac{d\ell(\mu)}{d\mu} = \frac{\sum_{i=1}^{n} y_i}{\mu^2} - \frac{n}{\mu}.
\]

(7.15)

By setting the result to zero and solving for \( \mu \) gives

\[
\frac{d\ell(\mu)}{d\mu} = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}.
\]

(7.16)

where \( \hat{\mu} \) is the MLE of \( \mu \). [Since \( \frac{d^2\ell}{d\mu^2} < 0 \) at \( \hat{\mu} = \bar{y} \), it is a maximum point.] In this case the maximum likelihood estimate \( \hat{\mu} \) is the sample mean \( \bar{y} \) of the the observations \( y \). Note \( \bar{y} \) is the maximum likelihood estimate (MLE) of \( \mu \), while the corresponding random variable \( \bar{Y} \) is the maximum likelihood estimator (MLE) of \( \mu \). Note also that it is common in statistics to use a hat \( \hat{\cdot} \) to indicate an estimate. For the example in section 7.2.3 on the time intervals, in hours, between failures of the air-conditioning equipment, the MLE of \( \mu \) is \( \hat{\mu} = \bar{y} = 64.125 \) hours, assuming an exponential distribution for \( Y \).

Scenario II: gamma distribution

For scenario II, the gamma distribution was assumed. Differentiating the log likelihood function, \( \ell(\mu, \sigma) \), of equation (7.13) with respect to \( \mu \) and \( \sigma \) gives respectively:

\[
\frac{d\ell(\mu, \sigma)}{d\mu} = \frac{\sum_{i=1}^{n} y_i - n\mu}{\sigma^2\mu^2}
\]

(7.17)

\[
\frac{d\ell(\mu, \sigma)}{d\sigma} = \frac{2}{\sigma^3} \left[ \left( \sum_{i=1}^{n} \frac{y_i}{\mu} \right) - \left( \sum_{i=1}^{n} \log(y_i) \right) + n\log(\mu) + n\log(\sigma^2) - n + n\psi \left( \frac{n}{\sigma^2} \right) \right]
\]

(7.18)

where \( \psi(x) = \frac{d}{dx} \log \Gamma() \) is the psi or digamma function. Setting the two equations (7.17) and (7.18) equal to 0 and solving them simultaneously gives the MLE for both \( \mu \) and \( \sigma \). Clearly \( \hat{\mu} = \bar{y} \) but \( \hat{\sigma} \) can not be obtained explicitly. The next section shows how \( \hat{\mu} \) and \( \hat{\sigma} \) can be obtained numerically.

Numerical maximization

In order to find the MLE numerically in R there are several options. Here only two options will be considered: i) the use of the general optimisation function \texttt{optim()} and ii) the use of a more specialised function \texttt{mle()} in the package \texttt{stats4}. Note though that the function \texttt{mle()} uses function \texttt{optim()} as its minimization engine. The functions \texttt{gamlss()}, \texttt{gamlssML()}, \texttt{histDist()} and \texttt{fitDist()} which are used in the rest of this book for fitting distributions to data are explained in detail in Chapter 8.
First define the log-likelihood function or more precisely minus the log-likelihood (since by default \texttt{optim()} minimises rather than maximises functions) and then call the function.

For scenario I:

\begin{verbatim}
logl <- function(mu) -sum(dEXP(aircond, mu=mu, log=TRUE))
optim(45, logl, method="Brent", lower=0.01, upper=1000)$par
[1] 64.125
optim(45, logl, method="L-BFGS-B", lower=0.01, upper=Inf)$par
[1] 64.12498
\end{verbatim}

Value 45 is used as a starting value for \( \mu \), while 0.01 and 1000 (or \( \text{Inf} \)) are used as the lower and upper limits. Note that in the definition of the function \( \text{logl} \) the \texttt{gamlss.family} function \texttt{dEXP()} with argument \( \text{log=TRUE} \) is used to get the log-likelihood of the exponential distribution. In the function \texttt{optim()}, since this is a one parameter minimization, option \texttt{method="Brent"}, which is recommended for such situations, is used. For multidimensional \( \theta \) as below the option \texttt{method="L-BFGS-B"} should be used.

For the scenario II:

\begin{verbatim}
loglgamma <- function(p) -sum(dGA(aircond, mu=p[1], sigma=p[2], log=TRUE))
optim(c(45,1), loglgamma, method="L-BFGS-B", lower=c(0.01, 0.01),
    upper=c(Inf, Inf))$par
[1] 64.122747 0.972409
\end{verbatim}

Both \texttt{optim()} and \texttt{mle()} methods allow restrictions on the values of the parameter space something which is important in this case since the \( \mu \) parameter in the exponential distribution and the \( \mu \) and \( \sigma \) parameters in the gamma distribution only takes positive values.

Now use the function \texttt{mle()}:

\begin{verbatim}
m1<-mle(logl, start=list(mu=45))
m1@fullcoef
 mu 64.06148

m2<-mle(logl, start=list(mu=45), method="Brent", lower=0.01, upper=1000)
m2@fullcoef
 mu 64.125
 m2@min
[1] 123.86

loglgamma1 <- function(mu, sigma) -sum(dGA(aircond, mu=mu, sigma=sigma, log=TRUE))
m3<-mle(loglgamma1, start=list(mu=45,sigma=1), lower=c(0.01, 0.01),
    upper=c(Inf, Inf), method="L-BFGS-B")
m3@fullcoef
 mu sigma 64.122747 0.972409
 m3@min
[1] 123.8364
\end{verbatim}

The function \texttt{mle()} creates S4 rather S3 objects so to obtain its components use the \texttt{splot} operator @ rather the conventional $ reserved for S3 objects, see Venables and Ripley [2000] or...
Chambers [2008] for the definition of S3 and S4 objects in R. Note that in the case of scenario I leaving the function \texttt{mle()} with its default value for method resulted in an MLE value of \( \hat{\mu} = 64.061 \) rather than the actual 64.125. This happens probably because the algorithm finished prematurely and highlights the problem that numerical methods occasionally need fine tuning.

Also note, from the above output, that for scenario I, \( -\ell(\hat{\mu}) = -\ell(64.125) = 123.86 \), hence the maximum log likelihood for the exponential distribution model is \(-123.86\). For scenario II, \( -\ell(\hat{\mu}, \hat{\sigma}) = -\ell(64.122, 0.972) = 123.83 \), so the maximum log likelihood of the gamma model is \(-123.83\). Hence here there is a very small difference in log-likelihood between the two models.

- how we will introduce the different parametrization of GAMLSS??

### 7.5 Statistical properties of MLE when the model is correct

Assume \( Y = (Y_1, Y_2, ..., Y_n) \) is a random sample of independently identically distributed random variables with probability (density) function \( f_P(y) \).

A parametric model family of probability (density) functions is given by \( f_Y(y|\theta) \) for a range of values of \( \theta \), where \( f_Y(y|\theta) \) is a known function, except for parameters \( \theta \).

Assume the model is correct, i.e. assume that \( f_P(y) \) belongs to the model family and \( f_P(y) = f_Y(y|\theta_T) \) for all \( y \) for a particular value \( \theta_T \) of \( \theta \), where \( \theta_T \) is the true value of \( \theta \).

Let \( \hat{\theta} \) be the maximum likelihood estimator of \( \theta \) from true model pdf \( f_Y(y|\theta) \) given the random sample \( Y \).

There are three basic properties of the MLE \( \hat{\theta} \), assuming certain conditions hold.

- **invariance**
- **consistency**
- **asymptotic normality**.

#### 7.5.1 Invariance

Invariance means that if \( \phi \) is a one to one transformation of parameter \( \theta \) say \( \phi = g(\theta) \), then the maximum likelihood estimators \( \hat{\phi} \) and \( \hat{\theta} \) are related by \( \hat{\phi} = g(\hat{\theta}) \). So it does not matter which parametrization is used, at the point of maximum the likelihood will be the same. Also the transformation from one parametrization to another does not affect the estimates. Note though that, while the MLE’s are invariant, their standard errors are not since they depend on the curvature of the likelihood function at the point of maximum (see 7.5.3). Quadratic shapes of the likelihood at the maximum give more precise standard error based confidence intervals and Wald tests for \( \theta_T \) (see section 7.5.3).

#### 7.5.2 Consistency

The maximum likelihood estimator \( \hat{\theta} \) is, under certain conditions, a (weakly) consistent estimator of the true parameter \( \theta_T \), i.e. \( \hat{\theta} \) converges in probability to \( \theta_T \) as \( n \to \infty \).

This means that for all \( \varepsilon > 0 \),

\[
\lim_{n \to \infty} p(|\hat{\theta} - \theta_T| > \varepsilon) = 0.
\]

(7.19)
7.5. STATISTICAL PROPERTIES OF MLE WHEN THE MODEL IS CORRECT

Sufficient conditions for weak consistency 7.19 to hold are given by Newey and McFadden (1994) Theorem 2.5. A derivation of and sufficient conditions for strong consistency (which implies weak consistency (7.19) is given by Wald (1949), and with less restrictive conditions by White (1982).

7.5.3 Asymptotic normality

Convergence in distribution

First we will define convergence in distribution. Let \((U_n; n = 1, 2, \ldots)\) be a sequence of random variables and \(U\) another random variable, with cumulative distribution functions \(F_{U_n}(u)\) for \(n = 1, 2, \ldots\) and \(F_U(u)\) respectively, then \(U_n\) converges in distribution to \(U\) as \(n \to \infty\), written \(U_n \xrightarrow{d} U\), means \(\lim_{n \to \infty} F_{U_n}(u) = F_U(u)\) for all continuity points of \(F_U(u)\).

Asymptotic distribution of MLE

Under certain conditions, \(\sqrt{n}(\hat{\theta} - \theta_T)\) converges in distribution to \(N_K(0, J(\theta_T)^{-1})\) as \(n \to \infty\), i.e.

\[
\sqrt{n}(\hat{\theta} - \theta_T) \xrightarrow{d} N_K(0, J(\theta_T)^{-1}),
\]

where \(J(\theta_T)\) is the (Fisher) expected information matrix for a single observation \(Y_i\), evaluated at \(\theta_T\), given by

\[
J(\theta_T) = -\mathbb{E}_{f_P} \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^T} \right]_{\theta_T}
\]

where \(\ell_i(\theta) = \log f_Y(Y_i | \theta)\). Note that expectation in equation (7.21) is taken over the true population distribution \(f_P(y_i) = f_Y(y_i | \theta_T)\) for \(Y_i\). Note also that the subscript \(\theta_T\) in (7.21) means that the quantity in square brackets is evaluated at \(\theta = \theta_T\).

An outline of the derivation of (7.20) is given in Appendix 7.9.2, Ripley [1996] page 32 or Claeskens and Hjort (2008) page 26-27. A more rigorous derivation of (7.20) with sufficient conditions is given by Cramer (1946) and with less restrictive conditions by White (1982). Sufficient conditions for (7.20) are also given by Newey and McFaddan (1994) Theorem 3.3.

Note also (7.20) should be interpreted in terms of the limit of probabilities associated with \(\sqrt{n}(\hat{\theta} - \theta_T)\) [and not in terms of the limit of moments of \(\sqrt{n}(\hat{\theta} - \theta_T)\)]. For example the mean (or variance) of \(\sqrt{n}(\hat{\theta} - \theta_T)\) does not necessarily converge to the mean (or variance) of the asymptotic distribution.

Informally, asymptotically as \(n \to \infty\),

\[
\hat{\theta} \sim N_K(\theta_T, n^{-1}J(\theta_T)^{-1}) = N_K(\theta_T, i(\theta_T)^{-1})
\]

where

\[
i(\theta_T) = -\mathbb{E}_{f_P} \left[ \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} \right]_{\theta_T} = \sum_{i=1}^{n} \mathbb{E}_{f_P} \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^T} \right]_{\theta_T} = -n \mathbb{E}_{f_P} \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^T} \right]_{\theta_T} = n J(\theta_T)
\]

is the (Fisher) expected information matrix of the \(n\) iid random variables \(Y\), evaluated at \(\theta_T\), where here \(\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta) = \sum_{i=1}^{n} \log f_Y(Y_i | \theta)\).
**Asymptotic efficiency**

For a single parameter $\theta$, the maximum likelihood estimator $\hat{\theta}$ of $\theta_T$ is asymptotically a more efficient estimator of $\theta_T$ than a wide class of alternative estimators. This means that for their asymptotic distributions, the ratio of the mean square error of the alternative estimator of $\theta_T$ to that of the MLE is greater than or equal to 1.

**Approximating the expected information matrix**

The expected information $i(\theta_T)$ is not always easy to derive analytically, therefore the observed information $I(\theta_T)$ is often used instead. The observed information $I(\theta_T)$ evaluated at $\theta = \theta_T$ is defined as

$$I(\theta_T) = -\sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^\top} |_{\theta = \theta_T},$$

where $\ell_i(\theta) = \log f_Y(y_i | \theta)$. Note $I(\theta_T)$ is equal to the negative of the Hessian matrix of the log likelihood function at $\theta_T$. The variance of the asymptotic distribution of $\hat{\theta}$ is then approximated by $I(\theta_T)^{-1}$ instead of $i(\theta_T)^{-1}$.

Of course the point, $\theta_T$ is unknown, so $\theta_T$ is estimated by $\hat{\theta}$ in both the expected $i(\theta_T)$ and observed $I(\theta_T)$ information, giving $i(\hat{\theta})$ and $I(\hat{\theta})$.

The gamlss summary command output uses the following approximate distribution for $\hat{\theta}$, for large $n$,

$$\hat{\theta} \sim N_K(\theta_T, I(\hat{\theta})^{-1}).$$

Let $\theta_T = (\theta_{T1}, \theta_{T2}, ..., \theta_{TK})^\top$ and $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K)^\top$. Hence, for $k = 1, 2, ..., K$, the estimated standard error, $\text{se}(\hat{\theta}_k) = \left[\hat{V}(\hat{\theta}_k)\right]^{1/2}$, of $\hat{\theta}_k$, equal to the square root of the estimated variance, $\hat{V}(\hat{\theta}_k)$, of $\hat{\theta}_k$, is calculated from the square root of the $k^{th}$ diagonal element of $I(\hat{\theta})^{-1}$.

**Standard error based confidence interval for a parameter**

A (standard error based) approximate $100(1 - \alpha)\%$ confidence interval for a single parameter $\theta$, e.g. $\theta_{Tk}$, is given by $\left[\hat{\theta} \pm z_{\alpha/2} \text{se}(\hat{\theta})\right]$, where $z_{\alpha/2}$ is the upper tail value of a standard normal NO(0,1) distribution corresponding to upper tail probability $\alpha/2$.

The standard error based confidence interval for $\theta$ is often much less accurate than the profile confidence interval for $\theta$ given in section 7.6.1.

**Standard error based Wald test for the value of a parameter**

A 100$\alpha$% significance level (standard error based) Wald test of $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$ is based on the Wald test statistic

$$Z = \frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})} \sim NO(0,1)$$

asymptotically as $n \to \infty$, if $H_0$ is true. Hence reject $H_0$ if the observed

$$|z| = \left|\frac{\hat{\theta} - \theta_0}{\text{se}(\hat{\theta})}\right| > z_{\alpha/2}.$$
7.5. STATISTICAL PROPERTIES OF MLE WHEN THE MODEL IS CORRECT

The Wald test for $\theta$ above is often much less accurate than the generalized likelihood ratio test for $\theta$ given in section 7.7.1.

Approximating the log likelihood function by a quadratic function

The log likelihood function can be approximated at the point of maximum $\hat{\theta}$ by a quadratic function:

$$
\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^\top \left[ \frac{\partial^2 \ell}{\partial \theta \partial \theta^\top} \right]_{\hat{\theta}} (\theta - \hat{\theta})
$$

because the second term is zero and where subscript $\hat{\theta}$ above means the quantities inside the square brackets are evaluated at $\theta = \hat{\theta}$. To demonstrate this the exponential distribution model for the aircond data is used:

```r
logl <- function(mu) -sum(dEXP(aircond, mu=mu, log=TRUE)) # log-likelihood
mu <- seq(40, 112, length=101) # mu
logL <- rep(0,101)
for (i in 1:101) logL[i] <- -logl(mu[i]) # getting the likelihood
mm <- optim(45, logl, method="L-BFGS-B", lower=0.01, upper=Inf)# optimise
hess <- optimHess(mm$par, logl) # the Hessian
qr <- -mm$value - 0.5*hess*(x-mm$par)^2 # quadratic approximation
plot(logL~x, type="l", xlab= expression(paste( mu)), ylab="log Likelihood")
lines(qr~x, col=2, lty=2)
lines(c(mm$par,mm$par), c(min(logL), -mm$value), lty=3, col=4)
points(mm$par, -mm$value, col=4)
```

From the above code note how the matrix $\left[ \frac{\partial^2 \ell}{\partial \theta \partial \theta^\top} \right]_{\hat{\theta}}$ is obtained using the function `optimHess()`. Figure 7.8 shows the log likelihood function of the data for the parameter $\mu$ together with the quadratic approximation of the log likelihood. It can be seen that the approximation is very good close to the MLE $\hat{\mu}$, but because of the 'skewness' of the log likelihood function it becomes less accurate for points away from the maximum. There are two important points to be made here. The first is that a quadratic shape log likelihood function is associated with the normal distribution. That is, the log likelihood function for $\mu$ of a normal distribution given any value of $\sigma$ is quadratic. The second point has to do with what happens to the shape of the log-likelihood as the number of observations in the sample increases. The log likelihood becomes closer to a quadratic shape as the sample size increases corresponding to the normal asymptotic distribution of the MLE. The shape of the log likelihood for a finite sample size mainly depends on how the probability (density) function is parameterized in the first place. As a general rule the closer the log likelihood is to a quadratic shape the better, first because the search for the maximum is easier, but also because the standard errors of the estimates are more accurate.

7.5.4 Air conditioning example continued: se based CI and Wald test

For the `aircond` example using the exponential distribution, then, from (7.15), the observed information at $\mu = \mu_T$ is

$$
I(\mu_T) = - \left[ \frac{d^2 \ell(\mu)}{d\mu^2} \right]_{\mu_T} = 2 \sum_{i=1}^n \frac{y_i}{\mu_T^2} - \frac{n}{\mu_T^2}
$$
The (Fisher’s) expected information matrix evaluated at \( \mu = \mu_T \) is

\[
i(\mu_T) = -E \left[ \frac{d^2 \ell(\mu)}{d\mu^2} \right]_{\mu_T} = 2\sum_{i=1}^{n} E \left[ \frac{1}{\mu_T^3} \frac{Y_i}{\mu_T^2} \right] - \frac{n}{\mu_T^2}
\]

\[
= 2\frac{n\mu_T}{\mu_T^2} - \frac{n}{\mu_T^2} \\
= \frac{n}{\mu_T^2}.
\]

where, in (7.24), \( \ell(\mu) \) is treated as a function of the random sample of iid random variables \( Y \), rather than the observed sample \( y \).

Hence \( J(\mu_T) = i(\mu_T)/n = 1/\mu_T^2 \) and as \( n \to \infty \)

\[
\sqrt{n}(\hat{\mu} - \mu_T) \overset{d}{\to} N \left( 0, J(\mu_T)^{-1} \right) = N \left( 0, \frac{\mu_T^2}{n} \right) \approx N \left( 0, \frac{\hat{\mu}^2}{n} \right).
\]

and, informally, asymptotically as \( n \to \infty \)

\[
\hat{\mu} \sim N \left( \mu_T, i(\mu_T)^{-1} \right) = N \left( \mu_T, \frac{\mu_T^2}{n} \right) \approx N \left( \mu_T, \frac{\hat{\mu}^2}{n} \right).
\]

Using R we obtain:

```r
> # estimated observed information (calculated explicitly)
> 2*sum(aircond)/mean(aircond)^3-(length(aircond)/mean(aircond)^2)
[1] 0.005836554
> # estimated observed information (calculated numerically)
> optimHess(mm$par, logl)
```


7.5. STATISTICAL PROPERTIES OF MLE WHEN THE MODEL IS CORRECT

\[
\begin{bmatrix}
0.005836558
\end{bmatrix}
\]

> # estimated expected information matrix
> (length(aircond)/mean(aircond)^2)

All values are very similar. Note how the numerical Hessian of negative of the log likelihood is calculated using the standard R function `optimHess()`. The estimated standard error for \( \hat{\mu} \) is given in this case by

\[
se(\hat{\mu}) = (i(\hat{\mu}))^{-1/2} = \hat{\mu}/\sqrt{n} = \sqrt{1/0.005836554} = 13.0895.
\]

An approximate 95% confidence interval for \( \mu_T \) is

\[
[\hat{\mu} \pm (1.96 \times se(\hat{\mu}))] = [64.125 \pm (1.96 \times 13.0895)] = [38.47, 89.78].
\]

Inference about a parameter in `gamlss`

The following R code show how the `gamlss()` function can be used to obtain confidence intervals for \( \mu \) (=\( \mu_T \)). Note however that the default link for \( \mu \) in the exponential `gamlss` family, `EXP`, is log. What this means, is that the parameter fitted in the predictor model is not \( \mu \) but \( \log(\mu) \) so the corresponding confidence interval is for \( \log(\mu) \). By exponentiating the resulting confidence interval for \( \log(\mu) \) we can obtain a confidence interval for \( \mu \).

For parameters defined in the range from zero to infinity (as \( \mu \) in the exponential example above) modeling the log of the parameter and constructing a confidence interval on the log scale and then transforming it generally produces more reliable confidence intervals.

> # fitting the model
> m1 <- gamlss(aircond~1, family=EXP)

GAMLSS-RS iteration 1: Global Deviance = 247.72
GAMLSS-RS iteration 2: Global Deviance = 247.72

> summary(m1)

Family: c("EXP", "Exponential")
Call: gamlss(formula = aircond ~ 1, family = EXP)
Fitting method: RS()

Mu link function: log
Mu Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| 4.161e+00 | 2.041e-01 | 2.038e+01 | 3.194e-16 |

No. of observations in the fit: 24
Degrees of Freedom for the fit: 1
Residual Deg. of Freedom: 23
at cycle: 2

Global Deviance: 247.72
AIC: 249.72
SBC: 250.8981

> # the estimate of log mu
> coef(m1)
(Intercept) 4.160834

> # the estimate for mu
> fitted(m1, "mu")[1]
1 64.125

> # the standard error for log mu
> vcov(m1, "se")
(Intercept) 0.2041241

> # 95% CI for log mu
> confint(m1)
2.5 % 97.5 %
(Intercept) 3.760758 4.56091

> # 95% CI for mu
> exp(confint(m1))
2.5 % 97.5 %
(Intercept) 42.98101 95.67052

Standard error based confidence interval for a parameter

The 95% approximate confidence interval for $\beta_T = \log \mu_T$ is given by $[\hat{\beta} \pm (1.96 * se(\hat{\beta}))] = [4.161 \pm (1.96 * 0.2041)] = (3.76, 4.56)$ and is given by confint(m1). Note that (3.76, 4.56), the 95% confidence interval (CI) for $\beta_T = \log \mu_T$, is symmetrical about the estimate $\hat{\beta} = 4.161$. However $(42.98, 95.67) = (\exp(3.76), \exp(4.56))$, the transformed 95% CI for $\mu$ is not symmetrical about the estimate $\hat{\mu} = \exp(\hat{\beta}) = 64.125$. [Note also that the resulting 95% CI $(42.98, 95.67)$ for $\mu$ is very different from and probably more reliable than the 95% for $\mu$ given earlier (38.47, 89.78).]

The following important points are needed here.

- Confidence intervals based directly on standard errors are symmetrical about the fitted parameters. This may not be a good idea, if the likelihood for the parameter is far from a quadratic shape, because the resulting confidence intervals are not reliable (i.e. their coverage, the % of confidence intervals that capture the true parameter value, may be far from the nominal % of the CI).

- Profile confidence intervals (see section 7.6.1) generally produce more reliable confidence intervals.
7.6. ELIMINATING NUISANCE PARAMETERS USING THE PROFILE LOG LIKELIHOOD

Standard error based Wald test for a parameter

A 100\% significance level (standard error based) Wald test \( H_0 : \beta_T = \beta_0 \) against \( H_1 : \beta_T \neq \beta_0 \) is based on the Wald test statistic

\[
Z = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} \sim NO(0, 1),
\]

asymptotically as \( n \to \infty \), if \( H_0 \) is true.

Hence reject \( H_0 \) at the 100\% significance level if the observed

\[
| z | = \left| \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} \right| > z_{\alpha/2}.
\]

For example for a 5\% significance level (standard error based) Wald test of \( H_0 : \mu_T = 100 \) against \( H_1 : \mu_T \neq 100 \), equivalent to \( H_0 : \beta_T = \log(100) \) against \( H_1 : \beta_T \neq \log(100) \), then reject \( H_0 \) since the observed

\[
| z | = \frac{4.161 - 4.605}{0.2041} = 2.175 > z_{0.025} = 1.96,
\]

where \( \hat{\beta} = 4.161 \) and \( se(\hat{\beta}) = 0.2041 \) are given by summary(m1) above and \( \beta_0 = \log(100) = 4.605 \).

Alternatively calculate the corresponding approximate p-value where \( p = P(|Z| > |z|) = P(|Z| > 2.175) = 0.0296 \leq 0.05 \), so reject \( H_0 \) at the 5\% significance level.

Testing the value of a parameter by using the generalized likelihood ratio test explained in section 7.7.1 is more reliable.

7.6 Eliminating nuisance parameters using the profile log likelihood

The problem of eliminating nuisance parameters is important in practice.

Let \( \theta = (\theta_1, \theta_2) \) be the set of parameters of the model. Let us assume that we are interested in parameters \( \theta_1 \), since for example they answer the scientific question we are looking at. The question is how we can eliminate \( \theta_2 \) now called the nuisance parameter and concentrate only on inference about \( \theta_1 \), the parameter(s) of interest. One answer to the question is to use profile log likelihood for inference about \( \theta_1 \).

7.6.1 Profile log likelihood function and profile confidence intervals

Let \( \ell(\theta) = \ell(\theta_1, \theta_2) \) be the likelihood function for parameters \( \theta = (\theta_1, \theta_2) \). The profile log likelihood function \( p\ell(\theta_1) \) for parameters \( \theta_1 \) is given by maximizing \( \ell(\theta_1, \theta_2) \) over \( \theta_2 \) for each value of \( \theta_1 \), i.e.

\[
p\ell(\theta_1) = \max_{\theta_2} \ell(\theta_1, \theta_2)
\]

for each value of \( \theta_1 \).

For a single parameter \( \theta = \theta_1 \), \( p\ell(\theta) = \max_{\theta_2} \ell(\theta, \theta_2) \), and \( p\ell(\theta) \) can be plotted against \( \theta \). Alternatively the profile Global Deviance \( pGD(\theta) = -2 * p\ell(\theta) \) can be plotted against \( \theta \).
A 100α% profile confidence interval for a single parameter θ includes all values θ₀ for which 
\( pGD(θ₀) < pGD(\hat{θ}) + \chi^2_{1,α} \) where \( \hat{θ} \) maximizes \( p(θ) \) over θ and hence minimizes \( pGD(θ) \) over θ and \( \chi^2_{1,α} \) is the upper tail value of a Chi-squared distribution with one degree of freedom with upper tail probability α%.

This 100α% profile confidence interval for a single parameter θ includes all values θ₀ of parameter θ that are accepted by a generalized likelihood ratio test (see section 7.7.1) of \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \) at the 100α% significance level.

This equals all values θ₀ for which \( GD₀ < GD₁ + \chi^2_{1,α} \), where \( GD₀ \) and \( GD₁ \) are the fitted global deviance = \( -2 \times \max(\log \text{likelihood}) \) under hypotheses \( H₀ \) and \( H₁ \) respectively, i.e.

\[
GD₀ = -2 \times \max_{\theta_2} [\ell(θ₀, θ_2)] = -2 \times p(θ₀) = pGD(θ₀)
\]

and

\[
GD₁ = -2 \times \max_{\theta_2} [\ell(θ, θ_2)] = -2 \times \max_{\theta} \left[ \max_{\theta_2} [\ell(θ, θ_2)] \right] = -2 \times \max_{\theta} p(θ) = -2 \times p(\hat{θ}) = pGD(\hat{θ}).
\]

Hence Λ = \( GD₀ - GD₁ = pGD(θ₀) - pGD(\hat{θ}) \sim \chi^2_1 \) approximately if \( H₀ \) is true. Hence \( H₀ : \theta = \theta_0 \) is accepted if \( GD₀ < GD₁ + \chi^2_{1,α} \), i.e if \( pGD(θ₀) < pGD(\hat{θ}) + \chi^2_{1,α} \).

### 7.6.2 Air conditioning example continued: profile confidence intervals

In the air conditioning example from section 7.2.3, assume the data are a random sample from a gamma distribution, \( GA(μ, σ) \). The profile likelihood \( p(μ) \) for μ is given by

\[
p(μ) = \max_{σ} \ell(μ, σ)
\]

for each value of μ. Similarly the profile likelihood \( p(σ) \) for σ is given by

\[
p(σ) = \max_{μ} \ell(μ, σ)
\]

for each value of σ.

The bottom left and right corners of Figure 7.5 show what the two dimensional likelihood looks like from the direction of the parameters \( \mu \) (theta1) and \( σ \) (theta2) respectively. They show the profile likelihood for each of the two parameters \( \mu \) and \( σ \) respectively.

The profile global deviance for \( μ \) is \( pGD(μ) = -2 \times p(μ) \) The profile global deviance plot of \( pGD(μ) \) against \( μ \) can be obtained in GAMLSS by:

\[
m2 <- \text{gamlss(aircond~1, family=GA)}
m2A <- \text{prof.dev(m2, "mu", min=30, max=120, step=0.1, type="l")}
\]

giving Figure 7.9 and following output

The Maximum Likelihood estimator is 64.125
with a Global Deviance equal to 247.6728
A 95\% Confidence interval is: ( 44.01499 , 98.59832 ).

The 95\% confidence interval for \( μ = μ_T \) from the profile deviance includes all values \( μ₀ \) that are accepted by a generalized likelihood ratio test (see section 7.7.1) of \( H₀ : μ = μ₀ \) against \( H₁ : μ \neq μ₀ \) at the 5% significance level.
The maximum likelihood estimate $\hat{\mu} = 64.125$ of $\mu$ corresponds to the minimum global deviance. The vertical dotted lines in Figure 7.9 mark the profile confidence interval for $\mu$, i.e. $(44.015, 98.60)$, given by all $\mu$ for which $pGD(\mu) < \min_\mu pGD(\mu) + 3.84$, since $\chi^2_{1,0.05} = 3.84$.

The profile global deviance for $\sigma$ is $pGD(\sigma) = -2 \cdot p\ell(\sigma)$ The profile global deviance plot of $pGD(\sigma)$ against $\sigma$ can be obtained in GAMLSS by:

```r
m2B <- prof.dev(m2,"sigma",min=0.6,max=1.7,step=0.01,type="l")
giving Figure 7.10 and the following output
```

The Maximum Likelihood estimator is 0.9724419
with a Global Deviance equal to 247.6728
A 95\% Confidence interval is: ( 0.7697473 , 1.271375 ).

The profile log likelihood function and profile confidence interval for the predictor log $\mu$ can be obtained using the prof.term command:

```r
x<-rep(1,length(aircond))
m2C <- quote(gamlss(aircond~1+offset(this*x), family=GA))
prof.term(m2C,min=3.6,max=4.8,step=0.01,start.prev=FALSE)
giving output
```

The Maximum Likelihood estimator is 4.160839
with a Global Deviance equal to 247.6728
A 95 \% Confidence interval is: ( 3.78453 , 4.591051 ).
Note that the model for the predictor $\log(\mu)$ in the `gamlss` function above is `-1+offset(this*x)' , where the `-1' removes fitting the constant in the predictor model (which is otherwise included by default), while `offset(this*x)' offsets constant `this' in the model, i.e. fits a predictor model with constant exactly equal to `this'. The profile log likelihood function is plotted against parameter `this', i.e. $\beta = \log(\mu)$. [Note in regression models x is an explanatory variable vector and `this' is the $\beta$ coefficient of x.]

Similarly the profile log likelihood function and profile confidence interval for the predictor $\log \sigma$ can be obtained by

```r
m2D <- quote(gamlss(aircond~1, sigma.fo=-1+offset(this*x), family=GA))
prof.term(m2D,min=-0.4,max=0.4,step=0.01,start.prev=FALSE)
```

giving output

The Maximum Likelihood estimator is -0.02794343
with a Global Deviance equal to 247.6728
A 95\% Confidence interval is: ( -0.2616835 , 0.2401039 ).

### 7.7 Model selection

#### 7.7.1 Testing between nested models using the generalized likelihood ratio test

Choosing between models, (we had called them scenarios in Section 7.2.3 is important because the GAMLSS models are flexible and therefore allow different possible scenarios to given single data set to be tried. We should be able to choose between those scenarios in a consistent way.
Global deviance
We shall refer to the quantity
\[
GD = -2 \log L(\hat{\theta}) = -2\ell(\hat{\theta})
\]
as the *Global Deviance* (or GD). It is a very important quantity for testing between different models.

generalized likelihood ratio test
Let \( M_0 \) and \( M_1 \) be two different models. Model \( M_0 \) is called *nested* within \( M_1 \) if it a special case of \( M_1 \). In this case \( M_0 \) is the simpler model while \( M_1 \) the more complicated one.

Two nested parametric GAMLSS models, \( H_0 : M_0 \) and \( H_1 : M_1 \), where \( M_1 \) is a submodel of \( M_0 \), with fitted global deviances \( GD_0 \) and \( GD_1 \) and error degrees of freedom \( df_{e0} \) and \( df_{e1} \) respectively may be compared using the (generalised likelihood ratio) test statistic \( \Lambda = GD_0 - GD_1 \) which has an asymptotic Chi-squared distribution under \( M_0 \), with degrees of freedom \( d = df_{e0} - df_{e1} \), (given that the usual conditions are satisfied). For each model \( M \) the error degrees of freedom \( df_e \) is defined by
\[
df_e = n - p,
\]
where \( p \) is the number of parameters in model \( M \). Hence, at the 100\( \alpha \)% significance level, reject \( H_0 \) if \( \Lambda \geq \chi^2_{d,\alpha} \), i.e. \( GD_0 \geq GD_1 + \chi^2_{d,\alpha} \), and accept \( H_0 \) if \( GD_0 < GD_1 + \chi^2_{d,\alpha} \).

