Distributions for Modelling Location, Scale and Shape: Using GAMLSS in R

Robert Rigby, Mikis Stasinopoulos, Gillian Heller and Fernanda De Bastiani

November 14, 2017

Contents

List of Figures

List of Tables

Preface

This is a book about statistical distributions and how they can be used in practical applications. It describes over 100 distributions available in the **R** package **gamlss.dist**, their properties, limitations and how they can be used in data applications.

Historically the distributions commonly used for modelling continuous and discrete count response variables were the normal and Poisson distributions respectively. Unfortunately, especially with the larger data sets often obtained these days, these distributions are often found to be inappropriate or to provide inadequate fits to the observed response variable distribution in many practical situations.

In practice, for a continuous response variable, the shape of its distribution can be highly positively or negatively skewed and/or highly platykurtic or leptokurtic (i.e. lighter or heavier tails than the normal distribution). Continuous variables can also have different ranges from that of the normal distribution, i.e. $(-\infty, \infty)$. The ranges $(0, \infty)$ and $(0, 1)$ are especially common.

Also in practice a discrete count response variable can have a distribution which is overdispersed or underdispersed (relative to the Poisson distribution), and/or highly positively skewed and/or leptokurtic and/or have an excess or reduced incidence of zero values.

There are also occasions where a response variable cannot be modelled with a continuous or discrete distribution but requires a mixture of a continuous and a discrete distribution, which is a mixed distribution, i.e. a continuous distribution with additional specific values with point probabilities. For example, in insurance claims, the claim amount can be modelled with a continuous distribution on $(0, \infty)$ if a claim is made and with a point probability at 0, if no claim is made. In this case the distribution should allow values in range $[0, \infty)$ which includes value 0. Another important example is a continuous distribution on $(0, 1)$ with additional point probabilities at 0 and 1, giving range $[0, 1]$ which includes 0 and 1.

This book is divided into the following parts.

- Parametric distributions
- The GAMLSS family of distributions
- Inference
- Advanced topics
- Distributions reference guide

The aims of the "Parametric distributions" part are:

- to introduce the basic idea of a distribution

- to introduce the properties of distributions.

The aims of the "The GAMLSS family of distributions" part are to introduce distributions implemented in the gamlss **R** packages and in particular:

- to investigate continuous distributions on ranges $(-\infty, \infty)$, $(0, \infty)$ and $(0, 1)$, including highly positively or negatively skewed and/or highly platykurtic or leptokurtic (i.e. lighter or heavier tails than the normal distribution),
- to investigate discrete distributions for counts, including overdispersed or underdispersed and/or highly positively skewed and/or leptokurtic and/or an excess or reduced incidence of zero values,
- to investigate mixed distributions with range $[0, \infty)$ including value 0, or range $[0, 1), (0, 1]$ or $[0, 1]$ i.e. the interval from 0 to 1 including value 0 or 1 or both,
- to investigate finite mixture distributions.

The aims of the "Inference" part are:

- to define the basic concepts of infrerence,
- to explain likelihood function,
- to explain the likelihood based statistical inference for distribution parameters.

The aims of the "Advanced topics" part are:

- to describe methods for generating new distributions,
- to investigate the heaviness of tails of distributions,
- to compare continuous distributions based on their moment and centile skewness and kurtosis.

The aim of the "Distributions reference guide" part of this book is to provide a reference guide to the continuous, discrete and mixed distributions available in the **gamlss.dist** package. In particular for each distribtion available in the package **gamlss.dist**, this provides:

- the range of the response variable
- the range of the parameters of the distribution
- the probability (density) function,
- the median and the mode
- the moments and hence the mean, variance, skewness and excess kurtosis
- the moment or probability generating function
- the cumulative distribution function
- the inverse cumulative distribution function.

Motivation

A theoretical distribution for a response variable is a representation (usually an approximation) of its population distribution. The theoretical distribution may be parametric (i.e. it has unknown parameters). A theoretical parametric distribution generally provides a simple parsimonious (and usually smooth) representation of the population distribution. It can be used for inference (e.g. estimation) for the centiles (or quantiles) and moments (e.g. mean) of the population distribution and other population measures. A better theoretical distribution model for a population distribution (of the responses variable) should provide better inference for population centiles and moments. Note however inference (e.g. estimation) for the moments can be particularly sensitive to misspecification of the theoretical distribution and especially to misspecification of the heaviness of the tail(s) of the population distribution.

In a regression situation a theoretical parametric conditional distribution for a response variable (given the values of the explanatory variables) provides a simple parsimonious representation (usually an approximation) of the corresponding conditional population distribution. This allows a potentially simple interpretation of how changing values of the explanatory variables affects the conditional distribution of the response variable. A better theoretical conditional distribution should provide better inference (e.g. estimation) for the centiles (or quantiles) and moments (e.g. mean) of the population conditional distribution and potentially better inference regarding the relationship of the response variable population conditional distribution to explanatory variables. Furthermore a flexible theoretical conditional distribution with say four distribution parameters potentially allows modelling of changes in the location, scale, skewness and kurtosis of the population conditional distribution of the response variable with explanatory variables.

An alternative approach to (conditional) population centile (or quantile) estimation using a theoretical parametric (conditional) distribution is quantile regression (Koenker, 2017). The advantages and disadvantages of the two approaches are discussed in Rigby et al. (2013) and Stasinopoulos et al. (2017), p451-452.

An alternative approach to estimation of the (conditional) population mean (and variance) using a theoretical parametric (conditional) distribution is generalized estimating equations (GEE), Ziegler [2011].

Part I

Parametric distributions

Chapter 1

Types of distributions

This chapter provides:

1. the definition of a statistical distribution
2. the implementation of distributions in **R**

1.1 Introduction

In this Chapter we explore the different types of theoretical parametric distributions. In Chapter 2 the properties and summary measures which can be derived for parametric distributions are considered.

Statistical modelling is the art of building parsimonious statistical models for a better understanding of the response variables of interest. A statistical model deals with uncertainties and therefore it contains a *stochastic* or *random* part, that is, a component which describes the uncertain nature of the response variable. The main instrument to deal with uncertainty in mathematics is the concept of *probability* and all statistical models incorporate some probabilistic assumptions by assuming that response variables of interest have a probabilistic distribution. This probabilistic or 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 is a simplification of the population and its behaviour. Essential to understanding the basic concepts of statistical modelling are the ideas of

- the *population*,
- the *sample* and
- the *model*

Those concepts are explained in detail in Chapter 9.

An *'experiment'*, in the statistical sense, is any action which has an outcome that cannot be predicted with certainty. For example, "toss a die and" record the score' is an experiment with possible outcomes $S = \{1, 2, 3, 4, 5, 6\}$. The set of all possible outcomes S is called the sample space of the experiment. Outcomes of the experiment are grouped into *events* (denoted as E). Events that correspond to a single outcome are called elementary events, while those corresponding to more than one outcome are called composite events. For example, the event "even score" corresponds to the outcomes $E = \{2, 4, 6\}$ and is a composite event. Probabilities, denoted here as $P()$, are functions defined on the events and have the properties:

- $P(E) \geqslant 0$
- $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ where E_1 and E_2 are events which have no common outcomes (i.e. are disjoint)
- $P(S) = 1$.

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 assigns a value to an outcame determined by chance. For example, let us consider the simple action of tossing a coin and observing whether the result is a head or a tail. If we assign $Y = 0$ for a tail and $Y = 1$ for a head, then the random variable Y can take values in the set $\{0, 1\}$. We say that Y has *range* $R_Y = \{0, 1\}$.

A *continuous* random variable Y has a range that is an interval of the real line. The most common continuous ranges R_Y are:

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

A *discrete* random variable Y has a range that is a distinct set of values. The number of values may be either finite or countably infinite. The most common discrete ranges R_Y are:

- *the binary values* : $\{0, 1\}$
- *the binomial values* : $\{0, 1, 2, \ldots, n\}$
- *the non-negative integer values* : $\{0, 1, 2, \ldots, \infty\}$.

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

1.1.1 Probability (density) function

The probability distribution of a discrete random variable Y is defined by the probabilities for all possible values of Y, and is called a probability function (pf). The probability function $P()$ assigns the probability that Y takes each particular value in its range. For example, in the coin toss, $P(Y = 0) = P(Y = 1) = 0.5$ if the coin is fair. A continuous distribution is defined by a probability density function (pdf).

Discrete random variables

For a discrete random variable Y, let $f(y) = P(Y = y)$, that is $f(y)$[1] represents the probability that Y is equal to specific value y. The function $f(y)$ is said to be a proper probability function of the discrete random variable Y if $f(y)$ is positive for all values of y within R_Y and if

$$\sum_{y \in R_Y} f(y) = \sum_{y \in R_Y} P(Y = y) = 1 \ .$$

For convenience, let $f(y) = P(Y = y) = 0$ for all y that are in the real line $\mathbb{R}$, but are not in R_Y.

Continuous random variables

Let Y be a continuous random variable. The function $f(y)$ is said to be a proper probability density function of the continuous variable Y if $f(y)$ is positive for all values of y within R_Y and if

$$\int_{R_Y} f(y) \, dy = 1 \ .$$

For convenience let $f(y) = 0$ for all y in the real line $\mathbb{R}$ not in R_Y. Then $f(y)$ is defined on $y \in \mathbb{R}$. Probabilities are given by areas under $f(y)$:

$$P(a \leq Y \leq b) = \int_a^b f(y) \, dy \ .$$

Note here the peculiarity that $P(Y = a) = \int_a^a f(y) \, dy = 0$, for any arbitrary value a. That is, the probability of a continuous random variable Y being

[1] There are two notational comments needed to be made here: (i) the notation $P(Y = y)$ is preferable to $f(y)$ for discrete random variables, since it emphasises the discreteness of the variable involved. In this book we will use both notations since it is sometimes convenient to use the same notation for discrete and continuous random variables; (ii) It is common practice to use the notation $f_Y(y)$ to emphasise that the f function refers to the random variable Y. We generally drop Y from the notation $f_Y(y)$ in this book for simplicity, but we use it when needed for clarification, as for example when more than one random variable is involved.

exactly equal to any specific value is equal to zero. (This can be circumvented by defining the probability on a small interval $(a-\Delta y, a+\Delta y)$ around a, where Δy has a small value. Then $P(Y \in (a-\Delta y, a+\Delta y)) = \int_{a-\Delta y}^{a+\Delta y} f(y)\ dy$ is properly defined.)

Example 1: Exponential distribution

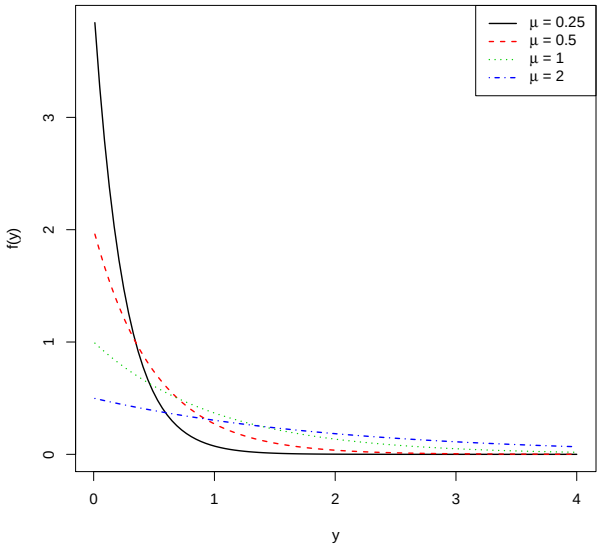


Figure 1.1: The exponential pdf with different values of μ

The function

$$f(y) = e^{-y} \qquad \text{for } y > 0$$

is a proper probability density function (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.

$$f(y) = \theta e^{-\theta y} \qquad \text{for } y > 0\ ,$$

where $\theta > 0$. Since $\int_0^\infty \theta e^{-\theta y} dy = 1$, the new function is still a pdf. Because $f(y)$ now contains the parameter θ, it is called a *parametric probability density function*. The key advantage of introducing the parameter θ is that by making the shape of the distribution more flexible it can be used for modelling a set of of historical observations. Finding a suitable estimated value for parameter θ for a given set of data is an example of *statistical inference*.

In order to emphasise the fact that the pdf depends on θ we write it as $f(y\,|\,\theta)$[2]. The notation $f(y\,|\,\theta)$ now represents, for different values of θ, a *family* of pdf's. Note that if instead of parameter θ we choose any other one-to-one function of θ, e.g. $\mu = 1/\theta$, the pdf family remains the same. For example

$$f(y\,|\,\theta) \quad = \quad \theta e^{-\theta y}\ , \qquad y > 0 \tag{1.1}$$

for $\theta > 0$, and

$$f(y\,|\,\mu) \quad = \quad \frac{1}{\mu} e^{-y/\mu}\ , \qquad y > 0 \tag{1.2}$$

for $\mu > 0$, define the same pdf family (the *exponential distribution*). Figure 1.1 shows $f(y\,|\,\mu)$ plotted against y, for each of four different values of the parameter μ, i.e. $\mu = (0.25, 0.5, 1, 2)$. As $y \to 0$, $f(y\,|\,\mu) \to \frac{1}{\mu}$ which decreases as μ increases. Note that (1.1) and (1.2) are different parametrizations of the same family. We shall see later that some parametrizations are preferable (as far as the interpretation of the model is concerned) than others in practice.

Next we give an example of using the exponential distribution in practice.

R data file: aircond in package **panel** of dimensions 24×1

source: Proschan [1963]

variables

`aircond` : the intervals, in service-hours, between failures of the air-conditioning equipment in a Boeing 720 aircraft.

The following data reported by Proschan [1963], provide observations of the interval, Y, 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 `rpanel` package, Bowman *et al,* 2007, and use only 24 observation from one of the aircraft:

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

A histogram of the data is shown in Figure 1.2(a). All observations are positive so we require a distribution defined on the positive line like the exponential distribution. In the **R** commands below we input the data, plot a histogram of the data and then fit an exponential distribution to the data using the **gamlss** package function `histDist()`. The exponential distribution is specified by the argument `family=EXP`. Note that the fitting [i.e. estimation of parameter μ in (1.2)] is performed using maximum likelihood estimation.

Figure 1.2

[2]Note that for simplicity in general we will drop the conditioning on θ unless the condition is important, for example in statistical inference.

```
aircond <- c(50, 44, 102,  72,  22,  39,  3, 15, 197, 188, 79,
             88, 46, 5,5, 36,  22, 139, 210, 97, 30, 23, 13, 14)
truehist(aircond, nbins=10, main="(a)")
m1 <- histDist(aircond, family=EXP, main="(b)")
fitted(m1, "mu")[1]
```

```
## [1] 64.125
```

R code on page 29

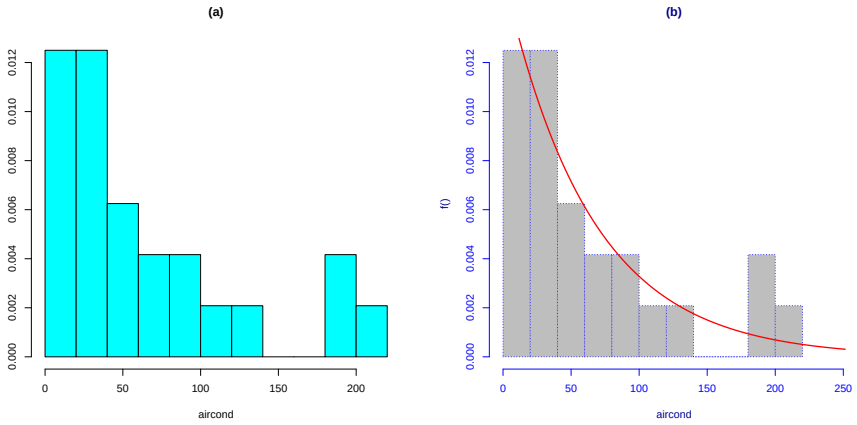


Figure 1.2: Showing (a) a histogram of the `aircond` data and (b) a histogram together with the fitted exponential distribution to the `aircond` data.

The GAMLSS implementation of the exponential distribution uses parametrization (1.2) of the exponential distribution and the fitted μ is given by the `fitted(m1, "mu")` command as $\hat{\mu} = 64.125$. The beauty of having fitted a theoretical distributions to the data is the fact that now we can calculate probabilities concerning Y, the interval in service-hours, between failures of the air-conditioning equipment.

Example 2: Poisson distribution

The function

$$P(Y = y \,|\, \mu) = \frac{e^{-\mu}\mu^y}{y!}\,, \qquad \text{for } y = 0, 1, 2, \ldots \tag{1.3}$$

is the probability function of one of the most popular discrete distributions, the *Poisson distribution*. Figure 1.3 plots $P(Y = y \,|\, \mu)$ against y for each of the values $\mu = (1, 5, 10)$, showing how the distribution shape changes with μ.

Next we give an example of using the Poisson distribution in practice.

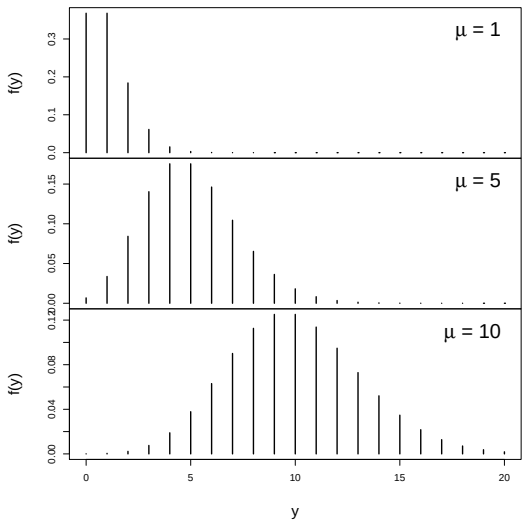


Figure 1.3: The Poisson probability function with different values of μ

R data file: prussian in package **pscl** of dimensions 280×3

source: Bortkiewicz 1898

variables

`y` : count of deaths

`year` : the year of observation

`corp` : corp of Prussian Army

In 1898 the Russian economist Ladislaus Bortkiewicz published a book with the title 'Das Gesetz der keinem Zahlen'. In the book he included a data example that became famous for illustrating the use of the Poisson distribution. The response variable here is the number of deaths by kicks of a horse, Y, in different years and at different cops (regiments) of the Prussian Army. Here we use the data with no explanatory variables and fit a Poisson distribution to the number of deaths. For the fitting we use the `histDist()` function. The Poisson distribution is spsecified by the argument `family=PO` which uses parametrization (1.3).

```
library(pscl)
library(gamlss)
h1 <- histDist(y,  family=PO, data=prussian)
```

Figure 1.4

```
fitted(h1)[1]
```

```
## [1] 0.7
```

R code on page 31

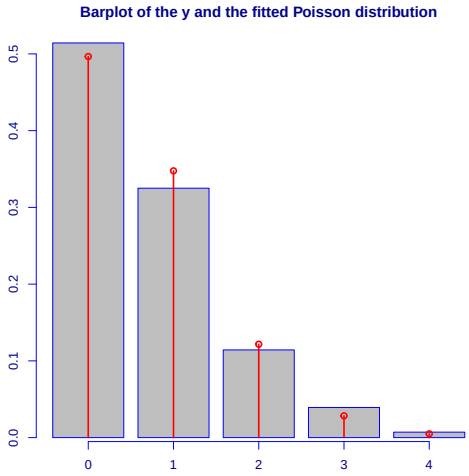


Figure 1.4: Showing a histogram together with the fitted Poisson distribution (vertical lines) for the `prussian` data.

The estimated parameter here is given by $\hat{\mu} = 0.700$ which in the case of the Poisson distribution is also the mean of the theoretical distribution. Note that no actual value of Y larger than 4 occurs, but because the Poisson distribution is defined on $0, 1, 2, \ldots$ we are able to estimate probabilities at values larger than 4. e.g. the probability that Y is equal to 5 is given by

```
dPO(5, mu=fitted(h1,"mu")[1])
```

```
## [1] 0.0006955091
```

a very small probability.

Summarising

- Any non-negative function $f(y) = P(Y = y)$ which sums over a discrete sample space to one can be a probability function (pf).
- Any non-negative function $f(y)$ which integrates over a continuous sample space to one can be a probability density function (pdf).
- Probability (density) functions may depend on more than one parameter. In general, we write a probability (density) function depending on

k parameters $\theta_1, \ldots, \theta_k$ as $f(y \,|\, \theta_1, \ldots, \theta_k)$, or $f(y \,|\, \boldsymbol{\theta})$. In the GAMLSS family of distributions, k can be 1, 2, 3 or 4, and parameters are denoted as $\boldsymbol{\theta}^\top = (\mu, \sigma, \nu, \tau)$.

- A one-to-one reparametrization of the parameters does not change the pf or pdf family.
- Parameters affect the shape of the distribution.

1.1.2 Cumulative distribution function

The *cumulative distribution function* (*cdf*) $F(y)$ is the probability of observing a value less than or equal to a specified value y. It is defined as

$$F(y) = P(Y \leq y) = \begin{cases} \sum_{w \leq y} f(w) = \sum_{w \leq y} P(Y = w) & \text{for discrete } Y \\ \int_{-\infty}^{y} f(w)\, dw & \text{for continuous } Y \ , \end{cases}$$

for $y \in \mathbb{R}$, where in the discrete case the sum is over all $w \in R_Y$ for which $w \leq y$.

The cdf is a nondecreasing function, with $\lim_{y \to -\infty} F(y) = 0$ and $\lim_{y \to \infty} F(y) = 1$. For a continuous random variable Y, $F(y)$ is a continuous function, often S-shaped; for discrete Y, $F(y)$ is a step function, with jumps at the values $y \in R_Y$.

Example 1: Exponential distribution

For the exponential distribution [using parametrization (1.2)], we have for $y > 0$,

$$F(y) = P(Y \leq y) = \int_0^y \frac{1}{\mu} e^{-w/\mu}\, dw = 1 - e^{-y/\mu} \ ,$$

giving the cdf

$$F(y) = \begin{cases} 0 \ , & y \leq 0 \\ 1 - e^{-y/\mu} \ , & y > 0 \ . \end{cases} \tag{1.4}$$

Figure 1.5 plot $f(y)$ and $F(y)$ given by (1.2) and (1.4) respectively, against y, for $\mu = 0.5$, demostrating the link between $f(y)$ and $F(y)$, e.g. $F(0.7) = P(Y \leq 0.7) = 0.75$.

The next question arising here is how the empirical (i.e. sample) cumulative distribution function compares with the fitted cdf. In general the fitted cdf is a smoother function than the empirical distribution function. We can see this in Figure 1.6 where the empirical distribution function of the `aircond` data

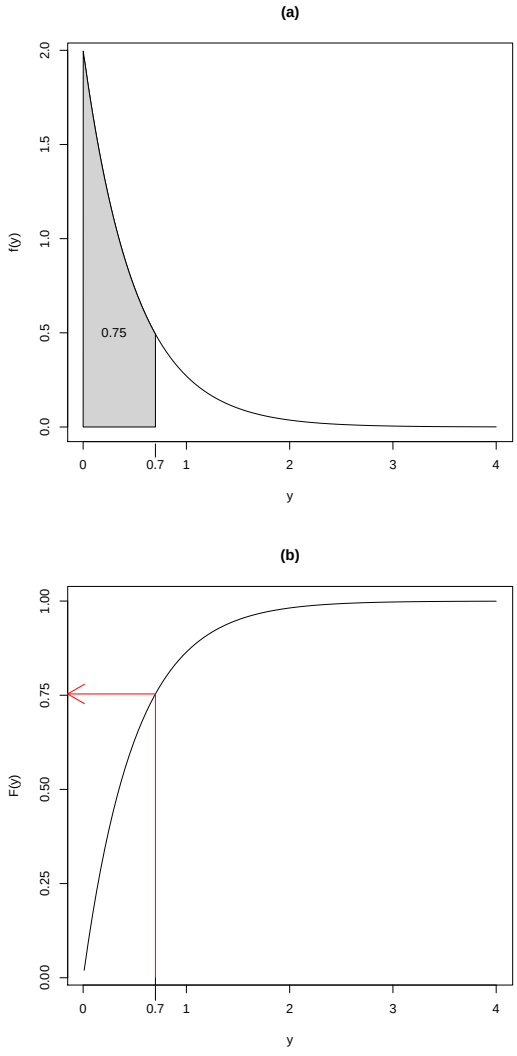
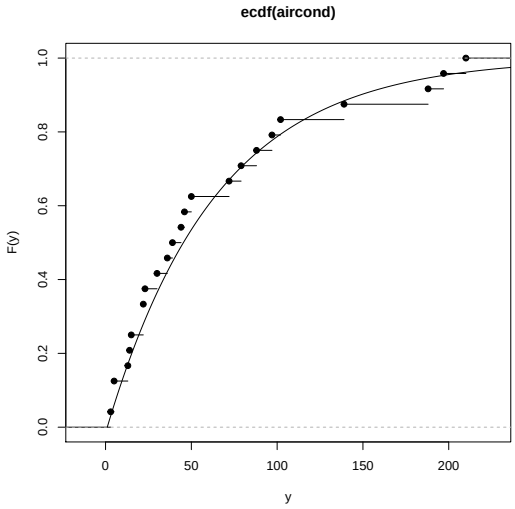


Figure 1.5: (a) The probability density function and (b) the cumulative distribution function, of the exponential distribution with $\mu = 0.5$.

is plotted together with the fitted cdf of the exponential distribution (i.e with $\hat{\mu} = 64.125$). Note that this empirical cdf is a step function with a step of height f/n (i.e. $f/24$) where f is the frequency at each distinct value of Y and n is the total number of observations.

```
Ecdf <- ecdf(aircond) # the ecdf
plot(Ecdf, ylab="F(y)", xlab="y")            # plot it
lines(pEXP(0:250, mu=64)) # plot the fitted cdf
```

Figure 1.6



R code on page 35

Figure 1.6: The fitted cumulative density function of the exponential distribution with $\hat{\mu} = 64.125$ (the continuous line), and the empirical (i.e. sample) cumululative function (the step function)

Example 2: Poisson distribution

The Poisson cdf is

$$F(y) = P(Y \leq y) = \begin{cases} 0 \ , & y < 0 \\ \displaystyle\sum_{w=0}^{\lfloor y \rfloor} \frac{e^{-\mu}\mu^w}{w!} \ , & y \geq 0 \ , \end{cases} \tag{1.5}$$

where $\lfloor y \rfloor$ denotes the largest integer less than or equal to y (the floor function). Figure 1.7 plots $P(Y = y)$ and $F(y)$, given by (1.3) and (1.5) respectively, against y, for $\mu = 2$, and shows

$$F(1.5) = P(Y \leq 1.5) = P(Y = 0) + P(Y = 1) = 0.406 \ .$$

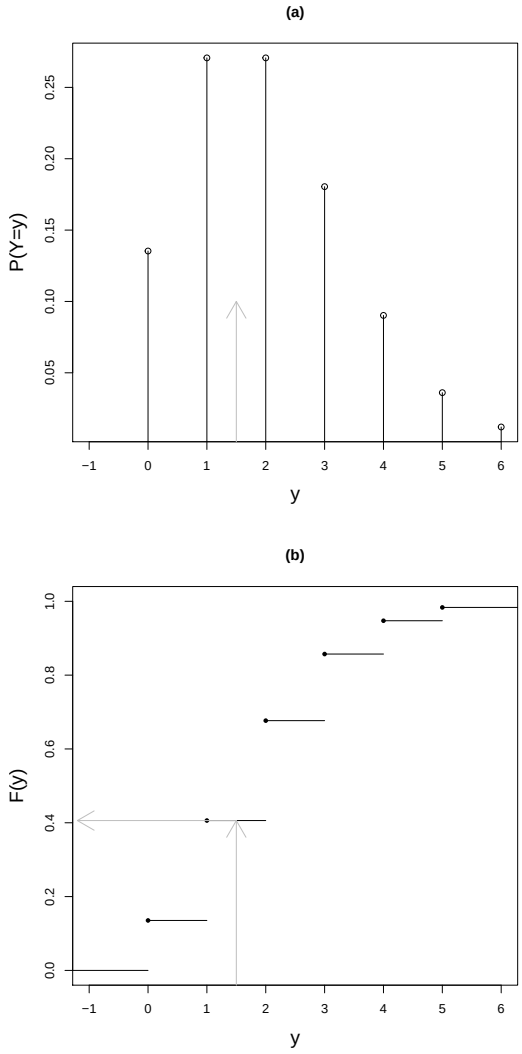
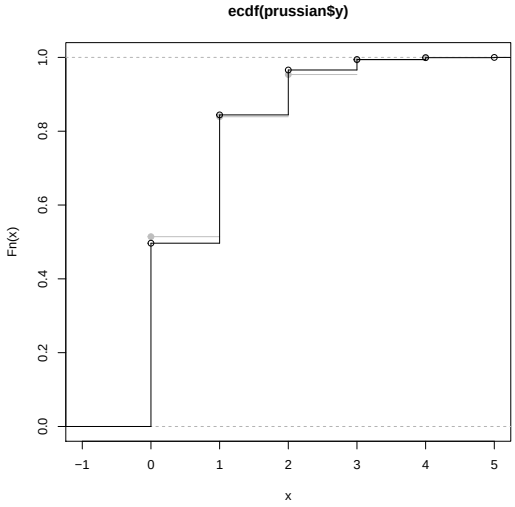


Figure 1.7: (a) The probability function of the Poisson distribution ($\mu = 2$), (b) the equivalend cumulative distribution function $P(Y = y)$ showing $F(1.5) = 0.406$. Note that while the probablility of observing 1.5 is zero, i.e. $P(Y = 1.5) = 0$, the cumulatitve distribution $F(1.5) = 0.406$ is well defined.

It will be of interest to compare the empirical distribution function of the `prussian` data set with the fitted cumulative Poisson function. Figure 1.8 shows the two cumulative functions which have very similar values indicating that the Poisson distribution model fits adequately the data. One advantage of using the fitted Poisson distribution is that we can calculate probabilities for values of Y not occuring the actual data.

```
Ecdf <- ecdf(prussian$y) # the ecdf
plot(Ecdf, col="gray")   # plot ecdf
cdf <- stepfun(0:6, c(0,pPO(0:6, mu=0.7)), f = 0)
plot(cdf, add=T)         # plot fitted Poisson
```

Figure 1.8



R code on page 37

Figure 1.8: The empirical cumulative distribution function (cdf) and fitted cdf for the Poisson distribution (with $\mu = 2$).

1.1.3 Inverse cumulative distribution function

For a continuous random variable Y, the *inverse* of the cumulative distribution function is $y_p = F^{-1}(p)$ for $0 < p < 1$. Hence y_p is defined by $Pr(Y \leq y_p) = p$, i.e. $F_Y(y_p) = p$.

For a discrete random variable Y, $y_p = F^{-1}(p)$ 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 $Pr(Y \leq y_p) \geq p$.

The inverse cumulative distribution function is very important for calculating the centile (or quantile) values of a variable, see Section 2.3.1.

1.1.4 Survival and hazard functions

The *survival function* $S(y)$ is the probability of Y 'surviving' beyond y, i.e.

$$S(y) = Pr(Y > y) = 1 - Pr(Y \leq y) = 1 - F(y) \ .$$

Alternative names for the survival function are the *complementary cumulative distribution function* (ccdf), the *tail distribution*, *exceedance*, or the *reliability function* (common in engineering).

In **R** the survival function can be obtained from the cumulative distribution function (see the `p` function in Section 1.2) by using the argument `lower.tail = FALSE`. for example

```
curve(pEXP(x, mu=1, lower.tail = TRUE), 0,10)
```

will plot the cdf of the exponential distribution, while

```
curve(pEXP(x, mu=1, lower.tail = FALSE), 0,10)
```

will plot the survival function.

For a continuous random variable Y, the *hazard function* $h(y)$ is the instantaneous likelihood of ('dying at') y given survival up to y, i.e.

$$h(y) = \lim_{\delta y \to 0} \frac{p\,(y < Y \leq y + \delta y \,|\, Y > y)}{\delta y} = \frac{f(y)}{S(y)} \ .$$

Both the survival function and the hazard function play an important role in medical statistics and in reliability theory.

The package `gamlss.dist` in **R** provides the facility of taking any GAMLSS family distribution, see Chapter 3, and creating its hazard function using the functions `gen.hazard()` or `hazardFun()`. For expample to generate the hazard function for the exponentian distribution use:

```
hEXP<-hazardFun(family = "EXP")
```

The hazard function of the exponential distribution has the characteristic that it is flat. That is, the instantaneous likelihood of 'dying' in the exponential distribution is constant. This can be verified by plotting the created hazard function `hEXP()`.

```
curve(hEXP(x, mu=0.7), 0,3)
```

Continuous distributions

R name	Distribution
`beta`	beta
`cauchy`	Cauchy
`chisq`	chi-squared
`exp`	exponential
`f`	F
`gamma`	gamma
`lnorm`	log-normal
`norm`	normal
`t`	Student's t
`unif`	uniform
`weibull`	Weibull

Discrete distributions

R name	Distribution
`binom`	binomial
`geom`	geometric
`hyper`	hypergeometric
`multinom`	multinomial
`nbinom`	negative binomial
`pois`	Poisson

Table 1.1: Standard distributions in R

1.2 Distributions in R: the d, p, q and r functions

The R statistical environment contains several popular distributions. R uses the `d`, `p`, `q` and `r` convention where

`d` is the probability (density) function;

`p` is the cumulative distribution function;

`q` is the inverse cumulative distribution function; and

`r` generates random numbers from the distribution.

For example `dnorm()`, `pnorm()` `qnorm()` and `rnorm()` are the pdf, cdf, inverse cdf and random number generating function of the normal distribution, respectively. Table 1.1 gives the distributions which are standard in R.

This book is about distributions in the `gamlss.dist` package. This package follows the `d`, `p`, `q` and `r` convention, but parametrizations can be different from those used in the standard R distributions. The main reason is that the `gamlss.dist` package is built to support the GAMLSS regression framework. In a regression model, interpretability of the distribution parameters, which are being modelled as functions of covariates, is important. For example, the gamma distribution in **gamlss.dist** is denoted $\texttt{GA}(\mu, \sigma)$ where μ is the mean of the distribution, while σ is the coefficient of variation. The gamma distribution in R is denoted here as $\texttt{gamma}(a, s)$, where a is the 'shape' and s is the 'scale', with mean as and coefficient of variation $a^{-0.5}$.

Graphical representation of the `d`, `p`, `q` and `r` functions in R can be accomplished easily. The following R code produces Figure 1.9, a plot of the pdf, cdf, inverse cdf and the histogram of a randomly generated sample from a gamma, $\texttt{GA}(\mu, \sigma)$,

distribution using the **gamlss.dist** package functions `dGA()`, `pGA()`, `qGA()` and `rGA()` respectively.

Figure 1.9

```
mu=3
sigma=.5
curve(dGA(y, mu, sigma), 0.01, 10, xname="y",ylab="f(y)",
      cex.lab=2) # pdf
curve(pGA(y, mu, sigma), 0.01, 10, xname="y", ylab="F(y)",
      cex.lab=2)  # cdf
curve(qGA(y, mu, sigma), 0.01, 1, xname="y", ylab="F(y)",
      cex.lab=2)  # cdf
y<-rGA(1000, mu, sigma) # random sample
hist(y,col="lightgray",main="", cex.lab=2)
```

Similarly, to create Figure 1.10, an example of a discrete distribution (the negative binomial), use using the **gamlss.dist** package functions `dNBI()`, `pNBI()`, `qNBI()` and `rNBI()` respectively:

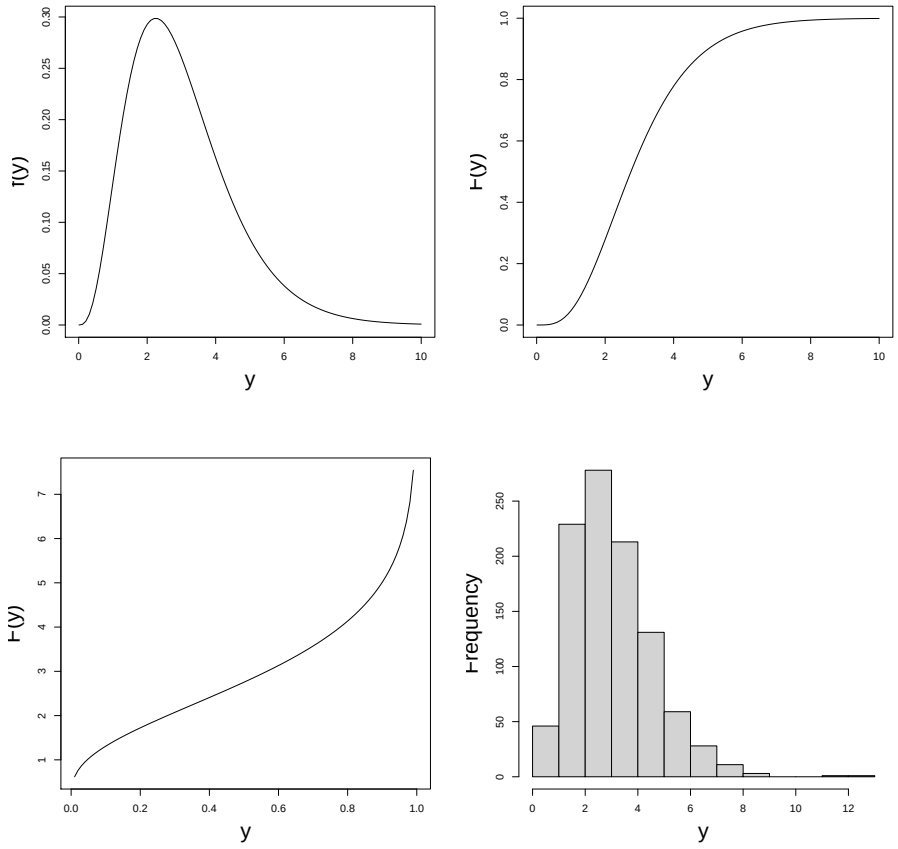
Figure 1.10

```
mu=8
sigma=.25
plot(function(y) dNBI(y, mu, sigma), from=0, to=30, n=30+1,
     type="h",xlab="y",ylab="f(y)",cex.lab=2)
cdf <- stepfun(0:29, c(0,pNBI(0:29,mu, sigma)), f = 0)
plot(cdf, xlab="y", ylab="F(y)", verticals=FALSE,
     cex.points=.8, pch=16, main="",cex.lab=2)
invcdf <- stepfun(seq(0.01,.99,length=39),
                  qNBI(seq(0.1,.99,length=40),mu, sigma), f = 0)
plot(invcdf, ylab="invcdf(x)", do.points=FALSE,verticals=TRUE,
     cex.points=.8, pch=16, main="",cex.lab=2  )
Ni <- rNBI(1000, mu, sigma)
hist(Ni,breaks=seq(min(Ni)-0.5,max(Ni)+0.5,by=1),col="lightgray",
     main="",cex.lab=2)
```

In Chapter 2 we discribe the properties of distributions. Chapter 3 describes all GAMLSS family distributions available in the **gamlss.dist** package.

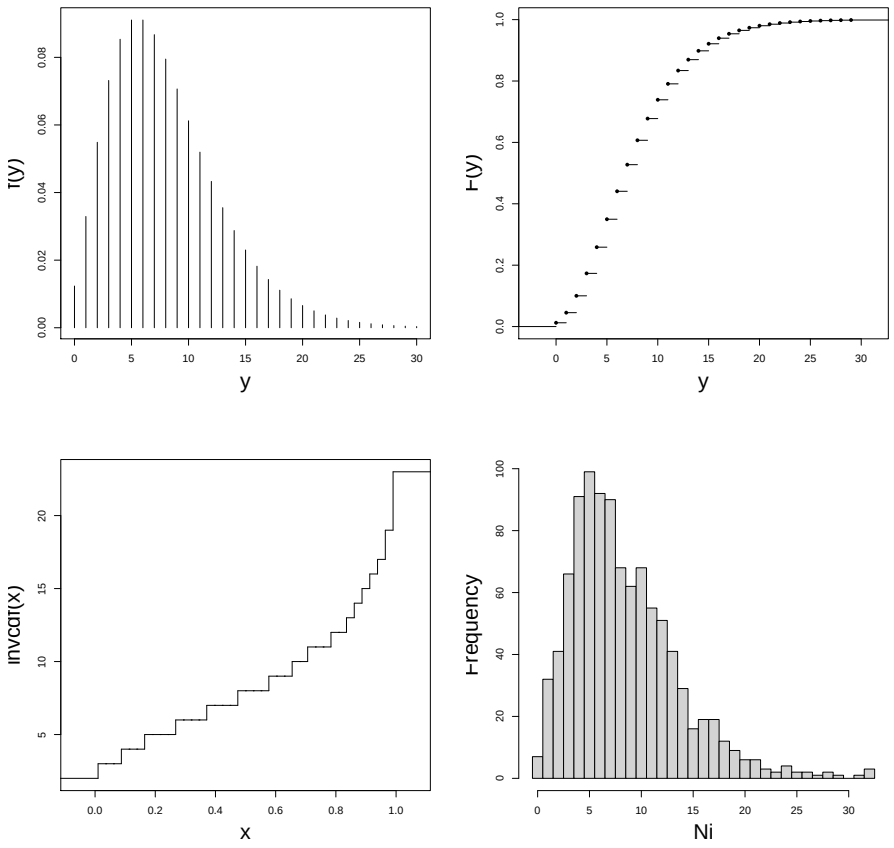
1.3 Bibliographic Notes

1.4 Exercises



R code on page 40

Figure 1.9: Probability function (pdf), cumulative distribution function (cdf), inverse cdf and histogram of a random sample of 1000 observations from a gamma distribution.



R code on page 40

Figure 1.10: Probability function, cumulative distribution function (cdf), inverse cdf and histogram of a random sample of 1000 observations from a negative binomial distribution

Chapter 2

Properties of distributions

This chapter provides an introduction to properties of the distributions:

1. moment based properties
2. moment, cumulant and probabiliy generating functions
3. centile based properties

2.1 Introduction

In this Chapter we are concerned with some general properties of the distributions including measures of location, scale, skewness and kurtosis. For parametric distributions it is common that these measures are related to some or all of the parameters.

It is assumed that the response variable Y comes from a population distribution which can be modelled by a theoretical probability (density) function $f(y \mid \boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta}$ can be any number of parameters, but in practice is usually limited to up to four, denoted as $\boldsymbol{\theta}^\top = (\mu, \sigma, \nu, \tau)$. Limiting $\boldsymbol{\theta}$ to four dimensions is not usually a serious restriction. The parameters μ, σ, ν and τ 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*, the average of Y,
- the *median*, the value of Y which cuts the distribution in two halves with probablilty at each size 0.50, and

- the *mode*, the value of Y which has the highest value in the probability (density) function.

Mikis: needs more explanation about the mean

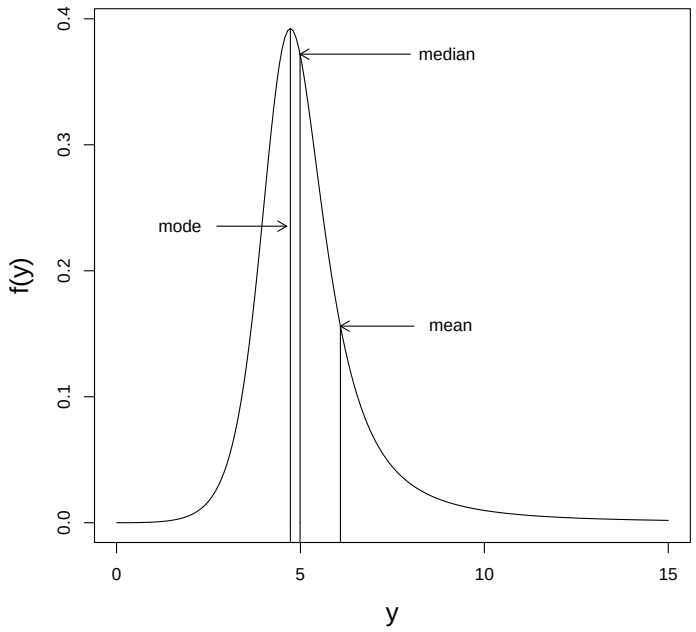


Figure 2.1: Showing the difference between the mean, median and mode of a positively skewed distribution

For symmetric distributions such as the normal distribution these three measures are identical, but this is not the case for a non-symmetrical distribution, as Figure 2.1 shows.

A *scale* parameter usually represents or is related to the 'spread' of the distribution. Occasionally it is the standard deviation or maybe the coefficient of variation of the distribution.

The skewness is a measure of asymmetry of the distribution of the variable. In general, a distribution with a heavier tail to the right than the left has positive skewness, while one with a heavier tail to the left has negative skewness, and a symmetric distribution has zero skewness. An example of a distribution that can be symmetric, positively or negatively skewed, is the skew-normal Type 2 distribution (SN2, Section **??**), shown in Figure 2.2(a).

Kurtosis is a measure of heavy tails. A distribution with heavy (i.e. fat) tails (relative to a normal distribution) will generally have high kurtosis (*leptokurtosis*), while a distribution with short (i.e. thin) tails will generally have low

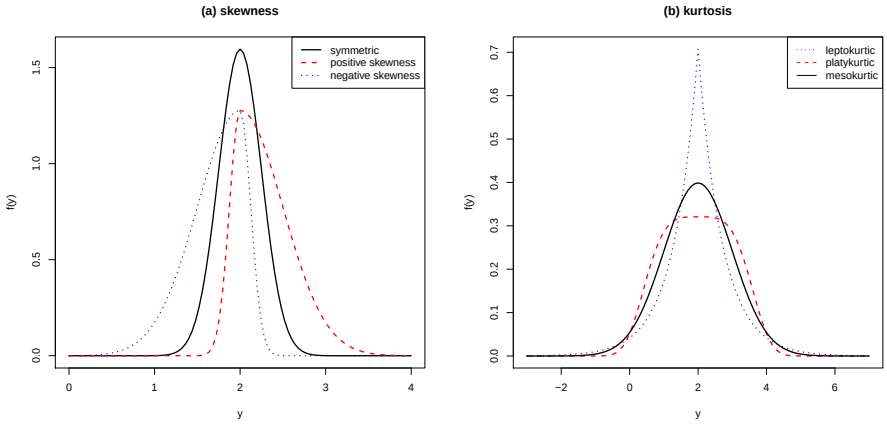


Figure 2.2: Showing (a) Skew-normal Type 2 distribution: symmetric, positively and negatively skewed (b) Power exponential distribution: leptokurtic, mesokurtic and platykurtic

kurtosis (*platykurtosis*). Note that leptokurtic and platykurtic distributions are judged by comparison to the normal distribution, which is called *mesokurtic*. Figure 2.2(b). shows an example of a distribution which can be either leptokurtic, mesokurtic or platykurtic. This is the power exponential (PE), see Section 14.3.2. Skewness and kurtosis measures can be either moment or centile based, see Sections 2.2 and 2.3 respectively.

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

In this section we omit conditioning on the parameter vector $\boldsymbol{\theta}$, for simplicity of notation. So, for example, we use $f(y)$ throughout, which has the meaning $f(y \mid \boldsymbol{\theta})$. The mean, variance and other distributional properties are also conditional on $\boldsymbol{\theta}$, but this conditioning is not explicitly denoted.

Mean or expected value

The population mean or expected value of Y is denoted by $\mathrm{E}(Y)$ and is given by

$$\mathrm{E}(Y) = \begin{cases} \int_{-\infty}^{\infty} y f(y)\, dy & \text{for a continuous random variable } Y \\ \sum_{y \in R_Y} y P(Y = y) & \text{for a discrete random variable } Y \ . \end{cases}$$

More generally the mean of a function $g(y)$ is given by

$$\mathrm{E}\left[g(Y)\right] = \begin{cases} \int_{-\infty}^{\infty} g(y) f(y)\, dy & \text{for continuous } Y \\ \sum_{y \epsilon R_Y} g(y) P(Y = y) & \text{for discrete } Y \ . \end{cases}$$

Properties of expectations

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

- $\mathrm{E}(a) = a$
- $\mathrm{E}(aY + b) = a\mathrm{E}(Y) + b$
- $\mathrm{E}(aY_1 + bY_2) = a\mathrm{E}(Y_1) + b\mathrm{E}(Y_2)$.

Variance and standard deviation

The population variance of Y is defined as

$$\mathrm{Var}(Y) = \begin{cases} \int_{-\infty}^{\infty} \left[y - \mathrm{E}(Y)\right]^2 f(y)\, dy & \text{for continuous } Y \\ \sum_{y \in R_Y} \left[y - \mathrm{E}(Y)\right]^2 P(Y = y) & \text{for discrete } Y \ . \end{cases}$$

The standard deviation of Y is $\mathrm{SD}(Y) = \sqrt{\mathrm{Var}(Y)}$. Both variance and standard deviation give a measure of the spread of the distribution about the mean.

Properties of variances

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

- $\mathrm{Var}(Y) = \mathrm{E}\left\{\left[Y - \mathrm{E}(Y)\right]^2\right\} = \mathrm{E}\left(Y^2\right) - \left[\mathrm{E}\left(Y\right)\right]^2$
- $\mathrm{Var}(a) = 0$
- $\mathrm{Var}(aY + b) = a^2\mathrm{Var}(Y)$
- if Y_1 and Y_2 are independent then $\mathrm{Var}(aY_1 + bY_2) = a^2\mathrm{Var}(Y_1) + b^2\mathrm{Var}(Y_2)$.

Moments

The kth population moment of Y about zero is given by

$$\mu_k{}' = \mathrm{E}\left(Y^k\right) \qquad \text{for } k = 1, 2, 3, \ldots$$

Hence $\mu_1{}' = \mathrm{E}(Y)$.

The kth central moment (or kth moment about the mean) of Y is given by

$$\mu_k = \mathrm{E}\left\{[Y - \mathrm{E}(Y)]^k\right\} \qquad \text{for } k = 2, 3, \ldots \ .$$

Hence $\mu_2 = \mathrm{Var}(Y)$. For symmetric distributions, all odd central moments are zero (if they exist), i.e. $\mu_k = 0$ for $k = 1, 3, 5, \ldots$. Note that $\mu_1' = \mathrm{E}(Y)$ should not be confused with the distribution parameter μ, which may or may not represent the population mean of Y.

Relationship between central moments and moments about zero

$$\begin{aligned} \mu_2 &= \mu_2{}' - \mu_1{}'^2 \\ \mu_3 &= \mu_3{}' - 3\mu_2{}'\mu_1{}' + 2\mu_1{}'^3 \\ \mu_4 &= \mu_4{}' - 4\mu_3{}'\mu_1{}' + 6\mu_2{}'\mu_1{}'^2 - 3\mu_1{}'^4 \end{aligned} \tag{2.1}$$

Relationship between moments about zero and central moments

$$\begin{aligned} \mu_2{}' &= \mu_2 + \mu_1{}'^2 \\ \mu_3{}' &= \mu_3 + 3\mu_2\mu_1{}' + \mu_1{}'^3 \\ \mu_4{}' &= \mu_4 + 4\mu_3\mu_1{}' + 6\mu_2\mu_1{}'^2 + \mu_1{}'^4 \end{aligned} \tag{2.2}$$

Moment based skewness and kurtosis

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

$$\gamma_1 = \sqrt{\beta_1} = \mu_3/(\mu_2)^{1.5} \ .$$

For symmetric distributions, $\mu_3 = 0$ and so $\gamma_1 = 0$. Negatively skewed distributions have $\mu_3 < 0$ and hence $\gamma_1 < 0$, and conversely $\gamma_1 > 0$ for positively skewed distributions.

The (moment based) kurtosis is defined as

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

For the normal distribution, $\beta_2 = 3$, which leads to the definition of the excess kurtosis:

$$\gamma_2 = \mu_4/(\mu_2)^2 - 3 \ ,$$

which measures the excess kurtosis relative to the normal distribution.

One of the problems with moment based properties of a distribution is the fact that for several distributions, moments do not exist. This is common for very fat-tailed (leptokurtic) distributions. The centile based properties described in Section 2.3.1 avoid this problem because they always exist.

Mikis: explain simply why this is the case

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

2.3.1 Centile and Quantile

Continuous

Mikis: are they the same? if not explain

For a continuous random variable Y, the $100p^{\text{th}}$ centile (or pth quantile) of Y is the value y_p such that

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

i.e. $F(y_p) = p$ and hence $y_p = F^{-1}(p)$, where $F^{-1}(\cdot)$ is the *inverse cumulative distribution function* evaluated at $0 \leq p \leq 1$. Hence the centile or quantile function is the inverse cumulative distribution function. For example the 5th 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. $F(y_{0.05}) = 0.05$, i.e. $y_{0.05} = F^{-1}(0.05)$.

Using the exponential cdf (1.4) example, to find y_p we solve $F(y_p) = p$ i.e.

$$1 - \exp(-y_p/\mu) = p \ ,$$

so

$$y_p = -\mu \log(1-p) \ .$$

In **R** in order to get the 60th centile from an exponential distribution with $\mu = 0.5$ use:

```
qEXP(.60, mu=0.5)
```

```
## [1] 0.4581454
```

Discrete

For discrete random variables, $F^{-1}(p)$ does not exist for all p. For example, the Poisson cdf graphed in Figure 1.7 has values (shown up to $6 \leq y < 7$):

$$F(y) = \begin{cases} 0 & y < 0 \\ 0.135 & 0 \leq y < 1 \\ 0.406 & 1 \leq y < 2 \\ 0.677 & 2 \leq y < 3 \\ 0.857 & 3 \leq y < 4 \\ 0.947 & 4 \leq y < 5 \\ 0.983 & 5 \leq y < 6 \\ 0.995 & 6 \leq y < 7 \ . \end{cases}$$

If, for example, we wanted the 60th centile, this would be $y_{0.6} = F^{-1}(0.6)$, i.e. the value of y at which $F(y) = 0.6$, which does not exist. In this case, $y_{0.6}$ is

defined as the value of y at which $F(y)$ first exceeds 0.6, which is $y_{0.6} = 2$. In general, for a discrete random variable Y, y_p is defined by $P(Y < y_p) < p$ and $P(Y \leq y_p) \geq p$, i.e.

$$y_p = \text{ smallest value of } Y \text{ for which } F(y) \geq p \ . \tag{2.3}$$

In **R** in order to get the 60th centile from a Poisson distribution with $\mu = 2$:

```
qPO(.60, mu=2)
```

```
## [1] 2
```

The 60th centile is illustrated for the exponential and Poisson distributions in Figure 2.3.

Figure ??

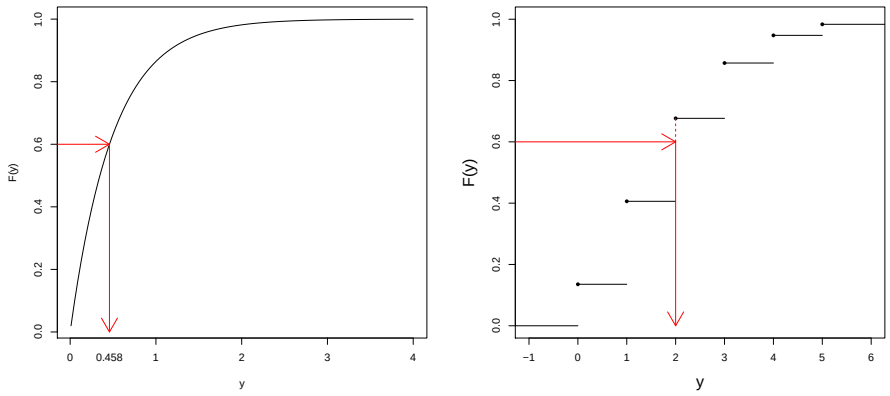


Figure 2.3: Shwoing the (a) 60th centile of the exponential ($\mu = 0.5$) distribution, $y_{0.6} = 0.458$ and (b) 60th centile of the Poisson ($\mu = 2$) distribution, $y_{0.6} = 2$

2.3.2 Median, first and third quartiles

For a continuous random variable Y, the median m is the value of Y for which $P(Y \leq m) = 0.5$. Similarly the first and third quartiles, Q_1 and Q_3, are given by $P(Y \leq Q_1) = 0.25$ and $P(Y \leq Q_3) = 0.75$. Hence

$$\begin{aligned} Q_1 &= y_{0.25} = F^{-1}(0.25) \\ m &= y_{0.5} = F^{-1}(0.5) \\ Q_3 &= y_{0.75} = F^{-1}(0.75) \ . \end{aligned}$$

Note the median $m = Q_2$. For a discrete random variable Y, the definitions of m, Q_1 and Q_3 follow (2.3), i.e.

$$m = y_{0.5} = \text{smallest value of } y \text{ for which } F(y) \geq 0.5 \ ,$$

with corresponding definitions for $Q1$ and $Q3$.

The *interquartile* range is $\text{IR} = Q_3 - Q_1$. The *semi-interquartile* range, $\text{SIR} = (Q_3 - Q_1)/2$, is a measure of spread.

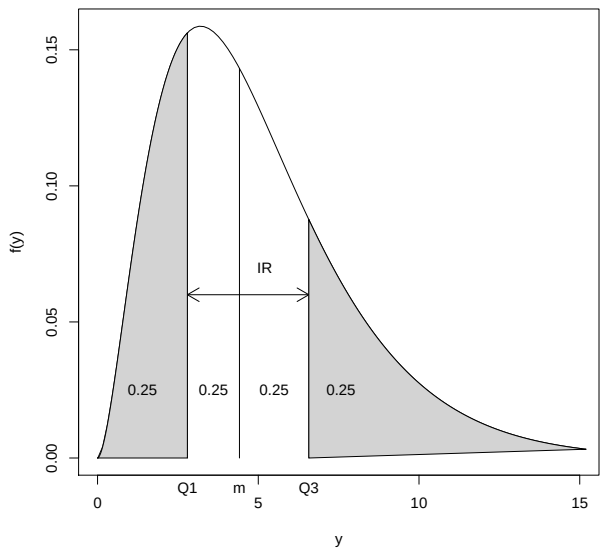


Figure 2.4: Showing how $Q1$, m (median), $Q3$ and the interquartile range of a continuous distribution are derived from $f(y)$.

Figure 2.4 shows how Q_1, m, and Q_3 are obtained for a continuous random variable, from its pdf.

2.3.3 Centile based skewness and kurtosis

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

Galton's skewness measure γ is defined as

$$\gamma = \frac{(Q_3 - Q_1)/2 - m}{\text{SIR}} \ .$$

A centile based kurtosis measure $t_{0.49}$ is given by Andrews et al. [1972] as

$$t_{0.49} = \frac{y_{0.99} - y_{0.01}}{\text{IR}} \ .$$

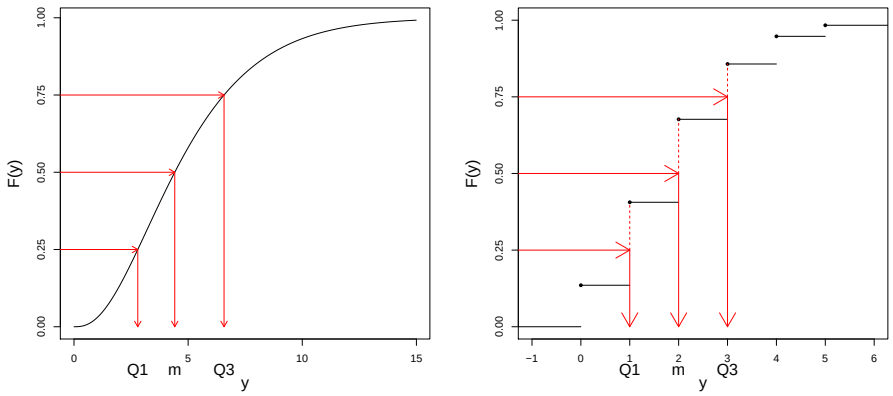


Figure 2.5: Showing how Q_1, the median (m) and Q_3 are derived for (a) a continuous and (b) a discrete distribution, from their cdf.

This measure has been scaled to be one for the normal distribution, giving $st_{0.49} = t_{0.49}/3.449$, Rosenberger and Gasko [1983].

2.4 Moment, cumulant and probability generating functions[1]

2.4.1 Moment generating function

The moment generating function of Y is given by

$$M_Y(t) = \mathrm{E}\left(e^{tY}\right) \tag{2.4}$$

provided it exists. It is called the moment generating function because setting $t = 0$ in its rth derivative with respect to t gives $E\left(Y^r\right)$, i.e.

$$\mathrm{E}\left(Y^r\right) = \left.\frac{d^{(r)}M_Y(t)}{dt^{(r)}}\right|_{t=0} . \tag{2.5}$$

Example

Let Y have the exponential distribution with mean μ, with probability density function $f(y) = \frac{1}{\mu}e^{-y/\mu}$ for $y > 0$. The moment generating function of Y is

[1]This is a more advanced topic and can be omitted by more practical readers.

given by:

$$\begin{aligned} M_Y(t) = \mathrm{E}\left(e^{tY}\right) &= \int_0^\infty e^{ty}\frac{1}{\mu}e^{-y/\mu}dy \\ &= \frac{1}{\mu}\int_0^\infty e^{y\left(\frac{1}{\mu}-t\right)}dy \\ &= (1-\mu t)^{-1} \ , \end{aligned}$$

which exists as long as $t < 1/\mu$. Hence from (2.5),

$$\mu_1{}' = \mathrm{E}\left(Y\right) = \left.\frac{dM_Y(y)}{dt}\right|_{t=0} = \left.\mu\left(1-\mu t\right)^{-2}\right|_{t=0} = \mu$$

and

$$\mu_2{}' = \mathrm{E}\left(Y^2\right) = \left.\frac{d^{(2)}M_Y(y)}{dt}\right|_{t=0} = \left.2\mu^2\left(1-\mu t\right)^{-3}\right|_{t=0} = 2\mu^2 \ .$$

Similarly $\mu_3{}' = \mathrm{E}\left(Y^3\right) = 6\mu^3$ and $\mu_4{}' = \mathrm{E}\left(Y^4\right) = 24\mu^4$. Hence

$$\mu_2 = \mathrm{Var}\left(Y\right) = \mu_2{}' - \mu_1{}'^2 = \mathrm{E}\left(Y^2\right) - \left[\mathrm{E}\left(Y\right)\right]^2 = 2\mu^2 - \mu^2 = \mu^2$$

$$\mu_3 = \mu_3{}' - 3\mu_2{}'\mu_1{}' + 2\mu_1{}'^3 = 6\mu^3 - 6\mu^3 + 2\mu^3 = 2\mu^3$$

$$\mu_4 = \mu_4{}' - 4\mu_3{}'\mu_1{}' + 6\mu_2{}'\mu_1{}'^2 - 3\mu_1{}'^4 = 24\mu^4 - 24\mu^4 + 12\mu^4 - 3\mu^4 = 9\mu^4 \ .$$

Hence the skewness of Y is

$$\gamma_1 = \mu_3/\mu_2^{1.5} = 2\mu^3/\mu^3 = 2 \ ,$$

and the kurtosis is

$$\beta_2 = \mu_4/\mu_2^2 = 9\mu^4/\mu^4 = 9$$

with excess kurtosis $\gamma_2 = \beta_2 = 6$.

2.4.2 Cumulant generating function

The cumulant generating function of Y is given by

$$K_Y(t) = \log M_Y(t)$$

provided it exists. It is called the cumulant generating function because setting $t = 0$ in its rth derivative with respect to t gives κ_r, the kth cumulant of Y:

$$\kappa_r = \left.\frac{d^{(r)}K_Y(t)}{dt^{(r)}}\right|_{t=0} \ . \tag{2.6}$$

Relationship between cumulants and moments

$$\begin{aligned} \kappa_1 &= \mu_1{}' = \mathrm{E}(Y) \\ \kappa_2 &= \mu_2 = \mathrm{Var}(Y) \\ \kappa_3 &= \mu_3 \\ \kappa_4 &= \mu_4 - 3\mu_2^2 \ . \end{aligned} \tag{2.7}$$

Example

Let $Y \sim \texttt{EXP}(\mu)$ then $K_Y(t) = \log(M_Y(t)) = -\log(1-\mu t)$, provided $t < 1/\mu$. Hence from (2.6) and (5.18),

$$\begin{aligned} \kappa_1 &= \mu(1-\mu t)^{-1}\Big|_{t=0} = \mu, \text{ so } \mathrm{E}(Y) = \mu \\ \kappa_2 &= \mu^2(1-\mu t)^{-2}\Big|_{t=0} = \mu^2, \text{ so } \mathrm{Var}(Y) = \mu^2 \\ \kappa_3 &= 2\mu^3(1-\mu t)^{-3}\Big|_{t=0} = 2\mu^3, \text{ so } \mu_3 = 2\mu^3 \\ \kappa_4 &= 6\mu^4(1-\mu t)^{-4}\Big|_{t=0} = 6\mu^4, \text{ so } \mu_4 = 9\mu^4 \ . \end{aligned}$$

The skewness of Y is $\gamma_1 = \kappa_2/\kappa_2^{1.5} = \mu_3/\mu_2^{1.5} = 2$ and the excess kurtosis is $\gamma_2 = \kappa_4/\kappa_2^2 = 6$.

2.4.3 Probability generating function

The probability generating function (PGF) of a discrete random variable Y is given by

$$G_Y(t) = \mathrm{E}\left(t^Y\right) \tag{2.8}$$

provided it exists. It is called the probability generating function because setting $t = 0$ in its rth derivative with respect to t gives

$$P(Y = r) = \frac{1}{r!}\frac{d^{(r)}G_Y(t)}{dt^{(r)}}\Big|_{t=0} \ . \tag{2.9}$$

Note that $G_Y(t) = M_Y(\log t)$ and $M_Y(t) = G_Y(e^t)$.

Example

Let Y have a Poisson distribution with mean μ, denoted $Y \sim PO(\mu)$, with probability function (18.3). The PGF of Y is given by

$$\begin{aligned} G_Y(t) = \mathrm{E}\left(t^Y\right) &= \sum_{y=0}^{\infty} t^y \frac{e^{-\mu}\mu^y}{y!} = e^{-\mu}\sum_{y=0}^{\infty}\frac{(t\mu)^y}{y!} \\ &= e^{-\mu}e^{\mu t} = e^{\mu(t-1)}. \end{aligned}$$

Hence from (2.9),

$$\begin{aligned} P(Y=0) &= G_Y(t)|_{t=0} = e^{\mu(t-1)}\Big|_{t=0} = e^{-\mu} \\ P(Y=1) &= \frac{dG_Y(t)}{dt}\Big|_{t=0} = \mu e^{\mu(t-1)}\Big|_{t=0} = e^{-\mu}\mu \\ P(Y=2) &= \frac{1}{2!}\frac{d^{(2)}G_Y(t)}{dt^{(2)}}\Big|_{t=0} = \frac{1}{2!}\mu^2\, e^{\mu(t-1)}\Big|_{t=0} = \frac{e^{-\mu}\mu^2}{2!}\ . \end{aligned}$$

2.4.4 Properties of MGF and PGF

- Let $Y = a + bY_1$ where a and b are constants, then $M_Y(t) = e^{at}M_{Y_1}(bt)$.
- Let $Y = aY_1 + bY_2$ where a and b are constants and Y_1 and Y_2 are independent random variables. Then $M_Y(t) = M_{Y_1}(at)M_{Y_2}(bt)$ for values of t for which $M_{Y_1}(at)$ and $M_{Y_1}(bt)$ exist. Also $G_Y(t) = G_{Y_1}(t^a)\,G_{Y_2}(t^b)$ for values of t for which $G_{Y_1}(t^a)$ and $G_{Y_2}(t^b)$ exist. In particular, if $Y = Y_1 + Y_2$ then $M_Y(t) = M_{Y_1}(t)M_{Y_2}(t)$ and $G_Y(t) = G_{Y_1}(t)\,G_{Y_2}(t)$, provided Y_1 and Y_2 are independent.
- The MGF and PGF are useful for finding the distribution of finite and countably infinite mixture distributions, see section ????
- The PGF is useful for finding the distribution of a stopped sum see section ?????

Example 1

Let $Y_1 \sim \texttt{EXP}(\mu)$, then $Y = bY_1$ for $b > 0$ has MGF given by $M_Y(t) = M_{Y_1}(bt) = (1 - b\mu t)^{-1}$ for $t < 1/b\mu$. Hence $Y \sim EXP(b\mu)$.

Example 2

Let $Y_1 \sim \texttt{PO}(\mu_1)$ and $Y_2 \sim \texttt{PO}(\mu_2)$ where Y_1 and Y_2 are independent, then $Y = Y_1 + Y_2$ has PGF given by $G_Y(t) = G_{Y_1}(e^t)\,G_{Y_2}(e^t) = e^{\mu_1(t-1)}e^{\mu_2(t-1)} = e^{(\mu_1+\mu_2)(t-1)}$. Hence $Y \sim \texttt{PO}(\mu_1 + \mu_2)$.

2.5 Bibliographic Notes

2.6 Exercises

Part II

The GAMLSS family of distributions

Chapter 3

The GAMLSS Family of Distributions

This chapter provides:

1. an introduction to different types of distributions within GAMLSS and in particular: continuous distributions discrete distributions and mixed distributions
2. information how to visualise different distributions
3. information about link functions withing **gamlss.dist**

Mikis: I think this chapter should be an updated version of Chapter 6 in the first book

3.1 Types of distribution within the GAMLSS family

One of the great advantages of the GAMLSS framework (and its associated distributions in package `gamlsss.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. Within the GAMLSS framework the probability (density) function $f(y\,|\,\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\mu, \sigma, \nu, \tau)$, 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 $f(y\,|\,\boldsymbol{\theta})$ and its first (and optionally expected second and cross) derivatives with respect to each of the parameters of $\boldsymbol{\theta}$ must be computable. Explicit derivatives are preferable, but numerical derivatives can be used (resulting in reduced computational speed).

Note that for consistency of notation, in the **R** implementation of **gamlss**, we

have called the $\boldsymbol{\theta}$ parameters as μ, σ, ν and τ. As we have mentioned in the previous chapter, the parameters μ, σ, ν and τ, represent in many distributions the location, scale, skewness and kurtosis parameters respectively, if such a one to one relationship exists. More generally μ, σ, ν and τ are just convenient names for the parameters.

This Chapter introduces all the available distributions within the current implementation of GAMLSS in **R**. Some of those distributions are *explicit* in the sense that their implementation exist within the package **gamlss.dist** while other are *generated* from existing explicit distribution.

We refer to both type of distributions as the *GAMLSS family*, to be consistent with the **R** implementation where the class of the distributions all those distributions is defined as `gamlss.family`.

The explicit `gamlss.family` distributions are subdivided in three distinct types according to the type of random variable modelled as a response. Those three distinct types of GAMLSS distributions are:

1. continuous `gamlss.family` distributions,
2. discrete `gamlss.family` distributions and
3. mixed `gamlss.family` distributions.

In the GAMLSS model, Y has probability (density) function $f(y\,|\,\boldsymbol{\theta})$ with up to four parameters, i.e. $\boldsymbol{\theta} = \mu$ or (μ, σ) or (μ, σ, ν) or (μ, σ, ν, τ).

In the next three sessions we consider the three explicit types of distribution in turn.

3.2 Continuous distributions in GAMLSS

Table 3.1 shows all the explicit continuous distribution within the `gamlss.family`. The columns of the table display the distribution name, the **R** implementation name, the range of Y and the default *link* functions of all available parameters.

Link functions were introduced by Nelder and Wedderburn [1972] for Generalized Linear Models, but are appropriate for all regression models since they guarantee that parameter estimates remain within the appropriate range. For example if a parameter θ has range $0 < \theta < \infty$, the logarithmic transformation

$$\log(\theta) = \eta$$

produces $-\infty < \eta < \infty$. In parameter estimation, if the logarithmic link is used, η is estimated and transformed back to θ as

$$e^{\eta} = \theta \ ,$$

which is guaranteed to be in the range $(0, \infty)$. For the logarithmic link, $\log(\theta)$ is the link function and e^{η} is the reverse link function. In general, the link function is denoted as $g(\theta) = \eta$, and the reverse link as $g^{-1}(\eta) = \theta$. Link functions $g(\cdot)$ have to be monotonic and differentiable. Table 3.1 gives the link functions used in `gamlss`.

For any `gamlss` distribution, each model parameter has its own link function, appropriate to its range, and these are denoted as

$$\begin{aligned} g_1(\mu) &= \eta_1 \\ g_2(\sigma) &= \eta_2 \\ g_3(\nu) &= \eta_3 \\ g_4(\tau) &= \eta_4 \ . \end{aligned}$$

Note that, for presentational reasons, in the columns displaying the link function of the parameters in Table 3.1, the `identity` and the `logshiftto2` links are abbreviated in as "indent." and "log-2" respectively.

Distribution	`gamlss` name	Range R_Y	Parameter link functions μ	σ	ν	τ
beta	`BE`	$(0,1)$	logit	logit	-	-
Box-Cox Cole-Green	`BCCG`	$(0,\infty)$	ident.	log	ident.	-
Box-Cox Cole-Green orig.	`BCCGo`	$(0,\infty)$	log	log	ident.	-
Box-Cox power exponential	`BCPE`	$(0,\infty)$	ident.	log	ident.	log
Box-Cox Power Expon. orig.	`BCPEo`	$(0,\infty)$	log	log	ident.	log
Box-Cox t	`BCT`	$(0,\infty)$	ident.	log	ident.	log
Box-Cox t orig.	`BCTo`	$(0,\infty)$	log	log	ident.	log
exponential	`EXP`	$(0,\infty)$	log	-	-	-
exponential Gaussian	`exGAUS`	$(-\infty,\infty)$	ident.	log	log	-
exponential gen. beta 2	`EGB2()`	$(-\infty,\infty)$	ident.	log	log	log
gamma	`GA`	$(0,\infty)$	log	log	-	-
generalised beta type 1	`GB1`	$(0,1)$	logit	logit	log	log
generalised beta type 2	`GB2`	$(0,\infty)$	log	log	log	log
generalised gamma	`GG`	$(0,\infty)$	log	log	ident.	-
generalised inv. Gaussian	`GIG`	$(0,\infty)$	log	log	ident.	-
generalised t	`GT`	$(-\infty,\infty)$	ident.	log	log	log
Gumbel	`GU`	$(-\infty,\infty)$	ident.	log	-	-
inverse Gamma	`IGAMMA`	$(0,\infty)$	log	log	-	-
inverse Gaussian	`IG`	$(0,\infty)$	log	log	-	-
Johnson's SU repar.	`JSU`	$(-\infty,\infty)$	ident.	log	ident.	log
Johnson's original SU	`JSUo`	$(-\infty,\infty)$	ident.	log	ident.	log
logistic	`LO`	$(-\infty,\infty)$	ident.	log	-	-
logit normal	`LOGITNO`	$(0,1)$	ident.	log	-	-
log normal	`LOGNO`	$(0,\infty)$	ident.	log	-	-
log normal 2	`LOGNO2`	$(0,\infty)$	log	log	-	-
log normal (Box-Cox)	`LNO`	$(0,\infty)$	ident.	log	fixed	-

NET	`NET`	$(-\infty, \infty)$	ident.	log	fixed	fixed
normal	`NO, NO2`	$(-\infty, \infty)$	ident.	log	-	-
normal family	`NOF`	$(-\infty, \infty)$	ident.	log	-	-
Pareto 2	`PARETO2`	$(0, \infty)$	log	log	-	-
Pareto 2 original	`PARETO2o`	$(0, \infty)$	log	log	-	-
Pareto 2 repar	`GP`	$(0, \infty)$	log	log	-	-
power exponential	`PE`	$(-\infty, \infty)$	ident.	log	log	-
reverse gen. extreme	`RGE`	$y > \mu - (\sigma/\nu)$	ident.	log	log	-
reverse Gumbel	`RG`	$(-\infty, \infty)$	ident.	log	-	-
sinh-arcsinh	`SHASH`	$(-\infty, \infty)$	ident.	log	log	log
sinh-arcsinh original	`SHASHo`	$(-\infty, \infty)$	ident.	log	ident.	log
sinh-arcsinh original 2	`SHASHo2`	$(-\infty, \infty)$	ident.	log	ident.	log
skew normal type 1	`SN1`	$(-\infty, \infty)$	ident.	log	ident.	-
skew normal type 2	`SN2`	$(-\infty, \infty)$	ident.	log	log	-
skew power exp. type 1	`SEP1`	$(-\infty, \infty)$	ident.	log	ident.	log
skew power exp. type 2	`SEP2`	$(-\infty, \infty)$	ident.	log	ident.	log
skew power exp. type 3	`SEP3`	$(-\infty, \infty)$	ident.	log	log	log
skew power exp. type 4	`SEP4`	$(-\infty, \infty)$	ident.	log	log	log
skew t type 1	`ST1`	$(-\infty, \infty)$	ident.	log	ident.	log
skew t type 2	`ST2`	$(-\infty, \infty)$	ident.	log	ident.	log
skew t type 3	`ST3`	$(-\infty, \infty)$	ident.	log	log	log
skew t type 3 repar	`SST`	$(-\infty, \infty)$	ident.	log	log	log-2
skew t type 4	`ST4`	$(-\infty, \infty)$	ident.	log	log	log
skew t type 5	`ST5`	$(-\infty, \infty)$	ident.	log	ident.	log
t Family	`TF`	$(-\infty, \infty)$	ident.	log	log	-
t Family repar	`TF2`	$(-\infty, \infty)$	ident.	log	log-2	-
Weibull	`WEI`	$(0, \infty)$	log	log	-	-
Weibull (PH)	`WEI2`	$(0, \infty)$	log	log	-	-
Weibull (μ the mean)	`WEI3`	$(0, \infty)$	log	log	-	-

Table 3.1: Continuous distributions implemented within the `gamlss.dist` package (with default link functions)

3.3 Discrete distributions in GAMLSS

Table 3.2 contains all the discrete distributions available in the `gamlss.family`. There are two types of discrete distribution according to their range:

- the *binomial* data type distributions with range $0, 1, \ldots, n$; and
- the *count* data type distributions with range $0, 1, \ldots, \infty$.

The binomial data type of distributions include the binomial, the beta binomial, and their inflated and zero-adjusted versions.

The rest of the distributions are count data type distributions, of which the basic distribution is the Poisson. Most of the other count data distributions are derived from the Poisson, see Chapter 5.

Distribution	`gamlss`	Range	Parameter link function			
	name	R_Y	μ	σ	ν	τ
beta binomial	`BB`	$\{0,1,\ldots,n\}$	logit	log	-	-
beta neg. binomial	`BNB`	$\{0,1,2,\ldots\}$	log	log	log.	-
binomial	`BI`	$\{0,1,\ldots,n\}$	logit	-	-	-
geometric	`GEOM`	$\{0,1,2,\ldots\}$	log	-	-	-
geometric (original)	`GEOMo`	$\{0,1,2,\ldots\}$	logit	-	-	-
logarithmic	`LG`	$1,2,\ldots,\infty$	logit	-	-	-
Delaporte	`DEL`	$\{0,1,2,\ldots\}$	log	log	logit	-
double Poisson	`DPO`	$\{0,1,2,\ldots\}$	log	log	-	-
Multinomial	`MULTIN`	$1,2,\ldots,n$	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}$	-
negative binomial type I	`NBI`	$\{0,1,2,\ldots\}$	log	log	-	-
negative binomial type II	`NBII`	$\{0,1,2,\ldots\}$	log	log	-	-
neg. binomial family	`NBF`	$\{0,1,2,\ldots\}$	log	log	ident.	-
Poisson	`PO`	$\{0,1,2,\ldots\}$	log	-	-	-
Poisson inv. Gaussian	`PIG`	$\{0,1,2,\ldots\}$	log	log	-	-
Poisson shifted GIG	`PSGIG`	$\{0,1,2,\ldots\}$	log	log	logit.	logit
Sichel	`SI`	$\{0,1,2,\ldots\}$	log	log	identity	-
Sichel (μ the mean)	`SICHEL`	$\{0,1,2,\ldots\}$	log	log	identity	-
Waring (μ the mean)	`WARING`	$\{0,1,2,\ldots\}$	log	log	-	-
Yule (μ the mean)	`YULE`	$\{0,1,2,\ldots\}$	log	-	-	-
zero alt. beta binomial	`ZABB`	$\{0,1,\ldots,n\}$	logit	log	logit	-
zero alt. beta neg. binom.	`ZABNB`	$\{0,1,2,\ldots\}$	log	log	ident.	logit
zero alt. binomial	`ZABI`	$\{0,1,\ldots,n\}$	logit	logit	-	-
zero alt. logarithmic	`ZALG`	$\{0,1,2,\ldots\}$	logit	logit	-	-
zero alt. neg. binomial	`ZANBI`	$\{0,1,2,\ldots\}$	log	log	logit	-
zero alt. neg. binom. fam.	`ZANBF`	$\{0,1,2,\ldots\}$	log	log	log	logit
zero alt. PIG	`ZAPIG`	$\{0,1,2,\ldots\}$	log	log	logit	-
zero alt. Sichel	`ZASICHEL`	$\{0,1,2,\ldots\}$	log	log	ident.	logit
zero alt. poisson	`ZAP`	$\{0,1,2,\ldots\}$	log	logit	-	-
zero alt. zipf	`ZAZIPF`	$\{0,1,2,\ldots\}$	log	logit	-	-
zero inf. beta binomial	`ZIBB`	$\{0,1,2,\ldots\}$	logit	log	logit	-
zero inf. beta neg. binom.	`ZIBNB`	$\{0,1,2,\ldots\}$	log	log	log	logit
zero inf. binomial	`ZIBI`	$\{0,1,\ldots,n\}$	logit	logit	-	-
zero inf. neg. binomial	`ZINBI`	$\{0,1,2,\ldots\}$	log	log	logit	-
zero inf. neg. binom. fam.	`ZINBF`	$\{0,1,2,\ldots\}$	log	log	log	logit
zero inf. poisson	`ZIP`	$\{0,1,2,\ldots\}$	log	logit	-	-
zero inf. poisson (μ the mean)	`ZIP2`	$\{0,1,2,\ldots\}$	log	logit	-	-
zero inf. PIG	`ZIPIG`	$\{0,1,2,\ldots\}$	log	log	logit	-

zero inf. Sichel	ZISICHEL	$\{0,1,2,\ldots\}$	log	log	ident.	logit
zipf	ZIPF	$\{0,1,2,\ldots\}$	log	-	-	-

Table 3.2: Discrete count distributions implemented within **gamlss.dist**, with default link functions.

More on the discrete distribution will be given in Chapter 5.

3.4 Mixed distributions in GAMLSS

Mixed distributions are a special case of finite mixture distributions described in Chapter 7. They are mixtures of continuous and discrete distributions, i.e. continuous distributions in which R_Y has been expanded to include some discrete values with non-zero probabilities. Table 3.3 shows the existing mixed distributions in `gamlss.family`.

Distribution	gamlss name	Range R_Y	Parameter link functions μ	σ	ν	τ
beta inflated (at 0)	BEOI	$[0,1)$	logit	log	logit	-
beta inflated (at 0)	BEINF0	$[0,1)$	logit	logit	log	-
beta inflated (at 1)	BEZI	$(0,1]$	logit	log	logit	-
beta inflated (at 1)	BEINF1	$(0,1]$	logit	logit	log	-
beta inflated (at 0 and 1)	BEINF	$[0,1]$	logit	logit	log	log
zero adjusted GA	ZAGA	$[0,\infty)$	log	log	logit	-
zero adjusted IG	ZAIG	$[0,\infty)$	log	log	logit	-

Table 3.3: Mixed distributions implemented within the `gamlss.dist` package

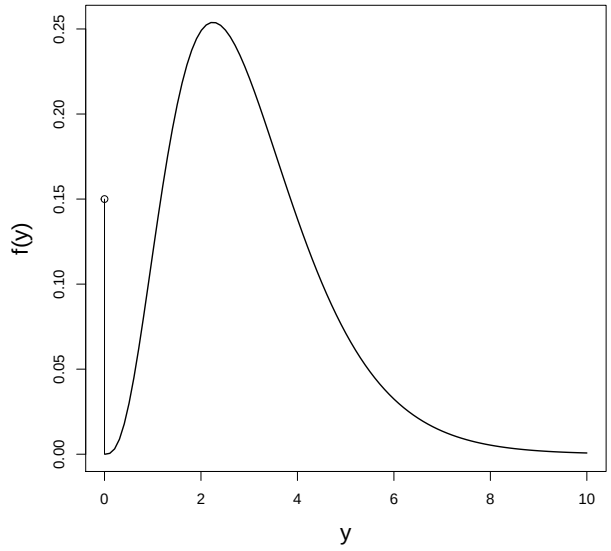
Note that in the range R_Y in Table 3.3, a square bracket indicates the end value is included in the interval, while a round bracket indicates the end value is not included. (This is consistent with conventional mathematical notation.) Hence, for example, $[0,1)$ includes 0 but not 1.

The ZAGA and ZAIG distributions are useful for modelling a response variable on the interval $[0,\infty)$, such as amount of insurance claim in a year. Most of the policies do not have any claims and therefore there is a high probability of the claim amount being zero, but for policies that do have claims, the distribution of the claim amount is defined on the positive real line $(0,\infty)$. The zero adjusted gamma (ZAGA) distribution is shown in Figure 3.1.

The beta inflated distributions are useful for modelling a proportion (or fractional) response variable Y on the interval 0 to 1, including 0, 1 or both.

Figure 3.1

```
p = 0.15
curve((1-p)*dGA(x,mu=3,sigma=0.5),0.01,10,xlab="y",ylab="f(y)",
      lwd=1.5,cex.lab=1.5)
segments(0,0,0,p)
points(0,p)
```



R code on page 62

Figure 3.1: The zero adjusted gamma distribution, an example of a mixed distribution

Extra mixed distributions can be defined using the package **gamlss.inf**. More details about mixed distributions are given in Chapter ??.

3.5 Generating GAMLSS family distributions

There are several ways to extend the `gamlss.family` distributions. This can be achieved by

- creating a new `gamlss.family` distribution;
- creating a *log* or *logit* version of a distribution from an existing continuous `gamlss.family` distribution on the real line;
- truncating an existing `gamlss.family`;
- using a censored version of an existing `gamlss.family`; and

- mixing different `gamlss.family` distributions to create a new finite mixture distribution.

These extensions are dealt with in the following sections.

3.5.1 New `gamlss.family` distribution

To create a new `gamlss.family` distribution is relatively simple if the probability (density) function of the distribution can be evaluated easily. To do that, a file of a current `gamlss.family` distribution, having the same number of distribution parameters, may be amended. Section 6.3 of Stasinopoulos et al. [2017] provides an example of how this can be done.

3.5.2 New log and logit versions from a continuous `gamlss.family` on $(-\infty, \infty)$

Any continuous random variable Z defined on $(-\infty, \infty)$ can be transformed by $Y = \exp(Z)$ to a random variable defined on $(0, \infty)$. A well-known example of this is the log-normal distribution, which is defined by $Y = \exp(Z)$ where Z is a normally distributed random variable. This is achieved in **gamlss** by using the function `gen.Family()` with the option `type="log"`. The following is an example in which we take a t family distribution, i.e. $Z \sim \texttt{TF}(\mu, \sigma, \nu)$ and apply an exponential transform $Y = \exp(Z)$ to give $Y \sim \log \texttt{TF}(\mu, \sigma, \nu)$, i.e. we create a log-$t$ family distribution on $(0, \infty)$. We then generate a random sample of 200 observations from the distribution and finally fit the distribution to the generated data.

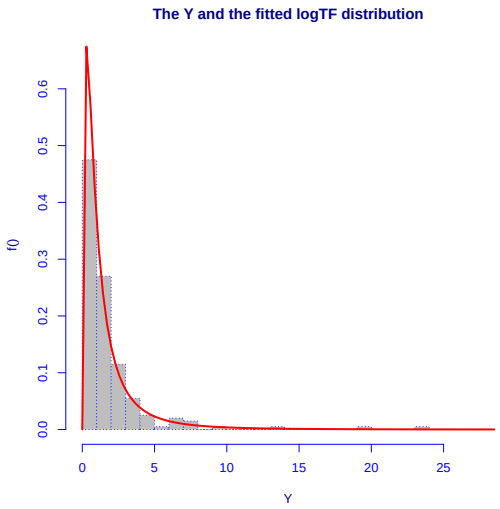
```
rm(list=ls())
library(gamlss)
```

```
# generate the distribution
gen.Family("TF", type="log")
```

```
## A  log  family of distributions from TF has been generated
##  and saved under the names:
##  dlogTF plogTF qlogTF rlogTF logTF
```

```
#generate 200 observations with df=nu=10
#(and default mu=0 and sigma=1)
set.seed(1434)
Y<- rlogTF(200, nu=10)
# fit the distribution
h1 <- histDist(Y, family=logTF, nbins=30, ylim=c(0,.65), line.wd=2.5)
```

Similarly take $Z \sim \texttt{TF}(\mu, \sigma, \nu)$ and apply the inverse logit transformation $Y = 1/(1 + \exp(-Z))$ to give $Y \sim \texttt{logitTF}(\mu, \sigma, \nu)$, i.e. a logit-$t$ family distribution



R code on page **??**

Figure 3.2: A fitted log-t distribution to 200 simulated observations.

on $(0, 1)$:

```
gen.Family("TF", type="logit")
```

```
## A  logit  family of distributions from TF has been generated
##  and saved under the names:
##  dlogitTF plogitTF qlogitTF rlogitTF logitTF
```

3.5.3 Truncating `gamlss.family` distributions

A truncated distribution is appropriate when the range of possible values of a variable Y is a subset of the range of an original distribution. Truncating existing `gamlss.family` distributions can be achieved using the package **gamlss.tr**. The function `gen.trun()` takes any `gamlss.family` distribution and generates the `d`, `p`, `q`, `r` and fitting functions for the specified truncated distribution. The truncation can be left, right or in both tails of the range of the response variable. For example, a t family distribution (`TF`) can be truncated in both tails, at and below 0 and at and above 100, as follows:

```
# generate the distribution
library(gamlss.tr)
gen.trun(par=c(0,100),family="TF", name="0to100", type="both")
```

```
## A truncated family of distributions from TF has been generated
##  and saved under the names:
```

Figure 3.3

```
##  dTF0to100 pTF0to100 qTF0to100 rTF0to100 TF0to100
## The type of truncation is both
##  and the truncation parameter is 0 100
```

```
Y <-rTF0to100(1000, mu=20 ,sigma=20, nu=5)
h1 <- histDist(Y, family=TF0to100, nbins=30, xlim=c(0,100),
               line.col="darkblue",line.wd=2.5)
```

R code on page 65

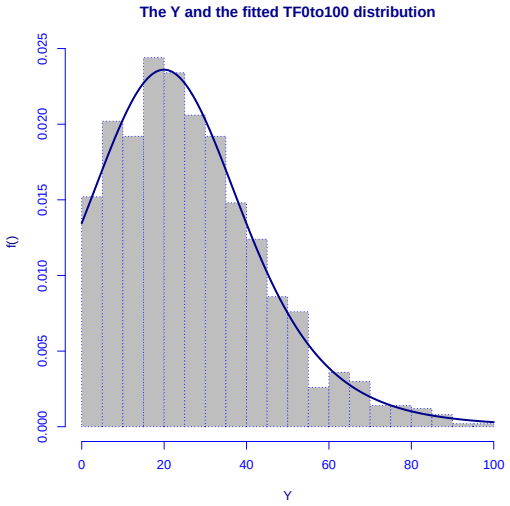


Figure 3.3: A truncated t distribution defined on $(0, 100)$, fitted to 1000 simulated observations

The `gen.trun()` function has the following main arguments:

`par:` a vector with one element for "left" or "right" truncation, or two elements for "both";

`family:` a `gamlss.family` distribution;

`name:` the name for the new truncated distribution;

`type:` `"left"`, `"right"` or `"both"` sides truncation.

For discrete count distributions,

- "left" truncation at the integer a means that the random variable can take values $\{a+1, a+2, \ldots\}$;
- "right" truncation at the integer b means that the random variable can take values up to but not including b, i.e. $\{\ldots, b-2, b-1\}$.
- "both" truncation at the integer interval (a, b) means the random variable can take values $\{a+1, a+2, \ldots, b-1\}$.

For a continuous random variable Y, $\text{Prob}(Y = a) = 0$ for any constant a and therefore the inclusion or not of an end point in the truncated range is not an issue. Hence "left" truncation at value a results in variable $Y > a$, "right" truncation at b results in $Y < b$, and "both" truncation results in $a < Y < b$.

Assume Y is obtained from Z by left truncation at and below a, and right truncation at and above b. If Z is a discrete count variable with probability function $\text{Prob}(Z = z)$, cdf $F_Z(\cdot)$ and a and b are integers $(a < b)$, then

$$\text{Prob}(Y = y) = \frac{\text{Prob}(Z = y)}{F_Z(b-1) - F_Z(a)} \ .$$

for $y = \{a+1, a+2, \ldots, b-1\}$. If Z is continuous with pdf $f_Z(z)$ and a and b are any values with $a < b$,

$$f_Y(y) = \frac{f_Z(y)}{F_Z(b) - F_Z(a)} \ .$$

for $a < y < b$.

3.6 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.

For example the pdf, cdf, inverse cdf and random generating functions of the gamma distribution which within the `gamlss.family` has the name `GA`, are given as `dGA`, `pGA`, `qGA` and `rGA` respectively.

There are several ways to explore the `gamlss.family` distributions. Here we suggest the following.

3.6.1 Using the distribution demos

A `gamlss.family` distribution can be displayed graphically in **R** using the `gamlss.demo` package. For example the following commands will load the **gamlss.demo** package and start the gamlss demos.

```
library(gamlss.demo)
gamlss.demo()
```

This will 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

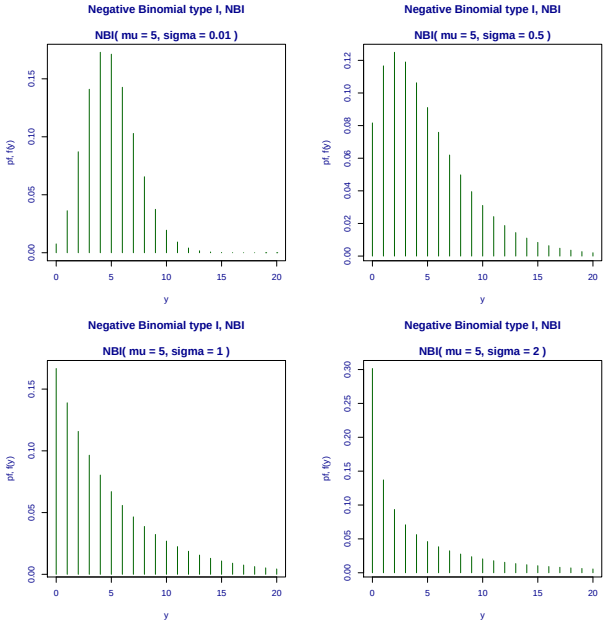
(currently from Tables 3.1, 5.1 and 3.3) to be displayed graphically and its parameters adjusted interactively.

3.6.2 Using the `pdf.plot()` function

An alternative method of graphically displaying the probability (density) functions is to use the `pdf.plot()` function. The following code produces Figure 3.4:

Figure 3.4

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



R code on page 68

Figure 3.4: Plotting the negative binomial distribution using the `pdf.plot()` function

3.6.3 Plotting the `d`, `p` and `r` functions of a distribution

Examples of plotting using the `d`, `p` and `r` functions are given in Figures 1.9 and 1.10.

3.7 Bibliographic Notes

Exercises

Chapter 4

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.

Mikis: Needs examples to show why those distribution are helpfull

4.1 Introduction

As discussed in Chapter 2, continuous distributions can be symmetric, negatively or positively skewed, and also mesokurtic, leptokurtic or platykurtic. In this chapter we study the GAMLSS family of continuous distributions in more detail. In particular we discuss the shapes of the distributions, and their flexibility in modelling data.

Distributions implemented in the **R** package **gamlss** have up to four parameters which we denote as μ, σ, ν and τ. In general, we use μ as a location parameter, σ as a scale parameter, while ν and τ may deal with skewness and/or kurtosis. However for some distributions these parameters have different interpretations. It is important to emphasise that μ and σ do not necessarily have their conventional meanings of the expected value, $\mathrm{E}(Y)$ and stadard deviation, $\mathrm{SD}(Y)$; respectively. For example, for the gamma distribution, $\texttt{GA}(\mu, \sigma)$, as parametrised in GAMLSS, the mean is given by the distribution parameter μ while the variance is given by $\sigma^2\mu^2$ therefore $\mathrm{SD}(Y) = \sigma\mu$. The relationship between the distribution parameters (μ, σ, ν and τ) and moment-based measurements (mean, variance, skewness and kurtosis) is a function of the properties of the specific distribution.

Tables 4.1 – 4.3 provide lists of the continuous GAMLSS family distributions of Y with ranges $\mathbb{R} = (-\infty, \infty)$, $\mathbb{R}^+ = (0, \infty)$ and $(0, 1)$, respectively. Many of these distributions can be generated by one (or more) of the methods described in Chapter 11. The probability density function and other properties of the distributions can be found in Part Two.

Within Tables 4.1 – 4.3, positive (negative) skewness is taken to indicate that either the moment based skewness $\gamma_1 > 0$ ($\gamma_1 < 0$) or, if γ_1 does not exist, then the centile based skewness $\gamma > 0$ ($\gamma < 0$). Leptokurtic (platykurtic) indicates that either the moment based excess kurtosis $\gamma_2 > 0$ ($\gamma_2 < 0$) or, if γ_2 does not exist, then the centile based kurtosis $st_{0.49} > 1$ ($st_{0.49} < 1$). Mesokurtic is taken to indicate $\gamma_2 = 0$. The measures γ_1, γ_2, γ and $st_{0.49}$ are defined in Sections 2.2 and 2.3.

In the skewness column, "both" 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. "Pos", "neg" and "sym" in the skewness column indicate positive skewness, negative skewness and symmetry, respectively. Brackets in the skewness or kurtosis columns indicate that the parameter cannot be modelled independently of the location and scale parameters.

4.2 Continuous distributions on $\mathbb{R}$

We now discuss the two-, three- and four-parameter distributions on $\mathbb{R}$, listed in Table 4.1. All of the distributions in Table 4.1, with the exception of the exGAUS distribution, are *location-scale* families of distributions with location parameter μ and scale parameter σ. For these distributions, if $Y \sim \mathcal{D}(\mu, \sigma, \nu, \tau)$ then $\varepsilon = (Y - \mu)/\sigma \sim \mathcal{D}(0, 1, \nu, \tau)$, i.e. $Y = \mu + \sigma\varepsilon$, so Y is a scaled and shifted version of the random variable ε. Then $Y' = a + bY \sim \mathcal{D}(a + b\mu, b\sigma, \nu, \tau)$. The location and scale parameters are called "location shift" and "scaling" parameters in the tables in Part 2. An advantage of a location-scale family of distributions is that the fitted GAMLSS model is invariant to a location and scale change in the unit of measurement of the random variable (e.g. temperature from °F to °C) if an identity link is used for μ and a log link for σ, and the predictor models for μ and $\log \sigma$ include a constant term. The model is invariant in the sense that the fitted model for Y is the same as the fitted model for Y' when it is converted back to the alternative unit of measurement.

(For details about the exception, the exGAUS distribution, see Section 14.3.1. A fitted exGAUS model is also invariant, provided it also has a log link for ν and the model for ν has a constant term.)

Distribution	gamlss name	p	Parameter range μ	σ	ν	τ	skewness	kurtosis	page
exponential Gaussian	exGAUS	3	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	-	+ve	-	
exponential gen. beta t2	EGB2	4	$\mathbb{R}$	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	both	lepto	
generalised t	GT	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$	0	lepto	
Gumbel	GU	2	$\mathbb{R}$	$\mathbb{R}^+$	-	-	-ve	-	
Johnson's original SU	JSUo	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	lepto	
Johnson's SU repar.	JSU	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	lepto	
logistic	LO	2	$\mathbb{R}$	$\mathbb{R}^+$	-	-	0	lepto	
NET	NET	2	$\mathbb{R}$	$\mathbb{R}^+$	fixed	fixed	0	lepto	
normal	NO	2	$\mathbb{R}$	$\mathbb{R}^+$	-	-	0	meso	
normal family	NOF	3	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	-	0	meso	
power exponential	PE	3	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	-	0	both	
reverse Gumbel	RG	2	$\mathbb{R}$	$\mathbb{R}^+$	-	-	+ve	-	
sinh-arcsinh	SHASH	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$	both	both	
sinh-arcsinh original	SHASHo	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	both	
sinh-arcsinh original 2	SHASHo2	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	both	
skew t type 1	ST1	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	lepto	
skew t type 2	ST2	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	lepto	
skew t type 3	ST3	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$	both	lepto	
skew t type 3 repar	SST	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$(2,\infty)$	both	lepto	
skew t type 4	ST4	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$	both	lepto	
skew t type 5	ST5	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	lepto	
skew normal type 1	SN1	3	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	-	both	-	
skew normal type 2	SN2	3	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	-	both	-	
skew power exp. t1	SEP1	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	both	
skew power exp. t2	SEP2	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$	both	both	
skew power exp. t3	SEP3	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$	both	both	
skew power exp. t4	SEP4	4	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$	both	both	
t Family	TF	3	$\mathbb{R}$	$\mathbb{R}^+$	$\mathbb{R}^+$	-	0	lepto	
t Family repar	TF	3	$\mathbb{R}$	$\mathbb{R}^+$	$(2,\infty)$	-	0	lepto	

Table 4.1: Continuous distributions on $\mathbb{R}$ implemented within the `gamlss.dist` package.

4.2.1 Two-parameter distributions on $\mathbb{R}$

The following are the two-parameter continuous distributions on $\mathbb{R}$, in `gamlss.dist`: Gumbel (`GU`(μ, σ)); logistic (`LO`(μ, σ)); normal (`NO`(μ, σ), `NO2`(μ, σ); and reverse Gumbel (`RG`(μ, σ). Two-parameter distributions are only able to model the location μ and scale σ of the distribution independently, while skewness and/or kurtosis are defined implicitly from those two parameters. For example, the normal distribution `NO`(μ, σ) has location parameter μ (its mean) and scale parameter σ (its standard deviation). The normal is a symmetric mesokurtic distribution, hence its (moment based) and excess kurtosis are fixed at zero and cannot be modelled. The logistic distribution (`LO`(μ, σ)) distribution is symmetric but has a higher kurtosis than the normal. For the `GU`(μ, σ) and `RG`$(\mu, \sigma$ distributions, the (moment based) skewness is a fixed negative and positive value respectively, and cannot be modelled. The `GU`(μ, σ) distribution is negatively skewed, while `RG`(μ, σ) is positively skewed. Figure 4.1 compares the `NO`$(0, 1)$, `GU`$(0, 1)$ and `RG`$(0, 1)$ distributions.

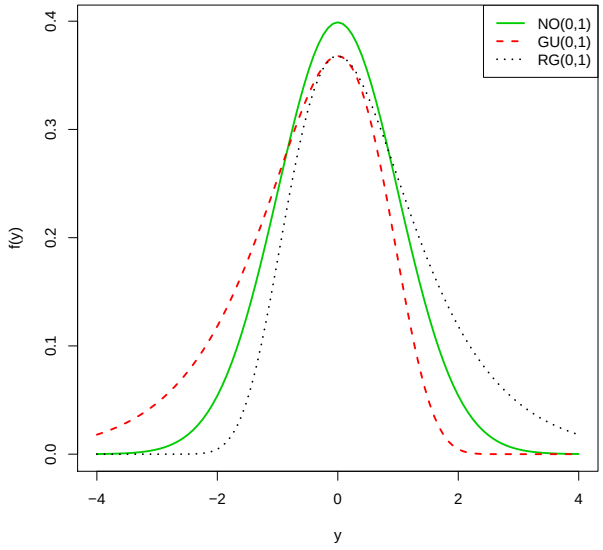


Figure 4.1: The `NO`(0,1), `GU`(0,1) and `RG`(0,1) distributions

4.2.2 Three-parameter distributions on $\mathbb{R}$

The following are the three-parameter continuous distributions on $\mathbb{R}$, in `gamlss.dist`: exponential Gaussian (`exGAUS`); normal family (`NOF`); power exponential (`PE`,

PE2); t family (TF, TF2); and skew normal (SN1, SN2). Three-parameter distributions are able to model either skewness or kurtosis in addition to location and scale. The skew normal type 1 and type 2 distributions and the exponential Gaussian distribution are able to model skewness. Figure 4.2 plots the $SN1(0, 1, \nu)$ distribution for $\nu = -100, -3, -1, 0, 1, 3, 100$. Changing the sign of ν reflects the distribution about $y = 0$.

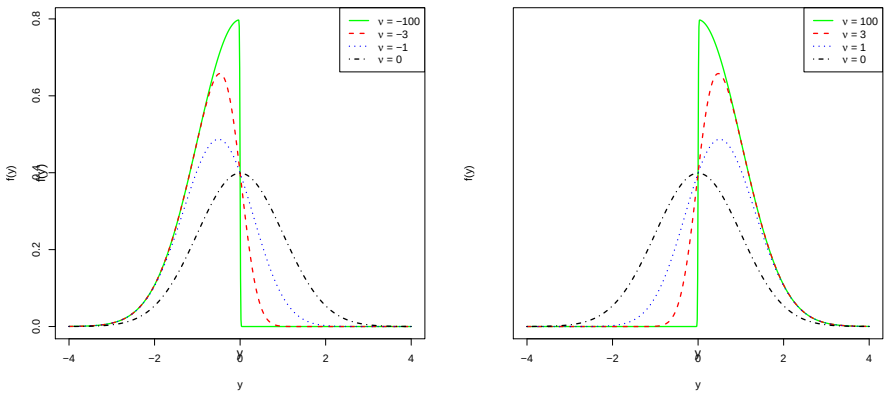


Figure 4.2: The skew normal type 1 distribution, $SN1(0, 1, \nu)$, for $\nu = 0, -1, -3, -100$ (left panel) and $\nu = 0, 1, 3, 100$ (right panel).

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 4.3 plots the power exponential distribution $\texttt{PE}(0, 1, \nu)$ for $\nu = 1, 2, 10, 1000$. The power exponential includes the Laplace (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 the power exponential can be more extremely kurtotic than the Laplace when $0 < \nu < 1$.

4.2.3 Four-parameter distributions on $\mathbb{R}$

The following are the four-parameter continuous distributions on $\mathbb{R}$, in gamlss.dist: exponential generalised beta type 2 (EGB2); generalised t (GT); Johnson's SU (JSU, JSUo); normal-exponential-t (NET); skew exponential power (SEP1 - SEP4); sinh-arcsinh (SHASH, SHASHo, SHASHo2); and skew t (ST1 - ST5).

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, which are both symmetric with ν and τ both modelling kurtosis. The EGB2, JSU, JSUo and ST1-ST5 are able to model skewness

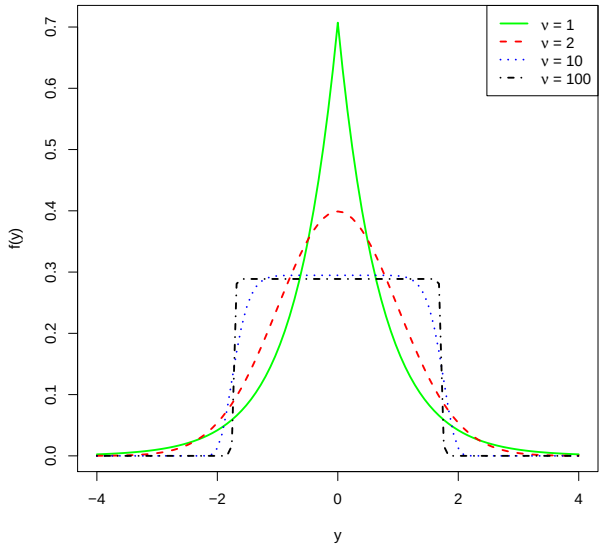


Figure 4.3: The power exponential distribution, PE$(0, 1, \nu)$, for ν=1, 2, 10 and 1000

and leptokurtosis. The `SHASH` and `SEP1-SEP4` distributions are able to model skewness and both leptokurtosis and platykurtosis.

Figure 4.4 shows `SEP1`$(0, 1, \nu, \tau)$ distributions. Changing the sign of ν reflects the distributions about the origin, changing the skewness from positive to negative. Setting $\nu = 0$ gives a symmetric distribution which is a reparametrised power exponential distribution.

4.3 Continuous distributions on $\mathbb{R}^+$

Table 4.2 gives the continuous distributions on $\mathbb{R}^+ = (0, \infty)$ in GAMLSS. Many of these are *scale* distributions with scale parameter μ (for fixed σ, ν and τ if the distribution has these parameters): `BCCG`, `BCCGo`, `BCPE`, `BCPEo`, `BCT`, `BCTo`,`EXP`, `GA`, `GB2`, `GG`, `GIG`, `IGAMMA`, `PARETO2`, `PARETO2o`, `WEI` and `WEI3`. For these distributions, if $Y \sim \mathcal{D}(\mu, \sigma, \nu, \tau)$ then $\varepsilon = (Y/\mu) \sim \mathcal{D}(1, \sigma, \nu, \tau)$ and $Y = \mu\varepsilon$ is a scaled version of ε. The distributions `IG`, `LOGNO`,and `WEI2` can be reparametrized to a scale family of distributions, but none of their parameters is a scale parameter.

The two-parameter distributions on $(0, \infty)$ are: gamma (`GA`); inverse gamma (`IGAMMA`); inverse Gaussian (`IG`); log-normal (`LOGNO`); and Weibull (`WEI`, `WEI2`, `WEI3`). These distributions are only able to model location and scale. The

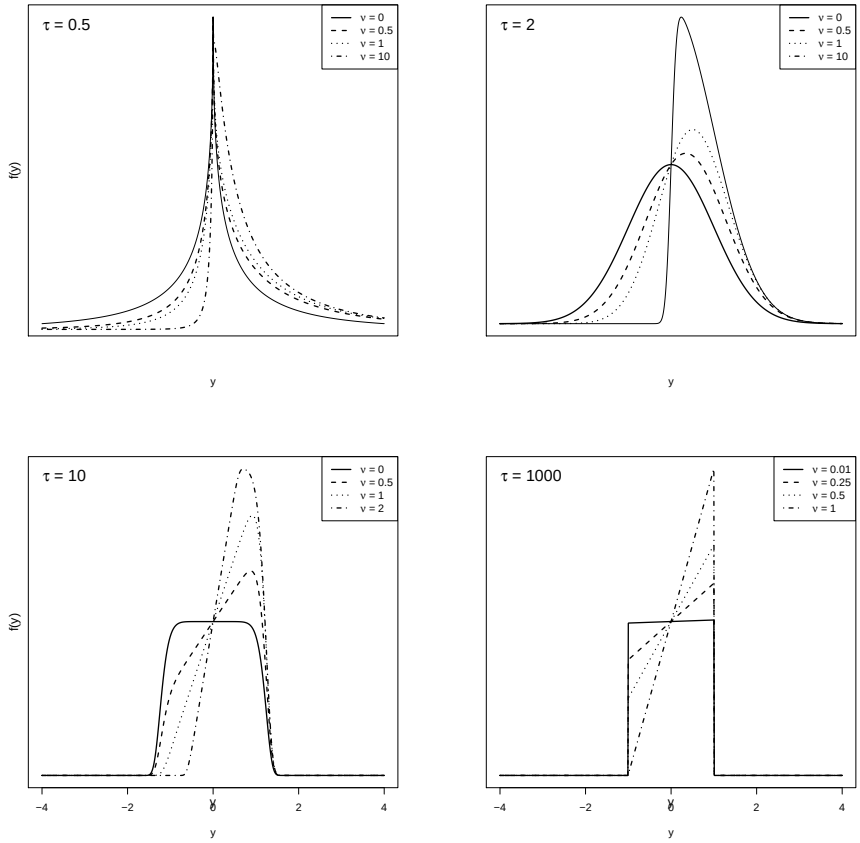


Figure 4.4: The skew exponential power type 1 distribution, $\texttt{SEP1}(0, 1, \nu, \tau)$

Distribution	gamlss name	p	Parameter range μ	σ	ν	τ	skewness	kurtosis	page
Box-Cox Cole-Green	BCCG	3	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}$	-			**??**
Box-Cox Cole-Green original	BCCGo								
Box-Cox power exp.	BCPE	4	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$			
Box-Cox power exp. original	BCPEo								
Box-Cox t	BCT	4	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}$	$\mathbb{R}^+$			
Box-Cox t original	BCTo								
exponential	EXP	1	$\mathbb{R}^+$	-	-	-			
gamma	GA	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
generalised beta type 2	GB2	4	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}^+$			
generalised gamma	GG	3	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}$	-			
generalised inv. Gaussian	GIG	3	$\mathbb{R}^+$	$\mathbb{R}^+$	$\mathbb{R}$	-			
inverse Gamma	IGAMMA	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
inverse Gaussian	IG	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
log normal	LOGNO	2	$\mathbb{R}$	$\mathbb{R}^+$	-	-			
log normal 2	LOGNO2	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
log normal (Box-Cox)	LNO	2	$\mathbb{R}$	$\mathbb{R}^+$	fixed	-			
Pareto 2	PARETO2	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
Pareto 2 original	PARETO2o	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
Pareto 2 repar	GP	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
Weibull	WEI	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
Weibull (μ the mean)	WEI3	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			
Weibull (PH)	WEI2	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-			

Table 4.2: Continuous distributions on $\mathbb{R}^+$ implemented within the `gamlss.dist` package.

R code on page 85

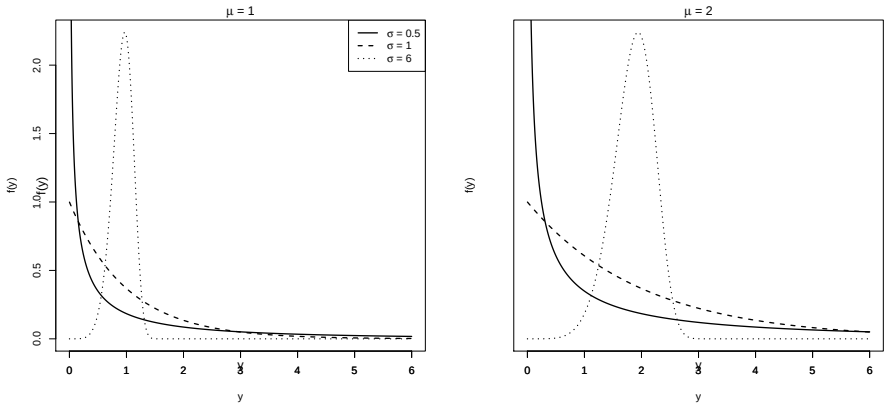
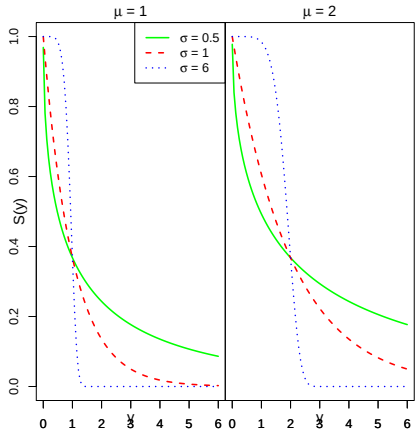


Figure 4.5: The Weibull distribution, WEI(μ, σ)

skewness and kurtosis of the distribution are determined by the location and scale.

The exponential (`EXP`), Weibull (`WEI`, `WEI2`, `WEI3`), and gamma (`GA`) distributions are widely used in survival and reliability analysis. Figure 4.5 shows the pdf's for the Weibull (`WEI`) distribution, for $\mu = 1, 2$ and $\sigma = 0.5, 1, 6$. The moment based skewness of this distribution is positive for $\sigma \leq 3.6023$ and negative otherwise. The survival and hazard functions, defined in Section 1.1.4, are important in survival and reliability analysis. These functions are shown in Figures 4.6 and 4.7, for the Weibull (WEI(μ, σ)) distribution. (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.)

The log-normal distribution is associated with geometric Brownian motion in finance.



R code on page 85

Figure 4.6: The WEI(μ, σ) survival function

4.3.1 Three-parameter distributions on $\mathbb{R}^+$

The three-parameter distributions on $(0, \infty)$ are: Box-Cox Cole and Green (`BCCG`); generalised gamma (`GG`, `GG2`); generalised inverse Gaussian (`GIG`); and log-normal family (`LNO`). Three-parameter distributions are able to model skewness in addition to location and scale.

The generalised gamma (`GG`) and generalised inverse Gaussian (`GIG`) distributions have been used in survival and reliability analysis. The LNO(μ, σ, ν) (`LNO`)

R code on page 85

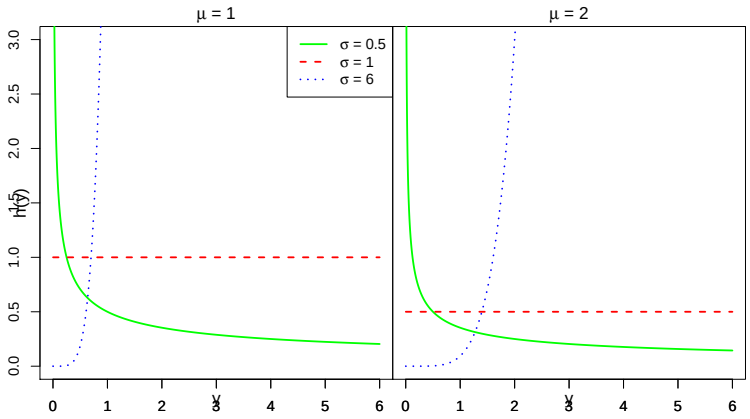


Figure 4.7: The WEI(μ, σ) hazard function

distribution is based on the standard Box-Cox transformation [**?**] which is defined as:

$$Z = \begin{cases} \frac{(Y^\nu - 1)}{\nu} & \text{if } \nu \neq 0 \\ \log(Y) & \text{if } \nu = 0 \end{cases} \tag{4.1}$$

and where the transformed variable Z is then assumed to have a truncated normal (NO(μ, σ)) distribution. The parameter ν (sometimes known as λ in the literature) is the skewness parameter. Note that $\nu = 0$ corresponds to the log-normal distribution and $\nu = 1$ to the normal distribution. The LNO(μ, σ, ν) is positively skewed for $\nu < 1$ and negatively skewed for $\nu > 1$. Figure 4.8 shows the `LNO(1, 1, ν)` distribution for $\nu = -1, 0, 1, 2$. A similar distribution is the `BCCG` distribution, which uses the Cole and Green [1992] transformation to normality (see section 15.4.1. The `BCCG` is used in the LMS method [Cole and Green, 1992], widely used in centile estimation.

Bob says why focus on LNO which is not a proper distribution? What should we do with this section?

The four-parameter distributions on $(0, \infty)$ are: Box-Cox power exponential (`BCPE`); Box-Cox power exponential (`BCT`); and generalised beta (`GB2`). 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 4.9 shows the BCT distribution for different values of each of the four parameters.

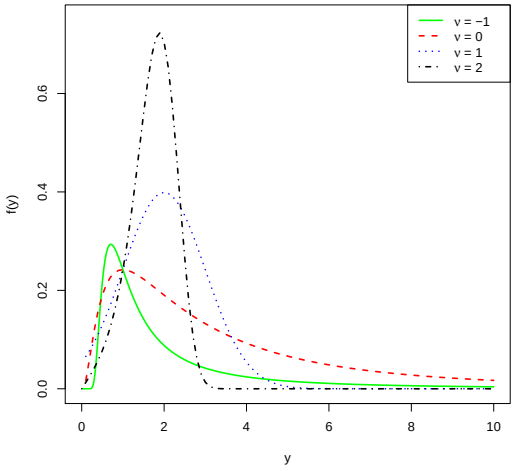


Figure 4.8: The log-normal family distribution, $LNO(1,1,\nu)$, for ν=-1, 0, 1, 2

Distribution	`gamlss` name	p	Parameter range μ	σ	ν	τ	skewness	kurtosis	page
beta	`BE`	2	$(0,1)$	$(0,1)$	-	-			
beta (original)	`BEo`	2	$\mathbb{R}^+$	$\mathbb{R}^+$	-	-	both	-	
generalised beta type 1	`GB1`	4	$(0,1)$	$(0,1)$	$\mathbb{R}^+$	$\mathbb{R}^+$			
logistic normal	`LOGITNO`	2	$(0,1)$	$\mathbb{R}^+$	-	-			

Table 4.3: Continuous distributions on $(0,1)$ implemented within the `gamlss.dist` package.

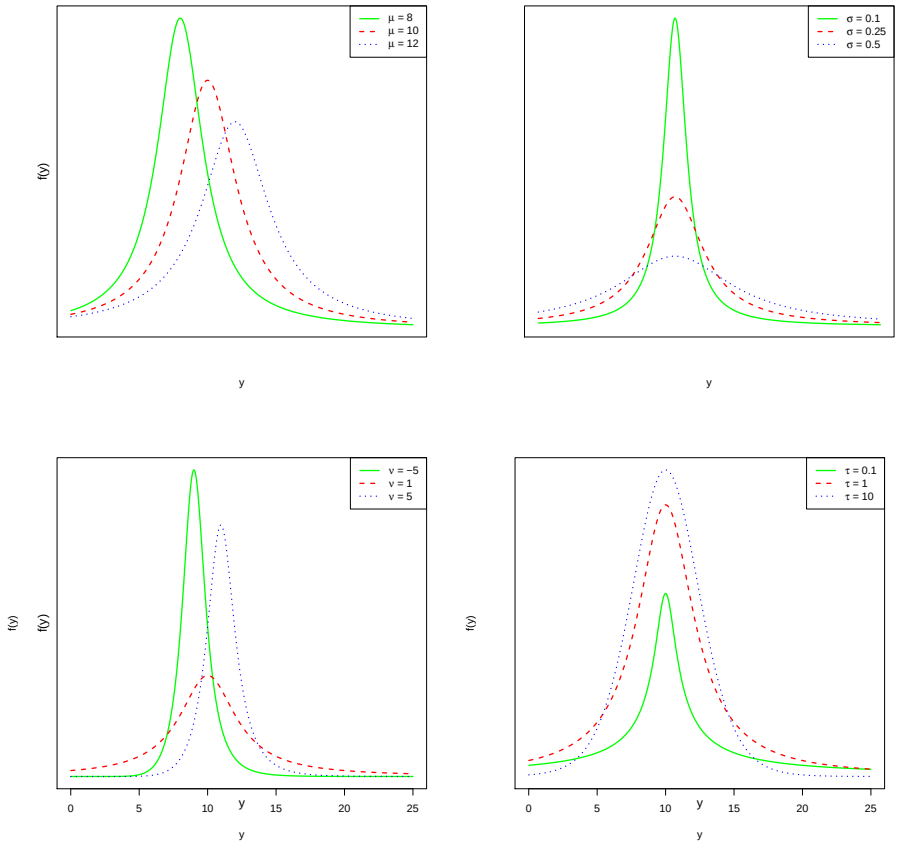


Figure 4.9: Box-Cox t Distribution $BCT(\mu, \sigma, \nu, \tau)$. Parameter values $\mu = 10$, $\sigma = 0.25$, $\nu = 1$, $\tau = 1$ (where not varying).

4.4 Continuous distributions on $(0,1)$

Table 4.3 gives the continuous distributions on $(0,1)$ in GAMLSS. 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 $\mathbb{R}$ can be transformed to a distribution on $(0,1)$, see section 11.5 and the example in section **??**. Figure 4.10 shows the pdf of the beta distribution $\text{BE}(\mu,\sigma)$, which for different (μ,σ) combinations can be unimodal, U-shaped or J-shaped. For $\mu = 0.5$ the distribution is symmetric; for $0 < \mu < 0.5$ it is positively skewed; and for $0.5 < \mu < 1$ it is negatively skewed.

Examples to be added

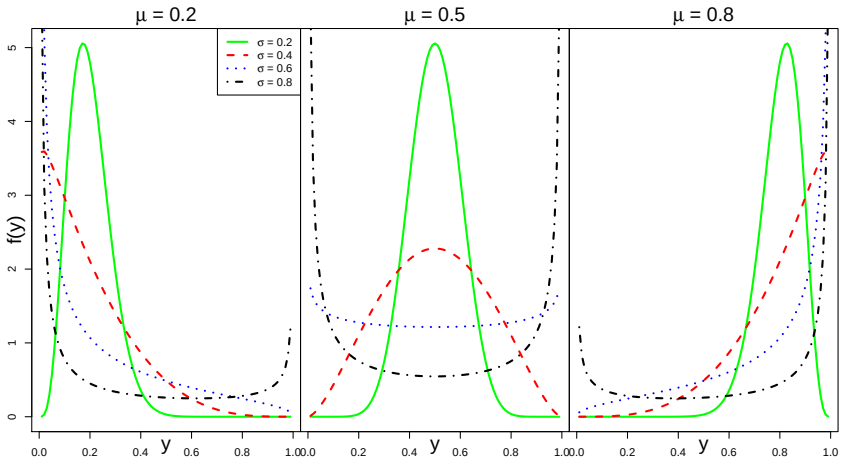


Figure 4.10: The beta distribution $BE(\mu,\sigma)$

4.5 Comparison of properties of continuous distributions

The choice of distribution for a particular response variable Y is based on how well it fits the data as judged by residual diagnostics, discussed in Chapter **??**. The fitted global deviance GDEV $= -2\log\hat{\ell}$, where $\hat{\ell} = \sum_{i=1}^{n} \log f(y_i \,|\, \hat{\theta})$ is the fitted likelihood function, and tests and information criteria (e.g. AIC or SBC) based on GDEV are useful for model comparisons. These measures are discussed fully in Chapter 9.3.

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 a simple explicit[1] formula for the mean, median or mode of Y may be desirable in a particular application. In addition to the range of the distribution, the following are properties of the distribution that may be relevant in choosing the model distribution:

- Explicit probability density function, cumulative distribution function and inverse cumulative distribution function;
- Explicit moment based measures of location, scale, skewness and kurtosis (i.e. population mean, standard deviation, γ_1, γ_2 respectively), see section 2.2;
- Explicit centiles and centile based measures of location, scale, skewness and kurtosis (i.e. median, semi-interquartile range, γ, $st_{0.49}$ repectively), see section 2.3;
- Continuity of $f(y \mid \mu, \sigma, \nu, \tau)$ and its derivatives with respect to y ;
- Continuity of the derivatives of $f(y \mid \mu, \sigma, \nu, \tau)$ with respect to μ, σ, ν and τ ;
- Flexibility in modelling skewness and kurtosis.

Many of the distributions of Tables 4.1, 4.2 and 4.3 can be generated by one (or more) of the methods described in Chapter 11, including univariate transformation, Azzalini type methods and splicing. 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 is that their 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 sometimes a lack of continuity of the second derivatives of the probability density function with respect to y and μ at the splicing point. However their mode and moment based measures are explicit and they can be flexible in modelling skewness and kurtosis.

[1] The term explicit indicates that the particular function or measure can be obtained using closed-form mathematical functions, i.e. not requiring numerical integration or numerical solution.

4.6 R code

R code for the pdf, survival function and hazard functions of the WEI(μ, σ) distribution, for $\mu = 1$ and $\sigma = 0.5, 1, 6$, is shown below.

Figure 4.5

```
cols=c("green","red","blue","black")
mu=1
sigma = c(0.5,1,6)
### Weibull pdf
curve(dWEI(x, mu=mu, sigma=sigma[3]),0.001,6, xlab="y", ylab="f(y)", type="n")
for(i in 1:length(sigma))
  curve(dWEI(x, mu=mu, sigma=sigma[i]),0.001,6,col=cols[i],
        lty=i,lwd=2,add=TRUE)

### Legend and title
lgnd <- c(bquote(paste(sigma," = ",.(sigma[1]))),
                          bquote(paste(sigma," = ",.(sigma[2]))),
                          bquote(paste(sigma," = ",.(sigma[3])))
                          )
legend("topright",legend=as.expression(lgnd),
       lty=1:length(sigma),col=cols[1:length(sigma)],lwd=2 )
title(bquote(paste(mu," = ",.(mu))))
```

Figure 4.6

```
### Weibull survival function
curve(pWEI(x, mu=mu, sigma=sigma[1],lower.tail=FALSE),0.001,6, ylim=c(0,1),
      xlab="y", ylab="f(y)", type="n")
for(i in 1:length(sigma))
  curve(pWEI(x, mu=mu, sigma=sigma[i],lower.tail=FALSE),0.001,6,col=cols[i],
        lty=i,lwd=2,add=TRUE)
```

A hazard function for WEI is generated and saved under the name: hWEI

I haven't been able to suppress this message given by gen.hazard - can you fix this?

Figure 4.7

```
### Weibull hazard function
gen.hazard("WEI")
curve(hWEI(x, mu=mu, sigma=sigma[1]),0.001,6, ylim=c(0,3),
      xlab="y", ylab="f(y)", type="n")
for(i in 1:length(sigma))
  curve(hWEI(x, mu=mu, sigma=sigma[i]),0.001,6,col=cols[i],
        lty=i,lwd=2,add=TRUE)
```

Exercises

Q1: Reproduce Figure 4.4.

The code is given in the text

Chapter 5

Discrete distributions for count data

This chapter provides explanation for:

1. different types of count data distributions within the GAMLSS family
2. how those distributions model overdispertion, zero inflations and heavy tails.

Mikis: Needs examples to show why those distribution are helpfull

5.1 Introduction

This chapter deals with discrete distributions for an unlimited count variable Y having range $R_Y = \{0, 1, 2, 3, \ldots, \infty\}$. There are discrete distributions with range $R_Y = \{1, 2, 3, \ldots, \infty\}$, e.g. the logarithmic distribution (`LG`), see Section 18.1.2, but these are not dealt with in this chapter. Discrete distributions with limited range $R_Y = \{0, 1, 2, 3, \ldots, N\}$, for known upper limit N, are dealt with in Chapter 6

5.1.1 Poisson distribution

For a count variable Y with range $R_Y = \{0, 1, 2, 3, \ldots, \infty\}$, the classical model for this variable is the Poisson distribution, which was first published by Poisson [1837] and has had a major influence on discrete distribution modelling for almost two centuries.

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

$$P(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \tag{5.1}$$

Mikis: we have the equation and plots of Poisson in the first Chapter

where $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$. However the Poisson distribution, $Y \sim \texttt{PO}(\mu)$ has only one parameter μ, which equals the mean or expected value, $E(Y)$, of Y. The variance $Var(Y) = \mu$, and moment based skewness $\sqrt{\beta_1} = \mu^{-\frac{1}{2}}$ and kurtosis $\beta_2 = 3+\mu^{-1}$ of the Poisson distribution all depend on μ, and so cannot be modelled independently of μ. The variance of the Poisson distribution equals the mean, i.e. $Var(Y) = E(Y)$, while the skewness $(\sqrt{\beta_1})$ and excess kurtosis $(\beta_2 - 3)$ both tend to zero as μ increases to infinity.

5.1.2 Limitations of the Poisson distribution

There are four major problems often encountered when modelling count data using the Poisson distribution.

- overdispersion (and underdispersion), dealt with in Section 5.2,
- long right tail (i.e. high positive skewness), dealt with in Sections 5.2 and 5.4,
- excess (or shortage) of zero values, dealt with in Section 5.3,
- variance-mean relationship, dealt with in Section 5.5.

Overdispersion is usually 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 variance of a Poisson distributed random variable is equal to its mean. That is, $Var(Y) = E(Y) = \mu$. Unfortunately very often for a count response variable $Var(Y) > E(Y)$, i.e. the distribution is overdispersed relative to a Poisson distribution. A less common problem occurs when $Var(Y) < E(Y)$, i.e. the distribution is underdispersed relative to a Poisson distribution.

The problem of *excess (or shortage) of zero values* occurs when the response variable has a higher (or lower) probability of a zero value than 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 fourth problem is the variance-mean relationship. For the Poisson distribution the variance equals (and hence increases with) the mean. However the variance may increase with a power of the mean.

5.1.3 Discrete distributions in gamlss

Discrete distributions in **gamlss** for an unlimited count variable Y have range $R_Y = (0, 1, 2, 3, \ldots, \infty)$, except for the logarithmic distribution (`LG`) which has range $R_Y = (1, 2, 3, \ldots, \infty)$. The details of the distributions are given in Part II of this book.

One parameter discrete count distributions in **gamlss**, the geometric, logarithmic, Poisson and Yule distributions, (i.e. `GEOM`, `LG`, `PO` and `YULE` rspectively) are only able to model the location of the distribution.

Two parameter discrete count distributions in **gamlss** are able to model the location and scale of the distribution (the negative binomial type I and type II, Poisson-inverse Gaussian and Waring distributions, i.e. `NBI`, `NBII`, `PIG` and `WARING` 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 and positive skewness (i.e. heavy 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 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).

Modelling count data in **gamlss** is dealt with in Section 5.2.3

5.2 Overdispersion and underdispersion

5.2.1 Introduction

Overdispersion and underdispersion (relative to a Poisson distribution) has been recognised for a long time as a potential problem within the distributions literature [Haight, 1967] and the literature of generalised linear models, Nelder and Wedderburn [1972]. Over the years several solutions to the problem of overdispersion have been suggested, see e.g. Consul [1989] and Dossou-Gbété and Mizère (2006).

Here we consider three approaches to modelling overdispersion:

(a) mixed distributions, dealt with in Sections 5.2.2 and 5.2.4,

(b) discretised continuous distributions, dealt with in Section 5.2.5,

(c) *ad-hoc* solutions, dealt with in Section 5.2.6.

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.

5.2.2 Mixed (or mixture) distributions

Mixed (or mixture) distributions account for overdispersion by assuming that the variable Y depends on a random variable γ whose distribution (apart from specific parameters) is known. [Note that γ can also be viewed as a random effect at the observation level,]This can also solve the problems of excess of zero values and long tails in the data. The methodology works like this.

Assume that the distribution of the response variable Y has a discrete probability function $P(Y = y|\gamma)$, conditional on a continuous random variable γ with probability density function $f_\gamma(\gamma)$. Then the marginal probability function of Y is given by

$$P(Y = y) = \int P(Y = y|\gamma) f_\gamma(\gamma) d\gamma. \tag{5.2}$$

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

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

$$P(Y = y) = \sum_{R_\gamma} P(Y = y|\gamma = \gamma_j) P(\gamma = \gamma_j). \tag{5.3}$$

where the summation is over $\gamma_j \epsilon R_\gamma$, where R_γ is the disrete range of γ. The resulting distribution of Y is called a discrete mixture of discrete distributions. When γ takes only a finite number of possible values, then the resulting distribution is called a finite mixture of discrete distributions.

Within the mixed (or mixture) distributions, category (a) in Section 5.2.1, we distinguish three different types:

(i) when $P(Y = y)$, the probability function of a continuous mixture of discrete distributions given by (5.2), exists explicitly. This is dealt with in Section 5.2.4,

(ii) when $P(Y = y)$, the probability function of a continuous mixture of discrete distributions given by (5.2), is not explicit, but is approximated by integrating out the variable γ using an approximation, e.g. Gaussian quadrature.

(iii) when $P(Y = y)$ is the probability function of a finite mixture of discrete distributions given by (5.3) where γ takes only a finite number of possible values.

5.2.3 Modelling count data in the gamlss package

The explicit distributions available in **gamlss** are given in Section 5.1.3. This includes explicit mixed Poisson distributions [category (a)(i) in Section 5.2.2] and also zero inflated and zero adjusted discrete count distributions in Section 5.3.

Models in category (a)(ii) in Section 5.2.2, in which the continuous mixture of discrete distributions is approximated by integrating out a normally distributed random variable using Gaussian quadrature, can be fitted using the package **gamlss.mx**, as can a finite mixture of discrete distributions in category (a)(iii). In fact package **gamlss.mx** fits type (a)(ii) and (a)(iii) models allowing a more general conditional distribution $P(Y = y|\gamma)$ to be used in equation (5.3) 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 ???.

Discretised continuous distributions in category (b) in Section 5.2.1 can be fitted in **gamlss** using the add-on package **gamlss.cens**. An example of this is given in Section ???.

The double Poisson distribution, a member of category (c) in Section 5.2.1, is an explicit distribution in **gamlss**.

5.2.4 Explicit continuous mixtures of Poisson distributions

These are discrete distributions in category (a)(i) in Section 5.2.2. Suppose, given that a continuous random variable γ takes the value γ, Y has a Poisson distribution with mean $\mu\gamma$, i.e. $Y|\gamma \sim PO(\mu\gamma)$, where $\mu > 0$, and suppose that γ has probability density function $f_\gamma(\gamma)$ defined on $(0, \infty)$, then the (marginal) distribution of Y is a continuously mixed Poisson distribution. Provided random variable γ has mean 1, then Y has mean μ. [The model can be considered as a multiplicative Poisson random effect model, provided the distribution of γ does not depend on μ.]

For example suppose $Y|\gamma \sim \texttt{PO}(\mu\gamma)$ where $\gamma \sim \texttt{GA}(1, \sigma^{1/2})$. The marginal distribution is $Y \sim \texttt{NBI}(\mu, \sigma)$.

To show this result

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

where

$$f_\gamma(\gamma) = \frac{\gamma^{1/\sigma-1}\exp\left(-\gamma/\sigma\right)}{\sigma^{(1/\sigma)}\Gamma(1/\sigma)} \ , \quad \gamma > 0 \ .$$

then equation (5.2) gives:

$$P(Y=y) = \int_0^\infty \frac{e^{-\mu\gamma}(\mu\gamma)^y}{y!} \cdot \frac{\gamma^{1/\sigma-1}\exp\left(-\gamma/\sigma\right)}{\sigma^{(1/\sigma)}\Gamma(1/\sigma)} \ d\gamma$$

$$= \frac{\Gamma(y+\frac{1}{\sigma})}{\Gamma(\frac{1}{\sigma})\Gamma(y+1)} \left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y \left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma} \tag{5.4}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$ and $\sigma > 0$, which is the probability function of the negative binomial type I, $\texttt{NBI}(\mu, \sigma)$, distribution. This is a genesis of the negative binomial distribution which is quite distinct from its classical development as the distribution of the number of failures until the k^{th} success in independent Bernoulli trials.

By using different distributions for γ, we can generate a variety of new (mixed Poisson) count distributions. Table 5.1 shows the mixed Poisson distributions currently available in **gamlss** package, together with their R name within **gamlss**, their corresponding mixing distributions for γ, where $Y|\gamma \sim \texttt{PO}(\mu\gamma)$, and their mean $E(Y)$ and variance $Var(Y)$.

The probability functions for all the distributions in Table 5.1 are given in the Part II of this book, [except for the shifted gamma (`SG`) distribution which is defined later in this Section].

Many previous 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.

Many of the mixed Poisson distribution given in Table 5.1 have mean equal to parameter μ. This allows easier interpretation of models for μ.

Specifically the following distributions with mean exactly equal to μ are considered below: the negative binomial type I and type II, Poisson-inverse Gaussian (PIG), Sichel and Delaporte distributions.

Distribution	**gamlss** name	Mixing distribution	$E(Y)$	$Var(Y)$
Delaporte	$\texttt{DEL}(\mu,\sigma,\nu)$	$\texttt{SG}(1,\sigma^{\frac{1}{2}},\nu)$	μ	$\mu+\sigma(1-\nu)^2\mu^2$
NB type I	$\texttt{NBI}(\mu,\sigma)$	$\texttt{GA}(1,\sigma^{\frac{1}{2}})$	μ	$\mu+\sigma\mu^2$
NB type II	$\texttt{NBII}(\mu,\sigma)$	$\texttt{GA}(1,\sigma^{\frac{1}{2}}\mu^{-\frac{1}{2}})$	μ	$\mu+\sigma\mu$
NB family	$\texttt{NBF}(\mu,\sigma,\nu)$	$\texttt{GA}(1,\sigma^{\frac{1}{2}}\mu^{\frac{\nu}{2}-1})$	μ	$\mu+\sigma\mu^\nu$
Poisson	$\texttt{PO}(\mu)$	-	μ	μ

PIG	$\texttt{PIG}(\mu,\sigma)$	$\texttt{IG}(1,\sigma^{\frac{1}{2}})$	μ	$\mu+\sigma\mu^2$
Sichel	$\texttt{SICHEL}(\mu,\sigma,\nu)$	$\texttt{GIG}(1,\sigma^{\frac{1}{2}},\nu)$	μ	$\mu+h(\sigma,\nu)\mu^2$
ZI NB	$\texttt{ZINBI}(\mu,\sigma,\nu)$	$\texttt{ZAGA}(1,\sigma^{\frac{1}{2}},\nu)$	$(1-\nu)\mu$	$(1-\nu)\left[\mu+(\sigma+\nu)\mu^2\right]$
ZI Poisson	$\texttt{ZIP}(\mu,\sigma)$	$\texttt{BI}(1,1-\sigma)$	$(1-\sigma)\mu$	$(1-\sigma)(\mu+\sigma\mu^2)$
ZI Poisson 2	$\texttt{ZIP2}(\mu,\sigma)$	$(1-\sigma)^{-1}\texttt{BI}(1,1-\sigma)$	μ	$\mu+\frac{\sigma}{(1-\sigma)}\mu^2$
ZI PIG	$\texttt{ZIPIG}(\mu,\sigma,\nu)$	$\texttt{ZAIG}(1,\sigma^{\frac{1}{2}},\nu)$	$(1-\nu)\mu$	$(1-\nu)\left[\mu+(\sigma+\nu)\mu^2\right]$

Table 5.1: Mixed Poisson distributions implemented in the **gamlss** package. ZI = zero-inflated; NB = negative binomial; PIG = Poisson-inverse Gaussian; $h(\sigma,\nu)=2\sigma(\nu+1)/c+1/c^2-1$, where c is defined below equation (**??**).

Negative binomial The negative binomial type I distribution, denoted $\texttt{NBI}(\mu,\sigma)$, is a continuously mixed Poisson distribution obtained as the marginal distribution of Y when $Y|\gamma\sim\texttt{PO}(\mu\gamma)$ and $\gamma\sim\texttt{GA}(1,\sigma^{1/2})$, i.e. γ has a gamma distribution with mean 1 and scale parameter $\sigma^{1/2}$.

The probability function of the negative binomial type I distribution, denoted $\texttt{NBI}(\mu,\sigma)$, is $P(Y=y|\mu,\sigma)$ given by equation (5.4).

The mean and variance of Y are given by $E(Y)=\mu$ and $Var(Y)=\mu+\sigma\mu^2$.

Figure **??** plots the negative binomial type I distribution, $\texttt{NBI}(\mu,\sigma)$, for $\mu=5$ and $\sigma=(0.01,0.5,1,2)$. 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 $\sigma=0.01$ is close to a Poisson, $\texttt{PO}(5)$, distribution which corresponds to $\mu=5$ and $\sigma\to 0$ in the $\texttt{NBI}(\mu,\sigma)$ distribution.

The negative binomial type II distribution, denoted $\texttt{NBII}(\mu,\sigma)$, is a mixed Poisson distribution obtained as the marginal distribution of Y when $Y|\gamma\sim\texttt{PO}(\mu\gamma)$ and $\gamma\sim\texttt{GA}(1,\sigma^{\frac{1}{2}}\mu^{-\frac{1}{2}})$. 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 $\texttt{NBII}(\mu,\sigma)$, is given by

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

for $y=0,1,2,3,\ldots$, where $\mu>0$ and $\sigma>0$. The mean and variance of Y are given by $E(Y)=\mu$ and $Var(Y)=(1+\sigma)\mu$.

The NBI and NBII models differ when there are explanatory variables for μ and/or σ. The extra σ 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 μ and even closer as μ increases.

The negative binomial family distribution, denoted $\texttt{NBF}(\mu, \sigma, \nu)$ is obtained by replacing σ by $\sigma\mu^{\nu-2}$ in the NBI distribution. This distribution has mean μ and variance $\sigma\mu^{\nu}$ and is dealt with in Section 5.5.

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

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

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

where $\alpha^2 = \frac{1}{\sigma^2} + \frac{2\mu}{\sigma}$, for $y = 0, 1, 2, ...$, where $\mu > 0$ and $\sigma > 0$ and

$$K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\{-\frac{1}{2}t(x + x^{-1})\}dx$$

is the modified Bessel function of the third kind. The mean and variance of Y are given by $E(Y) = \mu$ and $Var(Y) = \mu + \sigma\mu^2$.