### 7.7.2 Air conditioning example continued: GLR test
Consider the aircond data, here we are interested whether the exponential or the gamma distribution is appropriate for the data. Note that for \( \sigma = 1 \) the gamma reverts to the exponential, so the model with the exponential distribution is a submodel of the gamma. The null hypothesis is \( H_0 : \sigma = 1 \) against the alternative \( H_1 : \sigma \neq 1 \). Here we fit the exponential (\( H_0 \)) and gamma (\( H_1 \)) distributions and check the difference in global deviance against a chi-square distribution with one degree of freedom.

```r
> m1 <- gamlss(aircond~1, family=EXP)
GAMLSS-RS iteration 1: Global Deviance = 247.72
GAMLSS-RS iteration 2: Global Deviance = 247.72
> m2 <- gamlss(aircond~1, family=GA)
GAMLSS-RS iteration 1: Global Deviance = 247.6728
GAMLSS-RS iteration 2: Global Deviance = 247.6728

Test \( H_0 : \sigma = 1 \) (i.e. exponential model) \( H_1 : \sigma \neq 1 \) (i.e. gamma model)

Observed \( \Lambda = GD_0 - GD_1 = 247.72 - 247.6728 = 0.0472 \), where approximately \( \Lambda \sim \chi^2_1 \) if \( H_0 \) is true.

Since 0.0472 < \( \chi^2_{1,0.05} = 3.84 \), we accept \( H_0 \) : the exponential model at the 5% significance level.

Alternatively the p-value is \( p = P(\Lambda > 0.0472) = 1 - \text{pchisq}(0.0472, 1) = 0.828 > 0.05 \), so we accept the exponential model null hypothesis at the 5% significance level.

### 7.7.3 Model selection using the generalised Akaike information criterion
For comparing non-nested GAMLSS models, to penalize over-fitting the generalised Akaike Information Criterion (GAIC), Akaike [1983], can be used. This is obtained by adding to the
fitted global deviance a fixed penalty $k$ for each effective degree of freedom used in a model, i.e.

$$
\text{GAIC}(k) = \text{GD} + (k \times df),
$$

where $df$ denotes the total effective degrees of freedom used in the model and GD is the fitted global deviance. The model with the smallest value of the criterion GAIC($k$) is then selected. The Akaike information criterion (AIC), Akaike [1974], and the Schwartz Bayesian criterion (SBC), Schwarz [1978], are special cases of the GAIC($k$) criterion corresponding to $k = 2$ and $k = \log(n)$ respectively:

$$
\text{AIC} = \text{GD} + (2 \times df),
$$

$$
\text{SBC} = \text{GD} + (\log(n) \times df).
$$

The two criteria, AIC and SBC, are asymptotically justified as predicting the degree of fit in a new data set, i.e. approximations to the average predictive error. Justification for the use of SBC comes also as a crude approximation to Bayes factors, Raftery [1996, 1999]. In practice it is usually found that while the original AIC is very generous in model selection the SBC is too restrictive. Our experience is that a value of the penalty $k$ in the range $2.5 \leq k \leq 3$ works well for most data. Kin and Gu (2004) suggested using $k \approx 2.8$. Using GAIC($k$) allows different penalties $k$ to be tried for different modelling purposes. The sensitivity of the selected model to the choice of $k$ can also be investigated. A selection of different values of $k$ e.g. $k = 2, 2.5, 3, 3.5, 4$ could be used in turn to investigate the sensitivity or robustness of the model selection to the choice of the value of the penalty $k$. Claeskens and Hjort (2003) consider a focused information criterion (FIC) in which the criterion for model selection depends on the objective of the study, in particular on the specific parameter of interest.

### 7.8 Statistical properties of MLE when the model is mis-specified

#### 7.8.1 Graphical representation

In order to understand the concepts involved we shall use the schematic presentation of the population, sample and model in Figure 7.11. Figure 7.11 (similar to Figure 7.2) represents the population and the sample as points and the model as a line. In addition this figure shows two “directed” lines one from the population and the other from the sample to the model line respectively. The directed lines represent a form of minimal distance from the points to the line. The point $\theta_c$ on the model line represents the value of $\theta$ which is closest to the population distribution as measured by the Kullback-Liebler distance

$$
d[f_P(y), f_Y(y|\theta)] = \int [\log f_P(y) - \log f_Y(y|\theta)] f_P(y) dy,
$$

i.e. $\theta_c$ minimizes the Kullback-Liebler distance (7.25) over $\theta$.

Here is the confusion: traditional statistical books refer to $\theta_c$ as the ’true’ $\theta$ parameter value. We would like to emphasize that, when the model is mis-specified, this is, in general, incorrect. The model represents an assumption made about the population. A different model assumption will generate a different line and its equivalent point ’closest’ to the true population. We will refer to $\theta_c$ as the ’closest’ value for $\theta$ under the model $f_Y(y|\theta)$ to emphasize that $f_Y(y|\theta)$ is just one model among other possible ones.
7.8. STATISTICAL PROPERTIES OF MLE WHEN THE MODEL IS MIS-SPECIFIED

7.8.2 Properties of MLE under model mis-specification

Assume \( Y = (Y_1, Y_2, ..., Y_n) \) is a random sample of independently identically distributed random variables with probability (density) function \( f_P(y) \).

A parametric model family of probability (density) functions is given by \( f_Y(y|\theta) \) for a range of values of \( \theta \), where \( f_Y(y|\theta) \) is a known function, except for parameters \( \theta \).

Assume the model is incorrect, i.e. assume that \( f_P(y) \) does not belong to the model family \( f_Y(y|\theta) \) for any value of \( \theta \).

Let \( \hat{\theta} \) be the maximum likelihood estimator of \( \theta \) from model \( f_Y(y|\theta) \) given the random sample \( Y \).

There are three basic classical properties of the maximum likelihood estimator \( \hat{\theta} \), assuming certain conditions hold. The properties are invariance, consistency and asymptotic normality. White (1982) gives the derivations of and sufficient conditions for the properties to hold.

Invariance

The invariance property of \( \hat{\theta} \) given in section 7.5.1 still holds, except the true value \( \theta_T \) of \( \theta \) is replaced by the 'closest' value \( \theta_c \).

Consistency

The consistency property of \( \hat{\theta} \) given in section 7.5.2 still holds, except the true value \( \theta_T \) of \( \theta \) is replaced by the 'closest' value \( \theta_c \).

There are situations in which \( \theta_c \) still represents a true population distribution measure. In particular for a mis-specified Exponential Family distribution model, \( EF(\mu, \sigma) \) with mean parameter \( \mu \) and scale parameter \( \sigma \), then \( \mu_c \), the 'closest' value of \( \mu \) to the true population distribution, is equal to the true population mean. This follows from (7.25), by substituting

\[
\log f_Y(y|\theta) = \log f_Y(y|\mu, \sigma) \quad \text{for an Exponential Family distribution}
\]

and differentiating with respect to \( \mu \) giving

\[
\frac{\partial}{\partial \mu} d [f_P(y), f_Y(y|\mu, \sigma)] = - \int \frac{y-\mu}{\sigma^2 v(\mu)} f_P(y) dy + \frac{1}{\sigma^2 v(\mu)} [E_{f_P}(Y) - \mu].
\]

Setting this equal to zero and solving for \( \mu \) gives \( \mu_c = E_{f_P}(Y) \).

Asymptotic normality

Under certain conditions, \( \sqrt{n}(\hat{\theta} - \theta_c) \) converges in distribution to \( N_K(0, J(\theta_c)^{-1}K(\theta_c)J(\theta_c)^{-1}) \) i.e.

\[
\sqrt{n}(\hat{\theta} - \theta_c) \xrightarrow{d} N_K(0, J(\theta_c)^{-1}K(\theta_c)J(\theta_c)^{-1}),
\]

(7.27)

where \( J(\theta_c) \) is the (Fisher) expected information matrix for a single observation \( Y_i \), evaluated at \( \theta_c \), given by
Figure 7.11: A schematic presentation of the $\theta_c$, the 'closest' or 'best' value for $\theta$ under the model $f_Y(y|\theta)$, and the $\hat{\theta}$ the MLE. $\theta_c$ is the value of $\theta$ 'closest' to the population using the Kullback-Liebler distance (or risk function), while $\hat{\theta}$ is the value of $\theta$ closest to the sample using the empirical risk function. The solid (red) line represents confidence bounds for the parameter $\theta_c$. 
\[ J(\theta_c) = -E_{f_Y} \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^T} \right]_{\theta_c} \]  

(7.28)

where \( \ell_i(\theta) = \log f_Y(Y_i|\theta) \) and \( K(\theta_c) \) is the variance of the first derivative of the log likelihood function for a single observation \( Y_i \), evaluated at \( \theta_c \), given by

\[ K(\theta_c) = V_{f_Y} \left[ \frac{\partial \ell_i(\theta)}{\partial \theta} \right]_{\theta_c}. \]  

(7.29)

Note that the expectation and variance in equations (7.28) and (7.29) respectively is taken over the true population distribution \( f_Y(y_i|\theta) \) for \( Y_i \).

An outline of the derivation of (7.27) is given in Appendix 7.9.2, Ripley [1996] page 32 or Claeskens and Hjort (2008) page 26-27. A more rigorous derivation of (7.27) with sufficient conditions is given by White (1982).

Equation (7.27) shows that the asymptotic variance of the MLE is a function of the both the expected information matrix and the variance-covariance matrix of the first derivative of the log likelihood function, both for a single observation \( Y_i \) and both evaluated at \( \theta_c \).

Hence, informally, asymptotically as \( n \to \infty \),

\[ \hat{\theta} \sim N_K(\theta_c, n^{-1}J(\theta_c)^{-1}K(\theta_c)J(\theta_c)^{-1}) \]  

(7.30)

[Note if the population distribution belongs to the model parametric family of distributions, then \( \theta_c = \theta_T \), \( K(\theta_T) = J(\theta_T) \) and the asymptotic variance-covariance matrix of \( \hat{\theta} \) is just \( i(\theta_T)^{-1} = n^{-1}J(\theta_T)^{-1} \) and so \( \sqrt{n}(\hat{\theta} - \theta_T) \overset{d}{\to} N_K(0, i(\theta_T)^{-1}) \).]

**Approximating the asymptotic variance-covariance matrix of \( \hat{\theta} \)**

The asymptotic variance-covariance matrix of \( \hat{\theta} \) is given by \( n^{-1}J(\theta_c)^{-1}K(\theta_c)J(\theta_c)^{-1} \) from (7.30).

The matrices \( J(\theta) \) and \( K(\theta) \) are unknown expected and variance-covariance matrices respectively, which can be approximated by the corresponding sample mean and variance-covariance matrices, i.e.

\[ \hat{J}(\theta_c) = -\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^T} \right]_{\theta_c}, \]

and

\[ \hat{K}(\theta_c) = \frac{1}{n-K} \sum_{i=1}^{n} \left[ \frac{\partial \ell_i(\theta)}{\partial \theta} \frac{\partial \ell_i(\theta)}{\partial \theta^T} \right]_{\theta_c}. \]

where here \( \ell_i(\theta) = \log f_Y(y_i|\theta) \).

Of course the point \( \theta_c \) is unknown, so \( \theta_c \) is estimated by \( \hat{\theta} \) giving \( \hat{J}(\hat{\theta}) \) and \( \hat{K}(\hat{\theta}) \).

The output from the gamlss summary command with option robust=TRUE or robust=T uses the following approximate distribution for \( \theta \), for large \( n \),

\[ \hat{\theta} \sim N_K(\theta_c, n^{-1}\hat{J}(\hat{\theta})^{-1}\hat{K}(\hat{\theta})\hat{J}(\hat{\theta})^{-1}). \]

The estimated variance-covariance matrix of \( \hat{\theta} \), given by \( n^{-1}\hat{J}(\hat{\theta})^{-1}\hat{K}(\hat{\theta})\hat{J}(\hat{\theta})^{-1} \), is called a ‘sandwich’ estimate.

The following questions is worth asking here
• What happens if we overfit with more parameters than necessary in the distribution \( f(y|\theta) \) (do the unnecessary parameters effect the results?)

• those standard errors are the robust standard errors (sandwich) Shall we use them all the time?

### 7.8.3 Robust confidence intervals and tests

Let \( \theta_c = (\theta_{c1}, \theta_{c2}, ..., \theta_{cK})^T \) and \( \hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_K)^T \), where \( \theta_c \) is the value of parameter \( \theta \) which makes the model distribution \( f_Y(y|\theta) \) 'closest' to the true population distribution \( f_P(y) \).

First note that the usual (standard error based) confidence intervals and Wald tests for an individual parameter \( \theta_{ck} \) are, in general, not valid when the model is mis-specified (because they use an incorrect estimated standard error for \( \hat{\theta}_k \)). Also note that profile likelihood confidence intervals and generalized likelihood ratio tests are also not valid when the model is mis-specified, and can not be adapted to the mis-specified model situation.

Robust (standard error based) confidence intervals and Wald tests should be used when there is model mis-specification, as described below.

For \( k = 1, 2, ..., K \), the robust estimated standard error, \( se(\hat{\theta}_k) = \left[ \hat{V}(\hat{\theta}_k) \right]^{1/2} \), of \( \hat{\theta}_k \), [equal to the square root of the estimated variance, \( \hat{V}(\hat{\theta}_k) \), of \( \hat{\theta}_k \)], is calculated from the square root of the \( k^{th} \) diagonal element of \( n^{-1}J(\hat{\theta})^{-1}K(\hat{\theta})J(\hat{\theta})^{-1} \).

**Robust (standard error based) confidence interval for a parameter**

A robust (standard error based) approximate \( 100(1 - \alpha)\% \) confidence interval for a single parameter \( \theta \), e.g. \( \theta_{ck} \), is given by \( \left[ \hat{\theta} \pm z_{\alpha/2}se(\hat{\theta}) \right] \), where \( se(\hat{\theta}) \) is the robust estimated standard error from above.

**Robust (standard error based) Wald test for the value of a parameter**

A \( 100\alpha\% \) significance level (standard error based) Wald test of \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \) is based on the Wald test statistic

\[
Z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \sim NO(0,1)
\]
asymptotically as \( n \to \infty \), if \( H_0 \) is true, where \( se(\hat{\theta}) \) is the robust estimated standard error from above. Hence reject \( H_0 \) if the observed

\[
| z | = \left| \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \right| > z_{\alpha/2}.
\]

### 7.8.4 Air conditioning example continued: robust CI and test

Here we consider again the air conditioning example. If the population distribution does not belong to the model parametric family of distributions, then robust confidence intervals and tests should be used.

Assume a mis-specified exponential distribution model. Then \( \frac{dE}{d\mu} \) is still given by (7.15). Hence \( J(\mu) = \frac{1}{n} i(\mu) = \mu^{-2} \) from section 7.5.4, and
7.8. STATISTICAL PROPERTIES OF MLE WHEN THE MODEL IS MIS-SPECIFIED

\[ K(\mu_c) = V_{fp} \left[ \frac{d^2}{d\mu} \right]_{\mu_c} = V_{fp} \left[ \frac{Y_i}{\mu^2} - \frac{1}{\mu^4} \right] = \frac{V_{fp}(y)}{\mu_c^4} \]

and hence \( J(\mu_c)^{-1} K(\mu_c) J(\mu_c)^{-1} = V_{fp}(y) \) giving

\[ \sqrt{n}(\hat{\mu} - \mu_c) \xrightarrow{d} N(0, V_{fp}(y)) \approx N \left( 0, s_y^2 \right) , \]

where \( s_y^2 \) is the sample variance of \( Y \).

Informally, asymptotically as \( n \to \infty \),

\[ \hat{\mu} \sim N \left( \mu_c, \frac{s_y^2}{n} \right) . \]

Hence if the population distribution does not belong to the parametric exponential distribution family of distributions, then the estimated standard error of \( \hat{\mu} \) is given by \( se(\hat{\mu}) = \frac{s_y}{\sqrt{n}} = 12.7888 \). This is usually called a robust standard error, i.e. robust to misspecification of the true population distribution. An approximate robust 95% confidence interval for \( \mu_c \) is given by

\[ \left[ \hat{\mu} \pm 1.96 \times \frac{s_y}{\sqrt{n}} \right] = [64.125 \pm 1.96 \times 12.7888] = (39.06, 89.19) . \]

Note that the exponential distribution is an Exponential Family distribution, and hence \( \mu_c = E_{fp}(y) \), so the above is a robust confidence interval for \( E_{fp}(y) \), the true population mean.

The following R code shows how the `gamlss()` function can be used to obtain robust confidence intervals for \( \mu = \mu_c \). Note however that the default link for \( \mu \) in the exponential `gamlss` family, `EXP`, is log. What this means is that the parameter fitted in the model is not \( \mu \) but \( \log(\mu) \) so the corresponding confidence interval is for \( \log \mu \). By exponentiating the resulting confidence interval for \( \log \mu \) we can obtain a confidence interval for \( \mu \).

For parameters defined in the range from zero to infinity (as \( \mu \) in the exponential example above) modeling the log of the parameter and constructing a confidence interval on the log scale and then transforming it generally produces more reliable confidence intervals.

> # fitting the model
> m1 <- gamlss(aircond~1, family=EXP)
> summary(m1, robust=T)

Regression diagnostics have been calculated with the robust option.

** Robust calculations **

- **Family:** c("EXP", "Exponential")
- **Call:** gamlss(formula = aircond ~ 1, family = EXP)
- **Mu link function:** log
Mu Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| 4.161e+00 | 1.994e-01 | 2.086e+01 | 1.922e-16 |

No. of observations in the fit: 24
Degrees of Freedom for the fit: 1
Residual Deg. of Freedom: 23

Global Deviance: 247.72
AIC: 249.72
SBC: 250.8981

> # the estimate of log mu
> coef(m1)
(Intercept)
  4.160834

> # the estimate for mu
> fitted(m1, "mu")[1]
1
  64.125

> # the standard errors for log mu
> vcov(m1, "se", robust=T)
(Intercept)
  0.1994367

> # robust 95\% CI for log mu
> confint(m1, robust=T)
(Intercept) 2.5 % 97.5 %
0.769946 4.551723

> # robust 95\% CI for mu
> exp(confint(m1, robust=T))
(Intercept) 2.5 % 97.5 %
43.37771 94.7956

Note that using the robust=T option above does not affect the parameter estimates. It only affects the standard errors of the parameter estimates. In this example the distribution of the times to breakdown of the air condition system are close to an exponential distribution and so using robust standard errors makes little difference.

The robust 95\% confidence interval for \( \log \mu_c = \log E_{f_P}(y) \) is calculated by \([4.161 \pm (1.96 \times 0.1994)] = (4.16 \pm 0.39) = (3.77, 4.55)\) and is given by \( \text{confint}(m1, \text{robust=T}) \). Hence the robust 95\% confidence interval for \( \mu_c = E_{f_P}(y) \) is given by \((43.38, 94.80)\).
7.9 Appendix of Chapter 3

7.9.1 Entropy, risk and empirical risk functions

In order to be able to evaluate the performance of the parametric model \( f_Y(y|\theta) \) relatively to the the “true” distribution \( f_P(Y) \) we need to define a measurement of “discrepancy” (or “risk”) between the two distributions. To do that the well established concept of \( \text{entropy} \) is used. Entropy was introduced by the physicist Willard Gibbs as a way to define the “mixed-upness” of a system. Larger entropy is associated with more chaotic (more messy) systems.

Shannon (1948) in his seminal paper “A Mathematical Theory of Communications” defined the entropy of a discrete probabilistic system as \( H = -\sum_{i} p_i \log(p_i) \) where \( p_1, p_2, \ldots, p_n \) are the probabilities of the events of the system and where \( \sum_{i} p_i = 1 \). Applying this definition to the “true” population distribution \( f_P(Y) \), we can define the entropy of the population as:

\[
H_P = \sum_{i=1}^{D} P_i \log(1/P_i) = -\sum_{i=1}^{D} P_i \log(P_i) \\
= \sum_{i=1}^{D} f_P(Y_i) \log \left( \frac{1}{f_P(Y_i)} \right) = -\sum_{i=1}^{D} f_P(Y_i) \log (f_P(Y_i))
\]

(7.31)

Note that Shannon’s entropy is defined in terms of probabilities only. It has its minimum, at zero, if and only if all probabilities but one are zero (so we know that an event will occur for sure). It has its maximum at \( \log(D) \) when all the probabilities are equal to \( 1/D \). In this later case, a complete uninformative (chaotic) statistical system exits where any guess is as good as any other. Shannon’s entropy, has also the property that, if \( P_1 \) and \( P_2 \) are two different probabilistic systems then \( H_{P_1,P_2} \leq H_{P_1} + H_{P_2} \). That is, the uncertainty of joint events is less than or equal to the sum of the individual uncertainties, (equal only if the events are independent).

Kullback and Liebler (1951) defined the ratio of two theoretical densities \( f_{Y_1}(y)/f_{Y_2}(y) \) as the information in \( y \) for discrimination between two hypothesis \( H_1 \) and \( H_2 \). They also defined the quantity \( d[f_{Y_1}(y), f_{Y_2}(y)] = \int f_{Y_1}(y) \log \frac{f_{Y_1}(y)}{f_{Y_2}(y)} dy \) as the mean information for discrimination between \( H_1 \) and \( H_2 \). The mean information, provides a “distance” type of measure for measuring how far the distribution \( f_{Y_1}(y) \) is from the distribution \( f_{Y_2}(y) \). It is not a proper distance in the mathematical sense, since do not have the symmetric property of a proper distance. That is, \( d[f_{Y_1}(y), f_{Y_2}(y)] \neq d[f_{Y_2}(y), f_{Y_1}(y)] \). Nevertheless \( d \) is ideal for the purpose of judging how far the model distribution, \( f_Y(y|\theta) \), is relatively to the “true” distribution, \( f_P(Y) \). Now since for a continuous random variables \( Y \) the model distribution is defined in the real line while the population distribution is by nature discrete, we will have to do some kind of adjustment to apply the Kullback and Liebler mean information to define a distance measure between population and model. Define now the model probability of observing the event \( Y \) as

\[
Pr(Y \in \Delta_I) = \int_{\Delta_I} f_Y(y|\theta) dy \simeq f_Y(Y|\theta) \Delta_I 
\]

(7.32)

for a value of \( Y \) in the small interval \( \Delta_I \). The expected loss of information or risk function for

---

A quantity which has this property is the difference in entropy between \( f_{Y_1}(y) \) and \( f_{Y_2}(y) \) which is defined as \( d[f_{Y_1}(y), f_{Y_2}(y)] = \int [f_{Y_1} - f_{Y_2}] \log \frac{f_{Y_1}(y)}{f_{Y_2}(y)} dy \).
choosing the model distribution \( f_Y(y|\theta) \) instead of the “true” distribution \( f_P(Y) \) is defined as

\[
R_P(\theta) = \sum_{i=1}^D P_i \log \left( \frac{P_i}{f_Y(Y_i|\theta)\Delta_i} \right) \\
= \sum_{i=1}^D f_P(Y_i) \log \left( \frac{f_P(Y_i)}{f_Y(Y_i|\theta)\Delta_i} \right) \\
= \sum_{i=1}^D f_P(Y_i) \log (f_P(Y_i)) - \sum_{i=1}^D f_P(Y_i) \log (f_Y(Y_i|\theta)\Delta_i) \tag{7.33}
\]

The name “risk” is justified if we define the quantity \( \log \left( \frac{P_i}{f_Y(Y_i|\theta)\Delta_i} \right) \) as our loss function. The risk function in (7.33) is a function of the unknown parameter \( \theta \). Among the values of \( \theta \) there is a value, say \( \theta_0 \) which minimises the risk function, i.e. \( R_P(\theta_0) = \min_\theta \{ R_P(\theta) \} \), (see Figure 7.6 for a schematic presentation). At this value the model distribution \( f_Y(y|\theta) \) approximates the population distribution best. [Note that minimising the risk in (7.33) is equivalent of minimising the quantity \( -\sum_{i=1}^D f_P(Y_i) \log (f_Y(Y_i|\theta)\Delta_i) \) since the quantity \( \sum_{i=1}^D f_P((0:10) \ast 2 \ast p_i/10i) \log (f_P(Y_i)) \) in equation (7.33), (which sometimes is called conditional entropy), does not involve \( \theta \).] This is the best that model \( f_Y(y|\theta) \) can do for modelling the “true” distribution. Note however that different parametric models can do possible better than \( f_Y(y|\theta_0) \). A good practitioner should be aware of this and other parametric distributions families apart from the specific \( f_Y(y|\theta) \) should be explored. Traditional statistical books refer to \( \theta_0 \) as the “true” \( \theta \) parameter value. We will restrain from doing that and we will refer instead to \( \theta_0 \) as the “best” value for \( \theta \) under the model \( f_Y(y|\theta) \) to emphasise that \( f_Y(y|\theta) \) is just one model among other possible ones.

The problem with the risk function as defined above is that it requires the knowledge of the “true” population distribution \( f_P(Y) \). The empirical risk function does not have the same problem since it its definition involves only known quantities. The empirical risk function is defined as

\[
R_E(\theta) = \sum_{i=1}^n p_i \log \left( \frac{p_i}{f_Y(y_i|\theta)\Delta_i} \right) \tag{7.34}
= \sum_{i=1}^n \frac{1}{n} \log \left( \frac{1}{n} f_Y(y_i|\theta)\Delta_i \right) \\
= \frac{1}{n} \sum_{i=1}^n \log \left( \frac{1}{n f_Y(y_i|\theta)\Delta_i} \right) \\
= -\log(n) - \frac{1}{n} \sum_{i=1}^n \log (f_Y(y_i|\theta)\Delta_i)
\]

In the empirical risk function we have replaced the unknown population distribution with the empirical distribution (using the plug-in principle). Among all possible values of \( \theta \) there is one denoted here as \( \hat{\theta} \) minimising the empirical risk function, (or equivalently maximising the quantity \( \ell(\theta) = \sum_{i=1}^n \log (f_Y(y_i|\theta)\Delta_i) \) which is the log-likelihood). This is shown schematically in Figure 7.11 where the model with \( \hat{\theta} \) is closer to the sample that any other model within the \( f_Y(y|\theta) \) family. The values denoted as \( \hat{\theta} \) in Figure 7.11, is the maximum likelihood estimator, MLE, of \( \theta \). Maximisation of the log-likelihood function as a method of finding an estimator in
parametric models was propose by R. A. Fisher in 1905. The specific value of $\hat{\theta}$ is generally different from $\theta_0$ but we expect $R_E(\theta) \rightarrow R_P(\theta)$ as $n \rightarrow \infty$ and therefore $\hat{\theta} \rightarrow \theta_0$ as $n \rightarrow \infty$.

### 7.9.2 Asymptotic normality of MLE under model mis-specification

The asymptotic normality property of the maximum likelihood estimator $\hat{\theta}$ under model mis-specification was given by equation (7.27) in section 7.8.2.

**Proof:**

Let $U_n(\theta) = \sum_{i=1}^{n} \frac{\partial \ell_i(\theta)}{\partial \theta}$, the summation of the first derivatives of the log likelihood functions (called the score functions), where $\ell_i(\theta) = \log f_Y(Y_i|\theta)$.

The MLE $\hat{\theta}$ solves $U_n(\theta) = 0$. A Taylor expansion gives

$$U_n(\hat{\theta}) = U_n(\theta_c) + I_n(\hat{\theta})(\hat{\theta} - \theta_c) = 0$$  \tag{7.35}

where $\hat{\theta}$ is a value of $\theta$ which lies between $\hat{\theta}$ and $\theta_c$ and

$$I_n(\theta) = \sum_{i=1}^{n} \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^\top}.$$

Hence from (7.35),

$$\sqrt{n}(\hat{\theta} - \theta_c) = \left[ -\frac{1}{n} I_n(\hat{\theta}) \right]^{-1} n^{-\frac{1}{2}} U_n(\theta_c)$$  \tag{7.36}

Now by the law of large numbers

$$-\frac{1}{n} I_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^\top} \right]_{\theta_c} \frac{p}{\theta} - E_{f_P} \left[ \frac{\partial^2 \ell_i(\theta)}{\partial \theta \partial \theta^\top} \right]_{\theta_c} = J(\theta_c).$$  \tag{7.37}

Note $U_n(\theta_c) = \sum_{i=1}^{n} U_i(\theta_c)$ where $U_i(\theta_c) = \left[ \frac{\partial \ell_i(\theta)}{\partial \theta} \right]_{\theta_c}$ and $E_{f_P}[U_i(\theta_c)] = 0$ since

$$E_{f_P}[U_i(\theta_c)] = \int \frac{\partial}{\partial \theta} [\log f_Y(y_i|\theta)]_{\theta_c} f_P(y_i) dy_i = 0$$

because $\theta_c$ minimizes the Kullback-Liebler distance

$$d[f_P(y), f_Y(y|\theta)] = \int [\log f_P(y) - \log f_Y(y|\theta)] f_P(y) dy$$

with respect to $\theta$. Also $V_{f_P}[U_i(\theta_c)] = K(\theta_c)$. Hence $E_{f_P}[U_n(\theta_c)] = 0$ and $V_{f_P}[U_n(\theta_c)] = nK(\theta_c)$ and so by the central limit theorem,

$$n^{-\frac{1}{2}} U_n(\theta_c) \xrightarrow{d} N_K(0, K(\theta_c)).$$  \tag{7.38}

Hence applying Slutsky’s theorem to (7.36) using (7.37) and (7.38) gives

$$\sqrt{n}(\hat{\theta} - \theta_c) \xrightarrow{d} J(\theta_c)^{-1} N_K(0, K(\theta_c)) = N_K \left( 0, J(\theta_c)^{-1} K(\theta_c) J(\theta_c)^{-1} \right).$$  \tag{7.39}

[Note that if the true population distribution $f_P(y)$ belongs to the parametric family of distributions $f_Y(y|\theta)$, then we have $\theta_c = \theta_T$ and $K(\theta_T) = J(\theta_T)$, so $\sqrt{n}(\hat{\theta} - \theta_T) \xrightarrow{d} N_K \left( 0, J(\theta_T)^{-1} \right)$.]
7.10 Exercises for Chapter 3

7.10.1 Maximum Likelihood Estimation 1

Let $Y_i$ be a random variable having a binomial distribution with probability density function given by

$$f(Y_i) = \frac{n!}{(n_i - Y_i)!Y_i!}p^{Y_i}(1 - p)^{(n_i - Y_i)}$$

for $Y_i = 0, 1, \ldots, n_i$ and where $0 < p < 1$. Let $y = (y_1, y_2, \ldots, y_k)$ be a random sample of observations from the above distribution.

(a) Write down the likelihood function $L(p)$ and show that the log likelihood $\ell(p)$ for the parameter $p$ is given by:

$$\ell(p) = \sum_{i=1}^{k} y_i \log(p) + \sum_{i=1}^{k} (n_i - y_i) \log(1 - p)$$

(b) Find the maximum likelihood estimator pf $p$.

(c) Show that the expected information for the parameter $p$ is given by

$$\hat{i}(p) = -E\left[\frac{d^2\ell}{dp^2}\right] = \sum_{i=1}^{k} \frac{n_i}{p} + \sum_{i=1}^{k} \frac{n_i}{(1 - p)}$$

Note that $E(Y_i) = n_ip$.

7.10.2 Maximum Likelihood Estimation 2

A manufacturing process produce fibres of varying lengths. It is assumed that the length of a fibre is a continuous variable with pdf given by

$$f(Y) = \theta^{-2}Ye^{-\frac{Y}{\theta}}$$

for $Y > 0$ and where $\theta > 0$ is an unknown parameter. Suppose that $n$ randomly selected fibres have length $(y_1, y_2, \ldots, y_n)$. Find an expression for the MLE for $\theta$. 
Chapter 8

Fitting distributions to data: examples

This chapter shows how R and the package gamlss can be use for fitting a distribution to data. In particular it explains:

1. the gamlss()
2. the gamlssML()
3. the histDist()
4. the fitDist() and
5. the fitSmo()

functions. It also give several example of fitting a continuous distribution to data.

8.1 Fitting gamlss.family distributions

In this previous Chapter we have shown how to use the general optimisation function optim() and the more specialised function mle() to fit distributions in R. Here we described the functions available in the package gamlss which can be used for fitting gamlss.family distributions.

The GAMLSS models for fitting a (up to four parameters) distributions to an observed random sample is:

\[ Y_i | \mu, \sigma, \nu, \tau \sim D(\mu, \sigma, \nu, \tau) \]

\[
  g_1(\mu) = \eta_\mu = b_{0,1} \\
  g_2(\sigma) = \eta_\sigma = b_{0,2} \\
  g_3(\nu) = \eta_\nu = b_{0,3} \\
  g_4(\tau) = \eta_\tau = b_{0,4}
\]

(8.1)

independently for \( i = 1, 2, 3, 4 \) where the link functions \( g_i() \) for \( i = 1, 2, 3, 4 \) are fixed known functions, (selected for convenience so the that the parameters are denied in right range). The default link function for the gamlss.family distributions are given in Tables 1.1, 1.2 and 1.3 and by obtain in R by typing name of the distribution, i.e. NO().
8.1.1 The functions `gamlss()` and `gamlssML()`.

The function function `gamlss()` is the main model fitting function of the package `gamlss`. It uses the RS or the CG algorithm (which are also appropriate for regression type models, that is, when explanatory variables are available in the data), Rigby and Stasinopoulos [2005]. The function `gamlssML()` is specifically designed for fitting a `gamlss.family` distribution to a single data set when no explanatory variables are available. It uses the non linear maximisation algorithm in the R function `optim()` or `nlminb()` and it is relevant only when there are no explanatory variables. The following code shows how the two function can be used:

```r
> m1 <- gamlss(aircond~1, family=EXP)
GAMLSS-RS iteration 1: Global Deviance = 247.72
GAMLSS-RS iteration 2: Global Deviance = 247.72
> summary(m1)
*******************************************************************
Family: c("EXP", "Exponential")
Call: gamlss(formula = aircond ~ 1, family = EXP)
Fitting method: RS()
-------------------------------------------------------------------
Mu link function: log
Mu Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
4.161e+00  2.043e-01  2.036e+01  3.262e-16
-------------------------------------------------------------------
No. of observations in the fit: 24
Degrees of Freedom for the fit: 1
Residual Deg. of Freedom: 23
at cycle: 2
Global Deviance:  247.72
AIC:  249.72
SBC:  250.8981
*******************************************************************
> m2 <- gamlssML(aircond, family=EXP)
> summary(m2)
*******************************************************************
Family: c("EXP", "Exponential")
Call: gamlssML(y = aircond, family = EXP)
Fitting method: "nlminb"

Coefficient(s):
            Estimate Std. Error  t value Pr(>|t|)
```

```
8.1. FITTING GAMLLS.FAMILY DISTRIBUTIONS

eta.mu 4.160834 0.204124 20.3838 < 2.22e-16 ***
---
Signif. codes: 0 ^O***~O 0.001 ^O**~O 0.01 ^O*~O 0.05 ^O.~O 0.1 ^O ~O 1

Degrees of Freedom for the fit: 1 Residual Deg. of Freedom  23
Global Deviance: 247.72
AIC: 249.72
SBC: 250.898

There are two obvious differences in the two functions. The first is that gamlss uses a formula, aircond~1, as its first argument while gamlssML() uses only the y-variable. The second has to do with the output. The function gamlss() prints the iteration numbers by default while gamlssML() does not. The two functions have the following common arguments:

family: for appropriate gamlss.family distribution

data: the the data.frame.

weights: for prior weights

mu.start, sigma.start, nu.start, tau.start: for starting values for the parameters µ, σ, ν and τ respectively.

mu.fix, codesigma.fix, nu.fix, tau.fix: whether to fix the parameters µ, σ, ν and τ at specific values or not.

start.from: starting from a previous fitted model.

Settings for controlling the gamlss() fitting algorithm can be passed using the i.control and control arguments (type ?gamlss for more details). For just fitting distributions such controls are rarely needed. Control settings for the optim() or nlminb() within gamlssML() can be passed through the function by using the ‘...’ facility.

Both functions create S3 R objects with information stored about fitted models. The function gamlss() creates a gamlss object while gamlssML() a gamlssML object which is a subclass of a gamlss object. The generic functions AIC(), GAIC(), coef(), deviance(), extractAIC(), logLik(), print(), residuals(), summary(), vcov() and the simple functions wp() and dtop() can be used with both objects.

Important note on the parametrization.

Here we highlight an important feature of both the gamlss() and the gamlssML() concerning the parametrisation used in fitting procedure. Since the gamlss() function was created originally for regression type of models, link functions are used to ensure parameters remain within their proper range. For example, to ensure the µ parameter of the exponential distribution remains positive for all possible values of all the explanatory variables and their coefficients a log link function is used for modelling, i.e. \( \eta_\mu = \log(\mu) \). Link functions are one of the innovations introduced by Nelder and Wedderburn [1972] in their seminal “Generalised Linear Models” paper. A consequence of this is that both gamlss() and the gamlssML() functions fit the model on the predictor (link) scale rather than the actual parameter space. For the exponential distribution for example both functions fit \( \eta_\mu = \log(\mu) \) rather that \( \mu \). As far as
fitting distributions is concerned, because we are using MLE with its invariance property, we have $\hat{\mu} = \exp(\hat{\eta}_\mu)$ this does not creates a problem, It does though creates some confusion especially with the output in `summary()` because the estimates and their standard errors are in the predictor scale. The following code shows, at least in the `aircond` example where we are using an exponential distribution, that working on the predictor scale $\eta_\mu = \log(\mu)$ is beneficial in the sense that the log-likelihood shape is closer to a quadratic shape and therefore standard errors are more accurate.

```r
prof.dev(m1,"mu",min=40,max=120,step=0.5,type="l")
mod<-quote(gamlss(aircond~-1+offset(I(this*rep(1,length(aircond)))),
family =EXP))
prof.term(mod, min=3.5, max=5, step=.01, type="l")
```

Figure 8.1 plots the profile deviance of both $\mu$ (on the top) and $\eta_\mu$ (on the bottom) for the `aircond` data where an exponential distribution is assumed. The profile deviance for a parameter (eg. $\mu$ here) is obtained by fixing the parameter at each of a sequence of values (40 to 120 in steps of 0.5 here) fitting the rest of the distribution parameters to the data and plotting the fitted deviance (=minus twice the log-likelihood) against the parameters values. A % 95 confidence interval for the parameter includes all values of the parameter $\theta$ ($\mu$ here) for which the (profile deviance - maximum profile deviance) is less than $\chi^2_{0.05} = 3.841$, i.e. all values if the parameter $\theta$ which are accepted by a generalised likelihood ratio test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$.

It shows that the profile deviance of the $\eta_\mu$ is almost symmetric with a good quadratic like shape while for $\mu$ it is skew to the right. Using the estimate of the standard error we obtained earlier we can construct a 95% confidence interval for $\mu$ as $\hat{\mu} \pm 1.96 \times se(\hat{\mu}) \Rightarrow 64.125 \pm 1.96 \times 13.0895 \approx (38.46, 89.78)$ is a 95% CI using the $\mu$ scale. Using the $\eta_\mu$ scale we have $(\exp(4.160 - 1.96 \times 0.204), \exp(4.160 + 1.96 \times 0.204)) = (42.98, 95.67)$. Both intervals can be compared to the 95% confidence interval obtained using the profile deviance of Figure 8.1 which is $(44.06988, 98.44196)$.

### 8.1.2 The function `distHist()`

The specifically designed function `histDist()` will fit a `gamlss.family` distribution to a sample and automatically produce a graphical representation of the data and the fitted distribution. It can be applied to both continuous and discrete variables but not to mixed distribution. For example using `histDist(aircond, family=EXP)` will produce the plot in figure 8.2.

Note that the function `histDist()` has the following basic arguments: i) `y`: the vector containing the values of the $Y$ variable ii) `family`: for the appropriate `gamlss.family` distribution iii) `freq`: the observed frequencies corresponding to the values in `y` (usually appropriate for a discrete response variable). `freq` is used as prior `weights` in the fit. iv) `density`: whether to plot a nonparametric density plot, together with the parametric fitted distribution (only for continuous distributions). v) `nbins`: for controlling the number of bins in the histogram.

### 8.1.3 The function `fitDist()`

This function is design so the user can fit all available (from a list) `gamlss.family` distributions and select the one with the smallest generalised Akaike information criterion (GAIC). The function `fitDist` uses the function `gamlssML()` to fit all relevant parametric distributions to a data vector.

The function has the following arguments
8.1. FITTING GAMLSS.FAMILY DISTRIBUTIONS

Figure 8.1: Showing the global deviance for the parameter $\mu$ at the top and the parameter $\eta_\mu$ at the bottom.
The aircond and the fitted EXP distribution

Figure 8.2: Showing the plot produced by the function histDist() on the aircond using an exponential distribution fit.

y : the data vector

k : the penalty for the GAIC with default value k=2 the standard AIC

type : a list of distributions for fit

try.gamlss : if gamlssML() failed whether to try gamlss() instead. (For large data sets this will slow things up).

extra : a list of extra distribution which are not in the type list

data : the data frame where y can be found

The following lists of distributions are available as options in type:

"realAll": all the gamlss.family continuous distributions defined on the real line, \( \mathbb{R} \), or positive real line \( \mathbb{R}^+ \), i.e. realline, and realplus

"realline": all the continuous distributions in the real line \( \mathbb{R} \): GU, RG, LO, NET, TF, PE, SN1, SN2, SHASH, EGB2, JSU, SEP1, SEP2, SEP3, SEP4, ST1, ST2, ST3, ST4, ST5, GT

"realplus": all the continuous distributions in the positive real line \( \mathbb{R}^+ \): EXP, GA, IG, LNO, WEI, BCCGo, exGAUS, GG, GIG, BCTo, BCPEo

"real0to1": all the continuous distributions on the interval 0 to 1: BE, BEINF, BEINFO, BEINF1, BEOI, BEZI, GB1

"counts": All the discrete count distributions with infinite upper limit: PO, LG, NBI, NBII, PIG, DEL, SI, ZIP, ZIP2, ZAP, ZALG, ZINBI, ZANBI, ZIPIG

"binom": all the discrete count distributions from 0 to a finite upper limit: BI, BB, ZIBI, ZIBB, ZABI, ZABB
Here it how the function `fitDist()` can be used to find the best model under AIC searching all available `gamlss.family` distribution with domain on the positive real line:

```r
mod1<-fitDist(aircond, type="realplus")
```

The class of the object class `mod1` is `gamlssML` so all methods for this class will work. In addition the `mod1` has two extra components. i) `mod1$fit` which gives the GAIC of all fitted distributions and ii) `mod1$failed` which shows which distributions failed to fit.

**8.2 How to fit non-parametric distributions in R**

An alternative more flexible way of estimating the probability density function of a continuous response variable $Y$ (particularly in the absence of explanatory variables) is using non-parametric density function estimation, as described for example in Silverman (1986) or Wand and Jones (1995). This is a well established technique and in certain cases could be a preferable method of density function estimation. Nonparametric density function estimation has its own problems, notably the need to estimate the smoothing parameter and the inability to cope with discrete data. Here we concentrate on fitting parametric distributions to the data. Our approach is to fit different parametric distributions and choose between them using a generalised Akaike information criterion, (GAIC). For continuous data, nonparametric density estimation could be useful tool in helping us with this choice of the parametric distribution.

**8.2.1 The function fitSmo()**

**8.2.2 The function flexDist()**

**8.3 Examples of fitting continuous distributions to data**

**8.3.1 The Parzen data**

<table>
<thead>
<tr>
<th>Data summary:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>data file:</strong> parzen in package <code>gamlss.data</code> of dimensions 63 × 1</td>
</tr>
<tr>
<td><strong>source:</strong> Hand et al. (1994)</td>
</tr>
<tr>
<td><strong>variables</strong></td>
</tr>
<tr>
<td><code>snowfall</code> : the annual snowfall in Buffalo, NY (inches) for the 63 years, from 1910 to 1972 inclusive.</td>
</tr>
<tr>
<td><strong>purpose:</strong> to demonstrate the fitting of continuous distribution to a single variable.</td>
</tr>
<tr>
<td><strong>conclusion:</strong> the Weibull distribution seems to fit best</td>
</tr>
</tbody>
</table>
Fitting and display the model

The first data set is the snowfall data used by Parzen (1979) and also contained in Hand et al. (1994), data set 278. The data give the annual snowfall (inches) in Buffalo, NY for the 63 years, from 1910 to 1972 inclusive.

```r
> library(gamlss)
> data(parzen)
> names(parzen)

[1] "snowfall"
```

Here we use the function `fitDist()` to fit distributions to the data. We are using the default value for the argument `type = "realAll"`, meaning we are using all available continuous distributions. Also we try different information criteria: the AIC and the SBC.

```r
> mod1 <- fitDist(snowfall, data=parzen, )
> mod2 <- fitDist(snowfall, data=parzen, k=log(length(snowfall)))
> mod1$fit[1:9]

     WEI3    PE    GG   BCCGo   SN2     TF  exGAUS   SN1
579.9043 581.3779 581.9038 582.6824 582.7327 582.7331 582.7331 582.7331
     SEP4
582.8463

> mod2$fit[1:9]

     WEI3    PE    LO    GA    GG   BCCGo   SN2     TF
584.1905 587.8073 588.0162 588.1016 588.3332 589.1118 589.1621 589.1625
     exGAUS
589.1625

> mod1$failed

[[1]]
[1] "NET"

> mod2$failed

[[1]]
[1] "NET"
```

Using both criteria it is obvious that the best model is the one using the Weibull distribution. Next we refit the model using the command `histDist(parzen$snowfall, "WEI3", density=TRUE)`. Note that the option `density=TRUE` requests a non-parametric kernel density estimate to be included in the plot. The resulting Figure 8.3 shows a histogram of the snowfall together with the "best" fitted distribution (WEI3) and the non-parametric kernel density estimate.
8.3. EXAMPLES OF FITTING CONTINUOUS DISTRIBUTIONS TO DATA

> histDist(parzen$snowfall, "WEI3", density=TRUE)

Figure 8.3: A histogram of the Parzen’s snowfall data together with the fitted Weibull distribution (red) and a kernel density estimate (blue)
Checking the model

A check of the (normalised quantile) residuals provides a guide to the adequacy of the fit. The true (normalised quantile) residuals are defined as \( r_i = \Phi^{-1}(u_i) \) where \( \Phi^{-1} \) is the inverse cumulative distribution function of a standard normal variate and \( u_i = F_Y(y_i|\mu_i, \sigma_i, \nu_i, \tau_i) \). The true (normalised quantile) residuals are independent standard normal variables. We expect the fitted (normalised quantile) residuals \( \hat{r}_i \) to behave approximately as normally distributed variables (even though the original observations \( Y_i \) are not necessarily normal), so a normal Q-Q plot of the residuals is appropriate here. The \texttt{gamlss} package provides the functions i) \texttt{plot()} and ii) \texttt{wp()} and iii) \texttt{dtop()} for plotting QQ-plots.

Figure 8.4 shows the results of using \texttt{plot(mod1)}, while Figure 8.5 shows the result of using \texttt{wp(mod1)}. Figure 8.4 plots the normalised quantile residuals against the fitted values and the case number (i.e., the index number), together with their kernel density estimate and their normal QQ-plot. The worm plot is a detrended Q-Q plot (i.e., where the line in a Q-Q plot has been transformed horizontally). Both the QQ-plot, (in the right bottom corner of Figure 8.4 and the worm plot in Figure 8.5 indicate that there is no reason to worry about the inadequacy of the fit. In particular adequacy of the WEI3 distribution is indicated by deviations in the worm plot lying within the dashed confidence bands. Note that not all the plots in Figure 8.4 are useful in a regression type situation, especially the one at the left top of the picture.

Figure 8.4: The residual plot from the fitted Weibull model to the snowfall data
Testing hypotheses from the fitted model

There are several methods to check the reliability of the fitted parameters of the distribution. Standard errors for the fitted parameters are provided by two functions: i) the `summary()` and ii) by the `vcov()` function. In general the two values should be identical, since by default `summary` is the standard errors obtained by `vcov`. The standard errors obtained by `vcov()` are the ones obtained by inverting the full Hessian matrix and they do take into account the correlations between the distribution parameter estimates. Note that the function `vcov()` after a `gamlss()` object refits the final model one more time in order to obtain the Hessian matrix. Occasionally this could fail in which case `summary()` will use an alternative method called `qr`. This alternative method uses the QR decomposition of the individual distribution parameter estimation fits. The standard errors given by the `qr` method of `summary()` are not very reliable since they are the (conditional) standard errors obtained by assuming that the other distribution parameters are fixed at their maximum likelihood estimates. Next we use the `summary()` and the `vcov()` function.

```r
> m1<-gamlss(snowfall ~ 1, data=parzen, family=WEI3, trace=FALSE)
> summary(m1)
```

```
*******************************************************************
Family: c("WEI3", "Weibull type 3")
Call: gamlss(formula = snowfall ~ 1, family = WEI3, data = parzen,
            trace = FALSE)
Fitting method: RS()
```

Figure 8.5: The worm plot from the fitted Weibull model to the snowfall data
**Mu link function:** log

**Mu Coefficients:**

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| 4.387e+00 | 3.680e-02  | 1.192e+02 | 5.615e-74 |

---

**Sigma link function:** log

**Sigma Coefficients:**

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|---------|
| 1.344e+00 | 9.883e-02  | 1.360e+01 | 3.993e-20 |

---

No. of observations in the fit: 63

Degrees of Freedom for the fit: 2

Residual Deg. of Freedom: 61

at cycle: 3

Global Deviance: 575.9043

AIC: 579.9043

SBC: 584.1905

> vcov(m1, type="se")

(Intercept) (Intercept)

0.03680025 0.09883186

The fitted Weibull distribution model is given by $Y_i \sim \text{WEI}(\hat{\mu}, \hat{\sigma})$ where, since $\mu$ and $\sigma$ have log link functions, $\log(\hat{\mu}) = \hat{\beta}_{01} = 4.387$, so $\hat{\mu} = \exp(4.387) = 80.399$ and $\log(\hat{\sigma}) = \hat{\beta}_{02} = 1.344$, so $\hat{\sigma} = 3.834$. Note that $\hat{\mu}$ and $\hat{\sigma}$ are maximum likelihood estimates of $\mu$ and $\sigma$.

The standard errors obtained are 0.0368 for $\log(\hat{\mu}) = \hat{\beta}_{01}$ and 0.09883 for $\log(\hat{\sigma}) = \hat{\beta}_{02}$ respectively, using either the `summary()` or `vcov()` functions. Note that since the Weibull fitting function `WEI3()` uses the log link for $\mu$ and the log link for $\sigma$ the standard errors given are those for $\log(\hat{\mu}) = \hat{\beta}_{01}$ and for $\log(\hat{\sigma}) = \hat{\beta}_{02}$. For example, an approximate 95% confidence interval (CI) for $\log(\sigma) = \hat{\beta}_{02}$, using the `vcov()` results, will be

$$[1.344 - (1.96 \times 0.09883), 1.344 + (1.96 \times 0.09883)] = (1.1503, 1.5377).$$

Hence an approximate 95% CI confidence interval for $\sigma$ is given by $[\exp(1.1503), \exp(1.5377)] = (3.1591, 4.6539)$.

This can be compared with a profile deviance interval $(3.125, 4.616)$ obtained using the `prof.dev(m1, "sigma", min=3, max=4.8, step=.01, type="l")` function with the resulting profile evince plot for $\sigma$ given in Figure 8.6. (Note that `prof.dev()` works only with `gamlss` objects).

Also it can be compared with a bootstrap ‘Baisc’ CI given by $[\exp(1.160), \exp(1.491)] = (3.190, 4.442)$ obtained using the following R script. Note that the function `boot` is given in Venables and Ripley (2000) page 173.

> library(boot)
> set.seed(1453)
8.3. EXAMPLES OF FITTING CONTINUOUS DISTRIBUTIONS TO DATA

Figure 8.6: The profile deviance for the $\sigma$ parameter for fitted Weibull model

```r
> mod1<-gamlss(snowfall~1, data=parzen, family=WEI3, trace=FALSE)
> funB <- function(data, i)
+   { 
+     d<-data.frame(snowfall=data[i,])
+     coef(update(mod1, data=d),"sigma")
+   }
> (mod1.boot<-boot(parzen, funB, R=199))

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = parzen, statistic = funB, R = 199)

Bootstrap Statistics :
  original bias std. error
  t1* 1.343915 0.01359133 0.08991679

> boot.ci(mod1.boot, type=c("norm", "basic"))

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 199 bootstrap replicates

CALL :
boot.ci(boot.out = mod1.boot, type = c("norm", "basic"))

Intervals :
  Level       Normal     Basic
  95%  ( 1.154,  1.507 )  ( 1.160,  1.491 )
Calculations and Intervals on Original Scale

Some basic intervals may be unstable
8.3.2 The strengths of glass fibres data

<table>
<thead>
<tr>
<th>Data summary:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R data file:</strong> glass in package gamlss.dist of dimensions 63 × 1</td>
</tr>
<tr>
<td><strong>source:</strong> Smith and Naylor (1987)</td>
</tr>
<tr>
<td><strong>variables</strong></td>
</tr>
<tr>
<td>strength : the strength of glass fibres (the unit of measurement is not given).</td>
</tr>
<tr>
<td><strong>purpose:</strong> to demonstrate the fitting of a parametric distribution to the data.</td>
</tr>
<tr>
<td><strong>conclusion</strong> a SEP4 distribution fits adequately</td>
</tr>
</tbody>
</table>

The following data show the strength of glass fibres, measured at the National Physical Laboratory, England, see Smith and Naylor (1987), (the unit of measurement was not given in the paper). Here we fit different distributions to the data and we select the "best" model using first the Akaike information criterion, then the Schwartz Bayesian criterion.

We also demonstrate the use of the gen.trun() function in the package gamlss.tr designed to create a truncated distribution from an existing gamlss.family distribution. Here we generated a positive t distribution called TFtr and we fit it to the data together with the rest of the distributions. To obtain a positive t distribution we use the command gen.trun() with par=0 to truncate it at zero with left truncation by default:

```r
> # bring the truncated distributions package
> library(gamlss.tr)
> # create a truncated t
> gen.trun(par=0, family=TF)
```
A truncated family of distributions from TF has been generated and saved under the names: 
dTFtr pTFtr qTFtr rTFtr TFtr
The type of truncation is left and the truncation parameter is 0

Next we fit all available continuous distributions using AIC and SBC information criteria. The argument extra is used to add the truncated t distribution to the default list .realAll.

```r
> data(glass)
> m1<-fitDist(strength, data=glass, k=2, extra="TFtr") # AIC
> m2<-fitDist(strength, data=glass, k=log(length(strength)), extra="TFtr") # SBC
> m1$fit[1:8]
  SEP4  SEP3  SHASHo  SEP1  SEP2  JSU  BCPEo  ST2
27.65361 27.97321 28.01800 29.00072 29.05912 30.41380 31.17627 31.40100
> m2$fit[1:8]
  SEP4  SEP3  SHASHo  SEP1  SEP2  GU  WEI3  JSU
36.22615 36.54575 36.59054 37.57326 37.63166 38.19839 38.69995 38.98634
```

The best model for the glass fibre strength according to both the Akaike and Schwartz Bayesian information criteria is the SEP4 distribution. Our truncated t distribution, TFtr, did not fit well to this particular data set. The fitted SEP4 distribution together with the data is shown in Figure 8.8 obtained using the command:

```r
> h1 <- histDist(glass$strength, SEP4, nbins=13, main="SEP4 distribution",
+ method=mixed(20,50), trace=FALSE)
```

The fitted distribution has a spike at its mode. Distributions which involve the power exponential distribution, (eg. all the SEP’s), with values of the kurtosis parameter(s) (i.e. $\tau$ for SEP1, SEP2, SEP3 and both $\nu$ and $\tau$au for SEP4), less than or equal to 1 often have a discontinuity in the gradient, leading to a spike at the mode. This often results in a multimodal likelihood function (with respect to $\mu$, and leads to inferential problems. In the SEP4 distribution the parameters $\nu$ and $\tau$ adjust the kurtosis at the left and right side of the distribution respectively. The estimates of those two parameters are $\hat{\nu} = \exp(-0.1280) = 0.880$ and $\hat{\tau} = \exp(0.3183) = 1.375$ respectively, indicating possible problems with the inferential procedures since $\nu \leq 1$. Note we can extract the fitted coefficients using either of the functions coef() or fitted(), e.g. coef(msep4, "nu") or fitted(msep4, "nu")[1].

```r
> summary(m1)
```

```
**********************************************************
Family: c("SEP4", "skew exponential power type 4")
Call: gamlssML(y = y, family = DIST[i])
Fitting method: "nlminb"
Coefficient(s):
```

```r
> summary(m1)
```

```
**********************************************************
Family: c("SEP4", "skew exponential power type 4")
Call: gamlssML(y = y, family = DIST[i])
Fitting method: "nlminb"
Coefficient(s):
```
Figure 8.8: The strengths of glass fibres data and the fitted SEP4 distribution model

|        | Estimate   | Std. Error | t value | Pr(>|t|)    |
|--------|------------|------------|---------|-------------|
| eta.mu | 1.61000000 | 0.00436385 | 368.9400 | < 2.22e-16 *** |
| eta.sigma | -1.83908556 | 0.45814686 | -4.01418 | 5.9652e-05 *** |
| eta.nu | -0.33378926 | 0.20478263 | -1.62997 | 0.10311       |
| eta.tau | 0.06020469  | 0.33687863 | 0.17871  | 0.85816       |

---

Degrees of Freedom for the fit: 4 Residual Deg. of Freedom 59

Global Deviance: 19.6536

AIC: 27.6536

SBC: 36.2262

A worm plot of the residuals of the fitted model in Figure 8.9 shows that all the points are close to the horizontal line within the dashed confidence bands indicating that the SEP4 distribution provides an adequate fit to the data.
8.3. EXAMPLES OF FITTING CONTINUOUS DISTRIBUTIONS TO DATA

> wp(m1)

Figure 8.9: A worm plot of the residuals for the model using the SEP4 distribution fitted to the strengths of glass fibres data.

8.3.3 The tensile strength data: response on (0, 1)

<table>
<thead>
<tr>
<th>Data summary:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R data file:</strong> tensile in package gamlss.data of dimensions 30 × 1</td>
</tr>
<tr>
<td><strong>source:</strong> Hand et al. (1994)</td>
</tr>
<tr>
<td><strong>variables</strong></td>
</tr>
<tr>
<td>str : the strength of polyester fibres (the unit of measurement are not given).</td>
</tr>
<tr>
<td><strong>purpose:</strong> to demonstrate the fitting of a parametric distribution to the data.</td>
</tr>
<tr>
<td><strong>conclusion</strong> a truncated lognormal distribution fits best</td>
</tr>
</tbody>
</table>

These data come from Quesenberry and Hales (1980) and were also reproduced in Hand et al. (1994), data set 180, page 140. They contain measurements of tensile strength of polyester fibres and the authors were trying to check if they were consistent with the lognormal distribution. According to Hand et al. (1994) "these data follow from a preliminary transformation. If the lognormal hypothesis is correct, these data should have been uniformly distributed". Here we use them as an example of data from a variable restricted to the range (0, 1) and try to fit appropriate distributions. Note that apart from the beta (BE) and LOGITNO two parameter distributions, and the generalised beta type 1, (GB1), a four parameter distribution, there are no other distributions in the current version of GAMLSS software which are restricted to the range 0 to 1. So we create some using the gen.Family() function of the gamlss.dist package.
The distributions we create are the logit-\( t \), logit-SEP3 and logit-ST3 distributions. See section 11.5 for an explanation of logic transited distributions. First we fit the distributions and then we select the "best" using an Akaike information criterion.

```r
> library(gamlss)
> data(tensile)
> gen.Family("TF", "logit")
```

A logit family of distributions from TF has been generated and saved under the names:

\( dlogitTF \), \( plogitTF \), \( qlogitTF \), \( rlogitTF \), \( logitTF \)

```r
> gen.Family("ST3", "logit")
```

A logit family of distributions from ST3 has been generated and saved under the names:

\( dlogitST3 \), \( plogitST3 \), \( qlogitST3 \), \( rlogitST3 \), \( logitST3 \)

```r
> gen.Family("SEP3", "logit")
```

A logit family of distributions from SEP3 has been generated and saved under the names:

\( dlogitSEP3 \), \( plogitSEP3 \), \( qlogitSEP3 \), \( rlogitSEP3 \), \( logitSEP3 \)

```r
> mbe <- gamlss(str~1, data=tensile, family=BE, trace=FALSE)
> mlogitno <- gamlss(str~1, data=tensile, family=LOGITNO, trace=FALSE)
> mlogittf <- gamlss(str~1, data=tensile, family=logitTF, trace=FALSE)
> mlogitst3 <- gamlss(str~1, data=tensile, family=logitST3, trace=FALSE)
> mlogitsep3 <- gamlss(str~1, data=tensile, family=logitSEP3, trace=FALSE)
> mgb1 <- gamlss(str~1, data=tensile, family=GB1, method=mixed(10,100), trace=FALSE)
> GAIC(mbe, mlogitno, mlogittf, mlogitst3, mlogitsep3, mgb1)
```

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mlogitno</td>
<td>2</td>
<td>-4.2279342</td>
</tr>
<tr>
<td>mbe</td>
<td>2</td>
<td>-2.6101273</td>
</tr>
<tr>
<td>mlogittf</td>
<td>3</td>
<td>-2.2278772</td>
</tr>
<tr>
<td>mlogitsep3</td>
<td>4</td>
<td>-0.7301101</td>
</tr>
<tr>
<td>mlogitst3</td>
<td>4</td>
<td>-0.2946836</td>
</tr>
<tr>
<td>mgb1</td>
<td>4</td>
<td>0.2258434</td>
</tr>
</tbody>
</table>

The logit-normal distribution gives the best fit according to the Akaike information criterion. Figure 8.10 shows the fitted distributions. The plots in the figure were created using the following code:

```r
op <- par(mfrow = c(3, 2))
histDist(tensile$str, "LOGITNO", nbins=14, xlim=c(0,1), ylim=c(0,3), main="(a) logitNO")
histDist(tensile$str, "BE", nbins=14, xlim=c(0,1), ylim=c(0,3), main="(b) BE")
histDist(tensile$str, "logitTF", nbins=14, xlim=c(0,1), ylim=c(0,3), main="(c) logitTF")
```
histDist(tensile$str, "logitSEP3", nbins=14, xlim=c(0,1), ylim=c(0,3), main="(d) logitSEP3", n.cyc=50)
histDist(tensile$str, "logitST3", nbins=14, xlim=c(0,1), ylim=c(0,3), main="(d) logitST3", n.cyc=50)
histDist(tensile$str, "GB1", nbins=14, xlim=c(0,1), ylim=c(0,3), main="(e) GB1", n.cyc=50)
par(op)

Figure 8.10: Tensile strength data with fitted (a) logit-normal, (b) beta, (c) logit-t (d) logit-SEP3 (e) logit-TS3 and (f) generalised beta type distributions.

8.4 Exercises for Chapter 3

8.4.1 The Turkish stock exchange index

The Turkish stock exchange index, $I_i$, was recorded daily from 1/1/1988 to 31/12/1998. The daily returns, $ret = \log_e(I_{i+1}/I_i)$, were obtained for $i = 1, 2, \ldots, 2868$. 
(a) Input the data

```r
data(tse)
```

and plot the data sequentially using

```r
with(tse, plot(ret,type="l"))
```

(b) Fit each of the following distributions for ret using the command `histDist()` (and using different model names for later comparison):

- two parameter: normal \( NO(\mu, \sigma) \), i.e.  
  ```r
  mNO<-histDist(tse$ret,"NO")
  ```
- three parameter: t family \( TF(\mu, \sigma, \nu) \) and power exponential \( PE(\mu, \sigma, \nu) \)
- four parameter: Johnson Su \( JSU(\mu, \sigma, \nu, \tau) \), skew exponential power type 1 to 4 i.e. \( SEP1(\mu, \sigma, \nu, \tau) \), skew t type 1 to 5 i.e. \( ST1(\mu, \sigma, \nu, \tau) \) and sinh-arc-sinh \( SHASH(\mu, \sigma, \nu, \tau) \).

[Note that to ensure convergence you may need to increase the number of GAMLSS iterations using for example `n.cyc=100` or switch the algorithm from `RS()` to `CG()` after few iterations using the argument `method` i.e.

```r
method=mixed(10,100),
```

for both `gamlss()` and `histDist()` functions.

Also if you are using `histDist()` increase the default value of the argument `nbins` to 30 i.e.

```r
histDist(tse$ret, family=SEP4, nbins=30, n.cyc=100).
```

(c) Use the AIC command with each of the penalties \( k = 2, 3.8 \) and \( 7.96=\log(2868) \), [corresponding to criteria AIC, \( \chi^2_{0.05} \) and SBC respectively], in order to select a distribution model. Output the parameter estimates for your chosen model using the function `summary`. 
Chapter 9

Fitting Count Data

9.1 Examples: fitting a distribution

9.1.1 The computer failure data

<table>
<thead>
<tr>
<th>Data summary:</th>
</tr>
</thead>
<tbody>
<tr>
<td>R data file: <code>computer</code> in package <code>gamlss.dist</code> of dimensions 128 × 1</td>
</tr>
<tr>
<td>source: Hand et al. (1994)</td>
</tr>
<tr>
<td>variables</td>
</tr>
<tr>
<td>failure : the number of computers that broke down.</td>
</tr>
<tr>
<td>purpose: to demonstrate the fitting of a parametric discrete distribution to the data.</td>
</tr>
<tr>
<td>conclusion a PIG distribution fits best</td>
</tr>
</tbody>
</table>

The following data relate to DEC-20 computers which operated at the Open University in the 1980. They give the number of computers that broke down in each of the 128 consecutive weeks of operation, starting in late 1983, see Hand et al. (1994) page 109 data set 141. Here we use four different count data distributions and choose between them using the Akaike information criterion (AIC):

```r
> graphics.off()
> library(gamlss.dist)
> data(computer)
> op <- par(mfrow = c(2, 2))
> mPO <- histDist(computer$failure, "PO", main = "PO", trace = FALSE)
> mNBI <- histDist(computer$failure, "NBI", main = "NBI", trace = FALSE)
> mPIG <- histDist(computer$failure, "PIG", main = "PIG", trace = FALSE)
> mSI <- histDist(computer$failure, "SICHEL", main = "SICHEL",
+     trace = FALSE)
> AIC(mPO, mNBI, mPIG, mSI)
```

163
From the GAIC table above we conclude that the PIG model is the appropriate model. Now we refit the model and display a summary of the final model.

> mod1 <- gamlss(failure ~ 1, data = computer, family = PIG, trace = FALSE)
> summary(mod1)

Figure 9.1: The computer failure data fit with (a) Poisson, (b) negative binomial (c) Poisson inverse gaussian and (d) Sichel distributions respectively
Call: gamlss(formula = failure ~ 1, family = PIG, data = computer, trace = FALSE)

Fitting method: RS()

Mu link function: log
Mu Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 1.390 | 0.08355 | 16.64 | 1.058e-33 |

Sigma link function: log
Sigma Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -0.3489 | 0.2387 | -1.461 | 0.1464 |

No. of observations in the fit: 128
Degrees of Freedom for the fit: 2
Residual Deg. of Freedom: 126
at cycle: 3

Global Deviance: 632.4159
AIC: 636.4159
SBC: 642.12

Hence the fitted PIG model for the computer failure data is given by \( Y \sim PIG(\hat{\mu}, \hat{\sigma}) \) where \( \hat{\mu} = \exp(1.390) = 4.015 \) and \( \hat{\sigma} = \exp(-0.3489) = 0.7055 \), with fitted mean \( \hat{E}(Y) = \hat{\mu} = 4.015 \) and fitted variance \( \hat{V}(Y) = \hat{\mu} + \hat{\sigma}\hat{\mu}^2 = 15.39 \).

9.1.2 The lice data

Data summary:

R data file: lice in package gamlss.dist of dimensions 71 \times 2

source: Williams (1944)

variables

- head: the number of head lice
- freq: the frequency of prisoners with the number of head lice

purpose: to demonstrate the fitting of a parametric discrete distribution to the data.

conclusion: a SICHEL distributions fits best

The following data come from Williams (1944) and they are frequencies (freq) of prisoners
with number of head lice (head), for Hindu male prisoners in Cannamore, South India, 1937-1939. We fit four different distributions to head and choose between them using AIC:

```r
> library(gamlss.dist)
> con <- gamlss.control(trace = FALSE, n.cyc = 50)
> data(lice)
> mPO <- gamlss(head ~ 1, data = lice, family = PO, weights = freq,
+        trace = FALSE)
> mNBI <- gamlss(head ~ 1, data = lice, family = NBI, weights = freq,
+        trace = FALSE)
> mPIG <- gamlss(head ~ 1, data = lice, family = PIG, weights = freq,
+        trace = FALSE)
> mSI <- gamlss(head ~ 1, data = lice, family = SICHEL, weights = freq,
+        n.cyc = 50, trace = FALSE)
> AIC(mPO, mNBI, mPIG, mSI)

                      df   AIC
mSI       3 4646.214
mNBI      2 4653.687
mPIG      2 4756.275
mPO       1 29174.823
```

We conclude that the Sichel model explains the data best. The summary of the final fitted model is shown below:

```r
> summary(mSI)

*******************************************************************
Family: c("SICHEL", "Sichel")
Call:
gamlss(formula = head ~ 1, family = SICHEL, data = lice, weights = freq,
       n.cyc = 50, trace = FALSE)
Fitting method: RS()

---------------------------------------------------------------
Mu link function: log
Mu Coefficients:
  Estimate  Std. Error  t value  Pr(>|t|)   
(Intercept)  1.927      0.07952   24.23  5.965e-104
---------------------------------------------------------------
Sigma link function: log
Sigma Coefficients:
  Estimate  Std. Error  t value  Pr(>|t|)   
(Intercept)  4.806      0.20341   23.63  6.974e-100
---------------------------------------------------------------
Nu link function: identity
```

We conclude that the Sichel model explains the data best. The summary of the final fitted model is shown below:
9.1. EXAMPLES: FITTING A DISTRIBUTION

Nu Coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | -0.004787 | 0.01595 | -0.3002 | 0.7641 |

---

No. of observations in the fit: 1083
Degrees of Freedom for the fit: 3
Residual Deg. of Freedom: 1080
Global Deviance: 4640.214
AIC: 4646.214
SBC: 4661.177

Hence the fitted SICHEL distribution for the number of head lice ($Y = \text{head}$) is given by

$Y \sim \text{SICHEL}(\hat{\mu}, \hat{\sigma}, \hat{\nu})$ where $\hat{\mu} = \exp(1.927) = 6.869$ and $\hat{\sigma} = \exp(4.806) = 122.24$ and $\hat{\nu} = -0.0047$, with fitted mean $E(Y) = \hat{\mu} = 6.869$ and fitted variance $V(Y) = 432.25$, (obtained using the code VSICHEL(mSI)[1]). Figure 9.2 shows the fitted negative binomial and Sichel models created by the following R commands. Note that Figure 9.2 only plots the data and fitted distributions up to $y = 10$.

```R
> op <- par(mfrow = c(2, 1))
> m1 <- histDist(lice$head, "NBI", freq = lice$freq, xlim = c(0, + 10), main = "NBI distribution", trace = FALSE)
> m2 <- histDist(lice$head, "SICHEL", freq = lice$freq, xlim = c(0, + 10), main = "Sichel distribution", trace = FALSE, n.cyc = 50)
> par(op)
```

9.1.3 A stylometric application

Data summary:

R data file: stylo in package gamlss.dist of dimensions 64 x 2

source: Dr Mario Corina-Borja

variables

word: is the number of times a word appears in a single text

freq: the frequency of the number of times a word appears in a text

purpose: to demonstrate the fitting of a truncated discrete distribution to the data.

conclusion: the truncated SICHEL distributions fits best

The data are from a stylometric application, the discipline which tries to characterize the style of a text, Chappas and Corina-Borja (2006). Here the response variable, word, is the number of times a word appears in a single text. The variable freq, records the frequency of the word (i.e. frequency freq is the number of different words which occur exactly word times
Figure 9.2: The lice data with the fitted (a) negative binomial and (b) Sichel distributions respectively.
in the text). Possible values of word are \( y = 1, 2, 3, \ldots \), and since the objective here is to fit an appropriate distribution to the data we are looking for a zero truncated discrete distribution. In the specific data we are using, the maximum times that any word appears in the text is 64 (word=64), while the most frequent value of word is 1 with frequency 947. We first input and plot the data:

```r
> library(gamlss.dist)
> library(gamlss.tr)
> data(stylo)
> plot(freq ~ word, data = stylo, type = "h", xlim = c(0, 22),
>       xlab = "no of times", ylab = "frequencies", col = "blue")
```

![Graph](image)

Figure 9.3: The stylometric data: number of time a word appear in a text against the frequencies

Note that for plotting we restrict the upper x-limit to 22 since the most of the frequencies after that have zero values. We will now generate several truncated discrete distributions, using the function `gen.truc()` from the package `gamlss.tr`, to fit them to the data. Specifically we generate i) truncated Poisson, ii) truncated negative binomial type II iii) truncated Depalorte and iv) truncated Sichel. [Note that the truncated negative binomial type I model takes more than 300 iterations to converge and eventually give the same result as the truncated negative binomial type II.]

```r
> library(gamlss.tr)
> gen.trun(par = 0, family = PO, type = "left")
```

A truncated family of distributions from PO has been generated and saved under the names:

- dPOtr
- pPOtr
- qPOtr
- rPOtr
- POtr

The type of truncation is left and the truncation parameter is 0
CHAPTER 9. FITTING COUNT DATA

> gen.trun(par = 0, family = NBII, type = "left")
A truncated family of distributions from NBII has been generated and saved under the names:
dNBIItr pNBIItr qNBIItr rNBIItr NBIItr
The type of truncation is left and the truncation parameter is 0

> gen.trun(par = 0, family = DEL, type = "left")
A truncated family of distributions from DEL has been generated and saved under the names:
dDELtr pDELtr qDELtr rDELtr DELtr
The type of truncation is left and the truncation parameter is 0

> gen.trun(par = 0, family = SICHEL, type = "left", delta = 0.001)
A truncated family of distributions from SICHEL has been generated and saved under the names:
dSICHELtr pSICHELtr qSICHELtr rSICHELtr SICHELtr
The type of truncation is left and the truncation parameter is 0

We now fit the distributions to the data and choose between them using AIC:

> mPO <- gamlss(word ~ 1, weights = freq, data = stylo, family = POtr, + trace = FALSE)
> mNBII <- gamlss(word ~ 1, weights = freq, data = stylo, family = NBIItr, + n.cyc = 50, trace = FALSE)
> mDEL <- gamlss(word ~ 1, weights = freq, data = stylo, family = DELtr, + n.cyc = 50, trace = FALSE)
> mSI <- gamlss(word ~ 1, weights = freq, data = stylo, family = SICHELtr, + n.cyc = 50, trace = FALSE)
> GAIC(mPO, mNBII, mDEL, mSI)

<table>
<thead>
<tr>
<th>df</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>mSI</td>
<td>5148.454</td>
</tr>
<tr>
<td>mDEL</td>
<td>5160.581</td>
</tr>
<tr>
<td>mNBII</td>
<td>5311.627</td>
</tr>
<tr>
<td>mPO</td>
<td>9207.459</td>
</tr>
</tbody>
</table>

The best fitted model according to the AIC is the truncated Sichel model. The Depalorte model performed better than the negative binomial type II model. Figure 9.4 shows all of the fitted models above. The fit of the (zero truncated) Poisson distribution is shown in part (a) of the Figure 9.4. This is not a very good fit to the data and an improved fit is achieved by using the (truncated) negative binomial distribution II in part (b) and the Delaporte distribution in (c). The (truncated) Sichel in panel (d) is a superior fit according to both AIC and SBC. Figure 9.4 was produced using the following code:

> op <- par(mfrow = c(2, 2))
> tabley <- with(stylo, table(rep(word, freq)))
> mNO <- histDist(stylo$word, family = POtr, freq = stylo$freq, + main = "(b) Poisson", ylim = c(0, 0.65), xlim = c(1, 26),
9.1. EXAMPLES: FITTING A DISTRIBUTION

```r
+ trace = FALSE)
> mNBII <- histDist(stylo$word, family = NBIItr, freq = stylo$freq,
+ main = "(c) negative binomial II", ylim = c(0, 0.65), xlim = c(1, 26),
+ start.from = mNBII, trace = FALSE)
> mDEL <- histDist(stylo$word, family = DELtr, freq = stylo$freq,
+ main = "(c) Delaporte", ylim = c(0, 0.65), xlim = c(1, 26),
+ start.from = mDEL, trace = FALSE)
> mSI <- histDist(stylo$word, family = SICHELtr, freq = stylo$freq,
+ main = "(d) Sichel", ylim = c(0, 0.65), xlim = c(1, 26),
+ start.from = mSI, trace = FALSE)
> par(op)
```

![Graphs of distributions](image)

Figure 9.4: The stylometric data fits with (a) Poisson (b) negative binomial type II (c) Delaporte and (d) Sichel distributions respectively

```r
+ trace = FALSE)
> mNBII <- histDist(stylo$word, family = NBIItr, freq = stylo$freq,
+ main = "(c) negative binomial II", ylim = c(0, 0.65), xlim = c(1, 26),
+ start.from = mNBII, trace = FALSE)
> mDEL <- histDist(stylo$word, family = DELtr, freq = stylo$freq,
+ main = "(c) Delaporte", ylim = c(0, 0.65), xlim = c(1, 26),
+ start.from = mDEL, trace = FALSE)
> mSI <- histDist(stylo$word, family = SICHELtr, freq = stylo$freq,
+ main = "(d) Sichel", ylim = c(0, 0.65), xlim = c(1, 26),
+ start.from = mSI, trace = FALSE)
> par(op)
```
Chapter 10

Fitting Binomial data

10.1 Examples of fitting binomial data

10.1.1 The alveolar-bronchiolar adenomas data

<table>
<thead>
<tr>
<th>Data summary:</th>
</tr>
</thead>
<tbody>
<tr>
<td>R data file: alveolar in package gamlss.dist of dimensions 23 × 2</td>
</tr>
<tr>
<td>source: Tamura and Young (1987), and Hand et al. (1994)</td>
</tr>
<tr>
<td>variables</td>
</tr>
<tr>
<td>r : number of mice having alveolar-bronchiolar</td>
</tr>
<tr>
<td>n : total number of mice</td>
</tr>
<tr>
<td>purpose: to demonstrate the fitting of a binomial distribution to the data.</td>
</tr>
<tr>
<td>conclusion: a binomial distribution is adequate</td>
</tr>
</tbody>
</table>

Here we consider the alveolar-bronchiolar adenomas data used by Tamura and Young (1987) and also reproduced in Hand et al. (1994), data set 256. The data are the number of mice having alveolar-bronchiolar adenomas out of total numbers of mice (i.e. the "binomial denominator") in 23 independent groups.

For binomial type of data with no explainatory variables the histDist() can still be used but with a limited scope. The plot works fine if the 'binomial denominator' is constant for all observations. In this case we can plot a histogram of y against the number of events from zero to the constant 'binomial denominator', and then superimpose the fitted probabilities from the fitted binomial distribution. When the binomial denominator is not constant for all observations then histDist() plots a histogram of the proportions (which may be of some interest) and then indicates where the fitted proportion lies:

```r
> library(gamlss.dist)
> data(alveolar)
> alveolar$y <- with(alveolar, cbind(r, n - r))
> con <- gamlss.control(trace = F)
> m1 <- gamlss(y ~ 1, data = alveolar, family = BI, control = con)
```

173
> m2 <- gamlss(y ~ 1, data = alveolar, family = BB, control = con)
> GAIC(m1, m2)

        df   AIC
m1   1 73.1292
m2   2  75.0665

> m3 <- with(alveolar, histDist(y, "BI", xlim = c(0, 0.3)))

GAMLSS-RS iteration 1: Global Deviance = 71.1292
GAMLSS-RS iteration 2: Global Deviance = 71.1292

With the two models having similar deviances there is no support from the data to favour
the beta binomial model instead of the binomial one. Figure 10.1 shows the histogram of
proportion of mice having alveolar-bronchiolar adenomas and the red vertical line indicates the
fitted probability.

In order to demonstrate what would have happen in the plot if the binomial denominator
was the same for all observations we fix it in the above data to be 10.

> alveolar$yy <- with(alveolar, cbind(r, 10 - r))
> m1 <- gamlss(yy ~ 1, data = alveolar, family = BI, control = con)
> m2 <- gamlss(yy ~ 1, data = alveolar, family = BB, control = con)
> GAIC(m1, m2)

        df   AIC
m2  2 88.9301
m1  1 104.7267

> m3 <- histDist(alveolar$yy, "BB")

Figure 10.1: The proportion of alveolar-bronchiolar adenomas
10.1. EXAMPLES OF FITTING BINOMIAL DATA

GAMLSS-RS iteration 1: Global Deviance = 87.4166
GAMLSS-RS iteration 2: Global Deviance = 85.0155
GAMLSS-RS iteration 3: Global Deviance = 84.9312
GAMLSS-RS iteration 4: Global Deviance = 84.9301
GAMLSS-RS iteration 5: Global Deviance = 84.9301

Figure 10.2: The proportion and fitted distribution to the artificial alveolar-bronchiolar adenomas data

10.1.2 The first year student examination results data

Data summary:

R data file: students created here of dimensions 8 × 2

source: Karlis and Xekalaki (2008)

variables

\( r \): number of exams first year students passed out of 8 in total
\( freq \): the frequency (i.e. the number of students) for the number of exams passed

purpose: to demonstrate the fitting of a binomial distribution to the data when the binomial denominator is fixed.

conclusion: a beta binomial distribution is adequate

Here we demonstrate the fitting of a binomial type data response variable given that the binomial denominator if constant. The data shown in Table 10.1 (first used by Karlis and
Xekalaki, 2008) refer to the numbers of courses passed, \( r \), and their frequency, \( \text{freq} \), of a class of 65 first year students. The students enrolled for an 8 course during the year. The variable \( n.r \) in table 10.1 is defined as \( 8 - r \).

<table>
<thead>
<tr>
<th>( r )</th>
<th>( n.r )</th>
<th>( \text{freq} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>8.00</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>7.00</td>
</tr>
<tr>
<td>3</td>
<td>2.00</td>
<td>6.00</td>
</tr>
<tr>
<td>4</td>
<td>3.00</td>
<td>5.00</td>
</tr>
<tr>
<td>5</td>
<td>4.00</td>
<td>4.00</td>
</tr>
<tr>
<td>6</td>
<td>5.00</td>
<td>3.00</td>
</tr>
<tr>
<td>7</td>
<td>6.00</td>
<td>2.00</td>
</tr>
<tr>
<td>8</td>
<td>7.00</td>
<td>1.00</td>
</tr>
<tr>
<td>9</td>
<td>8.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 10.1: The first year student examination results data where the binomial denominator is constant at 8.

Now we create the data and fit a binomial and a beta binomial distribution. We then select a model using AIC.