Figure **??** plots $\texttt{PIG}(\mu, \sigma)$ for various (μ, σ) combinations. Note that the probability functions for $\sigma = 0.01$ are close to the Poisson distributions with the same means.

Sichel 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). This distribution is also known as the generalized inverse Gaussian Poisson (GIGP) distribution.

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

$$f_\gamma(\gamma) = \frac{c^\nu \gamma^{\nu-1}}{2K_\nu\left(\frac{1}{\sigma}\right)} \exp\left[-\frac{1}{2\sigma}\left(c\gamma + \frac{1}{c\gamma}\right)\right] \tag{5.5}$$

for $\gamma > 0$, where $\sigma > 0$ and $-\infty < \nu < \infty$, and $c = R_\nu(1/\sigma)$, $R_\lambda(t) = K_{\lambda+1}(t)/K_\lambda(t)$ and $K_\lambda(t)$ is the modified Bessel function of the third kind. The parameterization (5.5) of the GIG ensures that $E[\gamma] = 1$, which results in $E(Y) = \mu$.

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

$$P(Y=y|\mu,\sigma,\nu)=\frac{(\mu/c)^y K_{y+\nu}(\alpha)}{y!\,(\alpha\sigma)^{y+\nu}\,K_\nu\left(\frac{1}{\sigma}\right)} \tag{5.6}$$

for $y=0,1,2,3,\ldots$, where $\alpha^2=\sigma^{-2}+2\mu(c\sigma)^{-1}$.

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.

In the parametrization above μ is the mean of the Sichel distribution, while the two remaining parameters σ and ν 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. An alternative parameterisation of the Sichel distribution, denoted $\texttt{SI}(\mu,\sigma,\nu)$, is given in Section ??.

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

$$f_\gamma(\gamma)=\frac{(\gamma-\nu)^{\frac{1}{\sigma}-1}}{\sigma^{1/\sigma}(1-\nu)^{1/\sigma}\Gamma(1/\sigma)}\exp\left[-\frac{(\gamma-\nu)}{\sigma(1-\nu)}\right] \tag{5.7}$$

for $\gamma>\nu$, where $\sigma>0$ and $0\leq\nu<1$. This parameterization ensures that $E[\gamma]=1$, which results in $E(Y)=\mu$. Note that $\gamma=\nu+(1-\nu)Z$ where $Z\sim\texttt{GA}(1,\sigma^{\frac{1}{2}})$ and so γ has a lower bound of ν.

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

$$P(Y=y|\mu,\sigma,\nu)=\frac{e^{-\mu\nu}}{\Gamma(1/\sigma)}\left[1+\mu\sigma(1-\nu)\right]^{-1/\sigma}S \tag{5.8}$$

where

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

for $y=0,1,2,3,\ldots$, where $\mu>0$, $\sigma>0$ and $0\leq\nu<1$. The mean of Y is given by $E(Y)=\mu$ and the variance by $Var(Y)=\mu+\mu^2\sigma\,(1-\nu)^2$.

5.2.5 Discretised continuous distributions method

By discretised continuous distributions, category (b) solutions in Section 5.2.1, we refer to methods which use continuous distributions to create a discrete one. For example, let $F_W(w)$ be the cumulative distribution function of a continuous random variable W defined on $(0, \infty)$ 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 $P(Y = 0) = F_W(.5)$ 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 **g**amlss.cens package. One potential criticism of the above methods of generating discrete distributions is the fact that if the parameter μ_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 μ_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.

5.2.6 *Ad-hoc* methods

We refer to *ad-hoc* solutions, i.e category (c) in Section 5.2.1, 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 difficultly 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, miss-

ing 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 [1999] for a similar criticism of the *ad-hoc* methods.

The double Poisson distribution, a member of the EDE family, is implemented in **gamlss** as the `DPO`(μ, σ) distribution, a proper distribution summing to one, with probability function given by

$$P(Y = y|\mu, \sigma) = c(\mu, \sigma)\sigma^{-1/2}e^{-\mu/\sigma}\left(\frac{\mu}{y}\right)^{y/\sigma}\frac{e^{y/\sigma - y}y^y}{\Gamma(y+1)} \tag{5.9}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$ and $\sigma > 0$ and $c(\mu, \sigma)$ is a function of μ and σ defined to ensure that the discrete distribution is a proper distribution i.e. $\Sigma_{y=0}^{\infty} P(Y = y|\mu, \sigma) = 1$.

This distribution can model underdispersion (if $\sigma < 1$) as well as overdispersion (if $\sigma > 1$) relative to a Poisson distribution (if $\sigma = 1$).

5.3 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 5.3.1. A solution to excess and/or shortage of zero values in a particular discrete distribution is a zero adjusted (or altered) discrete distribution, dealt with in Section 5.3.2.

5.3.1 Zero inflated discrete distributions

A discrete count response variable Y can exhibit a greater probability of value zero than that of a particular discrete count distribution with range $R = (0, 1, 2, 3, \ldots, \infty)$, 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 discrete count 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

$$\begin{aligned} P(Y=0) &= p+(1-p)P(Y_1=0) \\ P(Y=y) &= (1-p)P(Y_1=y) \text{ if } y=1,2,3,\ldots \end{aligned} \tag{5.10}$$

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

Let the mean, variance and third and fourth cental moments of Y and Y_1 be denoted by $\mathrm{E}(Y)=\mu_{1Y}^{'}$, $\mathrm{Var}(Y)=\mu_{2Y}$, μ_{3Y}, μ_{4Y} and $\mathrm{E}(Y_1)=\mu_1^{'}$, $\mathrm{Var}(Y_1)=\mu_2$, μ_3, μ_4 respectively. Let $\mu_{rY}^{'}=\mathrm{E}(Y^r)$ and $\mu_r^{'}=\mathrm{E}(Y_1^r)$ and $\mu_{rY}=\mathrm{E}\{[Y-\mathrm{E}(Y)]^r\}$ and $\mu_r=\mathrm{E}\{[Y_1-\mathrm{E}(Y_1)]^r\}$.

Note that $\mathrm{E}(Y^r)=(1-p)\mathrm{E}(Y_1^r)$, i.e. $\mu_{rY}^{'}=(1-p)\mu_r^{'}$. Hence using equations (2.1) and (2.2), the mean, variance and third and fourth central moments aof Y are given by

$$\begin{aligned} \mu_{1Y}^{'} &= \mathrm{E}(Y)=(1-p)\mathrm{E}(Y_1)=(1-p)\mu_1^{'} \\ \mu_{2Y} &= \mathrm{Var}(Y)=(1-p)\mathrm{Var}(Y_1)+p(1-p)\left[\mathrm{E}(Y_1)\right]^2 \\ \mu_{3Y} &= (1-p)\left[\mu_3+3p\mu_2\mu_1^{'}+p(2p-1)\mu_1^{'3}\right] \\ \mu_{4Y} &= (1-p)\left[\mu_4+4p\mu_3\mu_1^{'}+6p^2\mu_2\mu_1^{'2}+p(1-3p+3p^2)\mu_1^{'4}\right]. \end{aligned} \tag{5.11}$$

where $\mu^{'}=\mathrm{E}(Y_1)$.

The cumulative distribution function (cdf) of $Y \sim$ `ZID` is given by

$$P(Y \le y)=p+(1-p)P(Y_1 \le y) \tag{5.12}$$

for $y=0,1,2,3,\ldots$.

The probability generating function (pgf) of $Y \sim$ `ZID` is given by

$$G_Y(t)=p+(1-p)G_{Y_1}(t) \tag{5.13}$$

where $G_{Y_1}(t)$ is the pgf of Y_1.

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

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

Hence the probability function (pf) of $Y \sim \texttt{ZIP}(\mu, \sigma)$ 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 $y = 1, 2, 3, \ldots$, where $\mu > 0$ and $0 < \sigma < 1$.

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 $Y|\gamma \sim \texttt{PO}(\mu\gamma)$ and $\gamma \sim \texttt{BI}(1, 1 - \sigma)$, i.e. $\gamma = 0$ with probability σ and $\gamma = 1$ with probability $1 - \sigma$. When $\gamma = 1$, $Y \sim \texttt{PO}(\mu)$, while when $\gamma = 0$, Y takes the value 0 with probability 1, since

$$P(Y = y|\gamma = 0) = \frac{e^{-0}0^y}{y!} = \begin{cases} 1, & y = 0 \\ 0, & y = 1, 2, 3, \ldots \end{cases}$$

as $0! = 1$, $0^0 = 1$ and $0^y = 0$, for $y = 1, 2, 3, \ldots$. Note however that γ has mean $1 - \sigma$ in this formulation and Y has mean $(1 - \sigma)\mu$.

Zero inflated Poisson type 2 parameterization A different parameterization of the zero inflated poisson distribution, denoted $\texttt{ZIP2}(\mu, \sigma)$, has pf given by

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

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

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

The mean of Y 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. $Y \sim \texttt{ZIP2}(\mu, \sigma)$, is the marginal distribution for Y where $Y|\gamma \sim \texttt{PO}(\mu\gamma)$ and $\gamma \sim (1 - \sigma)^{-1}\texttt{BI}(1, 1 - \sigma)$. Hence γ has mean 1 and hence Y has mean μ.

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

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

Hence the pf of $Y \sim \texttt{ZINBI}(\mu, \sigma, \nu)$ 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 $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$ and $Y_1 \sim \texttt{NBI}(\mu, \sigma)$.

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

Further zero inflated distributions

Other zero inflated distributions available in **gamlss.dist** are:

`ZIPIG`(μ, σ, ν)**:** zero inflated `PIG`,

`ZIBNB`(μ, σ, ν, τ)**:** zero inflated `BNB`,

`ZISICHEL`(μ, σ, ν, τ)**:** zero inflated `SICHEL`.

5.3.2 Zero adjusted (or altered) discrete distributions

A discrete count response variable Y can exhibit either a greater or less probability of value zero than that of a particular discrete count distribution with range $R = (0, 1, 2, 3, \ldots, \infty)$, denoted here by D.

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

Hence $Y = 0$ with probability p and $Y = Y_0$ with probability $(1 - p)$, where $Y_0 \sim \texttt{Dtr}$ and $0 < p < 1$.

Hence

$$P(Y = 0) = p$$

and

$$P(Y = y) = (1 - p)P(Y_0 = y)$$

if $y = 1, 2, 3, \ldots$, where $Y_0 \sim \texttt{Dtr}$.

Hence

$$P(Y = 0) = p$$

and

$$P(Y = y) = (1 - p)P(Y_0 = y) = \frac{(1 - p)P(Y_1 = y)}{1 - P(Y_1 = 0)} \tag{5.14}$$

if $y = 1, 2, 3, \ldots$, where $Y_1 \sim$ D.

Note that the probability that $Y = 0$ is exactly p, where $0 < p < 1$, so $P(Y = 0)$ can be greater than or less than $P(Y_1 = 0)$ and hence the distribution is called a 'zero adjusted' (or 'zero altered') distribution. If $p = P(Y_1 = 0)$ in (5.14) then $Y = Y_1 \sim$ D and ZAD becomes distribution D. If $p > P(Y_1 = 0)$ in (5.14) then the zero altered distribution ZAD has an inflated probability at 0 relative to D and is a reparameterization of the zero inflated distribution ZID in Section 5.3.1 (but because of the reparameterization, these models are, in general, different when the parameters are modelled by explanatory variables). However if $p < P(Y_1 = 0)$ in (5.14) then ZAD has a deflated probability at 0 and is different from (i.e. not a reparameterization of) ZID. Hence ZID is a reparameterized submodel of ZAD.

Let the mean, variance and third and fourth central momments of Y and Y_1 be denoted by $\mathrm{E}(Y) = \mu_{1Y}^{'}$, $\mathrm{Var}(Y) = \mu_{2Y}$, μ_{3Y}, μ_{4Y} and $\mathrm{E}(Y_1) = \mu_1^{'}$, $\mathrm{Var}(Y_1) = \mu_2$, μ_3, μ_4 respectively. Let $\mu_{tY}^{'} = \mathrm{E}(Y^r)$ and $\mu_r^{'} = \mathrm{E}(Y_1^r)$ and $\mu_{rY} = \mathrm{E}\{[Y - \mathrm{E}(Y)]^r\}$ and $\mu_r = \mathrm{E}\{[Y_1 - \mathrm{E}(Y_1)]^r\}$.

Note that $\mathrm{E}(Y^r) = c\mathrm{E}(Y_1^r)$, i.e. $\mu_{rY}^{'} = c\mu_r^{'}$, where $c = (1 - p)/(1 - p_0)$ and $p_0 = P(Y_1 = 0)$. Hence using equations (2.1) and (2.2), the mean, variance and third and fourth central moments aof Y are given by

$$\begin{aligned}
\mu_{1Y}^{'} &= \mathrm{E}(Y) = c\mathrm{E}(Y_1) = c\mu_1^{'} \\
\mu_{2Y} &= \mathrm{Var}(Y) = c\mathrm{Var}(Y_1) + c(1 - c)\left[\mathrm{E}(Y_1)\right]^2 \\
\mu_{3Y} &= c\left[\mu_3 + 3(1 - c)\mu_2\mu_1^{'} + (1 - 3c + 2c^2)\mu_1^{'3}\right] \\
\mu_{4Y} &= c\left[\mu_4 + 4(1 - c)\mu_3\mu_1^{'} + 6(1 - 2c + c^2)\mu_2\mu_1^{'2}(1 - 4c + 6c^2 - 3c^3)\mu_1^{'4}\right]
\end{aligned} \tag{5.15}$$

where $\mu_1^{'} = \mathrm{E}(Y_1)$. Note that equation (5.15) is also obtaind by replacing p by $(1 - c)$ in (5.11).

The cumulative distribution function (cdf) of $Y \sim$ ZAD is given by

$$\begin{aligned}
P(Y \leq 0) &= p \\
\\
P(Y \leq y) &= p + c[P(Y_1 \leq y) - P(Y_1 = 0)]
\end{aligned} \tag{5.16}$$

if $y = 1, 2, 3, \ldots$, where $c = (1 - p)/[1 - P(Y_1 = 0)]$.

The probability generating function (pgf) of $Y \sim$ ZAD is given by

$$G_Y(t) = (1 - c) + cG_{Y_1}(t) \tag{5.17}$$

where $G_{Y_1}(t)$ is the pgf of Y_1.

Zero adjusted (or altered) Poisson distribution A zero adjusted (or altered) Poisson distribution, denoted ZAP, i.e. $Y \sim \texttt{ZAP}(\mu, \sigma)$, is a discrete mixture of two components: value 0, with probability σ, and a distribution $\texttt{POtr}(\mu)$, (the Poisson distribution $\texttt{PO}(\mu)$ truncated at zero), with probability $1 - \sigma$.

Hence $Y = 0$ with probability σ and $Y = Y_0$ with probability $(1 - \sigma)$, where $Y_0 \sim \texttt{POtr}(\mu)$ and $0 < \sigma < 1$.

Hence

$$P(Y = 0|\mu, \sigma) = \sigma$$

$$P(Y = y|\mu, \sigma) = (1 - \sigma)P(Y_0 = y|\mu) = \frac{(1 - \sigma)P(Y_1 = y|\mu)}{1 - P(Y_1 = 0|\mu)}$$

if $y = 1, 2, 3, \ldots$, where $Y_0 \sim \texttt{POtr}(\mu)$ and $Y_1 \sim \texttt{PO}(\mu)$.

Hence the pf of $Y \sim \texttt{ZAP}(\mu, \sigma)$ 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 $0 < \sigma < 1$.

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

Zero adjusted (or altered) logarithmic distribution Let $Y = 0$ with probability σ and $Y = Y_1$, where $Y_1 \sim \texttt{LG}(\mu)$ a logarithmic distribution, with probability $(1 - \sigma)$, then Y has a zero adjusted logarithmic distribution, denoted by $\texttt{ZALG}(\mu, \sigma)$, with probability function given by

$$P(Y = 0|\mu, \sigma) = \sigma$$

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

if $y = 1, 2, 3, \ldots$, where $0 < \mu < 1$ and $0 < \sigma < 1$ and $\alpha = -\left[\log(1 - \mu)\right]^{-1}$. The mean and variance of of Y are given by $E(Y) = (1 - \sigma)\alpha\mu/(1 - \mu)$ and its variance by $Var(Y) = (1 - \sigma)\alpha\mu\left[1 - (1 - \sigma)\alpha\mu\right]/\left(1 - \mu\right)^2$.

Zero adjusted (or altered) negative binomial type 1 distribution A zero adjusted negative binomial type 1 distribution, denoted `ZANBI`, i.e. $Y \sim$ `ZANBI`(μ, σ, ν), is a discrete mixture of two components: value 0, with probability ν, and a distribution `NBItr`(μ, σ), [the negative binomial type 1 distribution `NBI`(μ, σ) truncated at zero], with probability $1 - \nu$.

Hence $Y = 0$ with probability ν and $Y = Y_0$ with probability $(1 - \nu)$, where $Y_0 \sim$ `NBItr`(μ, σ) and $0 < \nu < 1$.

Hence

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

$$P(Y = y|\mu, \sigma, \nu) = (1 - \nu)P(Y_0 = y|\mu, \sigma)$$

if $y = 1, 2, 3, \ldots$, where $Y_0 \sim$ `NBItr`(μ, σ).

Hence the pf of $Y \sim$ `ZANBI`(μ, σ, ν) 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 $Y_1 \sim$ `NBI`(μ, σ).

The mean of Y is given by

$$E(Y) = \frac{(1 - \nu)\, \mu}{1 - (1 + \mu\sigma)^{-1/\sigma}}$$

and the variance by

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

.

Further zero adusted distributions

Other zero adjusted distributions available in **gamlss.dist** are:

`ZAZIPF`(μ, σ)**:** zero adusted `ZIPF`,

`ZAPIG`(μ, σ, ν)**:** zero adusted `PIG`,

`ZABNB`(μ, σ, ν, τ)**:** zero adusted `BNB`,

`ZASICHEL`(μ, σ, ν, τ)**:** zero adusted `SICHEL`.

5.4 Comparison of the count distributions

Count distributions for Y can be compared using a diagram of their kurtosis against their skewness. Figure 5.1 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 5.1.

Figure **??** 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 5.1 but not labelled) from the skewness-kurtosis of the `ZIP` to that of the `NBI`. Figure **??** shows `ZINBI` distributions corresponding to four points along the `ZINBI` skewness-kurtosis curve from `ZIP` to `NBI`, where the extra probability ν 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 5.1 but not labelled) 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-parametrization of the `ZINBI` distribution, and hence has the same skewness-kurtosis curve (between `ZIP` and `NBI` in Figure 5.1 (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 5.1 for fixed mean 5 and fixed variance 30). Figure **??** shows `ZANBI` distributions corresponding to four points along the part of the `ZANBI` skewness-kurtosis curve which lies between `NBI` and `NBItr`, where the exact probability ν at $Y = 0$ decreases from 0.1667 [equal to that of the `NBI`(5,1) distribution] to close to zero.

The Sichel, Poisson-Tweedie (see Hougaard, P., Lee, M-L. T. and Whitmore [1997]) 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 Hougaarrd *et al.* (1997). It is a mixture of Poisson `PO`$(\mu\gamma)$ distributions, with a Tweedie mixing distribution for γ. The zero-inflated Poisson reciprocal Gamma (`ZIPRG`) distribution has three parameters and its skewness-kurtosis curve has the highest kurtosis for a given skewness.

Figure **??** shows four `SICHEL` distributions with increasing values of ν, all having mean 5 and variance 30, moving away from the `NBI`(5,1) distribution. Figure **??** shows four `DEL` distributions with increasing values of ν, 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 5.1, while the zero-inflated Sichel (ZISichel) covers the region between the ZIPRG and Sichel curves in Figure 5.1.

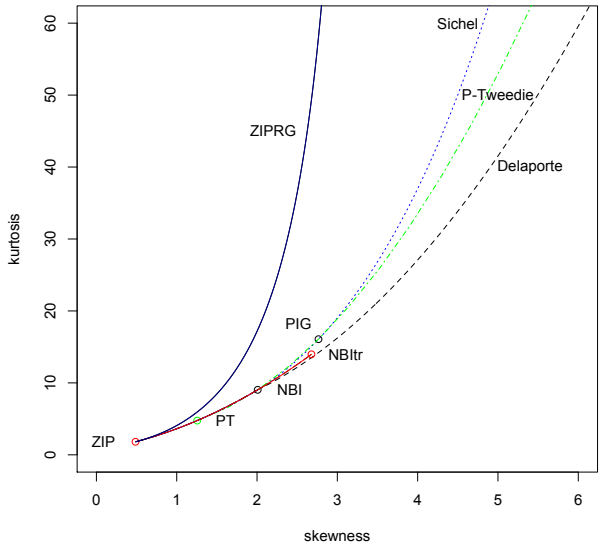


Figure 5.1: Skewness-kurtosis combinations for different distributions for Y (with fixed mean 5 and variance 30)

5.5 Families modelling the variance-mean relationship

Consider the mixed Poisson model defined by the marginal distribution of Y where $Y|\gamma \sim \texttt{PO}(\mu\gamma)$ and $\gamma \sim \texttt{D}(\sigma, \nu, \tau)$ for some distribution D where $E(\gamma) = 1$ and $Var(\gamma) = \upsilon(\sigma, \nu, \tau)$ for some function υ of the parameters σ, ν and τ of the mixing distribution D.

This leads to the mean of Y given by

$$E(Y) = E_\gamma\left[E(Y|\gamma)\right] = E_\gamma\left[\mu\gamma\right] = \mu$$

and a variance-mean relationship for Y given by $Var(Y) = \mu + \mu^2\,\upsilon(\sigma, \nu, \tau)$, since

$$Var(Y) = E_\gamma\left[Var(Y|\gamma)\right] + Var_\gamma\left[E(Y|\gamma)\right] = E_\gamma(\mu\gamma) + Var_\gamma(\mu\gamma) = \mu + \mu^2\upsilon(\sigma, \nu, \tau).$$

Hence in particular the negative binomial type I (NBI), the Poisson-inverse Gaussian (PIG), Sichel (SICHEL), Delaporte (DEL) and PSGIG distributions all have this quadratic variance-mean relationship. Alternative variance-mean relationships can be obtained by reparametrization.

For example the negative binomial type I (NBI) distribution has $E(Y) = \mu$ and $Var\,(Y) = \mu + \sigma\mu^2$. If σ is re-parametrized to σ/μ then $Var(Y) = (1 + \sigma)\mu$, giving the negative binomial type II (NBII) distribution. If σ is re-parametrized to $\sigma\mu$ then $Var(Y) = \mu + \sigma\mu^3$.

More generally, a family of re-parametrizations of the negative binomial type I distribution can be obtained by re-parametrizing σ to $\sigma\mu^{\nu-2}$, giving $Var(Y) = \mu + \sigma\mu^\nu$. This gives a three-parameter distribution model with parameters μ, σ and ν called the negative binomial family distribution and this is a distribution in **gamlss** denoted $\texttt{NBF}(\mu, \sigma, \nu)$, see Section ??? Note that for constant parameters μ, σ and ν, the distribution model is not identifiable. However if the parameters depend on explanatory variables then, in general, the model is identifiable.

Note that a family of re-parametrizations can be applied to other mixed Poisson distributions. In particular the Poisson-inverse Gaussian and Delaporte can be extended to reparametrization families using an extra parameter in a similar way to the negative binomial type I above.

Appendix: skewness and kurtosis for a mixture of Poisson distributions.

Let $Y|\gamma \sim \texttt{PO}(\mu\gamma)$ and γ 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 γ has mean 1, the cumulants of Y and γ are related by $E(Y) = \mu$, $Var(Y) = \mu + \mu^2 Var(\gamma,)$,

$$\kappa_{3Y} = \mu + 3\mu^2 Var(\gamma) + \mu^3\kappa_{3\gamma},$$

$$\kappa_{4Y} = \mu + 7\mu^2 Var(\gamma) + 6\mu^3\kappa_{3\gamma} + \mu^4\kappa_{4\gamma}, \tag{5.18}$$

where κ_{3Y} and κ_{4Y} are the third and fourth cumulants of Y

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

Sichel distribution

If Y has a $\texttt{SICHEL}(\mu, \sigma, \nu)$ distribution then the mean, variance skewness and kurtosis of Y are obtained using (5.18) from the cumulants of the mixing distribution $\gamma \sim \texttt{GIG}(1, \sigma^{1/2}, \nu)$, with pdf defined by (**??**), given by $E(\gamma) = 1$, $V(\gamma) = g_1$, $\kappa_{3\gamma} = g_2 - 3g_1$, and $\kappa_{4\gamma} = g_3 - 4g_2 + 6g_1 - 3g_1^2$, where

$$g_1 = 1/c^2 + 2\sigma(\nu+1)/c - 1,$$

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

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

where $g_2 = E(\gamma^2) - 1 = Var(\gamma)$, $g_3 = E(\gamma^3) - 1$ and $g_4 = E(\gamma^4) - 1$,

The cumulants for the Sichel distribution for Y are given by (5.18), from which the mean, variance, skewness and kurtosis of Y are obtained.

Exercises

- 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:

y	0	1	2	3	4	5	6	7
f	182	41	12	2	2	0	0	1

(a) Fit each of the following distributions for y to the data (using different model names e.g. `mPO` etc. for later comparison): `PO`(μ), `NBI`(μ,σ), `NBII`(μ,σ), `PIG`(μ,σ), `SICHEL`(μ,σ,ν), `DEL`(μ,σ,ν) and `ZIP`(μ,σ). [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.]

References: Gupta, P.L., Gupta, R.C. and Tripathi, R.C. (1996) Analysis of zero-adjusted count data. Computational Statistics and Data Analysis, 23, 207-218.

- 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:

y	0	1	2	3	4	5	5	7	10	12	15	20	30	37	50
male,no	541	19	17	16	3	6	5	2	6	1	0	3	1	0	0
male,yes	102	5	8	2	1	4	1	0	0	0	1	0	0	1	0
female,no	238	8	0	2	1	1	1	1	0	0	1	0	0	0	0
female,yes	103	6	4	2	0	1	0	0	0	0	0	0	0	0	1

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. buy using `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 `type<-gl(4,15)`.

(b) Fit each of the following distributions for y to the data (using different model names for later comparison): `PO`(μ), `NBI`(μ,σ), `NBII`(μ,σ), `PIG`(μ,σ), `SICHEL`(μ,σ,ν), `DEL`(μ,σ,ν) and `ZIP`(μ,σ), 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.

References: Heilbron, D.C. (1994) Zero-Altered and Other Regression Models for Count Data with Added Zeros. Biometrical Journal, 36, 531-547.

Chapter 6

Binomial data distributions

This chapter provides explanation for:

1. different types of binomial type data distributions within the GAMLSS family

Mikis: Needs examples to show why those distribution are helpfull

6.1 Available distributions

The binomial distribution is denoted BI(n, μ)) 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, μ, σ) 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 $n\mu(1-\mu)\left[1 + \sigma(n-1)/(1+\sigma)\right]$ and hence provides a model for overdispersed binomial data.

Chapter 7

Mixed distributions

This chapter provides explanation for:

1. Zero adjusted distributions defined on the interval $[0, \infty)$
2. Inflated distributions defined on $[0, 1]$

Mikis: stuff from the vignettes should be moved here

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.

7.1 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

$$f_Y(y) = \begin{cases} p & \text{if } y = 0 \\ (1-p) f_W(y) & \text{if } y > 0 \end{cases} \tag{7.1}$$

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).

7.1.1 Zero adjusted gamma distribution, ZAGA(μ, σ, ν)

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

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

The probability (density) function of the zero adjusted gamma distribution, denoted by (ZAGAμ,σ,ν), is given by

$$f_Y(y|\mu, \sigma, \nu) = \begin{cases} \nu & \text{if } y = 0 \\ (1-\nu) f_W(y|\mu, \sigma) & \text{if } y > 0 \end{cases} \tag{7.2}$$

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

The default link functions relating the parameters (μ, σ, ν) to the predictors (η_1, η_2, η_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 (7.2) is equivalent to a gamma distribution $GA(\mu, \sigma)$ model for $Y > 0$, together with a binary model for recoded variable Y_1 given by

$$Y_1 = \begin{cases} 0 & \text{if } Y > 0 \\ 1 & \text{if } Y = 0 \end{cases} \tag{7.3}$$

i.e.

$$p(Y_1 = y_1) = \begin{cases} (1-\nu) & \text{if } y_1 = 0 \\ \nu & \text{if } y_1 = 1 \end{cases} \tag{7.4}$$

The log likelihood function for the `ZAGA` model (7.2) is equal to the sum of the log likelihood functions of the gamma GA model and the binary BI model (7.4).

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 $Y > 0$.

7.1.2 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 (5.14) 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.

7.1.3 Example of fitting a response variable on zero and the positive real line

7.2 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

$$f_Y(y) = \begin{cases} p_0 & \text{if } y = 0 \\ (1 - p_0 - p_1) f_W(y) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases} \tag{7.5}$$

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.

7.2.1 Beta inflated distribution, BEINF(μ, σ, ν, τ)

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 $BE(\mu, \sigma)$ 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 BEINF(μ, σ, ν, τ), is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \begin{cases} p_0 & \text{if } y = 0 \\ (1 - p_0 - p_1) f_W(y|\mu, \sigma) & \text{if } 0 < y < 1 \\ p_1 & \text{if } y = 1 \end{cases} \tag{7.6}$$

for $0 \leq y \leq 1$, where $W \sim BE(\mu, \sigma)$ has a beta distribution with $0 < \mu < 1$ and $0 < \sigma < 1$ and $p_0 = \nu/(1 + \nu + \tau)$ 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 default link functions relating the parameters (μ, σ, ν, τ) 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 (7.6) 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

$$Y_1 = \begin{cases} 0 & \text{if } Y = 0 \\ 1 & \text{if } Y = 1 \\ 2 & \text{if } 0 < Y < 1 \end{cases} \tag{7.7}$$

i.e.

$$p(Y_1 = y_1) = \begin{cases} p_0 & \text{if } y_1 = 0 \\ p_1 & \text{if } y_1 = 1 \\ 1 - p_0 - p_1 & \text{if } y_1 = 2 \end{cases} \tag{7.8}$$

The log likelihood function for the `BEINF` model (7.6) is equal to the sum of the log likelihood functions of the beta BE model and the multinomial MN3 model (7.8).

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 $0 < Y < 1$.

7.2.2 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 (7.5) 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.

7.2.3 Beta inflated at 0 distribution, `BEINF0`(μ, σ, ν)

The beta inflated at 0 distribution is a mixture of two components: a discrete value 0 with probability p_0, and a beta `BE`(μ, σ) 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`(μ, σ, ν), is given by

$$f_Y(y|\mu, \sigma, \nu) = \begin{cases} p_0 & \text{if } y = 0 \\ (1 - p_0) f_W(y|\mu, \sigma) & \text{if } 0 < y < 1 \end{cases} \tag{7.9}$$

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 (μ, σ, ν) to the predictors (η_1, η_2, η_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 (7.9) 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

$$Y_1 = \begin{cases} 0 & \text{if } 0 < Y < 1 \\ 1 & \text{if } Y = 0 \end{cases} \tag{7.10}$$

i.e.

$$p(Y_1 = y_1) = \begin{cases} (1-\nu) & \text{if } y_1 = 0 \\ \nu & \text{if } y_1 = 1 \end{cases} \tag{7.11}$$

The log likelihood function for the `BEINF0`(μ, σ, ν) model (7.9) is equal to the sum of the log likelihood functions of the beta BE model and the binary `BI` model (7.14).

The `BEINF0`(μ, σ, ν) 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 $0 < Y < 1$.

7.2.4 Beta inflated at 1 distribution, `BEINF1`(μ, σ, ν)

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`(μ, σ, ν), is given by

$$f_Y(y|\mu,\sigma,\nu) = \begin{cases} p_1 & \text{if } y = 1 \\ (1-p_1) f_W(y|\mu,\sigma) & \text{if } 0 < y < 1 \end{cases} \tag{7.12}$$

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 (μ,σ,ν) to the predictors (η_1,η_2,η_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 (7.12) 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

$$Y_1 = \begin{cases} 0 & \text{if } 0 < Y < 1 \\ 1 & \text{if } Y = 1 \end{cases} \tag{7.13}$$

i.e.

$$p(Y_1 = y_1) = \begin{cases} (1-\nu) & \text{if } y_1 = 0 \\ \nu & \text{if } y_1 = 1 \end{cases} \tag{7.14}$$

The log likelihood function for the `BEINF1`(μ,σ,ν) model (7.12) is equal to the sum of the log likelihood functions of the gamma `BE` model and the binary `BI` model (7.14).

The `BEINF1`(μ,σ,ν) 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 $0 < Y < 1$.

7.2.5 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=FEV1/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 approriate 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`(μ, σ, ν) 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

$$f_Y(y|\mu,\sigma,\nu,\tau,p) = \begin{cases} p & \text{if } y = 1 \\ (1-p) f_W(y|\mu,\sigma,\nu,\tau) & \text{if } 0 < y < 1 \end{cases} \tag{7.15}$$

for $0 < y \leq 1$, where $W \sim logitSST(\mu, \sigma, \nu, \tau)$ 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 lht = log(height), i.e.

$$\begin{aligned} \mu &= \eta_1 = s(lht) \\ \log \sigma &= \eta_2 = s(lht) \\ \log \nu &= \eta_3 = s(lht) \\ \log \tau &= \eta_4 = s(lht) \\ \log\left(\frac{p}{1-p}\right) &= \eta_5 = s(lht) \end{aligned}$$

.

Model (7.15) 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 (7.13).

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

Chapter 8

Finite mixture distributions

This chapter covers finite mixtures within GAMLSS, in particular:

1. finite mixtures with no parameters in common and
2. finite mixtures with parameters in common.

8.1 Introduction to finite mixtures

This Chapter should be connected to the mixed distribution Chapter and also to the Chapter 7 from the first book. In fact do we need this Chapter or should be absorbed in the previous Chapter?

Part III

Inference

Chapter 9

The Likelihood function

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

1. basic concepts of statistical inference
2. the likelihood function

9.1 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.

9.1.1 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 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

$$f_P(y) = P(Y = y) \quad = \quad \begin{cases} P_I & \text{if} \quad y = Y_I \quad \text{for} \quad I = 1, 2, \ldots, D \\ 0 & \text{otherwise} \end{cases} \tag{9.1}$$

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 9.1 (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.

9.1.2 The sample

The 'true' population probability distribution $f_P(y)$ is usually *unknown* unless are we 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*.[1] The sample is denoted here as a vector $\mathbf{y}$ of length n. Let d be the number of distinct values of Y in the sample $\mathbf{y}$, and n_I the number of times that the distinct value y_I is 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 $\mathbf{y}$ given by

$$f_E(y) = \hat{P}(Y = y) \quad = \quad \begin{cases} p_I & \text{if} \quad y = y_I \quad \text{for} \quad I = 1, 2, \ldots, d \\ 0 & \text{otherwise} \end{cases} \tag{9.2}$$

The empirical probability function (epf), $f_E(y)$, can be plotted as a bar (or needle) plot as in Figure 9.1(a). Plotting the edf for a continuous variable Y is not a

[1] Strictly this is a random sample without replacement from a finite population.

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 **??** 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 district value appears in the sample. Figure 9.1 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 **??**. 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 9.1 (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.

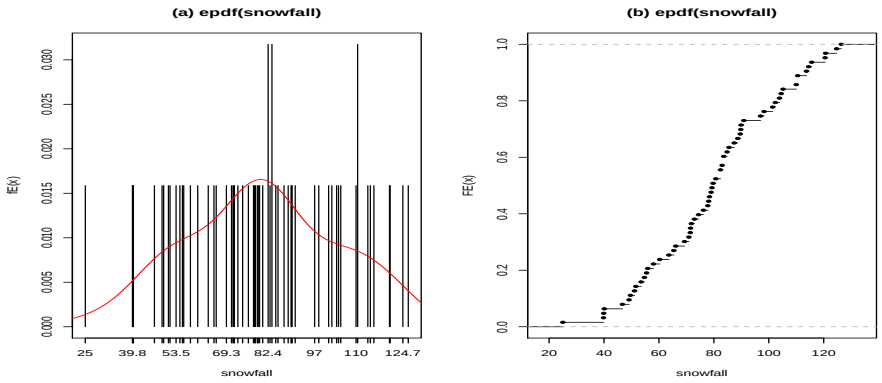


Figure 9.1: The empirical distribution and cumulative distribution function of the Parzen's snow fall data

9.1.3 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 generated from the population. The assumptions can be *parametric* assumptions or *non-parametric* assumptions leading to parametric or non-parametric modes of statistical inference respectively.

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 $f_Y(y|\boldsymbol{\theta})$ for a range of values of $\boldsymbol{\theta}$, and $f_P(y) = f_Y(y|\boldsymbol{\theta}_T)$ for all y for a particular unknown value $\boldsymbol{\theta}_T$ of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector of parameters.

The problem of the classical parametric statistical inference is then to determine possible values for the parameter $\boldsymbol{\theta}$ given the sample $\mathbf{y}$. Note that $\boldsymbol{\theta}$ is used here as generic parametric notation, since any one to one transformation of $\boldsymbol{\theta}$ (called a re-parametrization) will also perfectly define the theoretical distribution $f_Y(y|\boldsymbol{\theta})$. That is, $f_Y(y|\boldsymbol{\theta})$ and $f_Y(y|\boldsymbol{\phi})$ are equivalent assumptions if $\boldsymbol{\phi} = g(\boldsymbol{\theta})$ and the function $g()$ is a one to one transformation from the parameter space of $\boldsymbol{\theta}$ to the parameter space of $\boldsymbol{\phi}$. 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 $f_Y(y|\boldsymbol{\theta})$, (which can be continuous).
- By using a theoretical distribution $f_Y(y|\boldsymbol{\theta})$, 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 9.2 shows a schematic view[2] of the true population distribution, the empirical distribution and the the model $f_Y(y|\boldsymbol{\theta})$. Both population ("Population") and empirical ("Sample") distributions are denoted as points while the "Model" $f_Y(y|\boldsymbol{\theta})$ is represented as a line. Different values of the parameter $\boldsymbol{\theta}$ represent different points on the line. The different elliptical regions in Figure 9.2, 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 $f_Y(y|\boldsymbol{\theta})$, 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 $f_Y(y|\boldsymbol{\theta})$ is 'close' to $f_P(Y)$ and given that we choose $\boldsymbol{\theta}$ 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) $\boldsymbol{\theta}$, the parameter(s).

[2] In a schematic view we are trying to represent the concepts rather that correct mathematically defined spaces.

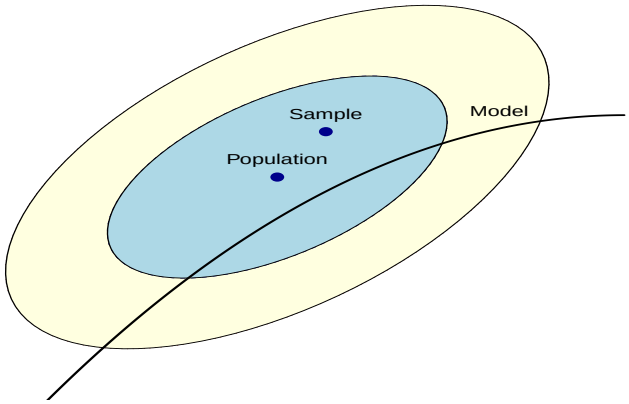


Figure 9.2: Showing a schematic plot of true population, the empirical distribution function and the model defined by $f(Y|\boldsymbol{\theta})$.

Classical parametric statistical inference assumes that the true population distribution $f_P(y)$ belongs to a particular family of probability (density) functions, given by $f_Y(y|\boldsymbol{\theta})$ for a range of values of $\boldsymbol{\theta}$, and $f_P(y) = f_Y(y|\boldsymbol{\theta})$ for all y for a particular value of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector of unknown parameters, i.e. the population distribution lies on the model line in Figure 9.2.

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

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

If $f_Y(y|\boldsymbol{\theta})$ is not a good approximation to the true population distribution for any $\boldsymbol{\theta}$, then this part of the classical statistical inference can be irrelevant to the particular problem. The dependance of inferential conclusions on the specification of the family of distributions $f_Y(y|\boldsymbol{\theta})$ for a range of values of $\boldsymbol{\theta}$ 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 9.1. 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 **??**).

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, ϑ, 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 ϑ 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 $\bar{y} = \hat{\mu} = \sum_{I=i}^{d} y_i p_i$ is the plug-in estimate of the population mean, $\mu_p = \sum_{I=i}^{D} Y_i P_i$.

9.2 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 ", (**?**) .

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 9.3.1 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 9.3.2 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 9.3.3 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 10.6.1 and **?**).

9.2.1 Definition of likelihood function

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

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be an observed *random* sample, from an assumed *discrete* population parametric probability function $f_Y(y|\boldsymbol{\theta})$ with a known functional form except for unknown parameter(s) $\boldsymbol{\theta}$. 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:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f_Y(y_i|\boldsymbol{\theta}). \tag{9.3}$$

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 $\mathbf{y}$ to $\boldsymbol{\theta}$. Given the observed values $\mathbf{y}$, the likelihood is **not** a function of the sample $\mathbf{y}$, since this has been observed and therefore is fixed, but a function of the parameter(s) $\boldsymbol{\theta}$.

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 (9.3) with respect to the parameter(s) $\boldsymbol{\theta}$. In practice it is more convenient to work with the logarithm of the likelihood. The log-likelihood is defined as

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_Y(y_i|\boldsymbol{\theta}). \tag{9.4}$$

In this book we deal with iid samples which justifies equation (9.3). More generally the definition of the the likelihood as the probability of observing the sample still holds but not (9.3). 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:

$$L(\boldsymbol{\theta}) = f_Y(y_1|\boldsymbol{\theta}) f_Y(y_2|(y_1, \boldsymbol{\theta}) f_Y(y_3|(y_1, y_2, \boldsymbol{\theta}) \dots f_Y(y_n|y_1, \dots, y_{n-1}, \boldsymbol{\theta}). \quad (9.5)$$

9.2.2 Clarification of the likelihood function for a continuous variable

Note that equation (9.3) is the exact probability of observing the data $\mathbf{y}$ given the parameters $\boldsymbol{\theta}$, 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 $\mathbf{y}$) can be defined as:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} P\left(y_i - \Delta_i < Y < y_i + \Delta_i | \boldsymbol{\theta}\right) = \prod_{i=1}^{n} \left[F_Y\left(y_i + \Delta_i | \boldsymbol{\theta}\right) - F_Y\left(y_i - \Delta_i | \boldsymbol{\theta}\right)\right] \quad (9.6)$$

where $F_Y()$ is the cumulative distribution function of Y. The definition of the likelihood in (9.6) is bounded above by one so cannot go to infinity, something which could happen if the definition (9.3) of the likelihood is used instead. Assume the Δ_i's are sufficiently small then

$$L(\boldsymbol{\theta}) \approx \prod_{i=1}^{n} f_Y(y_i|\boldsymbol{\theta}) \Delta_i = \left[\prod_{i=1}^{n} \Delta_i\right] \left[\prod_{i=1}^{n} f_Y(y_i|\boldsymbol{\theta})\right]. \quad (9.7)$$

Hence the log likelihood $\ell(\boldsymbol{\theta})$ is given approximately by:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_Y(y_i|\boldsymbol{\theta}) + \sum_{i=1}^{n} \log \Delta_i. \quad (9.8)$$

Clearly the second summation does not depend on $\boldsymbol{\theta}$ and hence when maximising $\ell(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$ 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 ∞. The problem is avoided by taking account of the measurement accuracy by using (9.6) instead of (9.7).

Within GAMLSS we have adopted the definition of the likelihood given in (9.3). Models can be maximised using (9.6) with the help of the package **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

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} [F_Y(y_{i2}|\boldsymbol{\theta}) - F_Y(y_{i1}|\boldsymbol{\theta})] \tag{9.9}$$

9.2.3 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 `rpanel` package, Bowman *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 5 36 22 139 210 97 30 23 13 14.

A histogram of the data is shown in Figure 9.3. All data points are positive so we require a distribution defined on the positive line. We are going to follow two different scenarios here.

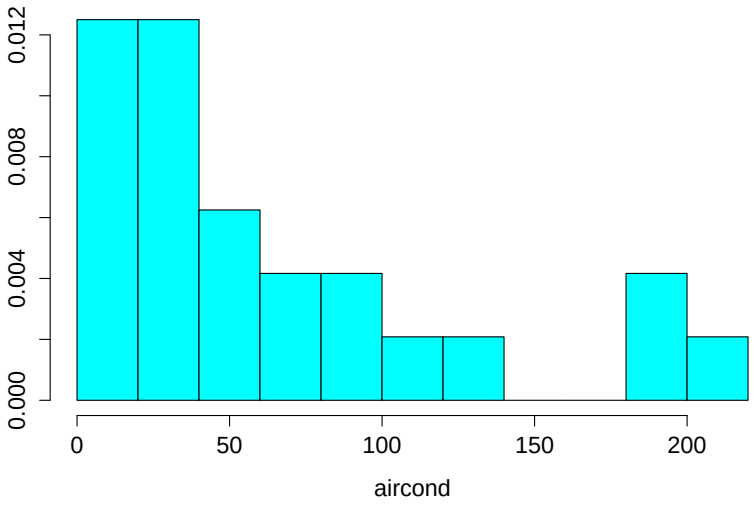


Figure 9.3: Showing a histogram of the air condition data.

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 (1.2) . Under this scenario the likelihood function is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\mu} e^{-y_i/\mu} \tag{9.10}$$

with log likelihood function given by

$$\begin{aligned} \ell(\mu) &= \sum_{i=1}^{n} \left\{ -\frac{y_i}{\mu} - \log(\mu) \right\} \\ &= -\frac{1}{\mu} \sum_{i=1}^{n} y_i - n \log(\mu) \end{aligned} \tag{9.11}$$

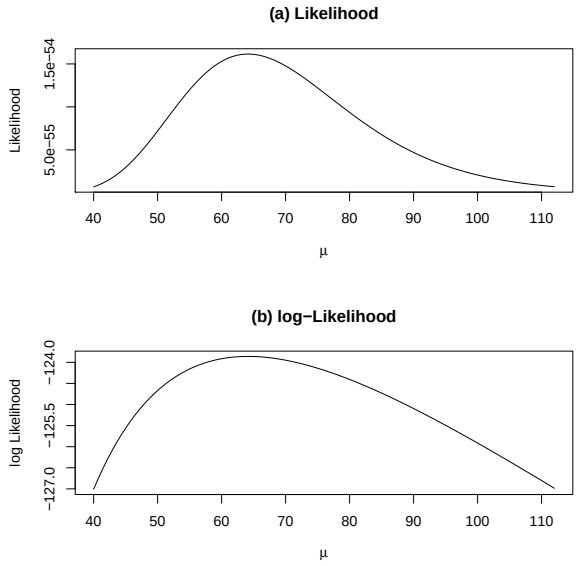


Figure 9.4: Showing the likelihood and log likelihood function for the air condition data assuming an exponential distribution.

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 (9.10) and (9.11) respectively for the parameter μ are shown in Figure 9.4 (a) and (b) respectively.

The following **R** code was used for the plot:

```
aircond <- c(50, 44, 102,  72,  22,  39,  3, 15, 197, 188, 79, 88, 46, 5,
                   5, 36,  22, 139, 210,  97,  30,  23,  13, 14)
truehist(aircond, nbins=10)
loglik <- function(theta, data) sum(dEXP(data, mu=theta, log=TRUE))
  logL <- rep(0,101)
     L <- rep(0,101)
    mu <- seq(40, 112, length=101)
for (i in 1:101)
{
logL[i] <- loglik(mu[i], data=aircond)
   L[i] <- exp(loglik(mu[i], data=aircond))
}
op <- par(mfrow=c(2,1))
plot(L~mu, type="l", xlab= expression(paste( mu)), ylab="Likelihood")
plot(logL~mu, type="l", xlab= expression(paste( mu)), ylab="log Likelihood")
par(op)
```

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

$$L(\mu,\sigma) = \prod_{i=1}^{n} \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y_i^{\frac{1}{\sigma^2}-1}\, e^{-y_i/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)} \tag{9.12}$$

with log Likelihood

$$\begin{aligned} \ell(\mu,\sigma) &= \sum_{i=1}^{n}\left[-\frac{1}{\sigma^2}\left(\log\sigma^2+\log\mu\right)+\left(\frac{1}{\sigma^2}-1\right)\log y_i - \frac{y_i}{\sigma^2\mu} - \log\Gamma\left(\frac{1}{\sigma^2}\right)\right] \\ &= -\frac{n}{\sigma^2}\left(\log\sigma^2+\log\mu\right) - n\log\Gamma\left(\frac{1}{\sigma^2}\right)+\left(\frac{1}{\sigma^2}-1\right)\sum_{i=1}^{n}\log y_i - \frac{\sum_{i=1}^{n} y_i}{\sigma^2\mu} \end{aligned} \tag{9.13}$$

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

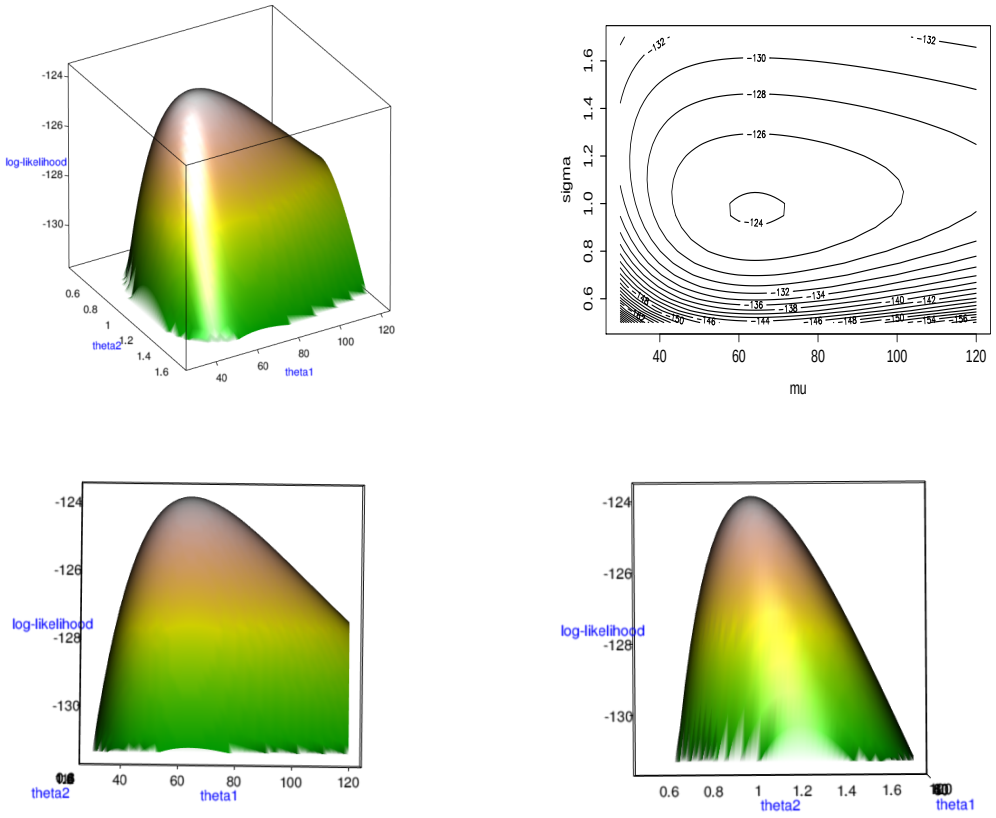


Figure 9.5: Showing different views of the log likelihood function for the air conditioning data assuming a gamma distribution: (a) top left corner shows the two dimensional log likelihood for μ (`theta1`) and σ (`theta2`) (b) top right corner shows a contour plot of the two dimensional log likelihood. (c) The bottom left the profile log likelihood for μ (`theta1`) (d) bottom right the profile log likelihood for σ (`theta2`) (see text for more details).

family. Note also that the gamma distribution used by (15.2) has polulation 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 μ and σ. Different views of the log-likelihood function of equation (9.13) for the μ and σ parameters are shown in Figure 9.5.

The top left corner of Figure 9.5 shows the two dimensional log likelihood function $\ell(\mu, \sigma)$ against μ (`theta1`) and σ (`theta2`). The top right corner shows a contour plot of the log likelihood function against μ and σ. The bottom left and right corners of Figure 9.5 show what the two dimensional log likelihood function looks like from the direction of the parameters μ and σ respectively. They show the *profile log likelihood functions* for each of the two parameters μ (`theta1`) and σ (`theta2`) respectively. See section 10.3.1 for the definition of the profile log likelihood function.

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

```
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 `rp.likelihood()` from the package `rpanel` of Bowman et al. [2007].

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

9.3 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 about the parameters, assuming repeated sampling of the data is allowed. Likelihood inference uses purely the likelihood for inference.

Sections 9.3.1, 9.3.2, and 9.3.3, 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) $\boldsymbol{\theta}$?
- 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 10 to 10.4.

9.3.1 Bayesian inference

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

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathbf{y}) &= \frac{L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \\ &\propto L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &\propto \text{Likelihood} \times \text{prior} \end{aligned} \tag{9.14}$$

where $\pi(\boldsymbol{\theta})$ is a prior distribution for the parameters $\boldsymbol{\theta}$ and where $\int_{\boldsymbol{\theta}} L(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is a constant to ensure that the posterior $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathbf{y})$ integrates to one. It is obvious from equation (9.14) that the only information coming from the data is contained in the likelihood function $L(\boldsymbol{\theta})$.

The following comments are related to Bayesian inference:

- By having a posterior distribution for all the parameters in $\boldsymbol{\theta}$, 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 $\boldsymbol{\theta}$ can be obtained. The parameters in $\boldsymbol{\theta}$ 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 $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathbf{y})$ 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 $\pi(\boldsymbol{\theta})$. Bayesian theory requires any prior information about $\boldsymbol{\theta}$ 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 $\pi(\boldsymbol{\theta})$ doest not have to be a proper distribution as long as the posterior $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathbf{y})$ 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].
- If we are interested only in one of the parameters in the vector $\boldsymbol{\theta}$, say θ_1 we can eliminate the rest of the parameters (sometimes called the *nuisance parameters*) by integrating them out i.e. $f_{\theta_1}(\theta_1) = \int_{\theta_2} \dots \int_{\theta_k} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\mathbf{y}) d\theta_2 \dots d\theta_k$. That is, Bayesian inference provides a consistent way of eliminating nuisance parameters.
- From the statistical modelling point of view, note that the likelihood $\mathrm{L}(\boldsymbol{\theta})$ 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.

9.3.2 Classical inference

Here we give an overview of how Classical inference answers the three question given at the start of section 9.3.

In section 10 maximum likelihood estimation is considered. An example is given in section 10.1.1. In section 10.2, 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 $f_Y(y|\boldsymbol{\theta})$ for a range of values of $\boldsymbol{\theta}$ and $f_P(y) = f_Y(y|\boldsymbol{\theta})$ for all y for a particular value of $\boldsymbol{\theta} = \boldsymbol{\theta}_T$, where $\boldsymbol{\theta}_T$ is the true value of $\boldsymbol{\theta}$]. Standard error based approximate confidence intervals and Wald tests for a parameter θ are given in section 10.2.3.

More accurate profile confidence intervals and generalized likelihood ratio tests for a parameter θ are given in sections 10.3.1 and 10.4.1 respectively. Section 10.4.3 investigates model selection using the generalized Akaike information criterion.

In section 10.5, the statistical properties of the maximum likelihood estimator (MLE) are given when the population model is mis-specified, i.e. $f_P(y) \neq$

$f_Y(y|\boldsymbol{\theta})$ for all y for any value of $\boldsymbol{\theta}$. Under model mis-specification, robust standard error based approximated confidence intervals and tests for $\boldsymbol{\theta}_c$ [where $\boldsymbol{\theta}_c$ is the value of $\boldsymbol{\theta}$ which makes $f_Y(y|\boldsymbol{\theta})$ *closest* to $f_P(y)$, as measured by the Kullback-Liebler distance] are also given in section 10.5.

9.3.3 Pure likelihood inference

The likelihood function provides a way for ordering possible values of the parameter(s) $\boldsymbol{\theta}$ given the data, but it is not a probability (density) function. For two different values of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the *likelihood ratio*

$$\frac{L(\boldsymbol{\theta}_1)}{L(\boldsymbol{\theta}_2)}$$

provides the ratio of how likely the data is, given the value of $\boldsymbol{\theta}_1$, relative to the value of $\boldsymbol{\theta}_2$. For example, a likelihood ratio of 10 means that the data is ten times more likely given $\boldsymbol{\theta}_1$ than given $\boldsymbol{\theta}_2$. This interpretation of the likelihood function is the basis of the pure likelihood theory approach in statical 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 (9.14).

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

$$L_s(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta})}{L(\hat{\boldsymbol{\theta}})}$$

is called the *standardized* likelihood function. The standardized likelihood can be used to create 'confidence' bounds for the $\boldsymbol{\theta}$ parameters. One can take the region of all values of $\boldsymbol{\theta}$ where $L_s(\boldsymbol{\theta}) > 0.10$ 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 $\boldsymbol{\theta}$ for which $L_s(\boldsymbol{\theta}) > 0.1466$ as an approximate 95% confidence region. The value 0.1466 is calculated by $0.1466 = \exp\left(-\chi^2_{10.05}/2\right) = \exp(-3.84/2)$ and is justified by section 10.4.1.

Figure 9.6 shows the standardized likelihood and log likelihood functions for μ 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 μ 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 μ are where this

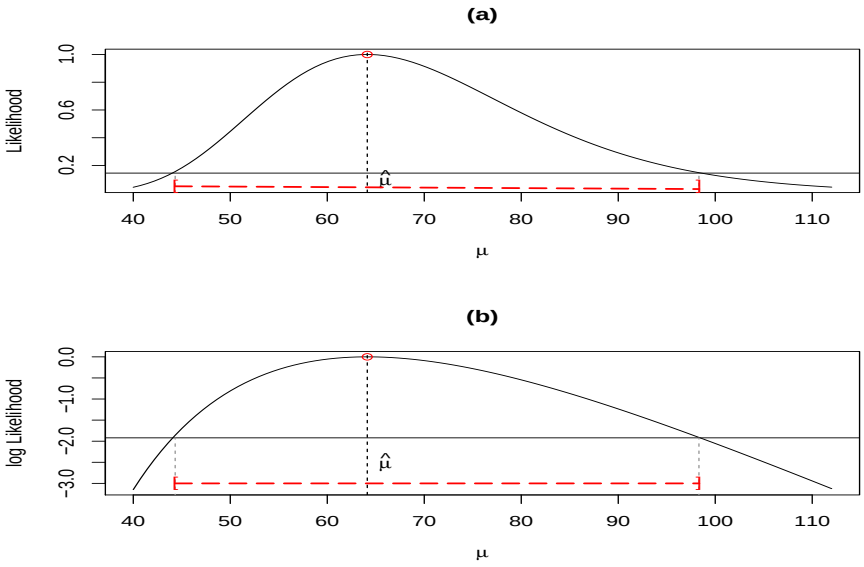


Figure 9.6: Showing for scenario I the likelihood and the log-likelihood function for the `aircond` data, the MLE estimator and confidence bounds for the parameters μ .

horizontal lines crosses the likelihood. The actual values for the bounds for μ are $[44.3, 98.3]$. The horizontal line on the log-likelihood is at $\log(0.1466) = -1.92$.

Figure 9.7 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.

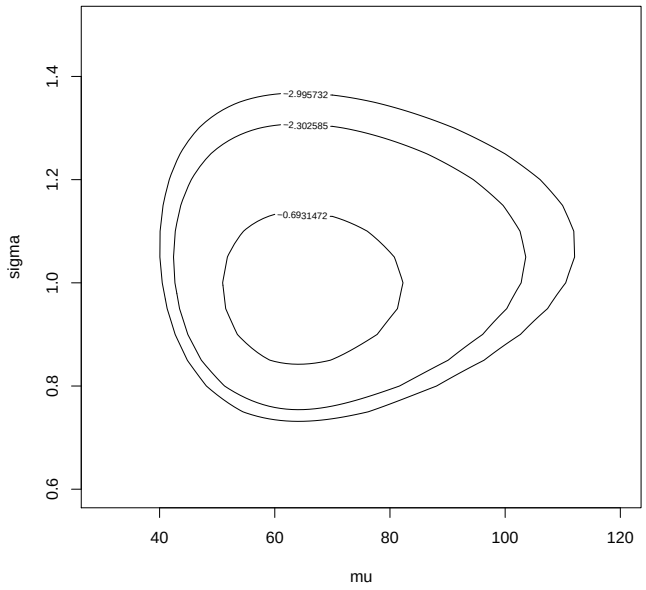


Figure 9.7: Scenario II log-likelihood based confidence bounds for parameter μ and σ at levels 0.5, 0.1 and 0.05 of the standardised likelihood.

Chapter 10

Maximum likelihood estimation

This chapter explains:

1. the maximum likelihood estimation,
2. properties of the aximum likelihood estimators and
3. how to obtain standard errors.

10.1 Introduction: Maximum likelihood estimation

The *maximum likelihood estimate* (MLe), is the value of $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, which maximizes the likelihood function $L(\boldsymbol{\theta})$ or equivalently maximizes the log-likelihood function $\ell(\boldsymbol{\theta})$. The maximum likelihood estimate is called a *point estimate* in Classical statistical inference. Point estimates are the solutions to the following inferential problem: If I have to guess the true value $\boldsymbol{\theta}_T$ of the parameters $\boldsymbol{\theta}$ what will this be? Point estimators are rather naive in the sense that even if we believe that there is a "true" parameter $\boldsymbol{\theta}_T$ our guess would inevitably be wrong. [Standard error based confidence intervals for a parameter θ are considered in section 10.2.3 while more reliable profile confidence intervals for θ are considered in section 10.3.1]

The maximum likelihood **estimate** (MLe) of $\boldsymbol{\theta}_T$ is a function of the observed sample $\mathbf{y}$, while the maximum likelihood **estimator** (MLE) of $\boldsymbol{\theta}_T$ is the same function of the random sample (of iid random variables) $\mathbf{Y}$. Hence the MLe of $\boldsymbol{\theta}_T$ is a vector of constant numerical values calculated from the observed $\mathbf{y}$, while the MLE of $\boldsymbol{\theta}_T$ is a vector of random variables which are functions of $\mathbf{Y}$.

Maximum likelihood estimates can be derived analytically or numerically. Analytical solutions are rather rare, so numerical computation of MLe's is frequently needed.

10.1.1 Air conditioning example continued: maximum likelihood estimation

Scenario I: exponential distribution

This demonstrates how to find the MLe analytically, using scenario I in section 9.2.3 where an exponential distribution was assumed. Differentiating the log likelihood function, $\ell(\mu)$, of equation (9.11) with respect to μ gives:

$$\frac{d\ell(\mu)}{d\mu} = \frac{\sum_{i=1}^{n} y_i}{\mu^2} - \frac{n}{\mu}. \tag{10.1}$$

By setting the result to zero and solving for μ gives

$$\frac{d\ell(\mu)}{d\mu} = 0 => \hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}. \tag{10.2}$$

where $\hat{\mu}$ is the MLe of μ. [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 $\mathbf{y}$. Note $\bar{y}$ is the maximum likelihood **estimate** (MLe)of μ, while the corresponding random variable $\bar{Y}$ is the maximum likelihood **estimator** (MLE) of μ. Note also that it is common in statistics to use a hat $\hat{}$ to indicate an estimate. For the example in section 9.2.3 on the time intervals, in hours, between failures of the air-conditioning equipment, the MLe of μ 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 (9.13) with respect to μ and σ gives respectively:

$$\frac{d\ell(\mu,\sigma)}{d\mu} = \frac{\sum_{i=1}^{n} y_i - n\mu}{\sigma^2\mu^2} \tag{10.3}$$

$$\frac{d\ell(\mu,\sigma)}{d\sigma} = \frac{2}{\sigma^3}\left[\left(\frac{\sum_{i=1}^{n} y_i}{\mu}\right) - \left(\sum_{i=1}^{n} log(y_i)\right) + nlog(\mu) + nlog(\sigma^2) - n + n\psi\left(\frac{n}{\sigma^2}\right)\right] \tag{10.4}$$

where $\psi(x) = \frac{d}{dx}\log\Gamma()$ is the psi or digamma function. Setting the two equations (10.3) and (10.4) equal to 0 and solving them simultaneously gives the

MLe for both μ and σ. 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 `optim()` and ii) the use of a more specialised function `mle()` in the package `stats4`. Note though that the function `mle()` uses function `optim()` as its minimization engine. The functions `gamlss()`, `gamlssML()`, `histDist()` and `fitDist()` which are used in the rest of this book for fitting distributions to data are explained in detail in Chapter **??**.

First define the log-likelihood function or more precisely minus the log-likelihood (since by default `optim()` minimises rather than maximises functions) and then call the function.

For scenario I:

```
 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
```

Value `45` is used as a starting value for μ, while `0.01` and `1000` (or `Inf`) are used as the lower and upper limits. Note that in the definition of the function `logl` the `gamlss.family` function `dEXP()` with argument `log=TRUE` is used to get the log-likelihood of the exponential distribution. In the function `optim()`, since this is a one parameter minimization, option `method="Brent"`, which is recommended for such situations, is used. For multidimensional $\boldsymbol{\theta}$ as below the option `method="L-BFGS-B"` should be used.

For the scenario II:

```
  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
```

Both `optim()` and `mle()` methods allow restrictions on the values of the parameter space something which is important in this case since the μ parameter in the exponential distribution and the μ and σ parameters in the gamma distribution only takes positive values.

Now use the function `mle()`:

```
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
```

The function `mle()` creates S4 rather S3 objects so to obtain its components use the `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 `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??

10.2 Statistical properties of MLE when the model is correct

Assume $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ 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 $f_Y(y|\boldsymbol{\theta})$ for a range of values of $\boldsymbol{\theta}$, where $f_Y(y|\boldsymbol{\theta})$ is a known function, except for parameters $\boldsymbol{\theta}$.

Assume the model is correct, i.e. assume that $f_P(y)$ belongs to the model family and $f_P(y) = f_Y(y|\boldsymbol{\theta})$ for all y for a particular value $\boldsymbol{\theta}_T$ of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}_T$ is the true value of $\boldsymbol{\theta}$.

Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$ from true model pdf $f_Y(y|\boldsymbol{\theta})$ given the random sample $\mathbf{Y}$.

There are three basic properties of the MLE $\hat{\boldsymbol{\theta}}$, assuming certain conditions hold.

- *invariance*
- *consistency*
- *asymptotic normality.*

10.2.1 Invariance

Invariance means that if $\boldsymbol{\phi}$ is a one to one transformation of parameter $\boldsymbol{\theta}$ say $\boldsymbol{\phi} = g(\boldsymbol{\theta})$, then the maximum likelihood estimators $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\theta}}$ are related by $\hat{\boldsymbol{\phi}} = g(\hat{\boldsymbol{\theta}})$. 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 10.2.3). Quadratic shapes of the likelihood at the maximum give more precise standard error based confidence intervals and Wald tests for $\boldsymbol{\theta}_T$ (see section 10.2.3).

10.2.2 Consistency

The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is, under certain conditions, a (weakly) consistent estimator of the true parameter $\boldsymbol{\theta}_T$, i.e. $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_T$ as $n \to \infty$.

This means that for all $\varepsilon > 0$,

$$\lim_{n\to\infty} p(|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T| > \varepsilon) = 0. \tag{10.5}$$

Sufficient conditions for weak consistency 10.5 to hold are given by Newey and McFaddon (1994) Theorem 2.5. A derivation of and sufficient conditions for strong consistency (which implies weak consistency (10.5) is given by Wald (1949), and with less restrictive conditions by White (1982).

10.2.3 Asymptotic normality

Convergence in distribution

First we will define convergence in distribution. Let $(U_n; n = 1, 2, ..)$ be a sequence of random variables and U another random variable, with cumulative distribution functions $F_{U_n}(u)$ for $n = 1, 2, ..$ 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, $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_T)$ converges in distribution to $N_K(0, \boldsymbol{J}(\boldsymbol{\theta}_T)^{-1})$ as $n \to \infty$, i.e.

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T) \xrightarrow{d} N_K(0, \boldsymbol{J}(\boldsymbol{\theta}_T)^{-1}), \tag{10.6}$$

where $\boldsymbol{J}(\boldsymbol{\theta}_T)$ is the (Fisher) expected information matrix for a single observation Y_i, evaluated at $\boldsymbol{\theta}_T$, given by

$$\boldsymbol{J}(\boldsymbol{\theta}_T) = -E_{f_P}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T} \tag{10.7}$$

where $\ell_i(\boldsymbol{\theta}) = \log f_Y(Y_i|\boldsymbol{\theta})$. Note that expectation in equation (10.7) is taken over the true population distribution $f_P(y_i) = f_Y(y_i|\boldsymbol{\theta}_T)$ for Y_i. Note also that the subscript $\boldsymbol{\theta}_T$ in (10.7) means that the quantity in square brackets is evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_T$.

An outline of the derivation of (10.6) is given in Appendix 10.6.2, Ripley [1996] page 32 or Claeskens and Hjort (2008) page 26-27. A more rigorous derivation of (10.6) with sufficient conditions is given by Cramer (1946) and with less restrictive conditions by White (1982). Sufficient conditions for (10.6) are also given by Newey and McFaddon (1994) Theoerm 3.3.

Note also (10.6) should be interpreted in terms of the limit of probabilities associated with $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_T)$ [and not in terms of the limit of moments of $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_T)$. For example the mean (or variance) of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T)$ does not necessarily converge to the mean (or variance) of the asymptotic distribution].

Informally, asymptotically as $n \to \infty$,

$$\hat{\boldsymbol{\theta}} \sim N_K(\boldsymbol{\theta}_T, n^{-1}\boldsymbol{J}(\boldsymbol{\theta}_T)^{-1}) = N_K(\boldsymbol{\theta}_T, \boldsymbol{i}(\boldsymbol{\theta}_T)^{-1}) \tag{10.8}$$

where

$$\boldsymbol{i}(\boldsymbol{\theta}_T) = -E_{f_P}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T} = -\sum_{i=1}^{n} E_{f_P}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T} = -nE_{f_P}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T} = n\boldsymbol{J}(\boldsymbol{\theta}_T)$$

is the (Fisher) expected information matrix of the n iid random variables $\mathbf{Y}$, evaluated at $\boldsymbol{\theta}_T$, where here $\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \ell_i(\boldsymbol{\theta}) = \sum_{i=1}^n \log f_Y(Y_i|\boldsymbol{\theta})$.

Asymptotic efficiency

For a single parameter θ, the maximum likelihood estimator $\hat{\theta}$ of θ_T is asymptotically a more efficient estimator of θ_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 θ_T to that of the MLE is greater than or equal to 1.

Approximating the expected information matrix

The expected information $\boldsymbol{i}(\boldsymbol{\theta}_T)$ is not always easy to derive analytically, therefore the *observed information* $\boldsymbol{I}(\boldsymbol{\theta}_T)$ is often used instead. The observed information $\boldsymbol{I}(\boldsymbol{\theta}_T)$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}_T$ is defined as

$$\boldsymbol{I}(\boldsymbol{\theta}_T) = -\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T} = -\sum_{i=1}^n \left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_T},$$

where here $\ell_i(\boldsymbol{\theta}) = \log f_Y(y_i|\boldsymbol{\theta})$. Note $\boldsymbol{I}(\boldsymbol{\theta}_T)$ is equal to the negative of the Hessian matrix of the log likelihood function at $\boldsymbol{\theta}_T$. The variance of the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is then approximated by $\boldsymbol{I}(\boldsymbol{\theta}_T)^{-1}$ instead of $\boldsymbol{i}(\boldsymbol{\theta}_T)^{-1}$.

Of course the point, $\boldsymbol{\theta}_T$ is unknown, so $\boldsymbol{\theta}_T$ is estimated by $\hat{\boldsymbol{\theta}}$ in both the expected $\boldsymbol{i}(\boldsymbol{\theta}_T)$ and observed $\boldsymbol{I}(\boldsymbol{\theta}_T)$ information, giving $\boldsymbol{i}(\hat{\boldsymbol{\theta}})$ and $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$.

The gamlss summary command output uses the following approximate distribution for $\hat{\boldsymbol{\theta}}$, for large n,

$$\hat{\boldsymbol{\theta}} \sim N_K(\boldsymbol{\theta}_T, \boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}).$$

Let $\boldsymbol{\theta}_T = (\theta_{T1}, \theta_{T2}, .., \theta_{TK})^\top$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, .., \hat{\theta}_K)^\top$. Hence, for $k = 1, 2, .., K$, the 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 $\boldsymbol{I}(\hat{\theta})^{-1}$.

Standard error based confidence interval for a parameter

A (standard error based) approximate $100(1-\alpha)\%$ confidence interval for a single parameter θ, e.g. θ_{Tk}, is given by $\left[\hat{\theta} \pm z_{\alpha/2} se(\hat{\theta})\right]$, 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 θ is often much less accurate than the profile confidence interval for θ given in section 10.3.1.

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

$$Z = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \sim NO(0, 1)$$

asymptotically as $n \to \infty$, if H_0 is true. Hence reject H_0 if the observed

$$\mid z \mid = \left| \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \right| > z_{\alpha/2}.$$

The Wald test for θ above is often much less accurate than the generalized likelihood ratio test for θ given in section 10.4.1.

Approximating the log likelihood function by a quadratic function

The log likelihood function can be approximated at the point of maximum $\hat{\boldsymbol{\theta}}$ by a quadratic function:

$$\begin{aligned} \ell(\boldsymbol{\theta}) &\approx \ell(\hat{\boldsymbol{\theta}}) + \left[\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right]_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \left[\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &= \ell(\hat{\boldsymbol{\theta}}) + +\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \left[\frac{\partial^2 \ell}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\hat{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \end{aligned} \tag{10.9}$$

because the second term is zero and where subscript $\hat{\boldsymbol{\theta}}$ above means the quantities inside the square brackets are evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. To demonstrate this the exponential distribution model for the `aircond` data is used:

```
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 $\left[\frac{\partial \ell^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\hat{\boldsymbol{\theta}}}$ is obtained using the function `optimHess()`. Figure 10.1 shows the log likelihood function of the data for the parameter μ 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 μ of a normal distribution given any value of σ 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.