```r
> r <- 0:8
> freq <- c(1, 4, 4, 8, 6, 8, 12, 13)
> y <- cbind(r, 8 - r)
> colnames(y) <- c("r", "n-r")
> students <- data.frame(y, freq)
> m1 <- gamlss(y ~ 1, weights = freq, data = students, family = BI,
+              trace = FALSE)
> m2 <- gamlss(y ~ 1, weights = freq, data = students, family = BB,
+              trace = FALSE)
> GAIC(m1, m2)

   df  AIC
m2  2 273.4987
m1  1 339.6467
```

The beta binomial model has a superior fit. This is also demonstrated in Figure 10.3 where the fitted probabilities from the two models are plotted. The figures were produced using the following code:

```r
> op <- par(mfrow = c(2, 1))
> m3 <- with(students, histDist(cbind(rep(y[, 1], freq), rep(y[, 2], freq)), family = BI, ylim = c(0, 0.3), xlab = "number of courses passed", + ylab = "probability", main = "(a) binomial"))
```

GAMLSS-RS iteration 1: Global Deviance = 337.6467
GAMLSS-RS iteration 2: Global Deviance = 337.6467

```r
> m4 <- with(students, histDist(cbind(rep(y[, 1], freq), rep(y[, 2], freq)), family = BB, ylim = c(0, 0.3), xlab = "number of courses passed", + ylab = "probability", main = "(b) beta binomial"))
```
10.1. EXAMPLES OF FITTING BINOMIAL DATA

GAMLSS-RS iteration 1: Global Deviance = 270.2592
GAMLSS-RS iteration 2: Global Deviance = 269.4993
GAMLSS-RS iteration 3: Global Deviance = 269.4987

> par(op)

Figure 10.3: The first year student examination results data: Fitted probabilities (a) binomial (b) beta binomial distributions
Chapter 11

Methods of generating Distributions

This chapter provides explanation for:

1. how new continuous distributions can be generated
2. how some of the distributions are interconnected
3. families of distributions

This chapter is more theoretical and can be omitted for a practical course.

11.1 Methods of generating continuous distributions

Here we examine how many of the distributions in Tables 2.1, 2.2 and 2.3 for the random variable $Y$ can be generated. Distribution families for $Y$ can be generated by one (or more) of the following methods:

1. Azzalini type methods
2. splicing distributions
3. a (continuous or finite) mixture of distributions
4. univariate transformation from a single random variable
5. transformation from two or more random variables
6. truncation distributions
7. systems of distributions

These methods are discussed next in Sections 11.2 to 11.8 respectively.
11.2 Distributions generated by Azzalini’s method

There are two Azzalini’s methods, the first was proposed in 1985 and the second in 2003. Those methods are described in sections 11.2.1 and 11.2.2 respectively.

Note that, as we have mentioned in section 2.5, an important disadvantage of distributions generated by Azzalini type methods are that their cumulative distribution function (cdf) is not explicitly available, but requires numerical integration. Their inverse cdf requires a numerical search and many integrations. Consequently both functions can be slow, particularly for large data sets. Centiles and centile based measures (e.g. the median) are not explicitly available. Moment based measures are usually complicated, if available. However they can be very flexible in modelling skewness and kurtosis.

11.2.1 Azzalini (1985) method

Lemma 1 of Azzalini (1985) proposed the following method of introducing skewness into a symmetric probability density function. Let \( f_{Z_1}(z) \) be a probability density function symmetric about \( z \) equals zero and let \( F_{Z_2}(z) \) be an absolutely continuous cumulative distribution function such that \( dF_{Z_2}(z)/dz \) is symmetric about zero. Then, for any real \( \nu \), \( f_Z(z) \) is a proper probability density function where

\[
f_Z(z) = 2f_{Z_1}(z)F_{Z_2}(\nu z).
\]  

(11.1)

Let \( Y = \mu + \sigma Z \) then

\[
f_Y(y) = \frac{2}{\sigma} f_{Z_1}(z) F_{Z_2}(\nu z)
\]  

(11.2)

where \( z = (y - \mu)/\sigma \). This allows the generation of families of skew distributions including Skew Normal type 1, SN1, Skew exponential power type 1, SEP1, and Skew t type 1, ST1, given below.

Skew Normal type 1 (SN1)

The skew normal type 1 family for \( -\infty < Y < \infty \), Azzalini (1985), denoted by SN1(\( \mu, \sigma, \nu \)), is defined by assuming \( Z_1 \) and \( Z_2 \) have standard normal, NO(0, 1), distributions in (11.2).

Consider \( Y \sim \text{SN1}(\mu, \sigma, \nu) \). First note that \( Z = (Y - \mu)/\sigma \sim \text{SN1}(0, 1, \nu) \) has pdf given by (11.1) where \( Z_1 \) and \( Z_2 \) have standard normal, NO(0, 1) distributions.

Figure 11.1(a) plots \( f_{Z_1}(z) \) of \( Z_1 \sim \text{NO}(0,1) \) against \( z \). Figure 11.1(b) plots \( 2 * F_{Z_2}(\nu z) \) of \( Z_2 \sim \text{NO}(0,1) \) against \( z \) for \( \nu = 0, 1, 2, 1000 \). Figure 11.1(c) plots the skew normal type 1 distribution, \( f_{Z_1}(z) \) of \( Z \sim \text{SN1}(0,1,\nu) \) against \( z \) for \( \nu = 0, 1, 2, 1000 \).

Clearly from equation (11.1) the pdf \( f_Z(z) \) is pdf \( f_{Z_1}(z) \) weighted by \( 2 * F_{Z_2}(\nu z) \) for each value for \( z \) for \( -\infty < z < \infty \). When \( \nu = 0 \) then \( 2 * F_{Z_2}(\nu z) = 1 \) for all \( z \), i.e. a constant weight 1, hence \( f_Z(z) = f_{Z_1}(z) \), a standard normal pdf. For \( \nu > 0 \), \( 2 * F_{Z_2}(\nu z) \) provides heavier weights for \( z > 0 \) then for \( z < 0 \) resulting in a positively skew \( f_Z(z) \).

As \( \nu \to \infty \), \( 2 * F_{Z_2}(\nu z) \) tends to a 0-1 step function resulting in a half normal distribution \( f_Z(z) \), the most positively skew distribution in the SN1 family. Switching from \( \nu \) to \(-\nu \) reflects \( F_Z(z) \) about \( z = 0 \) leading to negatively skew distributions. Finally \( Y = \mu + \sigma Z \), so the distribution of \( Y \sim \text{SN1}(\mu, \sigma, \nu) \) is a scaled and shifted version of the distribution of \( Z \sim \text{SN1}(0,1,\nu) \).

The code for creating figure 11.1 is the following:
11.2. DISTRIBUTIONS GENERATED BY AZZALINI’S METHOD

Figure 11.1: Azzalini’s method (a) pdf, $f_{Z_1}(z)$ of $Z_1 \sim NO(0,1)$ (b) 2*cdf, $2 \times F_{Z_2}(\nu z)$ of $Z_2 \sim NO(0,1)$ for $\nu=0, 1, 2, 1000$ (note that the solid line $\nu = 1000$ is a step function) (c) The skew normal type 1 distribution, $Z \sim SN1(0,1,\nu)$ for $\nu=0, 1, 2, 1000$. 
CHAPTER 11. METHODS OF GENERATING DISTRIBUTIONS

<table>
<thead>
<tr>
<th>Distributions of $Y$</th>
<th>Distribution of $Z_1$</th>
<th>Distribution of $Z_2$</th>
<th>$w(z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN1($\mu, \sigma, \nu$)</td>
<td>NO(0,1)</td>
<td>NO(0,1)</td>
<td>$\nu z$</td>
</tr>
<tr>
<td>SEP1($\mu, \sigma, \nu, \tau$)</td>
<td>PE2(0, $\tau^{1/\tau}$, $\tau$)</td>
<td>PE2(0, $\tau^{1/\tau}$, $\tau$)</td>
<td>$\nu z$</td>
</tr>
<tr>
<td>SEP2($\mu, \sigma, \nu, \tau$)</td>
<td>PE2(0, $\tau^{1/\tau}$, $\tau$)</td>
<td>NO(0,1)</td>
<td>$\nu (2/\tau)^{1/2} \text{sign}(z)</td>
</tr>
<tr>
<td>ST1($\mu, \sigma, \nu, \tau$)</td>
<td>TF(0,1,$\tau$)</td>
<td>TF(0,1,$\tau$)</td>
<td>$\nu z$</td>
</tr>
<tr>
<td>ST2($\mu, \sigma, \nu, \tau$)</td>
<td>TF(0,1,$\tau$)</td>
<td>TF(0,1,$\tau+1$)</td>
<td>$\nu \lambda^{1/2} z$</td>
</tr>
</tbody>
</table>

Table 11.1: Showing distributions generated by Azzalini type methods using equation (11.4)

```r
op <- par(mfrow=c(3,1))
z<-seq(-4,4,by=0.01)
plot(dNO(z, mu=0, sigma=1)~z, type="l",col="black",ylab="f_{Z_1}(z)"
plot(2*pNO(z, mu=0, sigma=0.001)~z, type="l",col="blue",ylab="2*F_{Z_2}(z)",lt=1)
lines(2*pNO(z, mu=0, sigma=0.5)~z, col="red", lt=2,lw=2)
lines(2*pNO(z, mu=0, sigma=1)~z, col="green",lt=4,lw=2)
lines(2*pNO(z, mu=0, sigma=10000)~z, col="black",lt=3,lw=2)

plot(dSEP1(z, mu=0, sigma=1, nu=1000, tau=2)~z, type="l",col="blue",
    ylab="f_{Z}(z)",lt=1)
lines(dSEP1(z, mu=0, sigma=1, nu=2, tau=2)~z, col="red", lt=2, lw=2)
lines(dSEP1(z, mu=0, sigma=1, nu=1, tau=2)~z, col="green",lt=4,lw=2)
lines(dSEP1(z, mu=0, sigma=1, nu=0, tau=2)~z, col="black",lt=3,lw=2)
par(op)
```

**Skew exponential power type 1 (SEP1)**

The skew exponential power type 1 family for $-\infty < Y < \infty$, Azzalini (1986), denoted by $\text{SEP1}(\mu, \sigma, \nu, \tau)$, is defined by assuming $Z_1$ and $Z_2$ have power exponential type 2, $\text{PE2}(0, \tau^{1/\tau}, \tau)$, distributions in (11.2). Azzalini (1986) called this distribution type I.

The skew normal type 1, $\text{SN1}(\mu, \sigma, \nu)$, is a special case of $\text{SEP1}(\mu, \sigma, \nu, \tau)$ obtained by setting $\tau = 2$. The flexibility of the SEP1 was demonstrated in figure 2.5.

**Skew t type 1 (ST1)**

The skew t type 1 family for $-\infty < Y < \infty$, Azzalini (1986), denoted by $\text{ST1}(\mu, \sigma, \nu, \tau)$, is defined by assuming $Z_1$ and $Z_2$ have Student t distributions with $\tau > 0$ degrees of freedom, i.e., $\text{TF}(0,1,\tau)$, in (11.2).

**11.2.2 Azzalini and Capitano (2003) method**

Equation (11.1) was generalised, in Azzalini and Capitano (2003) Proposition 1, to

$$f_Z(z) = 2f_{Z_1}(z)F_{Z_2}[w(z)] \quad (11.3)$$

where $w(z)$ is any odd function of $z$ i.e. $w(-z) = -w(z)$. Hence if $Y = \mu + \sigma z$ then

$$f_Y(y) = \frac{2}{\sigma}f_{Z_1}(z)F_{Z_2}[w(z)] \quad (11.4)$$
where \( z = (y - \mu)/\sigma \). This allows a wider generation of families of distributions than the Azzalini (1985) method, including the Skew exponential power type 2, SE2 and Skew t type 2, ST2 below. A summary of distributions generated by (11.4) is given in Table 11.1.

**Skew exponential power type 2 (SEP2)**

The skew exponential power type 2 family, denoted by SEP2(\( \mu, \sigma, \nu, \tau \)), Azzalini (1986) and DiCiccio and Monti (2004) is expressed in the form (11.4) by letting \( Z_1 \sim \text{PE2}(0, \tau^{1/\tau}, \tau) \), \( Z_2 \sim \text{NO}(0, 1) \) and \( w(z) = \nu(2/\tau)^{1/2} \text{sign}(z)|z|^{\tau/2} \). Azzalini (1986) developed a reparametrization of this distribution given by setting \( \nu = \text{sign}(\lambda)|\lambda|^{\tau/2} \) and called it type II. The skew normal type 1, SN1(\( \mu, \sigma, \nu \)), distribution is a special case of SEP2(\( \mu, \sigma, \nu, \tau \)) obtained by setting \( \tau = 2 \).

**Skew t type 2 (ST2)**

The skew t type 2 family, denoted by ST2(\( \mu, \sigma, \nu, \tau \)) is expressed in the form (11.4) by letting \( Z_1 \sim \text{TF}(0, 1, \tau), Z_2 \sim \text{TF}(0, 1, \tau + 1) \) and \( w(z) = \nu \lambda^{1/2} z \) where \( \lambda = (\tau + 1)/(\tau + z^2) \), Azzalini and Capitanio (2003). An alternative derivation of ST2 is given in Section 11.6.

### 11.3 Distributions generated by splicing

#### 11.3.1 Splicing using two components

Splicing has been used to introduce skewness into a symmetric distribution family. Let \( Y_1 \) and \( Y_2 \) have probability density functions that are symmetric about \( \mu \). A spliced distribution for \( Y \) may be defined by

\[
f_Y(y) = \pi_1 f_{Y_1}(y)I(y < \mu) + \pi_2 f_{Y_2}(y)I(y \geq \mu),
\]

where \( I() \) is an indicator variable taking value 1 of the condition is true and 0 otherwise. Ensuring that \( f_Y(y) \) is a proper probability density function requires \( (\pi_1 + \pi_2)/2 = 1 \). Ensuring continuity at \( y = \mu \) requires \( \pi_1 f_{Y_1}(\mu) = \pi_2 f_{Y_2}(\mu) \). Hence \( \pi_1 = 2/(1 + k) \) and \( \pi_2 = 2k/(1 + k) \) where \( k = f_{Y_1}(\mu)/f_{Y_2}(\mu) \) and

\[
f_Y(y) = \frac{2}{(1 + k)} \left\{ f_{Y_1}(y)I(y < \mu) + kf_{Y_2}(y)I(y \geq \mu) \right\}.
\]

A summary of distributions generated by (11.6) is given in Table 11.2.

#### 11.3.2 Splicing using two components with different scale parameters

A “scale-spliced” distribution for \( Y \) may be defined by assuming that probability density function \( f_Z(z) \) is symmetric about 0 and that \( Y_1 = \mu + \sigma Z/\nu \) and \( Y_2 = \mu + \sigma\nu Z \) in (11.6). Hence

\[
f_Y(y) = \frac{2}{(1 + k)} \left\{ \frac{\nu}{\sigma} f_Z(\nu z)I(y < \mu) + \frac{k}{\nu\sigma} f_Z(z/\nu)I(y \geq \mu) \right\}.
\]

for \( z = (y - \mu)/\sigma \) and where \( k = f_{Y_1}(\mu)/f_{Y_2}(\mu) = \nu^2 \). Hence

\[
f_Y(y) = \frac{2\nu}{\sigma(1 + \nu^2)} \left\{ f_Z(\nu z)I(y < \mu) + f_Z(z/\nu)I(y \geq \mu) \right\}.
\]
The formulation (11.8) was used by Fernandez, Osiewalski and Steel (1995) and Fernandez and Steel (1998). This allows the generation of "scale-spliced" families of distributions including Skew normal, \( \text{SN2} \), Skew exponential power type 3, \( \text{SEP3} \) and Skew t type 3, \( \text{ST3} \), below. The distribution of \( Y \) is symmetric for \( \nu = 1 \), positively skew for \( \nu > 1 \) and negatively skew for \( 0 < \nu < 1 \) (assuming \( Z \) has its mode at 0). Switching from \( \nu \) to \( 1/\nu \) reflects \( f_Y(y) \) about \( y = \mu \).

### Skew normal (\( \text{SN2} \))

A skew normal type 2 distribution (or two-piece normal distribution) for \( -\infty < Y < \infty \), denoted by \( \text{SN2}(\mu, \sigma, \nu) \), is defined by assuming \( Z \sim \text{NO}(0, 1) \) in (11.8) or equivalently \( Y_1 \sim \text{NO}(\mu, \sigma/\nu) \) and \( Y_2 \sim \text{NO}(\mu, \sigma) \) in (11.6), giving

\[
f_Y(y) = \frac{2\nu}{\sqrt{2\pi}\sigma(1 + \nu^2)} \left\{ \exp \left[ -\frac{1}{2} \left( \nu z \right)^2 \right] I(y < \mu) + \exp \left[ -\frac{1}{2} \left( \frac{z}{\nu} \right)^2 \right] I(y \geq \mu) \right\}
\]

(11.9)

where \( z = (y - \mu)/\sigma \). References to this distribution are given in Johnson \textit{et al.} (1994) p 173 and Jones and Faddy (2003). The earliest reference appears to be Gibbons and Mylroie (1973).

For example consider \( Y \sim \text{SN2}(0, 1, \nu) \) in (11.9), where \( Y_1 \sim \text{NO}(0, 1/\nu) \) and \( Y_2 \sim \text{NO}(0, \nu) \), i.e. a spliced two-piece normal distribution. Figure 11.2 plots \( Y \sim \text{SN2}(0,1,\nu) \) for \( \nu=1, 2, 3, 5 \). Switching from \( \nu \) to \( 1/\nu \) reflects \( f_Y(y) \) about \( y = 0 \).

![Figure 11.2: Splicing method \( Y \sim \text{SN2}(0,1,\nu) \) for \( \nu=1, 2, 3, 5 \). Switching from \( \nu \) to \( 1/\nu \) reflects \( f_Y(y) \) about \( y = 0 \).](image)

### Skew exponential power type 3 (\( \text{SEP3} \))

A skew exponential power type 3 distribution for \( -\infty < Y < \infty \), Fernandez, Osiewalski and Steel (1995), denoted by \( \text{SEP3}(\mu, \sigma, \nu, \tau) \), is defined by assuming \( Z \sim \text{PE2}(0, 2^{1/\tau}, \tau) \) in (11.8) or equivalently, \( Y_1 \sim \text{PE2}(\mu, \sigma 2^{1/\tau}/\nu, \tau) \) and \( Y_2 \sim \text{PE2}(\mu, \sigma /\nu 2^{1/\tau}, \tau) \) in (11.6). Note that the skew normal type 2 distribution, \( \text{SN2}(\mu, \sigma, \nu) \), is a special case of \( \text{SEP3}(\mu, \sigma, \nu, \tau) \) given by setting \( \tau = 2 \).
11.3. DISTRIBUTIONS GENERATED BY SPLICING

<table>
<thead>
<tr>
<th>Distributions of $Y$</th>
<th>Distribution of $Y_1$</th>
<th>Distribution of $Y_2$</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>SN2($\mu, \sigma, \nu$)</td>
<td>NO($\mu, \sigma/\nu$)</td>
<td>NO($\mu, \sigma\nu$)</td>
<td>Gibbons and Mylroie (1973)</td>
</tr>
<tr>
<td>SEP3($\mu, \sigma, \nu, \tau$)</td>
<td>PE2($\mu, \sigma^{2\nu/\tau}$, $\nu, \tau$)</td>
<td>PE2($\mu, \sigma^{2\nu/\tau}$, $\nu, \tau$)</td>
<td>Fernandez, Osiewolski and Steel (1995)</td>
</tr>
<tr>
<td>SEP4($\mu, \sigma, \nu, \tau$)</td>
<td>PE2($\mu, \sigma/\nu$, $\nu, \tau$)</td>
<td>PE2($\mu, \sigma/\nu$, $\nu, \tau$)</td>
<td>Jones (2005)</td>
</tr>
<tr>
<td>ST3($\mu, \sigma, \nu, \tau$)</td>
<td>TF($\mu, \sigma/\nu$, $\nu, \tau$)</td>
<td>TF($\mu, \sigma/\nu$, $\nu, \tau$)</td>
<td>Fernandez and Steel (1998)</td>
</tr>
<tr>
<td>ST4($\mu, \sigma, \nu, \tau$)</td>
<td>TF($\mu, \sigma/\nu$, $\nu, \tau$)</td>
<td>TF($\mu, \sigma/\nu$, $\nu, \tau$)</td>
<td>Rigby and Stasinopoulos (????)</td>
</tr>
</tbody>
</table>

Table 11.2: Showing distributions generated by splicing

**Skew $t$ type 3 (ST3)**

A *skew type 3* distribution for $-\infty < Y < \infty$, Fernandez and Steel (1998), denoted by ST3($\mu, \sigma, \nu, \tau$) is defined by assuming $Z \sim TF(0, 1, \tau) \equiv t_{\tau}$ in (11.8), or equivalently $Y_1 \sim TF(\mu, \sigma/\nu, \tau)$ and $Y_2 \sim TF(\mu, \sigma\nu, \tau)$ in (11.6). A reparametrization of ST3, in which $\mu$ and $\sigma$ are the mean and the standard deviation of $Y$ is given by Hansen (1994). Theodossiou (1998) extended the Hansen reparametrization to a five parameter skew generalised $t$ distribution.

**11.3.3 Splicing using two components with different shape parameters**

A “shape-spliced” distribution for $Y$ may be defined by assuming $Y_1$ and $Y_2$ in (11.6) have different shape parameters. This allows the generation of “shape-spliced” families of distributions, including Skew exponential power type 4 SEP4 and Skew $t$ type 4 ST4 below.

**Skew exponential power type 4 (SEP4)**

A *skew exponential power type 4* family for $-\infty < Y < \infty$, Jones (2005), denoted by SEP4($\mu, \sigma, \nu, \tau$), is defined by assuming $Y_1 \sim PE2(\mu, \sigma, \nu)$ and $Y_2 \sim PE2(\mu, \sigma, \tau)$ in (11.6). Note that $\mu$ is the mode of $Y$.

A similar distribution was used by Nandi and Mämpel (1995) who set $Y_1 \sim PE2(\mu, \sigma, \nu)$ and $Y_2 \sim PE2(\mu, \sigma/q, \tau)$ in (11.6), where $q = \Gamma[1 + (1/\tau)]/\Gamma[1 + (1/\nu)]$. However this distribution constrains both the median and mode of $Y$ to be $\mu$, which is perhaps rather restrictive.

**Skew $t$ type 4 (ST4)**

A *skew $t$ type 4* family for $-\infty < Y < \infty$, denoted by ST4($\mu, \sigma, \nu, \tau$), is defined by assuming $Y_1 \sim TF(\mu, \sigma, \nu)$ and $Y_2 \sim TF(\mu, \sigma, \tau)$ in (11.6).

**11.3.4 Splicing using three components**

Splicing has also been used to introduce robustness into the normal distribution, as in the NET distribution below.

**Normal-exponential-t (NET)**

The *normal-exponential-t* family for $-\infty < Y < \infty$, denoted by NET($\mu, \sigma, \nu, \tau$), Rigby and Stasinopoulos (1994), is defined by $Y = \mu + \sigma Z$, where $Z$ has a standard normal density function for $|Z| < \nu$, an exponential density function for $\nu \leq |Z| < \tau$, and a Student $t$ density function for $|z| \geq \tau$, given by

$$f_Z(z) = \pi_1 f_{Z_1}(z) I(|z| < \nu) + \pi_2 f_{Z_2}(z) I(\nu < |z| < \tau) + \pi_3 f_{Z_3}(z) I(|z| > \tau)$$  \hfill (11.10)
where $Z_1 \sim \text{NO}(0,1)$, $Z_2 \sim \text{EXP}(\nu)$, $Z_3 \sim \text{TF}(0,1,\nu\tau - 1) \equiv \text{t}_{\nu\tau - 1}$ and $\pi_1$, $\pi_2$ and $\pi_3$ are defined to ensure $f_Z(z)$ is a proper density function and to ensure continuity of $f_Z(z)$ at $\nu$ and $\tau$. In \texttt{gamiss()} parameters $\nu$ and $\tau$ are constants which are either chosen by the user or estimated using the function \texttt{prof.dev()}. For fixed $\nu$ and $\tau$, the NET distribution has bounded influence functions for both $\mu$ and $\sigma$, Rigby and Stasinopoulos (1994), and hence provides a robust method of estimating $\mu$ and $\sigma$ for contaminated normal data.

### 11.4 Distributions generated by a mixture of distributions

A distribution for $Y$ can be generated by assuming that a parameter $\gamma$ of a distribution for $Y$ itself comes from a distribution.

Assume that, given $\gamma$, $Y$ has conditional probability (density) function $f(y|\gamma)$ and marginally $\gamma$ has probability (density) function $f(\gamma)$. Then the marginal of $Y$ is given by

$$f_Y(y) = \begin{cases} \int f(y|\gamma)f(\gamma)d\gamma, & \text{if } \gamma \text{ is continuous,} \\ \sum f(y|\gamma)p(\gamma = \gamma_i), & \text{if } \gamma \text{ is discrete.} \end{cases} \tag{11.11}$$

The marginal distribution of $Y$ is called a continuous mixture distribution if $\gamma$ is continuous and a discrete (or finite) mixture distribution is $\gamma$ is discrete.

Discrete (or finite) mixture distributions are considered in detail in Chapter ???. Continuous mixture density functions may be explicitly defined if the integral in (11.11) is tractable. This is dealt with in this section. However the integral in (11.11) is often intractable (and so the density functions is not explicitly defined), but may be approximated, e.g. using Gaussian quadrature points. This is dealt with in Section ??, where the model is viewed as a random effect model at the observational level.

#### 11.4.1 Explicitly defined continuous mixture distributions

The marginal distribution of $Y$ will, in general, be continuous if the conditional distribution of $Y$ is continuous. A summary of explicit continuous mixture distributions for $Y$ generated by (11.11) is given in Table 11.3.

**Student $t$ family (TF)**

The \textit{Student $t$} family for $-\infty < Y < \infty$, denoted TF($\mu, \sigma, \nu$), may be generated from a continuous mixture by assuming $Y|\gamma \sim \text{NO}(\mu, \gamma)$ and $\gamma \sim \sqrt{\nu}\sigma\chi^{-1}_{\nu} \equiv \text{GG}(\sigma, [2\nu]^{-1/2}, -2)$ has a scale inverted Chi distribution (which is a special case of the generalised gamma distribution), Box and Tiao (1973).

**Generalised $t$ (GT)**

The \textit{generalised $t$} family for $-\infty < Y < \infty$, denoted GT($\mu, \sigma, \nu, \tau$), may be generated by assuming $Y|\gamma \sim \text{PE2}(\mu, \gamma, \tau)$ has a power exponential type 2 distribution and $\gamma \sim \text{GG2}(-\tau, \sigma\nu^{1/\tau}, \nu)$ has a generalised gamma type 2 distribution, McDonald (1991).
11.5. DISTRIBUTIONS GENERATED BY UNIVARIATE TRANSFORMATION

| Distributions of $Y$ | Distribution of $Y|\gamma$ | Distribution of $\gamma$ | References |
|---------------------|---------------------------|-------------------------|------------|
| TF($\mu, \sigma, \nu$) | NO($\mu, \gamma$) | GG($\sigma, [2\nu]^{-1/2}, -2$) | Box and Tiao (1973) |
| GT ($\mu, \sigma, \nu, \tau$) | PE2($\mu, \gamma, \tau$) | GG($-\tau, \sigma\nu^{1/\tau}, \nu$) | McDonald (1991) |
| GB2($\mu, \sigma, \nu, \tau$) | GG2($\sigma, \gamma, \nu$) | GG2($-\sigma, \mu, \tau$) | McDonald (1996) |
| EGB2($\mu, \sigma, \nu, \tau$) | EGG2($1/\sigma, \gamma, \nu$) | GG2($-1/\sigma, e^\mu, \tau$) | McDonald (1996) |

Table 11.3: Showing distributions generated by continuous mixtures

**Generalised Beta type 2 (GB2)**

The generalised beta type 2 family for $Y > 0$, denoted GB2($\mu, \sigma, \nu, \tau$), may be generated by assuming $Y|\gamma \sim$ GG2($\sigma, \gamma, \nu$) and $\gamma \sim$ GG2($-\sigma, \mu, \tau$), McDonald (1996).

**Exponential Generalised Beta type 2 (EGB2)**

The exponential generalised beta type 2 family for $-\infty < Y < \infty$, denoted EGB2($\mu, \sigma, \nu, \tau$) may be generated by assuming $Y|\gamma \sim$ EGG2($1/\sigma, \gamma, \nu$) has an exponential generalised gamma type 2 distribution and $\gamma \sim$ GG2($-1/\sigma, e^\mu, \tau$), McDonald (1996). [Note that the exponential generalised gamma type 2 distribution is defined by: if $Z \sim$ EGG2($\mu, \sigma, \nu$) then $e^Z \sim$ GG2($\mu, \sigma, \nu$)].

11.5. DISTRIBUTIONS GENERATED BY UNIVARIATE TRANSFORMATION

Maybe we should have this earlier. Also we need to define here the log-Normal and probably the logic normal as an exercise.

The following is the general rule applied when transforming from a random variable to another. Let $Z$ be a continuous random variable with known pdf defined on the space $A$. Let the new variable be $Y = g(Z)$, where the function $g()$ is a one-to-one transformation that maps the set $Z \in A$ onto the set $Y \in B$. Let the inverse of $g()$ be $z = g^{-1}(y) = h(y)$ with continuous and non-zero first derivative $\frac{dz}{dy} = h'(y)$ for all points in set $B$, then the pdf of $Y$ is given by:

$$f_Y(y) = f_Z(h(y)) |h'(y)| = f_Z(h(y)) \left| \frac{dz}{dy} \right|$$

(11.12)

Note that $F_Y(y) = F_Z(z)$ provided that the function $g()$ is a monotonic increasing function.

If the location parameter $\mu$ is the median for the distribution for $Z$ then $h(\mu)$ is the median for the distribution of $Y$ since $F_Z(Z > \mu) = F_Y(h(y) > h(\mu)) = 0.5$. Therefore if we want the location parameter for $Y$ to have some meaning we could reparametrise from the $\mu$ of $Z$, i.e. $\mu_Z$, by setting $\mu_Y = h(\mu_Z)$. Note that the GAMLSS algorithm requires the first and expected second derivative in the fitting process so would need to change:

$$\frac{d\ell_Y}{d\mu_Y} = \frac{d\ell_Z}{d\mu_Z} \frac{d\mu_Z}{d\mu_Y}$$

and

$$\frac{d^2\ell_Y}{d\mu_Y^2} = \frac{d^2\ell_Z}{d\mu_Z^2} \left( \frac{d\mu_Z}{d\mu_Y} \right)^2 + \frac{d\ell_Z}{d\mu_Z} \frac{d^2\mu_Z}{d\mu_Y^2}.$$
Example 1: the log family of distributions

Consider the case of \( Z \) defined on the range \((-\infty, +\infty)\) then \( Y = \exp(Z) \) i.e. \( Z = \log(Y) \) is defined on the positive real line \((0, +\infty)\). The pdf of \( Y \) will be:

\[
 f_Y(y) = f_Z(\log(y)) \left| \frac{1}{y} \right| \tag{11.13}
\]

since \( z = \log(y) \) and \( \frac{dz}{dy} = \frac{1}{y} \).

The classic example in this case is the log-normal distribution (LOGNO) where it is assumed that the random variable \( Z \) has a normal NO\( \mu, \sigma \) distribution. The \( Y \) has a log-normal distribution with a pdf given by

\[
 f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma^2}} \frac{1}{y} \exp \left\{ -\frac{[\log(y) - \mu]^2}{2\sigma^2} \right\} \tag{11.14}
\]

for \( Y > 0 \).

A different parametrisation could be obtained by setting \( \mu_Y = \exp(\mu) \) i.e. \( \mu = \log(\mu_Y) \) in 11.14.

Example 2: the logit family of distributions

Consider the case of \( Z \) defined on the range \((-\infty, +\infty)\) then \( Y = \frac{1}{1+\exp(-Z)} \) i.e. \( Z = \log\left(\frac{Y}{1-Y}\right) \) is defined on \((0, 1)\). The pdf of \( Y \) will be:

\[
 f_Y(y) = f_Z(\text{logit}(y)) \left| \frac{1}{y} + \frac{1}{1-y} \right| \tag{11.15}
\]

since \( z = \text{logit}(y) = \log\left(\frac{y}{1-y}\right) \) and \( \frac{dz}{dy} = \frac{1}{y} + \frac{1}{1-y} \).

Let \( Z \) have a normal NO\( (\mu, \sigma) \) distribution then \( Y \) has a logit-normal distribution with a pdf given by:

\[
 f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi \sigma^2 y(1-y)}} \exp \left\{ -\frac{\left[\log\left(\frac{y}{1-y}\right) - \mu\right]^2}{2\sigma^2} \right\} \tag{11.16}
\]

where \( 0 < y < 1 \). A different parametrisation could be obtained by setting \( \mu_Y = \frac{1}{1+\exp(-\mu)} \) i.e. \( \mu = \log\left(\frac{\mu_Y}{1-\mu_Y}\right) \).

Many three and four parameter families of continuous distribution for \( Y \) can be defined by assuming that a transformed variable \( Z \), obtained from \( Y \), has a simple well-known distribution. The parameters of the distribution of \( Y \) may come from parameters of the univariate transformation or from parameters of the distribution of \( Z \) or both. Below we consider distributions available in GAMLSS which can be obtained by a univariate transformation.

Box-Cox, Cole and Green (BCCG)

The Box-Cox Cole and Green family for \( Y > 0 \) used by Cole and Green (1992), denoted by BCCG\((\mu, \sigma, \nu)\), assumes that \( Z \) has a standard normal distribution, NO\((0, 1)\), with mean 0 and standard deviation 1, where

\[
 Z = \begin{cases} 
 \frac{1}{\sigma^\nu} \left[ \left( \frac{Y}{\mu} \right)^\nu - 1 \right], & \text{if } \nu \neq 0, \\
 \frac{1}{\sigma} \log\left( \frac{Y}{\mu} \right), & \text{if } \nu = 0. 
\end{cases} \tag{11.17}
\]
11.5. DISTRIBUTIONS GENERATED BY UNIVARIATE TRANSFORMATION

<table>
<thead>
<tr>
<th>Distributions of $Y$</th>
<th>Random variable $Z$</th>
<th>Transformation to $Z$</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCCG</td>
<td>NO(0, 1)</td>
<td>(11.17)</td>
<td>Cole and Green (1992)</td>
</tr>
<tr>
<td>BCPE</td>
<td>PE(0, 1, $\tau$)</td>
<td>(11.17)</td>
<td>Rigby and Stasinopoulos (2004)</td>
</tr>
<tr>
<td>BCT</td>
<td>TF(0, 1, $\tau$)</td>
<td>(11.17)</td>
<td>Rigby and Stasinopoulos (2006)</td>
</tr>
<tr>
<td>EGB2</td>
<td>F(2$\nu$, 2$\tau$)</td>
<td>(11.18)</td>
<td>Johnson et al. (1995) p.142</td>
</tr>
<tr>
<td>GB1</td>
<td>BE($\mu$, $\sigma$)</td>
<td>(11.19)</td>
<td>McDonald and Xu (1995)</td>
</tr>
<tr>
<td>GB2</td>
<td>F(2$\nu$, 2$\tau$)</td>
<td>($\tau/\nu$)(Y/$\mu$)$^\nu$</td>
<td>McDonald and Xu (1995)</td>
</tr>
<tr>
<td>GG</td>
<td>GA(1, $\nu$)</td>
<td>(Y/$\mu$)$^\nu$</td>
<td>Lopatatazidis and Green (2000)</td>
</tr>
<tr>
<td>JSUo</td>
<td>NO(0, 1)</td>
<td>(11.20)</td>
<td>Johnson (1949)</td>
</tr>
<tr>
<td>JSU</td>
<td>NO(0, 1)</td>
<td>(14.15)</td>
<td>Rigby and Stasinopoulos (2006)</td>
</tr>
<tr>
<td>PE</td>
<td>GA(1, $\nu^{1/2}$)</td>
<td>$\nu$</td>
<td>Nelson (1991)</td>
</tr>
<tr>
<td>SHASHo</td>
<td>NO(0, 1)</td>
<td>(11.22)</td>
<td>Jones and Pewsey (2009)</td>
</tr>
<tr>
<td>SHASH</td>
<td>NO(0, 1)</td>
<td>(11.23)</td>
<td>Jones (2005)</td>
</tr>
<tr>
<td>ST3</td>
<td>BEo($\alpha$, $\beta$)</td>
<td>(11.24)</td>
<td>Jones and Faddy (2003)</td>
</tr>
</tbody>
</table>

Table 11.4: Showing distributions generated by univariate transformation

Cole and Green (1992) were the first to model all three parameters of a distribution as nonparametric smooth functions of a single explanatory variable. Note that the parameterization above is different from and more orthogonal than the one used originally by Box and Cox (1964). Rigby and Stasinopoulos (2000) and Stasinopoulos et al. (2000) used the original parameterization, $Z = (Y^{\nu} - 1)/\nu$ (if $\nu \neq 0$) + log $(Y)$ (if $\nu = 0$) where $Z \sim \text{NO}(\mu, \sigma)$, to model the mean $\mu$ and the variance $\sigma^2$ of $Z$ as functions of explanatory variables for a constant $\nu$. They obtained the maximum likelihood estimate of the power parameter $\nu$ from its profile likelihood. This model for $Y > 0$ is denoted by $\text{LNO}\{\mu, \sigma, \nu\}$ where $\nu$ is fixed by the user in the GAMLSS software.

**Box-Cox Power Exponential (BCPE)**

The *Box-Cox power exponential* family for $Y > 0$, denoted by $\text{BCPE}(\mu, \sigma, \nu, \tau)$, is defined by assuming $Z$ given by (11.17) has a (truncated) standard Power Exponential distribution, $\text{PE}(0, 1, \tau)$, see Rigby and Stasinopoulos (2004). This distribution is useful for modelling (positive or negative) skewness combined with (lepto or platy) kurtosis in continuous data.

**Box-Cox t (BCT)**

The *Box-Cox t* family for $Y > 0$, denoted by $\text{BCT}(\mu, \sigma, \nu, \tau)$, is defined by assuming $Z$ given by (11.17) has a (truncated) standard $t$ distribution with $\tau$ degrees of freedom, i.e. $\text{TF}(0, 1, \tau)$, see Rigby and Stasinopoulos (2006).

**Exponential generalised beta type 2 (EGB2)**

The *exponential generalised beta type 2* family for $-\infty < Y < \infty$, denoted by $\text{EGB2}(\mu, \sigma, \nu, \tau)$, assumes that $\exp(Y)$ has a generalised beta type 2 distribution. This distribution was called the exponential generalised beta of the second kind by McDonald (1991) and was investigated by McDonald and Xu (1995). The distribution may also be defined by assuming the $Z$ has an $F$ distribution with degrees of freedom $2\nu$ and $2\tau$, i.e. $Z \sim F_{2\nu, 2\tau}$, where

$$Z = (\tau/\nu) \exp\left[\left(Y - \mu\right)/\sigma\right],$$

(11.18)
CHAPTER 11. METHODS OF GENERATING DISTRIBUTIONS

Johnson et al. (1995) p142. The distribution has also been called a generalised logistic distribution type IV, see Johnson et al. (1995) p142, who report its long history from Perks (1932). Note also that \( R = \exp \left( \frac{(Y - \mu)}{\sigma} \right) \) has a beta distribution of the second kind \( BE_2(\nu, \tau) \), Johnson et al. (1995) p248 and p325 and \( B = R/(1 + R) \) has an original beta \( BE_0(\nu, \tau) \) distribution.

**generalised Beta type 1 (GB1)**

The **generalised beta type 1** family for \( 0 < Y < 1 \), denoted by \( GB1(\mu, \sigma, \nu, \tau) \), is defined by assuming \( Z \) has a beta, \( BE(\mu, \sigma) \), distribution where

\[
Z = \frac{Y^\tau}{\nu + (1 - \nu)Y^\tau}
\]

(11.19)

where \( 0 < \mu < 1, 0 < \sigma < 1, \nu > 0 \) and \( \tau > 0 \). Note that GB1 always has range \( 0 < y < 1 \) and so is different from the generalised beta of the first kind, McDonald and Xu (1995), whose range depends on the parameters.

Note that for \( 0 < \nu \leq 1 \) only, \( GB1(\mu, \sigma, \nu, \tau) \) is a reparameterization of the submodel with range \( 0 < Y < 1 \) of the five parameter generalised beta, \( GB(a, b, c, p, q) \) distribution of McDonald and Xu (1995) given by

\[
GB1(\mu, \sigma, \nu, \tau) \equiv GB \left( \frac{\nu}{\nu + (1 - \nu)Y^\tau}, 1 - \nu, \mu (\sigma^{-2} - 1), (1 - \mu) (\sigma^{-2} - 1) \right).
\]

Note also that \( \tau = 1 \) in \( GB1(\mu, \sigma, \nu, \tau) \) gives a reparametrisation of the generalised 3 parameter beta distribution, \( G3B(\alpha_1, \alpha_2, \lambda) \), distribution, Pham-Gia and Duong (1989) and Johnson et al. (1995) p251, given by \( G3B(\alpha_1, \alpha_2, \lambda) = GB1 \left( \frac{\alpha_1}{\alpha_1 + \alpha_2}, (\alpha_1 + \alpha_2 - 1)^{-1/2}, 1/\lambda, 1 \right) \). Hence \( G3B(\alpha_1, \alpha_2, \lambda) \) is a reparameterized submodel of \( GB1(\mu, \sigma, \nu, \tau) \).

**generalised Beta type 2 (GB2)**

The **generalised beta type 2** family for \( Y > 0 \), McDonald (1996), denoted by \( GB2(\mu, \sigma, \nu, \tau) \), is defined by assuming \( Z \) has an \( F \) distribution with degrees of freedom \( 2\nu \) and \( 2\tau \), i.e. \( Z \sim F_{2\nu,2\tau} \), where \( Z = \frac{Y}{\mu} \nu (Y/\mu)^\nu \).

The distribution is also called the generalised beta distribution of the second kind. Note also that \( R = (Y/\mu)^\nu \) has a beta distribution of the second kind, \( BE_2(\nu, \tau) \), Johnson et al. (1995) p248 and p325 and \( B = R/(1 + R) \) has an original beta, \( BE_0(\nu, \tau) \), distribution.

**generalised Gamma (GG, GG2)**

The **generalised gamma** family for \( Y > 0 \), parameterized by Lopatatzidis and Green (2000), denoted by \( GG(\mu, \sigma, \nu) \), assumes that \( Z \) has a gamma \( GA(1, \sigma\nu) \) distribution with mean 1 and variance \( \sigma^2\nu^2 \), where \( Z = (Y/\mu)^\nu \). A reparametrisation of \( GG(\mu, \sigma, \nu) \), Johnson et al. (1995) p401, given by setting \( \mu = \alpha_2\alpha_3^{1/\alpha_1}, \sigma = (\alpha^2_1\alpha_3)^{-1/2} \) and \( \nu = \alpha_1 \), is denoted \( GG2(\alpha_1, \alpha_2, \alpha_3) \).

**Johnson Su (JSUo, JSU)**

The original Johnson Su family for \( -\infty < Y < \infty \), denoted by \( JSUo(\mu, \sigma, \nu, \tau) \), Johnson (1949), is defined by assuming

\[
Z = \nu + \tau \sinh^{-1}[\frac{(Y - \mu)}{\sigma}]
\]

(11.20)
has a standard normal distribution.

The reparameterized Johnson Su family, for $-\infty < Y < \infty$, denoted by JSU($\mu, \sigma, \nu, \tau$), has exact mean $\mu$ and standard deviation $\sigma$ for all values of $\nu$ and $\tau$, see Appendix 14.3.3 for details.

### Power Exponential (PE, PE2)

The **power exponential** family for $-\infty < Y < \infty$, denoted by PE($\mu, \sigma, \nu$) is defined by

$$f_Y(y) = \frac{\nu}{2\sigma\Gamma(1/\nu)} \exp\left\{ -\frac{|y - \mu|^{\nu}}{\nu \sigma} \right\}$$

(11.21)

where $c = [\Gamma(1/\nu)/\Gamma(3/\nu)]^{1/2}$, and $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$.

This parameterization, used by Nelson (1991), ensures that $\mu$ and $\sigma$ are the mean and standard deviation of $Y$ respectively for all $v > 0$. This distribution assumes that $Z = \nu \left| \frac{Y - \mu}{\sigma} \right|^{\nu}$ has a gamma $\text{GA}(1, v^{1/2})$ distribution. A reparametrization of $\text{PE}(\mu, \sigma, \nu)$ used by Nandi and Mämpel (1995), denoted by PE2($\alpha_1, \alpha_2, \alpha_3$), is given by setting $\alpha_1 = \theta$, $\alpha_2 = \phi^{2/3}$ and $\alpha_3 = 2/3$. Box and Tiao (1973) p 157 equations (3.2.3) and (2.2.5) are respectively reparameterizations of the Subbotin parameterization and (11.21) in which $\delta = 1 + \beta$ and $\nu = 2/(1 + \beta)$. The distribution is also called the exponential power distribution or the Box-Tiao distribution.

### Sinh-Arcsinh (SHASHo, SHASHo2, SHASH)

The original **sinh-arcsinh** family for $-\infty < Y < \infty$, Jones and Pewsey (2009), denoted by SHASHo($\mu, \sigma, \nu, \tau$), is defined by assuming that $Z$ has a standard normal distribution $\text{NO}(0, 1)$, where

$$Z = \sinh \left\{ \tau \sinh^{-1}(R) - \nu \right\}$$

(11.22)

where $R = (Y - \mu)/\sigma$.

Jones and Pewsey (2009) suggest a more stable re-parametrization SHASHo2 of SHASHo by setting $R = (Y - \mu)/(\tau \sigma)$ above.

The **sinh-arcsinh** family for $-\infty < Y < \infty$, Jones (2005), denoted by SHASH($\mu, \sigma, \nu, \tau$), is defined by assuming that $Z$ has a standard normal distribution $\text{NO}(0, 1)$, where

$$Z = \frac{1}{2} \left\{ \exp \left[ \tau \sinh^{-1}(R) \right] - \exp \left[ -\nu \sinh^{-1}(R) \right] \right\}$$

(11.23)

where $R = (Y - \mu)/\sigma$.

### Skew t type 5 (ST5)

The **skew t type 5** family for $-\infty < Y < \infty$, Jones and Faddy (2003), denoted by ST5($\mu, \sigma, \nu, \tau$), assumes that $Z$ has a beta $\text{BEo}(\alpha, \beta)$ distribution with $f_Z(z) = z^{\alpha-1} (1 - z)^{\beta-1} / B(\alpha, \beta)$ where

$$Z = \frac{1}{2} \left[ 1 + R/(\alpha + \beta + R^2)^{1/2} \right]$$

(11.24)

where $R = (Y - \mu)/\sigma$, $\alpha = \tau^{-1} \left[ 1 + \nu(2\tau + \nu^2)^{-1/2} \right]$ and $\beta = \tau^{-1} \left[ 1 - \nu(2\tau + \nu^2)^{-1/2} \right]$.  

---

**11.5. DISTRIBUTIONS GENERATED BY UNIVARIATE TRANSFORMATION**

191
11.6 Distributions generated by transformation from two or more random variables

Distributions can be generated from a function of two (or more) random variables.

**Student t family (TF)**

The Student t family for $-\infty < Y < \infty$ (e.g. Lange et al., 1989), denoted by TF($\mu, \sigma, \nu$), is defined by assuming that $Y = \mu + \sigma T$ where $T \sim t_\nu$ has a standard t distribution with $\nu$ degrees of freedom, defined itself by $T = Z(W/\nu)^{-1/2}$ where $Z \sim \text{NO}(0, 1)$ and $W \sim \chi^2_\nu \equiv \text{GA}(\nu, [2/\nu]^{1/2})$, a Chi-square distribution with $\nu$ degrees of freedom treated as a continuous parameter, and where $Z$ and $W$ are independent random variables.

**Skew t type 2 (ST2)**

The skew t type 2 family for $-\infty < Y < \infty$, Azzalini and Capitanio (2003), denoted ST2($\mu, \sigma, \nu, \tau$), is defined by assuming that $Y = \mu + \sigma T$ where $T = Z(W/\tau)^{-1/2}$ and $Z \sim \text{SN}(0, 1, \nu)$ has a skew normal type 1 distribution (see Section 11.2) and $W \sim \chi^2_\tau \equiv \text{GA}(\tau, [2/\tau]^{1/2})$ has a Chi-square distribution with $\tau > 0$ degrees of freedom, and where $Z$ and $W$ are independent random variables. Note that $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$ and $\tau > 0$.

The distribution ST2($\mu, \sigma, \nu, \tau$) is the one dimensional special case of the multivariate skew t used in R package Sn, Azzalini (2006).

An important special case of a function of two independent random variables is their sum, i.e. $Y = Z_1 + Z_2$. The probability density function of $Y$ is obtained by convolution, i.e.

$$f_Y(y) = \int_{-\infty}^{y} f_{Z_1}(z)f_{Z_2}(y-z)dz. \quad (11.25)$$

The following are two examples.

**Exponential Gaussian (exGAUS)**

If $Z_1 \sim \text{NO}(\mu, \sigma)$ and $Z_2 \sim \text{EXP}(\nu)$ in (11.25), then $Y = Z_1 + Z_2$ has an exponential Gaussian distribution, denoted by exGAUS($\mu, \sigma, \nu$), for $-\infty < Y < \infty$. The distribution has been also called a lagged normal distribution, Johnson et al. (1994), p 172.

**generalised Erlangian**

As pointed out by Johnson et al. (1994), p 172, the convolution of two or more exponential probability density functions with different mean parameters gives the generalised Erlangian distribution, while the convolution of a normal, $\text{NO}(\mu, \sigma)$, with a generalised Erlangian probability density function gives a generalised lagged normal distribution, see Davis and Kutner (1976).

11.7 Truncation distributions

A truncated distribution can be created from any distribution by restricting the range the values of its random variable $Y$. The are three types of truncation deepening on which size the truncation is performed. Let $c_l$ and $c_r$ so $c_l < c_r$ be constants defined within the range $R_Y$ of all possible values of the random variable $Y$. Then the resulting distribution is called:
1. **left** truncated distribution if \( c_l \leq Y \)
2. **right** truncated distribution if \( Y < c_r \) and
3. truncated in **both** size distribution if \( c_l \leq Y < c_r \)

Note that for continuous distributions the less or equal sign ‘\( \leq \)’ does not matter but it does in the definition of discrete truncated distributions where we took the convention that in the left truncation the values \( c_l \) is included in the range but not the value \( c_r \) in the right truncation.

In general the following results are relevant to left, right and both sizes truncation.

### 11.7.1 Left truncation

Let \( Y \) denote the random variable left truncated at \( c_l \), so \( c_l \leq Y \) and \( Y \) the original random variable with pdf \( f_{Y_o}(Y_o) \) and cdf \( F_{Y_o}(Y_o) \). Then the (probability) distribution function of the left truncated random variable \( Y \) is

\[
f_{Y}(y) = \frac{f_{Y_o}(Y_o)}{1 - F_{Y_o}(c_l)}
\] (11.26)

with commutative distribution function,

\[
F_{Y}(y) = \frac{F_{Y_o}(Y_o) - F_{Y_o}(c_l)}{1 - F_{Y_o}(c_l)}
\] (11.27)

and inverse commutative distribution function

\[
q = F_{Y_o}^{-1} \{ F_{Y_o}(c_l) + p [1 - F_{Y_o}(c_l)] \}.
\] (11.28)

Are there any results for how the truncated mean and variance are related with the original parameters?

Also for defined and calculate the likelihood function and its maximum the following results may be useful. The log-likelihood for one observation is defined as:

\[
\ell_Y = \log f_Y(y) = \log f_{Y_o}(y) + \log [1 - F_{Y_o}(c_l)]
\] (11.29)

For any parameter \( \theta \) in \((\mu, \sigma, \nu, \tau)\) the first and second derivatives are given by:

\[
\frac{\partial \ell_Y}{\partial \theta} = \frac{\partial \ell_{Y_o}}{\partial \theta} - \frac{\partial}{\partial \theta} \{ \log [1 - F_{Y_o}(c_l)] \}
\] (11.30)

and

\[
\frac{\partial^2 \ell_Y}{\partial \theta^2} = \frac{\partial^2 \ell_{Y_o}}{\partial \theta^2} - \frac{\partial^2}{\partial \theta^2} \{ \log [1 - F_{Y_o}(c_l)] \}
\] (11.31)

### 11.7.2 Right truncation

Let \( Y \) denote the random variable right truncated at \( c_r \), so \( Y < c_r \) and \( Y_o \) the original random variable with pdf \( f_{Y_o}(Y_o) \) and cdf \( F_{Y_o}(Y_o) \). Then the (probability) distribution function of \( Y \) is

\[
f_{Y}(y) = \frac{f_{Y_o}(Y_o)}{F_{Y_o}(c_r)}
\] (11.32)
with commutative distribution function

\[ F_Y(y) = \frac{F_{Y_o}(Y_o)}{F_{Y_o}(c_r)} \tag{11.33} \]

and inverse commutative distribution function

\[ q = F^{-1}_{Y_o} \{ p [F_{Y_o}(c_r)] \} \tag{11.34} \]

The log-likelihood for one observation is defined as:

\[ \ell_Y = \log f_Y(y) = \log f_{Y_o}(y) - \log [F_{Y_o}(c_r)] \tag{11.35} \]

For any parameter \( \theta \) in \((\mu, \sigma, \nu, \tau)\) the first and second derivatives are given by:

\[ \frac{\partial \ell_Y}{\partial \theta} = \frac{\partial \ell_{Y_o}}{\partial \theta} - \frac{\partial}{\partial \theta} \{ \log [F_{Y_o}(c_r)] \} \tag{11.36} \]

and

\[ \frac{\partial^2 \ell_Y}{\partial \theta^2} = \frac{\partial^2 \ell_{Y_o}}{\partial \theta^2} - \frac{\partial^2}{\partial \theta^2} \{ \log [F_{Y_o}(c_r)] \} \tag{11.37} \]

### 11.7.3 Both sizes truncation

Let \( Y \) denote the random variable left truncated at \( c_l \) and right truncated at \( c_r \), so \( c_l \leq Y < c_r \) and \( Y_o \) the original random variable with pdf \( f_{Y_o}(Y_o) \) and cdf \( F_{Y_o}(Y_o) \). Then the (probability) distribution function of \( Y \) is

\[ f_Y(y) = \frac{f_{Y_o}(Y_o)}{F_{Y_o}(c_r) - F_{Y_o}(c_l)} \tag{11.38} \]

with commutative distribution function

\[ F_Y(y) = \frac{F_{Y_o}(Y_o) - F_{Y_o}(c_l)}{F_{Y_o}(c_r) - F_{Y_o}(c_l)} \tag{11.39} \]

and inverse commutative distribution function

\[ q = F^{-1}_{Y_o} \{ p [F_{Y_o}(c_r) - F_{Y_o}(c_l)] + F_{Y_o}(c_r) \} \tag{11.40} \]

The log-likelihood for one observation is defined as:

\[ \ell_Y = \log f_Y(y) = \log f_{Y_o}(y) + \log [F_{Y_o}(c_r) - F_{Y_o}(c_l)] \tag{11.41} \]

For any parameter \( \theta \) in \((\mu, \sigma, \nu, \tau)\) the first and second derivatives are given by:

\[ \frac{\partial \ell_Y}{\partial \theta} = \frac{\partial \ell_{Y_o}}{\partial \theta} - \frac{\partial}{\partial \theta} \{ \log [F_{Y_o}(c_r) - F_{Y_o}(c_l)] \} \tag{11.42} \]

and

\[ \frac{\partial^2 \ell_Y}{\partial \theta^2} = \frac{\partial^2 \ell_{Y_o}}{\partial \theta^2} - \frac{\partial^2}{\partial \theta^2} \{ \log [F_{Y_o}(c_r) - F_{Y_o}(c_l)] \} \tag{11.43} \]
11.8 Systems of distributions

11.8.1 Pearson system

The Pearson system of probability density functions $f_Y(y|\theta)$, where $\theta^\top = (\theta_1, \theta_2, \theta_3, \theta_4)$, is defined by solutions of the equation:

$$\frac{d}{dy} f_Y(y|\theta) = -\frac{\theta_1 + y}{\theta_2 + \theta_3 y + \theta_4 y^2} \quad (11.44)$$

The solutions of (11.44) fall into one of seven families of distributions called Type I to Type VII. Type I, IV, and VI cover disjoint regions of the skewness-kurtosis $(\sqrt{\beta_1}, \beta_2)$ space, while the other four types are boundary types, see Johnson et al. (1994), Figure 12.2. Type I is a shifted and scaled beta BE($\mu, \sigma$) distribution, with the resulting arbitrary range defined by two extra parameters. Type II is a symmetrical form of type I. Type III is a shifted gamma distribution. Types IV and V are not well known distribution (probably because the constants of integration are intractable). Type VI is a generalization of the $F$ distribution. Type VII is a scaled $t$ distribution, i.e. $TF(0, \sigma, \nu)$.

11.8.2 Stable distribution system

Stable distributions are defined through their characteristic function, given by Johnson et al. (1994) p57. In general their probability density function cannot be obtained explicitly (except using complicated infinite summations). McDonald (1996) and Lambert and Lindsey (1999) discuss the application of stable distributions to modelling stock returns.

11.8.3 Exponential Family

The exponential family of distributions $EF(\mu, \phi)$ is defined by the probability (density) function $f_Y(y|\mu, \phi)$ of $Y$ having the form:

$$f_Y(y|\mu, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\} \quad (11.45)$$

where $E(Y) = \mu = b'(\theta)$ and $V(Y) = \phi V'(\mu)$ where the variance function $V(\mu) = b''[\theta(\mu)]$. The form of (11.45) includes many important distributions including the normal, Poisson, gamma, inverse Gaussian and Tweedie, (Tweedie, 1984), distributions having variance functions $V(\mu) = 1, \mu, \mu^2, \mu^3$ and $\mu^p$ for $p < 0$ or $p > 1$, respectively, and also binomial and negative binomial distributions with variance functions $V(\mu) = \frac{\mu(1-\mu)}{N}$ and $V(\mu) = \mu + \frac{\mu}{\phi}$ respectively.

The exponential family for $Y$ with mean $\mu$ and variance having the form $\phi \mu^\nu$ (where $\phi = \sigma^2$ and $\sigma$ is a scale parameter), McCullagh and Nelder (1989), does not transform to a simple well known distribution. This is also called the Tweedie family. The probability (density) function exists only for $\nu \leq 0$ or $\nu > 1$ and suffers from being intractable (except using complicated series approximations) except for specific values $\nu = 0, 2, 3$. Furthermore, in general for $1 < \nu < 2$, the distribution is a combination of a mass probability at $Y = 0$ together with a continuous distribution for $Y > 0$, (which cannot be modelled independently), which is inappropriate for a continuous dependent variable $Y$, see Gilchrist (2000). This distribution is not currently available in GAMLSS.
11.8.4 generalised inverse Gaussian family

This family was developed by Jorgensen (1982).
Chapter 12

Heaviness of tails of continuous distributions

This chapter concentrates on the behaviour of the tails, in particular:

1. classify the tails of continuous distributions
2. provides method for identify the tail of a given data

This chapter is more theoretical and can be omitted for a practical course. Having said that this chapter should be of interest when the focus of the analysis is on the 'extreme' values rather than on the 'central middle' part of the distribution (beyond mean regression models).

12.1 Introduction

It should be clear from the previous chapters that distributions occurring in statistical practice vary considerably. This is because some distributions are symmetrical and some are markedly skew. Some are mesokurtic (Normal distribution) and some are markedly leptokurtic or platykurtic.

In particular, heavy tailed data has been observed in many applications in economics, finance, and natural sciences where the 'extreme' observations (or outliers) are not mistakes, but an essential part of the distribution. As a result, there are occasions when the tail of the distribution is of primary importance in the statistical analysis. Value at risk (VaR) and Expected shortfall (ES) are well known concept in financial analysis and have to do with how the tail of the distribution of the data is behaving. The important point here is that understanding of the heaviness of tails of distributions to fit heavy tailed data results in robust modelling (rather than just robust estimation) where the interest is in both estimating the regression coefficients and in fitting the error distribution, as advocated in Lange et al. (1989).\(^1\)

This Chapter is divided in three different sections. The first is introducing the basic concepts on how tails are behaving. The secondly classifies all continuous gamlss.family distributions into categories according to their tail behaviour. The third section is more practical in the sense that it tries to give guidance how to determine the tail behaviour of a given data set.

12.2 Types of tails for continuous distributions

Traditionally when we investigate the behaviour of the tail of a continuous distribution we are concentrating on the logarithm of the distribution, \( \log f_Y(y|\theta) \), rather than the distribution, \( f_Y(y|\theta) \), itself. This is because the logarithmic scale exaggerates the tail behaviour. Figure 12.1 show the logarithm of the standardised normal, Cauchy and Laplace distributions. Below -2.5 and above 2.5 the behaviour in the tail of those three distribution varies considerably. The logarithm of the normal distribution is quadratic, the logarithm of the Laplace is linear while logarithm of the Cauchy shows that the tail of this distribution degreases lot slower that the previous two. This justifies that the ordering of the heaviness of the tail of a continuous distribution should be based on the log of the probability density function. But as we will see below the same ordering is applied to the actual probability density function and to the survival function of a continuous distribution. The R code for creating figure 12.1 is given below:

```r
curve(dNO(x, log=T), -5, 5, xlim=c(-10, 10), ylab="log(pdf)", lwd=2)
curve(dTF(x, nu=1, log=T), -10, 10, add=TRUE, col=2, lty=2, lwd=2) # Cauchy
curve(dPE(x, nu=1, log=T), -10, 10, add=TRUE, col=3, lty=3, lwd=2) # Laplace
legend("topright", legend=c("Normal", "Cauchy", "Laplace"), text.col = c(1,2,3), lty = c(1, 2, 3), merge = TRUE, bg = 'gray90')
```

Figure 12.1: Figure showing the log of the standardised version of the normal, Cauchy and Laplace distributions

For continuous distributions defined in the range 0 to \( \infty \) the same idea applies. Figure 12.2 shows the logarithm of the exponential, Pareto type II and log-Normal distribution against \( Y \). The log of tail of the exponential distribution is decreases linearly. Both the Pareto and the log-normal distributions have heavier tails that the exponential distribution but the Pareto
becomes heavier than the log-Normal for greater values of $Y$. The R code for creating figure 12.2 is:

```r
curve(dPARETO2(x, sigma=1, log=T), 2, 30, col=2, lty=2, lwd=2, ylab="log(pdf)", xlab="y")
curve(dEXP(x, log=T), 2, 30, add=T, lwd=2)
curve(dLOGNO(x, mu=1, sigma=1, log=T), 2, 30, add=TRUE, col=3, lty=3, lwd=2)
legend("topright", legend=c("exponential", "Pareto 2", "log Normal"), t
ext.col = c(1,2,3), lty = c(1, 2, 3), merge = TRUE, bg = "gray90")
```

![Figure 12.2: Figure showing the log of the standardised version of the normal, Cauchy and Laplace distributions](image)

Next we define what we mean by a heavy tail distribution.

**definition**

If random variables $Y_1$ and $Y_2$ have continuous probability density functions $f_{Y_1}(y)$ and $f_{Y_2}(y)$ and

$$\lim_{y \to \infty} f_{Y_1}(y) = \lim_{y \to \infty} f_{Y_2}(y) = 0$$

then

$$Y_2 \text{ has a heavier right tail than } Y_1 \Leftrightarrow \lim_{y \to \infty} [\log f_{Y_2}(y) - \log f_{Y_1}(y)] = \infty.$$  

Note that the resulting ordering of $\log f_Y(y)$ for the right tail of $Y$ results in the same ordering for the probability density function $f_Y(y)$, where

$$Y_2 \text{ has a heavier tail than } Y_1 \Leftrightarrow f_{Y_1}(y) = o[f_{Y_2}(y)] \text{ as } y \to \infty$$
CHAPTER 12. HEAVINESS OF TAILS OF CONTINUOUS DISTRIBUTIONS

by Lemma B1 in Appendix B. It also the same ordering as the standard ordering for the survivor function \( F_Y(y) = 1 - F_Y(Y) \) where \( F_Y(y) \) is the cumulative distribution function, where

\[
Y_2 \text{ has a heavier tail than } Y_1 \iff \bar{F}_{Y_1}(y) = o \left[ \bar{F}_{Y_2}(y) \right],
\]

by Lemma B2 in Appendix B. Similarly for the left tail of \( Y \).

three types of tails

There are three main forms for \( \log f_Y(y) \) for a tail of \( Y \), i.e. as \( y \to \infty \) (for the right tail) or as \( y \to -\infty \) (for the left tail), \( \log f_Y(y) \sim \)

Type I: \[-k_2 \log |y|^{k_1},\]

Type II: \[-k_4 |y|^{k_3},\]

Type III: \[-k_6 e^{k_5 |y|},\]

in decreasing order of heaviness of the tail. That is, type I has heavier tail that type II and type III, while type III has the lightest tails go all.

For type I, \(-k_2 \log |y|^{k_1}\), decreasing \( k_1 \) results in a heavier tail, while decreasing \( k_2 \) for fixed \( k_1 \) results in a heavier tail. Similarly for type II, \(-k_4 |y|^{k_3}\), with \((k_3, k_4)\) replacing \((k_1, k_2)\) and for type III, \(-k_6 e^{k_5 |y|}\), with \((k_5, k_6)\) replacing \((k_1, k_2)\). Important special cases are \(k_1 = 1, k_2 = 2, k_3 = 1, k_4 = 2\). Figure 12.3 shows tail behaviour of the three different types for \(k_1, k_3, k_5 = 1, 2, k_2, k_4, k_6 = 1, 2\).

12.3 Classification Tables

Tables 12.1 and 12.2 provide a summary of many important distributions on the real line and positive real line respectively (note the use of \( k_1, k_2, k_3 \) and \( k_4 \) to associate the distributions with the three types of tails given above).

Many of the distributions in Tables 1 and 2 have important special cases. For example, the generalized beta type 2 distribution, GB2(\( \mu, \sigma, \nu, \tau \)), also known as the generalized beta-prime distribution and the generalized beta of the second kind, includes special cases the Burr III (or Dagum) distribution when \( \tau = 1 \), the Burr XII (or Singh-Maddala) when \( \nu = 1 \), (Johnson et al., 1994, p 54), a form of Pearson type VI when \( \sigma = 1 \), (Johnson et al., 1995, p 248), the generalized Pareto distribution when \( \sigma = 1 \) and \( \nu = 1 \) and the log logistic when \( \nu = 1 \) and \( \tau = 1 \). The skew exponential power type 3 distribution, SEP3(\( \mu, \sigma, \nu, \tau \)) includes the skew normal type 2 when \( \tau = 2 \), (Johnson et al., 1994, p 173).

The parametrizations of the distributions (column 2 of Tables 12.1 and 12.2) are those used by the gamlss package in R [Stasinopoulos and Rigby, 2007] and given by Stasinopoulos et al. (2008) (available from website http://www.gamlss.org). This parameterization was chosen for consistency as the parameters for all distributions (up to four parameters) are defined as \( \mu \), \( \sigma \), \( \nu \) and \( \tau \). Note that \( \mu \) and \( \sigma \) are (usually location and scale) parameters and not, in general, the mean and standard deviation of the distribution, while \( \nu \) and \( \tau \) are usually skewness and kurtosis parameters. Some distributions are parameterized in two different ways, for example JSU and JSUo. For many distributions the left and right tails have the same asymptotic form for \( \log f_Y(y) \), otherwise the relevant tail is specified in the table, see e.g. Gumbel distribution. Some distributions have different tail forms dependent on a condition on one (or more) parameters, see e.g. the generalized gamma distribution.
12.3. CLASSIFICATION TABLES

![Graphs showing the shape of the tail for different types of distributions for $k_1, k_3, k_5 = 1, 2$, and $k_2, k_4, k_6 = 1, 2$. Smaller values in the $k$'s result in heavier tails.]

Figure 12.3: Figure showing the shape of the tail for different types of distributions for $k_1, k_3, k_5 = 1, 2$, and $k_2, k_4, k_6 = 1, 2$. Smaller values in the $k$’s result in heavier tails.
### Table 12.1: Left and right tail asymptotic form of the log of the probability density function for continuous distributions on the real line, where \( c = \frac{1}{k} \) if \( 1 < k < 6 \) and \( \gamma > 1 \).

<table>
<thead>
<tr>
<th>Parameter Range</th>
<th>Distribution name</th>
<th>Distribution Condition</th>
<th>Value of ( \gamma )</th>
<th>Value of ( \theta )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Lognormal Lognormal} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Gumbel} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Reverse Gumbel} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Gumbel} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Reverse Gumbel} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Logistic} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Laplace} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Skew exponential power type 4 SEP4} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Sinh-arcsinh} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Sinh-arcsinh} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Power exponential type 2 PE2} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Power exponential} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Johnson's SU original JSUo} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Johnson's SU} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{TF} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Stable} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
<tr>
<td>( 0 &lt; \rho, \infty &gt; \eta &gt; \infty &gt; -\infty )</td>
<td>( \text{Skew} )</td>
<td>( \text{Left tail} )</td>
<td>( \Gamma(\theta) )</td>
<td>( \text{Right tail} )</td>
</tr>
</tbody>
</table>

Here, where \( \gamma = \frac{1}{k} \) if \( 1 < k < 6 \) and \( \gamma > 1 \).
<table>
<thead>
<tr>
<th>Value of $k_1-k_6$</th>
<th>Distribution name</th>
<th>Distribution</th>
<th>Condition</th>
<th>Value of $k_1-k_6$</th>
<th>Parameter range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k_1 = 1$</td>
<td>Box-Cox Cole-Green</td>
<td>BCCG($\mu, \sigma, \nu$)</td>
<td>$\nu &lt; 0$</td>
<td>$k_2 =</td>
<td>\nu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BCPE($\mu, \sigma, \nu, \tau$)</td>
<td>$\nu &lt; 0$</td>
<td>$k_2 =</td>
<td>\nu</td>
</tr>
<tr>
<td></td>
<td>Box-Cox t</td>
<td>BCT($\mu, \sigma, \nu, \tau$)</td>
<td>$\nu \leq 0$</td>
<td>$k_2 =</td>
<td>\nu</td>
</tr>
<tr>
<td></td>
<td>Generalized beta type 2</td>
<td>GB2($\mu, \sigma, \nu, \tau$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_2 = \nu \tau + 1$</td>
<td>$\tau &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Generalized gamma</td>
<td>GG($\mu, \sigma, \nu$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_2 = (\sigma^2</td>
<td>\nu</td>
</tr>
<tr>
<td></td>
<td>Inverse gamma</td>
<td>IG($\mu, \sigma$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_2 = \sigma^{-2} + 1$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>log $t$</td>
<td>LOGT($\mu, \sigma, \nu$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_2 = 1$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Pareto Type 2</td>
<td>PA2o($\mu, \sigma$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_2 = \sigma + 1$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td>$k_1 = 2$</td>
<td>Box-Cox Cole-Green</td>
<td>BCCG($\mu, \sigma, \nu$)</td>
<td>$\nu = 0$</td>
<td>$k_2 = 0.5\sigma^{-2}$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Lognormal</td>
<td>LOGNO($\mu, \sigma$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_2 = 0.5\sigma^{-2}$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Log Weibull</td>
<td>LOGWEI($\mu, \sigma$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_2 = 1$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td>$1 \leq k_1 &lt; \infty$</td>
<td>Box-Cox power exponential</td>
<td>BCPE($\mu, \sigma, \nu, \tau$)</td>
<td>$\nu = 0, \tau &gt; 1$</td>
<td>$k_1 = \tau, k_2 = (c_1\sigma)^{-\tau}$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\nu = 0, \tau = 1$</td>
<td>$k_1 = 1, k_2 = 1 + (c_1\sigma)^{-\tau}$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\nu = 0, \tau &lt; 1$</td>
<td>$k_1 = 1, k_2 = 1$</td>
<td>$\sigma &gt; 0$</td>
</tr>
<tr>
<td>$0 &lt; k_3 &lt; \infty$</td>
<td>Box-Cox Cole-Green</td>
<td>BCCG($\mu, \sigma, \nu$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_3 = 2\nu, k_4 = \frac{2\mu^2\nu^2\sigma^2}{3}$</td>
<td>$\mu &gt; 0, \sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Box-Cox power exponential</td>
<td>BCPE($\mu, \sigma, \nu, \tau$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_3 = \nu \tau, k_4 = [c_1\mu^\nu \sigma \nu]^{-\tau}$</td>
<td>$\mu &gt; 0, \sigma &gt; 0, \tau &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Generalized gamma</td>
<td>GG($\mu, \sigma, \nu$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_3 = \nu, k_4 = \left[c_1\mu^\nu \sigma \nu^2\right]^{-1}$</td>
<td>$\mu &gt; 0, \sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Weibull</td>
<td>WEI($\mu, \sigma$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_3 = \sigma, k_4 = \mu^{-\sigma}$</td>
<td>$\mu &gt; 0, \sigma &gt; 0$</td>
</tr>
<tr>
<td>$k_3 = 1$</td>
<td>Exponential</td>
<td>EX($\mu$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_4 = \mu^{-1}$</td>
<td>$\mu &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Gamma</td>
<td>GA($\mu, \sigma$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_4 = \mu^{-1} \sigma^{-2}$</td>
<td>$\mu &gt; 0, \sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Generalized inverse Gaussian</td>
<td>GIG($\mu, \sigma, \nu$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_4 = 0.5c_2\mu^{-1} \sigma^{-2}$</td>
<td>$\mu &gt; 0, \sigma &gt; 0$</td>
</tr>
<tr>
<td></td>
<td>Inverse Gaussian</td>
<td>IG($\mu, \sigma$)</td>
<td>$\nu &gt; 0$</td>
<td>$k_4 = 0.5 \mu^{-\sigma} \sigma^{-2}$</td>
<td>$\mu &gt; 0, \sigma &gt; 0$</td>
</tr>
</tbody>
</table>

Table 12.2: Right tail asymptotic form of the log of the probability density function for continuous distributions on the positive real line, where $c_2 = \left[K_{\nu+1}(\frac{1}{\sigma^\nu})\right]\left[K_{\nu}(\frac{1}{\sigma})\right]^{-1}$ where $K_\lambda(t) = \frac{1}{2} \int_0^\infty x^{\lambda-1} \exp\left\{-\frac{1}{2} t (x + x^{-1})\right\} dx$
Note, for example, that all distribution tails with \( k_1 = 1 \) are heavier than those with \( k_1 = 2 \). Within the \( k_1 = 1 \) group a smaller \( k_2 \) has the heavier tail. Note from Table 12.1 that the stable distribution and the skew t type 3 distribution with degrees of freedom parameter \( 0 < \tau < 2 \) have the same range for \( k_2 \). Distribution tails with \( 0 < k_3 < \infty \) can be:

- heavier than the Laplace (two sided exponential) if \( 0 < k_3 < 1 \),
- lighter than the Laplace but heavier than the normal if \( 1 < k_3 < 2 \),
- lighter than the normal if \( k_3 > 2 \).

It should also be noted that although the tails of two distributions with the same combination of \( k_1 \) and \( k_2 \) values, are not necessarily equally heavy, a reduction in \( k_2 \), no matter how small, for either distribution will make it the heavier tail distribution. Similarly replacing \((k_1, k_2)\) by \((k_3, k_4)\) or \((k_5, k_6)\). Hence the important point is that the \( k \) values are dominant in determining the heaviness of the tail of the distribution\(^2\).

Distribution tails in Tables 1 and 2 can be split into four categories: ‘non-heavy’ tails \((k_3 \geq 1 \text{ or } 0 < k_3 < \infty)\), ‘heavy’ tail (i.e. heavier than any exponential distribution) but lighter than any ‘Paretian type’ tail \((k_1 > 1 \text{ and } 0 < k_3 < 1)\), ‘Paretian type’ tail \((k_1 = 1 \text{ and } k_2 > 1)\), and heavier than any ‘Paretian type’ tail \((k_1 = 1 \text{ and } k_2 = 1)\). These four categories correspond closely to mild, slow, wild (pre or proper) and extreme randomness of Mandlebrot (1997), as shown by Table 12.3.

<table>
<thead>
<tr>
<th>Type</th>
<th>Conditions</th>
<th>Mandlebrot’s randomness</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-heavy</td>
<td>((k_3 \geq 1 \text{ or } 0 &lt; k_3 &lt; \infty))</td>
<td>mild</td>
</tr>
<tr>
<td>heavy</td>
<td>((k_1 &gt; 1 \text{ and } 0 &lt; k_3 &lt; 1))</td>
<td>slow</td>
</tr>
<tr>
<td>Paretian</td>
<td>((k_1 = 1 \text{ and } k_2 &gt; 1))</td>
<td>wild</td>
</tr>
<tr>
<td>heavier than Paretian</td>
<td>((k_1 = 1 \text{ and } k_2 = 1))</td>
<td>extreme</td>
</tr>
</tbody>
</table>

Following Lemma B3 and Corollaries C1 and C2, Tables 12.1 and 12.2 also apply to the asymptotic form of the log of the survivor function, \( \log \bar{F}_Y(y) \), with the following changes:

(i) when \( k_1 = 1 \) and \( k_2 > 1 \) then \( k_2 \) is reduced by 1 (e.g., if \( k_2 = 3 \), then \( k_2 = 2 \)),

(ii) when \( k_1 = 1 \) and \( k_2 = 1 \) then \( \log \bar{F}_Y(y) = o \left( \log |y| \right) \) and specific asymptotic forms for \( \log \bar{F}_Y(y) \) for specific distributions are given in Table 12.4.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Asymptotic form of ( \log \bar{F}_Y(y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOGT((\mu, \sigma, \nu))</td>
<td>(-\nu \log(\log y))</td>
</tr>
<tr>
<td>BCT((\nu = 0))</td>
<td>(-\tau \log(\log y))</td>
</tr>
<tr>
<td>LOG WEI((\mu, \sigma)) for all (0 &lt; \sigma &lt; \infty)</td>
<td>(-\mu^{-\sigma}(\log y)^\sigma)</td>
</tr>
<tr>
<td>BCPE((\nu = 0)) for all (0 &lt; \tau &lt; \infty)</td>
<td>(-(c_1 \sigma)^{-\tau}(\log y)^\tau)</td>
</tr>
</tbody>
</table>

Table 12.4: Asymptotic form of \( \log \bar{F}_Y(y) \) as \( y \to \infty \).

\(^2\)If it is required to distinguish between the two distributions with the same \( k \) values the second order terms of \( \log f_Y(y) \) can be compared.
Note that the distributions having log survivor function upper tails in exactly the forms $-k_2(\log y)^{k_1}$, $-k_4y^{k_3}$ and $-k_6e^{k_5y}$ are the log Weibull (LOGWEI), the Weibull (WEI) and the Gumbel (GU), respectively.

12.4 Methods for choosing the appropriate tail

The substantive practical implications of ordering of distribution tails is in the development and selection of statistical distributions with tails appropriate for observations on a variable. This is particularly true, for example, for measures of market risk such as Value-at-Risk (VaR), which is heavily used by financial institutions. In the case of VaR estimates, the choice of the appropriate tails emphasize the need for a better understanding of market risk. The choice of an appropriate tail is particularly important when the tail integral above a specified quantile is needed, as is the case with the use of the expected shortfall (ES) in insurance. Underestimation of the VaR or ES can have important consequences, as was evidenced by the great recession of 2008.

Fat-tailed distributions are often defined in terms of higher than normal kurtosis (mesokurtosis). Important ways of distinguishing different distribution tails in practice are:

- the log survival function plot or log complementary cumulative distribution function (CCDF) plot,
- the log log survival function plot,
- fitting appropriate truncated distributions to the tail of the data.

The three methods are explained below.

12.4.1 Exploratory Method 1: log survival function plot

The log survival function plot, (or log CCDF plot), is given by plotting the logarithm of the survival, $\log[\bar{F}_Y(y)]$, against $\log y$. Note that if the upper tail of the survivor function $\bar{F}_Y(y)$ is asymptotically in the form:

- $-k_2(\log y)^{k_1}$,
- $-k_4y^{k_3}$,
- $-k_6e^{k_5y}$,

then the plot of $\log[\bar{F}_Y(y)]$ against $t = \log y$ will be asymptotically in the form $-k_2t^{k_1}$, $-k_4e^{k_3t}$ and $-k_6e^{k_5e^t}$ respectively corresponding to power, exponential and double-exponential relationships respectively. Table 12.5 summarise the relationships.

Table 12.5: Showing possible relationships of the $\log[\bar{F}_Y(y)]$ against $t = \log y$ for Method 1

<table>
<thead>
<tr>
<th>Type</th>
<th>CCDF</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>$-k_2t^{k_1}$</td>
<td>power</td>
</tr>
<tr>
<td>Type II</td>
<td>$-k_4e^{k_3t}$</td>
<td>exponential</td>
</tr>
<tr>
<td>Type III</td>
<td>$-k_6e^{k_5e^t}$</td>
<td>double-exponential</td>
</tr>
</tbody>
</table>
A sample version of the log survival plot, the empirical log survival plot, is given by plotting \( \log (1 - (i - 0.5)/n) \) against \( \log y(i) \), where \( y(i) \) is the \( i^{th} \) largest value of \( y \) in the sample (e.g., \( i = 1 \) for the largest observation and \( i = 2 \) for the second largest observation). The empirical sample log survival plot can be used to investigate the tail form of \( \bar{F}_Y(y) \). Typically we would include only certain percentage from the tail of the sample, for example 10%, of the right (or left) tail of the empirical survival function.

The exploratory Method 1 is widely used in the literature (e.g., Mandlebrot, 1997) to estimate the parameters \( \alpha \) of a Pareto type I distribution. A linear line is fitted to the empirical log survival function plot and the estimate of the slope of the fitted line is taken as an estimate of the parameter \( \alpha \) of the Pareto distribution. Note that this corresponds to the Type I where \( k_2 = \alpha \) and \( k_1 = 1 \) in Table 12.5.

The `gamlss` function `logSurv()` is design for exploring method 1. It plots the empirical log survival function against \( \log(y) \) for specified percentage of the tail and fits a linear, quadratic and exponential curve to the points of the plot. For distributions defined on the positive real line a good linear fit would indicate a Pareto type tail \( (k_1 = 1) \) (this is because the log-log plot of a distribution with power tails results in a linear relationship), a good quadratic fit a log-normal type tail \( (k_1 = 2) \) (similarly, this is because the log-log plot of a distribution with exponential tails results in a quadratic relationship) and good exponential fit a Weibull type tail \( (0 \leq k_3 < \infty) \) according to table 12.2 (not to be confused with the shapes of figure 12.2). Note that method one is only appropriate to investigate rather heavy tails and it is not very good to discriminate between different type of tails, as the method 2.

**Example**

To demonstrate the method, the total USA box office film revenue, which was recorded for 4031 films from 1988-1999, is used. Film revenues are highly skewed, in such a way that a small number of large revenue films coexist alongside considerably greater numbers of smaller revenue films. Moreover, the skewed nature of these distributions appears to be an empirical regularity, with Pokorny and Sedgwick [2010] dating this phenomenon back to at least the 1930s, making it an early example of a mass market long tail.

Figure 12.4 shows the sample CCDF plot (exploratory method 1) for the largest 10% of revenues (or the 403 most profitable films, which are denoted by F90) together with fitted linear, quadratic and exponential functions. The linear fit appears inadequate, hence \( k_1 = 1 \) (e.g. a Pareto distribution) appears inappropriate. The quadratic or exponential fits adequately suggesting \( k_1 = 2 \) or \( 0 < k_3 < \infty \) may be appropriate (see table 12.2 for the relevant distributions). Note that Voudouris et al. (2012) proposed the Box-Cox power exponential distribution, \( 0 < k_3 < \infty \), to fit the film dataset while they rejected the Pareto distribution as an appropriate distribution for the above sample.

### 12.4.2 Exploratory Method 2: log-log-Survival

Correspondingly, the upper tail of \( \log \{- \log \bar{F}_Y(y) \} \) is asymptotically as \( y \to \infty \) in the form:

- \( \log k_2 + k_1 \log [\log(y)] \),
- \( \log k_4 + k_3 \log y \),
- \( \log k_6 + k_5 y \).

Hence a plot of \( \log \{- \log \bar{F}_Y(y) \} \) against \( \log [\log(y)] \) or \( \log y \) or \( y \) will be asymptotically linear in each case (see figure 12.5 for an example).
12.4. METHODS FOR CHOOSING THE APPROPRIATE TAIL

Table 12.6 summarise the relationship between the different types of tails and model term needed to be fitted. The corresponding sample plot (e.g., figure 12.5) can be used to investigate the tail form of $F_Y(y)$. Note from Table 12.6 that method 2 provides estimates for all the parameters involved. For example for type I tail the fitted constant term $\hat{\beta}_0$ provides an estimate to $\hat{k}_2 = \exp(\hat{\beta}_0)$ while the fitted slope $\hat{\beta}_1$ provides an estimate for $k_1$. Therefore in this respect the method 2 is more general than method 1 where we be able to estimate only $k_2$ and $k_4$. Note although that for accurate estimates a large sample size (from the tail) may be required especially in the Type I case.

Another advantage of method 2 is that the response variable, (the empirical log $\{-\log\left[\bar{F}_Y(y)\right]\}$), in all types cases in Table 12.6 is the same allowing straightforward comparison of the three fitted regressions. This is implemented in the R function `loglogSurv()` which fits (at certain percentage of the tail) the empirical log $\{-\log\left[\bar{F}_Y(y)\right]\}$ against log(log($y$)), log $y$ and $y$ respectively and choses the best fit according to the the residuals sum of squares and report it. The functions `loglogSurv1()`, `loglogSurv2()` and `loglogSurv3()` only fit the equivalent model for type I, II, and III tail. For example, to fit a model for Type I tail, we set as $y = \log\left\{-\log\left[\bar{F}_Y(y)\right]\right\}$ and $x = \log(\log(y))$, while we assume the normal distribution for the error term.

Figure 12.5 plots log $\{-\log\left[\bar{F}_Y(y)\right]\}$ against log(log($y$)), log $y$ and $y$ respectively, (exploratory method 2), with the middle graph providing the best linear fit (error sum of squares equal 0.0909, see Table 12.7), with estimates $\hat{k}_3 = 0.561$ and $\hat{k}_4 = \exp(-8.917) = 0.000134$, suggesting a Box-Cox power exponential tail may be appropriate.
Table 12.6: Showing possible relationships of the log[$\hat{F}_Y(y)$] against $t = \log y$ for Method 1

<table>
<thead>
<tr>
<th>Type</th>
<th>$\log {-\log [\hat{F}_Y(y)]}$</th>
<th>Linear Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>$\log k_2 + k_1 \log [\log(y)]$</td>
<td>$\log [\log(y)]$</td>
</tr>
<tr>
<td>Type II</td>
<td>$\log k_3 + k_2 \log y$</td>
<td>$\log y$</td>
</tr>
<tr>
<td>Type III</td>
<td>$\log k_6 + k_5 y$</td>
<td>$y$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>slope</th>
<th>Error SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>type I</td>
<td>-28.2422</td>
<td>10.1806</td>
<td>0.12597</td>
</tr>
<tr>
<td>type II</td>
<td>-8.91724</td>
<td>0.560959</td>
<td>0.09090</td>
</tr>
<tr>
<td>type III</td>
<td>0.75172</td>
<td>5.697e-09</td>
<td>2.82237</td>
</tr>
</tbody>
</table>

Table 12.7: Estimated coefficients from exploratory method 2.

Figure 12.5: Exploratory Method 2 applied to the 90’s film revenues data
12.4.3 Exploratory Method 3: truncated distribution fitting

Truncated lognormal and Weibull distributions (see chapter 11 for a discussion on truncated distributions) were fitted to the largest 10% of revenues leading to reasonable fits in each case. Figure 12.6 provides a normal QQ plot for the normalised quantile residuals of Dunn and Smyth [1996] from the truncated Weibull fit to the largest 10% of revenues, indicating a reasonable fit in the upper tail. The estimated Weibull parameters were \( \hat{\mu} = 13467053 \) and \( \hat{\sigma} = 0.6476 \). Sequential fits of the truncated Weibull distribution to the largest \( r \) revenues, for \( r = 4, 5, 6, \ldots, 403 \), were followed by a plot of the parameter estimate \( \hat{\sigma} \) against \( r \) indicating that the fitted parameter \( \hat{\sigma} \) is relatively stable (Figure 12.7) indicating that the Weibull fit to the tail is relatively stable as \( r \) changes. This plot is analogous to the Hill plot (Hill [1975]).

![Normal Q–Q plot of Weibull 10\% fit](image)

Figure 12.6: QQ plot for the truncated Weibull.

Appendix 12.5

12.5.1 Lemma B1

Let the random variables \( Y_1 \) and \( Y_2 \) have probability density functions \( f_{Y_1}(y) \) and \( f_{Y_2}(y) \) respectively, then \( f_{Y_1}(y) = o(f_{Y_2}(y)) \) as \( y \to \infty \) \( \iff \) \( \lim_{y \to \infty} [\log f_{Y_2}(y) - \log f_{Y_1}(y)] = +\infty \). Similarly replacing \( y \to \infty \) by \( y \to -\infty \) for the left tail.

**Proof B1**

\[ f_{Y_1}(y) = o(f_{Y_2}(y)) \quad \text{as } y \to \infty \]
210

CHAPTER 12. HEAVINESS OF TAILS OF CONTINUOUS DISTRIBUTIONS

![Sequential plot of \( \hat{\sigma} \) for the truncated Weibull distribution](image)

Figure 12.7: Sequential plot of \( \hat{\sigma} \) for the truncated Weibull

\[
\lim_{y \to \infty} \left[ \frac{f_{Y_1}(y)}{f_{Y_2}(y)} \right] = 0 \quad \iff \quad \lim_{y \to \infty} \left[ \log \frac{f_{Y_2}(y)}{f_{Y_1}(y)} \right] = +\infty
\]

12.5.2 Lemma B2

Let random variables \( Y_1 \) and \( Y_2 \) have probability density functions \( f_{Y_1}(y) \) and \( f_{Y_2}(y) \), cumulative distribution functions \( F_{Y_1}(y) \) and \( F_{Y_2}(y) \) and survivor functions \( \bar{F}_{Y_1}(y) \) and \( \bar{F}_{Y_2}(y) \) respectively, then

\[
f_{Y_1}(y) = o \left[ f_{Y_2}(y) \right] \quad \text{as} \quad y \to \infty \quad \iff \quad F_{Y_1}(y) = o \left[ \bar{F}_{Y_2}(y) \right] \quad \text{as} \quad y \to \infty
\]

\[
f_{Y_1}(y) = o \left[ f_{Y_2}(y) \right] \quad \text{as} \quad y \to -\infty \quad \iff \quad F_{Y_1}(y) = o \left[ \bar{F}_{Y_2}(y) \right] \quad \text{as} \quad y \to -\infty
\]

provided \( F_{Y_1}(y) \) and \( F_{Y_2}(y) \) are differentiable and, as \( y \to \infty \) and as \( y \to -\infty \), \( \lim f_{Y_1}(y) = \lim f_{Y_2}(y) = 0 \) and \( \lim \left[ \frac{f_{Y_1}(y)}{f_{Y_2}(y)} \right] \) exists.

**Proof B2**

\[
f_{Y_1}(y) = o \left[ f_{Y_2}(y) \right] \quad \text{as} \quad y \to \infty
\]
<table>
<thead>
<tr>
<th>Distribution</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Box-Cox</td>
<td>Cole and Green [1992]</td>
</tr>
<tr>
<td>Box-Cox power exponential</td>
<td>Rigby and Stasinopoulos [2004]</td>
</tr>
<tr>
<td>Box-Cox t</td>
<td>Rigby and Stasinopoulos [2006]</td>
</tr>
<tr>
<td>Cauchy</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Exponential</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Exponential generalized beta type 2</td>
<td>McDonald and Xu [1995], McDonald [1996]</td>
</tr>
<tr>
<td>Gamma</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Generalized beta type 2</td>
<td>McDonald and Xu [1995], McDonald [1996]</td>
</tr>
<tr>
<td>Generalized gamma</td>
<td>Lopatatzidis and Green [2000]; Harter [1967]</td>
</tr>
<tr>
<td>Generalized inverse</td>
<td>Jørgensen [1997]; Jørgensen [1982]</td>
</tr>
<tr>
<td>Gaussian</td>
<td></td>
</tr>
<tr>
<td>Generalized t</td>
<td>McDonald and Newey [1988], McDonald [1991]</td>
</tr>
<tr>
<td>Gumbel</td>
<td>Crowder et al. 1991</td>
</tr>
<tr>
<td>Inverse Gamma</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Johnson's SU</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Johnson's SU Original</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Laplace</td>
<td>Johnson et al. (1995)</td>
</tr>
<tr>
<td>Lognormal</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Normal</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Pareto Type 2</td>
<td>Johnson et al. (1994)</td>
</tr>
<tr>
<td>Power exponential</td>
<td>Nelson [1991]</td>
</tr>
<tr>
<td>Reverse Gumbel</td>
<td>Johnson et al. (1995)</td>
</tr>
<tr>
<td>Sinh-arcsinh</td>
<td>Jones [2005]</td>
</tr>
<tr>
<td>Sinh-arcsinh original</td>
<td>Jones and Pewsey [2009]</td>
</tr>
<tr>
<td>Skew exponential power type 3</td>
<td>Fernandez et al. (1995)</td>
</tr>
<tr>
<td>Skew exponential power type 4</td>
<td>Jones [2005]</td>
</tr>
<tr>
<td>Skew t type 3</td>
<td>Fernandez and Steel [1998]</td>
</tr>
<tr>
<td>Skew t type 4</td>
<td>Stasinopoulos et al. (2008)</td>
</tr>
<tr>
<td>Stable</td>
<td>Nolan [2012]</td>
</tr>
<tr>
<td>t</td>
<td>Johnson et al. (1995)</td>
</tr>
<tr>
<td>Weibull</td>
<td>Johnson et al. (1994)</td>
</tr>
</tbody>
</table>

Table 12.8: References for continuous distributions

\[
\lim_{y \to \infty} \frac{f_{Y_1}(y)}{f_{Y_2}(y)} = 0 \\
\lim_{y \to \infty} \frac{\bar{F}_{Y_1}(y)}{\bar{F}_{Y_2}(y)} = 0 \quad \text{using L'Hopital's rule} \\
\bar{F}_{Y_1}(y) = o\left[\bar{F}_{Y_2}(y)\right]
\]

The proof follows similarly for the left tail as \(y \to -\infty\).
212  CHAPTER 12.  HEAVINESS OF TAILS OF CONTINUOUS DISTRIBUTIONS

12.5.3 Lemma B3

Let \( f_Y(y) \) and \( F_Y(y) \) be respectively the probability density function and cumulative distribution function of a random variable \( Y \). Then

\[
\frac{1}{F_Y(y)} \sim \left[ \frac{1}{f_Y(y)} \right]' \quad \text{as } y \to \infty
\]

provided \( \lim_{y \to \infty} f_Y(y) = 0 \) and \( \lim_{y \to \infty} \frac{f'_Y(y)}{f_Y(y)} \) exists, where \( ' \) indicates the derivative with respect to \( y \) and \( \bar{F}_Y(y) = 1 - F_Y(y) \).

Proof B3

\[
\lim_{y \to \infty} \frac{f_Y(y)}{F_Y(y)} = \lim_{y \to \infty} \frac{f'_Y(y)}{f_Y(y)} \quad \text{using L'Hopital's rule}
\]

\[
\therefore \quad \lim_{y \to \infty} \left\{ \frac{[f_Y(y)]^2}{F_Y(y)f'_Y(y)} \right\} = 1
\]

\[
\therefore \quad \frac{1}{F_Y(y)} \sim \frac{f'_Y(y)}{[f_Y(y)]^2} = \left[ \frac{1}{f_Y(y)} \right]'
\]

12.5.4 Corollary C1

If \( \log f_Y(y) \sim -g(y) \) as \( y \to \infty \) then \( \log \bar{F}_Y(y) \sim \left\{ \begin{array}{ll} -g(y) - \log g'(y) & \text{if } g(y) \sim -\log g'(y) \\ o[g(y)] & \text{if } g(y) \sim -\log g'(y) \end{array} \right. \)

Proof C1

\[
\log f_Y(y) = -g(y)[1 + o[1]]
\]

\[
\therefore \quad \frac{1}{f_Y(y)} = e^{g(y)[1+o[1]]}
\]

\[
\therefore \quad \left[ \frac{1}{f_Y(y)} \right]' \sim g'(y)e^{g(y)}
\]

since \( \frac{d}{dy} \{ g(y)[1 + o[1]] \} \sim g'(y) \) as \( \frac{d}{dy} [1 + o[1]] = o(1) \)

\[
\therefore \quad \bar{F}_Y(y)g'(y)e^{g(y)} \to 1 \quad \text{as } y \to \infty \text{ using Lemma B3}
\]

\[
\therefore \quad \log \bar{F}_Y(y) + g(y) + \log g'(y) \to 0 \quad \text{as } y \to \infty
\]

Hence result.

12.5.5 Corollary C2

As \( y \to \infty \) (or \( y \to -\infty \)),

(a) If \( \log f_Y(y) \sim -k_2(\log |y|)^{k_1} \) then \( \log F_Y(y) \sim \left\{ \begin{array}{ll} -k_2(\log |y|)^{k_1} & \text{if } k_1 > 1 \\ -(k_2 - 1) \log |y| & \text{if } k_1 = 1 \text{ and } k_2 > 1 \\ o(\log |y|) & \text{if } k_1 = k_2 = 1 \end{array} \right. \)
(b) If \( \log f_Y(y) \sim -k_4 |y|^{k_3} \) then \( \log F_Y(y) \sim -k^4 |y|^{k_3} \)
(c) If \( \log f_Y(y) \sim -k_6 e^{-k_5 |y|} \) then \( \log F_Y(y) \sim -k_6 e^{-k_5 |y|} \)

**Proof C2**

(a) From Corrolary C1, \( \log \bar{F}_Y(y) \sim -k_2 (\log |y|)^{k_1} \log \left[ \frac{k_1 k_2}{|y|} (\log |y|)^{k_1 - 1} \right] \) if \( k_1 \geq 1 \) and \( k_2 > 1 \) and \( \log \bar{F}_Y(y) \sim o(\log |y|) \) if \( k_1 = k_2 = 1 \).

(b) (c) From corrolary C1.
Chapter 13

Centile based comparisons of continuous distributions

This chapter compares the skewness and kurtosis properties of distributions, in particular:

1. compares the authorised domain of continuous distributions
2. provide useful guide in developing statistical models from a list of flexible theoretical distributions

This chapter is more theoretical and can be omitted for a practical course. Having said that the chapter provides important information in terms of selecting appropriately flexible distributions in terms of skewness and kurtosis.

Appendix 13.1 Introduction

Observable variables are often characterised by high skewness and kurtosis (not just kurtosis or skewness). Therefore, we need methods to compare distributions in terms of their flexibility in capturing simultaneously various degrees of skewness and kurtosis. It is important to note that if the data is heavy tailed but not skewed, then the classifications of chapter 12 should be used. The important point here is that by comparing different distributions (rather than focusing on an individual distribution), the proposed classification of theoretical distributions is a useful guide in developing models from a list of theoretical distributions when flexible statistical tools are used to analyse processes characterized by highly skew and kurtic data.

Although moment-based measures of skewness and kurtosis have traditionally been used to compare distributions, moment-based measures may not exist or may be unreliable (suffering from being affected by an extreme tail of the distribution, which may have negligible probability). Thus, we use centile-based measures of skewness and kurtosis to compare distributions. Distributions not included in the current comparison can be easily classified and their properties checked with the methods given here.

The centile-based measures of skewness and kurtosis are defined using the quantile function of the distribution of a random variable $Y$ given by $y_p = F_Y^{-1}(p)$ for $0 < p < 1$, where $F_Y^{-1}$ is the inverse cumulative distribution function of $Y$. 

215
A general centile based measure of skewness is given by MacGillivray [1986]:

\[ s_p = \frac{(y_p + y_{1 - p})/2 - y_{0.5}}{(y_{1 - p} - y_p)/2} \]  

(13.1)

for \(0 < p < 0.5\), i.e. the midpoint of a central \(100(1 - 2p)\)% interval for \(Y\) minus the median, divided by the half length of the central \(100(1 - 2p)\)% interval. Note that \(-1 \leq s_p \leq 1\).

One important case is \(p = 0.25\), giving Galton’s measure of skewness:

\[ s_{0.25} = \frac{(Q_1 + Q_3)/2 - m}{(Q_3 - Q_1)/2} \]  

(13.2)

i.e. the mid quartile \((Q_1 + Q_3)/2\) minus the median divided by the semi-quartile range \((Q_3 - Q_1)/2\), where \(Q_1 = y_{0.25}\) and \(Q_3 = y_{0.75}\). This can be considered as a measure of central skewness since it focuses on the skewness within the interquartile range for \(Y\).

A second important case is \(p = 0.01\), giving

\[ s_{0.01} = \frac{(y_{0.01} + y_{0.99})/2 - y_{0.5}}{(y_{0.99} - y_{0.01})/2} \]  

(13.3)

i.e. the midpoint of a central 98% interval for \(Y\) minus the median, divided by the half length of the central 98% interval for \(Y\). This can be considered as a measure of tail skewness since it focuses on skewness within a central 98% interval for \(Y\). A third important case is \(p = 0.001\) which measures extreme tail skewness.

Following Balanda and MacGillivray [1988], a general centile based measure of kurtosis is given by Andrews et al. [1972]:

\[ k_p = \frac{(y_{1 - p} - y_p)}{(Q_3 - Q_1)} \]  

(13.4)

for \(0 < p < 0.5\), i.e. the ratio of the length of a central \(100(1 - 2p)\)% interval for \(Y\) to its interquartile range. An important case is \(p = 0.01\), i.e. \(k_{0.01}\) [Andrews et al., 1972]. This has been scaled relative to a normal distribution for which \(k_{0.01} = 3.49\) giving

\[ s_k_{0.01} = \frac{(y_{0.99} - y_{0.01})}{3.49 (Q_3 - Q_1)}. \]  

(13.5)

Rosenberger and Gasko [1983]. Hence a normal distribution has \(s_k_{0.01} = 1\). To allow the full range of kurtosis to be plotted it is transformed to

\[ t s_k_{0.01} = \frac{s_k_{0.01} - 1}{s_k_{0.01}}. \]  

(13.6)

Note that \(t s_k_{0.01} \in (-2.49, 1)\), where \(t s_k_{0.01} \to 1\) corresponds to \(k_{0.01} \to \infty\) and \(t s_k_{0.01} \to -2.49\) corresponds to \(k_{0.01} \to 1\). Also \(t s_k_{0.01} = 0\) corresponds to \(s_k_{0.01} = 1\), e.g. a normal distribution, while \(t s_k_{0.01} = -1\) corresponds to \(s_k_{0.01} = 0.5\). See Balanda and MacGillivray [1988] for a review of kurtosis.