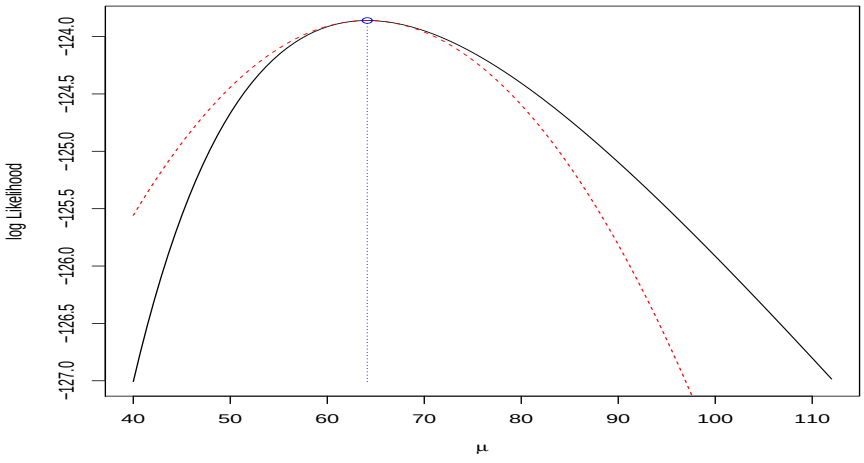


Figure 10.1: Showing the log likelihood of the `aircond` data together with the quadratic approximation of the log likelihood

10.2.4 Air conditioning example continued: se based CI and Wald test

For the `aircond` example using the exponential distribution, then, from (10.1), the observed information at $\mu = \mu_T$ is

$$\boldsymbol{I}(\mu_T) = -\left[\frac{d^2\ell(\mu)}{d\mu^2}\right]_{\mu_T} = 2\frac{\sum_{i=1}^n y_i}{\mu_T^3} - \frac{n}{\mu_T^2}.$$

The (Fisher's) expected information matrix evaluated at $\mu = \mu_T$ is

$$\begin{aligned} \boldsymbol{i}(\mu_T) = -E\left[\frac{d^2\ell(\mu)}{d\mu^2}\right]_{\mu_T} &= 2\frac{\sum_{i=1}^n E[Y_i]}{\mu_T^3} - \frac{n}{\mu_T^2} \\ &= 2\frac{n\mu_T}{\mu_T^3} - \frac{n}{\mu_T^2} \\ &= \frac{n}{\mu_T^2}, \end{aligned} \tag{10.10}$$

where, in (10.10), $\ell(\mu)$ is treated as a function of the random sample of iid random variables $\mathbf{Y}$, rather than the observed sample $\mathbf{y}$.

Hence $\boldsymbol{J}(\mu_T) = \boldsymbol{i}(\mu_T)/n = 1/\mu_T^2$ and as $n \to \infty$

$$\sqrt{n}(\hat{\mu} - \mu_T) \xrightarrow{d} N\left(0, \boldsymbol{J}(\mu_T)^{-1}\right) = N\left(0, \mu_T^2\right) \approx N\left(0, \hat{\mu}^2\right).$$

and, informally, asymptotically as $n \to \infty$

$$\hat{\mu} \sim N\left(\mu_T, \boldsymbol{i}(\mu_T)^{-1}\right) = N\left(\mu_T, \frac{\mu_T^2}{n}\right) \approx N\left(\mu_T, \frac{\hat{\mu}^2}{n}\right).$$

Using **R** we obtain:

```
 > # 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)
            [,1]
[1,] 0.005836558
> # estimated expected information matrix
> (length(aircond)/mean(aircond)^2)
[1] 0.005836554
```

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 $se(\hat{\mu}) = [\boldsymbol{i}(\hat{\mu})]^{-1/2} = \hat{\mu}/\sqrt{n} = \sqrt{(}1/0.005836554) = 13.0895$.

An approximate 95% confidence interval for μ_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 μ ($=\mu_T$). Note however that the default link for μ in the exponential gamlss family, `EXP`, is log. What this means, is that the parameter fitted in the predictor model is not μ 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 μ.

For parameters defined in the range from zero to infinity (as μ 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 μ 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 μ is very different from and probably more reliable than the 95% for μ 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 10.3.1) generally produce more reliable confidence intervals.

Standard error based Wald test for a parameter

A $100\alpha\%$ 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 $100\alpha\%$ significance level if the observed

$$\mid z \mid = \left| \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} \right| > z_{\alpha/2}.$$

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

$$\mid z \mid = \left| \frac{4.161 - 4.605}{0.2041} \right| = 2.175 > z_{0.025} = 1.96,$$

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 10.4.1 is more reliable.

10.3 Eliminating nuisance parameters using the profile log likelihood

The problem of eliminating nuisance parameters is important in practice.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the set of parameters of the model. Let us assume that we are interested in parameters $\boldsymbol{\theta}_1$, since for example they answer the scientific question we are looking at. The question is how we can eliminate $\boldsymbol{\theta}_2$ now called the *nuisance* parameter and concentrate only on inference about $\boldsymbol{\theta}_1$, the parameter(s) of *interest*. One answer to the question is to use *profile log likelihood* for inference about $\boldsymbol{\theta}_1$.

10.3.1 Profile log likelihood function and profile confidence intervals

Let $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the likelihood function for parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The profile log likelihood function $p\ell(\boldsymbol{\theta}_1)$ for parameters $\boldsymbol{\theta}_1$ is given by maximizing $\ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ over $\boldsymbol{\theta}_2$ for each value of $\boldsymbol{\theta}_1$, i.e.

$$p\ell(\boldsymbol{\theta}_1) = \max_{\boldsymbol{\theta}_2} \ell(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

for each value of $\boldsymbol{\theta}_1$.

For a single parameter $\theta = \theta_1$, $p\ell(\theta) = \max_{\boldsymbol{\theta}_2} \ell(\theta, \boldsymbol{\theta}_2)$, and $p\ell(\theta)$ can be plotted against θ. Alternatively the profile Global Deviance $pGD(\theta) = -2 * pl(\theta)$ can be plotted against θ.

A $100\alpha\%$ profile confidence interval for a single parameter θ includes all values θ_0 for which $pGD(\theta_0) < pGD(\hat{\theta}) + \chi^2_{1,\alpha}$ where $\hat{\theta}$ maximizes $p\ell(\theta)$ over θ and hence minimizes $pGD(\theta)$ over θ and $\chi^2_{1,\alpha}$ is the upper tail value of a Chi-squared distribution with one degree of freedom with upper tail probability $\alpha\%$.

This $100\alpha\%$ profile confidence interval for a single parameter θ includes all values θ_0 of parameter θ that are accepted by a generalized likelihood ratio test (see section 10.4.1) of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ at the $100\alpha\%$ significance level.

This equals all values θ_0 for which $GD_0 < GD_1 + \chi^2_{1,\alpha}$, where GD_0 and GD_1 are the fitted global deviance = -2*max(log likelihood) under hypotheses H_0 and H_1 respectively, i.e.

$$GD_0 = -2 * \max_{\boldsymbol{\theta}_2}[\ell(\theta_0, \boldsymbol{\theta}_2)] = -2 * p\ell(\theta_0) = pGD(\theta_0)$$

and

$$GD_1 = -2*\max_{\theta, \boldsymbol{\theta}_2}[\ell(\theta, \boldsymbol{\theta}_2)] = -2*\max_{\theta}\left[\max_{\boldsymbol{\theta}_2} \ell(\theta, \boldsymbol{\theta}_2)\right] = -2*\max_{\theta} p\ell(\theta) = -2*p\ell(\hat{\theta}) = pGD(\hat{\theta}).$$

Hence $\Lambda = GD_0 - GD_1 = pGD(\theta_0) - pGD(\hat{\theta}) \sim \chi^2_1$ approximately if H_0 is true. Hence $H_0 : \theta = \theta_0$ is accepted if $GD_0 < GD_1 + \chi^2_{1,\alpha}$, i.e if $pGD(\theta_0) < pGD(\hat{\theta}) + \chi^2_{1,\alpha}$.

10.3.2 Air conditioning example continued: profile confidence intervals

In the air conditioning example from section 9.2.3, assume the data are a random sample from a gamma distribution, $GA(\mu, \sigma)$. The profile likelihood $p\ell(\mu)$ for μ is given by

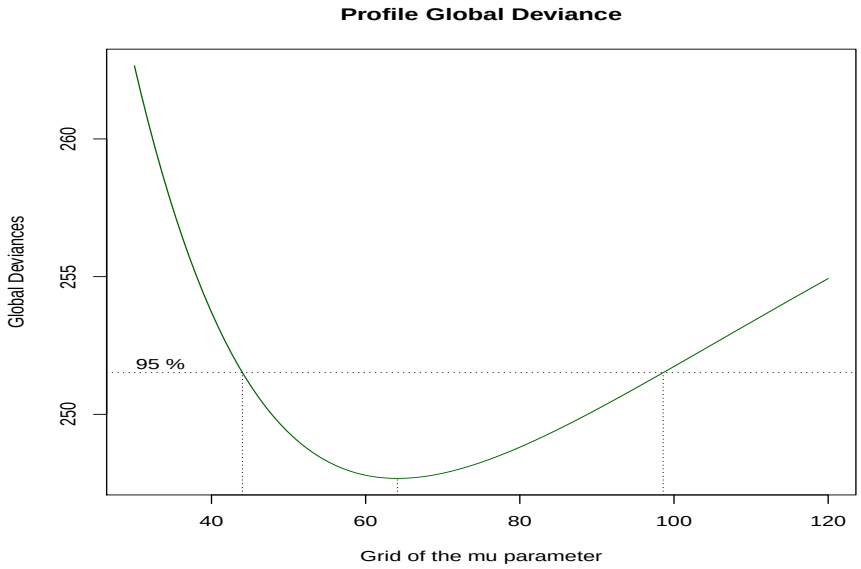


Figure 10.2: Profile global deviance plot for μ for gamma example

$$p\ell(\mu) = \max_{\sigma} \ell(\mu, \sigma)$$

for each value of μ. Similarly the profile likelihood $p\ell(\sigma)$ for σ is given by

$$p\ell(\sigma) = \max_{\mu} \ell(\mu, \sigma)$$

for each value of σ.

The bottom left and right corners of Figure 9.5 show what the two dimensional likelihood looks like from the direction of the parameters μ (theta1) and σ (theta2) respectively. They show the *profile likelihood* for each of the two parameters μ and σ respectively.

The profile global deviance for μ is $pGD(\mu) = -2 * p\ell(\mu)$ The profile global deviance plot of $pGD(\mu)$ against μ can be obtained in GAMLSS by:

```
m2 <- gamlss(aircond~1, family=GA)
m2A <- prof.dev(m2,"mu",min=30,max=120,step=0.1,type="l")
```

giving Figure 10.2 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 ($=\mu_T$) from the profile deviance includes all values μ_0 that are accepted by a generalized likelihood ratio test (see section 10.4.1) of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ at the 5% significance level.

The maximum likelihood estimate $\hat{\mu} = 64.125$ of μ corresponds to the minimum global deviance. The vertical dotted lines in Figure 10.2 mark the profile confidence interval for μ, i.e. (44.015 , 98.60), given by all μ for which $pGD(\mu) < min_\mu pGD(\mu) + 3.84$, since $\chi^2_{1,0.05} = 3.84$.

The profile global deviance for σ is $pGD(\sigma) = -2 * p\ell(\sigma)$ The profile global deviance plot of $pGD(\sigma)$ against σ can be obtained in GAMLSS by:

```
m2B <- prof.dev(m2,"sigma",min=0.6,max=1.7,step=0.01,type="l")
```

giving Figure 10.3 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:

```
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 β coefficient of x.]

Similarly the profile log likelihood function and profile confidence interval for the predictor $\log \sigma$ can be obtained by

```
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 ).
```

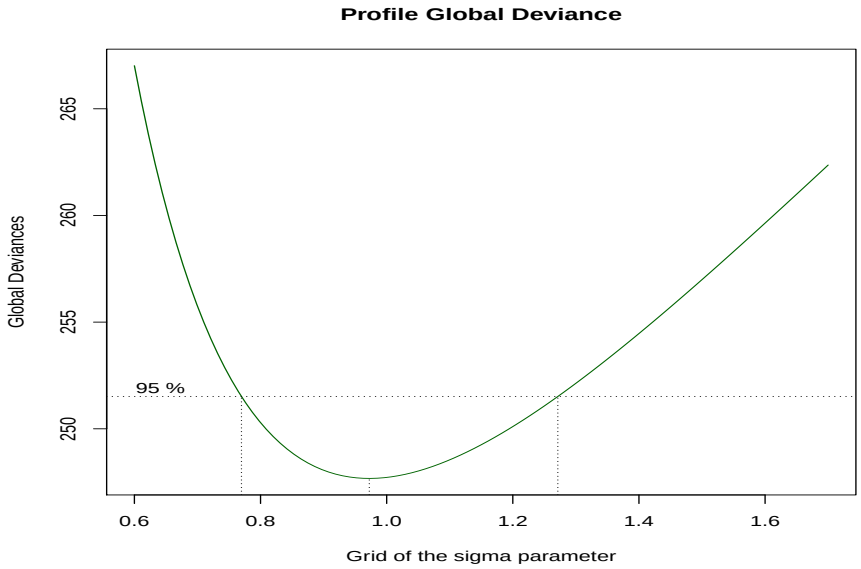


Figure 10.3: Profile global deviance plot for σ for gamma example

10.4 Model selection

10.4.1 Testing between nested models using the generalized likelihood ratio test

Choosing between models, (we had called them scenarios in Section 9.2.3 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

$$GD = -2\log L(\hat{\boldsymbol{\theta}}) = -2\ell(\hat{\boldsymbol{\theta}})$$

as the *Global Deviance* (or GD). It is a very important quantity for testing between different models.

generalized likelihood ratio test

Let $\mathcal{M}_0$ and $\mathcal{M}_1$ be two different models. Model $\mathcal{M}_0$ is called *nested* within $\mathcal{M}_1$ if it a special case of $\mathcal{M}_1$. In this case $\mathcal{M}_0$ is the simpler model while $\mathcal{M}_1$ the more complicated one.

Two nested parametric GAMLSS models, $H_0 : \mathcal{M}_0$ and $H_1 : \mathcal{M}_1$, where $\mathcal{M}_1$ is a submodel of $\mathcal{M}_0$, 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 $\mathcal{M}_0$, with degrees of freedom $d = df_{e0} - df_{e1}$, (given that the usual conditions are satisfied). For each model $\mathcal{M}$ the error degrees of freedom df_e is defined by $df_e = n - p$, where p is the number of parameters in model $\mathcal{M}$. 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 $GD_0 < GD_1 + \chi^2_{d,\alpha}$.

10.4.2 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.

```
> 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.

10.4.3 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.

10.5 Statistical properties of MLE when the model is mis-specified

10.5.1 Graphical representation

In order to understand the concepts involved we shall use the schematic presentation of the population, sample and model in Figure 10.4. Figure 10.4 (similar to Figure 9.2) 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 $\boldsymbol{\theta}_c$ on the model line represents the value of $\boldsymbol{\theta}$ which is closest to the population distribution as measured by the Kullback-Liebler distance

$$d\left[f_P(y), f_Y(y|\boldsymbol{\theta})\right] = \int \left[log f_P(y) - log f_Y(y|\boldsymbol{\theta})\right] f_P(y) dy, \tag{10.11}$$

i.e. $\boldsymbol{\theta}_c$ minimizes the Kullback-Liebler distance (10.11) over $\boldsymbol{\theta}$.

Here is the confusion: traditional statistical books refer to $\boldsymbol{\theta}_c$ as the 'true' $\boldsymbol{\theta}$ 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 $\boldsymbol{\theta}_c$ as the 'closest' value for $\boldsymbol{\theta}$ under the model $f_Y(y|\boldsymbol{\theta})$ to emphasize that $f_Y(y|\boldsymbol{\theta})$ is just one model among other possible ones.

10.5.2 Properties of MLE under model mis-specification

Assume $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ 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 $f_Y(y|\boldsymbol{\theta})$ for a range of values of $\boldsymbol{\theta}$, where $f_Y(y|\boldsymbol{\theta})$ is a known function, except for parameters $\boldsymbol{\theta}$.

Assume the model is incorrect, i.e. assume that $f_P(y)$ does not belong to the model family and $f_Y(y|\boldsymbol{\theta})$ for any value of $\boldsymbol{\theta}$.

Let $\hat{\boldsymbol{\theta}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}$ from model $f_Y(y|\boldsymbol{\theta})$ given the random sample $\mathbf{Y}$.

There are three basic classical properties of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, 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 $\hat{\boldsymbol{\theta}}$ given in section 10.2.1 still holds, except the true value $\boldsymbol{\theta}_T$ of $\boldsymbol{\theta}$ is replaced by the 'closest' value $\boldsymbol{\theta}_c$.

Consistency

The consistency property of $\hat{\boldsymbol{\theta}}$ given in section 10.2.2 still holds, except the true value $\boldsymbol{\theta}_T$ of $\boldsymbol{\theta}$ is replaced by the 'closest' value $\boldsymbol{\theta}_c$.

There are situations in which $\boldsymbol{\theta}_c$ still represents a true population distribution measure. In particular for a mis-specified Exponential Family distribution model, $EF(\mu, \sigma)$ with mean parameter μ and scale parameter σ, then μ_c, the 'closest' value of μ to the true population distribution, is equal to the true population mean. This follows from (10.11), by substituting $\log f_Y(y|\boldsymbol{\theta}) = \log f_Y(y|\mu, \sigma)$ for an Exponential Family distribution and differentiating with respect to μ giving

$$\frac{\partial}{\partial \mu} d\left[f_P(y), f_Y(y|\mu, \sigma)\right] = -\int \frac{y-\mu}{\sigma^2 v(\mu)} f_P(y) dy = \frac{1}{\sigma^2 v(\mu)} \left[E_{f_P}(Y) - \mu\right] \quad (10.12)$$

Setting this equal to zero and solving for μ gives $\mu_c = E_{f_P}(Y)$.

Hence for an Exponential Family distribution model the maximum likelihood estimator $\hat{\mu} = \bar{Y}$ of μ is a consistent estimator of the true population mean $E_{f_P}(Y)$. This does not, in general, hold for a non Exponential Family distribution model, Gourieroux *et al.* (1984). It does however hold in special cases, for example for a model with a t family distribution, when the population distribution is assumed to be symmetric.

Asymptotic normality

Under certain conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_c)$ converges in distribution to $N_K(0, \boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_c)\boldsymbol{J}(\boldsymbol{\theta}_c)^{-1})$ i.e.

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_c) \xrightarrow{d} N_K(0, \boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_c)\boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}), \quad (10.13)$$

where $\boldsymbol{J}(\boldsymbol{\theta}_c)$ is the (Fisher) expected information matrix for a single observation Y_i, evaluated at $\boldsymbol{\theta}_c$, given by

$$\boldsymbol{J}(\boldsymbol{\theta}_c) = -E_{f_P}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_c} \quad (10.14)$$

where $\ell_i(\boldsymbol{\theta}) = \log f_Y(Y_i|\boldsymbol{\theta})$ and $\boldsymbol{K}(\boldsymbol{\theta}_c)$ is the variance of the first derivative of the log likelihood function for a single observation Y_i, evaluated at $\boldsymbol{\theta}_c$, given by

$$\boldsymbol{K}(\boldsymbol{\theta}_c) = V_{f_P}\left[\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]_{\boldsymbol{\theta}_c}. \quad (10.15)$$

Note that the expectation and variance in equations (10.14) and (10.15) respectively is taken over the true population distribution $f_P(y_i)$ for Y_i.

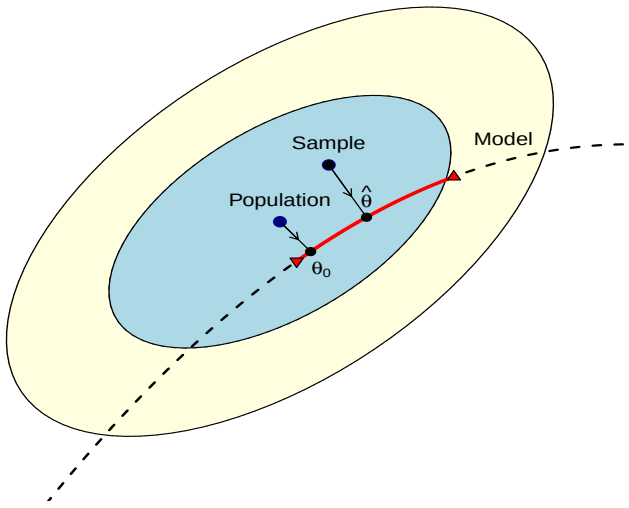


Figure 10.4: A schematic presentation of the $\boldsymbol{\theta}_c$, the 'closest' or 'best' value for $\boldsymbol{\theta}$ under the model $f_Y(y|\boldsymbol{\theta})$, and the $\hat{\boldsymbol{\theta}}$ the MLE. $\boldsymbol{\theta}_c$ is the value of $\boldsymbol{\theta}$ 'closest' to the population using the Kullback-Liebler distance (or risk function), while $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ closest to the sample using the empirical risk function. The solid (red) line represents confidence bounds for the parameter $\boldsymbol{\theta}_c$.

An outline of the derivation of (10.13) is given in Appendix 10.6.2, Ripley [1996] page 32 or Claeskens and Hjort (2008) page 26-27. A more rigorous derivation of (10.13) with sufficient conditions is given by White (1982).

Equation (10.13) 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 $\boldsymbol{\theta}_c$.

Hence, informally, asymptotically as $n \to \infty$,

$$\hat{\boldsymbol{\theta}} \sim N_K(\boldsymbol{\theta}_c, n^{-1}\boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_c)\boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}) \tag{10.16}$$

[Note if the population distribution belongs to the model parametric family of distributions, then $\boldsymbol{\theta}_c = \boldsymbol{\theta}_T$, $\boldsymbol{K}(\boldsymbol{\theta}_T) = \boldsymbol{J}(\boldsymbol{\theta}_T)$ and the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is just $\boldsymbol{i}(\boldsymbol{\theta}_T)^{-1} = n^{-1}\boldsymbol{J}(\boldsymbol{\theta}_T)^{-1}$ and so $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T) \xrightarrow{d} N_K(0, \boldsymbol{i}(\boldsymbol{\theta}_T)^{-1})$.]

Approximating the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$

The asymptotic variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by $n^{-1}\boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}\boldsymbol{K}(\boldsymbol{\theta}_c)\boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}$ from (10.16).

The matrices $\boldsymbol{J}(\boldsymbol{\theta})$ and $\boldsymbol{K}(\boldsymbol{\theta})$ are unknown expected and variance-covariance matrices respectively, which can be approximated by the corresponding sample mean and variance-covariance matrices, i.e.

$$\hat{\boldsymbol{J}}(\boldsymbol{\theta}_c) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right]_{\boldsymbol{\theta}_c}$$

and

$$\hat{\boldsymbol{K}}(\boldsymbol{\theta}_c) = \frac{1}{n-K}\sum_{i=1}^{n}\left[\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^T}\right]_{\boldsymbol{\theta}_c}.$$

where here $\ell_i\boldsymbol{\theta} = \log f_Y(y_i|\boldsymbol{\theta})$.

Of course the point $\boldsymbol{\theta}_c$ is unknown, so $\boldsymbol{\theta}_c$ is estimated by $\hat{\boldsymbol{\theta}}$ giving $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})$.

The output from the gamlss summary command with option robust=TRUE or robust=T uses the following approximate distribution for $\hat{\boldsymbol{\theta}}$, for large n,

$$\hat{\boldsymbol{\theta}} \sim N_K(\boldsymbol{\theta}_c, n^{-1}\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}).$$

The estimated variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, given by $n^{-1}\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}$, 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 $f(y|\boldsymbol{\theta})$ (do the unnecessary parameters effect the results?)
- those standard errors are the robust standard errors (sandwich) Shall we use them all the time?

10.5.3 Robust confidence intervals and tests

Let $\boldsymbol{\theta}_c = (\theta_{c1}, \theta_{c2}, .., \theta_{cK})^\top$ and $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, .., \hat{\theta}_K)^\top$, where $\boldsymbol{\theta}_c$ is the value of parameter $\boldsymbol{\theta}$ which makes the model distribution $f_Y(y|\boldsymbol{\theta})$ '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 θ_{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 $n^{-1}\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}\hat{\boldsymbol{K}}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{J}}(\hat{\boldsymbol{\theta}})^{-1}$.

Robust (standard error based) confidence interval for a parameter

A robust (standard error based) approximate $100(1-\alpha)\%$ confidence interval for a single parameter θ, e.g. θ_{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

$$| z | = \left| \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \right| > z_{\alpha/2}.$$

10.5.4 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 $\frac{d\ell}{d\mu}$ is still given by (10.1). Hence $\boldsymbol{J}(\mu_c) = \frac{1}{n}\boldsymbol{i}(\mu_c) = \mu_c^{-2}$ from section 10.2.4, and

$$\boldsymbol{K}(\mu_c) = V_{f_P}\left[\frac{d\ell_i}{d\mu}\right]_{\mu_c} = V_{f_P}\left[\frac{Y_i}{\mu^2} - \frac{1}{\mu}\right]_{\mu_c} = \frac{V_{f_P}(y)}{\mu_c^4}$$

and hence $\boldsymbol{J}(\mu_c)^{-1}\boldsymbol{K}(\mu_c)\boldsymbol{J}(\mu_c)^{-1} = V_{f_P}(y)$ giving

$$\sqrt{n}(\hat{\mu} - \mu_c) \xrightarrow{d} N\left(0, V_{f_P}(y)\right) \approx N\left(0, s_y^2\right),$$

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 $se(\hat{\mu}) = s_y/\sqrt{(}n) = 12.7888$. This is usually called a robust standard error, i.e. robust to mis-specification of the true population distribution. An approximate robust 95% confidence interval for μ_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 $\mu_c = E_{f_P}(y)$, so the above is a robust confidence interval for $E_{f_P}(y)$, 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 μ in the exponential gamlss family, `EXP`, is log. What this means is that the parameter fitted in the model is not μ 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 μ.

For parameters defined in the range from zero to infinity (as μ 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, robust=T)
******************************************************************
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   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
                      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 errors for log mu
> vcov(m1, "se", robust=T)
(Intercept)
  0.1994367
```

```
> # robust 95\% CI for log mu
> confint(m1, robust=T)
               2.5 %   97.5 %
(Intercept) 3.769946 4.551723

> # robust 95\% CI for mu
> exp(confint(m1, robust=T))
               2.5 %  97.5 %
(Intercept) 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 $[4.161 \pm (1.96 * 0.1994)] = (4.16 \pm 0.39) = (3.77, 4.55)$ and is given by `confint(m1, robust=T)`. Hence the robust 95% confidence interval for $\mu_c = E_{f_P}(y)$ is given by (43.38, 94.80).

10.6 Appendix of Chapter 3

10.6.1 Entropy, risk and empirical risk functions

In order to be able to evaluate the performance of the parametric model $f_Y(y|\boldsymbol{\theta})$ 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 *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 $H = -\sum_i^n p_i log(p_i)$ where $p_1, p_2, \ldots, p_n$ are the probabilities of the events of the system and where $\sum_i^n p_i = 1$. Applying this definition to the "true" population distribution $f_P(Y)$, we can define the entropy of the population as:

$$\begin{aligned} H_P &= \sum_{I=1}^{D} P_I log(1/P_I) = -\sum_{I=1}^{D} P_I log(P_I) \\ &= \sum_{I=1}^{D} f_P(Y_I) \log\left(\frac{1}{f_P(Y_I)}\right) = -\sum_{i=1}^{D} f_P(Y_I) \log\left(f_P(Y_I)\right) \end{aligned} \tag{10.17}$$

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 $d\left[f_{Y_1}(y), f_{Y_2}(y)\right] = \int f_{Y_1}(y) \log \frac{f_{Y_1}(y)}{f_{Y_2}(y)} dy$ 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, $d\left[f_{Y_1}(y), f_{Y_2}(y)\right] \neq d\left[f_{Y_2}(y), f_{Y_1}(y)\right]$. [1] Nevertheless d is ideal for the purpose of judging how far the model distribution, $f_Y(y|\boldsymbol{\theta})$, 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

$$Pr(Y \in \Delta_I) = \int_{\Delta_I} f_Y(y|\theta) dy \simeq f_Y(Y|\boldsymbol{\theta}) \Delta_I \tag{10.18}$$

for a value of Y in the small interval Δ_I. The *expected loss of information* or *risk* function for choosing the model distribution $f_Y(y|\boldsymbol{\theta})$ instead of the "true" distribution $f_P(Y)$ is defined as

$$\begin{aligned} R_P(\theta) &= \sum_{I=1}^{D} P_I \log\left(\frac{P_I}{f_Y(Y_I|\boldsymbol{\theta})\Delta_I}\right) \\ &= \sum_{I=1}^{D} f_P(Y_I) \log\left(\frac{f_P(Y_I)}{f_Y(Y_I|\boldsymbol{\theta})\Delta_I}\right) \\ &= \sum_{I=1}^{D} f_P(Y_I) \log\left(f_P(Y_I)\right) - \sum_{I=1}^{D} f_P(Y_I) \log\left(f_Y(Y_I|\boldsymbol{\theta})\Delta_I\right) \end{aligned} \tag{10.19}$$

The name "risk" is justified if we define the quantity $\log\left(\frac{P_I}{f_Y(Y_I|\boldsymbol{\theta})\Delta_I}\right)$ as our *loss* function. The risk function in (10.19) is a function of the unknown parameter $\boldsymbol{\theta}$. Among the values of $\boldsymbol{\theta}$ there is a value, say $\boldsymbol{\theta}_0$ which minimises the risk

[1] 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 $j\left[f_{Y_1}(y), f_{Y_2}(y)\right] = \int \left[f_{Y_1} - f_{Y_2}\right] \log \frac{f_{Y_1}(y)}{f_{Y_2}(y)} dy$

function, i.e. $R_P(\boldsymbol{\theta}_0) = \min_{\boldsymbol{\theta}} [R_P(\boldsymbol{\theta})]$, (see Figure 9.6 for a schematic presentation). At this value the model distribution $f_Y(y|\boldsymbol{\theta})$ approximates the population distribution best. [Note that minimising the risk in (10.19) is equivalent of minimising the quantity $-\sum_{I=1}^{D} f_P(Y_I) \log(f_Y(Y_I|\boldsymbol{\theta})\Delta_I)$ since the quantity $\sum_{I=1}^{D} f_P((0:10)*2*pi/10_I) \log(f_P(Y_I))$ in equation (10.19), (which sometimes is called conditional entropy), does not involve $\boldsymbol{\theta}$.] This is the best that model $f_Y(y|\boldsymbol{\theta})$ 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 $f_Y(y|\boldsymbol{\theta})$ should be explored. Traditional statistical books refer to $\boldsymbol{\theta}_0$ as the "true" $\boldsymbol{\theta}$ parameter value. We will restrain from doing that and we will refer instead to $\boldsymbol{\theta}_0$ as the "best" value for $\boldsymbol{\theta}$ under the model $f_Y(y|\boldsymbol{\theta})$ to emphasise that $f_Y(y|\boldsymbol{\theta})$ 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

$$\begin{aligned} R_E(\boldsymbol{\theta}) &= \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{f_Y(y_i|\boldsymbol{\theta})\Delta_i}\right) \\ &= \sum_{I=1}^{n} \frac{1}{n} \log\left(\frac{\frac{1}{n}}{f_Y(y_i|\boldsymbol{\theta})\Delta_i}\right) \\ &= \frac{1}{n}\sum_{I=1}^{n} \log\left(\frac{1}{n f_Y(y_i|\boldsymbol{\theta})\Delta_i}\right) \\ &= -\log(n) - \frac{1}{n}\sum_{I=1}^{n} \log(f_Y(y_i|\boldsymbol{\theta})\Delta_i) \end{aligned} \tag{10.20}$$

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 θ 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 10.4 where the model with $\hat{\theta}$ is closer to the sample that any other model within the $f_Y(y|\boldsymbol{\theta})$ family. The values denoted as $\hat{\theta}$ in Figure 10.4, is the maximum likelihood estimator, MLE, of θ. 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 $\hat{\boldsymbol{\theta}}$ is generally different from $\boldsymbol{\theta}_0$ but we expect $R_E(\theta) \to R_P(\theta)$ as $n \to \infty$ and therefor the $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ as $n \to \infty$.

10.6.2 Asymptotic normality of MLE under model mis-specification

The asymptotic normality property of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ under model mis-specification was given by equation (10.13) in section 10.5.2.

Proof:

Let $\mathbf{U}_n(\boldsymbol{\theta}) = \sum_{i=1}^n \frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$, the summation of the first derivatives of the log likelihood functions (called the score functions), where $\ell_i(\boldsymbol{\theta}) = \log f_Y(Y_i|\boldsymbol{\theta})$.

The MLE $\hat{\boldsymbol{\theta}}$ solves $\mathbf{U}_n(\hat{\boldsymbol{\theta}}) = 0$. A Taylor expansion gives

$$\mathbf{U}_n(\hat{\boldsymbol{\vartheta}}) = \mathbf{U}_n(\boldsymbol{\theta}_c) + \boldsymbol{I}_n(\tilde{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_c) = 0 \tag{10.21}$$

where $\tilde{\boldsymbol{\theta}}$ is a value of $\boldsymbol{\theta}$ which lies between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_c$ and

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = \sum_{i=i}^n \frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

Hence from (10.21),

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_c) = \left[-\frac{1}{n} \boldsymbol{I}_n(\tilde{\boldsymbol{\theta}}) \right]^{-1} n^{-\frac{1}{2}} \mathbf{U}_n(\boldsymbol{\theta}_c) \tag{10.22}$$

Now by the law of large numbers

$$-\frac{1}{n} \boldsymbol{I}_n(\tilde{\boldsymbol{\theta}}) = -\frac{1}{n} \sum_{i=i}^n \left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]_{\tilde{\boldsymbol{\theta}}} \xrightarrow{p} -E_{f_P} \left[\frac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]_{\boldsymbol{\theta}_c} = \boldsymbol{J}(\boldsymbol{\theta}_c). \tag{10.23}$$

Note $\mathbf{U}_n(\boldsymbol{\theta}_c) = \sum_{i=1}^n \mathbf{U}_i(\boldsymbol{\theta}_c)$ where $\mathbf{U}_i(\boldsymbol{\theta}_c) = \left[\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]_{\boldsymbol{\theta}_c}$ and $E_{f_P}[\mathbf{U}_i(\boldsymbol{\theta}_c)] = \mathbf{0}$ since

$$E_{f_P}[\mathbf{U}_i(\boldsymbol{\theta}_c)] = \int \frac{\partial}{\partial \boldsymbol{\theta}} [\log f_Y(y_i|\boldsymbol{\theta})]_{\boldsymbol{\theta}_c} f_P(y_i) dy_i = \mathbf{0}$$

because $\boldsymbol{\theta}_c$ minimizes the Kullback-Liebler distance

$$d[f_P(y), f_y(y|\boldsymbol{\theta})] = \int [\log f_P(y) - \log f_y(y|\boldsymbol{\theta})] f_P(y) dy$$

with respect to $\boldsymbol{\theta}$. Also $V_{f_P}[\mathbf{U}_i(\boldsymbol{\theta}_c)] = \boldsymbol{K}(\boldsymbol{\theta}_c)$. Hence $E_{f_P}[\mathbf{U}_n(\boldsymbol{\theta}_c)] = 0$ and $V_{f_P}[\mathbf{U}_n(\boldsymbol{\theta}_c)] = n\boldsymbol{K}(\boldsymbol{\theta}_c)$ and so by the central limit theorem,

$$n^{-\frac{1}{2}} \mathbf{U}_n(\boldsymbol{\theta}_c) \xrightarrow{d} N_K(0, \boldsymbol{K}(\boldsymbol{\theta}_c)). \tag{10.24}$$

Hence applying Slutsky's theorem to (10.22) using (10.23) and (10.24) gives

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_c) \xrightarrow{d} \boldsymbol{J}(\boldsymbol{\theta}_c)^{-1} N_K(\mathbf{0}, \boldsymbol{K}(\boldsymbol{\theta}_c)) = N_K\left(0, \boldsymbol{J}(\boldsymbol{\theta}_c)^{-1} \boldsymbol{K}(\boldsymbol{\theta}_c) \boldsymbol{J}(\boldsymbol{\theta}_c)^{-1}\right) \tag{10.25}$$

[Note that if the true population distribution $f_P(y)$ belongs to the parametric family of distributions $f_Y(y|\boldsymbol{\theta})$, then we have $\boldsymbol{\theta}_c = \boldsymbol{\theta}_T$ and $\boldsymbol{K}(\boldsymbol{\theta}_T) = \boldsymbol{J}(\boldsymbol{\theta}_T)$, so $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T) \xrightarrow{d} N_K\left(0, \boldsymbol{J}(\boldsymbol{\theta}_T)^{-1}\right)$.]

10.7 Exercises for Chapter 3

10.7.1 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 $\mathbf{y} = (y_1, y_2, \ldots, y_k)$ 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:

$$\ell(p) = \sum_1^k y_i \log(p) + \sum_1^k (n_i - y_i)\log(1-p)$$

(b) Find the maximum likelihood estimator pf p.

(c) Show that the expected information for the parameter p is given by

$$\boldsymbol{i}(p) = -E\left[\frac{d^2\ell}{dp^2}\right] = \frac{\sum_i^k n_i}{p} + \frac{\sum_1^k n_i}{(1-p)}$$

Note that $E(Y_i) = n_i p$.

10.7.2 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

$$f(Y) = \theta^{-2} Y e^{\frac{-Y}{\theta}}$$

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 θ.

,

Part IV

Advanced topics

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.

11.1 Methods of generating continuous distributions

Here we examine how many of the distributions in Tables 4.1, 4.2 and 4.3 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

There methods are discussed next in Sections 11.2 to 11.8 respectively.

11.2 Distributions generated by Azzalini's method

There are two Azzalili'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 4.5, 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 ν, $f_Z(z)$ is a proper probability density function where

$$f_Z(z) = 2 f_{Z_1}(z) F_{Z_2}(\nu z). \tag{11.1}$$

Let $Y = \mu + \sigma Z$ then

$$f_Y(y) = \frac{2}{\sigma} f_{Z_1}(z)\, F_{Z_2}(\nu z) \tag{11.2}$$

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(μ, σ, ν), is defined by assuming Z_1 and Z_2 have standard normal, NO$(0, 1)$, distributions in (11.2).

Consider $Y \sim$ SN1(μ, σ, ν). First note that $Z = (Y-\mu)/\sigma \sim$ 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 $Z_1 \sim$ NO(0,1) against z. Figure 11.1(b) plots $2 * F_{Z_2}(\nu z)$ of $Z_2 \sim$ NO(0,1) against z for ν=0, 1, 2, 1000. Figure 11.1(c) plots the skew normal type 1 distribution, $f_{Z_1}(z)$ of $Z \sim$ SN1(0,1,ν) against z for ν=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 ν 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 $Z \sim \text{SN1}(0, 1, \nu)$.

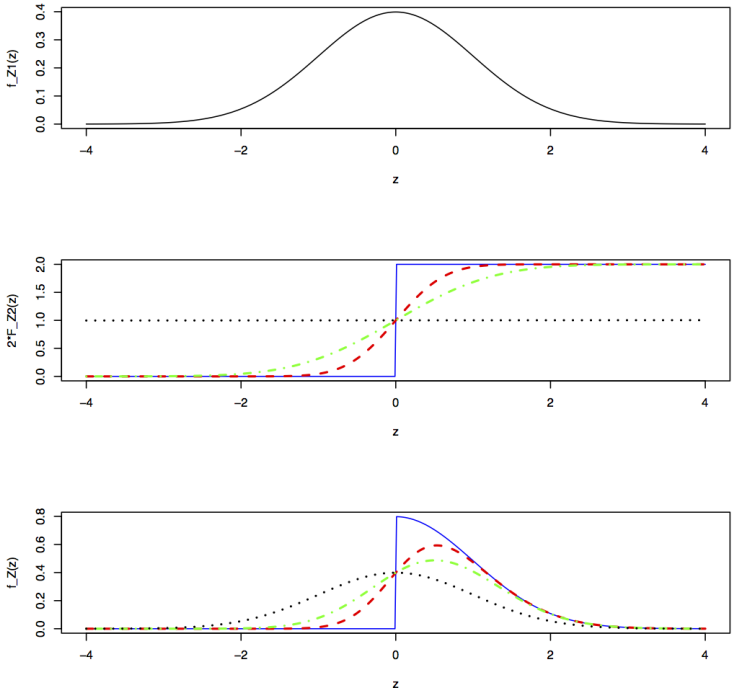


Figure 11.1: Azzalini's method (a) pdf, $f_{Z1}(z)$ of $Z_1 \sim$ NO(0,1) (b) 2*cdf, $2 * F_{Z_2}(\nu z)$ of $Z2 \sim$ NO(0,1) for ν=0, 1, 2, 1000 (note that the solid line $\nu = 1000$ is a step function) (c) The skew normal type 1 distribution, $Z \sim$ SN1(0,1,ν) for ν=0, 1, 2, 1000.

The code for creating figure 11.1 is the following:

```
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_Z1(z)")
```

Distributions of Y	Distribution of Z_1	Distribution of Z_2	$w(z)$
SN1(μ,σ,ν)	NO$(0,1)$	NO$(0,1)$	νz
SEP1(μ,σ,ν,τ)	PE2$\big(0,\tau^{1/\tau},\tau\big)$	PE2$\big(0,\tau^{1/\tau},\tau\big)$	νz
SEP2(μ,σ,ν,τ)	PE2$\big(0,\tau^{1/\tau},\tau\big)$	NO$(0,1)$	$\nu\left(2/\tau\right)^{1/2} sign(z)\lvert z\rvert^{\tau/2}$
ST1(μ,σ,ν,τ)	TF$(0,1,\tau)$	TF$(0,1,\tau)$	νz
ST2(μ,σ,ν,τ)	TF$(0,1,\tau)$	TF$(0,1,\tau+1)$	$\nu\lambda^{1/2}z$

Table 11.1: Showing distributions generated by Azzalini type methods using equation (11.4)

```
plot(2*pNO(z, mu=0, sigma=0.001)~z, type="l",col="blue",ylab="2*F_Z2(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 SEP1(μ,σ,ν,τ), is defined by assuming Z_1 and Z_2 have power exponential type 2, PE2$\big(0,\tau^{1/\tau},\tau\big)$, distributions in (11.2). Azzalini (1986) called this distribution type I.

The skew normal type 1, SN1(μ,σ,ν), is a special case of SEP1(μ,σ,ν,τ) obtained by setting $\tau = 2$. The flexibility of the SEP1 was demonstrated in figure 4.4.

Skew t type 1 (ST1)

The *skew t type 1* family for $-\infty < Y < \infty$, Azzalini (1986), denoted by ST1(μ,σ,ν,τ), is defined by assuming Z_1 and Z_2 have Student t distributions with $\tau > 0$ degrees of freedom, i.e., TF$(0,1,\tau)$, in (11.2).

11.2.2 Azzalini and Capitano (2003) method

Equation (11.1) was generalised, in Azzalini and Capitanio (2003) Proposition 1, to

$$f_Z(z) = 2f_{Z_1}(z)F_{Z_2}[w(z)] \tag{11.3}$$

where $w(z)$ is any odd function of z i.e. $w(-z) = -w(z)$. Hence if $Y = \mu + \sigma z$ then

$$f_Y(y) = \frac{2}{\sigma} f_{Z_1}(z)F_{Z_2}[w(z)] \tag{11.4}$$

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, `SEP2` 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(μ, σ, ν, τ), Azzalini (1986) and DiCiccio and Monti (2004) is expressed in the form (11.4) by letting $Z_1 \sim$ PE2$(0, \tau^{1/\tau}, \tau)$, $Z_2 \sim$ NO$(0,1)$ and $w(z) = \nu(2/\tau)^{1/2} sign(z)|z|^{\tau/2}$. Azzalini (1986) developed a reparametrization of this distribution given by setting $\nu = sign(\lambda)|\lambda|^{\tau/2}$ and called it type II. The skew normal type 1, SN1(μ, σ, ν), distribution is a special case of SEP2(μ, σ, ν, τ) obtained by setting $\tau = 2$.

Skew t type 2 (ST2)

The *skew t type 2* family, denoted by ST2(μ, σ, ν, τ) is expressed in the form (11.4) by letting $Z_1 \sim$ TF$(0, 1, \tau)$, $Z_2 \sim$ 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 μ. A spliced distribution for Y may be defined by

$$f_Y(y) = \pi_1 f_{Y_1}(y)I(y < \mu) + \pi_2 f_{Y_2}(y)I(y \geq \mu). \tag{11.5}$$

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

$$f_Y(y) = \frac{2}{(1+k)} \left\{ f_{Y_1}(y) I(y < \mu) + k f_{Y_2}(y) I(y \geq \mu) \right\}. \tag{11.6}$$

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

$$f_Y(y) = \frac{2}{(1+k)} \left\{ \frac{\nu}{\sigma} f_Z(\nu z) I(y < \mu) + \frac{k}{\nu\sigma} f_Z(z/\nu) I(y \geq \mu) \right\}. \tag{11.7}$$

for $z = (y-\mu)/\sigma$ and where $k = f_{Y_1}(\mu)/f_{Y_2}(\mu) = \nu^2$. Hence

$$f_Y(y) = \frac{2\nu}{\sigma(1+\nu^2)} \left\{ f_Z(\nu z) I(y < \mu) + f_Z(z/\nu) I(y \geq \mu) \right\}. \tag{11.8}$$

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, `SN2`, Skew exponential power type 3, `SEP3` and Skew t type 3, `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 ν to $1/\nu$ reflects $f_Y(y)$ about $y = \mu$.

Skew normal (`SN2`)

A *skew normal type 2* distribution (or two-piece normal distribution) for $-\infty < Y < \infty$, denoted by SN2(μ, σ, ν), is defined by assuming $Z \sim NO(0,1)$ in (11.8) or equivalently $Y_1 \sim$ NO$(\mu, \sigma/\nu)$ and $Y_2 \sim$ NO$(\mu, \sigma\nu)$ in (11.6), giving

$$f_Y(y) = \frac{2\nu}{\sqrt{2\pi}\sigma(1+\nu^2)} \left\{ \exp\left[-\frac{1}{2}(\nu z)^2 \right] I(y < \mu) + \exp\left[-\frac{1}{2}\left(\frac{z}{\nu}\right)^2 \right] I(y \geq \mu) \right\} \tag{11.9}$$

where $z = (y-\mu)/\sigma$. References to this distribution are given in Johnson *at al.* (1994) p 173 and Jones and Faddy (2003). The earliest reference appears to be Gibbons and Mylroie (1973).

For example consider $Y \sim SN2(0, 1, \nu)$ in (11.9), where $Y_1 \sim NO(0, 1/\nu)$ and $Y_2 \sim NO(0, \nu)$, i.e. a spliced two-piece normal distribution. Figure 11.2 plots $Y \sim$ SN2(0,1,ν) for ν=1, 2, 3, 5. Switching from ν to $1/\nu$ reflects $f_Y(y)$ about $y = 0$.

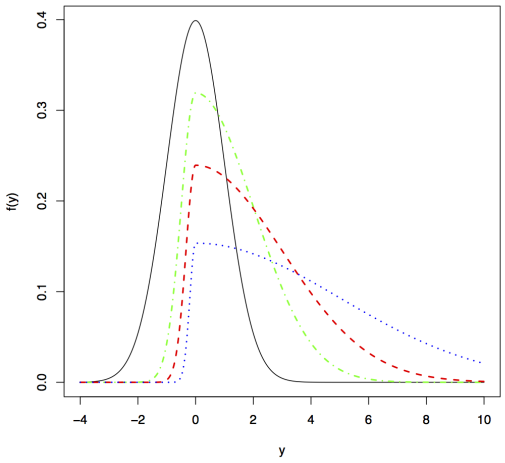


Figure 11.2: Splicing method $Y \sim$ SN2(0,1,ν) for ν=1, 2, 3, 5. Switching from ν to $1/\nu$ reflects $f_Y(y)$ about $y = 0$.

Skew exponential power type 3 (SEP3)

A *skew exponential power type 3* distribution for $-\infty < Y < \infty$, Fernandez, Osiewalski and Steel (1995), denoted by SEP3(μ, σ, ν, τ), is defined by assuming $Z \sim$ PE2$(0, 2^{1/\tau}, \tau)$ in (11.8) or equivalently, $Y_1 \sim$ PE2$(\mu, \sigma 2^{1/\tau}/\nu, \tau)$ and $Y_2 \sim$ PE2$(\mu, \sigma\nu 2^{1/\tau}, \tau)$ in (11.6). Note that the skew normal type 2 distribution, SN2(μ, σ, ν), is a special case of SEP3(μ, σ, ν, τ) given by setting $\tau = 2$.

Skew t type 3 (ST3)

A *skew type 3* distribution for $-\infty < Y < \infty$, Fernandez and Steel (1998), denoted by ST3(μ, σ, ν, τ) 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 μ and σ 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-

Distributions of Y	Distribution of Y_1	Distribution of Y_2	References
SN2(μ, σ, ν)	NO$(\mu, \sigma/\nu)$	NO$(\mu, \sigma\nu)$	Gibbons and Mylroie (1973)
SEP3(μ, σ, ν, τ)	PE2$\left(\mu, \sigma 2^{1/\tau}/\nu, \tau\right)$	PE2$\left(\mu, \sigma\nu 2^{1/\tau}, \tau\right)$	Fernandez, Osiewolski and Steel (1995)
SEP4(μ, σ, ν, τ)	PE2(μ, σ, ν)	PE2(μ, σ, τ)	Jones (2005)
ST3(μ, σ, ν, τ)	TF$(\mu, \sigma/\nu, \tau)$	TF$(\mu, \sigma\nu, \tau)$	Fernandez and Steel (1998)
ST4(μ, σ, ν, τ)	TF(μ, σ, ν)	TF(μ, σ, τ)	Rigby and Stasinopoulos (????)

Table 11.2: Showing distributions generated by splicing

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(μ, σ, ν, τ), is defined by assuming $Y_1 \sim$ PE2 (μ, σ, ν) and $Y_2 \sim$ PE2 (μ, σ, τ) in (11.6). Note that μ is the mode of Y.

A similar distribution was used by Nandi and Mämpel (1995) who set $Y_1 \sim$ PE2(μ, σ, ν) and $Y_2 \sim$ PE2$(\mu, \sigma/q, \tau)$ in (11.6), where $q = \Gamma\left[1 + (1/\tau)\right]/\Gamma\left[1 + (1/\nu)\right]$. However this distribution constrains *both* the median and mode of Y to be μ, which is perhaps rather restrictive.

Skew t type 4 (`ST4`)

A *skew t type 4* family for $-\infty < Y < \infty$, denoted by ST4(μ, σ, ν, τ), is defined by assuming $Y_1 \sim$ TF(μ, σ, ν) and $Y_2 \sim$ TF(μ, σ, τ) 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(μ, σ, ν, τ), 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

$$f_Z(z) = \pi_1 f_{Z_1}(z) I(|z| < \nu) + \pi_2 f_{Z_2}(z) I(\nu < |z| < \tau) + \pi_3 f_{Z_3}(z) I(|z| > \tau) \quad (11.10)$$

where $Z_1 \sim \text{NO}(0,1)$, $Z_2 \sim \text{EXP}(\nu)$, $Z_3 \sim \text{TF}(0,1,\nu\tau-1) \equiv t_{\nu\tau-1}$ and π_1, π_2 and π_3 are defined to ensure $f_Z(z)$ is a proper density function and to ensure continuity of $f_Z(z)$ at ν and τ. In `gamlss()` parameters ν and τ are constants which are either chosen by the user or estimated using the function `prof.dev()`. For fixed ν and τ, the NET distribution has bounded influence functions for both μ and σ, Rigby and Stasinopoulos (1994), and hence provides a robust method of estimating μ and σ 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 γ of a distribution for Y itself comes from a distribution.

Assume that, given γ, Y has conditional probability (density) function $f(y|\gamma)$ and marginally γ 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 γ is continuous and a discrete (or finite) mixture distribution is γ 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 expliocit continuous mixture distributions for Y generated by (11.11) is given in Table 11.3.

Student t family (`TF`)

The *Student t* family for $-\infty < Y < \infty$, denoted $\text{TF}(\mu,\sigma,\nu)$, may be generated from a continuous mixture by assuming $Y|\gamma \sim \text{NO}(\mu,\gamma)$ and $\gamma \sim \sqrt{\nu}\sigma\chi_\nu^{-1} \equiv$

Distributions of Y	Distribution of $Y\|\gamma$	Distribution of γ	References
TF(μ,σ,ν)	NO(μ,γ)	GG$(\sigma,[2\nu]^{-1/2},-2$	Box and Tiao (1973)
GT (μ,σ,ν,τ)	PE2(μ,γ,τ)	GG2$\big(-\tau,\sigma\nu^{1/\tau},\nu\big)$	McDonald(1991)
GB2(μ,σ,ν,τ)	GG2(σ,γ,ν)	GG2$(-\sigma,\mu,\tau)$	McDonald (1996)
EGB2(μ,σ,ν,τ)	EGG2$(1/\sigma,\gamma,\nu)$	GG2$(-1/\sigma,e^{\mu},\tau)$	McDonald (1996)

Table 11.3: Showing distributions generated by continuous mixtures

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 *generalised t* family for $-\infty < Y < \infty$, denoted GT(μ,σ,ν,τ), may be generated by assuming $Y|\gamma \sim$ PE2(μ,γ,τ) has a power exponential type 2 distribution and $\gamma \sim$ GG2$(-\tau,\sigma\nu^{1/\tau},\nu)$ has a generalised gamma type 2 distribution, McDonald (1991).

Generalised Beta type 2 (GB2)

The *generalised beta type 2* family for $Y > 0$, denoted GB2(μ,σ,ν,τ), may be generated by assuming $Y|\gamma \sim$ GG2(σ,γ,ν) 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(μ,σ,ν,τ) 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(μ,σ,ν) then $e^{Z} \sim$ GG2(μ,σ,ν).]

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:

$$\begin{aligned} f_Y(y) &= f_Z(h(y))\,|h'(y)| \\ &= f_Z(h(y))\left|\frac{dz}{dy}\right| \end{aligned} \tag{11.12}$$

Note that $F_Y(y) = F_Z(z)$ provided that the function $g()$ is a monotonic increasing function.

If the location parameter μ 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 μ of Z, i.e. μ_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

$$\frac{d^2\ell_Y}{d\mu_Y^2} = \frac{d^2\ell_Z}{d\mu_Z^2}\left(\frac{d\mu_Z}{d\mu_Y}\right)^2 + \frac{d\ell_Z}{d\mu_z}\frac{d^2\mu_Z}{d\mu_Y^2}.$$

$$E\left[\frac{d^2\ell_Y}{d\mu_Y^2}\right] = E\left[\frac{d^2\ell_Z}{d\mu_Z^2}\right]\left(\frac{d\mu_Z}{d\mu_Y}\right)^2$$

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 the this case is the log-normal distribution (`LOGNO`) where it is assumed that the random variable Z has a normal NOμ, σ 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 obtain by setting $\mu_Y = \exp(\mu)$ i.e. $\mu = \log(\mu_Y)$ in 11.14.

Distributions of Y	Random variable Z	Transformation to Z	References
BCCG	NO$(0,1)$	(11.17)	Cole and Green (1992)
BCPE	PE$(0,1,\tau)$	(11.17)	Rigby and Stasinopoulos (2004)
BCT	TF$(0,1,\tau)$	(11.17)	Rigby and Stasinopoulos (2006)
EGB2	F$(2\nu,2\tau)$	(11.18)	Johnson *et al.* (1995) p.142
GB1	BE(μ,σ)	(11.19)	McDonald and Xu (1995)
GB2	F$(2\nu,2\tau)$	$(\tau/\nu)\,(Y/\mu)^{\sigma}$	McDonald and Xu (1995)
GG	GA$(1,\sigma\nu)$	$(Y/\mu)^{\nu}$	Lopatatazidis and Green (2000)
JSUo	NO$(0,1)$	(11.20)	Johnson (1949)
JSU	NO$(0,1)$	(14.19)	Rigby and Stasinopoulos (2006)
PE	GA$(1,\nu^{1/2})$	$\nu\left\lvert\frac{Y-\mu}{c\sigma}\right\rvert^{\nu}$	Nelson (1991)
SHASHo	NO$(0,1)$	(11.22)	Jones and Pewsey (2009)
SHASH	NO$(0,1)$	(11.23)	Jones (2005)
ST3	BEo(α,β)	(11.24)	Jones and Faddy (2003)

Table 11.4: Showing distributions generated by univariate transformation

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:

$$f_Y(y) \quad = \quad f_Z(\text{logit}(y))\left|\frac{1}{y}+\frac{1}{1-y}\right| \tag{11.15}$$

since $z=\text{logit}(y)=\log(\frac{y}{1-y})$ and $\frac{dz}{dy}=\frac{1}{y}+\frac{1}{1-y}$. Let Z have a normal NO(μ,σ) distribution then Y has a logit-normal distribution with a pdf given by:

$$f_Y(y|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}y\,(1-y)}\;\exp\left\{-\frac{\left[\log(\frac{y}{1-y})-\mu\right]^2}{2\sigma^2}\right\} \tag{11.16}$$

where $0<y<1$. A different parametrisation could be obtained by setting $\mu_Y=\frac{1}{1+\exp(-\mu)}$ i.e. $\mu=\log(\frac{\mu_Y}{1-\mu_Y})$.

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(μ, σ, ν), assumes that Z has a standard normal distribution, NO$(0, 1)$, with mean 0 and standard deviation 1, where

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^{\nu} - 1\right], & \text{if } \nu \neq 0 \\ \frac{1}{\sigma}\log(\frac{Y}{\mu}), & \text{if } \nu = 0. \end{cases} \tag{11.17}$$

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 $Z \sim$ NO(μ, σ), to model the mean μ and the variance σ^2 of Z as functions of explanatory variables for a constant ν. They obtained the maximum likelihood estimate of the power parameter ν from its profile likelihood. This model for $Y > 0$ is denoted by LNO$\{\mu, \sigma, \nu\}$ where ν 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 BCPE(μ, σ, ν, τ), is defined by assuming Z given by (11.17) has a (truncated) standard Power Exponential distribution, 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 BCT(μ, σ, ν, τ), is defined by assuming Z given by (11.17) has a (truncated) standard t distribution with τ degrees of freedom, i.e. 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 EGB2(μ, σ, ν, τ), 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 2ν and 2τ, i.e. $Z \sim F_{2\nu,2\tau}$, where

$$Z = (\tau/\nu) \exp\left[(Y-\mu)/\sigma\right], \tag{11.18}$$

Johnson *et al.* (1995) p142. The distribution has also been called a generalised logistic distribution type IV, see Johnson *et al.* (1995) p 142, who report its long history from Perks (1932). Note also that $R = \exp\left[(Y-\mu)/\sigma\right]$ has a beta distribution of the second kind BE2(ν,τ), Johnson *et al.* (1995) p248 and p325 and $B = R/(1+R)$ has an original beta BEo(ν,τ) distribution.

generalised Beta type 1 (GB1)

The *generalised beta type 1* family for $0 < Y < 1$, denoted by GB1(μ,σ,ν,τ), is defined by assuming Z has a beta, BE(μ,σ), distribution where

$$Z = \frac{Y^\tau}{\nu + (1-\nu)\, Y^\tau} \tag{11.19}$$

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(μ,σ,ν,τ) 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

$$\text{GB1}(\mu,\sigma,\nu,\tau) \equiv \text{GB}\left(\tau, \nu^{1/\tau}, 1-\nu, \mu\left(\sigma^{-2}-1\right), (1-\mu)\left(\sigma^{-2}-1\right)\right).$$

Note also that $\tau = 1$ in GB1(μ,σ,ν,τ) gives a reparametrization 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$\left(\alpha_1/\left(\alpha_1+\alpha_2\right), \left(\alpha_1+\alpha_2-1\right)^{-1/2}, 1/\lambda, 1\right)$. Hence G3B$(\alpha_1,\alpha_2,\lambda)$ is a reparameterized submodel of GB1(μ,σ,ν,τ).

generalised Beta type 2 (GB2)

The *generalised beta type 2* family for $Y > 0$, McDonald (1996), denoted by GB2(μ,σ,ν,τ), is defined by assuming Z has an F distribution with degrees of freedom 2ν and 2τ, i.e. $Z \sim F_{2\nu,2\tau}$, where $Z = (\tau/\nu)\left(Y/\mu\right)^\sigma$.

The distribution is also called the generalised beta distribution of the second kind. Note also that $R = (Y/\mu)^\sigma$ has a beta distribution of the second kind, BE2(ν,τ), Johnson *et al.* (1995) p248 and p325 and $B = R/(1+R)$ has an original beta, BEo(ν,τ), distribution.

generalised Gamma (GG, GG2)

The *generalised gamma* family for $Y > 0$, parameterized by Lopatatzidis and Green (2000), denoted by $\text{GG}(\mu, \sigma, \nu)$, assumes that Z has a gamma $\text{GA}(1, \sigma\nu)$ distribution with mean 1 and variance $\sigma^2\nu^2$, where $Z = (Y/\mu)^\nu$. A reparametrization of $\text{GG}(\mu, \sigma, \nu)$, Johnson *et al.* (1995) p401, given by setting $\mu = \alpha_2\alpha_3^{1/\alpha_1}$, $\sigma = \left(\alpha_1^2\alpha_3\right)^{-1/2}$ and $\nu = \alpha_1$, is denoted $\text{GG2}(\alpha_1, \alpha_2, \alpha_3)$.

Johnson Su (JSUo, JSU)

The original *Johnson Su* family for $-\infty < Y < \infty$, denoted by $\text{JSUo}(\mu, \sigma, \nu, \tau)$, Johnson (1949), is defined by assuming

$$Z = \nu + \tau \sinh^{-1}[(Y - \mu)/\sigma] \tag{11.20}$$

has a standard normal distribution.

The *reparameterized Johnson Su* family, for $-\infty < Y < \infty$, denoted by $\text{JSU}(\mu, \sigma, \nu, \tau)$, has exact mean μ and standard deviation σ for all values of ν and τ, see Appendix 14.4.3 for details.

Power Exponential (PE, PE2)

The *power exponential* family for $-\infty < Y < \infty$, denoted by $\text{PE}(\mu, \sigma, \nu)$ is defined by

$$f_Y(y) = \frac{\nu}{2c\sigma\Gamma(1/\nu)} \exp\left\{-\left|\frac{y-\mu}{c\sigma}\right|^\nu\right\} \tag{11.21}$$

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 μ and σ are the mean and standard deviation of Y respectively for all $v > 0$. This distribution assumes that $Z = \nu\left|\frac{Y-\mu}{c\sigma}\right|^\nu$ has a gamma $\text{GA}(1, \nu^{1/2})$ distribution. A reparametrization of $\text{PE}(\mu, \sigma, \nu)$ used by Nandi and Mämpel (1995), denoted by $\text{PE2}(\alpha_1, \alpha_2, \alpha_3)$, is given by setting $\mu = \alpha_1$, $\sigma = \alpha_2/c$ and $\nu = \alpha_3$.

The Subbotin distribution, Subbotin (1923) and Johnson *et al.* (1995), p195, which uses as parameters (θ, ϕ, δ) is also a reparametrization of $\text{PE2}(\alpha_1, \alpha_2, \alpha_3)$ given by setting $\alpha_1 = \theta$, $\alpha_2 = \phi 2^{\delta/2}$ and $\alpha_3 = 2/\delta$. 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(μ, σ, ν, τ), is defined by assuming that Z has a standard normal distribution NO$(0, 1)$, where

$$Z = \sinh\left\{\tau \sinh^{-1}(R) - \nu\right\} \tag{11.22}$$

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(μ, σ, ν, τ), is defined by assuming that Z has a standard normal distribution NO$(0, 1)$, where

$$Z = \frac{1}{2}\left\{\exp\left[\tau \sinh^{-1}(R)\right] - \exp\left[-\nu \sinh^{-1}(R)\right]\right\} \tag{11.23}$$

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(μ, σ, ν, τ), assumes that Z has a beta BEo(α, β) distribution with $f_Z(z) = z^{\alpha-1}(1 - z)^{\beta-1}/B(\alpha, \beta)$ where

$$Z = \frac{1}{2}\left[1 + R/(\alpha + \beta + R^2)^{1/2}\right] \tag{11.24}$$

where $R = (Y-\mu)/\sigma$, $\alpha = \tau^{-1}\left[1 + \nu(2\tau + \nu^2)^{-1/2}\right]$ and $\beta = \tau^{-1}\left[1 - \nu(2\tau + \nu^2)^{-1/2}\right]$.

11.6 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(μ, σ, ν), is defined by assuming that $Y = \mu + \sigma T$ where $T \sim t_\nu$ has a standard t distribution with ν degrees of freedom, defined itself by $T = Z(W/\nu)^{-1/2}$ where $Z \sim$ NO$(0, 1)$ and $W \sim \chi^2_\nu \equiv$ GA$(\nu, [2/\nu]^{1/2})$, a Chi-square distribution with ν 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(μ, σ, ν, τ), 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(μ, σ, ν, τ) 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.

$$f_Y(y) = \int_{-\infty}^{y} f_{Z_1}(z) f_{Z_2}(y - z) dz. \tag{11.25}$$

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(μ, σ, ν), 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, $NO(\mu, \sigma)$, 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 convension 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_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 the left truncated random variable Y is

$$f_Y(y) = \frac{f_{Y_o}(Y_o)}{1 - F_{Y_o}(c_l)} \tag{11.26}$$

with commutative distribution function,

$$F_Y(y) = \frac{F_{Y_o}(Y_o) - F_{Y_o}(c_l)}{1 - F_{Y_o}(c_l)} \tag{11.27}$$

and inverse commutative distribution function

$$q = F_{Y_o}^{-1}\left\{F_{Y_o}(c_l) + p\left[1 - F_{Y_o}(c_l)\right]\right\}. \tag{11.28}$$

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:

$$\ell_Y = \log f_Y(y) = \log f_{Y_o}(y) + \log\left[1 - F_{Y_o}(c_l)\right] \tag{11.29}$$

For any parameter θ in $(\mu, \sigma.\nu, \tau)$ the first and second derivatives are given by:

$$\frac{\partial \ell_Y}{\partial \theta} = \frac{\partial \ell_{Y_o}}{\partial \theta} - \frac{\partial}{\partial \theta}\left\{\log\left[1 - F_{Y_o}(c_l)\right]\right\} \tag{11.30}$$

and

$$\frac{\partial^2 \ell_Y}{\partial \theta^2} = \frac{\partial^2 \ell_{Y_o}}{\partial \theta^2} - \frac{\partial^2}{\partial \theta^2}\left\{\log\left[1 - F_{Y_o}(c_l)\right]\right\} \tag{11.31}$$

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

$$f_Y(y) = \frac{f_{Y_o}(Y_o)}{F_{Y_o}(c_r)} \tag{11.32}$$

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

$$q = F_{Y_o}^{-1}\left\{p\left[F_{Y_o}(c_r)\right]\right\} \tag{11.34}$$

The log-likelihood for one observation is defined as:

$$\ell_Y = \log f_Y(y) = \log f_{Y_o}(y) - \log\left[F_{Y_o}(c_r)\right] \tag{11.35}$$

For any parameter θ in $(\mu, \sigma.\nu, \tau)$ the first and second derivatives are given by:

$$\frac{\partial \ell_Y}{\partial \theta} = \frac{\partial \ell_{Y_o}}{\partial \theta} - \frac{\partial}{\partial \theta}\left\{\log\left[F_{Y_o}(c_r)\right]\right\} \tag{11.36}$$

and

$$\frac{\partial^2 \ell_Y}{\partial \theta^2} = \frac{\partial^2 \ell_{Y_o}}{\partial \theta^2} - \frac{\partial^2}{\partial \theta}\left\{\log\left[F_{Y_o}(c_r)\right]\right\} \tag{11.37}$$

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

$$q = F_{Y_o}^{-1}\left\{p\left[F_{Y_o}(c_r) - F_{Y_o}(c_l)\right] + F_{Y_o}(c_r)\right\} \tag{11.40}$$

The log-likelihood for one observation is defined as:

$$\ell_Y = \log f_Y(y) = \log f_{Y_o}(y) + \log\left[F_{Y_o}(c_r) - F_{Y_o}(c_l)\right] \tag{11.41}$$

For any parameter θ in $(\mu, \sigma.\nu, \tau)$ the first and second derivatives are given by:

$$\frac{\partial \ell_Y}{\partial \theta} = \frac{\partial \ell_{Y_o}}{\partial \theta} - \frac{\partial}{\partial \theta}\left\{\log\left[F_{Y_o}(c_r) - F_{Y_o}(c_l)\right]\right\} \tag{11.42}$$

and

$$\frac{\partial^2 \ell_Y}{\partial \theta^2} = \frac{\partial^2 \ell_{Y_o}}{\partial \theta^2} - \frac{\partial^2}{\partial \theta}\left\{\log\left[F_{Y_o}(c_r) - F_{Y_o}(c_l)\right]\right\} \tag{11.43}$$

11.8 Systems of distributions

11.8.1 Pearson system

The Pearson system of probability density functions $f_Y(y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}^\top = (\theta_1, \theta_2, \theta_3, \theta_4)$, is defined by solutions of the equation:

$$\frac{d}{dy} f_Y(y|\boldsymbol{\theta}) = -\frac{\theta_1 + y}{\theta_2 + \theta_3 y + \theta_4 y^2} \tag{11.44}$$

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 $\left(\sqrt{\beta_1}, \beta_2\right)$ 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 $\mathrm{BE}(\mu, \sigma)$ 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. $\mathrm{TF}(0, \sigma, \nu)$.

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:

$$f_Y(y|\mu, \sigma) = \exp\left\{ \frac{y\theta - b(\theta)}{\phi} + c(y, \phi) \right\} \tag{11.45}$$

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 μ^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 μ and variance having the form $\phi\mu^\nu$ (where $\phi = \sigma^2$ and σ 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 import 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:

```
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')
```

For continuous distributions defined in the range 0 to ∞ 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:

```
curve(dPARETO2(x, sigma=1, log=T), 2, 30,  col=2, lty=2, lwd=2,
```

[1] Lange, K. L., Little, R. J. A., Taylor, J. M. G., 1989. Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84 (408), 881-896.

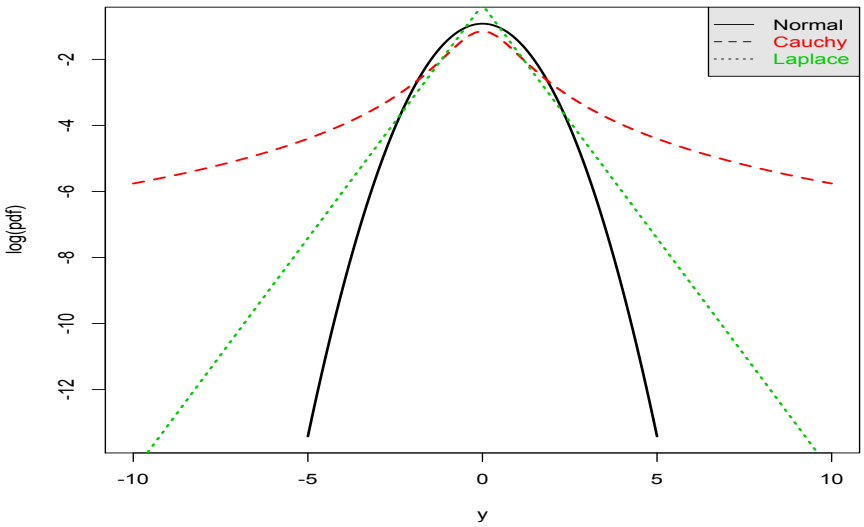


Figure 12.1: Figure showing the log of the standardised version of the normal, Cauchy and Laplace distributions

```
        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"),
        text.col = c(1,2,3),  lty = c(1, 2, 3),
        merge = TRUE, bg = 'gray90')
```

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

$$Y_2 \text{ has a heavier tail than } Y_1 \Leftrightarrow f_{Y_1}(y) = o\left[f_{Y_2}(y)\right] \text{ as } y \to \infty$$

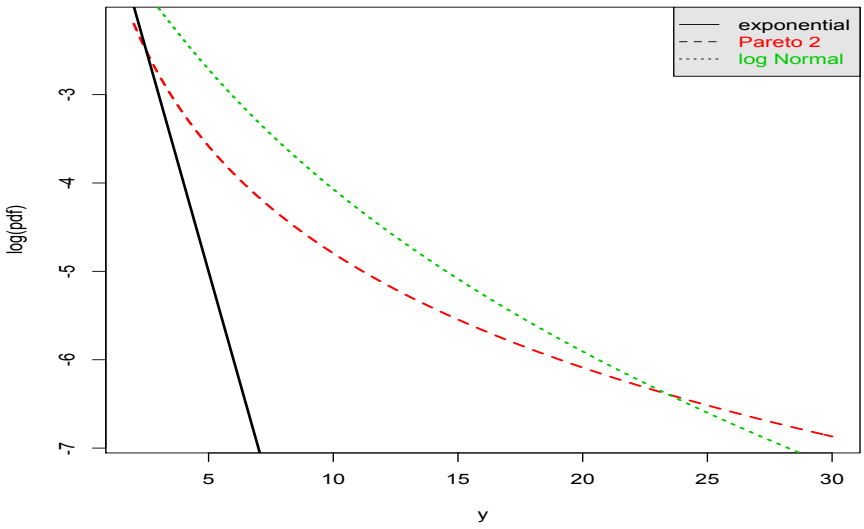


Figure 12.2: Figure showing the log of the standardised version of the normal, Cauchy and Laplace distributions

by Lemma B1 in Appendix B. It also the same ordering as the standard ordering for the *survivor* function $\bar{F_Y}(y) = 1 - F_Y(Y)$ where $F_Y(y)$ is the cumulative distribution function, where

$$Y_2 \text{ has a heaver tail than } Y_1 \Leftrightarrow \bar{F}_{Y_1}(y) = o\left[\bar{F}_{Y_2}(y)\right],$$

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: $\quad -k_2 \ (\log|y|)^{k_1},$

Type II: $\quad -k_4 \ |y|^{k_3},$

Type III: $\quad -k_6 \, \mathrm{e}^{k_5|y|},$

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 $k_1 = 1$, $k_1 = 2$, $k_3 = 1$ and $k_3 = 2$. Figure 12.3 shows tail behaviour of the three different types for $k_1, k_3, k_5 = 1, 2$, and $k_2, k_4, k_6 = 1, 2$.

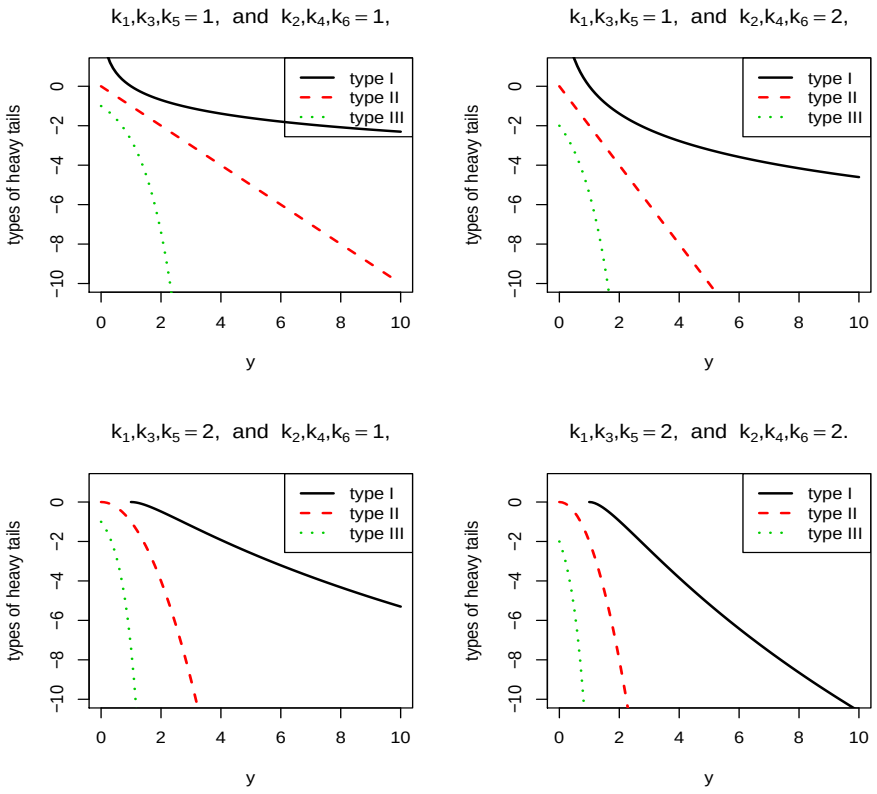


Figure 12.3: Figure showing the shape of the tail for different types of distributions for $k_1, k_3, k_5 = 1, 2$, and $k_2, k_4, k_6 = 1, 2$. Smaller values i the k's result heavier tails.

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(μ, σ, ν, τ), also known as

Value of k_1–k_6	**Distribution name**	**Distribution**	**Condition**	**Value of k_1–k_6**	**Parameter range**
$k_1 = 1$	Cauchy	CA(μ, σ)		$k_2 = 2$	
	Generalized t	GT(μ, σ, ν, τ)		$k_2 = \nu\tau + 1$	$\nu > 0, \tau > 0$
	Skew t type 3	ST3(μ, σ, ν, τ)		$k_2 = \tau + 1$	$\tau > 0$
	Skew t type 4	ST4(μ, σ, ν, τ)	right tail	$k_2 = \tau + 1$	$\tau > 0$
			left tail	$k_2 = \nu + 1$	$\nu > 0$
	Stable	SB(μ, σ, ν, τ)		$k_2 = \tau + 1$	$0 < \tau < 2$
	t	TF(μ, σ, ν)		$k_2 = \nu + 1$	$\nu > 0$
$k_1 = 2$	Johnson's SU	JSU(μ, σ, ν, τ)		$k_2 = 0.5\tau^2$	$\tau > 0$
	Johnson's SU original	JSUo(μ, σ, ν, τ)		$k_2 = 0.5\tau^2$	$\tau > 0$
$0 < k_3 < \infty$	Power exponential	PE(μ, σ, ν)		$k_3 = \nu$, $k_4 = (c_1\,\sigma)^{-\nu}$	$\sigma > 0$, $\nu > 0$
	Power exponential type 2	PE2(μ, σ, ν)		$k_3 = \nu$, $k_4 = \sigma^{-\nu}$	$\sigma > 0$, $\nu > 0$
	Sinh-arcsinh original	SHASHo(μ, σ, ν, τ)		$k_3 = \tau$	$\sigma > 0$, $\nu > 0$, $\tau > 0$
			right tail	$k_4 = \mathrm{e}^{-2\nu}\tau\,\sigma^{-2\tau}$	
			left tail	$k_4 = \mathrm{e}^{2\nu}\tau\,\sigma^{-2\tau}$	
	Sinh-arcsinh	SHASH(μ, σ, ν, τ)	right tail	$k_3 = 2\tau$, $k_4 = 2^{2\tau-3}\sigma^{-2\tau}$	$\sigma > 0$, $\tau > 0$
			left tail	$k_3 = 2\nu$, $k_4 = 2^{2\nu-3}\sigma^{-2\nu}$	$\sigma > 0$, $\nu > 0$
	Skew exponential power type 3	SEP3(μ, σ, ν, τ)		$k_3 = \tau$	$\sigma > 0$, $\nu > 0$, $\tau > 0$
			right tail	$k_4 = 0.5(\sigma\,\nu)^{-\tau}$	
			left tail	$k_4 = 0.5\sigma^{-\tau}\nu^{\tau}$	
	Skew exponential power type 4	SEP4(μ, σ, ν, τ)	right tail	$k_3 = \tau$, $k_4 = \sigma^{-\tau}$	$\tau > 0$
			left tail	$k_3 = \nu$, $k_4 = \sigma^{-\nu}$	$\nu > 0$
$k_3 = 1$	Exponential generalized beta type 2	EGB2(μ, σ, ν, τ)	$\sigma > 0$	$k_4 = \tau\,\sigma^{-1}$	$\tau > 0$
			$\sigma < 0$	$k_4 = \nu\,\lvert\sigma\rvert^{-1}$	$\nu > 0$
	Gumbel	GU(μ, σ)	left tail	$k_4 = \sigma^{-1}$	$\sigma > 0$
	Laplace	LA(μ, σ)		$k_4 = \sigma^{-1}$	$\sigma > 0$
	Logistic	LG(μ, σ)		$k_4 = \sigma^{-1}$	$\sigma > 0$
	Reverse Gumbel	RG(μ, σ)	right tail	$k_4 = \sigma^{-1}$	$\sigma > 0$
$k_3 = 2$	Normal	NO(μ, σ)		$k_4 = 0.5\sigma^{-2}$	$\sigma > 0$
$0 < k_5 < \infty$	Gumbel	GU(μ, σ)	right tail	$k_5 = \sigma^{-1}$, $k_6 = \mathrm{e}^{-\frac{\mu}{\sigma}}$	$-\infty < \mu < \infty$, $\sigma > 0$
	Reverse Gumbel	RG(μ, σ)	left tail	$k_5 = \sigma^{-1}$, $k_6 = \mathrm{e}^{\frac{\mu}{\sigma}}$	$-\infty < \mu < \infty$, $\sigma > 0$

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 $c_1^2 = \Gamma(\frac{1}{\nu})\left[\Gamma(\frac{3}{\nu})\right]^{-1}$

Value of k_1–k_6	**Distribution name**	**Distribution**	**Condition**	**Value of k_1–k_6**	**Parameter range**
$k_1 = 1$	Box-Cox Cole-Green	BCCG(μ, σ, ν)	$\nu < 0$	$k_2 = \lvert\nu\rvert + 1$	
	Box-Cox power exponential	BCPE(μ, σ, ν, τ)	$\nu < 0$	$k_2 = \lvert\nu\rvert + 1$	
	Box-Cox t	BCT(μ, σ, ν, τ)	$\nu \leq 0$	$k_2 = \lvert\nu\rvert + 1$	
			$\nu > 0$	$k_2 = \nu\,\tau + 1$	$\tau > 0$
	Generalized beta type 2	GB2(μ, σ, ν, τ)	$\sigma > 0$	$k_2 = \sigma\tau + 1$	$\tau > 0$
			$\sigma < 0$	$k_2 = \lvert\sigma\rvert\,\nu + 1$	$\nu > 0$
	Generalized gamma	GG(μ, σ, ν)	$\nu < 0$	$k_2 = \left(\sigma^2\,\lvert\nu\rvert\right)^{-1} + 1$	$\sigma > 0$
	Inverse gamma	IGA(μ, σ)		$k_2 = \sigma^{-2} + 1$	$\sigma > 0$
	log t	LOGT(μ, σ, ν)		$k_2 = 1$	
	Pareto Type 2	PA2o(μ, σ)		$k_2 = \sigma + 1$	$\sigma > 0$
$k_1 = 2$	Box-Cox Cole-Green	BCCG(μ, σ, ν)	$\nu = 0$	$k_2 = 0.5\sigma^{-2}$	$\sigma > 0$
	Lognormal	LOGNO(μ, σ)		$k_2 = 0.5\sigma^{-2}$	$\sigma > 0$
	Log Weibull	LOGWEI(μ, σ)	$\sigma > 1$	$k_1 = \sigma$, $k_2 = \mu^{-\sigma}$	
			$\sigma = 1$	$k_1 = 1$, $k_2 = \mu^{-\sigma} + 1$	
			$\sigma < 1$	$k_1 = 1$, $k_2 = 1$	
$1 \leq k_1 < \infty$	Box-Cox power exponential	BCPE(μ, σ, ν, τ)	$\nu = 0$, $\tau > 1$	$k_1 = \tau$, $k_2 = (c_1\sigma)^{-\tau}$	$\sigma > 0$
			$\nu = 0$, $\tau = 1$	$k_1 = 1$, $k_2 = 1 + (c_1\sigma)^{-\tau}$	
			$\nu = 0$, $\tau < 1$	$k_1 = 1$, $k_2 = 1$	$\sigma > 0$
$0 < k_3 < \infty$	Box-Cox Cole-Green	BCCG(μ, σ, ν)	$\nu > 0$	$k_3 = 2\nu$, $k_4 = \left[2\mu^{2\nu}\sigma^2\nu^2\right]^{-1}$	$\mu > 0$, $\sigma > 0$
	Box-Cox power exponential	BCPE(μ, σ, ν, τ)	$\nu > 0$	$k_3 = \nu\tau$, $k_4 = \left[c_1\mu^{\nu}\sigma\nu\right]^{-\tau}$	$\mu > 0$, $\sigma > 0$, $\tau > 0$
	Generalized gamma	GG(μ, σ, ν)	$\nu > 0$	$k_3 = \nu$, $k_4 = \left[\mu^{\nu}\sigma^2\nu^2\right]^{-1}$	$\mu > 0$, $\sigma > 0$
	Weibull	WEI(μ, σ)		$k_3 = \sigma$, $k_4 = \mu^{-\sigma}$	$\mu > 0$, $\sigma > 0$
$k_3 = 1$	Exponential	EX(μ)		$k_4 = \mu^{-1}$	$\mu > 0$
	Gamma	GA(μ, σ)		$k_4 = \mu^{-1}\sigma^{-2}$	$\mu > 0$, $\sigma > 0$
	Generalized inverse Gaussian	GIG(μ, σ, ν)		$k_4 = 0.5c_2\mu^{-1}\sigma^{-2}$	$\mu > 0$, $\sigma > 0$
	Inverse Gaussian	IG(μ, σ)		$k_4 = 0.5\mu^{-2}\sigma^{-2}$	$\mu > 0$, $\sigma > 0$

Table 12.2: Right tail asymptotic form of the log of the probability density function for continuous distributions on the positive real line, where $c_2 = \left[K_{\nu+1}\left(\frac{1}{\sigma^2}\right)\right]\left[K_{\nu}\left(\frac{1}{\sigma^2}\right)\right]^{-1}$ where $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\left\{-\frac{1}{2}\,t\left(x + x^{-1}\right)\right\}dx$

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(μ, σ, ν, τ) 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 μ, σ, ν and τ. Note that μ and σ are (usually location and scale) parameters and not, in general, the mean and standard deviation of the distribution, while ν and τ 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.

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 ($k_3 \geq 1 \,\text{or}\, 0 < \mathrm{k}_5 < \infty$), '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,

[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

wild (pre or proper) and extreme randomness of Mandlebrot (1997), as shown by Table 12.3.

Table 12.3: Mandlebrot's classification of randomness

Type	Conditions	Mandlebrot's rendomness
non-heavy	$(k_3 \geq 1 \text{ or } 0 < \mathrm{k}_5 < \infty)$	mild
heavy	$(k_1 > 1 \text{ and } 0 < k_3 < 1)$	slow
Paretian	$(k_1 = 1 \text{ and } k_2 > 1)$,	wild
heavier than Paretian	$(k_1 = 1 \text{ and } k_2 = 1)$	extreme

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 $\log \bar{F}_Y(y) = o\ (\log |y|)$ and specific asymptotic forms for $\log \bar{F}_Y(y)$ for specific distributions are given in Table 12.4.

Distribution	**Asymptotic form of** $\log \bar{F}_Y(y)$
LOGT(μ, σ, ν)	$-\nu \log(\log y)$
BCT$(\nu = 0)$	$-\tau \log(\log y)$
LOG WEI(μ, σ) for all $0 < \sigma < \infty$	$-\mu^{-\sigma}(\log y)^{\sigma}$
BCPE$(\nu = 0)$ for all $0 < \tau < \infty$	$-(c_1\sigma)^{-\tau}(\log y)^{\tau}$

Table 12.4: Asymptotic form of $\log \bar{F}_Y(y)$ as $y \to \infty$.

Note that the distributions having log survivor function upper tails in exactly the forms $-k_2(\log y)^{k_1}$, $-k_4 y^{k_3}$ and $-k_6 \mathrm{e}^{k_5 y}$ 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.

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.

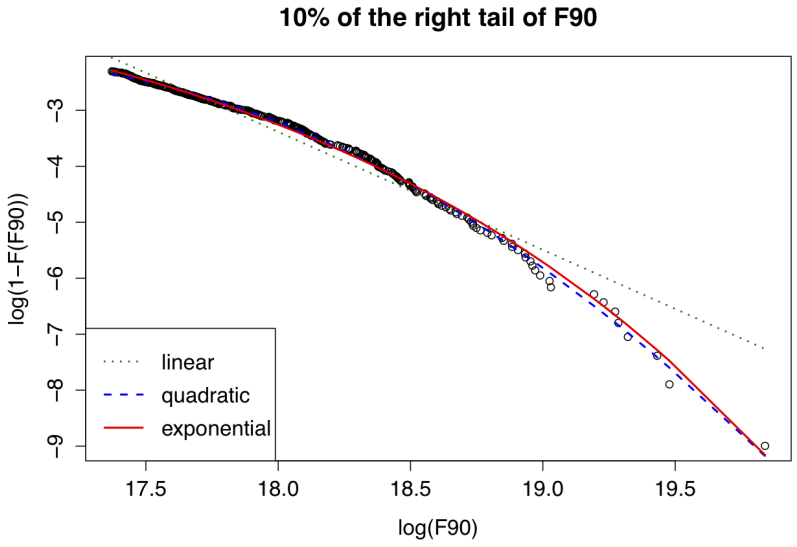


Figure 12.4: Exploratory Method 1 applied to the 90's film revenues data

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.

12.4.1 Exploratory Method 2: log-log-Survival

Correspondingly, the upper tail of $\log\left\{-\log\left[\bar{F}_Y(y)\right]\right\}$ 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 $\log\left\{-\log\left[\bar{F}_Y(y)\right]\right\}$ against $\log[\log(y)]$ or $\log y$ or y will be asymptotically linear in each case (see figure 12.5 for an example).

Table 12.5 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 $\bar{F}_Y(y)$. Note from Table 12.5 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 $\log\left\{-\log\left[\bar{F}_Y(y)\right]\right\}$), in all types cases in Table 12.5 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 $\log\left\{-\log\left[\bar{F}_Y(y)\right]\right\}$) 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 $\log\left\{-\log\left[\bar{F}_Y(y)\right]\right\}$ 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 12.6), with estimates $\hat{k}_3 = 0.561$

Table 12.5: Showing possible relationships of the $\log[\bar{F}_Y(y)]$ against $t = \log y$ for Method 1

Type	$\log\left\{-\log\left[\bar{F}_Y(y)\right]\right\}$	Linear Term
Type I	$\log k_2 + k_1 \log[\log(y)]$	$\log[\log(y)]$
Type II	$\log k_4 + k_3 \log y$	$\log y$
Type III	$\log k_6 + k_5 y$	y

and $\hat{k}_4 = \exp(-8.917) = 0.000134$, suggesting a Box-Cox power exponential tail may be appropriate.

	Intercept	slope	Error SS
type I	-28.2422	10.1806	0.12597
type II	-8.91724	0.560959	0.09090
type III	0.75172	5.697e-09	2.82237

Table 12.6: Estimated coefficients from exploratory method 2.

12.4.2 Exploratory Method 3: truncated distribution fitting

Truncated lognormal and Weibull distributions (see chapter **??** 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]).

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 $f_{Y_1}(y) = o\left[f_{Y_2}(y)\right]$ as $y \to \infty \Leftrightarrow \lim_{y\to\infty}\left[\log f_{Y_2}(y) - \log f_{Y_1}(y)\right] = +\infty$. Similarly replacing $y \to \infty$ by $y \to -\infty$ for the left tail.

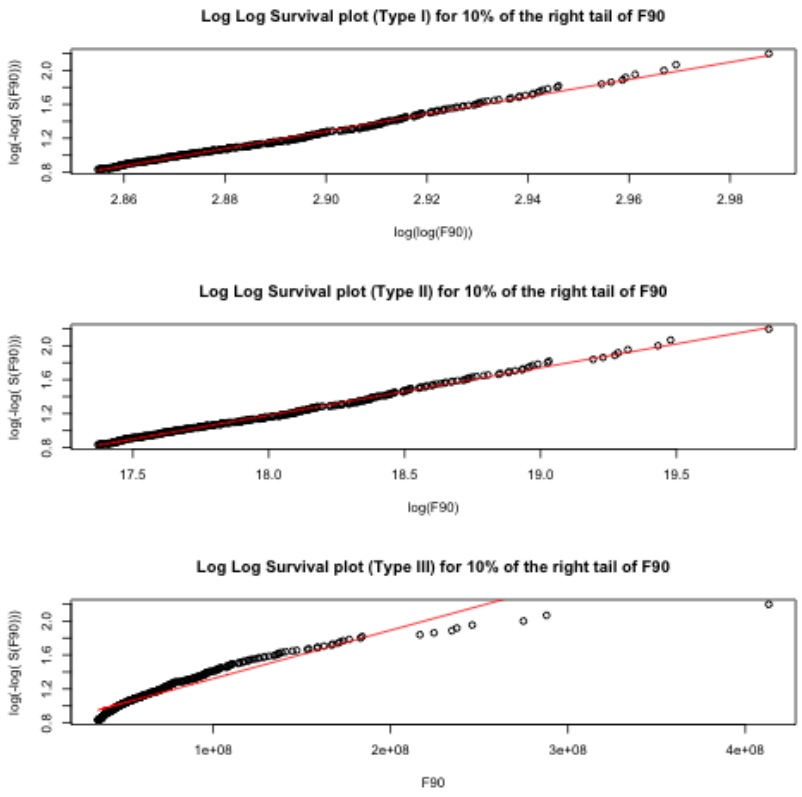


Figure 12.5: Exploratory Method 2 applied to the 90's film revenues data

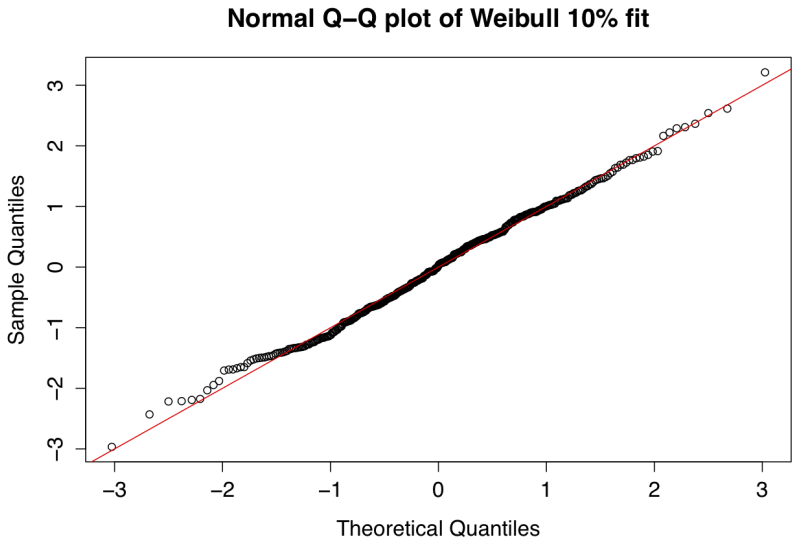


Figure 12.6: QQ plot for the truncated Weibull.

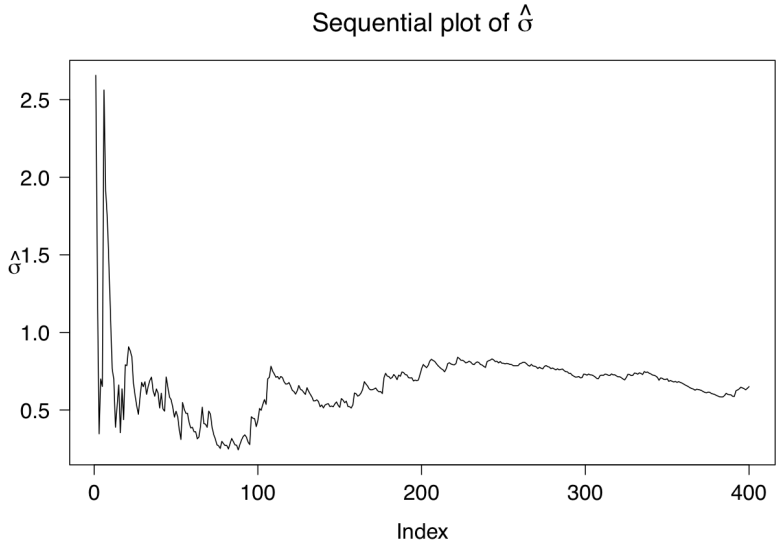


Figure 12.7: Sequencial plot of $\hat{\sigma}$ for the truncated Weibull

Box-Cox Cole-Green	Cole and Green [1992]
Box-Cox power exponential	Rigby and Stasinopoulos [2004]
Box-Cox t	**?**
Cauchy	Johnson *et al.* (1994) Johnson et al. [1994]
Exponential	Johnson *et al.* (1994) Johnson et al. [1994]
Exponential generalized beta type 2	McDonald and Xu [1995], McDonald [1996]
Gamma	Johnson et al. (1994) Johnson et al. [1994]
Generalized beta type 2	McDonald and Xu [1995], McDonald [1996]
Generalized gamma	Lopatatzidis and Green [2000]; Harter [1967]
Generalized inverse Gaussian	Jørgensen [1997]; Jørgensen [1982]
Generalized t	McDonald and Newey [1988], McDonald [1991]
Gumbel	Crowder *et al.* 1991 Crowder, M. J., Kimber, A. C., Smith R. L. and Sweeting [1991]
Inverse Gamma	Johnson *et al.* (1994)
Johnson's SU	Johnson *et al.* (1994) Johnson [1949]
Johnson's SU Original	Johnson *et al.* (1994) Johnson [1949]
Laplace	Johnson *et al.* (1995) Johnson et al. [1995]
Lognormal	Johnson*et al.*(1994)Johnson et al. [1994]
Normal	Johnson *et al.* (1994) Johnson et al. [1994]
Pareto Type 2	Johnson *et al.* (1994) Johnson et al. [1994]
Power exponential	Nelson [1991]
Power exponential type 2	Nelson [1991]; Johnson *et al.* (1995) Johnson et al. [1995]
Reverse Gumbel	Johnson *et al.* (1995) Johnson et al. [1995]
Sinh-arcsinh	Jones [2005]
Sinh-arcsinh original	Jones and Pewsey [2009]
Skew exponential power type 3	Fernandez *et al.* (1995) Fernandez et al. [1995]
Skew exponential power type 4	Jones [2005]
Skew t type 3	Fernandez and Steel [1998]
Skew t type 4	Stasinopoulos *et al.*(2008)
Stable	Nolan [2012]
t	Johnson *et al.* (1995) Johnson et al. [1995]
Weibull	Johnson *et al.* (1994) Johnson et al. [1994]

Table 12.7: References for continuous distributions

Proof B1 $f_{Y_1}(y) = o\left[f_{Y_2}(y)\right]$ as $y \to \infty$

$$\Leftrightarrow \quad \lim_{y\to\infty}\left[\frac{f_{Y_1}(y)}{f_{Y_2}(y)}\right] = 0$$

$$\Leftrightarrow \quad \lim_{y\to\infty}\left[\log\frac{f_{Y_2}(y)}{f_{Y_1}(y)}\right] = +\infty$$

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

$$\begin{aligned} f_{Y_1}(y) = o\left[f_{Y_2}(y)\right] \text{ as } y \to \infty \quad &\Leftrightarrow \bar{F}_{Y_1}(y) = o\left[\bar{F}_{Y_2}(y)\right] \text{ as } y \to \infty \\ f_{Y_1}(y) = o\left[f_{Y_2}(y)\right] \text{ as } y \to -\infty \quad &\Leftrightarrow F_{Y_1}(y) = o\left[F_{Y_2}(y)\right] \text{ as } y \to -\infty \end{aligned}$$

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 $\lim \left[\dfrac{f_{Y_1}(y)}{f_{Y_2}(y)}\right]$ exists.

Proof B2 $\quad f_{Y_1}(y) = o\left[f_{Y_2}(y)\right] \quad$ as $\quad y \to \infty$

$$\begin{aligned} \Leftrightarrow \quad & \lim_{y\to\infty} \frac{f_{Y_1}(y)}{f_{Y_2}(y)} = 0 \\ \Leftrightarrow \quad & \lim_{y\to\infty} \frac{\bar{F}_{Y_1}(y)}{\bar{F}_{Y_2}(y)} = 0 \qquad \text{using L'Hopital's rule} \\ \Leftrightarrow \quad & \bar{F}_{Y_1}(y) = o\left[\bar{F}_{Y_2}(y)\right] \end{aligned}$$

The proof follows similarly for the left tail as $y \to -\infty$.

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

$$\frac{1}{\bar{F}_Y(y)} \sim \left[\frac{1}{f_Y(y)}\right]' \quad \text{as } y \to \infty$$

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 $\quad \lim_{y\to\infty} \frac{f_Y(y)}{\bar{F}_Y(y)} = \lim_{y\to\infty} \frac{f_Y'(y)}{f_Y(y)} \quad$ using L'Hopital's rule

$$\therefore \qquad \lim_{y\to\infty} \left\{ \frac{[f_Y(y)]^2}{\bar{F}_Y(y) f_Y'(y)} \right\} = 1$$

$$\therefore \qquad \frac{1}{\bar{F}_Y(y)} \sim \frac{f_Y^{'}(y)}{[f_Y(y)]^2} = \left[\frac{1}{f_Y(y)}\right]^{'}$$

12.5.4 Corrolary C1

$$\text{If } \log f_y(y) \sim -g(y) \text{ as } y \to \infty \text{ then } \log \bar{F}_Y(y) \sim \begin{cases} -g(y) - \log g^{'}(y) & \text{if } \; g(y) \nsim -\log g^{'}(y) \\ o\left[g(y)\right] & \text{if } \; g(y) \sim -\log g^{'}(y) \end{cases}$$

Proof C1 $\qquad \log f_Y(y) = -g(y)[1 + o\left[1\right]]$

$$\therefore \qquad \frac{1}{f_Y(y)} = \mathrm{e}^{g(y)[1+o[1]]}$$

$$\therefore \qquad \left[\frac{1}{f_Y(y)}\right]^{'} \sim g^{'}(y)\mathrm{e}^{g(y)}$$

since $\qquad \frac{d}{dy}\{g(y)[1 + o\left[1\right]]\} \sim g^{'}(y)$ as $\frac{d}{dy}[1 + o\left[1\right]] = o(1)$

$$\therefore \qquad \bar{F}_Y(y)\, g^{'}(y)\mathrm{e}^{g(y)} \to 1 \qquad \text{as } y \to \infty \text{ using Lemma B3}$$

$$\therefore \qquad \log \bar{F}_Y(y) + g(y) + \log g^{'}(y) \to 0 \qquad \text{as } y \to \infty$$

Hence result.

12.5.5 Corrolary C2

As $y \to \infty$ (or $y \to -\infty$),

(a) $\text{If } \log f_Y(y) \sim -k_2(\log|y|)^{k_1} \text{ then } \log F_Y(y) \sim \begin{cases} -k_2(\log|y|)^{k_1} & \text{if } \; k_1 > 1 \\ -(k_2 - 1)\log|y| & \text{if } \; k_1 = 1 \text{ and } \; k_2 > 1 \\ o(\log|y|) & \text{if } \; k_1 = k_2 = 1 \end{cases}$

(b) If $\log f_Y(y) \sim -k_4\left|y\right|^{k_3}$ then $\log F_Y(y) \sim -k^4\left|y\right|^{k_3}$

(c) If $\log f_Y(y) \sim -k_6\mathrm{e}^{-k_5|y|}$ then $\log F_Y(y) \sim -k_6\mathrm{e}^{-k_5|y|}$

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 $k \geq 1$ 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.

A general centile based measure of skewness is given by MacGillivray [1986]:

$$s_p = \frac{(y_p + y_{1-p})/2 - y_{0.5}}{(y_{1-p} - y_p)/2} \tag{13.1}$$

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 $100(1-2p)\%$. Note that $-1 \leq s_p \leq 1$.

One important case is $p = 0.25$, giving Galton's measure of skewness:

$$s_{0.25} = \frac{(Q_1 + Q_3)/2 - m}{(Q_3 - Q_1)/2} \tag{13.2}$$

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

$$s_{0.01} = \frac{(y_{0.01} + y_{0.99})/2 - y_{0.5}}{(y_{0.99} - y_{0.01})/2} \tag{13.3}$$

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]:

$$k_p = \frac{(y_{1-p} - y_p)}{(Q_3 - Q_1)} \tag{13.4}$$

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

$$s\,k_{0.01} = \frac{(y_{0.99} - y_{0.01})}{3.49\,(Q_3 - Q_1)}\,, \tag{13.5}$$

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

$$t\,s\,k_{0.01} = \frac{s\,k_{0.01} - 1}{s\,k_{0.01}}\,. \tag{13.6}$$

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 $t\,s\,k_{0.01}$ 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.

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 $t\,s\,k_{0.01}$

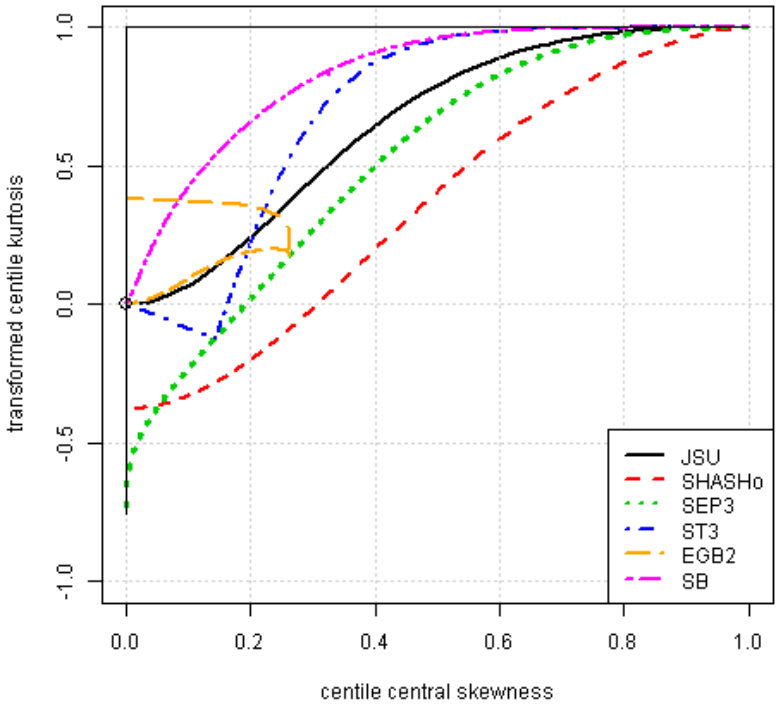


Figure 13.1: The upper boundary of centile central skewness against the transformed centile kurtosis for six distributions on the real line.

(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, ν and τ 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 τ was decreased below 0.001 and ν increased above 100, respectively. Note also that for a fixed τ, ν affects the centile skewenss 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 ν and τ 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.

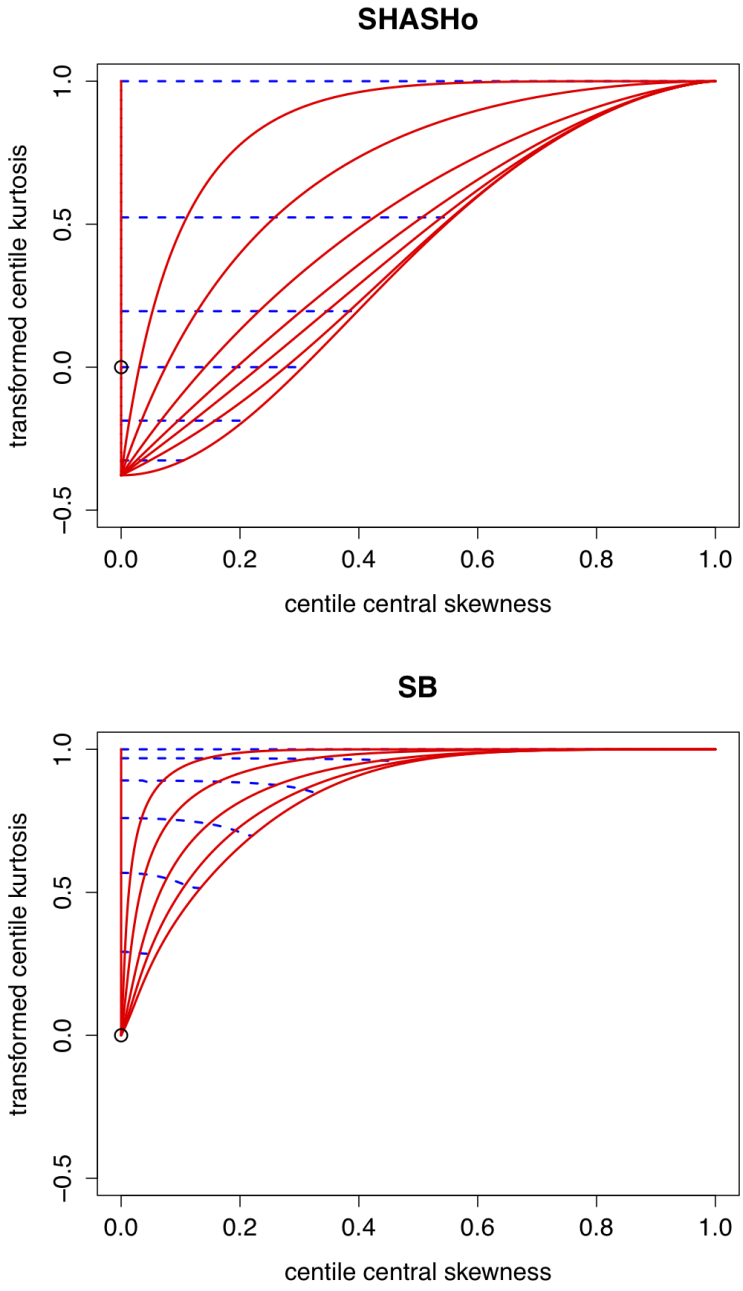


Figure 13.2: Contours of centile central skewness against the transformed centile kurtosis for constant values of ν and τ for the SHASHo and SB distributions.

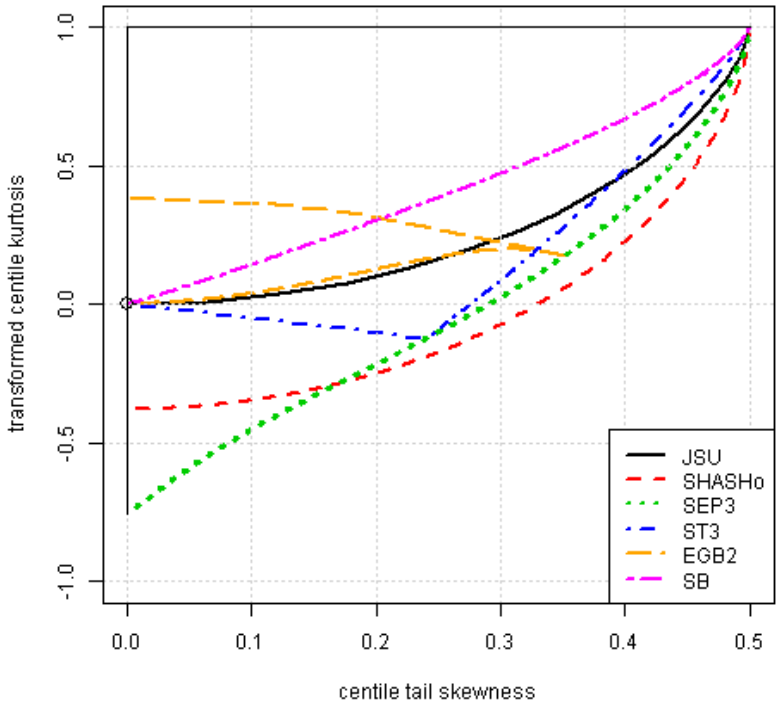


Figure 13.3: The upper boundary of centile tail skewness against the transformed centile kurtosis for six distributions on the real line.

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

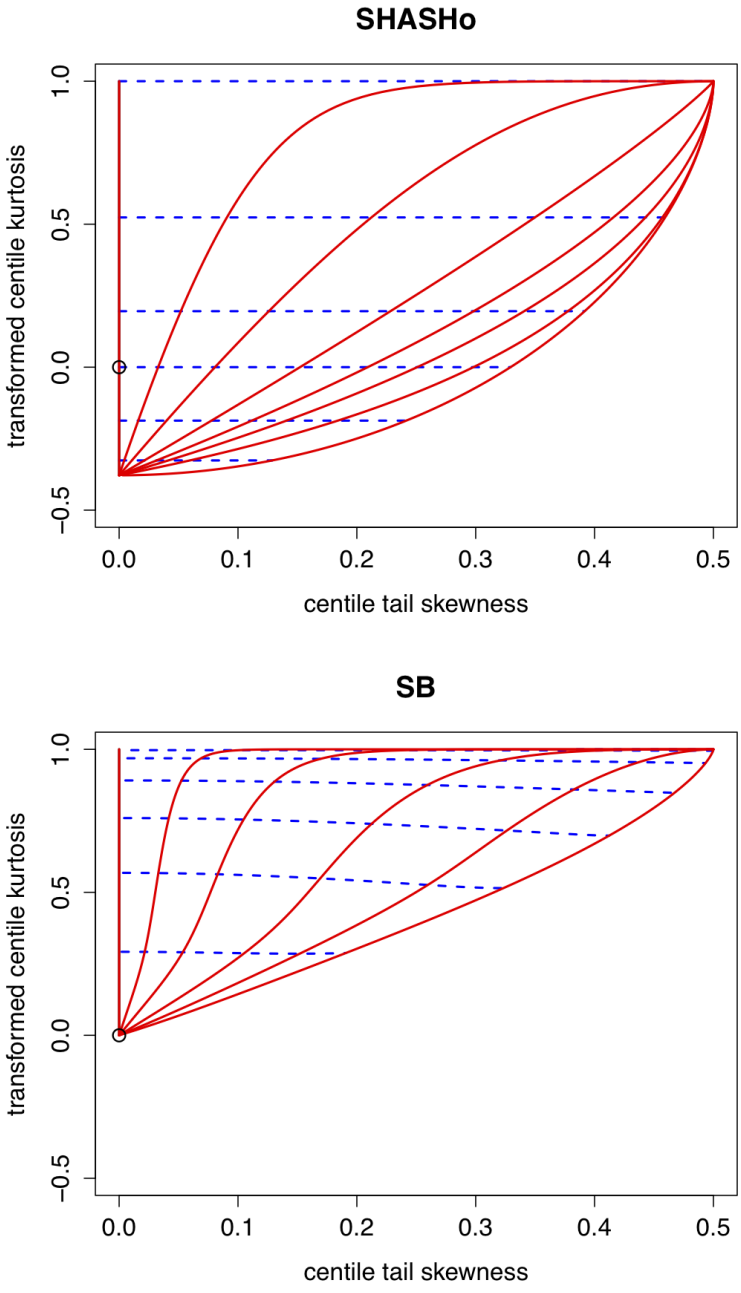


Figure 13.4: Contours of centile tail skewness against the transformed centile kurtosis for constant values of ν and τ for the SHASHo and SB distributions.

Part V

Distributions references guide

Chapter 14

Continuous distributions on $(-\infty, \infty)$

14.1 Location-scale family of distributions

A continuous random variable Y defined on $(-\infty, \infty)$ is said to have a location-scale family of distributions with location shift parameter θ_1 and scaling parameter θ_2 (for fixed values of all other parameters of the distribution) if

$$Z = \frac{(Y - \theta_1)}{\theta_2},$$

has a cumulative distribution function (cdf) which does not depend on θ_1 or θ_2. Hence

$$F_Y(y) = F_Z\left(\frac{y - \theta_1}{\theta_2}\right),$$

and

$$f_Y(y) = \frac{1}{\theta_2} f_Z\left(\frac{y - \theta_1}{\theta_2}\right),$$

so $F_Y(y)$ and $\theta_2 f_Y(y)$ only depend on y, θ_1 and θ_2 through the function $z = (y - \theta_1)/\theta_2$. Note $Y = \theta_1 + \theta_2 Z$.

Example: Let Y have a Gumbel distribution, $Y \sim \texttt{GU}(\mu, \sigma)$, then Y has a location-scale family of distributions with location shift parameter μ and scaling parameter σ, since $F_Y(y) = 1 - \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]$ and hence $F_Z(z) = 1 - \exp\left[-\exp\left(z\right)\right]$ does not depend on either μ or σ. Note $Z \sim \texttt{GU}(0, 1)$.

All distributions with range $(-\infty, \infty)$ in **gamlss.dist** are location-scale families of distributions with location shift parameter μ and scaling parameter σ, except for $\texttt{NO2}(\mu, \sigma)$, and $\texttt{exGAUS}(\mu, \sigma, \nu)$.

14.2 Continuous two parameter distributions on $(-\infty, \infty)$

14.2.1 Gumbel distribution, GU(μ, σ)

Table 14.1: Gumbel distribution

GU(μ, σ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mode, location shift par.
σ	$0 < \sigma < \infty$, scaling parameter
Distribution measures	
mean	$\mu - \gamma\sigma \approx \mu - 0.57722\sigma$
median	$\mu - 0.36611\sigma$
mode	μ
variance	$\pi^2\sigma^2/6 \approx 1.64493\sigma^2$
skewness	-1.13955
excess kurtosis	2.4
MGF	$e^{\mu t}\Gamma(1+\sigma t)$, for $\sigma\lvert t\rvert < 1$
pdf	$\frac{1}{\sigma}\ \exp\left[\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right]$
cdf	$1 - \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]$
Inverse cdf (y_p)	$\mu + \sigma\log[-\log(1-p)]$

The pdf of the Gumbel distribution (or reverse extreme value distribution), denoted by $GU(\mu, \sigma)$, is defined by

$$f_Y(y|\mu,\sigma) = \frac{1}{\sigma}\ \exp\left[\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right] \tag{14.1}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$ and $\sigma > 0$.

Note if $Y \sim$ GU(μ, σ) and $W = -Y$ then $W \sim$ RG$(-\mu, \sigma)$, from which the results in Table 14.1 were obtained. The Gumbel distribution is appropriate for moderately negative skew data.

14.2.2 Logistic distribution, LO(μ, σ)

The pdf of the logistic distribution, denoted by LO(μ, σ), is given by

$$f_Y(y|\mu,\sigma) = \frac{1}{\sigma}\left\{\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}\ \left\{1 + \exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{-2} \tag{14.2}$$

Table 14.2: Logistic distribution

$\texttt{LO}(\mu, \sigma)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift par.
σ	$0 < \sigma < \infty$, scaling parameter
Distribution measures	
mean [a]	μ
median	μ
mode	μ
variance [a]	$\pi^2\sigma^2/3$
skewness [a]	0
excess kurtosis [a]	1.2
MGF	$e^{\mu t}B(1-\sigma t, 1+\sigma t)$
pdf [a]	$\frac{1}{\sigma}\left\{\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}\left\{1+\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{-2}$
cdf [a]	$\left\{1+\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{-1}$
Inverse cdf (y_p)	$\mu + \sigma \log\left(\frac{p}{1-p}\right)$
Reference	[a] Johnson et al. [1995], Chapter 23, p115-117, with $\alpha = \mu$ and $\beta = \sigma$.

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$ and $\sigma > 0$. The logistic distribution is appropriate for a moderately kurtotic response variable distribution. The $\texttt{LO}(\mu, \sigma)$ distribution is symmetric about $y = \mu$.

14.2.3 Normal (or Gaussian) distribution, $\texttt{NO}(\mu, \sigma)$, $\texttt{NO2}(\mu, \sigma)$

First parameterization, $\texttt{NO}(\mu, \sigma)$

Table 14.3: Normal distribution

$\texttt{NO}(\mu, \sigma)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift par.
σ	$0 < \sigma < \infty$, standard deviation, scaling parameter
Distribution measures	
mean	μ
median	μ
mode	μ
variance	σ^2
skewness	0
excess kurtosis	0
MGF	$\exp\left[\mu t + \frac{1}{2}\sigma^2 t^2\right]$
pdf	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$
cdf	$\Phi[(y-\mu)/\sigma]$
Inverse cdf (y_p)	$\mu + \sigma z_p$ where $z_p = \Phi^{-1}(p)$
Reference	Johnson et al. [1994] Chapter 13, p80-89.

The normal distribution is the default distribution of the argument `family` of the function `gamlss()`. The parameterization used for the normal (or Gaussian) probability density function (pdf), denoted by $\texttt{NO}(\mu, \sigma)$, is

$$f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \tag{14.3}$$

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 $Var(Y) = \sigma^2$, so μ is the mean and σ is the standard deviation of Y. The $\texttt{NO}(\mu, \sigma)$ distribution is symmetric about $y = \mu$.

Second parameterization, $\texttt{NO2}(\mu, \sigma)$

The $\texttt{NO2}(\mu, \sigma)$ distribution is a parameterization of the normal distribution where μ represents the mean and σ represents the variance of Y, with pdf given

Table 14.4: Normal distribution - second reparameterization

NO2(μ, σ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift par.
σ	$0 < \sigma < \infty$, variance
Distribution measures	
mean	μ
median	μ
mode	μ
variance	σ
skewness	0
excess kurtosis	0
MGF	$\exp\left[\mu t + \frac{1}{2}\sigma t^2\right]$
pdf	$(1/\sqrt{2\pi\sigma})\exp[-(y-\mu)^2/(2\sigma)]$
cdf	$\Phi[(y-\mu)/\sigma^{1/2}]$
Inverse cdf (y_p)	$\mu + \sigma^{1/2} z_p$ where $z_p = \Phi^{-1}(p)$
Reference	Reparametrize σ to $\sigma^{1/2}$ in $NO(\mu, \sigma)$

by $f_Y(y|\mu,\sigma) = (1/\sqrt{2\pi\sigma})\exp[-(y-\mu)^2/(2\sigma)]$. The NO2$(\mu, \sigma)$ distribution is symmetric about $y = \mu$.

14.2.4 Reverse Gumbel distribution, RG(μ, σ)

The reverse Gumbel distribution, which is also called is the **type I extreme value distribution** is a special case of the generalized extreme value distribution, see Johnson et al. [1995] p2, p11-13 and p75-76. The pdf of the reverse Gumbel distribution, denoted by RG(μ, σ), is defined by

$$f_Y(y|\mu,\sigma) = \frac{1}{\sigma}\ \exp\left\{-\left(\frac{y-\mu}{\sigma}\right) - \exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\} \tag{14.4}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$ and $\sigma > 0$.

Note that if $Y \sim$ RG(μ, σ) and $W = -Y$, then $W \sim GU(-\mu, \sigma)$. The reverse Gumbel distribution is appropriate for moderately positive skew data.

Since the reverse Gumbel distribution is the type I extreme value distribution, it is the reparameterized limiting distribution of the standardized maximum of a sequence of independent and identically distributed random variables from an 'exponential type distribution' including the exponential and gamma.

Table 14.5: Reverse Gumbel distribution

RG(μ, σ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
Distribution measures	
mean	$\mu + \gamma\sigma \approx \mu + 0.57722\sigma$
median	$\mu + 0.36611\sigma$
mode	μ
variance	$\pi^2\sigma^2/6 \approx 1.64493\sigma^2$
skewness	1.13955
excess kurtosis	2.4
MGF	$e^{\mu t}\Gamma(1 - \sigma t)$, for $\sigma\lvert t\rvert < 1$
pdf	$\frac{1}{\sigma}\ \exp\left\{-\left(\frac{y-\mu}{\sigma}\right) - \exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}$
cdf	$\exp\left\{-\exp\left[-\left(\frac{y-\mu}{\sigma}\right)\right]\right\}$
Inverse cdf (y_p)	$\mu - \sigma\log[-\log(p)]$
Reference	Johnson et al. [1995] Chapter 22, p2, p11-13, with $\xi = \mu$ and $\theta = \sigma$.
Note	$\gamma \approx 0.57722$ is Euler's constant.

14.3 Continuous three parameter distributions on $(-\infty, \infty)$

14.3.1 Exponential Gaussian distribution, exGAUS(μ, σ, ν)

Table 14.6: Exponential Gaussian distribution

exGAUS(μ, σ, ν)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean of normal comp., location shift par.
σ	$0 < \sigma < \infty$, standard deviation of normal component
ν	$0 < \nu < \infty$, mean of exponential component
Distribution measures	
mean [a]	$\mu + \nu$
median	–
mode	–
variance [a]	$\sigma^2 + \nu^2$
skewness [a]	$2\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-1.5}$
excess kurtosis [a]	$6\left(1 + \frac{\sigma^2}{\nu^2}\right)^{-2}$
MGF	$(1 - \nu t)^{-1} \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$
pdf [a]	$\frac{1}{\nu} \exp\left[\frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2}\right] \Phi\left(\frac{y-\mu}{\sigma} - \frac{\sigma}{\nu}\right)$
cdf	–
Inverse cdf (y_p)	–
Reference	[a]Lovison and Schindler [2014]

The pdf of the ex-Gaussian distribution, denoted by exGAUS(μ, σ, ν), is defined as

$$f_Y(y|\mu, \sigma, \nu) = \frac{1}{\nu} \exp\left[\frac{\mu - y}{\nu} + \frac{\sigma^2}{2\nu^2}\right] \Phi\left(\frac{y - \mu}{\sigma} - \frac{\sigma}{\nu}\right) \tag{14.5}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$, and where Φ is the cdf of the standard normal distribution. Note $Y = Y_1 + Y_2$ where $Y_1 \sim N(\mu, \sigma^2)$ and $Y_2 \sim EX(\nu)$ are independent.This distribution has also been called the lagged normal distribution, Johnson et al. [1994], p172. See also Davis and Kutner [1976].

Note that if $Y \sim$ exGAUS(μ, σ, ν) then $Y_1 = a + bY \sim$ exGAUS$(a + b\mu, b\sigma, b\nu)$. So $Z = Y - \mu \sim$ exGAUS$(0, \sigma, \nu)$. Hence, for fixed σ and ν, the exGAUS(μ, σ, ν) distribution is a location family of distributions with location parameter μ. Also exGAUS(μ, σ, ν) is a location-scale family of distributions. If exGAUS(μ, σ, ν)

is reparameterized by setting $\alpha_1 = \sigma + \nu$ and $\alpha_2 = \sigma/\nu$ then the resulting $\texttt{exGAUS2}(\mu, \alpha_1, \alpha_2)$ distribution is a location-scale family of distributions, for fixed α_2, with location paramter μ and scale paramter α_1, since if $Y \sim \texttt{exGAUS2}(\mu, \alpha_1, \alpha_2)$ then $Z = (Y - \mu)/\alpha_1 \sim \texttt{exGAUS}(0, 1, \alpha_2)$.

Normal family (of variance-mean relationships), $\texttt{NOF}(\mu, \sigma, \nu)$

Table 14.7: Normal family (of variance-mean relationships) distribution

$\texttt{NOF}(\mu, \sigma, \nu)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$0 < \mu < \infty$, mean, median, mode, location shift par.
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$
Distribution measures	
mean	μ
median	μ
mode	μ
variance	$\sigma^2\mu^\nu$
skewness	0
excess kurtosis	0
MGF	$\exp\left[\mu t + \frac{1}{2}\sigma^2\mu^\nu t^2\right]$
pdf	$\frac{1}{\sqrt{2\pi}\sigma\mu^{\nu/2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2\mu^\nu}\right]$
cdf	$\Phi[(y-\mu)/(\sigma\mu^{\nu/2})]$
Inverse cdf (y_p)	$\mu + \sigma\mu^{\nu/2} z_p$ where $z_p = \Phi^{-1}(p)$
Reference	Reparametrize σ to $\sigma\mu^{\nu/2}$ in $NO(\mu, \sigma)$

The function $\texttt{NOF}(\mu, \sigma, \nu)$ defines a normal distribution family with three parameters. The third parameter ν allows the variance of the distribution to be proportional to a power of the mean. The mean of $\texttt{NOF}(\mu, \sigma, \nu)$ is equal to μ while the variance is equal to $Var(Y) = \sigma^2\mu^\nu$, so the standard deviation is $\sigma\mu^{\nu/2}$. The pdf of the normal, $\texttt{NOF}(\mu, \sigma, \nu)$, distribution is

$$f(y|\mu, \sigma, \nu) = \frac{1}{\sqrt{2\pi}\sigma\mu^{\nu/2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2\mu^\nu}\right] \tag{14.6}$$

for $-\infty < y < \infty$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$. The $\texttt{NOF}(\mu, \sigma, \nu)$ distribution is symmetric about $y = \mu$.

The function $\texttt{NOF}(\mu, \sigma, \nu)$ 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 Ruppert [1988] but here a proper likelihood is maximized. The ν 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 σ and ν parameters, the `mixed()` method argument is essential in the `gamlss()` fitting function. Alternatively a constant ν can be estimated from its profile function, obtained using **gamlss** package function `prof.dev()`.

14.3.2 Power Exponential distribution, PE(μ, σ, ν), PE2(μ, σ, ν)

First parameterization, PE(μ, σ, ν)

Table 14.8: Power exponential distribution

PE(μ, σ, ν)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift par.
σ	$0 < \sigma < \infty$, standard deviation, scaling parameter
ν	$0 < \nu < \infty$, kurtosis parameter
Distribution measures	
mean	μ
median	μ
mode	μ
variance	σ^2
skewness	0
excess kurtosis	$\frac{\Gamma\left(5\nu^{-1}\right)\Gamma\left(\nu^{-1}\right)}{\left[\Gamma\left(3\nu^{-1}\right)\right]^2} - 3$
MGF	–
pdf	$\frac{\nu \exp\left[-\lvert z\rvert^{\nu}\right]}{2c\sigma\Gamma\left(\frac{1}{\nu}\right)}$ where $c^2 = \Gamma(1/\nu)[\Gamma(3/\nu)]^{-1}$ and $z = (y-\mu)/(c\sigma)$
cdf	$\frac{1}{2}\left[1 + \frac{\gamma\left(\nu^{-1}, \lvert z\rvert^{\nu}\right)}{\Gamma\left(\nu^{-1}\right)}\text{sign}(y-\mu)\right]$
Inverse cdf (y_p)	–
Reference	Reparametrize σ to $c\sigma$ in PE2(μ, σ, ν)

The pdf of the power exponential family distribution, denoted by PE(μ, σ, ν), is defined by

$$f_Y(y|\mu, \sigma, \nu) = \frac{\nu \exp\left[-\lvert z\rvert^{\nu}\right]}{2c\sigma\Gamma\left(\frac{1}{\nu}\right)} \tag{14.7}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$ and where $z = (y-\mu)/(c\sigma)$ and $c^2 = \Gamma(1/\nu)[\Gamma(3/\nu)]^{-1}$. Note $\texttt{PE}(\mu, \sigma, \nu) = \texttt{PE2}(\mu, c\sigma, \nu)$.

In this parameterization given by (14.7), used by Nelson [1991], $E(Y) = \mu$ and $Var(Y) = \sigma^2$. The $\texttt{PE}(\mu, \sigma, \nu)$ distribution includes the Laplace (i.e. two sided exponential) and normal $NO(\mu, \sigma)$ distributions, as special cases $\nu = 1$ and $\nu = 2$ respectively, while the uniform distribution is the limiting distribution as $\nu \to \infty$.

The $\texttt{PE}(\mu, \sigma, \nu)$ distribution is symmetric about $y = \mu$. The power exponential distribution is suitable for leptokurtic as well as platykurtic data.

Second parameterization, $\texttt{PE2}(\mu, \sigma, \nu)$

Table 14.9: Second parametrization of power exponential distribution

$\texttt{PE2}(\mu, \sigma, \nu)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift par.
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, kurtosis parameter
Distribution measures	
mean	μ
median	μ
mode	μ
variance [a]	$\frac{\sigma^2}{c^2}$ where $c^2 = \Gamma(1/\nu)[\Gamma(3/\nu)]^{-1}$
skewness	0
excess kurtosis [a]	$\frac{\Gamma\left(5\nu^{-1}\right)\Gamma\left(\nu^{-1}\right)}{\left[\Gamma\left(3\nu^{-1}\right)\right]^2} - 3$
MGF	–
pdf [a]	$\frac{\nu \exp[-\lvert z\rvert^{\nu}]}{2\sigma\Gamma\left(\frac{1}{\nu}\right)}$
cdf	$\frac{1}{2}\left[1 + \frac{\gamma\left(\nu^{-1}, \lvert z\rvert^{\nu}\right)}{\Gamma\left(\nu^{-1}\right)}\text{sign}(y-\mu)\right]$ where $z = \frac{y-\mu}{\sigma}$
Inverse cdf (y_p)	–
Reference	[a] Johnson et al. [1995] Section 24.6, p195-196, equation (24.83), reparametrized by $\theta = \mu$, $\phi = 2^{-1/\nu}\sigma$ and $\delta = 2/\nu$ and hence $\mu = \theta$, $\sigma = \phi 2^{\delta/2}$ and $\nu = 2/\delta$.

An alternative parameterization, the power exponential type 2 distribution,

denoted by $\texttt{PE2}(\mu, \sigma, \nu)$, has pdf defined by

$$f_Y(y|\mu, \sigma, \nu) = \frac{\nu \exp[-|z|^\nu]}{2\sigma\Gamma\left(\frac{1}{\nu}\right)} \tag{14.8}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$ and where $z = (y - \mu)/\sigma$. Note $\texttt{PE2}(\mu, \sigma, \nu) = \texttt{PE}(\mu, \sigma/c, \nu)$.

This is a re-parametrization of a version by Subbotin [1923] given in Johnson et al. [1995] Section 24.6, p195-196, equation (24.83).The cdf of Y is given by $F_Y(y) = \frac{1}{2}[1 + F_S(s)\text{sign}(z)]$ where $S = |Z|^\nu$ has a gamma distribution with pdf $f_S(s) = s^{(1/\nu)-1}\exp(-s)/\Gamma\left(\frac{1}{\nu}\right)$ and $Z = (Y - \mu)/\sigma$, from which the cdf result in Table 14.9 was obtained.

The $\texttt{PE2}(\mu, \sigma, \nu)$ distribution is symmetric about $y = \mu$.

The $\texttt{PE2}(\mu, \sigma, \nu)$ distribution includes the (reparameterized) normal and Laplace (i.e. two sided exponential) distributions as special cases when $\nu = 2$ and $\nu = 1$ respectively, and the uniform (or rectangular) distribution as a limiting case as $\nu \to \infty$.

14.3.3 Skew normal type 1 distribution, $\texttt{SN1}(\mu, \sigma, \nu)$

The pdf of a skew normal type 1 distribution, denoted by $\texttt{SN1}(\mu, \sigma, \nu)$, is defined by

$$f_Y(y|\mu, \sigma, \nu) = \frac{2}{\sigma}\phi(z)\Phi(\nu z) \tag{14.9}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \nu < \infty$, and where $z = (y - \mu)/\sigma$ and ϕ and Φ are the pdf and cdf of a standard normal $\texttt{NO}(0, 1)$ variable, Azzalini [1985]. The skew normal type 1 distribution is a special case of the skew exponential power type 1 distribution where $\tau = 2$, i.e. $\texttt{SN1}(\mu, \sigma, \nu) = \texttt{SEP1}(\mu, \sigma, \nu, 2)$.

Note if $Y \sim \texttt{SN1}(\mu, \sigma, \nu)$ then $-Y \sim \texttt{SN1}(-\mu, \sigma, -\nu)$. The $\texttt{SN1}(\mu, \sigma, \nu)$ distribution includes the normal $\texttt{NO}(\mu, \sigma)$ as a special case when $\nu = 0$, and the half normal distribution as a limiting case as $\nu \to \infty$ and the reflected half normal (i.e. $Y < \mu$) as a limiting case as $\nu \to -\infty$.

14.3.4 Skew normal type 2 distribution, $\texttt{SN2}(\mu, \sigma, \nu)$

The pdf of a skew normal type 2 distribution, denoted by $\texttt{SN2}(\mu, \sigma, \nu)$, is defined by

Table 14.10: Skew normal type 1 distribution

SN1(μ, σ, ν)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, skewness parameter
Distribution measures	
mean [a]	$\mu + \sigma\nu[2(1+\nu^2)^{-1}\pi^{-1}]^{1/2}$
median	–
mode	–
variance [a]	$\sigma^2[1 - 2\nu^2(1+\nu^2)^{-1}\pi^{-1}]$
skewness [a]	$\frac{1}{2}(4-\pi)\left[\frac{\pi}{2}(1+\nu^{-2})-1\right]^{-3/2}$ sign(ν)
excess kurtosis [a]	$2(\pi-3)\left[\frac{\pi}{2}(1+\nu^{-2})-1\right]^{-2}$
MGF [a]	$2\exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)\Phi\left(\frac{\sigma\nu t}{\sqrt{1+\nu^2}}\right)$
pdf [a]	$\frac{2}{\sigma}\phi(z)\Phi(\nu z)$ where $z = (y-\mu)/\sigma$ and ϕ and Φ are the pdf and cdf of NO$(0,1)$
cdf	–
Inverse cdf (y_p)	–
Reference	[a] From Azzalini [1985], p172 and p174, where $\lambda = \nu$ and here $Y = \mu + \sigma Z$

Table 14.11: Skew normal type 2 distribution

SN2(μ, σ, ν)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, skewness parameter
Distribution measures	
mean [a]	$\mu + \sigma E(Z) = \mu + \sigma \frac{\sqrt{2}}{\sqrt{\pi}}(\nu - \nu^{-1})$ where $Z = (Y - \mu)/\sigma$
median [a2]	$\begin{cases} \mu + \frac{\sigma}{\nu}\Phi^{-1}\left(\frac{1+\nu^2}{2}\right) & \text{if } \nu \leq 1 \\ \mu + \sigma\nu\Phi^{-1}\left(\frac{3\nu^2-1}{4\nu^2}\right) & \text{if } \nu > 1 \end{cases}$
mode	μ
variance [a]	$\sigma^2 Var(Z) = \sigma^2\{(\nu^2 + \nu^{-2} - 1) - [E(Z)]^2\}$
skewness [a]	$\begin{cases} \mu_{3Y}/[Var(Y)]^{1.5} \text{ where} \\ \mu_{3Y} = \sigma^3\mu_{3Z} = \sigma^3\{\mu'_{3Z} - 3Var(Z)E(Z) - [E(Z)]^3\} \\ \text{where } \mu'_{3Z} = E(Z^3) = \frac{2\sqrt{2}(\nu^4 - \nu^{-4})}{\sqrt{\pi}(\nu + \nu^{-1})} \end{cases}$
excess kurtosis [a]	$\begin{cases} \{\mu_{4Y}/[Var(Y)]^2\} - 3 \text{ where} \\ \mu_{4Y} = \sigma^4\mu_{4Z} = \sigma^4\{\mu'_{4Z} - 4\mu'_{3Z}E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\} \\ \text{where } \mu'_{4Z} = E(Z^4) = \frac{3(\nu^5 + \nu^{-5})}{(\nu + \nu^{-1})} \end{cases}$
pdf [a]	$\begin{cases} \frac{c}{\sigma}\exp\left[-\frac{1}{2}(\nu z)^2\right], & \text{if } y < \mu \\ \frac{c}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{z}{\nu}\right)^2\right], & \text{if } y \geq \mu \\ \text{where } z = (y-\mu)/\sigma \text{ and } \quad c = \frac{\sqrt{2}\nu}{\sqrt{\pi}(1+\nu^2)} \end{cases}$
cdf [a2]	$\begin{cases} \frac{2\Phi[\nu(y-\mu)/\sigma]}{(1+\nu^2)}, & \text{if } y < \mu \\ \frac{2}{(1+\nu^2)}\left[1 + 2\nu^2\left\{\Phi[(y-\mu)/(\sigma\nu)] - \frac{1}{2}\right\}\right], & \text{if } y \geq \mu \\ \text{where } \Phi \text{ is the cdf of NO}(0,1) \end{cases}$
Inverse cdf [a2] (y_p)	$\begin{cases} \mu + \frac{\sigma}{\nu}\Phi^{-1}\left(\frac{p(1+\nu^2)}{2}\right), & \text{if } p \leq (1+\nu^2)^{-1} \\ \mu + \sigma\nu\Phi^{-1}\left(\frac{p(1+\nu^2)-1+\nu^2}{2\nu^2}\right), & \text{if } p > (1+\nu^2)^{-1} \\ \text{where } \Phi^{-1} \text{ is the inverse cdf of NO}(0,1) \end{cases}$
Reference	[a] Set $\tau = 2$ in SEP3(μ, σ, ν, τ) [a2] Set $\tau \to \infty$ in ST3(μ, σ, ν, τ)

$$f_Y(y|\mu,\sigma\,\nu,\tau) = \begin{cases} \dfrac{c}{\sigma}\exp\left[-\dfrac{1}{2}(\nu z)^2\right], & \text{if } y < \mu \\ \dfrac{c}{\sigma}\exp\left[-\dfrac{1}{2}\left(\dfrac{z}{\nu}\right)^2\right], & \text{if } y \geq \mu \end{cases} \tag{14.10}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$, and where $z = (y-\mu)/\sigma$ and $c = \sqrt{2}\nu/[\sqrt{\pi}(1+\nu^2)]$. This distribution is also called the two piece normal distribution, Johnson et al. [1994], Section 10.3, p173.

The skew normal type 2 distribution is a special case of the skew exponential power type 3 distribution where $\tau = 2$, i.e. $\texttt{SN2}(\mu,\sigma,\nu) = \texttt{SEP3}(\mu,\sigma,\nu,2)$. Also $\texttt{SN2}(\mu,\sigma,\nu)$ is a limiting case of $\texttt{ST3}(\mu,\sigma,\nu,\tau)$ as $\tau \to \infty$. Note if $Y \sim \texttt{SN2}(\mu,\sigma,\nu)$ then $-Y \sim \texttt{SN2}(-\mu,\sigma,\nu^{-1})$. The $\texttt{SN2}(\mu,\sigma,\nu)$ distribution includes the normal $\texttt{NO}(\mu,\sigma)$ as a special case when $\nu = 1$, and is positively skew if $\nu > 1$ and negatively skew if $\nu < 1$.

14.3.5 t family distribution, $\texttt{TF}(\mu,\sigma,\nu)$

The pdf of the t family distribution, denoted by $\texttt{TF}(\mu,\sigma,\nu)$, is defined by

$$f_Y(y|\mu,\sigma,\nu) = \frac{1}{\sigma B\left(1/2,\nu/2\right)\nu^{1/2}}\left[1+\frac{(y-\mu)^2}{\sigma^2\nu}\right]^{-(\nu+1)/2} \tag{14.11}$$

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. Note that $T = (Y-\mu)/\sigma$ has a standard t distribution with ν degrees of freedom with pdf given by Johnson et al. [1995], p 363, equation (28.2).

The $\texttt{TF}(\mu,\sigma,\nu)$ distribution is symmetric about $y = \mu$. The t family distribution is suitable for modelling leptokurtic data, that is, data with higher kurtosis than the normal distribution.

See Lange et al. [1989] for modelling a response variable using the $\texttt{TF}(\mu,\sigma,\nu)$ distribution, including parameter estimation.

14.3.6 t family type 2 distribution, $\texttt{TF2}(\mu,\sigma,\nu)$

The pdf of the t family type 2 distribution, denoted by $\texttt{TF2}(\mu,\sigma,\nu)$, is defined by

$$f_Y(y|\mu,\sigma,\nu) = \frac{1}{\sigma B\left(1/2,\nu/2\right)(\nu-2)^{1/2}}\left[1+\frac{(y-\mu)^2}{\sigma^2(\nu-2)}\right]^{-(\nu+1)/2} \tag{14.12}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 2$, where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function. Note that since $\nu > 2$, the mean $E(Y) = \mu$ and variance $Var(Y) = \sigma^2$ of $Y \sim \texttt{TF2}(\mu,\sigma,\nu)$ are always defined and finite.

Table 14.12: t family distribution

TF(μ, σ, ν)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, kurtosis parameter (degrees of freedom parameter)
Distribution measures	
mean	$\begin{cases} \mu, & \text{if } \nu > 1, \\ \text{undefined}, & \text{if } \nu \leq 1 \end{cases}$
median	μ
mode	μ
variance [a]	$\begin{cases} \dfrac{\sigma^2 \nu}{v-2}, & \text{if } \nu > 2 \\ \infty, & \text{if } \nu \leq 2 \end{cases}$
skewness	$\begin{cases} 0, & \text{if } \nu > 3 \\ \text{undefined}, & \text{if } \nu \leq 3 \end{cases}$
excess kurtosis [a]	$\begin{cases} \dfrac{6}{v-4}, & \text{if } \nu > 4 \\ \infty, & \text{if } \nu \leq 4 \end{cases}$
MGF	–
pdf [a]	$\dfrac{1}{\sigma B\left(1/2, \nu/2\right) \nu^{1/2}} \left[1 + \dfrac{(y-\mu)^2}{\sigma^2 \nu}\right]^{-(\nu+1)/2}$
cdf [a]	$\dfrac{1}{2} + \dfrac{B\left(1/2, \nu/2, z^2/[\nu + z^2]\right)}{2B(1/2, \nu/2)} \text{sign}(z)$ where $z = \dfrac{(y-\mu)}{\sigma}$
Inverse cdf (y_p)	$\begin{cases} \mu + \sigma t_{\nu,p} \text{ where } t_{\nu,p} \text{ is the } p \text{ quantile or } 100p \text{ centile value of } t_\nu, \\ \text{i.e. } P(T < t_{\nu,p}) = p \text{ where } T \sim t_\nu \end{cases}$
Reference	[a]Obtained from Johnson et al. [1995], Section 28.2, p363-365.

Table 14.13: t family type 2 distribution

TF2(μ, σ, ν)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift parameter
σ	$0 < \sigma < \infty$, standard deviation, scaling parameter
ν	$2 < \nu < \infty$, kurtosis parameter (degrees of freedom parameter)
Distribution measures	
mean	μ
median	μ
mode	μ
variance	σ^2
skewness	$\begin{cases} 0, & \text{if } \nu > 3 \\ \text{undefined}, & \text{if } 2 < \nu \leq 3 \end{cases}$
excess kurtosis	$\begin{cases} \dfrac{6}{v-4}, & \text{if } \nu > 4 \\ \infty, & \text{if } 2 < \nu \leq 4 \end{cases}$
MGF	–
pdf	$\dfrac{1}{\sigma B\left(1/2, \nu/2\right)\left(\nu-2\right)^{1/2}}\left[1+\dfrac{(y-\mu)^2}{\sigma^2(\nu-2)}\right]^{-(\nu+1)/2}$
cdf	$\dfrac{1}{2}+\dfrac{B\left(1/2, \nu/2, z^2/[\nu+z^2]\right)}{2B(1/2, \nu/2)}\text{sign}(z)$ where $z=\dfrac{(y-\mu)\nu^{1/2}}{\sigma(\nu-2)^{1/2}}$
Inverse cdf (y_p)	$\mu+\sigma(\nu-2)^{1/2}\nu^{-1/2}t_{\nu,p}$
Reference	Reparametrize σ to $\sigma(\nu-2)^{1/2}\nu^{-1/2}$ in TF(μ, σ, ν).

The TF2(μ, σ, ν) distribution is symmetric about $y = \mu$. The t family type 2 distribution is suitable for modelling leptokurtic data, that is, data with higher kurtosis than the normal distribution.

14.4 Continuous four parameter distributions on $(-\infty, \infty)$

14.4.1 Exponential generalized beta type 2 distribution, EGB2(μ, σ, ν, τ)

Table 14.14: Exponential generalized beta type 2 distribution

EGB2(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$
τ	$0 < \tau < \infty$
Distribution measures	
mean [a2]	$\mu + \sigma[\psi(\nu) - \psi(\tau)]$
median	–
mode	$\mu + \sigma \log\left(\frac{\nu}{\tau}\right)$
variance [a2]	$\sigma^2[\psi^{(1)}(\nu) + \psi^{(1)}(\tau)]$
skewness [a2]	$\frac{\psi^{(2)}(\nu) - \psi^{(2)}(\tau)}{[\psi^{(1)}(\nu) + \psi^{(1)}(\tau)]^{1.5}}$
excess kurtosis[a2]	$\frac{\psi^{(3)}(\nu) + \psi^{(3)}(\tau)}{[\psi^{(1)}(\nu) + \psi^{(1)}(\tau)]^2}$
MGF [a]	$\frac{e^{\mu t} B(\nu + \sigma t, \tau - \sigma t)}{B(\nu, \tau)}$
pdf [a]	$e^{\nu z}\{\lvert\sigma\rvert B(\nu, \tau)[1 + e^z]^{\nu+\tau}\}^{-1}$, where $z = (y - \mu)/\sigma$
cdf	$\frac{B(\nu, \tau, c)}{B(\nu, \tau)}$, where $c = \{1 + \exp[-(y - \mu)/\sigma]\}^{-1}$
Inverse cdf (y_p)	–
Reference	[a] McDonald and Xu [1995], p140-141, p150-151, where $\delta = \mu, \sigma = \sigma, p = \nu$ and $q = \tau$. [a2] McDonald [1996], p436-437, where $\delta = \mu, \sigma = \sigma$, $p = \nu$ and $q = \tau$
Note	$\psi^{(r)}(x) = d^{(r)}\psi(x)/dx^{(r)}$ the r^{th} derivative of the psi function $\psi(x)$

The pdf of the exponential generalized beta type 2 distribution, denoted by

$\texttt{EGB2}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = e^{\nu z}\{|\sigma|B(\nu, \tau)[1 + e^z]^{\nu+\tau}\}^{-1} \tag{14.13}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$, and where $z = (y - \mu)/\sigma$, McDonald and Xu [1995], p141, equation (3.3). Note that McDonald and Xu (1995) allow $\sigma < 0$, however this is unnecessary since $\texttt{EGB2}(\mu, -\sigma, \nu, \tau) = \texttt{EGB2}(\mu, \sigma, \tau, \nu)$. So we assume $\sigma > 0$ and $|\sigma|$ can be replaced by σ in (14.13).

Note also that if $Y \sim \texttt{EGB2}(\mu, \sigma, \nu, \tau)$, then $-Y \sim \texttt{EGB2}(-\mu, \sigma, \tau, \nu)$.

The EGB2 distribution is also called the type IV generalized logistic distribution, Johnson *et al.* (1995), Section 23.10, p142.

If $Y \sim \texttt{EGB2}(\mu, \sigma, \nu, \tau)$ then $R = \{1 + \exp[-(Y - \mu)/\sigma]\}^{-1} \sim BEo(\nu, \tau)$ from which the cdf in Table 14.14 is obtained.

14.4.2 Generalized t distribution, $\texttt{GT}(\mu, \sigma, \nu, \tau)$

This pdf of the generalized t distribution, denoted by $\texttt{GT}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma\, \nu, \tau) = \tau \left\{2\sigma\nu^{1/\tau} B\left(1/\tau, \nu\right)\left[1 + |z|^\tau/\nu\right]^{\nu+(1/\tau)}\right\}^{-1} \tag{14.14}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$, and where $z = (y - \mu)/\sigma$, McDonald and Newey [1988] p430, equation (2.1) where $p = \nu$ and $q = \tau$.

See also Butler et al. [1990] and McDonald [1991].

The $\texttt{GT}(\mu, \sigma, \nu, \tau)$ distribution is symmetric about $y = \mu$.

14.4.3 Johnson SU distribution $\texttt{JSUo}(\mu, \sigma, \nu, \tau)$, $\texttt{JSU}(\mu, \sigma, \nu, \tau)$

First parametrization, $\texttt{JSUo}(\mu, \sigma, \nu, \tau)$

This is the original parameterization of the Johnson S_u distribution, Johnson [1949].

The pdf of the original Johnson's S_u, denoted by $\texttt{JSUo}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma\, \nu, \tau) = \frac{\tau}{\sigma(s^2 + 1)^{1/2}\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right] \tag{14.15}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$ and $\tau > 0$, and where

$$z = \nu + \tau \sinh^{-1}(s) = \nu + \tau \log\left[s + (s^2 + 1)^{1/2}\right], \tag{14.16}$$

Table 14.15: Generalized t distribution

$\texttt{GT}(\mu, \sigma, \nu, \tau)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, first kurtosis parameter
τ	$0 < \tau < \infty$, second kurtosis parameter
Distribution measures	
mean	$\begin{cases} \mu, & \text{if } \nu\tau > 1 \\ \text{undefined}, & \text{if } \nu\tau \le 1 \end{cases}$
median	μ
mode	μ
variance [a2]	$\begin{cases} \dfrac{\sigma^2 \nu^{2/\tau} B(3\tau^{-1}, \nu - 2\tau^{-1})}{B(\tau^{-1}, \nu)}, & \text{if } \nu\tau > 2 \\ \infty, & \text{if } \nu\tau \le 2 \end{cases}$
skewness	$\begin{cases} 0, & \text{if } \nu\tau > 3 \\ \text{undefined}, & \text{if } \nu\tau \le 3 \end{cases}$
excess kurtosis [a2]	$\begin{cases} \{\mu_4/[Var(Y)]^2\} - 3, \text{ where } \mu_4 = \dfrac{\sigma^4 \nu^{4/\tau} B(5\tau^{-1}, \nu - 4\tau^{-1})}{B(\tau^{-1}, \nu)}, & \text{if } \nu\tau > 4 \\ \infty, & \text{if } \nu\tau \le 4 \end{cases}$
MGF	–
pdf [a]	$\tau \left\{ 2\sigma \nu^{1/\tau} B\left(1/\tau, \nu\right) \left[1 + \lvert z \rvert^{\tau}/\nu\right]^{\nu + (1/\tau)} \right\}^{-1}$, where $z = (y - \mu)/\sigma$
cdf	–
Inverse cdf (y_p)	–
Reference	[a] McDonald and Newey [1988],p430, equation (2.1), where $q = \nu$ and $p = \tau$. [a2] McDonald [1991], p 274, where $q = \nu$ and $p = \tau$.

Table 14.16: Johnson SU distribution

JSUo(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$-\infty < \nu < \infty$ skewness parameter
τ	$0 < \tau < \infty$ kurtosis parameter
Distribution measures	
mean [a]	$\mu - \sigma\omega^{1/2}\sinh(\nu/\tau)$, where $w = \exp(1/\tau^2)$
median	$\mu - \sigma\sinh(\nu/\tau)$
mode	–
variance [a]	$\frac{1}{2}\sigma^2(\omega - 1)[\omega\cosh(2\nu/\tau) + 1]$
skewness [a]	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = -\frac{1}{4}\sigma^3\omega^{1/2}(\omega - 1)^2[\omega(\omega + 2)\sinh(3\nu/\tau) + 3\sinh(\nu/\tau)]$
excess kurtosis [a]	$\frac{\mu_4}{[Var(Y)]^2} - 3$ where $\mu_4 = \frac{1}{8}\sigma^4(\omega - 1)^2[\omega^2(\omega^4 + 2\omega^3 + 3\omega^2 - 3)\cosh(4\nu/\tau) + 4\omega^2(\omega + 2)\cosh(2\nu/\tau) + 3(2\omega + 1)]$
MGF	–
cdf	$\Phi(\nu + \tau\sinh^{-1}[(y - \mu)/\sigma])$, where Φ is the cdf of NO$(0, 1)$
Inverse cdf (y_p)	$\mu + \sigma\sinh[(z_p - \nu)/\tau]$ where $z_p = \Phi^{-1}(p)$
Reference	[a] Johnson [1949], p163, equation (37), p152 and p162 where $\xi = \mu, \lambda = \sigma, \gamma = \nu, \delta = \tau$ and $x = y, y = s, z = z$.

where $s = (y-\mu)/\sigma$, from Johnson [1949], p162, equation (33) and p152, [where $\xi = \mu, \lambda = \sigma, \gamma = \nu, \delta = \tau$ and $x = y, y = s, z = z$].

Hence $s = \sinh[(z-\nu)/\tau] = \frac{1}{2}\{\exp[(z-\nu)/\tau] - \exp[-(z-\nu)/\tau]\}$ and $y = \mu + \sigma s$. Note that $Z \sim \texttt{NO}(0,1)$, where $Z = \nu + \tau \sinh^{-1}[(Y-\mu)/\sigma]$, from which the results for the cdf, inverse cdf and median in Table 14.16 are obtained.

Also

$$E\left[\frac{(Y-\mu)^r}{\sigma^r}\right] = \frac{1}{2^r}\sum_{j=0}^{r}(-1)^{r-j}C_j^r \exp\left[\frac{1}{2\tau^2}(r-2j)^2 + \frac{\nu}{\tau}(r-2j)\right] \quad (14.17)$$

for $r = 1, 2, 3, ..,$ where $C_j^r = r!/[j!(r-j)!]$

The parameter ν determines the skewness of the distribution with $\nu > 0$ indicating negative skewness and $\nu < 0$ positive skewness. The parameter τ determines the kurtosis of the distribution. τ 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 leptokurtotic.

Second parametrization, $\texttt{JSU}(\mu, \sigma, \nu, \tau)$

This is a reparameterization of the original Johnson S_u distribution, Johnson [1949], so that parameters μ and σ are the mean and the standard deviation of the distribution.

The $\texttt{JSU}(\mu, \sigma, \nu, \tau)$ is given by re-parameterizing $\texttt{JSUo}(\mu_1, \sigma_1, \nu_1, \tau_1)$ to $\mu = \mu_1 - \sigma_1\omega^{1/2}\sinh(\nu_1/\tau_1)$, $\sigma = \sigma_1/c$, $\nu = -\nu_1$ and $\tau = \tau_1$, where $\omega = \exp(1/\tau^2)$ and c is defined below in equation (14.20).

Hence $\mu_1 = \mu - c\sigma\omega^{1/2}\sinh(\nu/\tau)$, $\sigma_1 = c\sigma$, $\nu_1 = -\nu$ and $\tau_1 = \tau$.

The pdf of the Johnson's S_u, denoted by $\texttt{JSU}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma\, \nu, \tau) = \frac{\tau}{c\sigma(s^2+1)^{1/2}\sqrt{2\pi}} \exp\left[-\frac{1}{2}z^2\right] \quad (14.18)$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$, $\tau > 0$, and where

$$z = -\nu + \tau \sinh^{-1}(s) = -\nu + \tau \log\left[s + (s^2+1)^{1/2}\right], \quad (14.19)$$

$$s = \frac{y - \mu + c\sigma w^{1/2}\sinh(\nu/\tau)}{c\sigma},$$

$$c = \left\{\frac{1}{2}(w-1)\left[w\cosh(2\nu/\tau) + 1\right]\right\}^{-1/2}, \quad (14.20)$$

Table 14.17: Second parametrization Johnson SU distribution

JSU(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, location shift parameter
σ	$0 < \sigma < \infty$, standard deviation, scaling parameter
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean	μ
median	$\begin{cases} \mu_1 + c\sigma \sinh(\nu/\tau) \text{ where} \\ \mu_1 = \mu - c\sigma\omega^{1/2} \sinh(\nu/\tau) \text{ and } \omega = \exp(1/\tau^2) \\ \text{and } c \text{ is given by equation (2.20)} \end{cases}$
Variance	σ^2
skewness	$\begin{cases} \dfrac{\mu_3}{[Var(Y)]^{1.5}} \text{ where} \\ \mu_3 = \dfrac{1}{4} c^3\sigma^3\omega^{1/2}(\omega-1)^2[\omega(\omega+2)\sinh(3\nu/\tau) + 3\sinh(\nu/\tau)] \end{cases}$
excess kurtosis	$\begin{cases} \dfrac{\mu_4}{[Var(Y)]^2} - 3 \text{ where} \\ \mu_4 = \dfrac{1}{8} c^4\sigma^4(\omega-1)^2[\omega^2(\omega^4+2\omega^3+3\omega^2-3)\cosh(4\nu/\tau) \\ +4\omega^2(\omega+2)\cosh(2\nu/\tau) + 3(2\omega+1)] \end{cases}$
MGF	–
cdf	$\Phi(-\nu + \tau \sinh^{-1}[(y-\mu_1)/(c\sigma)])$
Inverse cdf (y_p)	$\mu_1 + c\sigma \sinh[(z_p+\nu)/\tau]$ where $z_p = \Phi^{-1}(p)$
Reference	Reparameterize $\mu_1, \sigma_1, \nu_1, \tau_1$ in JSUo$(\mu_1, \sigma_1, \nu_1, \tau_1)$ by letting $\mu_1 = \mu - c\sigma\omega^{1/2}\sinh(\nu/\tau), \sigma = c\sigma, \nu_1 = -\nu$ and $\tau_1 = \tau$ to give JSU(μ, σ, ν, τ).

where $w = \exp(1/\tau^2)$. Note that $Z \sim \texttt{NO}(0, 1)$. Here $E(Y) = \mu$ and $Var(Y) = \sigma^2$.

The parameter ν determines the skewness of the distribution with $\nu > 0$ indicating positive skewness and $\nu < 0$ negative. The parameter τ determines the kurtosis of the distribution. τ 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 leptokurtic.

14.4.4 Normal-Exponential-t distribution, $\texttt{NET}(\mu, \sigma, \nu, \tau)$

Table 14.18: NET distribution

$\texttt{NET}(\mu, \sigma, \nu, \tau)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, median, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, first kurtosis parameter (fixed constant)
τ	$\max(\nu, \nu^{-1}) < \tau < \infty$, ssecond kurtosis parameter(fixed constant)
Distribution measures	
mean	$\begin{cases} \mu, & \text{if } \nu\tau > 2 \\ \text{undefined}, & \text{if } \nu\tau \le 2 \end{cases}$
median	μ
mode	μ
Variance	$\begin{cases} 2\sigma^2 c \left\{ \sqrt{2\pi}\left[\Phi(\nu) - 0.5\right] + \left(\frac{2}{\nu} + \frac{2}{\nu^3}\right)\exp(-\nu^2/2) + \frac{(\nu^2\tau^2 + 4\nu\tau + 6)}{\nu^3(\nu\tau - 3)}\exp(-\nu\tau + \nu^2/2)\right\} & \text{if } \nu\tau > 3 \\ \infty, & \text{if } \nu\tau \le 3 \end{cases}$
skewness	$\begin{cases} 0 & \text{if } \nu\tau > 4 \\ \text{undefined}, & \text{if } \nu\nu \le 4 \end{cases}$
excess kurtosis	$\begin{cases} - & \text{if } \nu\tau > 5 \\ \infty & \text{if } \nu\tau \le 5 \end{cases}$
MGF	–
pdf	$\frac{c}{\sigma}\begin{cases} \exp\left\{-\frac{z^2}{2}\right\}, & \text{if } \quad \lvert z\rvert \le \nu \\ \exp\left\{-\nu\lvert z\rvert + \frac{\nu^2}{2}\right\}, & \text{if } \quad \nu < \lvert z\rvert \le \tau \\ \exp\left\{-\nu\tau\log\left(\frac{\lvert z\rvert}{\tau}\right) - \nu\tau + \frac{\nu^2}{2}\right\}, & \text{if } \quad \lvert z\rvert > \tau \end{cases}$
cdf	see equation (14.22)
Inverse cdf (y_p)	–
Reference	

The NET distribution is a four parameter continuous distribution, although in **gamlss** it is used as a two parameter distribution with the other two of its parameters fixed as constants. The NET distribution is symmetric abouts its mode μ. It was introduced by Rigby and Stasinopoulos [1994] as a robust method of fitting the location and 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 $\texttt{NET}(\mu, \sigma, \nu, \tau)$. It is normal up to ν, exponential with mean $1/\nu$ from ν to τ and a Student-t with $(\nu\tau - 1)$ degrees of freedom type tail after τ. In **gamlss** the first two parameters, μ and σ can be modelled. Parameters ν and τ may be chosen as fixed constants by the user. [Alternatively estimates of constants for ν and τ can be obtained using the **gamlss** function **prof.dev()**.]

The pdf of the normal exponential t distribution, denoted by $\texttt{NET}(\mu, \sigma, \nu, \tau)$, is given by Rigby and Stasinopoulos [1994] and defined by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{c}{\sigma}\begin{cases} \exp\left\{-\frac{z^2}{2}\right\}, & \text{if} \quad |z| \le \nu \\ \exp\left\{-\nu|z| + \frac{\nu^2}{2}\right\}, & \text{if} \quad \nu < |z| \le \tau \\ \exp\left\{-\nu\tau \log\left(\frac{|z|}{\tau}\right) - \nu\tau + \frac{\nu^2}{2}\right\}, & \text{if} \quad |z| > \tau \end{cases} \tag{14.21}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, $\tau > \max(\nu, \nu^{-1})$ and where $z = (y-\mu)/\sigma$ and $c = (c_1 + c_2 + c_3)^{-1}$, where $c_1 = \sqrt{2\pi}[2\Phi(\nu) - 1]$, $c_2 = \frac{2}{\nu}\exp\left\{-\frac{\nu^2}{2}\right\}$ and $c_3 = \frac{2}{\nu(\nu\tau-1)}\exp\left\{-\nu\tau + \frac{\nu^2}{2}\right\}$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The cdf of $Y \sim \texttt{NET}(\mu, \sigma, \nu, \tau)$ is given by $F_Y(y) = F_Z(z)$ where $Z = (Y-\mu)/\sigma$, $z = (y-\mu)/\sigma$ and

$$F_Z(z) = \begin{cases} \frac{c\tau^{\nu\tau}|z|^{-\nu\tau+1}}{(\nu\tau - 1)}\exp(-\nu\tau + \nu^2/2) & \text{if } z < -\tau \\ \frac{c}{\nu(\nu\tau-1)}\exp(-\nu\tau + \nu^2/2) + \frac{c}{\nu}\exp(-\nu|z| + \nu^2/2) & \text{if } -\tau \le z < -\nu \\ \frac{c}{\nu(\nu\tau-1)}\exp(-\nu\tau + \nu^2/2) + \frac{c}{\nu}\exp(-\nu^2/2) & \\ \quad + c\sqrt{2\pi}[\Phi(z) - \Phi(-\nu)] & \text{if } -\nu \le z \le 0 \\ 1 - F_Z(z) & \text{if } z > 0. \end{cases} \tag{14.22}$$

14.4.5 Sinh-Arcsinh, $\texttt{SHASH}(\mu, \sigma, \nu, \tau)$

The pdf of the sinh-arcsinh distribution Jones [2005], denoted by $\texttt{SHASH}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu,\sigma\,\nu,\tau) = \frac{c}{\sqrt{2\pi}\sigma(1+z^2)^{1/2}}\exp\left(-\frac{1}{2}r^2\right) \tag{14.23}$$

Table 14.19: Sinh-Arcsinh distribution

SHASH(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, median, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, left tail heaviness parameter
τ	$0 < \tau < \infty$, right tail heaviness parameter
Distribution measures	
median	μ
pdf	$\begin{cases} \frac{c}{\sigma\sqrt{2\pi}(1+z^2)^{1/2}} \exp\left(-\frac{1}{2}r^2\right) \\ \text{where } \tau = \frac{1}{2}\left\{\exp[\tau \sinh^{-1}(z)] - \exp[-\nu \sinh^{-1}(z)]\right\} \\ \text{and } z = (y-\mu)/\sigma \text{ and } c = \frac{1}{2}\left\{\tau \exp[\tau \sinh^{-1}(z)] + \nu \exp[-\nu \sinh^{-1}(z)]\right\} \end{cases}$
cdf	$\Phi(\tau)$ where Φ is the cdf of NO$(0,1)$
Reference	Jones and Pewsey [2009], page 777 with $(\xi, \eta, \gamma, \delta, x, z)$ replaced by $(\mu, \sigma, \nu, \tau, z, r)$.

where

$$r = \frac{1}{2}\left\{\exp\left[\tau \sinh^{-1}(z)\right] - \exp\left[-\nu \sinh^{-1}(z)\right]\right\} \tag{14.24}$$

and

$$c = \frac{1}{2}\left\{\tau \exp\left[\tau \sinh^{-1}(z)\right] + \nu \exp\left[-\nu \sinh^{-1}(z)\right]\right\}$$

and $z = (y-\mu)/\sigma$ for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$. [Note $\sinh^{-1}(z) = \log(u)$ where $u = z + \left(z^2+1\right)^{1/2}$. Hence $r = \frac{1}{2}\left(u^\tau - u^{-\nu}\right)$.] Note that $R \sim$ NO$(0,1)$ where R is obtained from (14.24). Hence μ is the median of Y (since $y = \mu$ gives $z = 0, u = 1, \sinh^{-1}(z) = 0$ and hence $r = 0$). Note also R is the normalized quantile residual (or Z score).

The parameter ν controls the left tail heaviness, with the left tail being heavier than the normal distribution if $\nu < 1$ and lighter if $\nu > 1$. Similarly τ controls the right tail.

14.4.6 Sinh-Arcsinh SHASHo(μ, σ, ν, τ)

The original sinh-arcsinh distribution, developed by Jones and Pewsey [2009] is denoted by SHASHo(μ, σ, ν, τ), with pdf given by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{\tau c}{\sigma\sqrt{2\pi}(1+z^2)^{1/2}} \exp\left(-\frac{1}{2}r^2\right) \tag{14.25}$$

Table 14.20: Sinh-Arcsinh original distribution

SHASHo(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	$\mu + \sigma E(Z) = \mu + \sigma \sinh(\nu/\tau) P_{1/\tau}$ where $Z = (Y-\mu)/\sigma$ and $P_q = \frac{e^{1/4}}{(8\pi)^{1/2}} \left[K_{(q+1)/2}(0.25) + K_{(q-1)/2}(0.25)\right]$ and $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\left\{-\frac{1}{2}t(x+x^{-1})\right\} dx$ is the modified Bessel function of the second kind
median	$\mu + \sigma \sinh(\nu/\tau) = \mu + \frac{\sigma}{2}[e^{\nu/\tau} - e^{-\nu/\tau}]$
mode	μ only when $\nu = 0$
Variance, $Var(Y)$	$\sigma^2 Var(Z) = \frac{\sigma^2}{2}[\cosh(2\nu/\tau)P_{2/\tau} - 1] - \sigma^2[\sinh(\nu/\tau)P_{1/\tau}]^2$.
skewness	$\mu_{3Y}/[Var(Y)]^{1.5}$ where $\mu_{3Y} = \sigma^3\mu_{3Z} = \sigma^3\{\mu'_{3Z} - 3Var(Z)E(Z) - [E(Z)]^3\}$ where $\mu'_{3Z} = E(Z^3) = \frac{1}{4}[\sinh(3\nu/\tau)P_{3/\tau} - 3\sinh(\nu/\tau)P_{1/\tau}]$
excess kurtosis	$\{\mu_{4Y}/[Var(Y)]^2\} - 3$ where $\mu_{4Y} = \sigma^4\mu_{4Z} = \sigma^4\{\mu'_{4Z} - 4\mu'_{3Z}E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\}$ where $\mu'_{4Z} = E(Z^3) = \frac{1}{8}[\cosh(4\nu/\tau)P_{4/\tau} - 4\cosh(2\nu/\tau)P_{2/\tau} + 3]$
pdf	$\frac{\tau c}{\sigma\sqrt{2\pi}(1+z^2)^{1/2}} e^{-r^2/2}$ where $r = \sinh[\tau \sinh^{-1}(z) - \nu]$ and $z = (y-\mu)/\sigma$ and $c = \cosh[\tau \sinh^{-1}(z) - \nu]$
cdf	$\Phi(r)$, where Φ is the cdf of NO$(0,1)$
Inverse cdf (y_p)	$\mu + \sigma \sinh\left\{\frac{\nu}{\tau} + \frac{1}{\tau}\sinh^{-1}\left[\Phi^{-1}(p)\right]\right\}$ where Φ^{-1} is the inverse cdf of NO$(0,1)$
Reference	Jones and Pewsey [2009], page 762-764, with $(\xi, \eta, \gamma, \delta, x, z)$ replaced by $(\mu, \sigma, \nu, \tau, z, r)$.

where

$$r = \sinh[\tau \sinh^{-1}(z) - \nu]$$

and

$$c = \cosh[\tau \sinh^{-1}(z) - \nu]$$

and $z = (y-\mu)/\sigma$ for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$ and $\tau > 0$. [Note $\sinh^{-1}(z) = \log(\mu)$ where $u = z + (z^2+1)^{1/2}$. Hence $z = \frac{1}{2}(u - u^{-1})$.] Note that $R = \sinh[\tau \sinh^{-1}(Z) - \nu)] \sim \texttt{NO}(0,1)$ where $Z = (Y-\mu)/\sigma$. Hence R is the normalized quantile residual (or z-score).

Note that ν is a skewness parameter with $\nu > 0$ and $\nu < 0$ corresponding to positive and negative skewness respectively. Also τ is a kurtosis parameter with $\tau < 1$ and $\tau > 1$ corresponding to heavier and lighter tails than the normal distribution, Jones and Pewsey [2009], p762.

14.4.7 Sinh-Arcsinh original type 2 distribution, $\texttt{SHASHo2}(\mu, \sigma, \nu, \tau)$

Jones and Pewsey [2009], page 768, suggest reparameterizing $\texttt{SHASHo}(\mu, \sigma, \nu, \tau)$ in order to provide a more orthogonal parameterization, $\texttt{SHASHo2}(\mu, \sigma, \nu, \tau)$, with pdf given by (14.25) with σ replaced by $\sigma\tau$. They say that this solved numerical problems encountered in the original parameterization i.e. $\texttt{SHASHo}(\mu, \sigma, \nu, \tau)$ when $\tau > 1$. The summary table for $\texttt{SHASHo2}(\mu, \sigma, \nu, \tau)$ is given by replacing σ by $\sigma\tau$ in Table 14.20.

14.4.8 Skew Exponential Power type 1 distribution, $\texttt{SEP1}(\mu, \sigma, \nu, \tau)$

The pdf of the skew exponential power type 1 distribution, denoted by $\texttt{SEP1}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{2}{\sigma}\, f_{Z_1}(z)\, F_{Z_1}(\nu z) \tag{14.26}$$

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 $Z_1 \sim \texttt{PE2}(0, \tau^{1/\tau}, \tau)$, a power exponential type 2 distribution with $f_{Z_1}(z) = \alpha^{-1} \exp[-|z|^\tau/\tau]$, where $\alpha = 2\tau^{(1/\tau)-1}\Gamma(1/\tau)$. The $\texttt{SEP1}(\mu, \sigma, \nu, \tau)$ distribution was introduced by Azzalini [1986] as his type I distribution.

The skew normal type 1 distribution, denoted by $\texttt{SN1}(\mu, \sigma, \nu)$, is a special case of $\texttt{SEP1}(\mu, \sigma, \nu, \tau)$ given by $\tau = 2$.

Table 14.21: Skew Exponential Power type 1 distribution

SEP1(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	$\begin{cases} \mu + \sigma E(Z) = \mu + \dfrac{\sigma \sinh(\nu)\tau^{1/\tau}\Gamma(2\tau^{-1})B(\tau^{-1}, 2\tau^{-1}, \nu^\tau/[1+\nu^\tau])}{\Gamma(\tau^{-1})B(\tau^{-1}, 2\tau^{-1})} \\ \text{where } Z = (Y-\mu)/\sigma \end{cases}$
median	–
mode	–
Variance, $Var(Y)$	$\sigma^2 Var(Z) = \sigma^2 \left\{ \dfrac{\tau^{2/\tau}\Gamma(3\tau^{-1})}{\Gamma(\tau^{-1})} - [E(Z)]^2 \right\}$
skewness	$\begin{cases} \mu_{3Y}/[Var(Y)]^{1.5} \text{ where} \\ \mu_{3Y} = \sigma^3 \mu_{3Z} = \sigma^3 \{\mu_{3Z}^{'} - 3Var(Z)E(Z) - [E(Z)]^3\} \\ \text{where } \mu_{3Z}^{'} = E(Z^3) = \dfrac{\text{sign}(\nu)\tau^{3/\tau}\Gamma(4\tau^{-1})B(\tau^{-1}, 4\tau^{-1}, \nu^\tau/[1+\nu^\tau])}{\Gamma(\tau^{-1})B(\tau^{-1}, 4\tau^{-1})} \end{cases}$
excess kurtosis	$\begin{cases} \{\mu_{4Y}/[Var(Y)]^2\} - 3 \text{ where} \\ \mu_{4Y} = \sigma^4 \mu_{4Z} = \sigma^4 \{\mu_{4Z}^{'} - 4\mu_{3Z}^{'} E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\} \\ \text{where } \mu_{4Z}^{'} = E(Z^4) = \dfrac{\tau^{4/\tau}\Gamma(5\tau^{-1})}{\Gamma(\tau^{-1})} \end{cases}$
pdf	$\begin{cases} \dfrac{2}{\sigma} f_{Z_1}(z) F_{Z_1}(\nu z) \\ \text{where } z = (y-\mu)/\sigma \text{ and } Z_1 \sim \text{PE2}(0, \tau^{1/\tau}, \tau) \end{cases}$
cdf	–
Inverse cdf (y_p)	–
Reference	Azzalini [1986], page 202-203, with (λ, ω) replaced by (ν, τ) giving the pdf and moments of Z.

Table 14.22: Skew Exponential Power type 2 distribution

SEP2(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	$\begin{cases} \mu + \sigma E(Z) \\ \text{where } Z = (Y-\mu)/\sigma \text{ and } E(Z) \text{ is given by (14.28)} \end{cases}$
median	–
mode	–
Variance, $Var(Y)$	$\sigma^2 Var(Z) = \sigma^2 \left\{ \frac{\tau^{2/\tau}\Gamma(3\tau^{-1})}{\Gamma(\tau^{-1})} - [E(Z)]^2 \right\}$
skewness	$\begin{cases} \mu_{3Y}/[Var(Y)]^{1.5} \text{ where} \\ \mu_{3Y} = \sigma^3 \mu_{3Z} = \sigma^3 \{\mu_{3Z}^{'} - 3Var(Z)E(Z) - [E(Z)]^3\} \\ \text{where } \mu_{3Z}^{'} \text{ is given by (14.29)} \end{cases}$
excess kurtosis	$\begin{cases} \{\mu_{4Y}/[Var(Y)]^2\} - 3 \text{ where} \\ \mu_{4Y} = \sigma^4 \mu_{4Z} = \sigma^4 \{\mu_{4Z}^{'} - 4\mu_{3Z}^{'} E(Z) + 6Var(Z)[E(Z)]^2 \\ +3[E(Z)]^4\} \\ \text{where } \mu_{4Z}^{'} = E(Z^4) = \frac{\tau^{4/\tau}\Gamma(5\tau^{-1})}{\Gamma(\tau^{-1})} \end{cases}$
MGF	–
pdf	$\begin{cases} \frac{2}{\sigma} f_{Z_1}(z)\Phi(\omega) \\ \text{where } z = (y-\mu)/\sigma \text{ and } Z_1 \sim \text{PE2}(0, \tau^{1/\tau}, \tau) \text{ and} \\ \omega = \text{sign}(z)\lvert z\rvert^{\tau/2}\nu\sqrt{2/\tau} \end{cases}$
cdf	–
Inverse cdf (y_p)	–
Reference	DiCiccio and Monti [2004], page 439-440, with (λ, α) replaced by (ν, τ) giving the pdf and moments of Z.

14.4.9 Skew Exponential Power type 2 distribution, SEP2(μ, σ, ν, τ)

The pdf of the skew exponential power type 2 distribution, denoted by SEP2(μ, σ, ν, τ), is defined by

$$f_Y(y|\mu, \sigma\, \nu, \tau) = \frac{2}{\sigma} f_{Z_1}(z)\ \Phi(\omega) \tag{14.27}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$, and $\tau > 0$, and where $z = (y - \mu)/\sigma$ and $\omega = \text{sign}(z)|z|^{\tau/2}\nu\sqrt{2/\tau}$ and f_{Z_1} is the pdf of $Z_1 \sim$ PE2$(0, \tau^{1/\tau}, \tau)$ and $\Phi(\omega)$ is the cdf of a standard normal variable evaluated at ω.

This distribution was developed by DiCiccio and Monti [2004] (having been introduced by Azzalini [1986] as his type II distribution). The parameter ν determines the skewness of the distribution with $\nu > 0$ indicating positive skewness and $\nu < 0$ negative. The parameter τ determines the kurtosis of the distribution, with $\tau > 2$ for platykurtic data and $\tau < 2$ for leptokurtic.

Here $E(Y) = \mu + \sigma E(Z)$ where $Z = (Y - \mu)/\sigma$ and

$$E(Z) = \frac{2\tau^{1/\tau}\nu}{\sqrt{\pi}\Gamma\left(\tau^{-1}\right)\left(1+\nu^2\right)^{(2/\tau)+(1/2)}} \sum_{n=0}^{\infty} \frac{\Gamma\left(2\tau^{-1}+n+1/2\right)}{(2n+1)!!} \left(\frac{2\nu^2}{1+\nu^2}\right)^n \tag{14.28}$$

where $(2n+1)!! = 1.3.5 \ldots (2n+1)$, DiCiccio and Monti [2004], p439. Also $\mu_{3Z}^{'} = E(Z^3)$ is given by

$$E(Z^3) = \frac{2\tau^{3/\tau}\nu}{\sqrt{\pi}\Gamma\left(\tau^{-1}\right)\left(1+\nu^2\right)^{(4/\tau)+(1/2)}} \sum_{n=0}^{\infty} \frac{\Gamma\left(4\tau^{-1}+n+1/2\right)}{(2n+1)!!} \left(\frac{2\nu^2}{1+\nu^2}\right)^n. \tag{14.29}$$

For $\tau = 2$ the SEP2(μ, σ, ν, τ) distribution is the skew normal type 1 distribution, Azzalini (1985), denoted by SN1(μ, σ, ν), while for $\nu = 0$ and $\tau = 2$ the SEP2(μ, σ, ν, τ) distribution is the normal distribution, NO(μ, σ).

14.4.10 Skew Exponential Power type 3 distribution, SEP3(μ, σ, ν, τ)

This is a "spliced-scale" distribution with pdf, denoted by SEP3(μ, σ, ν, τ), defined by

$$f_Y(y|\mu, \sigma\, \nu, \tau) = \begin{cases} \dfrac{c}{\sigma} \exp\left[-\dfrac{1}{2}|\nu z|^{\tau}\right] & \text{if } y < \mu \\ \dfrac{c}{\sigma} \exp\left[-\dfrac{1}{2}\left|\dfrac{z}{\nu}\right|^{\tau}\right] & \text{if } y \geq \mu, \end{cases} \tag{14.30}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, and $\tau > 0$, and where $z = (y - \mu)/\sigma$ and $c = \nu\tau/\left[(1+\nu^2)2^{1/\tau}\Gamma\left(1/\tau\right)\right]^{-1}$, Fernandez et al. [1995]. Note that μ is the mode of Y.

Table 14.23: Skew Exponential Power type 3 distribution

SEP3(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	$\begin{cases} \mu + \sigma E(Z) = \mu + \dfrac{\sigma 2^{1/\tau}\Gamma(2\tau^{-1})(\nu - \nu^{-1})}{\Gamma(\tau^{-1})} \\ \text{where } Z = (Y-\mu)/\sigma \end{cases}$
median	–
mode	μ
Variance, $Var(Y)$	$\sigma^2 Var(Z) = \sigma^2 \left\{ \dfrac{2^{2/\tau}\Gamma(3\tau^{-1})(\nu^2 + \nu^{-2} - 1)}{\Gamma(\tau^{-1})} - [E(Z)]^2 \right\}$
skewness	$\begin{cases} \mu_{3Y}/[Var(Y)]^{1.5} \text{ where} \\ \mu_{3Y} = \sigma^3 \mu_{3Z} = \sigma^3 \{\mu_{3Z}^{'} - 3Var(Z)E(Z) - [E(Z)]^3\} \\ \text{where } \mu_{3Z}^{'} = E(Z^3) = \dfrac{2^{3/\tau}\Gamma(4\tau^{-1})(\nu^4 - \nu^{-4})}{\Gamma(\tau^{-1})(\nu + \nu^{-1})} \end{cases}$
excess kurtosis	$\begin{cases} \{\mu_{4Y}/[Var(Y)]^2\} - 3 \text{ where} \\ \mu_{4Y} = \sigma^4 \mu_{4Z} = \sigma^4 \left\{ \mu_{4Z}^{'} - 4\mu_{3Z}^{'} E(Z) + 6Var(Z)[E(Z)]^2 \right. \\ \left. +3[E(Z)]^4 \right\} \\ \text{where } \mu_{4Z}^{'} = E(Z^4) = \dfrac{2^{4/\tau}\Gamma(5\tau^{-1})(\nu^5 + \nu^{-5})}{\Gamma(\tau^{-1})(\nu + \nu^{-1})} \end{cases}$
pdf	$\begin{cases} \dfrac{c}{\sigma} \exp\left[-\dfrac{1}{2} \lvert \nu z \rvert^{\tau} \right] & \text{if } y < \mu \\ \dfrac{c}{\sigma} \exp\left[-\dfrac{1}{2} \left\lvert \dfrac{z}{\nu} \right\rvert^{\tau} \right] & \text{if } y \geq \mu, \\ \text{where } z = (y-\mu)/\sigma \text{ and } \quad c = \nu\tau \left[(1+\nu^2) 2^{1/\tau} \Gamma(1/\tau) \right]^{-1} \end{cases}$
cdf	$\begin{cases} \dfrac{1}{1+\nu^2} \left[\dfrac{\Gamma(\tau^{-1}, \alpha_1)}{\Gamma(\tau^{-1})} \right], & \text{if } y < \mu \\ 1 - \dfrac{\nu^2 \Gamma(\tau^{-1}, \alpha_2)}{(1+\nu^2)\Gamma(\tau^{-1})}, & \text{if } y \geq \mu \\ \text{where } \alpha_1 = \dfrac{\nu^2(\mu - y)^{\tau}}{2\sigma^{\tau}} \text{ and } \quad \alpha_2 = \dfrac{(y-\mu)^{\tau}}{2\sigma^{\tau}\nu^{\tau}} \end{cases}$
Inverse cdf (y_p)	–
Reference	Fernandez et al. [1995], p1333, equation (8) and (12), with (γ, q) replaced by (ν, τ) giving the pdf and moments of Z, respectively.

The skew normal type 2 (or two-piece normal) distribution, Johnson et al. [1994], p173, denoted by $\texttt{SN2}(\mu, \sigma, \nu)$, is a special case of $\texttt{SEP3}(\mu, \sigma, \nu, \tau)$ given by $\tau = 2$. The PE2 distribution is a reparameterized special case of SEP3 given by $\texttt{PE2}(\mu, \sigma, \nu) = \texttt{SEP3}(\mu, \sigma 2^{-1/\nu}, 1, \nu)$.

14.4.11 Skew Exponential Power type 4 distribution, $\texttt{SEP4}(\mu, \sigma, \nu, \tau)$

Table 14.24: Skew Exponential Power type 4 distribution

$\texttt{SEP4}(\mu, \sigma, \nu, \tau)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, left tail heaviness parameter
τ	$0 < \tau < \infty$, right tail heaviness parameter
Distribution measures	
mean, $E(Y)$	$\begin{cases} \mu + \sigma E(Z) = \mu + \sigma c \left[\tau^{-1}\Gamma\left(2\tau^{-1}\right) - \nu^{-1}\Gamma\left(2\nu^{-1}\right)\right] \\ \text{where } Z = (Y-\mu)/\sigma \text{ and } c = \left[\Gamma\left(1+\tau^{-1}\right) + \Gamma\left(1+\nu^{-1}\right)\right]^{-1} \end{cases}$
median	–
mode	μ
Variance, $Var(Y)$	$\sigma^2 Var(Z) = \sigma^2 \left\{ c\left[\tau^{-1}\Gamma\left(3\tau^{-1}\right) + \nu^{-1}\Gamma\left(3\nu^{-1}\right)\right] - [E(Z)]^2 \right\}$
skewness	$\begin{cases} \mu_{3Y}/[Var(Y)]^{1.5} \text{ where} \\ \mu_{3Y} = \sigma^3 \mu_{3Z} = \sigma^3 \{\mu_{3Z}^{'} - 3Var(Z)E(Z) - [E(Z)]^3\} \\ \text{where } \mu_{3Z}^{'} = E(Z^3) = c\left[\tau^{-1}\Gamma\left(4\tau^{-1}\right) - \nu^{-1}\Gamma\left(4\nu^{-1}\right)\right] \end{cases}$
excess kurtosis	$\begin{cases} \{\mu_{4Y}/[Var(Y)]^2\} - 3 \text{ where} \\ \mu_{4Y} = \sigma^4 \mu_{4Z} = \sigma^4 \{\mu_{4Z}^{'} - 4\mu_{3Z}^{'} E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\} \\ \text{where } \mu_{4Z}^{'} = E(Z^4) = c\left[\tau^{-1}\Gamma\left(5\tau^{-1}\right) + \nu^{-1}\Gamma\left(5\nu^{-1}\right)\right] \end{cases}$
pdf [a]	$\begin{cases} \frac{c}{\sigma}\exp\left[-\lvert z\rvert^{\nu}\right] & \text{if } y < \mu \\ \frac{c}{\sigma}\exp\left[-\lvert z\rvert^{\tau}\right] & \text{if } y \geq \mu, \\ \text{where } z = (y-\mu)/\sigma \text{ and} & c = \left[\Gamma\left(1+\tau^{-1}\right) + \Gamma\left(1+\nu^{-1}\right)\right]^{-1} \end{cases}$
cdf	$\begin{cases} \frac{c}{\nu}\Gamma(\nu^{-1}, \lvert z\rvert^{\nu}), & \text{if } y < \mu \\ 1 - \frac{c}{\tau}\Gamma(\tau^{-1}, \lvert z\rvert^{\tau}), & \text{if } y \geq \mu \end{cases}$
Reference	[a] Jones [2005]

This is a "spliced-shap" distribution with pdf, denoted by $\texttt{SEP4}(\mu, \sigma, \nu, \tau)$, defined by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \begin{cases} \frac{c}{\sigma}\exp\left[-|z|^{\nu}\right], & \text{if } y < \mu \\ \frac{c}{\sigma}\exp\left[-|z|^{\tau}\right], & \text{if } y \geq \mu \end{cases} \tag{14.31}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, and $\tau > 0$, and where $z = (y - \mu)/\sigma$ and $c = \left[\Gamma\left(1 + \tau^{-1}\right) + \Gamma\left(1 + \nu^{-1}\right)\right]^{-1}$, Jones [2005]. Note that μ is the mode of Y.

14.4.12 Skew Student t distribution, $\texttt{SST}(\mu, \sigma, \nu, \tau)$

Wurtz et al. [2006] reparameterized the ST3 distribution [Fernandez and Steel, 1998] so that in the new parameterization μ is the mean and σ is the standard deviation. They called this the skew student t distribution (denoted SST here).

Let $Z_0 \sim \texttt{ST3}(0, 1, \nu, \tau)$ and $Y = \mu + \sigma\left(\frac{Z_0 - m}{s}\right)$ where

$$m = E(Z_0) = \frac{2\tau^{1/2}(\nu - \nu^{-1})}{(\tau - 1)B\left(1/2, \tau/2\right)} \tag{14.32}$$

for $\tau > 1$, and

$$s^2 = V(Z_0) = E(Z_0^2) - m^2 = \frac{\tau}{(\tau - 2)}(\nu^2 + \nu^{-2} - 1) - m^2 \tag{14.33}$$

for $\tau > 2$.

Hence $Y = \mu_0 + \sigma_0 Z_0$, where $\mu_0 = \mu - \sigma m/s$ and $\sigma_0 = \sigma/s$, and so $Y \sim \texttt{ST3}(\mu_0, \sigma_0, \nu, \tau)$ with $E(Y) = \mu$ and $Var(Y) = \sigma^2$ for $\tau > 2$. Let $Y \sim \texttt{SST}(\mu, \sigma, \nu, \tau) = \texttt{ST3}(\mu_0, \sigma_0, \nu, \tau)$, for $\tau > 2$.

Hence the pdf of the skew student t distribution,denoted by $Y \sim \texttt{SST}(\mu, \sigma, \nu, \tau)$, is defined by

$$f_Y(y|, \mu, \sigma, \nu, \tau) = \begin{cases} \dfrac{c}{\sigma_0}\left[1 + \dfrac{\nu^2 z^2}{\tau}\right]^{-(\tau+1)/2} & \text{if } y < \mu_0 \\ \dfrac{c}{\sigma_0}\left[1 + \dfrac{z^2}{\nu^2 \tau}\right]^{-(\tau+1)/2} & \text{if } y \geq \mu_0, \end{cases}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0, \nu > 0$ and $\tau > 2$ and where $\mu_0 = \mu - \sigma m/s$ and $\sigma_0 = \sigma/s$ and $z = (y - \mu_0)/\sigma_0$, $c = 2\nu\left[(1 + \nu^2)B\left(1/2, \tau/2\right)\tau^{1/2}\right]^{-1}$ and m and s are obtained from equations (14.32) and (14.33), respectively.

Note that $E(Y) = \mu$ and $Var(Y) = \sigma^2$ and the moment based skewness and excess kurtosis of $Y \sim \texttt{SST}(\mu, \sigma, \nu, \tau)$ are the same as for $\texttt{ST3}(\mu_0, \sigma_0, \nu, \tau)$, and hence the same as for $\texttt{ST3}(0, 1, \nu, \tau)$, depending only on ν and τ.

Note also that in **gamlss.dist** the default link function for τ is a shifted log link function, $\log(\tau - 2)$, which ensures that τ is always in its valid range, i.e. $\tau > 2$.

Table 14.25: Skew Student t distribution

$\texttt{SST}(\mu, \sigma, \nu, \tau)$	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mean, location shift parameter
σ	$0 < \sigma < \infty$, standard deviation, scaling parameter
ν	$0 < \nu < \infty$, skewness parameter
τ	$2 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	μ
median	$\begin{cases} \mu_0 + \frac{\sigma_0}{\nu} t_{\frac{(1+\nu^2)}{4},\tau} & \text{if } \nu \leq 1 \\ \mu_0 + \frac{\sigma_0}{\nu} t_{\frac{(3\nu^2-1)}{4\nu^2},\tau} & \text{if } \nu > 1 \end{cases}$
mode	μ_0
Variance, $Var(Y)$	σ^2
skewness	Equal to skewness of $\texttt{ST3}(0, 1, \nu, \tau)$
excess kurtosis	Equal to excess kurtosis of $\texttt{ST3}(0, 1, \nu, \tau)$
pdf	$\begin{cases} \frac{c}{\sigma_0}\left[1 + \frac{\nu^2 z^2}{\tau}\right]^{-(\tau+1)/2} & \text{if } y < \mu_0 \\ \frac{c}{\sigma_0}\left[1 + \frac{z^2}{\nu^2 \tau}\right]^{-(\tau+1)/2} & \text{if } y \geq \mu_0, \\ \text{where } z = (y-\mu_0)/\sigma_0 \text{ and } & c = 2\nu\left[(1+\nu^2)B\left(1/2, \tau/2\right)\tau^{1/2}\right]^{-1} \\ \text{and } \mu_0 \text{ and } \sigma_0 \text{ are defined in} & \text{Section 1.3.12} \end{cases}$
cdf	$\begin{cases} \frac{2}{(1+\nu^2)} F_T[\nu(y-\mu_0)/(\sigma_0)], & \text{if } y < \mu_0 \\ \frac{1}{(1+\nu^2)}\left[1 + 2\nu^2\left\{F_T[(y-\mu_0)/(\sigma_0\nu)] - \frac{1}{2}\right\}\right], & \text{if } y \geq \mu_0 \\ \text{where } T \sim t_\tau \end{cases}$
Inverse cdf ($y_p = F_Y^{-1}(p)$)	$\begin{cases} \mu_0 + \frac{\sigma_0}{\nu} t_{\left[\frac{p(1+\nu^2)}{2}\right],\tau} & \text{if } p \leq (1+\nu^2)^{-1} \\ \mu_0 + \sigma_0 \nu t_{\left[\frac{p(1+\nu^2)-1+\nu^2}{2\nu^2}\right],\tau} & \text{if } p > (1+\nu^2)^{-1} \\ \text{where } t_{\alpha,\tau} = F_T^{-1}(\alpha) \text{ and } T \sim t_\tau \end{cases}$

14.4.13 Skew t type 1 distribution, ST1(μ, σ, ν, τ)

The pdf of the skew t type 1 distribution, denoted by ST1(μ, σ, ν, τ), is defined by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \begin{cases} \frac{c}{\sigma_0}\left[1+\frac{\nu^2 z^2}{\tau}\right]^{-(\tau+1)/2} & \text{if } y < \mu_0 \\ \frac{c}{\sigma_0}\left[1+\frac{z^2}{\nu^2\tau}\right]^{-(\tau+1)/2} & \text{if } y \geq \mu_0, \end{cases} \tag{14.34}$$

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 $Z \sim$ TF$(0, 1, \tau) = t_\tau$, a t distribution with $\tau > 0$ degrees of freedom, with τ treated as a continuous parameter. This distribution is in the form of a type I distribution of Azzalini (1986). [No summary table is given for ST1.]

14.4.14 Skew t type 2 distribution, ST2(μ, σ, ν, τ)

The pdf of the skew t type 2 distribution, denoted by ST2(μ, σ, ν, τ), is defined by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{2}{\sigma} f_{Z_1}(z) F_{Z_2}(\omega) \tag{14.35}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $-\infty < \nu < \infty$, and $\tau > 0$, and where $z = (y-\mu)/\sigma$, $\omega = \nu\lambda^{1/2} z$ and $\lambda = (\tau+1)/(\tau+z^2)$ and f_{Z_1} is the pdf of $Z_1 \sim$ TF$(0, 1, \tau) = t_\tau$ and F_{Z_2} is the cdf of $Z_2 \sim$ TF$(0, 1, \tau+1) = t_{\tau+1}$. This distribution is the univariate case of the multivariate skew t distribution introduced by Azzalini and Capitanio [2003], p380, equation (26).

14.4.15 Skew t type 3 distribution, ST3(μ, σ, ν, τ)

This is a "spliced-scale" distribution, denoted by ST3(μ, σ, ν, τ), with pdf defined by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \begin{cases} \frac{c}{\sigma}\left[1+\frac{\nu^2 z^2}{\tau}\right]^{-(\tau+1)/2} & \text{if } y < \mu \\ \frac{c}{\sigma}\left[1+\frac{z^2}{\nu^2\tau}\right]^{-(\tau+1)/2} & \text{if } y \geq \mu \end{cases} \tag{14.36}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$, and $\tau > 0$, and where $z = (y-\mu)/\sigma$ and $c = 2\nu[(1+\nu^2)B(1/2, \tau/2)\tau^{1/2}]^{-1}$, Fernandez and Steel [1998], p362, equation (13), with (γ, ν) replaced by (ν, τ).

Note that μ is the mode of Y.

The moments of Y in Table 14.27 are obtained using Fernandez and Steel [1998], p360, equation (5).

Table 14.26: Skew Student t type 2 distribution

ST2(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	$\mu + \sigma E(Z)$ where $Z = (Y-\mu)/\sigma$ and $E(Z) = \frac{\nu \tau^{1/2} \Gamma([\tau-1]/2)}{(1+\nu^2)^{1/2} \pi^{1/2} \Gamma(\tau/2)}$, for $\tau > 1$,
median	–
mode	–
Variance, $Var(Y)$	$\sigma^2 Var(Z) = \sigma^2 \left\{ \left(\frac{\tau}{\tau-2} \right) - [E(Z)]^2 \right\}$, for $\tau > 2$
skewness	$\mu_{3Y}/[Var(Y)]^{1.5}$ where $\mu_{3Y} = \sigma^3 \mu_{3Z} = \sigma^3 \{\mu'_{3Z} - 3Var(Z)E(Z) - [E(Z)]^3\}$ where $\mu'_{3Z} = E(Z^3) = \frac{\tau(3-\delta^2)}{(\tau-3)} E(Z)$ for $\tau > 3$ and $\delta = \nu(1+\nu^2)^{-1/2}$,
excess kurtosis	$\{\mu_{4Y}/[Var(Y)]^2\} - 3$ where $\mu_{4Y} = \sigma^4 \mu_{4Z} = \sigma^4 \{\mu'_{4Z} - 4\mu'_{3Z} E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\}$ where $\mu'_{4Z} = E(Z^4) = \frac{3\tau^2}{(\tau-2)(\tau-4)}$ for $\tau > 4$
pdf	$\frac{2}{\sigma} f_{Z_1}(z) F_{Z_2}(\omega)$ where $z = (y-\mu)/\sigma, \omega = \nu \lambda^{1/2} z, \lambda = (\tau+1)/(\tau+z^2)$ and $Z_1 \sim t_\tau$ and $Z_2 \sim t_{\tau+1}$
cdf	–
Inverse cdf	–
Reference	Azzalini and Capitanio [2003], p380, equation(26) and p 382 with dimension $d = 1$ and $(\xi, \omega, \alpha, \nu)$ and Ω replaced by (μ, σ, ν, τ) and σ^2 respectively, giving the pdf of Y and moments of Z.

Table 14.27: Skew Student t type 3 distribution

ST3(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	$\mu + \sigma E(Z) = \mu + \frac{2\sigma\tau^{1/2}(\nu - \nu^{-1})}{(\tau-1)B(1/2, \tau/2)}$, for $\tau > 1$ where $Z = (Y - \mu)/\sigma$
median	$\mu + \frac{\sigma}{\nu} t_{\frac{(1+\nu^2)}{4}, \tau}$ if $\nu \leq 1$ $\mu + \sigma\nu t_{\frac{(3\nu^2-1)}{4\nu^2}, \tau}$ if $\nu > 1$
mode	–
Variance, $Var(Y)$	$\sigma^2 Var(Z) = \sigma^2 \left\{ \left(\frac{\tau}{\tau-2} \right) (\nu^2 + \nu^{-2} - 1) - [E(Z)]^2 \right\}$ for $\tau > 2$
skewness	$\mu_{3Y}/[Var(Y)]^{1.5}$ where $\mu_{3Y} = \sigma^3 \mu_{3Z} = \sigma^3 \{\mu'_{3Z} - 3Var(Z)E(Z) - [E(Z)]^3\}$ where $\mu'_{3Z} = E(Z^3) = \frac{4\tau^{3/2}\left(\nu^4 - \nu^{-4}\right)}{(\tau-1)(\tau-3)B(1/2, \tau/2)(\nu + \nu^{-1})}$, for $\tau > 3$
excess kurtosis	$\{\mu_{4Y}/[Var(Y)]^2\} - 3$ where $\mu_{4Y} = \sigma^4 \mu_{4Z} = \sigma^4 \{\mu'_{4Z} - 4\mu'_{3Z} E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\}$ where $\mu'_{4Z} = E(Z^4) = \frac{3\tau^2\left(\nu^5 + \nu^{-5}\right)}{(\tau-2)(\tau-4)(\nu + \nu^{-1})}$, for $\tau > 4$
pdf [a]	$\frac{c}{\sigma}\left[1 + \frac{\nu^2 z^2}{\tau}\right]^{-(\tau+1)/2}$ if $y < \mu$ $\frac{c}{\sigma}\left[1 + \frac{z^2}{\nu^2 \tau}\right]^{-(\tau+1)/2}$ if $y \geq \mu$ where $z = (y - \mu)/\sigma$ and $c = 2\nu[(1 + \nu^2)B(1/2, \tau/2)\tau^{1/2}]^{-1}$
cdf	$\frac{2}{(1+\nu^2)} F_T[\nu(y - \mu)/\sigma]$ if $y < \mu$ $\frac{1}{(1+\nu^2)}\left[1 + 2\nu^2\left\{F_T[(y - \mu)/(\sigma\nu)] - \frac{1}{2}\right\}\right]$ if $y \geq \mu$
Inverse cdf ($y_p = F_Y^{-1}(p)$)	$\mu + \frac{\sigma}{\nu} t_{\left[\frac{p(1+\nu^2)}{2}\right], \tau}$ if $p \leq (1 + \nu^2)^{-1}$ $\mu + \sigma\nu t_{\left[\frac{p(1+\nu^2)-1+\nu^2}{2\nu^2}\right], \tau}$ if $p > (1 + \nu^2)^{-1}$ where $t_{\alpha,\tau} = F_T^{-1}(\alpha)$ and $T \sim t_\tau$
Reference	[a] From Fernandez and Steel [1998], p362, equation (13), with (γ, ν) replaced by (ν, τ).

14.4.16 Skew t type 4 distribution, $\texttt{ST4}(\mu, \sigma, \nu, \tau)$

This is a "spliced-shape" distribution, denoted by $\texttt{ST4}(\mu, \sigma, \nu, \tau)$, with pdf defined by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \begin{cases} \dfrac{c}{\sigma}\left[1+\dfrac{z^2}{\nu}\right]^{-(\nu+1)/2} & \text{if } y < \mu \\ \dfrac{c}{\sigma}\left[1+\dfrac{z^2}{\tau}\right]^{-(\tau+1)/2} & \text{if } y \geq \mu \end{cases} \tag{14.37}$$

for $-\infty < y < \infty$, where $-\infty < \mu < \infty$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$, and where $z = (y-\mu)/\sigma$ and $c = 2\left[\nu^{1/2}B\left(1/2, \nu/2\right) + \tau^{1/2}B\left(1/2, \tau/2\right)\right]^{-1}$.

14.4.17 Skew t type 5 distribution, $\texttt{ST5}(\mu, \sigma, \nu, \tau)$

The pdf of the skew t distribution type 5, denoted by $\texttt{ST5}(\mu, \sigma, \nu, \tau)$, 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 $z = (y-\mu)/\sigma$ and $c = \left[2^{a+b-1}(a+b)^{1/2}B(a,b)\right]^{-1}$ and $\nu = (a-b)/\left[ab(a+b)\right]^{1/2}$ and $\tau = 2/(a+b)$. Hence $a = \tau^{-1}[1+\nu(2\tau+\nu^2)^{-1/2}]$ and $b = \tau^{-1}[1-\nu(2\tau+\nu^2)^{-1/2}]$.

From Jones and Faddy [2003] p160, equation (2), if $B \sim BEo(a,b)$ then

$$Z = \frac{(a+b)^{1/2}(2B-1)}{2[B(1-B)]^{1/2}} \sim \texttt{ST5}(0,1,\nu,\tau)$$

Hence $B = \dfrac{1}{2}\left[1+Z(a+b+Z^2)^{-1/2}\right] \sim BEo(a,b)$, from which the cdf of Y is obtained.

Table 14.28: Skew Student t type 4 distribution

ST4(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, mode, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$0 < \nu < \infty$, left tail heaviness parameter
τ	$0 < \tau < \infty$, right tail heaviness parameter
Distribution measures	
mean $E(Y)$	$\mu + \sigma E(Z) = \mu + \sigma c \left[\frac{\tau}{(\tau-1)} - \frac{\nu}{(\nu-1)}\right]$ for $\nu > 1$ and $\tau > 1$ where $Z = (Y-\mu)/\sigma$ and $c = 2[\nu^{1/2}B(1/2, \nu/2) + \tau^{1/2}B(1/2, \tau/2)]^{-1}$
median	$\mu + \sigma t_{\frac{(1+k)}{4}, \nu}$ if $k \leq 1$; $\mu + \sigma t_{\frac{(3k-1)}{4k}, \tau}$ if $k > 1$; where $k = \frac{\tau^{1/2}B(1/2, \tau/2)}{\nu^{1/2}B(1/2, \nu/2)}$
mode	μ
Variance $Var(Y)$	$\sigma^2 Var(Z) = \sigma^2\{E(Z^2) - [E(Z)]^2\}$ where $E(Z^2) = \frac{c\tau^{3/2}B(1/2, \tau/2)}{2(\tau-2)} + \frac{c\nu^{3/2}B(1/2, \nu/2)}{2(\nu-2)}$ for $\nu > 2$ and $\tau > 2$
skewness	$\mu_{3Y}/[Var(Y)]^{1.5}$ where $\mu_{3Y} = \sigma^3\mu_{3Z} = \sigma^3\{\mu'_{3Z} - 3Var(Z)E(Z) - [E(Z)]^3\}$ where $\mu'_{3Z} = E(Z^3) = 2c\left[\frac{\tau^2}{(\tau-1)(\tau-3)} - \frac{\nu^2}{(\nu-1)(\nu-3)}\right]$ for $\nu > 3$ and $\tau > 3$
excess kurtosis	$\{\mu_{4Y}/[Var(Y)]^2\} - 3$ where $\mu_{4Y} = \sigma^4\mu_{4Z} = \sigma^4\{\mu'_{4Z} - 4\mu'_{3Z}E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\}$ where $\mu'_{4Z} = E(Z^4) = 3 + 3c\left[\frac{\tau^{1/2}B(1/2, \tau/2)}{(\tau-4)} + \frac{\nu^{1/2}B(1/2, \nu/2)}{(\nu-4)}\right]$ for $\nu > 4$ and $\tau > 4$
pdf [a]	$\frac{c}{\sigma}\left[1 + \frac{z^2}{\nu}\right]^{-(\nu+1)/2}$ if $y < \mu$; $\frac{c}{\sigma}\left[1 + \frac{z^2}{\tau}\right]^{-(\tau+1)/2}$ if $y \geq \mu$; where $z = (y-\mu)/\sigma$
cdf	$\frac{2}{(1+k)}F_{T_1}(z)$, if $y < \mu$; $\frac{1}{(1+k)}\left\{1 + 2k\left[F_{T_2}(z) - \frac{1}{2}\right]\right\}$, if $y \geq \mu$; where $T_1 \sim t_\nu$ and $T_2 \sim t_\tau$ and $z = (y-\mu)/\sigma$
Inverse cdf $(y_p = F_Y^{-1}(p))$	$\mu + \sigma t_{\left[\frac{p(1+k)}{2}\right], \nu}$ if $p \leq (1+k)^{-1}$; $\mu + \sigma t_{\left[\frac{p(1+k)-1}{2k} + \frac{1}{2}\right], \tau}$ if $p > (1+k)^{-1}$; where $t_{\alpha,\tau} = F_T^{-1}(\alpha)$ and $T \sim t_\tau$

Table 14.29: Skew Student t type 5 distribution

ST5(μ, σ, ν, τ)	
Ranges	
Y	$-\infty < y < \infty$
μ	$-\infty < \mu < \infty$, location shift parameter
σ	$0 < \sigma < \infty$, scaling parameter
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$, kurtosis parameter
Distribution measures	
mean, $E(Y)$	$\left\{ \begin{array}{l} \mu + \sigma E(Z) = \mu + \dfrac{\sigma(a+b)^{1/2}(a-b)\Gamma(a-1/2)\Gamma(b-1/2)}{2\Gamma(a)\Gamma(b)} \\ \text{for } a > 1/2 \text{ and } b > 1/2 \text{ where } Z = (Y-\mu)/\sigma \end{array} \right.$
median	–
mode	$\mu + \dfrac{\sigma(a+b)^{1/2}(a-b)}{(2a+1)^{1/2}(2b+1)^{1/2}}$
Variance, $Var(Y)$	$\left\{ \begin{array}{l} \sigma^2 Var(Z) = \sigma^2 \left\{ \dfrac{(a+b)[(a-b)^2+a+b-2]}{4(a-1)(b-1)} - [E(Z)]^2 \right\} \\ \text{for } a > 1 \text{ and } b > 1 \end{array} \right.$
skewness	$\left\{ \begin{array}{l} \mu_{3Y}/[Var(Y)]^{1.5} \text{ where} \\ \mu_{3Y} = \sigma^3 \mu_{3Z} = \sigma^3 \{\mu'_{3Z} - 3Var(Z)E(Z) - [E(Z)]^3\} \\ \text{where } \mu'_{3Z} = E(Z^3) = \dfrac{(a+b)^{3/2}\Gamma(a-3/2)\Gamma(b-3/2)}{4\Gamma(a)\Gamma(b)} \left[2a^3+6a^2-23a \right. \\ \left. +2b^3+6b^2-23b-6a^2b-6ab^2+24ab-3\right] \\ \text{for } a > 3/2 \text{ and } b > 3/2 \end{array} \right.$
excess kurtosis	$\left\{ \begin{array}{l} \{\mu_{4Y}/[Var(Y)]^2\} - 3 \text{ where} \\ \mu_{4Y} = \sigma^4 \mu_{4Z} = \sigma^4 \{\mu'_{4Z} - 4\mu'_{3Z}E(Z) + 6Var(Z)[E(Z)]^2 + 3[E(Z)]^4\} \\ \text{where } \mu'_{4Z} = E(Z^4) = \dfrac{(a+b)^{3/2}}{16(a-1)(a-2)(b-1)(b-2)} \left[a^4-2a^3-a^2 \right. \\ \left. +2a+b^4-2b^3-b^2+2b+2(a-2)(b-2)(3a-2a^2-2b^2-a-b-3)\right] \\ \text{for } a > 2 \text{ and } b > 2 \end{array} \right.$
pdf	$\left\{ \begin{array}{l} f_Y(y\|\mu,\sigma,\nu,\tau) = \dfrac{c}{\sigma}\left[1+\frac{z}{(a+b+z^2)^{1/2}}\right]^{a+1/2}\left[1-\frac{z}{(a+b+z^2)^{1/2}}\right]^{b+1/2} \\ \text{where } z = (y-\mu)/\sigma, c = \left[2^{a+b-1}(a+b)^{1/2}B(a,b)\right]^{-1} \\ a = \tau^{-1}[1+d] \text{ and } b = \tau^{-1}[1-d] \text{ where } d = \nu(2\tau+\nu^2)^{-1/2} \end{array} \right.$
cdf	$\left\{ \begin{array}{l} \dfrac{B(a,b,r)}{B(a,b)} \\ \text{where } r = \dfrac{1}{2}[1+z(a+b+z^2)^{-1/2} \text{ and } z = (y-\mu)/\sigma \end{array} \right.$
Reference	Jones and Faddy [2003], p159, equation (1) reparameterized as on p164 with (q,p) replaced by (ν,τ), and p162 equations (4b) and (5) giving the pdf, moments and mode of Z respectively.

Chapter 15

Continuous distributions on $(0, \infty)$

15.1 Scale family of distributions with scaling parameter θ

A continuous random variable Y, defined on $(0, \infty)$, is said to have a scale family of distributions with scaling parameter θ (for fixed values of all other parameters of the distribution) if

$$Z = \frac{Y}{\theta},$$

has a cumulative distribution function (cdf) which does not depend on θ. Hence

$$F_Y(y) = F_Z\left(\frac{y}{\theta}\right)$$

and

$$f_Y(y) = \frac{1}{\theta} f_Z\left(\frac{y}{\theta}\right),$$

so $F_Y(y)$ and $\theta f_Y(y)$ only depend on y and θ through the function $z = y/\theta$. Note $Y = \theta Z$.

Example: Let Y have a Weibull distribution, $Y \sim \texttt{WEI}(\mu, \sigma)$, then Y has a scale family of distributions with scaling parameter μ (for a fixed value of σ), since $F_Y(y) = 1 - \exp\left[-\left(\frac{y}{\mu}\right)^{\sigma}\right]$ and hence $F_Z(z) = 1 - \exp\left[-(z)^{\sigma}\right]$ does not depend on μ. Note $Z \sim \texttt{WEI}(1, \sigma)$.

All distributions in gamlss.dist with range $(0, \infty)$ are scale families of distributions with scaling parameter μ, except for $\texttt{IG}(\mu, \sigma)$, $\texttt{LOGNO}(\mu, \sigma)$, $\texttt{WEI2}(\mu, \sigma)$ and $\texttt{LNO}(\mu, \sigma, \nu)$.

15.2 Continuous one parameter distributions on $(0, \infty)$

15.2.1 Exponential distribution, EXP(μ)

Table 15.1: Exponential distribution

EXP(μ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, mean, scaling parameter
Distribution measures	
mean	μ
median	$\mu \log 2$
mode	$\to 0$
variance	μ^2
skewness	2
excess kurtosis	6
MGF	$(1 - \mu t)^{-1}$ for $t < 1/\mu$
pdf	$(1/\mu) \exp(-y/\mu)$
cdf	$1 - \exp(-y/\mu)$
Inverse cdf (y_p)	$-\mu \log(1 - p)$
Reference	Set $\sigma = 1$ in GA(μ, σ), or Johnson et al. [1994] Chapter 19, p494-499.

This is the only one parameter continuous distribution in the **gamlss** packages. The exponential distribution is appropriate for moderately positive skew data. The probability density function of the exponential distribution, denoted by EXP(μ), is defined by

$$f_Y(y|\mu) = \frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right) \tag{15.1}$$

for $y > 0$, where $\mu > 0$. Hence $E(Y) = \mu$ and $Var(Y) = \mu^2$.

Table 15.2: Gamma distribution

$\texttt{GA}(\mu, \sigma)$	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, mean, scaling parameter
σ	$0 < \sigma < \infty$, coefficient of variation
Distribution measures	
mean [a]	μ
median	–
mode	$\begin{cases} \mu(1-\sigma^2) & \text{if } \sigma^2 < 1 \\ \to 0 & \text{if } \sigma^2 \geq 1 \end{cases}$
variance [a]	$\sigma^2\mu^2$
skewness [a]	2σ
excess kurtosis [a]	$6\sigma^2$
MGF	$(1-\mu\sigma^2 t)^{-1/\sigma^2}$ for $t < (\mu\sigma^2)^{-1}$
pdf [a,a2]	$\frac{y^{1/\sigma^2-1}e^{-y/(\sigma^2\mu)}}{(\sigma^2\mu)^{1/\sigma^2}\Gamma(1/\sigma^2)}$
cdf	$\frac{\gamma(\sigma^{-2}, y\mu^{-1}\sigma^{-2})}{\Gamma(\sigma^{-2})}$
Inverse cdf (y_p)	–
Reference	[a] Derived from McCullagh and Nelder [1989] p287 reparameterized by $\mu = \mu$ and $\nu = 1/\sigma^2$ [a2] Johnson et al. [1994] p343, Equation (17.23) reparameterized by $\alpha = 1/\sigma^2$ and $\beta = \mu\sigma^2$.
Note	$\gamma(a, x) = \int_0^x t^{a-1}e^{-t}dt$, the incomplete gamma function.

15.3 Continuous two parameter distributions on $(0, \infty)$

15.3.1 Gamma distribution, GA(μ, σ)

The gamma distribution is appropriate for positively skew data. The pdf of the gamma distribution, denoted by GA(μ, σ), is defined by

$$f_Y(y|\mu, \sigma) = \frac{y^{1/\sigma^2 - 1} e^{-y/(\sigma^2 \mu)}}{(\sigma^2 \mu)^{1/\sigma^2} \Gamma(1/\sigma^2)} \tag{15.2}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Here $E(Y) = \mu$ and $Var(Y) = \sigma^2 \mu^2$ and $E(Y^r) = \mu^r \sigma^{2r} \Gamma\left(\frac{1}{\sigma^2} + r\right) / \Gamma\left(\frac{1}{\sigma^2}\right)$ for $r > -1/\sigma^2$.

This a reparameterization of Johnson et al. [1994] p343, equation (17.23), obtained by setting $\alpha = 1/\sigma^2$ and $\beta = \mu\sigma^2$. Hence $\mu = \alpha\beta$ and $\sigma^2 = 1/\alpha$.

15.3.2 Inverse gamma distribution, IGAMMA(μ, σ)

The pdf of the inverse gamma distribution, denoted by IGAMMA(μ, σ), is defined by

$$f_Y(y|\mu, \sigma) = \frac{\mu^\alpha (\alpha + 1)^\alpha y^{-(\alpha+1)}}{\Gamma(\alpha)} \exp\left[-\frac{\mu(\alpha + 1)}{y}\right]$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$ and where $\alpha = 1/\sigma^2$. The inverse gamma (IGAMMA) distribution is a reparameterized special case of the generalized gamma (GG) distribution given by IGAMMA(μ, σ) = GG$([1 + \sigma^2]\mu, \sigma, -1)$.

15.3.3 Inverse Gaussian distribution, IG(μ, σ)

The inverse Gaussian distribution is appropriate for highly positive skew data. The pdf of the inverse Gaussian distribution, denoted by IG(μ, σ) is defined by

$$f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left[-\frac{1}{2\mu^2\sigma^2 y} (y - \mu)^2\right] \tag{15.3}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Hence $E(Y) = \mu$ and $Var(Y) = \sigma^2 \mu^3$. This is a re-parametrization of Johnson *et al.* (1994) p 261 equation (15.4a), obtained by setting $\sigma^2 = 1/\lambda$. Note that the inverse Gaussian distribution is a re-parameterized special case of the generalized inverse Gaussian distribution, given by IG(μ, σ) = GIG$(\mu, \sigma\mu^{1/2}, -0.5)$.

Note also that if $Y \sim$ IG(μ, σ) then $Y_1 = aY \sim$ IG$(a\mu, a^{-1/2}\sigma)$. Hence IG(μ, σ) is a scale family of distributions, but neither μ nor σ are scaling parameters. The

Table 15.3: Gamma distribution

IGAMMA(μ, σ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, mode, scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean [a]	$\begin{cases} \frac{(1+\sigma^2)\mu}{(1-\sigma^2)}, & \text{if } \sigma^2 < 1 \\ \infty, & \text{if } \sigma^2 \geq 1 \end{cases}$
median	–
mode [a]	μ
variance [a]	$\begin{cases} \frac{(1+\sigma^2)^2\mu^2\sigma^2}{(1-\sigma^2)^2(1-2\sigma^2)}, & \text{if } \sigma^2 < 1/2 \\ \infty, & \text{if } \sigma^2 \geq 1/2 \end{cases}$
skewness [a]	$\begin{cases} \frac{4\sigma(1-2\sigma^2)^{1/2}}{(1-3\sigma^2)}, & \text{if } \sigma^2 < 1/3 \\ \infty, & \text{if } \sigma^2 \geq 1/3 \end{cases}$
excess kurtosis [a]	$\begin{cases} \frac{3\sigma^2(10-22\sigma^2)}{(1-3\sigma^2)(1-4\sigma^2)}, & \text{if } \sigma^2 > 1/4 \\ \infty, & \text{if } \sigma^2 \leq 1/4 \end{cases}$
MGF	–
pdf [a]	$\frac{\mu^\alpha(\alpha+1)^\alpha y^{-(\alpha+1)}}{\Gamma(\alpha)} \exp\left[-\frac{\mu(\alpha+1)}{y}\right]$, where $\alpha = 1/\sigma^2$
cdf [a]	$\frac{\Gamma\left(\alpha, \frac{\mu[\alpha+1]}{y}\right)}{\Gamma(\alpha)}$
Inverse cdf (y_p)	–
Reference	[a] Set $\mu = (1+\sigma^2)\mu$, $\sigma = \sigma$ and $\nu = -1$ (so $\theta = 1/\sigma^2$) in GG(μ, σ, ν).
Note	$\Gamma(a, x) = \int_x^\infty t^{a-1}e^{-t}dt$, the complement of the incomplete gamma function.

Table 15.4: Inverse Gaussian distribution

IG(μ, σ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, mean, NOT a scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean [a]	μ
median	–
mode [a]	$\frac{-3\mu^2\sigma^2 + \mu(9\mu^2\sigma^4 + 4)^{1/2}}{2}$
variance [a]	$\sigma^2\mu^3$
skewness [a]	$3\mu^{1/2}\sigma$
excess kurtosis[a]	$15\mu\sigma^2$
MGF [a]	$\exp\left\{\frac{1}{\mu\sigma^2}\left[1 - \left(1 - 2\mu^2\sigma^2 t\right)^{1/2}\right)\right]\right\}$ for $t < (2\mu^2\sigma^2)^{-1}$
pdf [a]	$(2\pi\sigma^2 y^3)^{-1/2} \exp\left[-\frac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right]$
cdf [a]	$\Phi\left[(\sigma^2 y)^{-1/2}\left(\frac{y}{\mu} - 1\right)\right] + e^{2(\mu\sigma^2)^{-1}}\Phi\left[-(\sigma^2 y)^{-1/2}\left(\frac{y}{\mu} + 1\right)\right]$
Inverse cdf (y_p)	–
Reference	[a] Johnson et al. [1994] Chapter 15, p261-263 and p268 with Equation (15.4a) reparametrized by $\mu = \mu$ and $\lambda = 1/\sigma^2$, or set $\mu = \mu, \sigma = \sigma\mu^{1/2}$ and $\nu = -1/2$ in GIG(μ, σ, ν).

shape of the $\texttt{IG}(\mu, \sigma)$ distribution depends only on the value of $\sigma^2\mu$. [So if $\sigma^2\mu$ is fixed, then changing μ changes the scale but not the shape of the $\texttt{IG}(\mu, \sigma)$ distribution. If σ in the distribution $\texttt{IG}(\mu, \sigma)$ is reparameterized by setting $\alpha = \sigma\mu^{1/2}$, then for fixed α the resulting $IG2(\mu, \alpha)$ distribution for Y is a scale family of distributions with scaling parameter μ, since $Z = Y/\mu \sim IG2(1, \alpha)$. Note $IG2$ is not currently available in **gamlss**.]

15.3.4 Log normal distribution, $\texttt{LOGNO}(\mu, \sigma)$

Table 15.5: Log normal distribution

$\texttt{LOGNO}(\mu, \sigma)$	
Ranges	
Y	$0 < y < \infty$
μ	$-\infty < \mu < \infty$, NOT a scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean [a]	$e^{\mu+\sigma^2/2}$
median [a]	e^{μ}
mode [a]	$e^{\mu-\sigma^2}$
variance [a]	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
skewness [a]	$(e^{\sigma^2}-1)^{1/2}(e^{\sigma^2}+2)$
excess kurtosis [a]	$e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-6$
MGF	–
pdf [a2]	$\frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{y}\exp\left\{-\frac{[\log(y)-\mu]^2}{2\sigma^2}\right\}$
cdf	$\Phi\left(\frac{\log y-\mu}{\sigma}\right)$
Inverse cdf (y_p)	$e^{\mu+\sigma z_p}$ where $z_p = \Phi^{-1}(p)$
Reference	[a] Johnson et al. [1994] Chapter 14, p208-213 [a2] Johnson et al. [1994] Chapter 14, p208, Equation (14.2)' where $\xi = \mu, \sigma = \sigma$ and $\theta = 0$.

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

$$f_Y(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{y}\exp\left\{-\frac{[\log(y)-\mu]^2}{2\sigma^2}\right\} \tag{15.4}$$

for $y > 0$, where $-\infty < \mu < \infty$ and $\sigma > 0$. Note that $\log Y \sim NO(\mu, \sigma)$. Also if $Y \sim \texttt{LOGNO}(\mu, \sigma)$ then $Y_1 = aY \sim \texttt{LOGNO}(\mu + loga, \sigma)$ and $Z = Z/e^{\mu} \sim \texttt{LOGNO}(0, \sigma)$. So the $\texttt{LOGNO}(\mu, \sigma)$ is a scale family of distribution with scaling parameter e^{μ}. Hence μ itself is not a scaling parameter.

15.3.5 Pareto type 2 original distribution, PARETO2o(μ, σ)

Table 15.6: Pareto type 2 original distribution

PARETO2o(μ, σ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean	$\begin{cases} \dfrac{\mu}{(\sigma-1)}, & \text{if } \sigma > 1 \\ \infty, & \text{if } \sigma \leq 1 \end{cases}$
median	$\mu(2^{1/\sigma} - 1)$
mode	$\to 0$
variance [a]	$\begin{cases} \dfrac{\sigma\mu^2}{(\sigma-1)^2(\sigma-2)}, & \text{if } \sigma > 2 \\ \infty, & \text{if } \sigma \leq 2 \end{cases}$
skewness [a]	$\begin{cases} \dfrac{2(\sigma+1)(\sigma-2)^{1/2}}{(\sigma-3)\sigma^{1/2}}, & \text{if } \sigma > 3 \\ \infty, & \text{if } \sigma \leq 3 \end{cases}$
excess kurtosis [a]	$\begin{cases} \dfrac{3(\sigma-2)(3\sigma^2+\sigma+2)}{\sigma(\sigma-3)(\sigma-4)} - 3 & \text{if } \sigma > 4 \\ \infty, & \text{if } \sigma \leq 4 \end{cases}$
MGF	–
pdf [a2]	$\dfrac{\sigma\mu^\sigma}{(y+\mu)^{\sigma+1}}$
cdf [a]	$1 - \dfrac{\mu^\sigma}{(y+\mu)^\sigma}$
Inverse cdf (y_p)	$\mu[(1-p)^{-1/\sigma} - 1]$
Reference	[a]Johnson et al. [1994] Sections 20.3 and 20.4 p574-579. [a2] Johnson et al. [1994] Equation (20.3), p574 with $k = \mu, a = \sigma$ and $x = y + \mu$, so $X = Y + \mu$.

The pdf of the Pareto type 2 original distribution, denoted by PARETO2o(μ, σ), is defined by

$$f_Y(y|\mu, \sigma) = \frac{\sigma\mu^\sigma}{(y+\mu)^{\sigma+1}} \tag{15.5}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. [This was called the Pareto distribution of the second kind or Lomax distribution by Johnson et al. [1994], Section 20,3, p575, Equation (20.4), where $C = \mu$ and $a = \sigma$.]

15.3.6 Pareto type 2 distribution, PARETO2(μ, σ)

Table 15.7: Pareto type 2 distribution

PARETO2(μ, σ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$ scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean	$\begin{cases} \frac{\mu\sigma}{(1-\sigma)}, & \text{if } \sigma < 1 \\ \infty, & \text{if } \sigma \geq 1 \end{cases}$
median	$\mu(2^\sigma - 1)$
mode	$\to 0$
variance	$\begin{cases} \frac{\sigma^2\mu^2}{(1-\sigma)^2(1-2\sigma)}, & \text{if } \sigma < 1/2 \\ \infty, & \text{if } \sigma \geq 1/2 \end{cases}$
skewness	$\begin{cases} \frac{2(1+\sigma)(1-2\sigma)^{1/2}}{(1-3\sigma)}, & \text{if } \sigma < 1/3 \\ \infty, & \text{if } \sigma \geq 1/3 \end{cases}$
excess kurtosis	$\begin{cases} \frac{3(1-2\sigma)(2\sigma^2+\sigma+3)}{(1-3\sigma)(1-4\sigma)} - 3, & \text{if } \sigma < 1/4 \\ \infty, & \text{if } \sigma \geq 1/4 \end{cases}$
MGF	–
pdf	$\frac{\sigma^{-1}\mu^{1/\sigma}}{(y+\mu)^{(1/\sigma)+1}}$
cdf	$1 - \frac{\mu^{1/\sigma}}{(y+\mu)^{1/\sigma}}$
Inverse cdf (y_p)	$\mu[(1-p)^{-\sigma} - 1]$
Reference	Set σ to $1/\sigma$ in PARETO2o(μ, σ).

The pdf of the Pareto type 2 distribution, denoted by PARETO2(μ, σ), is defined by

$$f_Y(y|\mu, \sigma) = \frac{\sigma^{-1}\mu^{1/\sigma}}{(y+\mu)^{(1/\sigma)+1}} \tag{15.6}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. Note PARETO2 is given by re-parameterizing σ to $1/\sigma$ in PARETO2o, i.e. PARETO2(μ, σ) = PARETO2o$(\mu, 1/\sigma)$.

15.3.7 Weibull distribution, WEI(μ, σ), WEI2(μ, σ), WEI3(μ, σ)

First parametrization, WEI(μ, σ)

Table 15.8: Weibull distribution

WEI(μ, σ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean	$\mu\Gamma(\sigma^{-1}+1)$
median	$\mu(\log 2)^{1/\sigma}$
mode	$\begin{cases} \mu(1-\sigma^{-1})^{1/\sigma}, & \text{if } \sigma > 1 \\ \to 0, & \text{if } \sigma \leq 1 \end{cases}$
variance	$\mu^2\{\Gamma(2\sigma^{-1}+1) - [\Gamma(\sigma^{-1}+1)]^2\}$
skewness	$\begin{cases} \frac{\mu_3}{[Var(Y)]^{1.5}} \text{ where} \\ \mu_3 = \mu^3\{\Gamma(3\sigma^{-1}+1) - 3\Gamma(2\sigma^{-1}+1)\Gamma(\sigma^{-1}+1) \\ \quad +2[\Gamma(\sigma^{-1}+1)]^3\} \end{cases}$
excess kurtosis	$\begin{cases} \frac{\mu_4}{[Var(Y)]^2} - 3 \text{ where} \\ \mu_4 = \mu^4\{\Gamma(4\sigma^{-1}+1) - 4\Gamma(3\sigma^{-1}+1)\Gamma(\sigma^{-1}+1) \\ \quad +6\Gamma(2\sigma^{-1}+1)[\Gamma(\sigma^{-1}+1)]^2 - 3[\Gamma(\sigma^{-1}+1)]^4\} \end{cases}$
MGF	–
pdf [a]	$\frac{\sigma y^{\sigma-1}}{\mu^\sigma} \exp\left[-\left(\frac{y}{\mu}\right)^\sigma\right]$
cdf	$1 - \exp\left[-\left(\frac{y}{\mu}\right)^\sigma\right]$
Inverse cdf (y_p)	$\mu[-\log(1-p)]^{1/\sigma}$
Reference	Johnson et al. [1994] Chapter 21, p628-632. [a] Johnson et al. [1994] Equation (21.3), p629, with $\alpha = \mu, c = \sigma$ and $\xi_0 = 0$.

There are three versions of the two parameter Weibull distribution implemented into the **gamlss** package. The first, denoted by WEI(μ, σ), has the following parametrization

$$f_Y(y|\mu,\sigma) = \frac{\sigma y^{\sigma-1}}{\mu^\sigma} \exp\left[-\left(\frac{y}{\mu}\right)^\sigma\right] \tag{15.7}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$, see Johnson et al. [1994] p629.

Note that the (moment based) skewness is positive for $\sigma \leq 3.6023$ and negative for $\sigma \geq 3.6024$ (to 4 decimal places).

Second parametrization, `WEI2`(μ, σ)

Table 15.9: Second parametrization of Weibull distribution

`WEI2`(μ, σ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, NOT a scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean	$\mu^{-1/\sigma}\Gamma(\sigma^{-1}+1)$
median	$\mu^{-1/\sigma}(\log 2)^{1/\sigma}$
mode	$\begin{cases} \mu^{-1/\sigma}(1-\sigma^{-1})^{1/\sigma}, & \text{if } \sigma > 1 \\ \to 0, & \text{if } \sigma \leq 1 \end{cases}$
variance	$\mu^{-2/\sigma}\{\Gamma(2\sigma^{-1}+1) - [\Gamma(\sigma^{-1}+1)]^2\}$
skewness	$\begin{cases} \frac{\mu_3}{[Var(Y)]^{1.5}} \text{ where} \\ \mu_3 = \mu^{-3/\sigma}\{\Gamma(3\sigma^{-1}+1) - 3\Gamma(2\sigma^{-1}+1)\Gamma(\sigma^{-1}+1) \\ \quad +2[\Gamma(\sigma^{-1}+1)]^3\} \end{cases}$
excess kurtosis	$\begin{cases} \frac{\mu_4}{[Var(Y)]^2} - 3 \text{ where} \\ \mu_4 = \mu^{-4/\sigma}\{\Gamma(4\sigma^{-1}+1) - 4\Gamma(3\sigma^{-1}+1)\Gamma(\sigma^{-1}+1) \\ \quad +6\Gamma(2\sigma^{-1}+1)[\Gamma(\sigma^{-1}+1)]^2 - 3[\Gamma(\sigma^{-1}+1)]^4\} \end{cases}$
MGF	–
pdf	$\mu\sigma y^{\sigma-1}\exp(-\mu y^\sigma)$
cdf	$1 - \exp(-\mu y^\sigma)$
Inverse cdf (y_p)	$\mu^{-1/\sigma}[-\log(1-p)]^{1/\sigma}$
Reference	Set $\mu = \mu^{-1/\sigma}$ in `WEI`(μ, σ).

The second parametrization of the Weibull distribution, denoted by `WEI2`(μ, σ), is defined as

$$f_Y(y|\mu, \sigma) = \sigma\mu y^{\sigma-1}\exp(-\mu y^\sigma) \tag{15.8}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$. The parametrization (15.8) gives the usual proportional hazards Weibull model. Note `WEI2`(μ, σ) = `WEI`$(\mu^{-1/\sigma}, \sigma)$. [Hence μ in the `WEI2`(μ, σ) distribution can be reparameterized to $\alpha = \mu^{-1/\sigma}$ giving the scale family of distributions `WEI`(α, σ) with scaling parameter α.]

In this second parameterisation of the Weibull distribution, `WEI2`(μ, σ), the two parameters μ and σ are highly correlated, so the `RS` method of fitting is very slow and therefore the `CG()` or `mixed()` method of fitting should be used.

Table 15.10: Third parametrization of Weibull distribution

WEI3(μ, σ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, mean, scaling parameter
σ	$0 < \sigma < \infty$
Distribution measures	
mean	μ
median	$\beta(\log 2)^{1/\sigma}$ where $\beta = \mu[\Gamma(\sigma^{-1}+1)]^{-1}$
mode	$\begin{cases} \beta(1-\sigma^{-1})^{1/\sigma}, & \text{if } \sigma > 1 \\ \to 0, & \text{if } \sigma \leq 1 \end{cases}$
variance	$\beta^2\{\Gamma(2\sigma^{-1}+1) - [\Gamma(\sigma^{-1}+1)]^2\}$
skewness	$\begin{cases} \dfrac{\mu_3}{[Var(Y)]^{1.5}} \text{ where} \\ \mu_3 = \beta^3\{\Gamma(3\sigma^{-1}+1) - 3\Gamma(2\sigma^{-1}+1)\Gamma(\sigma^{-1}+1) \\ \quad +2[\Gamma(\sigma^{-1}+1)]^3\} \end{cases}$
excess kurtosis	$\begin{cases} \dfrac{\mu_4}{[Var(Y)]^2} - 3 \text{ where} \\ \mu_4 = \beta^4\{\Gamma(4\sigma^{-1}+1) - 4\Gamma(3\sigma^{-1}+1)\Gamma(\sigma^{-1}+1) \\ \quad +6\Gamma(2\sigma^{-1}+1)[\Gamma(\sigma^{-1}+1)]^2 \\ \quad -3[\Gamma(\sigma^{-1}+1)]^4\} \end{cases}$
MGF	–
pdf	$\dfrac{\sigma y^{\sigma-1}}{\beta^\sigma} \exp\left\{-\left(\dfrac{y}{\beta}\right)^\sigma\right\}$
cdf	$1 - \exp\left[-\left(\dfrac{y}{\beta}\right)^\sigma\right]$
Inverse cdf (y_p)	$\beta[-\log(1-p)]^{1/\sigma}$
Reference	Set μ to β in WEI(μ, σ), where $\beta = \mu[\Gamma(\sigma^{-1}+1)]^{-1}$.

Third parameterization $\texttt{WEI3}(\mu, \sigma)$

This is a parameterization of the Weibull distribution where μ is the mean of the distribution. This parameterization of the Weibull distribution, denoted by $\texttt{WEI3}(\mu, \sigma)$, is defined as

$$f_Y(y|\mu, \sigma) = \frac{\sigma y^{\sigma-1}}{\beta^\sigma} \exp\left\{-\left(\frac{y}{\beta}\right)^\sigma\right\} \tag{15.9}$$

for $y > 0$, where $\mu > 0$ and $\sigma > 0$ and where $\beta = \mu/\Gamma(\frac{1}{\sigma} + 1)$.

The parametrization (15.9) gives the usual accelerated lifetime Weibull model.

Note $\texttt{WEI3}(\mu, \sigma) = \texttt{WEI}(\beta, \sigma)$.

15.4 Continuous three parameter distribution on $(0, \infty)$

15.4.1 Box-Cox Cole and Green distribution, $\texttt{BCCG}(\mu, \sigma, \nu)$, $\texttt{BCCGo}(\mu, \sigma, \nu)$

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 by $\texttt{BCCG}(\mu, \sigma, \nu)$, defined through the transformed random variable Z given by

$$Z = \begin{cases} \frac{1}{\sigma\nu}\left[\left(\frac{Y}{\mu}\right)^\nu - 1\right], & \text{if} \quad \nu \neq 0 \\ \frac{1}{\sigma}\log(\frac{Y}{\mu}), & \text{if} \quad \nu = 0 \end{cases} \tag{15.10}$$

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^ν to be real for all ν) 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. Note that

$$Y = \begin{cases} \mu(1 + \sigma\nu Z)^{1/\nu}, & \text{if } \nu \neq 0 \\ \mu \exp(\sigma Z), & \text{if } \nu = 0 \end{cases} \tag{15.11}$$

See Figure 15.1 for a plot of the relationship between Y and Z for $\mu = 1, \sigma = 0.2$ and $\nu = -2$. Note that for this case $-\infty < Z < 2.5$.

Table 15.11: Box-Cox Cole and Green distribution distribution

BCCG(μ, σ, ν)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, median[a], scaling parameter
σ	$0 < \sigma < \infty$, approximate coefficient of variation
ν	$-\infty < \nu < \infty$, skewness parameter
Distribution measures	
mean	–
median [a]	μ
mode [a2]	$\begin{cases} \mu\omega^{1/\nu}, & \text{if } \nu \neq 0 \\ \text{where } \omega = \{1 + [1 + 4\sigma^2\nu(\nu-1)]^{1/2}\}/2 & \\ \mu e^{-\sigma^2}, & \text{if } \nu = 0 \end{cases}$
variance	–
skewness	–
excess kurtosis	–
MGF	–
pdf	$\dfrac{y^{\nu-1}\exp\left(-\frac{1}{2}z^2\right)}{\mu^\nu\sigma\sqrt{2\pi}\Phi\left(\frac{1}{\sigma\lvert\nu\rvert}\right)}$
cdf [a]	$\Phi(z)$ where z is given by (1.11)
Inverse cdf [a] (y_p)	$\begin{cases} \mu(1+\sigma\nu z_p)^{1/\nu}, & \text{if } \nu \neq 0 \\ \mu\exp(\sigma z_p) & \text{if } \nu = 0 \\ \text{where } z_p = \Phi^{-1}(p), & \end{cases}$
Reference	Set $\tau = 2$ in BCPE(μ, σ, ν, τ)
Notes	[a] Provided $\Phi\left(-\frac{1}{\sigma\lvert\nu\rvert}\right)$ is negligible for $\nu \neq 0$.
	[a2] If $0 < \nu < 1$ there is a second mode $\to 0$.

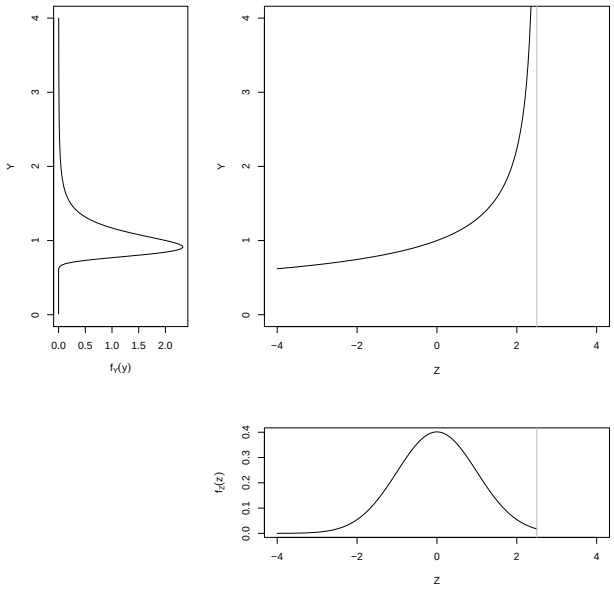


Figure 15.1: Relationship between Y and Z for the Box-Cox and Green distribution fir $\mu = 1, \sigma = 0.2$ and $\nu = -2$ (top right plot), $f_Z(z)$ (bottom plot), and $f_Y(y)$ (top left plot).

Hence the pdf of Y is given by

$$f_Y(y) = \frac{y^{\nu-1} \exp\left(-\frac{1}{2}z^2\right)}{\mu^\nu \sigma \sqrt{\pi} \Phi\left(\frac{1}{\sigma|\nu|}\right)} \tag{15.12}$$

for $y > 0$, where z is given by (15.10) and $\Phi()$ is the cumulative distribution function (cdf) of a standard normal distribution. [The distribution $BCGo(\mu, \sigma, \nu)$ also has pdf given by (15.12). It differs from $\texttt{BCCG}(\mu, \sigma, \nu)$ only in having a default log link function for μ, instead of the default identity link function for μ in $\texttt{BCCG}(\mu, \sigma, \nu)$.]

The exact cdf of Y is given by

$$F_Y(y) = \frac{\Phi(z) - \Phi\left(-\frac{1}{\sigma|\nu|}\right) (\text{ if } \nu > 0)}{\Phi\left(\frac{1}{\sigma|\nu|}\right)}$$

where z is given by (15.10). Hence $F_Y(y) \approx \Phi(z)$ provided the truncation probability $\Phi\left(-\frac{1}{\sigma|\nu|}\right)$ is negligible, and so $\Phi\left(\frac{1}{\sigma|\nu|}\right) \approx 1$. Also if the truncation probability $\Phi\left(-\frac{1}{\sigma|\nu|}\right)$ is negligible, the variable Y has median μ.

The exact inverse cdf (or quantile or centile) y_p of Y defined by $P(Y \leq y_p) = p$ is given by $y_p = \mu\left(1 + \sigma\nu z_T\right)^{1/\nu}$ and $z_T = \Phi^{-1}\left[p\Phi\left(\frac{1}{\sigma|\nu|}\right) + \Phi\left(-\frac{1}{\sigma|\nu|}\right) (\text{ if } \nu > 0)\right]$. Hence $z_T \approx \Phi^{-1}(p)$ provided the truncation probability $\Phi\left(-\frac{1}{\sigma|\nu|}\right)$ is negligible.

The parameterization in (15.10) was used by Cole and Green [1992] who assumed a standard normal distribution for Z and assumed that the truncation probability was negligible.

15.4.2 Generalized gamma distribution, $\texttt{GG}(\mu, \sigma, \nu)$

First parameterization, $\texttt{GG}(\mu, \sigma, \nu)$

The specific parameterization of the generalized gamma distribution used here and denoted by $\texttt{GG}(\mu,\sigma,\nu)$ was used by Lopatatzidis and Green [2000], with pdf defined by

$$f_Y(y|\mu, \sigma, \nu) = \frac{|\nu|\theta^\theta z^\theta \exp\{-\theta z\}}{\Gamma(\theta) y} \tag{15.13}$$

Table 15.12: Generalized gamma distribution

GG(μ, σ, ν)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, scaling parameter
σ	$0 < \sigma < \infty$
ν	$-\infty < \nu < \infty, \quad \nu \neq 0$
Distribution measures	
mean	$\begin{cases} \dfrac{\mu\Gamma(\theta + \frac{1}{\nu})}{\theta^{1/\nu}\Gamma(\theta)}, & \text{if } \{\nu > 0\} \text{ or } \{\nu < 0 \text{ and } \sigma^2\lvert\nu\rvert < 1\}, \\ & \text{where } \theta = 1/(\sigma^2\nu^2) \\ \infty, & \text{if } \nu < 0 \text{ and } \sigma^2\lvert\nu\rvert \geq 1 \end{cases}$
median	–
mode	$\begin{cases} \mu(1 - \sigma^2\nu)^{1/\nu}, & \text{if } \sigma^2\nu < 1 \\ \to 0, & \text{if } \sigma^2\nu \geq 1 \end{cases}$
variance	$\begin{cases} \dfrac{\mu^2}{\theta^{2/\nu}\left[\Gamma(\theta)\right]^2}\left\{\Gamma\left(\theta + \dfrac{2}{\nu}\right)\Gamma(\theta) - \left[\Gamma\left(\theta + \dfrac{1}{\nu}\right)\right]^2\right\}, & \text{if } \{\nu > 0\} \\ & \text{or } \{\nu < 0 \text{ and } \sigma^2\lvert\nu\rvert < 1/2\} \\ \infty, & \text{if } \{\nu < 0 \text{ and } \sigma^2\lvert\nu\rvert \geq 1/2\} \end{cases}$
skewness	$\begin{cases} \dfrac{\mu_3}{[Var(Y)]^{1.5}}, & \text{if } \{\nu > 0\} \text{ or } \{\nu < 0 \text{ and } \sigma^2\lvert\nu\rvert < 1/3\} \\ \text{where } \mu_3 = & \dfrac{\mu^3}{\theta^{3/\nu}[\Gamma(\theta)]^3}\left\{\Gamma\left(\theta + \dfrac{3}{\nu}\right)[\Gamma(\theta)]^2 - 3\Gamma\left(\theta + \dfrac{2}{\nu}\right) \times \right. \\ & \left. \Gamma\left(\theta + \dfrac{1}{\nu}\right)\Gamma(\theta) + 2\left[\Gamma\left(\theta + \dfrac{1}{\nu}\right)\right]^3\right\} \\ \infty, & \text{if } \{\nu < 0 \text{ and } \sigma^2\lvert\nu\rvert \geq 1/3\} \end{cases}$
excess kurtosis	$\begin{cases} \dfrac{\mu_4}{[Var(Y)]^2} - 3, & \text{if } \{\nu > 0\} \text{ or } \{\nu < 0 \text{ and } \sigma^2\lvert\nu\rvert < 1/4\} \\ \text{where } \mu_4 = & \dfrac{\mu^4}{\theta^{4/\nu}[\Gamma(\theta)]^4}\left\{\Gamma\left(\theta + \dfrac{4}{\nu}\right)[\Gamma(\theta)]^3 - 4\Gamma\left(\theta + \dfrac{3}{\nu}\right)\Gamma\left(\theta + \dfrac{1}{\nu}\right)[\Gamma(\theta)]^2 \right. \\ & \left. +6\Gamma\left(\theta + \dfrac{2}{\nu}\right)\left[\Gamma\left(\theta + \dfrac{1}{\nu}\right)\right]^2\Gamma(\theta) - 3\left[\Gamma\left(\theta + \dfrac{1}{\nu}\right)\right]^4\right\} \\ \infty, & \text{if } \{\nu < 0 \text{ and } \sigma^2\lvert\nu\rvert \geq 1/4\} \end{cases}$
MGF	–
pdf	$\dfrac{\lvert\nu\rvert\theta^\theta z^\theta \exp\{-\theta z\}}{\Gamma(\theta)y}$, where $z = (y/\mu)^\nu$ and $\theta = 1/(\sigma^2\nu^2)$
cdf	$\begin{cases} \gamma\left[\theta, \theta\left(\dfrac{y}{\mu}\right)^\nu\right]/\Gamma(\theta) & \text{if } \nu > 0 \\ \Gamma\left[\theta, \theta\left(\dfrac{y}{\mu}\right)^\nu\right]/\Gamma(\theta) & \text{if } \nu < 0 \end{cases}$
Inverse cdf (y_p)	–

for $y > 0$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, $\nu \neq 0$, and where $z = (y/\mu)^\nu$ and $\theta = 1/(\sigma^2\nu^2)$. Note that $Z = (Y/\mu)^\nu \sim \texttt{GA}(1, \sigma\nu)$ from which the results in Table 1.11 are obtained. Note $E(Y^r) = \mu^r\Gamma(\theta + \frac{r}{\nu})/\left[\theta^{r/\nu}\Gamma(\theta)\right]$ if $\theta > -r/\nu$.

Second parameterization, $GG2(\mu, \sigma, \nu)$

Note that $GG2(\mu, \sigma, \nu)$ is not currently implemented in **gamlss**.

A second parameterization, given by Johnson *et al.*, (1995), p401, denoted here by $GG2(\mu, \sigma, \nu)$, is defined as

$$f_Y(y|\mu, \sigma, \nu) = \frac{|\mu|y^{\mu\nu-1}}{\Gamma(\nu)\sigma^{\mu\nu}} \exp\left\{-\left(\frac{y}{\sigma}\right)^\mu\right\} \tag{15.14}$$

for $y > 0$, where $-\infty < \mu < \infty$, $\sigma > 0$ and $\nu > 0$.

The moments of $Y \sim GG2(\mu, \sigma, \nu)$ can be obtained from those of $\texttt{GG}(\mu, \sigma, \nu)$ since $\texttt{GG}(\mu, \sigma, \nu) = GG2(\nu, \mu\theta^{-1/\nu}, \theta)$ and $GG2(\mu, \sigma, \nu) = \texttt{GG}(\sigma\nu^{1/\mu}, \left[\mu^2\nu\right]^{-1/2}, \mu)$.

15.4.3 Generalized inverse Gaussian distribution, $\texttt{GIG}(\mu, \sigma, \nu)$

Table 15.13: Generalized inverse Gaussian distribution

$\texttt{GIG}(\mu, \sigma, \nu)$	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, mean, scaling parameter
σ	$0 < \sigma < \infty$
ν	$-\infty < \nu < \infty$
Distribution measures	
mean [a]	μ
median	–
mode [a2]	$\frac{\mu}{b}\{(\nu-1)\sigma^2 + [(\nu-1)^2\sigma^4 + 1]^{1/2}\}$
variance [a]	$\mu^2\left[\frac{2\sigma^2}{b}(\nu+1) + \frac{1}{b^2} - 1\right]$
skewness [a]	$\mu_3/[Var(Y)]^{1.5}$ where $\mu_3 = \mu^3\left[2 - \frac{6\sigma^2}{b}(\nu+1) + \frac{4}{b^2}(\nu+1)(\nu+2)\sigma^4 - \frac{2}{b^2} + \frac{2\sigma^2}{b^3}(\nu+2)\right]$
excess kurtosis [a]	$\frac{k_4}{[Var(Y)]^2}$ where $k_4 = \mu^4\{-6 + 24\frac{\sigma^2}{b}(\nu+1) + \frac{4}{b^2}[2 - \sigma^4(\nu+1)(7\nu+11)] + 4\frac{\sigma^2}{b^3}[2\sigma^4(\nu+1)(\nu+2)(\nu+3) - 4\nu - 5] + \frac{1}{b^4}[4\sigma^4(\nu+2)(\nu+3) - 2]\}$
MGF [a2]	$\left(1 - \frac{2\mu\sigma^2 t}{b}\right)^{-\nu/2}\left[K_\nu\left(\frac{1}{\sigma^2}\right)\right]^{-1} K_\nu\left[\frac{1}{\sigma^2}\left(1 - \frac{2\mu\sigma^2 t}{b}\right)^{1/2}\right]$
pdf [a]	$\left(\frac{b}{\mu}\right)^\nu\left[\frac{y^{\nu-1}}{2K_\nu\left(\frac{1}{\sigma^2}\right)}\right]\exp\left[-\frac{1}{2\sigma^2}\left(\frac{by}{\mu} + \frac{\mu}{by}\right)\right]$
cdf	–
Inverse cdf (y_p)	–
Reference	[a] Jørgensen [1982] page 6, Equation (2.2), reparameterized by $\eta = \mu/b$, $\omega = 1/\sigma^2$ and $\lambda = \nu$, and pages 15-17. [a2] Jørgensen [1982] page 1, Equation (1.1) reparameterized by $\chi = \mu/(\sigma^2 b)$ and $\psi = b/(\sigma^2\mu)$, with page 7, Equation (2.6), and page 12, Equation (2.9) where t is replaced by $-t$.
Note	$b = \left[K_{\nu+1}\left(1/\sigma^2\right)\right]\left[K_\nu\left(1/\sigma^2\right)\right]^{-1}$

The parameterization of the generalized inverse Gaussian distribution, denoted by $\texttt{GIG}(\mu, \sigma, \nu)$, is defined as

$$f_Y(y|\mu, \sigma, \nu) = \left(\frac{b}{\mu}\right)^{\nu} \left[\frac{y^{\nu-1}}{2K_\nu\left(\frac{1}{\sigma^2}\right)}\right] \exp\left[-\frac{1}{2\sigma^2}\left(\frac{by}{\mu} + \frac{\mu}{by}\right)\right] \qquad (15.15)$$

for $y > 0$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, where $b = \left[K_{\nu+1}\left(1/\sigma^2\right)\right]\left[K_\nu\left(1/\sigma^2\right)\right]^{-1}$ and$K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1}\exp\{-\frac{1}{2}t(x + x^{-1})\}dx$.

$\texttt{GIG}(\mu, \sigma, \nu)$ is a reparameterization of the generalized inverse Gaussian distribution of Jørgensen [1982]. Note also that $\texttt{GIG}(\mu, \sigma, -0.5) = \texttt{IG}(\mu, \sigma\mu^{-1/2})$ a reparameterization of the inverse Gaussian distribution.

15.4.4 Log normal family (i.e. original Box-Cox), $LNO(\mu, \sigma, \nu)$

The **gamlss** function $LNO(\mu, \sigma, \nu)$ allows the use of the Box-Cox power transformation approach, Box and Cox [1964], where a transformation is applied to Y in order to remove skewness, given by $Z = (Y^\nu - 1)/\nu(\textbf{if } \nu \neq 0) + \log(Y)(\textbf{if } \nu = 0)$. The transformed variable Z is then assumed to have a normal $NO(\mu, \sigma)$ distribution. The resulting distribution of Y is denoted by $LNO(\mu, \sigma, \nu)$. [Strictly this is not a proper distribution since $Y > 0$ and hence strictly Z should have a truncated normal distribution.] We have the resulting three parameter distribution

$$f_Y(y|\mu, \sigma, \nu) = \frac{y^{\nu-1}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] \qquad (15.16)$$

for $y > 0$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, and where $z = (y^\nu - 1)/\nu(\textbf{if } \nu \neq 0) + \log(y)(\textbf{if } \nu = 0)$. When $\nu = 0$, this results in the distribution in equation (15.4). The distribution in (15.16) can be fitted for fixed ν only, e.g. $\nu = 0.5$, using the following arguments of `gamlss()`: `family=LNO, nu.fix=TRUE, nu.start=0.5`. If ν is unknown, it can be estimated from its profile likelihood. Alternatively instead of (15.16), the more orthogonal parameterization of (15.16) given by the `BCCG` distribution in Section 15.4.1 can be used.

Table 15.14: Box-Cox t distribution

BCT(μ, σ, ν, τ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, median[a], scaling parameter
σ	$0 < \sigma < \infty$, approximate coefficient of variation
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$ kurtosis parameter
Distribution measures	
mean	–
median[a]	μ
mode	–
variance	–
skewness	–
excess kurtosis	–
MGF	–
pdf	$\dfrac{y^{\nu-1} f_T(z)}{\mu^{\nu}\sigma F_T\left(\dfrac{1}{\sigma\lvert\nu\rvert}\right)}$ where $T \sim t_\tau = TF(0, 1, \tau)$
cdf[a]	$F_T(z)$ where $T \sim t_\tau$ and z is given by (15.10)
Inverse cdf [a] (y_p)	$\begin{cases} \mu(1+\sigma\nu t_{p,\tau})^{1/\nu}, & \text{if } \nu \neq 0 \\ \mu \exp(\sigma t_{p,\tau}), & \text{if } \nu = 0 \\ \text{where } t_{p,\tau} = F_T^{-1}(p) \text{ and } T \sim t_\tau \end{cases}$
Reference	Rigby and Stasinopoulos [2006]
Notes	[a] Provided $F_T\left(-\dfrac{1}{\sigma\lvert\nu\rvert}\right)$ is neglibible for $\nu \neq 0$.

15.5 Continuous four parameter distributions on $(0, \infty)$

15.5.1 Box-Cox t distribution, $\texttt{BCT}(\mu, \sigma, \nu, \tau)$, $BCTo(\mu, \sigma, \nu, \tau)$

Let Y be a positive random variable having a Box-Cox t distribution, **?** denoted by $\texttt{BCT}(\mu, \sigma, \nu, \tau)$, defined through the transformed random variable Z given by (15.10), 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 $\texttt{BCT}(\mu, \sigma, \nu, \tau)$ random variable, is given by

$$f_Y(y|\mu, \sigma, \nu, \tau) = \frac{y^{\nu-1} f_T(z)}{\mu^\nu \sigma F_T\left(\frac{1}{\sigma|\nu|}\right)} \tag{15.17}$$

for $y > 0$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, and where z is given by (15.10) 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 $\tau > 0$, ie $T \sim t_\tau = TF(0, 1, \tau)$, see Section 2.2.5. Hence

$$f_T(z) = \frac{1}{B\left(\frac{1}{2}, \frac{\tau}{2}\right)\tau^{1/2}} \left[1 + \frac{z^2}{\tau}\right]^{-(\tau+1)/2)}.$$

[The distribution $BCTo(\mu, \sigma, \nu, \tau)$ also has pdf given by (15.17). It differs from $\texttt{BCT}(\mu, \sigma, \nu, \tau)$ only in having default log link function for μ, instead of the default identity link function for μ in $\texttt{BCT}(\mu, \sigma, \nu, \tau)$.] Note that $BCCG(\mu, \sigma, \nu)$ is a limiting distribution of $\texttt{BCT}(\mu, \sigma, \nu, \tau)$ as $\tau \to \infty$. The exact cdf of Y is given by

$$F_Y(y) = \frac{F_T(z) - F_T\left(-\frac{1}{\sigma|\nu|}\right) \text{ if } \nu > 0}{F_T\left(\frac{1}{\sigma|\nu|}\right)}, \tag{15.18}$$

where z is given by (15.10). The exact inverse cdf y_p of Y, defined by $P(Y \leq y_p) = p$, is given by $y_p = \mu(1 + \sigma\nu z_T)^{1/\nu}$, for $\nu \neq 0$, and $y_p = \mu \exp(\sigma z_T)$ for $\nu = 0$, where

$$z_T = F_T\left[pF_T\left(\frac{1}{\sigma|\nu|}\right) + F_T\left(-\frac{1}{\sigma|\nu|}\right) \text{ if } \nu > 0\right], \tag{15.19}$$

Rigby and Stasinopoulos [2006].

If the truncation probability $F_T(-\frac{1}{\sigma|\nu|})$ is negligible, then $F_Y(y) = F_T(z)$ and $z_T = F_T^{-1}(p) = t_{p,\tau}$ in (15.19), so the variable Y has median μ. The mean of Y is finite if $\nu < -1$ and also for $\tau > 1/\nu$ if $\nu > 0$. The variance of Y is finite if $\nu < -2$ and also for $\tau > 2/\nu$ if $\nu > 0$.

15.5.2 Box-Cox power exponential distribution, BCPE(μ, σ, ν, τ), BCPEo(μ, σ, ν, τ)

Table 15.15: Box-Cox power exponential distribution

BCPE(μ, σ, ν, τ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, median[a], scaling parameter
σ	$0 < \sigma < \infty$, approximate coefficient of variation
ν	$-\infty < \nu < \infty$, skewness parameter
τ	$0 < \tau < \infty$ kurtosis parameter
Distribution measures	
mean	–
median [a]	μ
mode	–
variance	–
skewness	–
excess kurtosis	–
MGF	–
pdf	$\frac{y^{\nu-1} f_T(z)}{\mu^\nu \sigma F_T\left(\frac{1}{\sigma\lvert\nu\rvert}\right)}$ where $T \sim PE(0,1,\tau)$
cdf [a]	$F_T(z)$ where $T \sim PE(0,1,\tau)$ and z is given by (15.10)
Inverse cdf [a] (y_p)	$\begin{cases} \mu\left[1+\sigma\nu F_T^{-1}(p)\right]^{1/\nu}, & \text{if } \nu \neq 0 \\ \mu \exp\left[\sigma F_T^{-1}(p)\right], & \text{if } \nu = 0 \\ \text{where } T \sim PE(0,1,\tau) \end{cases}$
Reference	Rigby and Stasinopoulos [2004]
Notes	[a] Provided $F_T\left(-\frac{1}{\sigma\lvert\nu\rvert}\right)$ is neglibible for $\nu \neq 0$.

Let Y be a positive random variable having a Box-Cox power exponential distribution, Rigby and Stasinopoulos [2004], denoted by BCPE(μ, σ, ν, τ), defined through the transformed random variable Z given by (15.10), 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 BCPE(μ, σ, ν, τ) random variable, is given by (15.17), where $f_T(t)$ and $F_T(t)$ are respectively the pdf and cumulative distribution function of a variable T having a standard power exponential distribution, $T \sim PE(0, 1, \tau)$, see Section 2.2.2. Hence

$$f_T(z) = \frac{\tau \exp\left[-\left|\frac{z}{c}\right|^\tau\right]}{2c\Gamma\left(\frac{1}{\tau}\right)}$$

where $c^2 = \Gamma(1/\tau)[\Gamma(3/\tau)]^{-1}$. Note that $BCCG(\mu, \sigma, \nu) = \texttt{BCPE}(\mu, \sigma, \nu, 2)$ is a special case of $\texttt{BCPE}(\mu, \sigma, \nu, \tau)$ when $\tau = 2$. [The $BCPEo(\mu, \sigma, \nu, \tau)$ distribution also has the same pdf as $\texttt{BCPE}(\mu, \sigma, \nu, \tau)$ distribution. It differs from $\texttt{BCPE}(\mu, \sigma, \nu, \tau)$ only in having default log link function for μ, instead of the default identity link function for μ in $\texttt{BCPE}(\mu, \sigma, \nu, \tau)$.]

The exact cdf of Y is given by (15.18) where $T \sim PE(0, 1, \tau)$. The exact inverse cdf y_p of Y is given by $y_p = \mu(1 + \sigma\nu z_T)^{1/\nu}$ for $\nu \neq 0$, and $y_p = \mu \exp(\sigma z_T)$ for $\nu = 0$, where z_T is given by (15.19) and $T \sim PE(0, 1, \tau)$, Rigby and Stasinopoulos [2004].

If the truncation probability $F_T(-\frac{1}{\sigma|\nu|})$ is negligible, then $F_Y(y) = F_T(z)$ and $z_T = F_T^{-1}(p)$ in (15.19) where $T \sim PE(0, 1, \tau)$, so the variable Y has median μ.

15.5.3 Generalized beta type 2 distribution, $\texttt{GB2}(\mu, \sigma, \nu, \tau)$

This pdf of the generalized beta type 2 distribution, denoted by $\texttt{GB2}(\mu, \sigma, \nu, \tau)$, is defined by

$$\begin{aligned} f_Y(y|\mu, \sigma, \nu, \tau) &= |\sigma| y^{\sigma\nu - 1} \left\{ \mu^{\sigma\nu} B(\nu, \tau) \left[1 + (y/\mu)^{\sigma}\right]^{\nu+\tau} \right\}^{-1} \\ &= \frac{\Gamma(\nu + \tau)}{\Gamma(\nu)\Gamma(\tau)} \frac{|\sigma|(y/\mu)^{\sigma\nu}}{y\left[1 + (y/\mu)^{\sigma}\right]^{\nu+\tau}} \end{aligned} \tag{15.20}$$

for $y > 0$, where $\mu > 0$, $\sigma > 0$, $\nu > 0$ and $\tau > 0$, McDonald [1984] page 648, Equation (3). Note that McDonald and Xu (1995) allow for $\sigma < 0$, however this is unnecessary since $\texttt{GB2}(\mu, -\sigma, \nu, \tau) = \texttt{GB2}(\mu, \sigma, \tau, \nu)$. So we assume $\sigma > 0$, and $|\sigma|$ can be replaced by σ in (15.20). The GB2 distribution is also considered by McDonald and Xu [1995] page 136-140 and McDonald [1996] page 433-435.

Note that if $Y \sim \texttt{GB2}(\mu, \sigma, \nu, \tau)$ then $Y_1 = [1 + (Y/\mu)^{-\sigma}]^{-1} \sim BEo(\nu, \tau)$ from which the cdf of Y in Table 1.13 is obtained.

Note $E(Y^r) = \mu^r B(\nu + r\sigma^{-1}, \tau - r\sigma^{-1})/B(\nu, \tau)$ for $\sigma > r\tau^{-1}$, McDonald and Xu [1995] page 136, Equation (2.8), from which the mean, variance, skewness and excess kurtosis in Table 1.13 are obtained

Setting $\sigma = 1$ in (15.20) gives a form of the Pearson type VI distribution:

$$f_Y(y|\mu, \nu, \tau) = \frac{\Gamma(\nu + \tau)}{\Gamma(\nu)\Gamma(\tau)} \frac{\mu^{\tau} y^{\nu-1}}{(y + \mu)^{\nu+\tau}}. \tag{15.21}$$

Setting $\nu = 1$ in (15.20) gives the *Burr* XII (or Singh-Maddala) distribution:

$$f_Y(y|\mu, \sigma, \tau) = \frac{\tau\sigma(y/\mu)^{\sigma}}{y\left[1 + (y/\mu)^{\sigma}\right]^{\tau+1}}. \tag{15.22}$$

Table 15.16: Generalized Beta type 2 distribution

GB2(μ, σ, ν, τ)	
Ranges	
Y	$0 < y < \infty$
μ	$0 < \mu < \infty$, scaling parameter
σ	$0 < \sigma < \infty$
ν	$0 < \nu < \infty$
τ	$0 < \tau < \infty$
Distribution measures	
mean	$\begin{cases} \mu \dfrac{B(\nu+\sigma^{-1}, \tau-\sigma^{-1})}{B(\nu,\tau)}, & \text{if } \tau > \sigma^{-1} \\ \infty, & \text{if } \tau \leq \sigma^{-1} \end{cases}$
median	–
mode	$\begin{cases} \mu \left(\dfrac{\sigma\nu - 1}{\sigma\tau + 1}\right)^{1/\sigma}, & \text{if } \nu > \sigma^{-1} \\ \to 0, & \text{if } \nu \leq \sigma^{-1} \end{cases}$
variance	$\begin{cases} \dfrac{\mu^2\{B(\nu+2\sigma^{-1}, \tau-2\sigma^{-1})B(\nu,\tau) - [B(\nu+\sigma^{-1}, \tau-\sigma^{-1})]^2\}}{[B(\nu,\tau)]^2}, & \text{if } \tau > 2\sigma^{-1} \\ \infty, & \text{if } \tau \leq 2\sigma^{-1} \end{cases}$
skewness	$\begin{cases} \dfrac{\mu_3}{[Var(Y)]^{1.5}}, & \text{if } \tau > 3\sigma^{-1} \\ \text{where } \mu_3 = \mu^3 \dfrac{1}{[B(\nu,\tau)]^3} \{B(\nu+3\sigma^{-1}, \tau-3\sigma^{-1})[B(\nu,\tau)]^2 \\ \quad -3B(\nu+2\sigma^{-1}, \tau-2\sigma^{-1})B(\nu+\sigma^{-1}, \tau-\sigma^{-1})B(\nu,\tau) \\ \quad +2[B(\nu+\sigma^{-1}, \tau-\sigma^{-1})]^3\} \\ \infty, & \text{if } \tau \leq 3\sigma^{-1} \end{cases}$
excess kurtosis	$\begin{cases} \dfrac{\mu_4}{[Var(Y)]^2} - 3, & \text{if } \tau > 4\sigma^{-1}, \\ \text{where } \mu_4 = \mu^4 \dfrac{1}{[B(\nu,\tau)]^4} \{B(\nu+4\sigma^{-1}, \tau-4\sigma^{-1})[B(\nu,\tau)]^3 \\ \quad -4B(\nu+3\sigma^{-1}, \tau-3\sigma^{-1})B(\nu+\sigma^{-1}, \tau-\sigma^{-1})[B(\nu,\tau)]^2 \\ \quad +6B(\nu+2\sigma^{-1}, \tau-2\sigma^{-1})[B(\nu+\sigma^{-1}, \tau-\sigma^{-1})]^2 B(\nu,\tau) \\ \quad -3[B(\nu+\sigma^{-1}, \tau-\sigma^{-1})]^4\} \\ \infty, & \text{if } \tau \leq 4\sigma^{-1} \end{cases}$
MGF	–
pdf[a]	$\lvert\sigma\rvert y^{\sigma\nu-1} \left\{\mu^{\sigma\nu} B(\nu,\tau) \left[1 + (y/\mu)^{\sigma}\right]^{\nu+\tau}\right\}^{-1}$
cdf	$\dfrac{B(\nu,\tau,c)}{B(\nu,\tau)}$ where $c = 1/[1 + (y/\mu)^{-\sigma}]$
Inverse cdf (y_p)	–
Reference	[a] McDonald [1984] page 648, Equation (3) where $b = \mu$, $a = \sigma$, $p = \nu$ and $q = \tau$.
Note	$B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function, $B(a,b,x) = \int_0^x t^{a-1}(1-t)^{b-1}dt$ is the incomplete beta function.

Setting $\tau = 1$ in (15.20) gives the Burr III (or Dagum) distribution

$$f_Y(y|\mu, \sigma, \nu) = \frac{\nu\sigma(y|\mu)^{\sigma\nu}}{y[1 + (y/\mu)^{\sigma}]^{\nu+1}}$$

Setting $\sigma = 1$ and $\nu = 1$ in (15.20) gives the Pareto type 2 original distribution, PARETO2o(μ, τ). Setting $\nu = 1$ and $\tau = 1$ in (15.20) gives the log logistic distribution. Other special cases and limiting cases are given in McDonald and Xu [1995] pages 136 and 139.

Chapter 16

Mixed distributions on $[0, \infty)$ including 0

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

Table 16.1: Zero adjusted gamma distribution

$ZAGA(\mu, \sigma, \nu)$	
Ranges	
Y	$0 \leq y < \infty$,
μ	$0 < \mu < \infty$, mean of gamma component
σ	$0 < \sigma < \infty$, coefficient of variation of gamma component
ν	$0 < \nu < 1$, exact probability that $Y = 0$
Distribution measures	
mean	$(1-\nu)\mu$
median	–
mode	$\begin{cases} 0, & \text{if } \sigma \geq 1 \\ 0 \text{ and } \mu(1-\sigma^2), & \text{if } \sigma^2 < 1 \end{cases}$
variance	$(1-\nu)\mu^2(\sigma^2+\nu)$
skewness	$\begin{cases} \dfrac{\mu_3}{[Var(Y)]^{1.5}} \text{ where} \\ \mu_3 = \mu^3(1-\nu)\left(2\sigma^4 + 3\nu\sigma^2 + 2\nu^2 - \nu\right) \end{cases}$
excess kurtosis	$\begin{cases} \dfrac{\mu_4}{[Var(Y)]^2} - 3, \text{ where} \\ \mu_4 = \mu^4(1-\nu)\left[6\sigma^6 + 3\sigma^4 + 8\nu\sigma^4 + 6\nu^2\sigma^2 + \nu(1-3\nu+3\nu^2)\right] \end{cases}$
MGF	$\nu + (1-\nu)(1-\mu\sigma^2 t)^{-1/\sigma^2}$ for $t < (\mu\sigma^2)^{-1}$
pdf	$\begin{cases} \nu, & \text{if } y = 0 \\ (1-\nu)\left[\dfrac{1}{(\sigma^2\mu)^{1/\sigma^2}} \dfrac{y^{1/\sigma^2-1}\, e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}\right], & \text{if } y > 0 \end{cases}$
cdf	$\begin{cases} \nu, & \text{if } y = 0 \\ \nu + (1-\nu)\dfrac{\gamma(\sigma^{-2}, y\mu^{-1}\sigma^{-2})}{\Gamma(\sigma^{-2})}, & \text{if } y > 0 \end{cases}$
Inverse cdf (y_p)	–
Reference	Obtained from equations in Mixed Distributions Chapter, where $Y_1 \sim GA(\mu, \sigma)$

The zero adjusted gamma distribution is appropriate when the response variable Y takes values from zero to infinity including zero, i.e. $[0,\infty)$. Here $Y=0$ with non zero probability ν and $Y\sim GA(\mu,\sigma)$ with probability $(1-\nu)$. The mixed continuous-discrete probability (density) function of the zero adjusted gamma distribution, denoted by $ZAGA(\mu,\sigma,\nu)$, is given (informally) by

$$f_Y(y|\mu,\sigma,\nu)=\begin{cases}\nu, & \text{if } y=0\\ (1-\nu)\left[\dfrac{1}{(\sigma^2\mu)^{1/\sigma^2}}\dfrac{y^{1/\sigma^2-1}\ e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}\right], & \text{if } y>0\end{cases} \tag{16.1}$$

for $0\le y<\infty$, where $\mu>0$, $\sigma>0$ and $0<\nu<1$.

16.1.1 Zero adjusted Inverse Gaussian distribution, $ZAIG(\mu,\sigma,\nu)$

Table 16.2: Zero adjusted inverse Gaussian distribution

$ZAIG(\mu,\sigma,\nu)$	
Ranges	
Y	$0\le y<\infty$
μ	$0<\mu<\infty$, mean of inverse Gaussian component
σ	$0<\sigma<\infty$
ν	$0<\nu<1$, exact probability that $Y=0$
Distribution measures	
mean	$(1-\nu)\mu$
median	–
mode	0 and $\dfrac{-3\mu^2\sigma^2+\mu(9\mu^2\sigma^4+4)^{1/2}}{2}$
variance	$(1-\nu)\mu^2(\nu+\mu\sigma^2)$
skewness	$\begin{cases}\mu_3/[Var(Y)]^{1.5}\text{ where}\\ \mu_3=\mu^3(1-\nu)\left[3\mu^2\sigma^4+3\mu\sigma^2\nu+2\nu^2-\nu\right]\end{cases}$
excess kurtosis	$\begin{cases}\mu_4/[Var(Y)]^2-3,\text{ where}\\ \mu_4=\mu^4(1-\nu)\left[3\mu^2\sigma^4+15\mu^3\sigma^6+\nu(1+12\mu^2\sigma^4)+\nu^2(6\mu\sigma^2-3)+3\nu^3\right]\end{cases}$
MGF	$\nu+(1-\nu)\exp\left\{\dfrac{1}{\mu\sigma^2}\left[1-\left(1-2\mu^2\sigma^2t\right)^{1/2}\right]\right\}$ for $t<(2\mu^2\sigma^2)^{-1}$
pdf	$\begin{cases}\nu, & \text{if } y=0\\ (1-\nu)\dfrac{1}{\sqrt{2\pi\sigma^2y^3}}\ \exp\left[-\dfrac{1}{2\mu^2\sigma^2y}\ (y-\mu)^2\right], & \text{if } y>0\end{cases}$
cdf	$\nu+(1-\nu)\Phi\left[(\sigma^2y)^{-1/2}\left(\dfrac{y}{\mu}-1\right)\right]+e^{2(\mu\sigma^2)^{-1}}\Phi\left[-(\sigma^2y)^{-1/2}\left(\dfrac{y}{\mu}+1\right)\right]$
Inverse cdf (y_p)	–
Reference	Obtained from equations in Mixed Distributions chapter, where $Y_1\sim IG(\mu,\sigma)$.

The zero adjusted inverse Gaussian distribution is appropriate when the response variable Y takes values from zero to infinity including zero, i.e. $[0, \infty)$. Here $Y = 0$ with non zero probability ν and $Y \sim IG(\mu, \sigma)$ with probability $(1 - \nu)$. The mixed continuous-discrete probability (density) function of the zero adjusted inverse Gaussian distribution, denoted by $ZAIG(\mu, \sigma, \nu)$, is given (informally) by

$$f_Y(y|\mu, \sigma, \nu) = \begin{cases} \nu, & \text{if } y = 0 \\ (1-\nu)\dfrac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left[-\dfrac{1}{2\mu^2\sigma^2 y}(y-\mu)^2\right], & \text{if } y > 0 \end{cases} \quad (16.2)$$

for $0 \leq y < \infty$, where $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$.

Chapter 17

Continuous and mixed distributions on $[0, 1]$

17.1 Continuous two parameter distributions on $(0, 1)$ excluding 0 and 1

17.1.1 Beta distribution, BE(μ, σ), BEo(μ, σ)

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(μ, σ)

The original parameterization of the beta distribution, denoted by BEo(μ, σ), has pdf given by $f_Y(y|\mu, \sigma) = \frac{1}{B(\mu,\sigma)}\ y^{\mu-1}(1-y)^{\sigma-1}$ 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(μ, σ)

In the second parameterization of the beta distribution below the parameters μ and σ are location and scale parameters that relate to the mean and standard deviation of Y. The pdf of the beta distribution, denoted by BE(μ, σ), is defined

by

$$f_Y(y|\mu,\sigma) = \frac{1}{B(\alpha,\beta)}\, y^{\alpha-1}(1-y)^{\beta-1} \tag{17.1}$$

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 (μ,σ) and (α,β) is given by $\mu = \alpha/(\alpha+\beta)$ and $\sigma = (\alpha+\beta+1)^{-1/2}$.] Hence $\texttt{BE}(\mu,\sigma) = \texttt{BEo}(\alpha,\beta)$. In this parameterization, the mean of Y is $E(Y) = \mu$ and the variance is $Var(Y) = \sigma^2\mu(1-\mu)$.

17.2 Continuous four parameter distributions on $(0,1)$ excluding 0 and 1

17.2.1 Generalized Beta type 1 distribution, $\texttt{GB1}(\mu,\sigma,\nu,\tau)$

The generalized beta type 1 distribution is defined by assuming $Z = Y^\tau/[\nu + (1-\nu)Y^\tau] \sim \texttt{BE}(\mu,\sigma) = \texttt{BEo}(\alpha,\beta)$. Hence, the pdf of generalized beta type 1 distribution, denoted by $\texttt{GB1}(\mu,\sigma,\nu,\tau)$, is given by

$$f_Y(y|\mu,\sigma,\nu,\tau) = \frac{\tau\nu^\beta y^{\tau\alpha-1}(1-y^\tau)^{\beta-1}}{B(\alpha,\beta)[\nu+(1-\nu)y^\tau]^{\alpha+\beta}} \tag{17.2}$$

for $0 < y < 1$, where $0 < \mu < 1$, $0 < \sigma < 1$, $\nu > 0$ and $\tau > 0$, and where $\alpha = \mu(1-\sigma^2)/\sigma^2$ and $\beta = (1-\mu)(1-\sigma^2)/\sigma^2$, for $\alpha > 0$ and $\beta > 0$. Hence $\texttt{GB1}(\mu,\sigma,\nu,\tau)$ has adopted parameters $\mu = \alpha/(\alpha+\beta)$, $\sigma = (\alpha+\beta+1)^{-1/2}$, ν and τ.

For $0 < \nu < 1$, $GB1(\mu,\sigma,\nu,\tau)$ is a re-parameterized submodel of the generalized beta (GB) distribution of McDonald and Xu (1995) equation (2.8) where $GB1(\mu,\sigma,\nu,\tau) = GB(\tau,\nu^{1/\tau},1-\nu,\alpha,\beta)$. [Note that $GB1$ is different from the generalized beta of the first kind of McDonald and Xu (1995).] The generalized three parameter beta (GB3) of Pham-Gia and Duong (1989) and Johnson *et al.* (1995), Section 25.7, p251 is a re-parameterized submodel of $GB1$ given by $G3B(\alpha_1,\alpha_2,\lambda) = GB1(\alpha_1(\alpha_1+\alpha_2)^{-1},(\alpha_1+\alpha_2+1)^{-1},\lambda^{-1},1)$. The beta $\texttt{BE}(\mu,\sigma)$ distribution is a submodel of $\texttt{GB1}(\mu,\sigma,\nu,\tau)$ where $\nu = 1$ and $\tau = 1$, i.e. $\texttt{BE}(\mu,\sigma) = \texttt{GB1}(\mu,\sigma,1,1)$.

17.3 Mixed distributions on $[0,1)$, $(0,1]$ or $[0,1]$, i.e. including 0, 1 or both.

17.3.1 Beta inflated distribution $BEINF(\mu,\sigma,\nu,\tau)$, $BEINF0(\mu,\sigma,\nu)$, $BEINF1(\mu,\sigma,\nu)$

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

$$f_Y(y|\mu,\sigma,\nu,\tau)=\begin{cases} p_0 & \text{if } y=0 \\ (1-p_0-p_1)\frac{1}{B(\alpha,\beta)}\, y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0<y<1 \\ p_1 & \text{if } y=1 \end{cases} \quad (17.3)$$

for $0\le y\le 1$, where $\alpha=\mu(1-\sigma^2)/\sigma^2$, $\beta=(1-\mu)(1-\sigma^2)/\sigma^2$, $p_0=\nu(1+\nu+\tau)^{-1}$, $p_1=\tau(1+\nu+\tau)^{-1}$ 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 $E(y)=\frac{\tau+\mu}{(1+\nu+\tau)}$.

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

$$f_Y(y|\mu,\sigma,\nu)=\begin{cases} p_0 & \text{if } y=0 \\ (1-p_0)\frac{1}{B(\alpha,\beta)}\, y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0<y<1 \end{cases} \quad (17.4)$$

for $0\le y<1$, 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 $\sigma=(\alpha+\beta+1)^{-1/2}$, $\nu=p_0/1-p_0$. Hence $0<\mu<1$, $0<\sigma<1$, $\nu>0$. Note that for $BEINF0(\mu,\sigma,\nu)$, $E(y)=\frac{\mu}{(1+\nu)}$.

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

$$f_Y(y|\mu,\sigma,\nu)=\begin{cases} (1-p_1)\frac{1}{B(\alpha,\beta)}\, y^{\alpha-1}(1-y)^{\beta-1} & \text{if } 0<y<1 \\ p_1 & \text{if } y=1 \end{cases} \quad (17.5)$$

for $0<y\le 1$, where $\alpha=\mu(1-\sigma^2)/\sigma^2$, $\beta=(1-\mu)(1-\sigma^2)/\sigma^2$, $p_1=\nu(1+\nu)^{-1}$ 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 $E(y)=\frac{\nu+\mu}{(1+\nu)}$.

For different parametrizations of the $BEINF0(\mu, \sigma, \nu)$ and $BEINF1(\mu, \sigma, \nu)$ distributions see $BEZI(\mu, \sigma, \nu)$ and $BEOI(\mu, \sigma, \nu)$ distributions contributed to **gamlss** by Raydonal Ospina, Ospina and Ferrari (2010).

Chapter 18

Count data distributions

For the tables of this section we use the following notation:

PGF: probability generating function

pf: probability function

cdf: cumulative distribution function

inverse cdf inverse cumulative distribution function

Distribution	`gamlss` name	Range R_Y	Parameter link function μ	σ	ν	τ
geometric	`GEOM`	$\{0,1,2,\ldots\}$	log	-	-	-
geometric (original)	`GEOMo`	$\{0,1,2,\ldots\}$	logit	-	-	-
logarithmic	`LG`	$\{1,2,3,\ldots\}$	logit	-	-	-
Poisson	`PO`	$\{0,1,2,\ldots\}$	log	-	-	-
Yule (μ the mean)	`YULE`	$\{0,1,2,\ldots\}$	log	-	-	-
zipf	`ZIPF`	$\{0,1,2,\ldots\}$	log	-	-	-
negative binomial type I	`NBI`	$\{0,1,2,\ldots\}$	log	log	-	-
negative binomial type II	`NBII`	$\{0,1,2,\ldots\}$	log	log	-	-
Poisson inverse Gaussian	`PIG`	$\{0,1,2,\ldots\}$	log	log	-	-
Waring (μ the mean)	`WARING`	$\{0,1,2,\ldots\}$	log	log	-	-
zero alt. logarithmic	`ZALG`	$\{0,1,2,\ldots\}$	logit	logit	-	-
zero alt. Poisson	`ZAP`	$\{0,1,2,\ldots\}$	log	logit	-	-
zero alt. zipf	`ZAZIPF`	$\{0,1,2,\ldots\}$	log	logit	-	-
zero inf. Poisson	`ZIP`	$\{0,1,2,\ldots\}$	log	logit	-	-
zero inf. Poisson (μ the mean)	`ZIP2`	$\{0,1,2,\ldots\}$	log	logit	-	-
generalised Poisson	`GPO`	$\{0,1,2,\ldots\}$	log	log	-	-
double Poisson	`DPO`	$\{0,1,2,\ldots\}$	log	log	-	-
beta neg. binomial	`BNB`	$\{0,1,2,\ldots\}$	log	log	log	-
neg. binomial family	`NBF`	$\{0,1,2,\ldots\}$	log	log	ident.	-
Delaporte	`DEL`	$\{0,1,2,\ldots\}$	log	log	logit	-

Sichel	`SI`	$\{0, 1, 2, \ldots\}$	log	log	ident.	-
Sichel (μ the mean)	`SICHEL`	$\{0, 1, 2, \ldots\}$	log	log	ident.	-
zero alt. neg. binomial	`ZANBI`	$\{0, 1, 2, \ldots\}$	log	log	logit	-
zero alt. PIG	`ZAPIG`	$\{0, 1, 2, \ldots\}$	log	log	logit	-
zero inf. neg. binomial	`ZINBI`	$\{0, 1, 2, \ldots\}$	log	log	logit	-
zero inf. PIG	`ZIPIG`	$\{0, 1, 2, \ldots\}$	log	log	logit	-
zero alt. neg. binom. fam.	`ZANBF`	$\{0, 1, 2, \ldots\}$	log	log	log	logit
zero alt. beta neg. binom.	`ZABNB`	$\{0, 1, 2, \ldots\}$	log	log	ident.	logit
zero alt. Sichel	`ZASICHEL`	$\{0, 1, 2, \ldots\}$	log	log	ident.	logit
zero inf. neg. binom. fam.	`ZINBF`	$\{0, 1, 2, \ldots\}$	log	log	log	logit
zero inf. beta neg. binom.	`ZIBNB`	$\{0, 1, 2, \ldots\}$	log	log	log	logit
zero inf. Sichel	`ZISICHEL`	$\{0, 1, 2, \ldots\}$	log	log	ident.	logit
Poisson shifted GIG	`PSGIG`	$\{0, 1, 2, \ldots\}$	log	log	logit	logit

Table 18.1: Discrete count distributions implemented within **gamlss.dist**, with default link functions.

18.1 Count data one parameter distributions

18.1.1 Geometric distribution.

First parametrization GEOM(μ)**.**

There are two parametrizations of the geometric distribution in the **gamlss.dist** package: GEOM(μ) and GEOMo(μ).

The probability function (pf) of the geometric distribution, GEOM(μ), is given by

$$P(Y = y|\mu) = \frac{\mu^y}{(\mu+1)^{y+1}} \tag{18.1}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$. Figure 18.1 shows the shapes of the geometric, GEOM(μ), distribution for different values of $\mu = 1, 2, 5$. Table 18.2 gives the mean, variance, skewness γ_1 and excess kurtosis γ_2 as defined in Section 2.2.

The geometric distribution has an explicit cumulative distribution function (cdf) given by $F_Y(y|\mu) = P(Y \leq y|\mu) = 1 - \left(\frac{\mu}{\mu+1}\right)^{y+1}$, for $y = 0, 1, 2, 3, \ldots$, and an explicit inverse cdf (or quantile or centile function) y_p given in Table 18.2 where y_p is defined for a discrete distribution in Section **??**. The mode of Y is always at 0, while the mean of Y is μ. The median of Y, $m = y_{0.5}$ is also given in Table 18.2. The probability generating function (PGF) of $Y \sim$ GEOM(μ) is given by $G_Y(t) = [1 + \mu(1-t)]^{-1}$ for $|t| < (1+\mu)\mu^{-1}$. For large values of y (and all y), $P(Y = y|\mu) = q \exp\{y \log[\mu/(\mu+1)]\}$ where $q = (\mu+1)^{-1}$, i.e. an exponential right tail.

Second parametrization GEOMo(μ)**.**

For the original parametrization of the geometric distribution, GEOMo(μ), replace μ in (18.1) by $(1-\mu)/\mu$ giving

$$P(Y = y|\mu) = (1-\mu)^y \mu$$

for for $y = 0, 1, 2, 3, \ldots$, and $0 < \mu < 1$. Hence μ in this case is the probability of observing $Y = 0$. Other characteristics of the GEOMo(μ) distribution are given in Table 18.3.