In sections 13.2 and 13.3 we compare plots of transformed centile kurtosis (13.6) against each of centile central skewness respectively (13.2) and centile tail skewness (13.3) for commonly used heavy tailed distributions on the real line. The following distributions on the real line are considered: exponential generalized beta type 2 (EGB2), Johnson’s SU (JSU), sinh-arcsinh original (SHASHo), skew exponential power type 3 (SEP3), skew t type 3 (ST3) and stable (SB). See Table A1 in Appendix A for references.
13.2 Transformed (centile) kurtosis against (centile) central skewness

Here we investigate the relationship between transformed kurtosis \( tsk_{0.01} \) given by (13.6) against positive central skewness \( s_{0.25} \in (0, 1) \) given by (13.3). For each of the six distributions the boundary of central skewness is plotted against transformed kurtosis in Figure 13.1. The vertical line at central skewness equals zero and the horizontal line at transformed kurtosis equals one form the outer boundaries of each of the six regions of the distributions. The corresponding plot for negative central skewness is a mirror image around the vertical origin axis.

![Graph showing the relationship between transformed centile kurtosis and central centile skewness for different distributions.](image)

Figure 13.1: The upper boundary of centile central skewness against the transformed centile kurtosis for six distributions on the real line.

Note that the normal distribution is plotted at the point \((0, 0)\) in Figure 13.1. Transformed kurtosis below 0 can be considered as ‘platykurtic’ while above 0 can be considered ‘leptokurtic’.
Clearly the EGB2, JSU and SB distributions do not allow ‘platykurtic’ distributions, while SEP3 allows the lowest kurtosis (most ‘platykurtic’) distributions for a fixed low central skewness \( s_{0.25} < 0.05 \) and SHASHo allows the lowest kurtosis distributions for a fixed high central skewness \( s_{0.25} > 0.05 \).

The SHASHo distribution is generally the most versatile covering the largest range of central skewness \( s_{0.25} \) for a given value of transformed kurtosis \( t s k_{0.01} \) (provided \( t s k_{0.01} > -0.36 \)). The SEP3 distribution is most versatile for \( t s k_{0.01} < -0.36 \) and second most versatile for \( t s k_{0.01} > -0.36 \). The JSU and ST3 distributions have more restricted central skewness for a given transformed kurtosis. The EGB2 distribution is more restricted in central skewness and transformed kurtosis, with the transformed kurtosis at or moderately above that of the normal distribution. The stable distribution is restrictive in central skewness for a given transformed kurtosis, with the transformed kurtosis generally much higher than the normal distribution. The range of possible central skewness increases with the transformed kurtosis for all distributions (except EGB2).

Figure 13.2 (a) and (b) show the transformed kurtosis against central skewness for the SHASHo and SB distributions respectively, showing contours for different values of each of the skewness and kurtosis parameters, \( \nu \) and \( \tau \) respectively, of the distribution, while keeping the other parameter constant. The SHASHo was chosen because of its flexibility, while SB was chosen because its moment based kurtosis-skewness plot is not possible. For the SHASHo distributions in Figure 13.2(a) the horizontal contours correspond to \( \tau = 0.001, 0.5, 0.75, 1, 1.5, 3 \) from top to bottom while the ‘vertical’ contours correspond to \( \nu = 0, 0.1, 0.25, 0.5, 0.75, 1, 1.5, 100 \) from left to right. Note that \( \tau = 0.001 \) and \( \nu = 100 \) effectively correspond to the limits \( \tau = 0 \) and \( \nu = \infty \) as no change in the contours was observed as \( \tau \) was decreased below 0.001 and \( \nu \) increased above 100, respectively. Note also that for a fixed \( \tau, \nu \) affects the centile skewness only. For the stable SB distribution in Figure 13.2(b) the ‘horizontal’ contours correspond to \( \tau = 0.001, 0.75, 1, 1.25, 1.5, 1.75 \) while the ‘vertical’ contours correspond to \( \nu = 0, 0.1, 0.25, 0.5, 0.75, 1 \) from left to right. Note that \( \tau = 0.001 \) effectively corresponds to the limit \( \tau = 0 \).

13.3 Transformed (centile) kurtosis against (centile) tail skewness

Section 3.2 is amended to replace the central skewness \( s_{0.25} \) given by (13.3) with the tail skewness \( s_{0.01} \) given by (13.4). Figures 13.3 and 13.4 correspond to Figures 13.1 and 13.2. The contour values of \( \nu \) and \( \tau \) in Figure 13.4 are the same as used in Figure 13.2. Note that the range of tail skewness for the six distributions is now more restricted to \((0, 0.5)\) instead of \((0, 1)\) for the central skewness. However the general comments about the kurtosis-skewness relationship for the six distributions still apply.
13.3. Transformed (Centile) Kurtosis Against (Centile) Tail Skewness

Figure 13.2: Contours of centile central skewness against the transformed centile kurtosis for constant values of $\nu$ and $\tau$ for the SHASHo and SB distributions.
Figure 13.3: The upper boundary of centile tail skewness against the transformed centile kurtosis for six distributions on the real line.
13.4 Conclusions

The boundary of (centile) central and tail skewness against the transformed (centile) kurtosis is also given for six important four parameter distributions on the real line. Overall the sinh-arcsinh (SHASHo) is the most flexible distribution in modelling the skewness and kurtosis. However its tails are not as heavy as the stable (SB) or skew $t$ type 3 (ST3). Hence the SHASHo and SEP3 are flexible enough to model data which can exhibit a wide range of skewness and kurtosis, while the SB and ST3 are more appropriate to model data with high kurtosis and low skewness. The EGB2 is only appropriate for mild leptokurtosis and low skewness.
Figure 13.4: Contours of centile tail skewness against the transformed centile kurtosis for constant values of $\nu$ and $\tau$ for the SHASHo and SB distributions.
Part II

A reference guide
Chapter 14

Continuous distributions on $\mathbb{R}$

14.1 Continuous two parameter distributions on $\mathbb{R}$

14.1.1 Normal (or Gaussian) distribution (NO, NO2, NOF)

First parameterization (NO)

The normal distribution is the default of the argument family of the function `gamlss()`. The parameterization used for the normal (or Gaussian) probability density function (pdf), denoted by $\text{NO}(\mu, \sigma)$, is

$$f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \quad (14.1)$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$ and $\sigma > 0$. The mean of $Y$ is given by $E(Y) = \mu$ and the variance of $Y$ by $\text{Var}(Y) = \sigma^2$, so $\mu$ is the mean and $\sigma$ is the standard deviation of $Y$.

Second parameterization (NO2)

$\text{NO2}(\mu, \sigma)$ is a parameterization of the normal distribution where $\mu$ represents the mean and $\sigma$ represents the variance of $Y$, i.e. $f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$.

Normal family (of variance-mean relationships) (NOF)

The function $\text{NOF}(\mu, \sigma, \nu)$ defines a normal distribution family with three parameters. The third parameter $\nu$ allows the variance of the distribution to be proportional to a power of the mean. The mean of $\text{NOF}(\mu, \sigma, \nu)$ is equal to $\mu$ while the variance is equal to $\text{Var}(Y) = \sigma^2|\mu|^{\nu}$, so the standard deviation is $\sigma|\mu|^{\nu/2}$. The parametrization of the normal distribution given in the function $\text{NOF}(\mu, \sigma, \nu)$ is

$$f(y|\mu, \sigma, \nu) = \frac{1}{\sqrt{2\pi\sigma|\mu|^{\nu/2}}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2|\mu|^{\nu}}\right] \quad (14.2)$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \nu < \infty$.

The function $\text{NOF}(\mu, \sigma, \nu)$ is appropriate for normally distributed regression type models where the variance of the response variable is proportional to a power of the mean. Models of
this type are related to the "pseudo likelihood" models of Carroll and Rubert (1987) but here a proper likelihood is maximized. The \(\nu\) parameter here is not designed to be modelled against explanatory variables but is a constant used as a device allowing us to model the variance mean relationship. Note that, due to the high correlation between the \(\sigma\) and \(\nu\) parameters, the \texttt{mixed()} method argument is essential in the \texttt{gamlss()} fitting function. Alternatively \(\nu\) can be estimated from its profile function, obtained using \texttt{gamlss} package function \texttt{prof.dev()}. 

14.1.2 Logistic distribution (LO)

The logistic distribution is appropriate for moderately kurtotic data. The parameterization of the logistic distribution, denoted here as \(\text{LO}(\mu, \sigma)\), is given by

\[
f_Y(y|\mu,\sigma) = \frac{1}{\sigma} \left\{ \exp \left[ - \left( \frac{y - \mu}{\sigma} \right) \right] \right\} \left\{ 1 + \exp \left[ - \left( \frac{y - \mu}{\sigma} \right) \right] \right\}^{-2}
\]

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty\) and \(\sigma > 0\), with \(E(Y) = \mu\) and \(\text{Var}(Y) = \pi^2\sigma^2/3\), Johnson et al. (1995) p 116.

14.1.3 Gumbel distribution (GU)

The Gumbel distribution is appropriate for moderately negative skew data. The pdf of the Gumbel distribution (or extreme value or Gompertz), denoted by \(\text{GU}(\mu, \sigma)\), is defined by

\[
f_Y(y|\mu,\sigma) = \frac{1}{\sigma} \exp \left[ \frac{y - \mu}{\sigma} \right] - \exp \left( \frac{y - \mu}{\sigma} \right)
\]

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty\) and \(\sigma > 0\), with \(E(Y) = \mu - \gamma\sigma \simeq \mu - 0.57722\sigma\) and \(\text{Var}(Y) = \pi^2\sigma^2/6 \simeq 1.64493\sigma^2\). See Crowder et al. (1991) p 17.

14.1.4 Reverse Gumbel distribution (RG)

The reverse Gumbel distribution, which is also called the type I extreme value distribution is a special case of the generalized extreme value distribution, [see Johnson et al. (1995) p 2 and p 75]. The reverse Gumbel distribution is appropriate for moderately positive skew data. The pdf of the reverse Gumbel distribution, denoted by \(\text{RG}(\mu, \sigma)\) is defined by

\[
f_Y(y|\mu,\sigma) = \frac{1}{\sigma} \exp \left\{ - \left( \frac{y - \mu}{\sigma} \right) - \exp \left( \frac{y - \mu}{\sigma} \right) \right\}
\]

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty\) and \(\sigma > 0\), with \(E(Y) = \mu + \gamma\sigma \simeq \mu + 0.57722\sigma\) and \(\text{Var}(Y) = \pi^2\sigma^2/6 \simeq 1.64493\sigma^2\). [Note that if \(Y \sim \text{RG}(\mu, \sigma)\) and \(W = -Y\), then \(W \sim \text{GU}(\mu, \sigma)\).]

14.2 Continuous three parameter distributions on \(\mathbb{R}\)

14.2.1 Exponential Gaussian distribution (exGAUS)

The pdf of the ex-Gaussian distribution, denoted by \(\text{exGAUS}(\mu, \sigma)\), is defined as

\[
f_Y(y|\mu,\sigma,\nu) = \frac{1}{\nu} \exp \left( \frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2} \right) \Phi \left( \frac{y - \mu - \sigma}{\nu} \right)
\]

(14.6)
for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$, and where $\Phi$ is the cdf of the standard normal distribution. Since $Y = Y_1 + Y_2$ where $Y_1 \sim N(\mu, \sigma^2)$ and $Y_2 \sim EX(\nu)$ are independent, the mean of $Y$ is given by $E(Y) = \mu + \nu$ and the variance is given by $\text{Var}(Y) = \sigma^2 + \nu^2$. This distribution has also been called the lagged normal distribution, Johnson et al., (1994), p172.

### 14.2.2 Power Exponential distribution (PE, PE2)

#### First parameterization (PE)

The power exponential distribution is suitable for leptokurtic as well as platykurtic data. The pdf of the power exponential family distribution, denoted by PE$(\mu, \sigma, \nu)$, is defined by

$$f_Y(y|\mu, \sigma, \nu) = \frac{\nu \exp[-|\frac{y}{c}|]}{2\sigma \Gamma\left(\frac{1}{\nu}\right)}$$

(14.7)

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$ and where $z = (y - \mu)/\sigma$ and $c^2 = \Gamma(1/\nu)[\Gamma(3/\nu)]^{-1}$.

In this parameterization, used by Nelson (1991), $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2$. Note that $\nu = 1$ and $\nu = 2$ correspond to the Laplace (i.e. two sided exponential) and normal distributions respectively, while the uniform distribution is the limiting distribution as $\nu \to \infty$.

The cdf of $Y$ is given by $F_Y(y) = \frac{1}{2}[1 + S(s|\mu, \sigma, \nu)]$ where $S = |z/c|^\nu$ has a gamma distribution with pdf $f_S(s) = s^{1/\nu} \exp(-s)/\Gamma\left(\frac{1}{\nu}\right)$.

#### Second parameterization (PE2)

An alternative parameterization, the power exponential type 2 distribution, denoted by PE2$(\mu, \sigma, \nu)$, is defined by

$$f_Y(y|\mu, \sigma, \nu) = \frac{\nu \exp[-|z|]}{2\sigma \Gamma\left(\frac{1}{\nu}\right)}$$

(14.8)

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$ and where $z = (y - \mu)/\sigma$. Here $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2/c^2$, where $c^2 = \Gamma(1/\nu)[\Gamma(3/\nu)]^{-1}$.

See also Johnson et al., 1995, volume 2, p195, equation (24.83) for a re-parameterized version by Subbotin (1923).

### 14.2.3 t family distribution (TF)

The $t$ family distribution is suitable for modelling leptokurtic data, that is, data with higher kurtosis than the normal distribution. The pdf of the $t$ family distribution, denoted here as TF$(\mu, \sigma, \nu)$, is defined by

$$f_Y(y|\mu, \sigma, \nu) = \frac{1}{\sigma B\left(\frac{1}{2}, \frac{\nu}{2}\right) \nu^{\frac{\nu+1}{2}}} \left[1 + \frac{(y - \mu)^2}{\sigma^2 \nu}\right]^{-\frac{\nu+1}{2}}$$

(14.9)

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$, where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function. The mean and variance of $Y$ are given by $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2\nu/(\nu-2)$ when $\nu > 2$. Note that $T = (Y - \mu)/\sigma$ has a standard $t$ distribution with $\nu$ degrees of freedom, given by Johnson et al. (1995), p 363, equation (28.2).
14.3 Continuous four parameter distributions on $\mathbb{R}$

14.3.1 Exponential Generalized Beta type 2 distribution (EGB2)

The pdf of the exponential generalized beta type 2 distribution, denoted by $EGB_2(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = e^{\nu z} \left\{ |\sigma| B(\nu, \tau) \left[ 1 + e^{z} \right]^{\nu+\tau} \right\}^{-1}$$

(14.10)

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $-\infty < \sigma < \infty$, $\nu > 0$ and $\tau > 0$, and where $z = (y - \mu) / \sigma$, McDonald and Xu (1995), equation (3.3). Here $E(Y) = \mu + \sigma [\Psi(\nu) - \Psi(\tau)]$ and $Var(Y) = \sigma^2 [\Psi^{(1)}(\nu) + \Psi^{(1)}(\tau)]$, from McDonald (1996), p437.

14.3.2 Generalized $t$ distribution (GT)

This pdf of the generalized $t$ distribution, denoted by $GT(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \tau \left\{ 2\sigma \nu^{1/\tau} B(1/\tau, \nu) [1 + |z|^{\nu}/\nu^{\nu+1/\tau}] \right\}^{-1}$$

(14.11)

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$, and where $z = (y - \mu) / \sigma$, McDonald (1991) and McDonald and Newey (1988) Here $E(Y) = \mu$ and $Var(Y) = \sigma^2 \nu^{1/\tau} B(3/\tau, \nu^2) / B(1/\tau, \nu)$, from McDonald (1991) p274.

14.3.3 Johnson SU distribution (JSUo, JSU)

First parameterization (JSUo)

This is the original parameterization of the Johnson $S_u$ distribution, Johnson (1949). The parameter $\nu$ determines the skewness of the distribution with $\nu > 0$ indicating negative skewness and $\nu < 0$ positive skewness. The parameter $\tau$ determines the kurtosis of the distribution. $\tau$ should be positive and most likely in the region above 1. As $\tau \to \infty$ the distribution approaches the normal density function. The distribution is appropriate for leptokurtotic data.

The pdf of the original Johnson’s $S_u$, denoted here as $JSUo(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{\tau}{\sigma} \frac{1}{(r^2 + 1)^{\frac{1}{2}}} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} z^2 \right\}$$

(14.12)

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$ and $\tau > 0$, and where

$$z = \nu + \tau \sinh^{-1}(r) = \nu + \tau \log \left[ r + (r^2 + 1)^{\frac{1}{2}} \right],$$

(14.13)

where $r = (y - \mu) / \sigma$. Note that $Z \sim NO(0,1)$. Here $E(Y) = \mu - \sigma \omega^{1/2} \sinh(\nu/\tau)$ and $Var(Y) = \sigma^2 (\omega^{1/2} + 1) [\omega \cosh(2\nu/\tau) + 1]$, where $\omega = \exp(1/\tau^2)$.

Second parameterization (JSU)

This is a reparameterization of the original Johnson $S_u$ distribution, Johnson (1949), so that parameters $\mu$ and $\sigma$ are the mean and the standard deviation of the distribution. The parameter $\nu$ determines the skewness of the distribution with $\nu > 0$ indicating positive skewness and $\nu < 0$ negative. The parameter $\tau$ determines the kurtosis of the distribution. $\tau$ should be positive and
most likely in the region above 1. As \( \tau \to \infty \) the distribution approaches the normal density function. The distribution is appropriate for leptokurtic data.

The pdf of the Johnson’s \( S_u \), denoted here as \( JSU(\mu, \sigma, \nu, \tau) \), is defined by

\[
 f_Y(y|\mu, \sigma, \nu, \tau) = \frac{\tau}{c\sigma} \cdot \frac{1}{(y^2 + 1)^{\frac{1}{2}}} \cdot \exp \left[ -\frac{1}{2} \left( z^2 - 1 \right) \right]
\]

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty\), \(\sigma > 0\), \(-\infty < \nu < \infty\), \(\tau > 0\), and where

\[
 z = -\nu + \tau \sinh^{-1}(r) = -\nu + \tau \log \left[ r + (r^2 + 1)^{\frac{1}{2}} \right],
\]

\[
 r = \frac{y - (\mu + c\sigma w^{\frac{1}{2}} \sinh \Omega)}{c\sigma},
\]

\[
 c = \left\{ \frac{1}{2} (w - 1) [w \cosh (2\Omega) + 1] \right\}^{-\frac{1}{2}},
\]

\[
 w = \exp(1/\tau^2) \text{ and } \Omega = -\nu/\tau.
\]

Note that \( Z \sim \text{NO}(0,1) \). Here \( E(Y) = \mu \) and \( \text{Var}(Y) = \sigma^2 \).

### 14.3.4 Normal-Exponential-\( t \) distribution (NET)

The NET distribution is a four parameter continuous distribution, although in \texttt{gamlss} it is used as a two parameter distribution with the other two of its parameters fixed. It was introduced by Rigby and Stasinopoulos (1994) as a robust method of fitting the mean and the scale parameters of a symmetric distribution as functions of explanatory variables. The NET distribution is the abbreviation of the Normal-Exponential-Student-\( t \) distribution and is denoted by \( \text{NET}(\mu, \sigma, \nu, \tau) \), for given values for \( \nu \) and \( \tau \). It is normal up to \( \nu \), exponential from \( \nu \) to \( \tau \) and Student-\( t \) with \((\nu\tau - 1)\) degrees of freedom after \( \tau \). Fitted parameters are the first two parameters, \( \mu \) and \( \sigma \). Parameters \( \nu \) and \( \tau \) may be chosen and fixed by the user. Alternatively estimates of the third and forth parameters can be obtained, using the \texttt{gamlss} function \texttt{prof.dev()}. 

The pdf of the normal exponential \( t \) distribution, denoted here as \( \text{NET}(\mu, \sigma, \nu, \tau) \), is given by Rigby and Stasinopoulos (1994) and defined by

\[
 f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \cdot \exp \left\{ \begin{array}{ll} 
 -\frac{z^2}{\tau} & \text{when } |z| \leq \nu \\
 -\nu|z| + \frac{v^2}{\tau} & \text{when } \nu < |z| \leq \tau \\
 -\nu\tau \log \left( \frac{|z|}{\tau} \right) - \nu\tau + \frac{v^2}{\tau} & \text{when } |z| > \tau 
\end{array} \right. \]

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty\), \(\sigma > 0\), \(\nu > 1\), \(\tau > \nu\) \footnote{since NET involves the Student-\( t \) distribution with \((\nu\tau-1)\) degrees of freedom}, and where \( z = (y - \mu)/\sigma \) and \( c = (c_1 + c_2 + c_3)^{-1} \), where \( c_1 = \sqrt{2\pi} \left[ 1 - 2\Phi(-\nu) \right] \), \( c_2 = \frac{2}{\nu} \exp \left\{ -\frac{v^2}{\tau} \right\} \) and \( c_3 = \frac{2}{(\nu\tau - 1)\nu} \exp \left\{ -\nu\tau + \frac{v^2}{\tau} \right\} \), where \( \Phi(\cdot) \) is the cumulative distribution function of the standard normal distribution. Here \( \mu \) is the mean of \( Y \).
14.3.5 Sinh-Arcs sinh (SHASH)

The pdf of the Sinh-Arcsinh distribution, denoted by $\text{SHASH}(\mu, \sigma, \nu, \tau)$, Jones(2005), is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sqrt{2\pi \sigma (1+r^2)^{1/2}}} e^{-z^2/2}$$  \hspace{1cm} (14.17)

where

$$z = \frac{1}{2} \left\{ \tau \sinh^{-1}(r) - \exp \left[ -\nu \sinh^{-1}(r) \right] \right\}$$

and

$$c = \frac{1}{2} \left\{ \tau \exp \left[ \tau \sinh^{-1}(r) \right] + \nu \exp \left[ -\nu \sinh^{-1}(r) \right] \right\}$$

and $r = (y - \mu)/\sigma$ for $-\infty < y < \infty$, where $-\infty < \mu < +\infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$.

Note $\sinh^{-1}(r) = \log(u)$ where $u = r + (r^2 + 1)^{1/2}$. Hence $z = \frac{1}{2} (u^\tau - u^{-\nu})$. Note that $Z \sim \text{NO}(0, 1)$. Hence $\mu$ is the median of $Y$.

14.3.6 Sinh-Arcs sinh (SHASHo)

14.3.7 Sinh-Arcs sinh (SHASHo2)

14.3.8 Skew Exponential Power type 1 distribution (SEP1)

The pdf of the skew exponential power type 1 distribution, denoted by $\text{SEP1}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{\sigma} f_{Z_1}(z) F_{Z_1}(\nu z)$$  \hspace{1cm} (14.18)

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$ and $\tau > 0$, and where $z = (y - \mu)/\sigma$ and $f_{Z_1}$ and $F_{Z_1}$ are the pdf and cdf of $Z_1 \sim \text{PE2}(0, 1^{1/\tau}, \tau)$, a power exponential type 2 distribution with $f_{Z_1}(z) = \alpha^{-1} \exp \left[ -|z|^{\tau} \right]$, where $\alpha = 2\tau^{(1/\tau) - 1} \Gamma(1/\tau)$. This distribution was introduced by Azzalini (1986) as his type I distribution.

Here $E(Y) = \mu + \sigma E(Z)$ and $\text{Var}(Y) = \sigma^2 \{ E(Z^2) - [E(Z)]^2 \}$ where $Z = (Y - \mu)/\sigma$ and $E(Z) = \text{sign}(\nu)\tau^{1/\tau} \left[ \Gamma \left( \frac{3}{2} \right) / \Gamma \left( \frac{1}{2} \right) \right] p\text{BEo} \left( \frac{\nu^2}{1+\nu^2}, \frac{1}{2}, \frac{3}{2} \right)$, and $E(Z^2) = \tau^3/\tau \Gamma \left( \frac{3}{2} \right) / \Gamma \left( \frac{1}{2} \right)$, where $p\text{BEo}(q, a, b)$ is the cdf of an original beta distribution $\text{BEo}(a, b)$ evaluated at $q$, Azzalini (1986), p202-203.

The skew normal type 1 distribution, denoted by $\text{SN1}(\mu, \sigma, \nu)$, a special case of $\text{SEP1}(\mu, \sigma, \nu, \tau)$ given by $\tau = 2$, has mean and variance given by $E(Y) = \mu + \sigma \text{sign}(\nu) \left\{ 2\nu^2 / [\pi (1 + \nu^2)] \right\}^{1/2}$ and $\text{Var}(Y) = \sigma^2 \left\{ 1 - 2\nu^2 / [\pi (1 + \nu^2)] \right\}$, Azzalini (1985), p174. Note that SN1 is not currently implemented as a specific distribution, but can be obtained by fixing $\tau = 2$ in SEP1 using the arguments tau.start=2, tau.fix=TRUE in gamlss().

14.3.9 Skew Exponential Power type 2 distribution (SEP2)

The pdf of the skew exponential power type 2 distribution, denoted by $\text{SEP2}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{\sigma} f_{Z_1}(z) \Phi(\omega)$$  \hspace{1cm} (14.19)
for \(-\infty < y < 0\), where \(-\infty < \mu < \infty\), \(\sigma > 0\), \(-\infty < \nu < 0\), and \(\tau > 0\), and where 
\(z = (y - \mu)/\sigma\) and \(\omega = \text{sign}(z)|z|^{\nu}/\sqrt{\tau}\) and \(f_{Z_1}\) is the pdf of \(Z_1 \sim \text{PE2}(0, \tau^{1/\tau}, \tau)\) and \(\Phi(\omega)\) is the cdf of a standard normal variable evaluated at \(\omega\).

This distribution was introduced by Azzalini (1986) as his type II distribution and was further developed by DiCiccio and Monti (2004). The parameter \(\nu\) determines the skewness of the distribution with \(\nu > 0\) indicating positive skewness and \(\nu < 0\) negative. The parameter \(\tau\) determines the kurtosis of the distribution, with \(\tau > 2\) for platykurtic data and \(\tau < 2\) for leptokurtic.

Here \(E(Y) = \mu + \sigma E(Z)\) and \(\text{Var}(Y) = \sigma^2 V(Z)\) where

\[
E(Z) = \frac{2\tau^{1/\tau}\nu}{\sqrt{\pi\Gamma\left(\frac{1}{\tau}\right)}} \left(1 + \frac{\nu^2}{2}\right) \left(\frac{\nu^2}{1 + \nu^2}\right) n
\]

(14.20)

and \(E(Z^2) = \tau^{1/\tau}\Gamma\left(\frac{3}{\tau}\right)/\Gamma\left(\frac{1}{\tau}\right)\), where \((2n + 1)! = 1.35...(2n - 1)\), DiCiccio and Monti (2004), p439.

For \(\tau = 2\) the \(\text{SEP2}(\mu, \sigma, \nu, \tau)\) distribution is the skew normal type 1 distribution, Azzalini (1985), denoted by \(\text{SN1}(\mu, \sigma, \nu, \tau)\), while for \(\nu = 1\) and \(\tau = 2\) the \(\text{SEP2}(\mu, \sigma, \nu, \tau)\) distribution is the normal density function, \(\text{NO}(\mu, \sigma)\).

### 14.3.10 Skew Exponential Power type 3 distribution (SEP3)

This is a "spliced-scale" distribution with pdf, denoted by \(\text{SEP3}(\mu, \sigma, \nu, \tau)\), defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \left\{ \exp \left[ -\frac{1}{2} \left( \frac{\nu^2}{\nu} \right) I(y < \mu) + \exp \left[ -\frac{1}{2} \left( \frac{\nu^2}{\nu} \right) \right] I(y \geq \mu) \right\}
\]

(14.21)

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty\), \(\sigma > 0\), \(\nu > 0\), and \(\tau > 0\), and where \(z = (y - \mu)/\sigma\) and \(c = \nu\tau/[\Gamma(1 + \nu^2)]\), Fernandez, Osiewalski and Steel (1995). Note that \(I()\) is an indicator function, where \(I(u) = 1\) if \(u\) is true and \(I(u) = 0\) if \(u\) is false.

Note that \(\mu\) is the mode of \(Y\). Here \(E(Y) = \mu + \sigma E(Z)\) and \(\text{Var}(Y) = \sigma^2 V(Z)\) where \(E(Z) = 2\nu^{1/\tau}\Gamma\left(\frac{2}{\nu^2}\right) I(y < \mu) + \exp \left[ -\frac{1}{2} \left( \frac{\nu^2}{\nu} \right) \right] I(y \geq \mu)\) and \(E(Z^2) = 2\nu^{2/\tau}\Gamma\left(\frac{3}{\nu^2}\right) I(y < \mu) + \exp \left[ -\frac{1}{2} \left( \frac{\nu^2}{\nu} \right) \right] I(y \geq \mu)\), Fernandez, Osiewalski and Steel (1995), p1333, eqns. (12) and (13).

The skew normal type 2 distribution, Johnson et al. (1994) p173, denoted by \(\text{SN2}(\mu, \sigma, \nu, \tau)\), (or two-piece normal) is a special case of \(\text{SEP3}(\mu, \sigma, \nu, \tau)\) given by \(\tau = 2\).

### 14.3.11 Skew Exponential Power type 4 distribution (SEP4)

This is a "spliced-shape" distribution with pdf, denoted by \(\text{SEP4}(\mu, \sigma, \nu, \tau)\), defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \left\{ \exp[-|z|^\nu] I(y < \mu) + \exp[-|z|^\nu] I(y \geq \mu) \right\}
\]

(14.22)

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty\), \(\sigma > 0\), \(\nu > 0\), and \(\tau > 0\), and where \(z = (y - \mu)/\sigma\) and \(c = \left[ \Gamma\left(1 + \frac{1}{\nu}\right) \right]^{-1}\), Jones (2005). Note that \(\mu\) is the mode of \(Y\).

Here \(E(Y) = \mu + \sigma E(Z)\) and \(\text{Var}(Y) = \sigma^2 V(Z)\) where \(E(Z) = c \left[ \frac{1}{\nu^2} \Gamma\left(\frac{2}{\nu^2}\right) - \frac{1}{2} \Gamma\left(\frac{3}{\nu^2}\right) \right]\) and \(E(Z^2) = c \left[ \frac{1}{\nu^2} \Gamma\left(\frac{3}{\nu^2}\right) + \frac{1}{2} \Gamma\left(\frac{5}{\nu^2}\right) \right]\).
14.3.12 Skew t type 1 distribution (ST1)
The pdf of the skew t type 1 distribution, denoted by \( \text{ST1}(\mu, \sigma, \nu, \tau) \), is defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{\sigma} f_Z(z) F_{Z_1}(\nu z)
\]
(14.23)
for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty, \sigma > 0, -\infty < \nu < \infty\), and \(\tau > 0\), and where
\(z = (y - \mu)/\sigma\) and \(f_Z\), and \(F_{Z_1}\) are the pdf and cdf of \(Z \sim TF(0,1,\tau)\), a \(t\) distribution with \(\tau > 0\) degrees of freedom, with \(\tau\) treated as a continuous parameter. This distribution is in the form of a type I distribution of Azzalini (1986).

14.3.13 Skew t type 2 distribution (ST2)
The pdf of the skew t type 2 distribution, denoted by \( \text{ST2}(\mu, \sigma, \nu, \tau) \), is defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{\sigma} f_Z(z) F_{Z_2}(w)
\]
(14.24)
for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty, \sigma > 0, -\infty < \nu < \infty\), and \(\tau > 0\), and where
\(z = (y - \mu)/\sigma, w = \nu \lambda^{1/2} z\) and \(\lambda = (\tau + 1)/(\tau + z^2)\) and \(f_Z\), is the pdf of \(Z_1 \sim TF(0,1,\tau)\) and \(F_{Z_2}\) is the cdf of \(Z_2 \sim TF(0,1,\tau+1)\). This distribution is the univariate case of the multivariate skew \(t\) distribution introduced by Azzalini and Capitanio (2003).

Here the mean and variance of \(Y\) are given by \(E(Y) = \mu + \sigma E(Z)\) and \(Var(Y) = \sigma^2 V(Z)\) where \(E(Z) = \nu \tau^{1/2} \Gamma\left(\frac{\tau+1}{2}\right)/\left[\pi^{1/2}(1+\nu^2)^{1/2} \Gamma\left(\frac{\tau}{2}\right)\right]\) for \(\tau > 1\) and \(E(Z^2) = \tau/(\tau - 2)\) for \(\tau > 2\), Azzalini and Capitanio (2003), p382.

14.3.14 Skew t type 3 distribution (ST3)
This is a "spliced-scale" distribution with pdf, denoted by \( \text{ST3}(\mu, \sigma, \nu, \tau) \), defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \left\{ 1 + \frac{z^2}{\tau} \left[ \nu^2 I(y < \mu) + \frac{1}{\nu^2} I(y \geq \mu) \right] \right\}
\]
(14.25)
for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty, \sigma > 0, \nu > 0\), and \(\tau > 0\), and where \(z = (y - \mu)/\sigma\) and \(c = 2\nu/\left[\sigma (1 + \nu^2) B\left(\frac{1}{2}, \frac{\nu}{\nu+1}\right) \tau^{1/2}\right]\), Fernandez and Steel (1998).

Note that \(\mu\) is the mode of \(Y\). The mean and variance of \(Y\) are given by \(E(Y) = \mu + \sigma E(Z)\) and \(Var(Y) = \sigma^2 V(Z)\) where \(E(Z) = 2\nu^{1/2}(\nu^2 - 1)/[(\tau - 1)B\left(\frac{1}{2}, \frac{\nu}{\nu+1}\right) \nu]\) and \(E(Z^2) = \tau (\nu^2 + \frac{1}{\nu^2}) / [(\tau - 2) (\nu + \frac{1}{\nu})]\), Fernandez and Steel (1998), p360, eqn. (5).

14.3.15 Skew t type 4 distribution (ST4)
This is a "spliced-shape" distribution with pdf, denoted by \( \text{ST4}(\mu, \sigma, \nu, \tau) \), defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \left\{ 1 + \frac{z^2}{\nu} \left[ 2 \nu B\left(\frac{1}{2}, \frac{\nu}{\nu+1}\right) + B\left(\frac{1}{2}, \frac{\tau}{\tau+1}\right) \right] I(y < \mu) + \left[ 1 + \frac{z^2}{\tau} \right] \left( \frac{1}{\nu^2} - \frac{1}{\nu+1} \right) I(y \geq \mu) \right\}
\]
(14.26)
for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty, \sigma > 0, \nu > 0\) and \(\tau > 0\), and where \(z = (y - \mu)/\sigma\) and \(c = 2\nu B\left(\frac{1}{2}, \frac{\tau}{\tau+1}\right) + B\left(\frac{1}{2}, \frac{\tau}{\tau+1}\right)\).

Here \(E(Y) = \mu + \sigma E(Z)\) and \(Var(Y) = \sigma^2 V(Z)\) where \(E(Z) = c \left[ \nu B\left(\frac{1}{2}, \frac{\nu}{\nu+1}\right) \right]\), provided \(\nu > 1\) and \(\tau > 1\), and \(E(Z^2) = \frac{1}{\nu} \left[ \nu B\left(\frac{1}{2}, \frac{\nu}{\nu+1}\right) / (\nu - 2) \right] + \left[ \nu B\left(\frac{1}{2}, \frac{\nu}{\nu+1}\right) / (\nu - 2) \right]\), provided \(\nu > 2\) and \(\tau > 2\).
14.3. CONTINUOUS FOUR PARAMETER DISTRIBUTIONS ON \( \mathbb{R} \)

14.3.16 Skew \( t \) type 5 distribution (ST5)

The pdf of the skew \( t \) distribution type 5, denoted by \( \text{ST5}(\mu, \sigma, \nu, \tau) \), Jones and Faddy (2003), is defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{c}{\sigma} \left[ 1 + \frac{z}{(a + b + z^2)^{1/2}} \right]^{a+1/2} \left[ 1 - \frac{z}{(a + b + z^2)^{1/2}} \right]^{b+1/2}
\]

for \(-\infty < y < \infty\), where \(-\infty < \mu < \infty, \sigma > 0, -\infty < \nu < \infty\) and \(\tau > 0\), and where

\[
z = \frac{(y - \mu)}{\sigma}
\]

and

\[
c = \left[ 2^{a+b-1}(a+b)^{1/2}B(a,b) \right]^{-1}
\]

and

\[
v = (a-b)/(ab(a+b))^{1/2}
\]

and \(\tau = 2/(a+b)\).

Here \(E(Y) = \mu + \sigma E(Z)\) where

\[
E(Z) = (a - b)(a + b)^{1/2} \frac{(a - \frac{1}{2}) \Gamma(a - \frac{1}{2})}{[2\Gamma((a)(b))}
\]

and

\[
Var(Y) = \sigma^2 V(Z)\text{ where } E(Z^2) = (a + b) \frac{(a - b)^2 + a + b - 2}{[4(a - 1)(b - 1)]}.
\]

Jones and Faddy (2003), p162.
Chapter 15

Continuous distributions in $\mathbb{R}^+$

15.1 Continuous one parameter distribution in $\mathbb{R}^+$

15.1.1 Exponential distribution (EXP)

This is the only one parameter continuous distribution in $\texttt{gamlss}$ packages. The exponential distribution is appropriate for moderately positive skew data. The parameterization of the exponential distribution, denoted here as $\texttt{EXP}(\mu)$, is defined by

$$f_Y(y|\mu) = \frac{1}{\mu} \exp \left\{ -\frac{y}{\mu} \right\}$$

for $y > 0$, where $\mu > 0$ and where $E(Y) = \mu$ and $\text{Var}(Y) = \mu^2$.

15.2 Continuous two parameter distribution in $\mathbb{R}^+$

15.2.1 Gamma distribution (GA)

The gamma distribution is appropriate for positively skew data. The pdf of the gamma distribution, denoted by $\texttt{GA}(\mu, \sigma)$, is defined by

$$f_Y(y|\mu, \sigma) = \frac{1}{(\mu^2\sigma)^{1/\sigma^2}} \frac{y^{\frac{1}{\sigma^2}-1} e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Here $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2\mu^2$. This a reparameterization of Johnson et al. (1994) p 343 equation (17.23) obtained by setting $\sigma^2 = 1/\alpha$ and $\mu = \alpha\beta$.

15.2.2 Log Normal distribution (LOGNO, LNO)

Log Normal distribution (LOGNO)

The log-normal distribution is appropriate for positively skew data. The pdf of the log-normal distribution, denoted by $\texttt{LOGNO}(\mu, \sigma)$, is defined by

$$f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{y} \exp \left\{ -\frac{[\log(y) - \mu]^2}{2\sigma^2} \right\}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Here $E(Y) = \mu$ and $\text{Var}(Y) = \sigma^2\mu^2$. This a reparameterization of Johnson et al. (1994) p 343 equation (17.23) obtained by setting $\sigma^2 = 1/\alpha$ and $\mu = \alpha\beta$. 

235
for \( y > 0 \), where \( \mu > 0 \) and \( \sigma > 0 \). Here \( E(Y) = \omega^{1/2}e^{\mu} \) and \( \text{Var}(Y) = \omega(\omega - 1)e^{2\mu} \), where \( \omega = \exp(\sigma^2) \).

**Log normal family (i.e. original Box-Cox) (LNO)**

The `gamlss` function `LNO(\mu, \sigma, \nu)` allows the use of the Box-Cox power transformation approach, Box and Cox (1964), where the transformation \( Y(\nu) \) is applied to \( Y \) in order to remove skewness, where \( Z = (Y^\nu - 1)/\nu(\text{if } \nu \neq 0) + \log(Y)(\text{if } \nu = 0) \). The transformed variable \( Z \) is then assumed to have a normal NO(\( \mu, \sigma \)) distribution. The resulting distribution for \( Y \) is denoted by `LNO(\mu, \sigma, \nu)`. When \( \nu = 0 \), this results in the distribution in equation (15.3). For values of \( \nu \) different from zero we have the resulting three parameter distribution

\[
 f_Y(y|\mu, \sigma, \nu) = \frac{y^{\nu-1}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z - \mu)^2}{2\sigma^2}\right] \tag{15.4}
\]

for \( y > 0 \), where \( \mu > 0 \), \( \sigma > 0 \) and \(-\infty < \nu < \infty\), and where \( z = (y^\nu - 1)/\nu(\text{if } \nu \neq 0) + \log(y)(\text{if } \nu = 0) \). The distribution in (15.4) can be fitted for fixed \( \nu \) only, e.g. \( \nu = 0.5 \), using the following arguments of `gamlss()`: `family=LNO, nu.fix=TRUE, nu.start=0.5`. If \( \nu \) is unknown, it can be estimated from its profile likelihood. Alternatively instead of (15.4), the more orthogonal parameterization of (15.4) given by the BCCG distribution in Section 15.3.1 can be used.

**15.2.3 Inverse Gaussian distribution (IG)**

The inverse Gaussian distribution is appropriate for highly positive skew data. The pdf of the inverse Gaussian distribution, denoted by IG(\( \mu, \sigma \)) is defined by

\[
 f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2y^3}} \exp\left[-\frac{1}{2\mu^2\sigma^2y} (y - \mu)^2\right] \tag{15.5}
\]

for \( y > 0 \), where \( \mu > 0 \) and \( \sigma > 0 \) with \( E(Y) = \mu \) and \( \text{Var}(Y) = \sigma^2\mu^3 \). This is a reparameterisation of Johnson et al. (1994) p 261 equation (15.4a), obtained by setting \( \sigma^2 = 1/\lambda \).

**15.2.4 (WEI, WEI2, WEI3)**

First parameterization (WEI)

There are three version of the two parameter Weibull distribution implemented into the `gamlss` package. The first, denoted by `WEI(\mu, \sigma)` has the following parameterisation

\[
 f_Y(y|\mu, \sigma) = \frac{\mu y^{\sigma-1}}{\mu^\sigma} \exp\left[-\left(\frac{y}{\mu}\right)^\sigma\right] \tag{15.6}
\]

for \( y > 0 \), where \( \mu > 0 \) and \( \sigma > 0 \), [see Johnson et al. (1994) p629]. The mean and the variance of \( Y \) in this parameterisation (15.6) of the two parameter Weibull are given by \( E(Y) = \mu \Gamma\left(\frac{\mu}{\sigma} + 1\right) \) and \( \text{Var}(Y) = \mu^2 \left\{ \Gamma\left(\frac{\mu}{\sigma} + 1\right) - \left[\Gamma\left(\frac{\mu}{\sigma} + 1\right)\right]^2 \right\} \), from Johnson et al. (1994) p632. Although the parameter \( \mu \) is a scale parameter, it also affects the mean of \( Y \). The median of \( Y \) is \( m_Y = \mu(\log 2)^{1/\sigma} \), see Johnson et al. (1994), p630.
15.3. CONTINUOUS THREE PARAMETER DISTRIBUTION IN $\mathbb{R}^+$

Second parameterisation (WEI2)

The second parameterisation of the Weibull distribution, denoted by $\text{WEI2}(\mu, \sigma)$, is defined as

$$f_Y(y|\mu, \sigma) = \sigma y^\sigma - 1 \exp\left\{-\left(\frac{y}{\beta}\right)\right\}$$

(15.7)

for $y > 0$, where $\mu > 0$ and $\sigma > 0$, Johnson et al. (1994), p686. The mean of $Y$ in this parameterization (15.7) is $E(Y) = \mu^{-1/\sigma} \Gamma\left(\frac{1}{\sigma} + 1\right)$ and the variance of $Y$ is $Var(Y) = \mu^{-2/\sigma} \left\{\Gamma\left(\frac{2}{\sigma} + 1\right) - \left[\Gamma\left(\frac{1}{\sigma} + 1\right)\right]^2\right\}$.

In the second parameterisation of the Weibull distribution the two parameters $\mu$ and $\sigma$ are highly correlated, so the RS method of fitting is very slow and therefore the CG() method of fitting should be used.