```
disc1("GEOM", mu=c(1,2,5), miny=1, maxy=20)
```

Figure 18.1

Table 18.2: Geometric distribution

GEOM(μ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$, mean
Distribution measures	
mean	μ
median	$\begin{cases} \lfloor \alpha \rfloor, & \text{if } \alpha \text{ is not an integer} \\ \alpha - 1, & \text{if } \alpha \text{ is an integer} \end{cases}$ where $\alpha = \frac{-\log 2}{\log\left[\mu(\mu+1)^{-1}\right]}$
mode	0
variance	$\mu + \mu^2$
skewness	$(1+2\mu)\left(\mu+\mu^2\right)^{-0.5}$
excess kurtosis	$6 + (\mu + \mu^2)^{-1}$
PGF	$[1+\mu(1-t)]^{-1}$ for $\lvert t \rvert < (1+\mu)\mu^{-1}$
pf	$\frac{\mu^y}{(\mu+1)^{y+1}}$
cdf	$1 - \left(\frac{\mu}{\mu+1}\right)^{y+1}$
Inverse cdf (y_p)	$\begin{cases} \lfloor \alpha_p \rfloor, & \text{if } \alpha_p \text{ is not an integer} \\ \alpha_p - 1, & \text{if } \alpha_p \text{ is an integer} \end{cases}$ where $\alpha_p = \frac{\log(1-p)}{\log\left[\mu(\mu+1)^{-1}\right]}$
Reference	set $\sigma = 1$ in NBI(μ, σ)
Note	$\lfloor \alpha \rfloor$ is the largest integer less than or equal to α, i.e. the floor function.

Table 18.3: Geometric distribution (original)

GEOMo(μ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < 1$
Distribution measures	
mean	$(1-\mu)\mu^{-1}$
median	$\begin{cases} \lfloor \alpha \rfloor, & \text{if } \alpha \text{ is not an integer} \\ \alpha - 1, & \text{if } \alpha \text{ is an integer} \end{cases}$ where $\alpha = \frac{-\log 2}{\log(1-\mu)}$
mode	0
variance	$(1-\mu)\,\mu^{-2}$
skewness	$(2-\mu)\,(1-\mu)^{-0.5}$
excess kurtosis	$6 + \mu^2\,(1-\mu)^{-1}$
PGF	$\mu\,[1-(1-\mu)t]^{-1}$ for $\lvert t \rvert < (1-\mu)^{-1}$
pf	$\mu(1-\mu)^y$
cdf	$1-(1-\mu)^{y+1}$
Inverse cdf (y_p)	$\begin{cases} \lfloor \alpha_p \rfloor, & \text{if } \alpha_p \text{ is not an integer} \\ \alpha_p - 1, & \text{if } \alpha_p \text{ is an integer} \end{cases}$ where $\alpha_p = \frac{\log(1-p)}{\log(1-\mu)}$
Reference	Reparametrize μ to $(1-\mu)\mu^{-1}$ in GEOM(μ)

R code on page 305

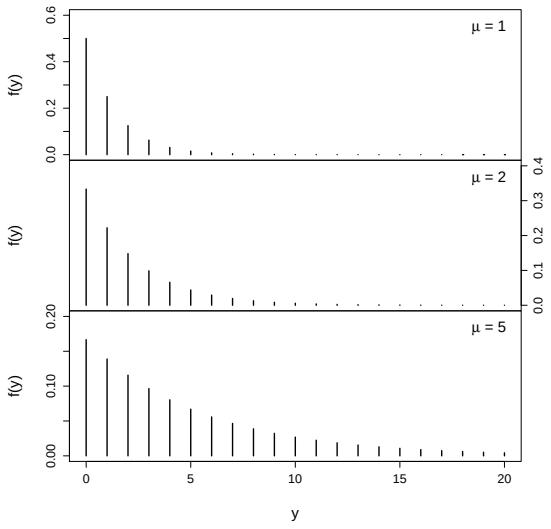


Figure 18.1: The geometric, GEOM(μ), distribution with $\mu = 1, 2, 5$.

18.1.2 Logarithmic distribution, LG(μ).

The probability function of the logarithmic distribution, denoted by LG(μ), is given by

$$P(Y = y|\mu) = \frac{\alpha \mu^y}{y} \tag{18.2}$$

for $y = 1, 2, 3, \ldots$, where $\alpha = -\left[\log(1-\mu)\right]^{-1}$ for $0 < \mu < 1$. Note that the range of Y starts from 1. The mean and variance and other properties of Y are given in Table 18.4. For large y (and all y), $P(Y = y|\mu) =\propto \exp\{y \log \mu - \log y\}$. Hence $\log P(Y = y|\mu) \sim y \log \mu$. for large y.

[When a response variable Y has range $R_Y = \{0, 1, 2, \ldots\}$ a possible model distribution is the zero adjusted logarithmic distribution ZALG(μ, σ) discussed in Section 18.2.8.]

Plots of the logarithmic, LG(μ), distribution for different values of the μ parameter are given in Figure 18.2.

Table 18.4: Logarithmic distribution

LG(μ)	
Ranges	
Y	$1, 2, 3 \dots$
μ	$0 < \mu < 1$
Distribution measures	
mean[a]	$\alpha\mu(1-\mu)^{-1}$ where $\alpha = -[\log(1-\mu)]^{-1}$
mode[a]	1
variance[a]	$\alpha\mu(1-\alpha\mu)(1-\mu)^{-2}$
skewness[a]	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = \alpha\mu(1+\mu-3\alpha\mu+2\alpha^2\mu^2)(1-\mu)^{-3}$
excess kurtosis	$\frac{k_4}{[Var(Y)]^2}$ where $k_4 = \alpha\mu[1+4\mu+\mu^2-\alpha\mu(7+4\mu)+12\alpha^2\mu^2-6\alpha^3\mu^3](1-\mu)^{-4}$
PGF[a]	$\frac{\log(1-\mu t)}{\log(1-\mu)}$
pf [a]	$\frac{\alpha\mu^y}{y}$
Reference	[a] Johnson et al. [2005], section 7.1, p 302-307, parametrized by $\theta = \mu$