Third parameterization (WEI3)

This is a parameterization of the Weibull distribution where $\mu$ is the mean of the distribution. This parameterization of the Weibull distribution, denoted by $\text{WEI3}(\mu, \sigma)$, is defined as

$$f_Y(y|\mu, \sigma) = \frac{\sigma}{\beta} \left(\frac{y}{\beta}\right)^{\sigma - 1} \exp\left\{-\left(\frac{y}{\beta}\right)^\sigma\right\}$$

(15.8)

for $y > 0$, where $\mu > 0$ and $\sigma > 0$ and where $\beta = \mu/\Gamma\left(\frac{1}{\sigma} + 1\right)$. The mean of $Y$ is given by $E(Y) = \mu$ and the variance $Var(Y) = \mu^2 \left\{\Gamma\left(\frac{2}{\sigma} + 1\right) / \Gamma\left(\frac{1}{\sigma} + 1\right)\right\}^2 - 1$.

15.3 Continuous three parameter distribution in $\mathbb{R}^+$

15.3.1 Box-Cox Cole and Green distribution (BCCG)

The Box-Cox Cole and Green distribution is suitable for positively or negatively skew data. Let $Y > 0$ be a positive random variable having a Box-Cox Cole and Green distribution, denoted here as $\text{BCCG}(\mu, \sigma, \nu)$, defined through the transformed random variable $Z$ given by

$$Z = \begin{cases} \frac{1}{\sigma\nu} \left[\left(\frac{Y}{\mu}\right)^\nu - 1\right], & \text{if } \nu \neq 0 \\ \frac{1}{\nu} \log\left(\frac{Y}{\mu}\right), & \text{if } \nu = 0 \end{cases}$$

(15.9)

for $0 < Y < \infty$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, and where the random variable $Z$ is assumed to follow a truncated standard normal distribution. The condition $0 < Y < \infty$ (required for $Y^\nu$ to be real for all $\nu$) leads to the condition $-1/(\sigma\nu) < Z < \infty$ if $\nu > 0$ and $-\infty < Z < -1/(\sigma\nu)$ if $\nu < 0$, which necessitates the truncated standard normal distribution for $Z$.

Hence the pdf of $Y$ is given by

$$f_Y(y) = \frac{y^{\nu-1} \exp\left(-\frac{1}{2}z^2\right)}{\mu^{\nu} \sigma^{\nu} \sqrt{2\pi} \Phi\left(\frac{1}{\sigma\nu}\right)}$$

(15.10)

where $z$ is given by (15.9) and $\Phi()$ is the cumulative distribution function (cdf) of a standard normal distribution.

If the truncation probability $\Phi\left(-\frac{1}{\sigma\nu}\right)$ is negligible, the variable $Y$ has median $\mu$. The parameterization (15.9) was used by Cole and Green (1992) who assumed a standard normal distribution for $Z$ and assumed that the truncation probability was negligible.
15.3.2 Generalized gamma distribution (GG, GG2)

First parameterization (GG)

The specific parameterization of the generalized gamma distribution used here and denoted by \( \text{GG}(\mu, \sigma, \nu) \) was used by Lopatatzidis and Green (2000), and is defined as

\[
f_Y(y|\mu, \sigma, \nu) = \frac{|\nu|y^{\nu-1} \exp \left\{ -\frac{y}{\sigma} \right\}}{\Gamma(\nu)\sigma^\nu} \quad (15.11)
\]

for \( y > 0 \), where \( \mu > 0 \), \( \sigma > 0 \) and \( -\infty < \nu < \infty \) and where \( z = (y/\mu)^\nu \) and \( \theta = 1/(\sigma^2 \nu^2) \).

The mean and variance of \( Y \) are given by \( E(Y) = \mu \Gamma \left( \frac{\nu}{2} + \frac{\nu}{\theta} \right) / \left[ \theta^{1/\nu} \Gamma(\theta) \right] \) and \( \text{Var}(Y) = \mu^2 \left\{ \Gamma(\theta) \Gamma \left( \theta + \frac{2}{\nu} \right) - \left[ \Gamma \left( \theta + \frac{1}{\theta} \right) \right]^2 \right\} / \left\{ \theta^{2/\nu} \Gamma(\theta)^2 \right\} \). Note that GG2 is not currently implemented in gamlass.

Second parameterisation (GG2)

A second parameterisation, given by Johnson et al., (1995), p401, denoted by \( \text{GG2}(\mu, \sigma, \nu) \), is defined as

\[
f_Y(y|\mu, \sigma, \nu) = \frac{\nu \exp \left\{ -\frac{y}{\sigma} \right\}}{\Gamma(\nu)\sigma^\nu} \quad (15.12)
\]

for \( y > 0 \), where \( -\infty < \mu < \infty \), \( \sigma > 0 \) and \( \nu > 0 \).

The mean and variance of \( Y \sim \text{GG2}(\mu, \sigma, \nu) \) can be obtained from those of \( \text{GG}(\mu, \sigma, \nu) \) since \( \text{GG}(\mu, \sigma, \nu) \equiv \text{GG2}(\nu, \mu^{-1/\nu}, \theta) \) and \( \text{GG2}(\mu, \sigma, \nu) \equiv \text{GG}(\nu^{1/\mu}, [\sigma^2 \nu]^{-1/2}, \mu) \).

15.3.3 Generalised inverse Gaussian distribution (GIG)

The parameterization of the generalized inverse Gaussian distribution, denoted by \( \text{GIG}(\mu, \sigma, \nu) \), is defined as

\[
f_Y(y|\mu, \sigma, \nu) = \left( \frac{c}{\mu} \right)^\nu \left[ \frac{y^{\nu-1}}{2\Gamma(\nu)\sigma^\nu} \right] \exp \left\{ -\frac{1}{2\sigma^2} \left( \frac{cy}{\mu} + \frac{\mu}{cy} \right) \right\} \quad (15.13)
\]

for \( y > 0 \), where \( \mu > 0 \), \( \sigma > 0 \) and \( -\infty < \nu < \infty \), where \( c = \left[ K_{\nu+1} \left( 1/\sigma^2 \right) \right] \left[ K_{\nu} \left( 1/\sigma^2 \right) \right]^{-1} \) and \( K_\lambda(t) = \frac{1}{2} \int_0^\infty x^{\lambda-1} \exp \left\{ -\frac{1}{2} t(x + x^{-1}) \right\} dx \).

Here \( E(Y) = \mu \) and \( \text{Var}(Y) = \mu^2 \left[ 2\sigma^2 (\nu + 1)/c + 1/c^2 - 1 \right] \). \( \text{GIG}(\mu, \sigma, \nu) \) is a reparameterization of the generalized inverse Gaussian distribution of Jorgensen (1982). Note also that \( \text{GIG}(\mu, \sigma, -0.5) \equiv \text{IG}(\mu, \sigma^{-1/2}) \) a reparameterization of the inverse Gaussian distribution.

15.3.4 Zero adjusted Gamma distribution (ZAGA)

The zero adjusted Gamma distribution is appropriate when the response variable \( Y \) takes values from zero to infinity including zero, i.e. \( [0, \infty) \). Hence \( Y = 0 \) has non zero probability \( \nu \). The pdf of the zero adjusted Gamma distribution, denoted by \( \text{ZAGA}(\mu, \sigma, \nu) \), is defined by

\[
f_Y(y|\mu, \sigma, \nu) = \begin{cases} \nu & \text{if } y = 0 \\ \left( 1 - \nu \right) \left[ \frac{1}{(\sigma^2 \mu)^{1/\nu}} \frac{y^{\nu-1} e^{-y/(\sigma^2 \mu)}}{\Gamma(1/\nu)} \right] & \text{if } y > 0 \end{cases} \quad (15.14)
\]

for \( 0 \leq y < \infty \), where \( 0 < \nu < 1 \), \( \mu > 0 \) and \( \sigma > 0 \) with \( E(Y) = (1 - \nu)\mu \) and \( \text{Var}(Y) = (1 - \nu)\mu^2 (\nu + \sigma^2) \).
15.3.5 Zero adjusted Inverse Gaussian distribution (ZAIG)

The zero adjusted inverse Gaussian distribution is appropriate when the response variable \( Y \) takes values from zero to infinity including zero, i.e. \([0, \infty)\). Hence \( Y = 0 \) has non zero probability \( \nu \). The pdf of the zero adjusted inverse Gaussian distribution, denoted by \( \text{ZAIG}(\mu, \sigma, \nu) \), is defined by

\[
f_Y(y|\mu, \sigma, \nu) = \begin{cases} 
\nu & \text{if } y = 0 \\
(1-\nu) \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp \left[ \frac{-1}{2\mu^2\sigma^2} (y - \mu)^2 \right] & \text{if } y > 0 
\end{cases} \tag{15.15}
\]

for \( 0 \leq y < \infty \), where \( 0 < \nu < 1 \), \( \mu > 0 \) and \( \sigma > 0 \) with \( \text{E}(Y) = (1-\nu)\mu \) and \( \text{Var}(Y) = (1-\nu)\mu^2(\nu + \mu \sigma^2) \).

15.4 Continuous four parameter distribution in \( \mathbb{R}^+ \)

15.4.1 Box-Cox \( t \) distribution (BCT)

Let \( Y \) be a positive random variable having a Box-Cox \( t \) distribution, Rigby and Stasinopoulos (2006), denoted by \( \text{BCT}(\mu, \sigma, \nu, \tau) \), defined through the transformed random variable \( Z \) given by (15.9), where the random variable \( Z \) is assumed to follow a truncated \( t \) distribution with degrees of freedom, \( \tau > 0 \), treated as a continuous parameter.

The pdf of \( Y \), a \( \text{BCT}(\mu, \sigma, \nu, \tau) \) random variable, is given by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = \frac{y^{\nu-1}f_T(z)}{\mu^\nu \sigma F_T(\frac{1}{\sigma \nu})} \tag{15.16}
\]

for \( y > 0 \), where \( \mu > 0 \), \( \sigma > 0 \) and \( -\infty < \nu < \infty \), and where \( z \) is given by (15.9) and \( f_T(t) \) and \( F_T(t) \) are respectively the pdf and cumulative distribution function of a random variable \( T \) having a standard \( t \) distribution with degrees of freedom parameter \( \tau > 0 \), i.e. \( T \sim t_\tau \equiv T \mathcal{F}(0, 1, \tau) \). If the truncation probability \( F_T(-\frac{1}{\sigma \nu}) \) is negligible, the variable \( Y \) has median \( \mu \).

15.4.2 Box-Cox power exponential distribution (BCPE)

Let \( Y \) be a positive random variable having a Box-Cox power exponential distribution, Rigby and Stasinopoulos (2004), denoted by \( \text{BCPE}(\mu, \sigma, \nu, \tau) \), defined through the transformed random variable \( Z \) given by (15.9), where the random variable \( Z \) is assumed to follow a truncated standard power exponential distribution with power parameter, \( \tau > 0 \), treated as a continuous parameter.

The pdf of \( Y \), a \( \text{BCPE}(\mu, \sigma, \nu, \tau) \) random variable, is given by (15.16), where \( f_T(t) \) and \( F_T(t) \) are respectively the pdf and cumulative distribution function of a random variable \( T \) having a standard power exponential distribution, \( T \sim \text{PE}(0, 1, \tau) \). If the truncation probability \( F_T(-\frac{1}{\sigma \nu}) \) is negligible, the variable \( Y \) has median \( \mu \).

15.4.3 Generalized Beta type 2 distribution (GB2)

This pdf of the generalized beta type 2 distribution, denoted by \( \text{GB2}(\mu, \sigma, \nu, \tau) \), is defined by

\[
f_Y(y|\mu, \sigma, \nu, \tau) = |\sigma| y^{\sigma \nu - 1} \left\{ \mu^{\sigma \nu} B(\nu, \tau) \left[ 1 + (y/\mu)^\sigma \right]^{\nu + \tau} \right\}^{-1} \\
= \frac{\Gamma(\nu + \tau)}{\Gamma(\nu) \Gamma(\tau)} \frac{\sigma(y/\mu)^{\sigma \nu}}{y [1 + (y/\mu)^\sigma]^{\nu + \tau}} \tag{15.17}
\]
for \( y > 0 \), where \( \mu > 0 \), \( -\infty < \sigma < \infty \), \( \nu > 0 \) and \( \tau > 0 \), McDonald and Xu (1995), equation (2.7). The mean and variance of \( Y \) are given by 

\[
E(Y) = \mu B(\nu + \frac{1}{\sigma}, \tau - \frac{1}{\sigma}) / B(\nu, \tau)
\]

for \( -\nu < \frac{1}{\sigma} < \tau \) and 

\[
E(Y^2) = \mu^2 B(\nu + \frac{2}{\sigma}, \tau - \frac{2}{\sigma}) / B(\nu, \tau)
\]

for \( -\nu < \frac{2}{\sigma} < \tau \), McDonald (1996), p434. Note the by setting \( \nu = 1 \) in 15.17 we obtain the Burr distribution:

\[
f_Y(y | \mu, \sigma, \tau) = \frac{\tau \sigma (y/\mu)^\sigma}{y [1 + (y/\mu)\sigma]^\tau + 1}.
\]

(15.18)

By setting \( \sigma = 1 \) in 15.17 we obtain the Generalized Pareto distribution:

\[
f_Y(y | \mu, \nu, \tau) = \frac{\Gamma(\nu + \tau)}{\Gamma(\nu)\Gamma(\tau)} \frac{(\mu \tau y^{-1})^{\nu-1}}{(y + \mu)^{\nu+\tau}}.
\]

(15.19)
Chapter 16

Continuous distribution in \( \mathbb{R}[0, 1] \)

16.1 Continuous two parameter distribution in \( \mathbb{R}[0, 1] \)

16.1.1 Beta distribution (BE, BEo)

The beta distribution is appropriate when the response variable takes values in a known restricted range, excluding the endpoints of the range. Appropriate standardization can be applied to make the range of the response variable (0,1), i.e. from zero to one excluding the endpoints. Note that \( 0 < Y < 1 \) so values \( Y = 0 \) and \( Y = 1 \) have zero density under the model.

First parameterization (BEo)

The original parameterization of the beta distribution, denoted by \( BEo(\mu, \sigma) \), has pdf given by

\[
f_Y(y|\mu, \sigma) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1-y)^{\beta-1}
\]

for \( 0 < y < 1 \), with parameters \( \mu > 0 \) and \( \sigma > 0 \). Here \( E(Y) = \mu/(\mu + \sigma) \) and \( Var(Y) = \mu\sigma(\mu + \sigma)^{-2}(\mu + \sigma + 1)^{-1} \).

Second parameterization (BE)

In the second parameterization of the beta distribution below the parameters \( \mu \) and \( \sigma \) are location and scale parameters that relate to the mean and standard deviation of \( Y \). The pdf of the beta distribution, denoted by \( BE(\mu, \sigma) \), is defined by

\[
f_Y(y|\mu, \sigma) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1-y)^{\beta-1}
\]

for \( 0 < y < 1 \), where \( \alpha = \mu(1-\sigma^2)/\sigma^2 \) and \( \beta = (1-\mu)(1-\sigma^2)/\sigma^2 \), \( \alpha > 0 \), and \( \beta > 0 \) and hence \( 0 < \mu < 1 \) and \( 0 < \sigma < 1 \). [Note the relationship between parameters \( (\mu, \sigma) \) and \( (\alpha, \beta) \) is given by \( \mu = \alpha/(\alpha + \beta) \) and \( \sigma = (\alpha + \beta + 1)^{-1/2} \).] In this parameterization, the mean of \( Y \) is \( E(Y) = \mu \) and the variance is \( Var(Y) = \sigma^2 \mu(1-\mu) \).

16.1.2 Beta inflated distribution (BEINF, BEINF0, BEINF1)

The beta inflated distribution is appropriate when the response variable takes values in a known restricted range including the endpoints of the range. Appropriate standardization can be applied to make the range of the response variable \( [0,1] \), i.e. from zero to one including the endpoints. Values zero and one for \( Y \) have non zero probabilities \( p_0 \) and \( p_1 \) respectively. The
probability (density) function of the inflated beta distribution, denoted by BEINF($\mu, \sigma, \nu, \tau$) is defined by

$$f_Y(y | \mu, \sigma, \nu, \tau) = \begin{cases} p_0 \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} & \text{if } y = 0 \\ (1 - p_0 - p_1) \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases}$$

(16.2)

for $0 \leq y \leq 1$, where $\alpha = \mu(1-\sigma^2)/\sigma^2$, $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, $p_0 = \nu(1+\nu+\tau)^{-1}$, $p_1 = \tau(1+\nu+\tau)^{-1} \text{ so } \alpha > 0$, $\beta > 0$, $0 < p_0 < 1$, $0 < p_1 < 1 - p_0$. Hence BEINF($\mu, \sigma, \nu, \tau$) has parameters $\mu = \alpha/(\alpha + \beta)$ and $\sigma = (\alpha + \beta + 1)^{-1/2}$, $\nu = p_0/p_2$, $\tau = p_1/p_2$ where $p_2 = 1 - p_0 - p_1$. Hence $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$ and $\tau > 0$. Note that $E(y) = \frac{\mu + \nu}{(1+\nu+\tau)}$.

The probability (density) function of the inflated at zero beta distribution, denoted by BEINF0($\mu, \sigma, \nu$) is defined by

$$f_Y(y | \mu, \sigma, \nu) = \begin{cases} p_0 \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} & \text{if } y = 0 \\ (1 - p_0) \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0 < y < 1 \end{cases}$$

(16.3)

for $0 \leq y < 1$, where $\alpha = \mu(1-\sigma^2)/\sigma^2$, $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, $p_0 = \nu(1+\nu)^{-1}$, so $\alpha > 0$, $\beta > 0$, $0 < p_0 < 1$. Hence BEINF0($\mu, \sigma, \nu$) has parameters $\mu = \alpha/(\alpha + \beta)$ and $\sigma = (\alpha + \beta + 1)^{-1/2}$, $\nu = p_0/p_2$. Hence $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$. Note that for BEINF0($\mu, \sigma, \nu$), $E(y) = \frac{\mu}{(1+\nu+\tau)}$.

The probability (density) function of the inflated beta distribution, denoted by BEINF1($\mu, \sigma, \nu$) is defined by

$$f_Y(y | \mu, \sigma, \nu) = \begin{cases} (1 - p_1) \frac{1}{B(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases}$$

(16.4)

for $0 < y \leq 1$, where $\alpha = \mu(1-\sigma^2)/\sigma^2$, $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, $p_1 = \tau(1+\tau)^{-1} \text{ so } \alpha > 0$, $\beta > 0$, $0 < p_1 < 1$. Hence BEINF1($\mu, \sigma, \nu$) has parameters $\mu = \alpha/(\alpha + \beta)$ and $\sigma = (\alpha + \beta + 1)^{-1/2}$, $\nu = p_1/(1 - p_2)$. Hence $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$. Note that $E(y) = \frac{\nu + \mu}{(1+\nu+\tau)}$.

For different parametrization of the BEINF0($\mu, \sigma, \nu$) and BEINF1($\mu, \sigma, \nu$) distributions see also BEZI($\mu, \sigma, \nu$) and BEOI($\mu, \sigma, \nu$) distributions contributed to gamlss by Raydonal Ospina, Ospina and Ferrari (2010).

### 16.1.3 Generalized Beta type 1 distribution (GB1)

The generalized beta type 1 distribution is defined by assuming $Z = Y^\tau/[\nu + (1 - \nu)Y^\tau] \sim BE(\mu, \sigma)$. Hence, the pdf of generalized beta type 1 distribution, denoted by $GB1(\mu, \sigma, \nu, \tau)$, is given by

$$f_Y(y | \mu, \sigma, \nu, \tau) = \frac{\tau^\nu \beta^\nu y^{\nu-1}(1-y)^{\alpha-1} B(\alpha, \beta)}{\nu + (1 - \nu) y^\tau}$$

(16.5)

for $0 < y < 1$, where $\alpha = \mu(1-\sigma^2)/\sigma^2$ and $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, $\alpha > 0$ and $\beta > 0$. Hence, $GB1(\mu, \sigma, \nu, \tau)$ has adopted parameters $\mu = \alpha/(\alpha + \beta)$, $\sigma = (\alpha + \beta + 1)^{-1/2}$, $\nu$ and $\tau$, where $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$ and $\tau > 0$. The beta $BE(\mu, \sigma)$ distribution is a special case of $GB1(\mu, \sigma, \nu, \tau)$ where $\nu = 1$ and $\tau = 1$. 
Chapter 17

Count data distributions

17.1 Count data one parameter distributions

17.1.1 Poisson distribution (PO)

Poisson distribution

The probability function (pf) of the Poisson distribution, denoted here as \( \text{PO}(\mu) \), is given by

\[
p_Y(y|\mu) = P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!}
\]  

(17.1)

where \( y = 0, 1, 2, \ldots \), where \( \mu > 0 \).

The mean, variance, skewness and kurtosis of \( Y \) are given by \( E(Y) = \mu \), \( \text{Var}(Y) = \mu \), \( \sqrt{\beta_1} = \mu^{-0.5} \) and \( \beta_2 = 3 + \mu^{-1} \) respectively. [See Johnson et al. (1993), p 151.]

The probability generating function of \( Y \) is given by \( G_Y(t) = e^{\mu(t-1)} \).

Note that the Poisson distribution has the property that \( E[Y] = \text{Var}[Y] \) and that \( \beta_2 - \beta_1 - 3 = 0 \). The index of dispersion, that is, the ratio \( \text{Var}[Y]/E[Y] \) is equal to one for the Poisson distribution. For \( \text{Var}[Y] > E[Y] \) we have overdispersion and for \( \text{Var}[Y] < E[Y] \) we have underdispersion or repulsion. The distribution is skew for small values of \( \mu \), but almost symmetric for large \( \mu \) values.

17.1.2 Logarithmic distribution (LG)

The probability function of the logarithmic distribution, denoted here as \( \text{LG}(\mu) \), is given by

\[
p_Y(y|\mu) = \frac{\alpha \mu^y}{y}, \quad \text{for} \quad y = 1, 2, \ldots
\]  

(17.2)

where \( \alpha = -[\log(1 - \mu)]^{-1} \) for \( 0 < \mu < 1 \). Note that the range of \( Y \) starts from 1. The mean and variance of \( Y \) are given by \( E(Y) = \frac{\alpha \mu}{(1-\mu)} \) and \( \text{Var}(Y) = \frac{\alpha \mu (1-\alpha \mu)}{(1-\mu)^2} \); see Johnson et al. (2005) p.302-342.
17.2 Count data two parameters distributions

17.2.1 Negative Binomial distribution (NBI, NBII)

First parameterization: Negative Binomial type I (NBI)

The probability function of the negative binomial distribution type I, denoted here as \( \text{NBI}(\mu, \sigma) \), is given by

\[
p_Y(y|\mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left( \frac{\sigma \mu}{1 + \sigma \mu} \right)^{y} \left( \frac{1}{1 + \sigma \mu} \right)^{1/\sigma}
\]

for \( y = 0, 1, 2, \ldots \), where \( \mu > 0, \sigma > 0 \).

The above parameterization is equivalent to that used by Anscombe (1950) except he used \( \alpha = 1/\sigma \), as pointed out by Johnson et al. (1993), p 200, line 5.

The mean, variance, skewness and kurtosis of \( Y \) are given by

\[
E(Y) = \mu, \quad \text{Var}(Y) = \mu + \sigma \mu^2, \quad \sqrt{\beta_1} = \frac{(1 + 2\sigma)}{(\mu + \sigma \mu)^{0.5}} \quad \text{and} \quad \beta_2 = 3 + \frac{(1 + 6\sigma + 6\sigma^2)}{(\mu + \sigma \mu)^{-1}}
\]

respectively.

The probability generating function of \( Y \) is given by

\[
g_Y(t) = \left[ 1 + \sigma \mu (1 - t) \right]^{-1/\sigma}.
\]

Second parameterization: Negative Binomial type II (NBII)

The probability function of the negative binomial distribution type II, denoted here as \( \text{NBII}(\mu, \sigma) \), is given by

\[
p_Y(y|\mu, \sigma) = \frac{\Gamma(y + \mu/\sigma)}{\Gamma(\mu/\sigma)\Gamma(y + 1 + \mu/\sigma)} (\mu + \sigma \mu)^{-y}
\]

for \( y = 0, 1, 2, \ldots \), where \( \mu > 0 \) and \( \sigma > 0 \).

This parameterization was used by Evans (1953) as pointed out by Johnson et al. (1993) p 200 line 7.

The mean, variance, skewness and kurtosis of \( Y \) are given by

\[
E(Y) = \mu, \quad \text{Var}(Y) = \mu + \sigma \mu, \quad \sqrt{\beta_1} = \frac{(1 + 2\sigma)}{(\mu + \sigma \mu)^{0.5}} \quad \text{and} \quad \beta_2 = 3 + \frac{(1 + 6\sigma + 6\sigma^2)}{(\mu + \sigma \mu)^{-1}}
\]

respectively.

The probability generating function of \( Y \) is given by

\[
g_Y(t) = [1 + \sigma (1 - t)]^{-\mu/\sigma}.
\]

17.2.2 Poisson-inverse Gaussian distribution (PIG)

The probability function of the Poisson-inverse Gaussian distribution, denoted by \( \text{PIG}(\mu, \sigma) \), is given by

\[
p_Y(y|\mu, \sigma) = \frac{\left( \frac{2\alpha}{\pi} \right)^{\frac{1}{2}} \mu^{\frac{1}{2}} \mu^{\frac{1}{2}} K_{y - \frac{1}{2}}(\alpha)}{(\alpha \sigma)^{y}}
\]

where \( \alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma} \), for \( y = 0, 1, 2, \ldots, \infty \) where \( \mu > 0 \) and \( \sigma > 0 \) and \( K_{y - \frac{1}{2}}(\alpha) = \frac{1}{\pi} \int_{0}^{\infty} x^{y-1} \exp\left\{-\frac{1}{2}t(x + x^{-1})\right\} dx \) is the modified Bessel function of the third kind.
The mean, variance, skewness and kurtosis of $Y$ are given by $E(Y) = \mu$, $Var(Y) = \mu + \sigma \mu^2$, $\sqrt{\beta_1} = (1 + 3\mu \sigma + 3\mu^2 \sigma^2) (1 + \mu \sigma)^{-1.5} \mu^{-0.5}$ and $\beta_2 = 3 + (1 + 7\mu \sigma + 18\mu^2 \sigma^2 + 15\mu^3 \sigma^3) (1 + \mu \sigma)^{-2} \mu^{-1}$ respectively.

The probability generating function of $Y$ is given by

$$G_Y(t) = \left(\frac{2q}{\pi}\right)^{\frac{1}{2}} e^{1/\sigma} K_{-1/2}(q)$$

where $q^2 = \sigma^{-2} + 2\mu(1-t)\sigma^{-1}$.

Note that the above parameterization was used by Dean, Lawless and Willmot (1989). It is also a special case of the **gamlss.family** distribution $SI(\mu, \sigma, \nu)$ when $\nu = -\frac{1}{2}$.

### 17.2.3 Zero inflated poisson (ZIP, ZIP2)

#### First parameterization (ZIP)

Let $Y = 0$ with probability $\sigma$ and $Y \sim Po(\mu)$ with probability $(1 - \sigma)$, then $Y$ has a zero inflated Poisson distribution, denoted by **ZIP**(\(\mu, \sigma\)), given by

$$p_Y(y|\mu, \sigma) = \begin{cases} 
\sigma + (1 - \sigma)e^{-\mu}, & \text{if } y = 0 \\
(1 - \sigma)\frac{\mu^y}{y!} e^{-\mu}, & \text{if } y = 1, 2, 3, \ldots 
\end{cases}$$

(17.3)

See Johnson *et al* (1993), p 186, equation (4.100) for this parametrization. This parametrization was also used by Lambert (1992).

The mean, variance, skewness and kurtosis of $Y$ are given by $E(Y) = (1 - \sigma)\mu$, $Var(Y) = \mu (1 - \sigma)(1 + \mu \sigma)$, $\sqrt{\beta_1} = \mu_3/[Var(Y)]^{1.5}$ and $\beta_2 = \mu_4/[Var(Y)]^2$, where

$$\mu_3 = \mu(1 - \sigma)\left[1 + 3\mu \sigma + \mu^2 \sigma(2\sigma - 1)\right]$$

and

$$\mu_4 = \mu(1 - \sigma)\left[1 + \mu(3 + 4\sigma) + 6\mu^2 \sigma^2 + \mu^3 \sigma(1 + 3\sigma - 3\sigma^3)\right]$$

(17.4)

respectively.

The probability generating function of $Y$ is given by

$$G_Y(t) = \sigma + (1 - \sigma)e^{\mu(t-1)}.$$  

#### Second parameterization (ZIP2)

A different parameterization of the zero inflated poisson distribution, denoted by **ZIP2**(\(\mu, \sigma\)), is given by

$$p_Y(y|\mu, \sigma) = \begin{cases} 
\sigma + (1 - \sigma)e^{-\left(\frac{\mu}{1 - \sigma}\right)}, & \text{if } y = 0 \\
(1 - \sigma)\frac{\mu^y}{y!(1 - \sigma)^y} e^{-\left(\frac{\mu}{1 - \sigma}\right)}, & \text{if } y = 1, 2, 3, \ldots 
\end{cases}$$

(17.5)

The mean of $Y$ in (17.5) is given by $E(Y) = \mu$ and the variance by $Var(Y) = \mu + \mu^2 \sigma/(1 - \sigma)$. 
The mean, variance, skewness and kurtosis of $Y$ are given by $E(Y) = \mu$, $Var(Y) = \mu + \mu^2 \frac{\alpha}{(1-\sigma)}$, $\sqrt{\beta_1} = \mu \beta_3/|Var(Y)|^{1.5}$ and $\beta_2 = \mu_4/|Var(Y)|^2$, where

$$\mu_3 = \mu \left[ 1 + \frac{3\mu \sigma}{(1-\sigma)} + \mu^2 \sigma(2\sigma - 1) \frac{\mu^2 \sigma(2\sigma - 1)}{(1-\sigma)^2} \right]$$

and

$$\mu_4 = \mu \left[ 1 + \frac{\mu(3 + 4\sigma)}{(1-\sigma)} + \frac{6\mu^2 \sigma^2}{(1-\sigma)^2} + \frac{\mu^3 (1 + 3\sigma - 3\sigma^3)}{(1-\sigma)^3} \right]$$

(17.6)

respectively.

The probability generating function of $Y$ is given by

$$G_Y(t) = \sigma + (1-\sigma)e^{\mu(t-1)/(1-\sigma)}.$$ 

### 17.2.4 Zero altered (or adjusted) poisson (ZAP)

Let $Y = 0$ with probability $\sigma$ and $Y \sim POtr(\mu)$ with probability $(1-\sigma)$, where $POtr(\mu)$ is a Poisson truncated at zero distribution, then $Y$ has a zero adjusted Poisson distribution, denoted by $ZAP(\mu, \sigma)$, given by

$$p_Y(y|\mu, \sigma) = \begin{cases} 
\sigma, & \text{if } y = 0 \\
\frac{(1-\sigma)\mu^y}{y(1-e^{-\mu})}, & \text{if } y = 1, 2, 3, \ldots 
\end{cases}$$

(17.7)

The mean of $Y$ in this parametrization is given by $E(Y) = (1-\sigma)\mu/(1-e^{-\mu})$ and its variance by $Var(Y) = (1-\sigma)\mu^2/(1-e^{-\mu}) - [E(Y)]^2$.

Let $B = (1-\sigma)/(1-e^{-\mu})$.

The mean, variance, skewness and kurtosis of $Y$ are given by $E(Y) = B\mu$, $Var(Y) = B\mu + B\mu^2 - B^2\mu^2$, $\sqrt{\beta_1} = \mu_3/|Var(Y)|^{1.5}$ and $\beta_2 = \mu_4/|Var(Y)|^2$, where

$$\mu_3 = B\mu \left[ 1 + 3\mu(1-B) + \mu^2(1-3B+2B^2) \right]$$

and

$$\mu_4 = B\mu \left[ 1 + \mu(7 - 4B) + 6\mu^2(1 - 2B + B^2) + \mu^3(1 - 4B + 6B^2 - 3B^3) \right]$$

(17.8)

respectively.

The probability generating function of $Y$ is given by

$$G_Y(t) = (1-B) + Be^{\mu(t-1)}.$$ 

### 17.2.5 Zero altered (or adjusted) logarithmic (ZALG)

Let $Y = 0$ with probability $\sigma$ and $Y \sim LG(\mu)$, a logarithmic distribution with probability $(1-\sigma)$, then $Y$ has a zero altered (adjusted) logarithmic distribution, denoted by $ZALG(\mu, \sigma)$, with probability function given by

$$p_Y(y|\mu, \sigma) = \begin{cases} 
\sigma, & \text{if } y = 0 \\
(1-\sigma)\alpha^y y, & \text{if } y = 1, 2, 3, \ldots 
\end{cases}$$

(17.9)

where $\alpha = \frac{[log(1-\mu)]^{-1}}{log(1-\sigma)}$ for $0 < \mu < 1$ and $0 < \sigma < 1$. The mean and variance of $Y$ are given by $E(Y) = (1-\sigma)\alpha \mu/(1-\mu)$ and its variance by $Var(Y) = \frac{(1-\sigma)\alpha \mu/[1-(1-\sigma)\alpha \mu]}{(1-\mu)^2}$. 
17.3 Count data three parameters distributions

17.3.1 Delaporte distribution (DEL)

The probability function of the Delaporte distribution, denoted by \( \text{DEL}(\mu, \sigma, \nu) \), is given by

\[
p_Y(y | \mu, \sigma, \nu) = \frac{e^{-\mu y}}{\Gamma(1/\sigma)} \left[ 1 + \mu \sigma (1 - \nu) \right]^{-1/\sigma} S
\]

where

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

for \( y = 0, 1, 2, \ldots, \infty \) where \( \mu > 0, \sigma > 0 \) and \( 0 < \nu < 1 \). This distribution is a reparameterization of the distribution given by Wimmer and Altmann (1999) p 515-516 where \( \alpha = \mu \nu, k = 1/\sigma \) and \( \rho = [1 + \mu \sigma (1 - \nu)]^{-1} \).

The mean, variance, skewness and kurtosis of \( Y \) are given by \( E(Y) = \mu \), \( \text{Var}(Y) = \mu + \mu^2 \sigma (1 - \nu)^2 \), \( \sqrt{\text{Var}(Y)} = \mu \left[ 1 + 3 \mu \sigma (1 - \nu)^2 + 2 \mu^2 \sigma^2 (1 - \nu)^3 \right] / \left[ \text{Var}(Y) \right]^{1.5} \) and

\[
\beta_2 = 3 + \left\{ \mu \left[ 1 + 7 \mu \sigma (1 - \nu)^2 + 12 \mu^2 \sigma^2 (1 - \nu)^3 + 6 \mu^3 \sigma^3 (1 - \nu)^4 \right] / \left[ \text{Var}(Y) \right]^2 \right\}
\]

respectively.

The probability generating function of \( Y \) is given by

\[
G_Y(t) = e^{\mu \nu (t-1)} \left[ 1 + \mu \sigma (1 - \nu)(1 - t) \right]^{-1/\sigma}.
\]

17.3.2 Sichel distribution (SI, SICHEL)

First parameterization (SI)

The probability function of the first parameterization of the Sichel distribution, denoted by \( \text{SI}(\mu, \sigma, \nu) \), is given by

\[
p_Y(y | \mu, \sigma, \nu) = \frac{\mu^y K_{y+\nu}(\alpha)}{(\alpha \sigma)^{y+\nu} \Gamma(\frac{1}{\sigma})}
\]

where \( \alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma} \), for \( y = 0, 1, 2, \ldots, \infty \) where \( \mu > 0, \sigma > 0 \) and \( -\infty < \nu < \infty \) and \( K_\lambda(t) = \frac{1}{\sqrt{\pi}} \int_0^\infty \lambda^{-1} \exp(-\frac{1}{2}t(x + x^{-1})) dx \) is the modified Bessel function of the third kind. Note that the above parameterization is different from Stein, Zucchini and Juritz (1988) who use the above probability function but treat \( \mu, \alpha \) and \( \nu \) as the parameters. Note that \( \sigma = [(\mu^2 + \alpha^2)^{\frac{1}{2}} - \mu]^{-1} \).

Second parameterization (SICHEL)

The second parameterization of the Sichel distribution, Rigby, Stasinopoulos and Akantziliotou (2008), denoted by \( \text{SICHEL}(\mu, \sigma, \nu) \), is given by

\[
p_Y(y | \mu, \sigma, \nu) = \frac{(\mu/c)^y K_{y+\nu}(\alpha)}{y! (\alpha \sigma)^{y+\nu} K_{\nu}(\frac{1}{\sigma})}
\]

for \( y = 0, 1, 2, \ldots, \infty \), where \( \alpha^2 = \sigma^{-2} + 2\mu(c\sigma)^{-1}, c = K_{\nu+1}(1/\sigma)/K_{\nu}(1/\sigma) \) and \( K_\lambda(t) = \frac{1}{\sqrt{\pi}} \int_0^\infty \lambda^{-1} \exp(-\frac{1}{2}t(x + x^{-1})) dx \) is the modified Bessel function of the third kind.
The mean and variance of $Y$ are given by $E(Y) = \mu$ and $Var(Y) = \mu + \mu^2 \left[ 2\sigma(\nu + 1)/c + 1/c^2 - 1 \right]$ respectively. The skewness and kurtosis of $Y$ are given in the Appendix to Chapter 6.

The probability generating function of $Y$ is given by

$$G_Y(t) = \frac{K_\nu(q)}{(q\sigma)^rK_{\nu}(1/\sigma)}$$

where $q^2 = \sigma^{-2} + 2\mu(1-t)/(c\sigma)^{-1}$.

### 17.3.3 Zero inflated negative binomial distribution (ZINBI)

Let $Y = 0$ with probability $\nu$ and $Y \sim NBI(\mu, \sigma)$, with probability $(1 - \nu)$, then $Y$ has a zero inflated negative binomial distribution, denoted by $\text{ZINBI}(\mu, \sigma, \nu)$, with probability function given by

$$p_Y(y|\mu, \sigma, \nu) = \begin{cases} \nu + (1 - \nu)p_Y(0|\mu, \sigma), & \text{if } y = 0 \\ (1 - \nu)p_Y(y|\mu, \sigma), & \text{if } y = 1, 2, 3, \ldots \end{cases} \quad (17.13)$$

for $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$, where $Y_1 \sim NBI(\mu, \sigma)$ so

$$p_{Y_1}(0|\mu, \sigma) = (1 + \sigma \mu)^{-\frac{1}{\sigma}}$$

and

$$p_{Y_1}(y|\mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y + 1)} \left( \frac{\sigma \mu}{1 + \sigma \mu} \right)^y \left( \frac{1}{1 + \sigma \mu} \right)^{1/\sigma}$$

for $y = 0, 1, 2, 3, \ldots$.

The mean, variance, skewness and kurtosis of $Y$ are given by $E(Y) = (1 - \nu)\mu$, $Var(Y) = \mu(1 - \nu)[1 + (\sigma + \nu)\mu]$, $\sqrt{\beta_1} = \mu_3/[Var(Y)]^{1.5}$ and $\beta_2 = \mu_4/[Var(Y)]^2$, where

$$\mu_3 = \mu(1 - \nu) \{ 1 + 3\mu(\sigma + \nu) + \mu^2 [(1 + \sigma)(2\sigma + 3\nu - 2) + 2(1 - \nu)^2] \}$$

and

$$\mu_4 = \mu(1 - \nu) \{ 1 + \mu(3 + 7\sigma + 4\nu) + 6\mu^2 [(1 + \sigma)(2\sigma + 2\nu - 1) + (1 - \nu)^2] + \mu^3 [(1 + \sigma)(1 + 2\sigma)(3\sigma + 4\nu - 3) + 3(1 - \nu)^2(1 + 2\sigma + \nu)] \}$$

respectively.

The probability generating function of $Y$ is given by

$$G_Y(t) = \nu + (1 - \nu) [1 + \mu \sigma(1-t)]^{-1/\sigma}.$$

### 17.3.4 Zero altered (or adjusted) negative binomial distribution (ZANBI)

Let $Y = 0$ with probability $\nu$ and $Y \sim NBI_{tr}(\mu, \sigma)$, with probability $(1 - \nu)$, where $NBI_{tr}(\mu, \sigma)$ is a negative binomial truncated at zero distribution, then $Y$ has a zero altered (or adjusted) negative binomial distribution, denoted by $\text{ZANBI}(\mu, \sigma, \nu)$, with probability function given by

$$p_Y(y|\mu, \sigma, \nu) = \begin{cases} \nu, & \text{if } y = 0 \\ (1 - \nu)p_{Y_1}(y|\mu, \sigma) \left[ 1 - p_{Y_1}(0|\mu, \sigma) \right]^{-1}, & \text{if } y = 1, 2, 3, \ldots \end{cases} \quad (17.14)$$
for $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$ where $Y_1 \sim NBI(\mu, \sigma)$ so
\[
p_{Y_1}(0|\mu, \sigma) = (1 + \sigma \mu)^{-1/\sigma}
\]
and
\[
p_{Y_1}(y|\mu, \sigma) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(y + 1)} \left(\frac{\sigma \mu}{1 + \sigma \mu}\right)^y \left(\frac{1}{1 + \sigma \mu}\right)^{1/\sigma}
\]
for $y = 0, 1, 2, 3, \ldots$

Let $p_0 = p_{Y_1}(0|\mu, \sigma) = (1 + \sigma \mu)^{-1/\sigma}$ and let $B = (1 - \nu)/(1 - p_0)$

The mean, variance, skewness and kurtosis of $Y$ are given by
\[
E(Y) = B \mu, \quad Var(Y) = B \mu \left[1 + \mu(1 + \sigma) - B^2 \mu^2\right],
\]
\[
\sqrt{\beta_1} = \frac{\mu^3}{[Var(Y)]^{1.5}} \quad \text{and} \quad \beta_2 = \frac{\mu_4}{[Var(Y)]^2},
\]
where
\[
\mu_3 = B \mu \left[1 + 3\mu(1 + \sigma - B) + \mu^2 \left[(1 + \sigma)(1 + 2\sigma - 3B) + 2B^2\right]\right]
\]
and
\[
\mu_4 = B \mu \left[1 + \mu \left[7(1 + \sigma) - 4B\right] + 6\mu^2 \left[(1 + \sigma)(1 + 2\sigma - 2B(1 + \sigma) + B^2)\right] + \mu^3 \left[(1 + \sigma)(1 + 2\sigma - 3B) + 2B^2\right]\right](1 + \sigma - 3B^2)\right] \}
\]
respectively.

The probability generating function of $Y$ is given by
\[
G_Y(t) = (1 - B) + B \left[1 + \mu \sigma (1 - t)\right]^{-1/\sigma}.
\]

### 17.3.5 Zero inflated Poisson inverse Gaussian distribution (ZIPIG)

Let $Y = 0$ with probability $\nu$ and $Y \sim PIG(\mu, \sigma)$, with probability $(1 - \nu)$, then $Y$ has a zero inflated Poisson inverse Gaussian distribution, denoted by ZIPIG($\mu, \sigma, \nu$), with probability function given by
\[
p_Y(y|\mu, \sigma, \nu) = \begin{cases} 
\nu + (1 - \nu) p_{Y'}(0|\mu, \sigma), & \text{if } y = 0 \\
(1 - \nu) p_{Y'}(y|\mu, \sigma), & \text{if } y = 1, 2, 3, \ldots
\end{cases}
\]  \tag{17.15}
for $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$, where $Y' \sim PIG(\mu, \sigma)$. The mean of $Y$ is given by $E(Y) = (1 - \nu) \mu$ and the variance by $Var(Y) = (1 - \nu) \mu \left[1 + (\sigma + \nu) \mu\right]$. 
Chapter 18

Binomial type distributions

18.1 Binomial type data one parameter distributions

18.1.1 The Binomial distribution (BI)

The probability function of the binomial distribution, denoted here as BI(n, µ), is given by

\[ p_Y(y|n, µ) = P(Y = y|n, µ) = \frac{n!}{y!(n-y)!} \mu^y (1-\mu)^{n-y} \]

for \( y = 0, 1, 2, ..., n \), where \( 0 < \mu < 1 \), (and \( n \) is a known positive integer), with \( E(Y) = n\mu \) and \( Var(Y) = n\mu(1-\mu) \). See Johnson et al. (1993), p 105 where \( \mu = p \).

18.2 Binomial type data two parameters distributions

18.2.1 Beta Binomial distribution (BB)

The probability function of the beta binomial distribution denoted here as BB(n, µ, σ) is given by

\[ p_Y(y|\mu, \sigma) = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma\left(\frac{1}{\sigma}y + \frac{\mu}{\sigma}\right)\Gamma\left[n + \frac{(1-\mu)}{\sigma} - y\right]}{\Gamma\left(n + \frac{1}{\sigma}\right)\Gamma\left(\frac{\mu}{\sigma}\right)\Gamma\left(\frac{1-\mu}{\sigma}\right)} \tag{18.1} \]

for \( y = 0, 1, 2, ..., n \), where \( 0 < \mu < 1 \) and \( \sigma > 0 \) (and \( n \) is a known positive integer). Note that \( E(Y) = n\mu \) and \( Var(Y) = n\mu(1-\mu) \left[ 1 + \frac{\sigma}{n-1} \right] \).

The binomial BI(n,µ) distribution is the limiting distribution of BB(n,µ,σ) as \( \sigma \to 0 \). For \( \mu = 0.5 \) and \( \sigma = 0.5 \), BB(n,µ,σ) is a uniform distribution.

18.2.2 Zero altered (or adjusted) binomial (ZABI)

Let \( Y = 0 \) with probability \( \sigma \) and \( Y \sim BItr(n, \mu) \) with probability \((1-\sigma)\), where BItr(n,µ) is a Binomial truncated at zero distribution, then \( Y \) has a zero altered (or adjusted) binomial
distribution, denoted by $\text{ZABI}(n, \mu, \sigma)$, given by

$$p_Y(y|n, \mu, \sigma) = \begin{cases} 
\sigma, & \text{if } y = 0 \\
\frac{(1-\sigma)y^y(1-\mu)^{n-y}}{[1-(1-\mu)^n]^y}, & \text{if } y = 1, 2, 3, \ldots
\end{cases}$$

(18.2)

For $0 < \mu < 1$, and $0 < \sigma < 1$. The mean and variance of $Y$ are given by

$$E(Y) = \frac{(1 - \sigma) n \mu}{[1 - (1 - \mu)^n]}$$

and

$$Var(Y) = \frac{n \mu (1 - \sigma) (1 - \mu + n \mu)}{[1 - (1 - \mu)^n]} - [E(Y)]^2$$

respectively.

### 18.2.3 Zero inflated binomial (ZIBI)

Let $Y = 0$ with probability $\sigma$ and $Y \sim BI(n, \mu)$ with probability $(1 - \sigma)$, then $Y$ has a zero inflated binomial distribution, denoted by $\text{ZIBI}(n, \mu, \sigma)$, given by

$$p_Y(y|n, \mu, \sigma) = \begin{cases} 
\nu, & \text{if } y = 0 \\
\frac{(1-\nu)p_Y(y|n, \mu, \sigma)}{[1-p_Y(0|n, \mu, \sigma)]}, & \text{if } y = 1, 2, 3, \ldots
\end{cases}$$

(18.3)

For $0 < \mu < 1$, and $0 < \sigma < 1$. The mean and variance of $Y$ are given by

$$E(Y) = (1 - \sigma) n \mu$$

and

$$Var(Y) = n \mu (1 - \sigma) [1 - \mu + n \mu \sigma]$$

respectively.

### 18.3 Binomial type data three parameters distributions

#### 18.3.1 Zero altered (or adjusted) beta binomial (ZABB)

Let $Y = 0$ with probability $\nu$ and $Y \sim BBtr(n, \mu, \sigma)$ with probability $(1-\nu)$, where $BBtr(n, \mu, \sigma)$ is a beta binomial truncated at zero distribution, then $Y$ has a zero altered (or adjusted) beta binomial distribution, denoted by $\text{ZABB}(n, \mu, \sigma, \nu)$, given by

$$p_Y(y|n, \mu, \sigma, \nu) = \begin{cases} 
\nu, & \text{if } y = 0 \\
\frac{(1-\nu)p_Y(y|n, \mu, \sigma)}{[1-p_Y(0|n, \mu, \sigma)]}, & \text{if } y = 1, 2, 3, \ldots
\end{cases}$$

(18.4)

where $Y' \sim BB(n, \mu, \sigma)$. For $0 < \mu < 1$, $\sigma > 0$ and $0 < \nu < 1$. The mean and variance of $Y$ are given by

$$E(Y) = \frac{(1 - \nu) n \mu}{[1 - p_Y(0|n, \mu, \sigma)]}$$
18.4 Binomial type data three parameters distributions

18.4.1 Zero inflated beta binomial (ZIBB)

Let $Y = 0$ with probability $\nu$ and $Y \sim BB(n, \mu, \sigma)$ with probability $(1 - \nu)$, then $Y$ has a zero inflated beta binomial distribution, denoted by $ZIBB(n, \mu, \sigma, \nu)$, given by

$$p_Y(y|n, \mu, \sigma, \nu) = \begin{cases} 
\nu + (1 - \nu) p_{Y'}(0|n, \mu, \sigma), & \text{if } y = 0 \\
(1 - \nu) p_{Y'}(y|n, \mu, \sigma), & \text{if } y = 1, 2, 3, \ldots 
\end{cases}$$

(18.5)

For $0 < \mu < 1$, $\sigma > 0$ and $0 < \nu < 1$ where $Y' \sim BB(n, \mu, \sigma)$. The mean and variance of $Y$ are given by

$$E(Y) = (1 - \nu) n\mu$$

and

$$Var(Y) = (1 - \nu) n\mu (1 - \mu) \left[ 1 + \frac{\sigma}{1 + \sigma} (n - 1) \right] + \nu (1 - \nu) n^2 \mu^2$$

respectively.
Bibliography


A. Azzalini. Further results on a class of distributions which includes the normal ones. Statistica, 46:199:208, 1986.