```
disc1("LG", mu=c(.1,.5,.9), miny=1, maxy=20)
```

Figure 18.2

R code on page 309

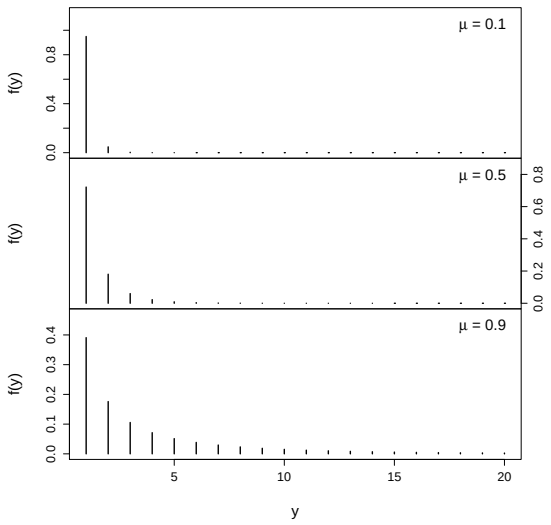


Figure 18.2: The logarithmic, $\texttt{LG}(\mu)$, distribution with $\mu = .1, .5, .9$.

18.1.3 Poisson distribution, $\texttt{PO}(\mu)$.

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

$$P(Y = y|\mu) \quad = \quad \frac{e^{-\mu}\mu^y}{y!} \tag{18.3}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$.

The mean, variance, skewness and kurtosis of Y are given by $E(Y) = \mu$, $Var(Y) = \mu$, $\sqrt{\beta}_1 = \mu^{-0.5}$ and $\beta_2 = 3 + \mu^{-1}$ respectively.

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 $E[Y] = Var[Y]$ and that $\beta_2 - \beta_1 - 3 = 0$. The coefficient of variation of the distribution is given by $\mu^{-0.5}$. The index of dispersion, that is, the ratio $Var[Y]/E[Y]$ is equal to one for the Poisson distribution. For distributions with $Var[Y] > E[Y]$ we have overdispersion relative to the Poisson distribution and for $Var[Y] < E[Y]$ we have underdispersion. The distribution is positively skew for small values of μ, but almost symmetric for large μ values.

For large y, $P(Y = y) \sim q \exp\left[y(\log \mu + 1) - (y + 0.5) \log y\right]$ where $q = e^{-\mu}\sqrt{2\pi}$,

(using Stirling's approximation to $y!$), which decreases faster than an exponential $\exp(-y)$. Note that $\log P(Y = y) \sim -y \log y$ for large y.

Table 18.5: Poisson distribution

PO(μ)		
Ranges		
Y	$0, 1, 2, 3 \ldots$	
μ	$0 < \mu < \infty$, mean	
Distribution measures		
mean	μ	
mode	$\begin{cases} \lfloor \mu \rfloor, & \text{if } \mu \text{ is not an integer} \\ \mu - 1 \text{ and } \mu, & \text{if } \mu \text{ is an integer} \end{cases}$	
variance	μ	
skewness	$\mu^{-0.5}$	
excess kurtosis	μ^{-1}	
PGF	$e^{\mu(t-1)}$	
pf	$\frac{e^{-\mu}\mu^y}{y!}$	
cdf	$\frac{\Gamma(y+1,\mu)}{\Gamma(y)}$	
Reference	Johnson et al. [2005] sections 4.1, 4.3, 4.4, p 156, p 161-165, p307	
Note	$\lfloor \alpha \rfloor$ is the largest integer less than or equal to α $\Gamma(\alpha, x) = \int_x^\infty t^{\alpha-1} e^{-t} dt$ is the complement of the incomplete gamma function	

```
disc1("PO", mu=c(1,5,10), miny=1, maxy=22)
```

Figure 18.3

R code on page 311

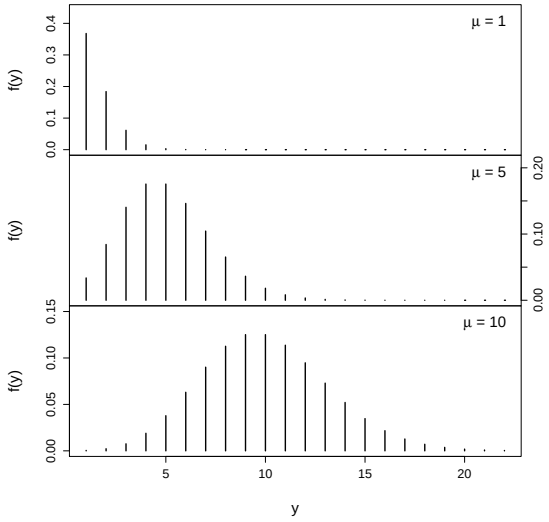


Figure 18.3: The Poisson, PO(μ), distribution with $\mu = 1, 5, 10$.

18.1.4 Yule distribution, YULE(μ).

The probability function of the Yule distribution, denoted by YULE(μ), is given by

$$P(Y = y|\mu) = (\mu^{-1} + 1)B(y + 1, \mu^{-1} + 2) \tag{18.4}$$

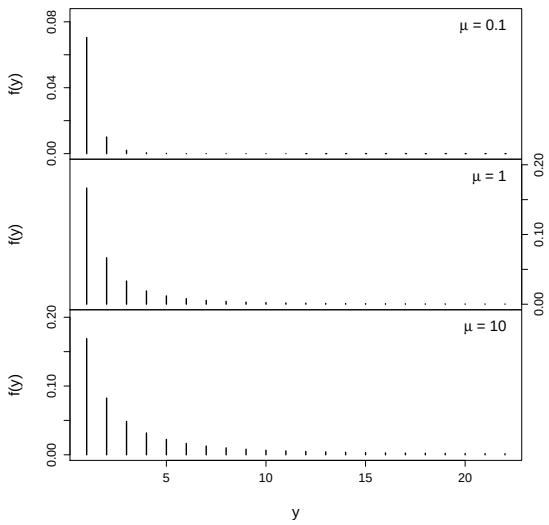
for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$. Note that the parametrization of the YULE(μ) distribution only includes Yule distribution with a finite mean μ. The YULE(μ) distribution is a special case of the WARING(μ, σ) where $\mu = \sigma$. For large y, $P(Y = y|\mu) \sim qy^{-(\mu^{-1}+2)}$ where $q = (\mu^{-1} + 1)\Gamma(\mu^{-1} + 2)$, [using equation (1.32) of Johnson et al. [2005], p 8], and hence the YULE(μ) distribution has a heavy tail, especially for large μ .

Figure 18.4

```
disc1("YULE", mu=c(.1,1,10), miny=1, maxy=22)
```

Table 18.6: Yule distribution

YULE(μ)	
Ranges	
Y	$0, 1, 2, 3 \dots$
μ	$0 < \mu < \infty$, mean
Distribution measures	
mean	μ
mode	0
variance	$\begin{cases} \mu(\mu+1)^2(1-\mu)^{-1}, & \text{if } \mu < 1 \\ \infty, & \text{if } \mu \geq 1 \end{cases}$
skewness	$\begin{cases} (2\mu+1)^2(1-\mu)^{0.5}(\mu+1)^{-1}(1-2\mu)^{-1}\mu^{-0.5}, & \text{if } \mu < 1/2 \\ \infty, & \text{if } \mu \geq 1/2 \end{cases}$
excess kurtosis	$\begin{cases} \frac{\left(1+11\mu+18\mu^2-6\mu^3-36\mu^4\right)}{\mu(\mu+1)(1-2\mu)(1-3\mu)} & \text{if } \mu < 1/3 \\ \infty, & \text{if } \mu \geq 1/3 \end{cases}$
PGF	$(\mu+1)(2\mu+1)^{-1}{}_2F_1(1, 1, 3+\mu^{-1}; t)$
pf	$(\mu^{-1}+1)B(y+1, \mu^{-1}+2)$
cdf	$1-(y+1)B(y+1, 2+\mu^{-1})$
Reference	Let $\sigma = \mu$ in WARING(μ, σ)



R code on page 312

Figure 18.4: The Yule, YULE(μ), distribution with $\mu = 0.1, 1, 10$.

18.1.5 Zipf distribution, ZIPF(μ).

The probability function of the zipf distribution, denoted by ZIPF(μ), is given by

$$P(Y = y|\mu) = \frac{y^{-(\mu+1)}}{\zeta(\mu+1)} \tag{18.5}$$

for $y = 1, 2, 3, \ldots$, where $\mu > 0$ and $\zeta(b) = \Sigma_{i=1}^{\infty} i^{-b}$ is the (Reimann) zeta function. Note that the range of Y starts from 1.

This distribution is also known as the (Riemann) zeta distribution or the discrete Pareto distribution. For large y (and all y), $P(Y = y|\mu) = qy^{-(\mu+1)}$ where $q = [\zeta(\mu+1)]^{-1}$, so the ZIPF(μ) distribution has a very heavy tail especially for μ close to 0. It is suitable for very heavy tail count data, as can be seen from Figure 18.5.

$E(Y^r) = \zeta(\mu - r + 1)/\zeta(\mu + 1)$ provided that $\mu > r$, Johnson et al. [2005] p 528. Hence this gives the mean, variance, skewness and excess kurtosis of Y in Table 18.7, using Section 2.2, equation (2.1). The mean $E(Y)$ increases as μ decreases and it is infinity if $\mu \leq 1$.

Figure 18.5

```
disc1("ZIPF", mu=c(.1,1,2), miny=1, maxy=20)
```

Table 18.7: ZIPF distribution

ZIPF(μ)	
Ranges	
Y	$1, 2, 3 \ldots$
μ	$0 < \mu < \infty$
Distribution measures	
mean	$\begin{cases} \zeta(\mu)/\zeta(\mu+1), & \text{if } \mu > 1 \\ \infty, & \text{if } \mu \leq 1 \end{cases}$
mode	1
variance	$\begin{cases} \left\{\zeta(\mu+1)\zeta(\mu-1) - [\zeta(\mu)]^2\right\}/[\zeta(\mu+1)]^2, & \text{if } \mu > 2 \\ \infty, & \text{if } \mu \leq 2 \end{cases}$
skewness	$\begin{cases} \mu_3/[Var(Y)]^{1.5} \quad \text{where} \\ \mu_3 = \{[\zeta(\mu+1)]^2\zeta(\mu-2) - 3\zeta(\mu+1)\zeta(\mu)\zeta(\mu-1) + \\ \qquad 2[\zeta(\mu)]^3\}/[\zeta(\mu+1)]^3, & \text{if } \mu > 3 \\ \infty, & \text{if } \mu \leq 3 \end{cases}$
excess kurtosis	$\begin{cases} \{\mu_4/[Var(Y)]^2\} - 3 \quad \text{where} \\ \mu_4 = \{[\zeta(\mu+1)]^3\zeta(\mu-3) - 4[\zeta(\mu+1)]^2\zeta(\mu)\zeta(\mu-2) + \\ \qquad 6\zeta(\mu+1)[\zeta(\mu)]^2\zeta(\mu-1) - 3[\zeta(\mu)]^4\}/[\zeta(\mu+1)]^4, & \text{if } \mu > 4 \\ \infty, & \text{if } \mu \leq 4 \end{cases}$
PGF[a]	$t\Phi(t, \mu+1, 1)/\Phi(1, \mu+1, 1)$
pf [a]	$[y^{(\mu+1)}\zeta(\mu+1)]^{-1}$
Reference	[a]Johnson et al. [2005] sections 11.2.20, p 527-528, where $\rho = \mu$
Notes	$\zeta(b) = \sum_{i=1}^{\infty} i^{-b}$ is the Riemann zeta function $\Phi(a, b, c) = \sum_{i=0}^{\infty} \frac{a^i}{(i+c)^b}$, for $c \neq 0, -1, -2$, is the Lerch function

R code on page 314

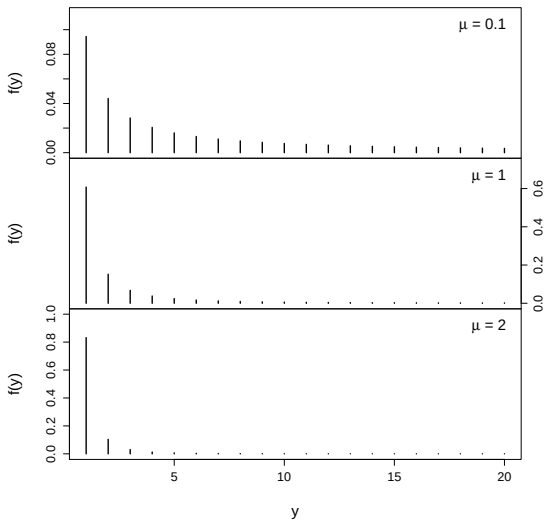


Figure 18.5: The ZIPF(μ) distribution with $\mu = 0.1, 1, 2$.

18.2 Count data two parameters distributions.

18.2.1 Double Poisson DPO(μ, σ).

The double Poisson distribution, denoted by DPO(μ, σ), has probability function giving by:

$$P\left(Y = y|\mu, \sigma\right) = c(\mu, \sigma)\sigma^{-1/2}e^{-\mu/\sigma}\left(\frac{\mu}{y}\right)^{y/\sigma}\frac{e^{y/\sigma - y}y^{y}}{y!} \tag{18.6}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$ and $\sigma > 0$, where $c(\mu, \sigma)$ is a 'normalizing constant' (ensuring that the distribution probabilities sum to one) given by

$$c(\mu, \sigma) = \left[\sum_{y=0}^{\infty}\sigma^{-1/2}e^{-\mu/\sigma}\left(\frac{\mu}{y}\right)^{y/\sigma}\frac{e^{y/\sigma - y}y^{y}}{y!}\right]^{-1}, \tag{18.7}$$

Put the proper reference in the bibliography

[obtained from Lindsey (1995), p 131, reparametrized by $\nu = \mu$ and $\psi = 1/\sigma$].

The double Poisson distribution, DPO(μ, σ), is a special case of the double exponential family of Efron [1986]. The DPO(μ, σ) distribution has approximate mean μ and approximate variance $\sigma\mu$. The DPO(μ, σ) distribution is a PO(μ) distribution if $\sigma = 1$. It is an overdispersed Poisson distribution if $\sigma > 1$

and it is underdispersed Poisson if $\sigma < 1$. Unlike some other implementations **gamlss.dist** approximates $c(\mu, \sigma)$ using a finite sum [with a very large number of terms, $3 \times \max(y)$] rather than a potentially less accurate functional approximation of $c(\mu, \sigma)$. For large y, $\log P(Y = y \,|\, \mu, \sigma) \sim -y \log(y)/\sigma$.

Table 18.8: Double Poisson distribution

DPO(μ, σ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$
σ	$0 < \sigma < \infty$
Distribution measures	
pf [a]	$c(\mu, \sigma)\sigma^{-1/2} e^{-\mu/\sigma} \left(\frac{\mu}{y}\right)^{y/\sigma} \frac{e^{y/\sigma - y} y^y}{y!}$
	where
	$c(\mu, \sigma) = \left[\sum_{y=0}^{\infty} \sigma^{-1/2} e^{-\mu/\sigma} \left(\frac{\mu}{y}\right)^{y/\sigma} \frac{e^{y/\sigma - y} y^y}{y!}\right]^{-1}$
Reference	[a] Lindsey [1995] p 131, reparametrized by $\nu = \mu$ and $\psi = 1/\sigma$

```
disc2R("DPO", mu=c(1,5), sigma=c(.5,1,2), miny=0, maxy=20)
```

Figure 18.6

18.2.2 Generalized Poisson GPO(μ, σ).

The probability function of the generalized Poisson distribution, Consul and Jain [1973], Consul [1989], Poortema [1999] and Johnson et al. [2005] page 336-339, is given by :

$$P(Y = y|\theta, \lambda) = \frac{\theta(\theta + \lambda y)^{y-1} \, e^{-(\theta+\lambda y)}}{y!} \tag{18.8}$$

for $y = 0, 1, 2, ...$ where $\theta > 0$ and $0 \leq \lambda \leq 1$. [Note λ is greater than 0 to ensure that all probabilites $P(Y = y|\theta, \lambda)$ are positive for $y = 0, 1, \ldots.$] The generalized Poisson distribution was derived by Consul and Jain [1973] as an approximation of a generalized negative binomial distribution. Consul [e.g. Consul [1989]] has extensively studied the generalized Poisson distribution, therefore sometimes it is called *Consul's generalized Poisson distribution.*

R code on page 317

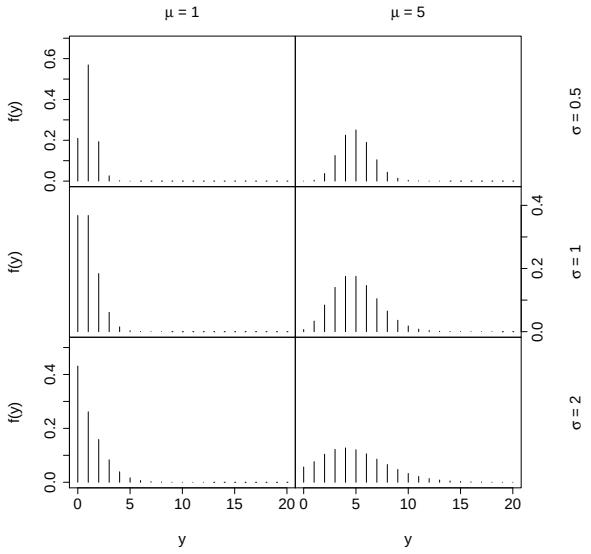


Figure 18.6: The double Poisson, DPO(μ, σ), distribution with $\mu = 1, 5$ and $\sigma = .5, 1, 2$.

The mean and the variance of the generalized Poisson distribution are given by

$$\mathrm{E}(Y) = \frac{\theta}{1-\lambda}$$

and

$$\mathrm{Var}(Y) = \frac{\theta}{(1-\lambda)^3}$$

(see Johnson *et al.* (1993), pages 337-339).

A better parametrization of the generalized Poisson distribution for regression type of models is given by reparameterizing (18.8) by setting $\mu = \theta/(1-\lambda)$ and $\sigma = \lambda/\theta$ [and hence $\theta = \mu/(1+\sigma\mu)$ and $\lambda = \sigma\mu/(1+\sigma\mu)$], giving the probability function for $\texttt{GPO}(\mu, \sigma)$:

$$P(Y = y|\mu, \sigma) = \left(\frac{\mu}{1+\sigma\mu}\right)^y \frac{(1+\sigma y)^{y-1}}{y!} \exp\left[\frac{-\mu\,(1+\sigma y)}{1+\sigma\mu}\right] \qquad (18.9)$$

for $y = 0, 1, 2, ...$, where $\mu > 0$ and $\sigma > 0$. For large y, $\log P(Y = y \,|\, \mu, \sigma) \sim -y \log y$, as for a $\texttt{PO}(\mu)$ distribution.

Figure 18.7 shows the generalized Poisson, $\texttt{GPO}(\mu, \sigma)$, distribution for $\mu = 1, 5$ and $\sigma = .01, .5, 1$.

Table 18.9: Generalised Poisson distribution

$\texttt{GPO}(\mu, \sigma)$	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$, mean
σ	$0 < \sigma < \infty$, dispersion parameter
Distribution measures	
mean	μ
variance	$\mu\,(1+\sigma\mu)^2$
skewness	$(1+3\sigma\mu)\,/\mu^{0.5}$
excess kurtosis	$\left(1+10\sigma\mu+15\sigma^2\mu^2\right)/\mu$
pf	$\left(\frac{\mu}{1+\sigma\mu}\right)^y \frac{(1+\sigma y)^{y-1}}{y!} \exp\left[\frac{-\mu(1+\sigma y)}{1+\sigma\mu}\right]$
Reference	Johnson et al. [2005] p 336-339 reparameterized by $\theta = \mu/(1+\sigma\mu)$ and $\lambda = \sigma\mu/(1+\sigma\mu)$ and hence $\mu = \theta/(1-\lambda)$ and $\sigma = \lambda/\theta$

Figure 18.7
```
disc2R("GPO", mu=c(1,5), sigma=c(.01,.5,1), miny=0, maxy=20)
```

R code on page 320

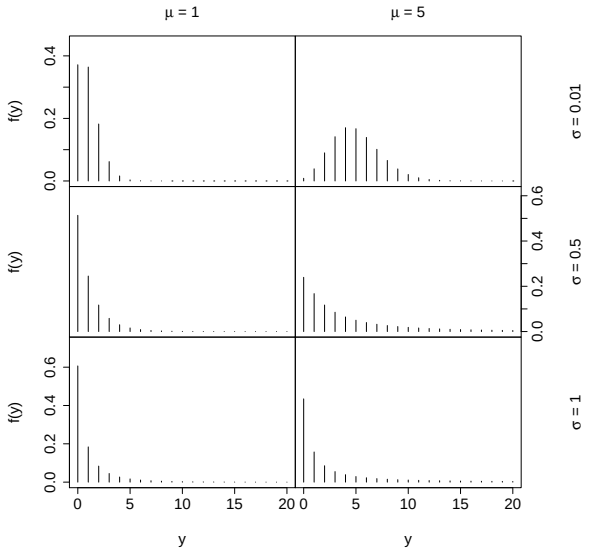


Figure 18.7: The generalised Poisson, $\texttt{GPO}(\mu, \sigma)$, distribution with $\mu = 1, 5$ and $\sigma = .01, .5, 1$.

18.2.3 Negative binomial distribution, $\texttt{NBI}(\mu, \sigma)$, $\texttt{NBII}(\mu, \sigma)$.

Negative binomial type I, $\texttt{NBI}(\mu, \sigma)$.

The probability function of the negative binomial distribution type I, denoted by $\texttt{NBI}(\mu, \sigma)$, 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 $y = 0, 1, 2, 3 \ldots$, where $\mu > 0$, $\sigma > 0$.

The above parametrization is equivalent to that used by Anscombe [1950] except he used $\alpha = 1/\sigma$, as pointed out by Johnson et al. [2005] p 209. The Poisson, $\texttt{PO}(\mu)$, distribution is a limiting distribution of $\texttt{NBI}(\mu, \sigma)$ as $\sigma \to 0$.

[Note that the original parametrization of the negative binomial distribution,

$\texttt{NBo}(k,\pi)$, is given by setting $\mu = k(1-\pi)/\pi$ and $\sigma = 1/k$ giving

$$P(Y=y|\pi) = \frac{\Gamma(y+k)}{\Gamma(y+1)\Gamma(k)}(1-\pi)^y\pi^k$$

for $y = 0, 1, 2, 3\ldots$, where $k > 0$ and $0 < \pi < 1$, Johnson *et al.* (2005), p 209. Hence $\pi = 1/(1+\mu\sigma)$ and $k = 1/\sigma$. The $\texttt{NBo}(k,\pi)$ distribution is not available in the package **gamlss.dist**].

For large y, $P(Y = y|\mu,\sigma) \sim q\exp\left[-y\log\left(1+\frac{1}{\mu\sigma}\right)+\left(\frac{1}{\sigma}-1\right)\log y\right]$ where $q = \left[\Gamma\left(\frac{1}{\sigma}\right)(1+\sigma\mu)^{1/\sigma}\right]^{-1}$, essentially an exponential tail. Note $\log P(Y = y|\mu,\sigma) \sim -y\log\left(1+\frac{1}{\mu\sigma}\right)$. See $\texttt{NBII}(\mu,\sigma)$ below with variance $Var(Y) = \mu+\sigma\mu$ for an alternative parametrization of the $\texttt{NBI}(\mu,\sigma)$. Also see $\texttt{NBF}(\mu,\sigma,\nu)$ in Section 18.3.3 with variance $Var(Y) = \mu+\sigma\mu^\nu$ for a family of reparametrizations of the $\texttt{NBI}(\mu,\sigma)$.

Table 18.10: Negative binomial Type I distribution

$\texttt{NBI}(\mu,\sigma)$	
Ranges	
Y	$0, 1, 2, 3\ldots$
μ	$0 < \mu < \infty$, mean
σ	$0 < \sigma < \infty$, dispersion
Distribution measures	
mean	μ
mode	$\begin{cases}\lfloor(1-\sigma)\mu\rfloor, & \text{if } (1-\sigma)\mu \text{ is not an integer and } \sigma<1\\ (1-\sigma)\mu-1 \text{ and } (1-\sigma)\mu, & \text{if } (1-\sigma)\mu \text{ is an integer and } \sigma<1\\ 0, & \text{if } \sigma\geq 1\end{cases}$
variance	$\mu+\sigma\mu^2$
skewness	$(1+2\mu\sigma)(\mu+\sigma\mu^2)^{-0.5}$
excess kurtosis	$6\sigma+(\mu+\sigma\mu^2)^{-1}$
PGF	$[1+\mu\sigma(1-t)]^{-1/\sigma}$
pf	$\frac{\Gamma(y+\sigma^{-1})}{\Gamma(\sigma^{-1})\Gamma(y+1)}\left(\frac{\sigma\mu}{1+\sigma\mu}\right)^y\left(\frac{1}{1+\sigma\mu}\right)^{1/\sigma}$
cdf	$1-\frac{B(y+1,\sigma^{-1},\mu\sigma(1+\mu\sigma)^{-1})}{B(y+1,\sigma^{-1})}$
Reference	Johnson et al. [2005], sections 5.1 to 5.5, p 209-217, reparameterized by $p = 1/(1+\mu\sigma)$ and $k = 1/\sigma$ and hence $\mu = k(1-p)/p$ and $\sigma = 1/k$
Note	$\lfloor\alpha\rfloor$ is the largest integer less than or equal to α $B(\alpha,\beta,x) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,dt$ is the incomplete beta function

Figure 18.8

```
disc2("NBI", mu=c(1,2,5), sigma=c(.1, 2), miny=0, maxy=20)
```

R code on page 322

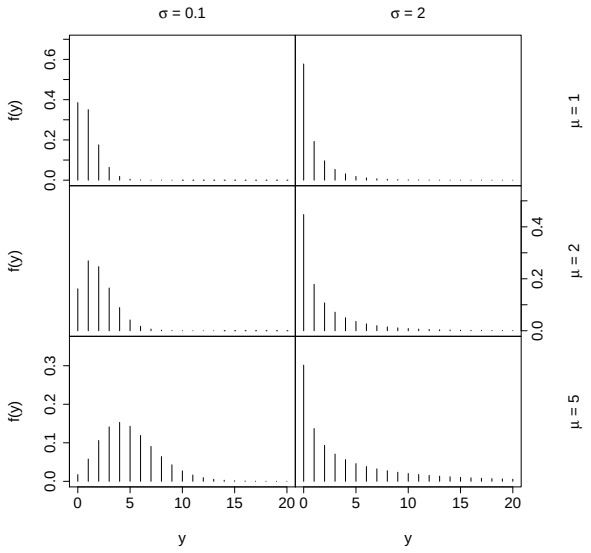


Figure 18.8: The negative binomial type I, NBI(μ, σ), distribution with $\mu = 1, 2, 5$ and $\sigma = .1, 2$.

Negative binomial type II, NBII(μ, σ)**.**

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

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

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$ and $\sigma > 0$.

This parametrization was used by Evans [1953] as pointed out by Johnson *et al* (2005) p 209, and is obtained by re-parametrizing σ to σ/μ in NBI(μ, σ).

Figure 18.9

Table 18.11: Negative binomial Type II distribution

NBII(μ,σ)	
Ranges	
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$, mean
σ	$0<\sigma<\infty$, dispersion
Distribution measures	
mean	μ
mode	$\begin{cases} \lfloor \mu-\sigma \rfloor, & \text{if } (\mu-\sigma) \text{ is not an integer and } \sigma<\mu \\ (\mu-\sigma-1) \text{ and } (\mu-\sigma), & \text{if } (\mu-\sigma) \text{ is an integer and } \sigma<\mu \\ 0, & \text{if } \sigma \geq \mu \end{cases}$
variance	$\mu+\sigma\mu$
skewness	$(1+2\sigma)(\mu+\sigma\mu)^{-0.5}$
excess kurtosis	$6\sigma\mu^{-1}+(\mu+\sigma\mu)^{-1}$
PGF	$[1+\sigma(1-t)]^{-\mu/\sigma}$
pf	$\frac{\Gamma(y+\mu\sigma^{-1})}{\Gamma(\mu\sigma^{-1})\Gamma(y+1)}\left(\frac{\sigma}{1+\sigma}\right)^{y}\left(\frac{1}{1+\sigma}\right)^{\mu/\sigma}$
cdf	$1-\frac{B(y+1,\mu\sigma^{-1},\sigma(1+\sigma)^{-1})}{B(y+1,\mu\sigma^{-1})}$
Reference	reparameterize σ to σ/μ in NBI(μ,σ)

```
disc2("NBII", mu=c(1,2,5), sigma=c(.1, 2), miny=0, maxy=20)
```

R code on page 322

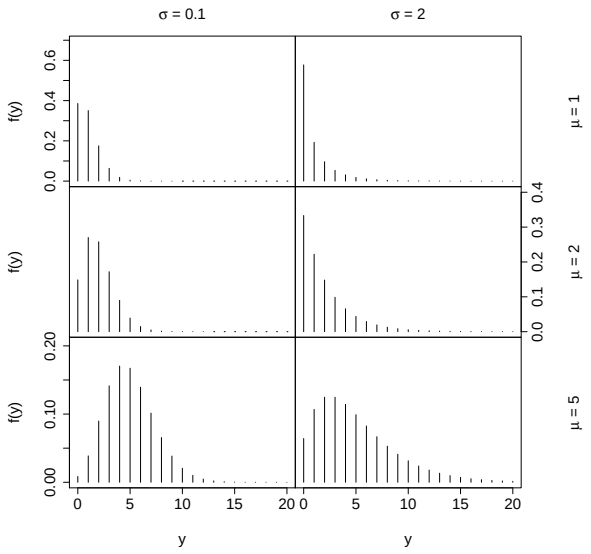


Figure 18.9: The negative binomial type II, `NBII`(μ, σ), distribution with $\mu = 1, 2, 5$ and $\sigma = .1, 2$.

18.2.4 Poisson-inverse Gaussian distribution, PIG(μ, σ).

18.2.5 First parametrization, PIG(μ, σ).

The probability function of the Poisson-inverse Gaussian distribution, denoted by PIG(μ, σ), is given by

$$P(Y = y|\mu, \sigma) = \left(\frac{2\alpha}{\pi}\right)^{1/2} \frac{\mu^y e^{1/\sigma} K_{y-\frac{1}{2}}(\alpha)}{y!(\alpha\sigma)^y} \tag{18.10}$$

for $y = 0, 1, 2, 3, \ldots$, where $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$, where $\mu > 0$ and $\sigma > 0$ and $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\{-\frac{1}{2}t(x + x^{-1})\}dx$ is the modified Bessel function of the third kind. Note that $K_{-1/2}(t) = K_{1/2}(t) = [\pi/(2t)]^{1/2}e^{-t}$ and $K_{\lambda+1}(t) = (2\lambda/t)K_\lambda(t) + K_{\lambda-1}(t)$. Note also that $\sigma = \left[\left(\mu^2 + \alpha^2\right)^{0.5} - \mu\right]^{-1}$.

Note that the above parametrization was used by Dean, Lawless and Willmot (1989). It is also a special case of the gamlss.family distributions SI(μ, σ, ν) and SICHEL(μ, σ, ν) when $\nu = -1/2$. The Poisson, PO(μ), distribution is a limiting distribution of PIG(μ, σ) as $\sigma \to 0$.

For large y, $P(Y = y|\mu, \sigma) \sim r \exp\left[-y \log\left(1 + \frac{1}{2\mu\sigma}\right) - \frac{3}{2}\log y\right]$, where r does not depend on y, i.e. essentially an exponential tail. Note that $\log P(Y = y|\mu, \sigma) \sim -y \log\left(1 + \frac{1}{2\mu\sigma}\right)$ for large y.

Table 18.12: Poisson inverse Gaussian distribution

PIG(μ, σ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$, mean
σ	$0 < \sigma < \infty$, dispersion
Distribution measures	
mean	μ
variance	$\mu + \sigma\mu^2$
skewness	$(1 + 3\mu\sigma + 3\mu^2\sigma^2)(1 + \mu\sigma)^{-1.5}\mu^{-0.5}$
excess kurtosis	$(1 + 7\mu\sigma + 18\mu^2\sigma^2 + 15\mu^3\sigma^3)(1 + \mu\sigma)^{-2}\mu^{-1}$
PGF[a]	$e^{1/\sigma - q}$ where $q^2 = \sigma^{-2} + 2\mu(1 - t)\sigma^{-1}$
pf [a]	$\left(\frac{2\alpha}{\pi}\right)^{1/2} \frac{\mu^y e^{1/\sigma} K_{y-\frac{1}{2}}(\alpha)}{y!(\alpha\sigma)^y}$ where $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$
Reference	Set $\nu = -\frac{1}{2}$ (and hence $c = 1$) in SICHEL(μ, σ, ν), see also [a] Dean et al. [1989]

Figure 18.10

```
disc2("PIG", mu=c(1,2,5), sigma=c(.1,  2), miny=0, maxy=20)
```

R code on page 326

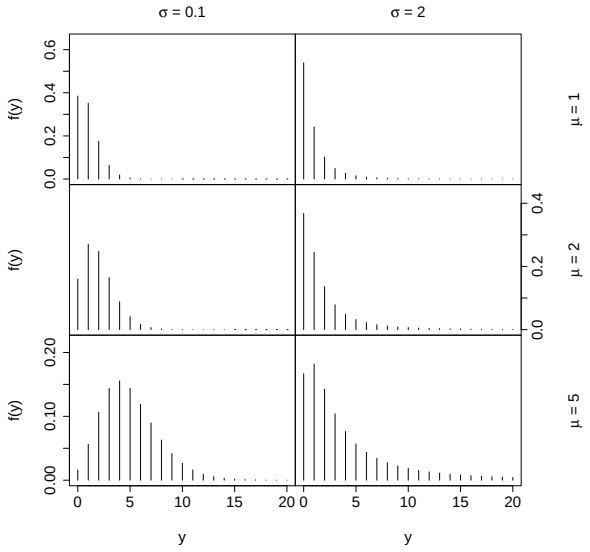


Figure 18.10: The Poisson inverse Gaussian, PIG(μ, σ), distribution with $\mu = 1, 2, 5$ and $\sigma = .1, 2$.

,

18.2.6 Second parametrization, PIG2(μ, σ).

To be added soon

18.2.7 Waring distribution, WARING(μ, σ).

The probability function of the Waring distribution, denoted by WARING(μ, σ), is given by

$$P(Y = y|\mu, \sigma) = \frac{B(y + \mu\sigma^{-1}, \sigma^{-1} + 2)}{B(\mu\sigma^{-1}, \sigma^{-1} + 1)} \tag{18.11}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$ and $\sigma > 0$. Note that the parametrisation of WARING(μ, σ) only includes Waring distributions with a finite mean μ.

The Waring distribution is also called the beta geometric distribution.

The $\texttt{WARING}(\mu, \sigma)$ distribution is a reparametrization of the distribution given in Wimmer and Altmann [1999] p 643, where $b = \sigma^{-1} + 1$ and $n = \mu\sigma^{-1}$. Hence $\mu = n(b-1)^{-1}$ and $\sigma = (b-1)^{-1}$. It is also a reparameterisation of the distribution given in equation 16.131 in Johnson et al. [2005] p 290, where $\alpha = \mu\sigma^{-1}$ and $c = (\mu+1)\sigma^{-1} + 1$.

The $\texttt{WARING}(\mu, \sigma)$ distribution is a special case of the beta negative binomial $BNB(\mu, \sigma, \nu)$ distribution where $\nu = 1$. It can can be derived as a beta mixture of geometric distributions, by assuming $Y|\pi \sim \texttt{GEOMo}(\pi)$, where $\pi \sim BEo(b, n)$ where $b = \sigma^{-1} + 1$ and $n = \mu\sigma^{-1}$. Hence the $\texttt{WARING}(\mu, \sigma)$ distribution can be considered as an overdispersed geometric distribution.

For large y,

$$P(Y = y, \mu, \sigma) \sim q y^{-(\sigma^{-1}+2)}$$

where

$$q = \left(\sigma^{-1} + 1\right) \Gamma\left(\left[\mu + 1\right]\sigma^{-1} + 1\right) / \Gamma\left(\mu\sigma^{-1}\right)$$

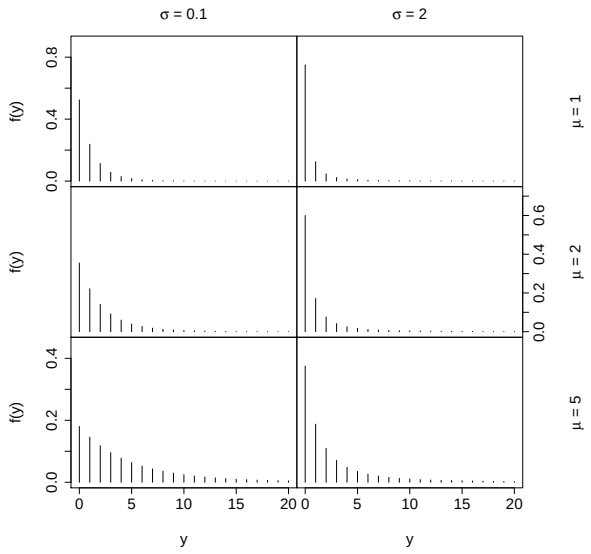
and hence the $\texttt{WARING}(\mu, \sigma)$ has a heavy right tail, especially for large σ.

```
disc2("WARING", mu=c(1,2,5), sigma=c(.1, 2), miny=0, maxy=20)
```

Figure 18.11

Table 18.13: Waring distribution

WARING(μ, σ)	
Ranges	
Y	$0, 1, 2, 3 \dots$
μ	$0 < \mu < \infty$, mean
σ	$0 < \sigma < \infty$
Distribution measures	
mean	μ
mode	0
variance	$\begin{cases} \mu(\mu+1)(1+\sigma)(1-\sigma)^{-1}, & \text{if } \sigma < 1 \\ \infty, & \text{if } \sigma \geq 1 \end{cases}$
skewness	$\begin{cases} \frac{(2\mu+1)(1+2\sigma)(1-\sigma)^{1/2}}{\mu^{1/2}(\mu+1)^{1/2}(1+\sigma)^{1/2}(1-2\sigma)}, & \text{if } \sigma < 1/2 \\ \infty, & \text{if } \sigma \geq 1/2 \end{cases}$
excess kurtosis	$\begin{cases} \beta_2 - 3 \text{ where } \beta_2 = \frac{(1-\sigma)\left\{1+7\sigma+6\sigma^2+\mu(\mu+1)[9+21\sigma+18\sigma^2]\right\}}{\mu(\mu+1)(1+\sigma)(1-2\sigma)(1-3\sigma)}, & \text{if } \sigma < 1/3 \\ \infty, & \text{if } \sigma \geq 1/3 \end{cases}$
PGF [a]	$\frac{(\sigma+1)}{(\mu+\sigma+1)}\ {}_2F_1(\mu\sigma^{-1}, 1; [\mu+2\sigma+1]\sigma^{-1}; t)$
pf [a,a2]	$\frac{B(y+\mu\sigma^{-1}, \sigma^{-1}+2)}{B(\mu\sigma^{-1}, \sigma^{-1}+1)}$
cdf [a2]	$1 - \frac{\Gamma(y+\mu\sigma^{-1}+1)\Gamma([\mu+1]\sigma^{-1}+1)}{\Gamma(y+[\mu+1]\sigma^{-1}+2)\Gamma(\mu\sigma^{-1})}$
Reference	Set $\nu = 1$ in $BNB(\mu, \sigma, \nu)$ [a] Wimmer and Altmann [1999], p. 643, reparametrized by $b = \sigma^{-1} + 1$ and $n = \mu\sigma^{-1}$ [a2] http://reference.wolfram.com/language/ref/WaringYuleDistribution.html reparametrised by $\alpha = \sigma^{-1} + 1$ and $\beta = \mu\sigma^{-1}$



R code on page 327

Figure 18.11: The Waring, `WARING`(μ, σ), distribution with $\mu = 1, 2, 5$ and $\sigma = .1, 2$.

18.2.8 Zero adjusted (or altered) logarithmic, ZALG(μ, σ).

Let $Y = 0$ with probability σ and $Y = Y_0$ where $Y_0 \sim$ LG(μ), a logarithmic distribution with probability $(1 - \sigma)$. Then Y has a zero adjusted (or altered) logarithmic distribution, denoted by ZALG(μ, σ), with probability function given by

$$P(Y = y|\mu, \sigma) = \begin{cases} \sigma, & \text{if } y = 0 \\ (1-\sigma)(\alpha\mu^y)/y, & \text{if } y = 1, 2, 3, \ldots \end{cases} \tag{18.12}$$

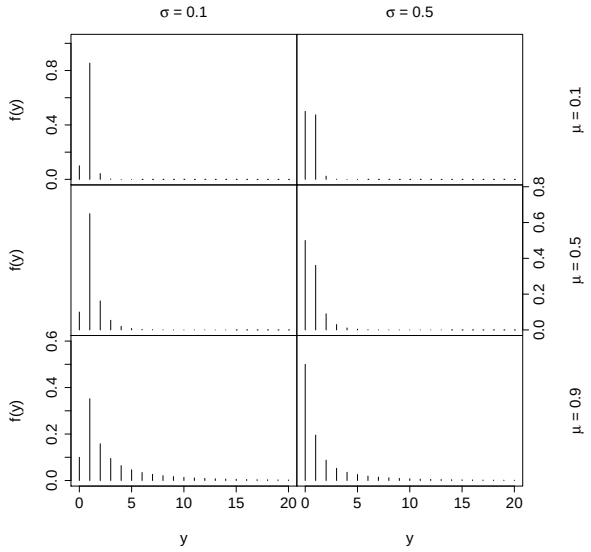
where $\alpha = -[log(1 - \mu)]^{-1}$ for $0 < \mu < 1$ and $0 < \sigma < 1$, see Johnson et al. [2005] p355. For large (and all) y, $P(Y = y|\mu, \sigma) = (1-\sigma)\alpha \exp(y \log \mu - \log y)$, so $\log P(Y = y|\mu, \sigma) \sim y \log \mu$.

Table 18.14: Zero adjusted logarithmic distribution

ZALG(μ, σ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < 1$
σ	$0 < \sigma < 1$, the probability that Y = 0
Distribution measures	
mean	$(1 - \sigma)\alpha\mu(1 - \mu)^{-1}$ where $\alpha = -[\log(1 - \mu)]^{-1}$
mode	$\begin{cases} 0, & \text{if } \sigma > \frac{\alpha\mu}{(1+\alpha\mu)} \\ 0 \text{ and } 1, & \text{if } \sigma = \frac{\alpha\mu}{(1+\alpha\mu)} \\ 1, & \text{if } \sigma < \frac{\alpha\mu}{(1+\alpha\mu)} \end{cases}$
variance	$(1 - \sigma)\alpha\mu[1 - (1 - \sigma)\alpha\mu](1 - \mu)^{-2}$
skewness	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = (1 - \sigma)\alpha\mu[1 + \mu - 3(1 - \sigma)\alpha\mu + 2(1 - \sigma)^2\alpha^2\mu^2](1 - \mu)^{-3}$
excess kurtosis	$\frac{k_4}{[Var(Y)]^2}$ where $k_4 = (1 - \sigma)\alpha\mu[1 + 4\mu + \mu^2 - (1 - \sigma)\alpha\mu(7 + 4\mu) + 12(1 - \sigma)^2\alpha^2\mu^2 - 6(1 - \sigma)^3\alpha^3\mu^3](1 - \mu)^{-4}$
PGF	$\sigma + (1 - \sigma)\frac{\log(1-\mu t)}{\log(1-\mu)}$
pf	$\begin{cases} \sigma, & \text{if } y = 0 \\ \frac{(1-\sigma)\alpha\mu^y}{y}, & \text{if } y = 1, 2, 3 \ldots \end{cases}$

Figure 18.12

```
disc2("ZALG", mu=c(.1,.5,.9), sigma=c(.1, .5), miny=0, maxy=20)
```



R code on page **??**

Figure 18.12: The zero adjusted logarithmic, ZALG(μ, σ), distribution, with $\mu = .1, .5, .9$ and $\sigma = .1, .5$.

18.2.9 Zero adjusted (or altered) Poisson, ZAP(μ, σ).

Let $Y = 0$ with probability σ and $Y = Y_0$ where $Y_0 \sim POtr(\mu)$ with probability $(1 - \sigma)$, where $POtr(\mu)$ is a Poisson truncated at zero distribution, then Y has a zero adjusted Poisson distribution, denoted by $ZAP(\mu, \sigma)$, with probability function given by

$$P(Y = y|\mu, \sigma) = \begin{cases} \sigma, & \text{if } y = 0 \\ (ce^{-\mu}\mu^y)/y!, & \text{if } y = 1, 2, 3, \ldots \end{cases} \tag{18.13}$$

for $\mu > 0$ and $0 < \sigma < 1$, where $c = (1 - \sigma)/(1 - e^{-\mu})$.

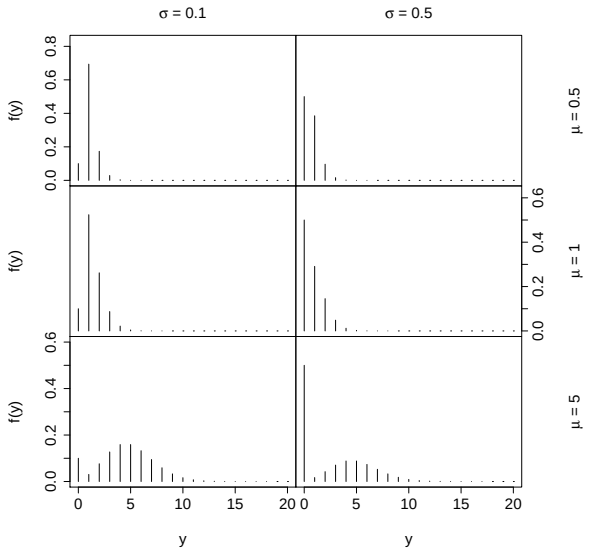
For large y, $P(Y = y|\mu, \sigma) \sim q \exp[y(\log \mu + 1) - (y + 0.5) \log y]$. where $q = ce^{-\mu}\sqrt{2\pi}$, so $\log P(Y = y|\mu, \sigma) \sim -y \log y$.

Table 18.15: Zero adjusted Poisson distribution

ZAP(μ, σ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$, mean of Poisson component before truncated$at0$
σ	$0 < \sigma < \infty$, exact probability that $Y = 0$
Distribution measures	
mean [a,a2]	$c\mu$ where $c = (1 - \sigma)/(1 - e^{-\mu})$
variance [a,a2]	$c\mu + c\mu^2 - c^2\mu^2$
skewness [a,a2]	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = c\mu[1 + 3\mu(1 - c) + \mu^2(1 - 3c + 2c^2)]$
excess kurtosis [a,a2]	$\frac{\mu_4}{[Var(Y)]^2} - 3$ where $\mu_4 = c\mu[1 + \mu(7 - 4c) + 6\mu^2(1 - 2c + c^2) + \mu^3(1 - 4c + 6c^2 - 3c^3)]$
PGF [a,a2]	$(1 - c) + ce^{\mu(t-1)}$
pf [a,a2]	$\begin{cases} \sigma, & \text{if } y = 0 \\ \frac{ce^{-\mu}\mu^y}{y!}, & \text{if } y = 1, 2, 3, \ldots \end{cases}$
cdf [a2]	$\begin{cases} \sigma, & \text{if } y = 0 \\ \sigma + c\left[\frac{\Gamma(y+1,\mu)}{\Gamma(y)} - e^{-\mu}\right], & \text{if y=1,2,3, } \ldots \end{cases}$
Reference	[a] let $\sigma \to 0$ in ZANBI(μ, σ, ν) and then set ν to σ [a2] obtained from equations (5.14), (5.15), (5.16), and (5.17), where $Y_1 \sim$ PO(μ)

```
disc2("ZAP", mu=c(.5,1,5), sigma=c(.1, .5), miny=0, maxy=20)
```

Figure 18.13



R code on page **??**

Figure 18.13: The zero adjusted Poisson, ZAP(μ, σ), distribution with $\mu = .5, 1, 5$ and $\sigma = .1, .5$.

18.2.10 Zero adjusted (or altered) zipf, ZAZIPF(μ, σ).

Let $Y = 0$ with probability σ and $Y = Y_1$ with probability $(1 - \sigma)$, where $Y_1 \sim ZIPF(\mu)$, a zipf distribution, then Y has a zero adjusted (or altered) zipf distribution, denoted by $ZAZIPF(\mu, \sigma)$, with probability function given by:

$$P(Y = y|\mu, \sigma) = \begin{cases} \sigma, & \text{if } y = 0 \\ (1 - \sigma)y^{-(\mu+1)}/\zeta(\mu + 1), & \text{if } y = 1, 2, 3, \ldots \end{cases} \quad (18.14)$$

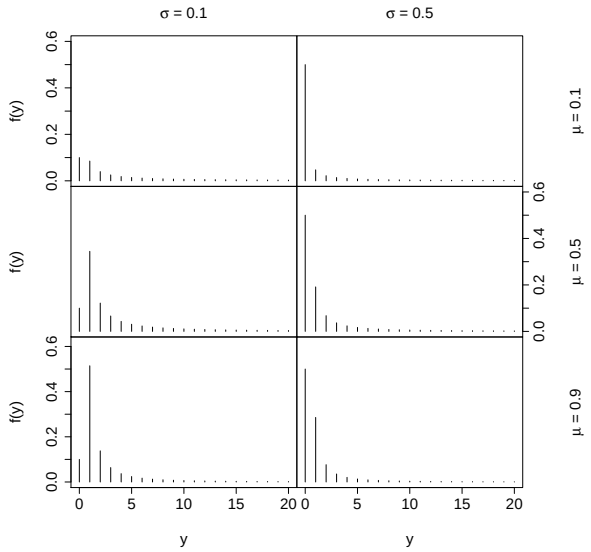
for $\mu > 0$ and $0 < \sigma < 1$.

Figure 18.14

Table 18.16: Zero adjusted zipf distribution

ZAZIPF(μ, σ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$
σ	$0 < \sigma < 1$, the probability that Y = 0
Distribution measures	
mean[a]	$\begin{cases} b(1-\sigma), & \text{if} \quad \mu > 1 \\ \infty, & \text{if } \mu \leq 1 \\ \text{where} \quad b = \zeta(\mu)/\zeta(\mu+1) & \end{cases}$
mode	$\begin{cases} 0, & \text{if} \quad \sigma > [1+\zeta(\mu+1)]^{-1} \\ 0 \quad \text{and} \quad 1, & \text{if} \quad \sigma = [1+\zeta(\mu+1)]^{-1} \\ 1, & \text{if} \quad \sigma < [1+\zeta(\mu+1)]^{-1} \end{cases}$
variance[a]	$\begin{cases} (1-\sigma)[\zeta(\mu-1)/\zeta(\mu+1)] - (1-\sigma)^2 b^2, & \text{if} \quad \mu > 2 \\ \infty, & \text{if} \quad \mu \leq 2 \end{cases}$
skewness[a]	$\begin{cases} \mu_3/[Var(Y)]^{1.5} \quad \text{where} & \\ \mu_3 = \{[(1-\sigma)\zeta(\mu-2) - 3b(1-\sigma)^2\zeta(\mu-1)]/\zeta(\mu+1)\} + & \\ \quad 2b^3(1-\sigma)^3, & \text{if } \mu > 3 \\ \infty, & \text{if } \mu \leq 3 \end{cases}$
excess kurtosis[a]	$\begin{cases} \{\mu_4/[Var(Y)]^2\} - 3 \quad \text{where} & \\ \mu_4 = \{[(1-\sigma)\zeta(\mu-3) - 4b(1-\sigma)^2\zeta(\mu-2)] + & \\ \quad 6b^2(1-\sigma)^3\zeta(\mu-1)]/\zeta(\mu+1)\} - 3b^4(1-\sigma)^4, & \text{if } \mu > 4 \\ \infty, & \text{if } \mu \leq 4 \end{cases}$
PGF	$\sigma + (1-\sigma)t\Phi(t, \mu+1, 1)/\Phi(1, \mu+1, 1)$
pf	$\begin{cases} \sigma, & \text{if} \quad y = 0 \\ (1-\sigma)y^{-(\mu-1)}/\zeta(\mu+1), & \text{if} \quad y = 1, 2, 3, \ldots \end{cases}$
Reference	[a]Obtained using equation (2.1), where $Y_1 \sim$ ZIPF(μ)
Notes	$\zeta(b) = \sum_{i=1}^{\infty} i^{-b}$ is the Riemann zeta function
	$\Phi(a, b, c) = \sum_{i=0}^{\infty} \frac{a^i}{(i+c)^b}$, for $c \neq 0, -1, -2$, is the Lerch function

```
disc2("ZAZIPF", mu=c(.1,.5,.9), sigma=c(.1, .5), miny=0, maxy=20)
```



R code on page **??**

Figure 18.14: The zero adjusted zipf, ZAZIPF(μ, σ), distribution with $\mu = .1, .5, .9$ and $\sigma = .1, .5$.

18.2.11 Zero inflated Poisson, ZIP(μ, σ) and ZIP2(μ, σ).

First parametrization, ZIP(μ, σ)

Let $Y = 0$ with probability σ and $Y = Y_1$ with probability $(1 - \sigma)$, where $Y_1 \sim PO(\mu)$. Then Y has a zero inflated Poisson distribution, denoted by $ZIP(\mu, \sigma)$, given by

$$P(Y = y|\mu, \sigma) = \begin{cases} \sigma + (1-\sigma)e^{-\mu}, & \text{if } y = 0 \\ (1-\sigma)e^{-\mu}\mu^y/y!, & \text{if } y = 1, 2, 3, \ldots \end{cases} \quad (18.15)$$

for $\mu > 0$ and $0 < \sigma < 1$.

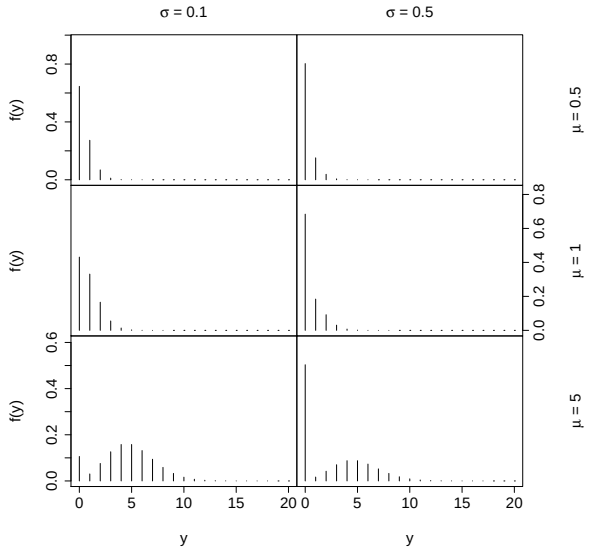
See Johnson et al. [2005] section 4.10.3, p 193 for this parametrization. This parametrization was also used by Lambert [1992]. For large y, $P(Y = y|\mu, \sigma) \sim q \exp\left[y\left(\log\mu + 1\right) - \left(y + 0.5\right)\log y\right]$. where $q = (1 - \sigma)\ e^{-\mu}\sqrt{2\pi}$, so $\log P(Y = y|\mu, \sigma) \sim -y \log y$.

Table 18.17: Zero inflated Poisson distribution

ZIP(μ, σ)	
	Ranges
Y	$0, 1, 2, 3 \dots$
μ	$0 < \mu < \infty$, mean of Poisson component
σ	$0 < \sigma < 1$, inflated (i.e. extra) probability that $Y = 0$
	Distribution measures
mean^{a,a_2}	$(1-\sigma)\mu$
variance^{a,a_2}	$\mu(1-\sigma)(1+\mu\sigma)$
skewness^{a,a_2}	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = \mu(1-\sigma)[1+3\mu\sigma+\mu^2\sigma(2\sigma-1)]$
excess kurtosis^{a,a_2}	$\frac{k_4}{[Var(Y)]^2}$ where $k_4 = \mu(1-\sigma)[1+7\mu\sigma-6\mu^2\sigma+12\mu^2\sigma^2+$ $\mu^3\sigma(1-6\sigma+6\sigma^2)]$
PGF^{a,a_2}	$\sigma + (1-\sigma)e^{\mu(t-1)}$
pf ^{a}	$\begin{cases} \sigma + (1-\sigma)e^{-\mu}, & \text{if} \quad y = 0 \\ (1-\sigma)\frac{e^{-\mu}\mu^y}{y!}, & \text{if} \quad y = 1, 2, 3, \dots \end{cases}$
cdf^{a,a_2}	$\sigma + \frac{(1-\sigma)\Gamma(y+1,\mu)}{\Gamma(y)}$
Reference	^{a} let $\sigma \to 0$ in ZINBI(μ, σ, ν) and then set ν to σ ^{a_2} obtained from equations (5.10), (5.11), (5.12) and (5.13), where $Y_1 \sim$ PO(μ)

```
disc2("ZIP", mu=c(.5,1,5), sigma=c(.1, .5), miny=0, maxy=20)
```

Figure 18.15



R code on page 337

Figure 18.15: The zero inflated Poisson, $ZIP(\mu, \sigma)$, distribution with $\mu = .5, 1, 5$ and $\sigma = .1, .5$.

18.2.12 Second parametrization, ZIP2(μ, σ)

A different parametrization of the zero inflated Poisson distribution, denoted by $ZIP2(\mu, \sigma)$, has a probability function given by

$$P(Y = y|\mu, \sigma) = \begin{cases} \sigma + (1-\sigma)e^{-\left(\frac{\mu}{1-\sigma}\right)}, & \text{if } y = 0 \\ \mu^y \{y!(1-\sigma)^{y-1}\}^{-1} e^{-\left(\frac{\mu}{1-\sigma}\right)}, & \text{if } y = 1, 2, 3, \ldots \end{cases} \quad (18.16)$$

The ZIP2(μ, σ) distribution is given by re-parametrizing μ to $\mu/(1-\sigma)$ in the ZIP(μ, σ) distribution.

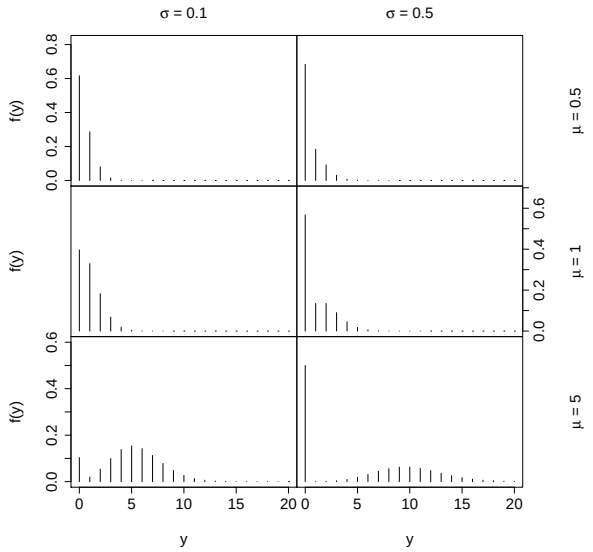
For large y, $P(Y = y|\mu, \sigma) \sim q \exp\left[y\left(\log(\frac{\mu}{1-\sigma}) + 1\right) - (y + 0.5)\log y\right]$, where $q = \sqrt{2\pi}(1-\sigma)\ e^{-\mu/(1-\sigma)}$, so $\log P(Y = y|\mu, \sigma) \sim -y \log y$.

Table 18.18: Zero inflated Poisson distribution type 2

ZIP2(μ, σ)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$, mean
σ	$0 < \sigma < 1$, inflated (i.e. extra) probability that $Y = 0$
Distribution measures	
mean	μ
variance	$\mu \left[1 + \frac{\mu\sigma}{(1-\sigma)}\right]$
skewness	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = \mu \left[1 + \frac{3\mu\sigma}{(1-\sigma)} + \frac{\mu^2\sigma(2\sigma-1)}{(1-\sigma)^2}\right]$
excess kurtosis	$\frac{k_4}{[Var(Y)]^2}$ where $k_4 = \mu[1 + \frac{7\mu\sigma}{(1-\sigma)} - \frac{6\mu^2\sigma}{(1-\sigma)^2} + \frac{12\mu^2\sigma^2}{(1-\sigma)^2} + \frac{\mu^3\sigma}{(1-\sigma)^3}(1 - 6\sigma + 6\sigma^2)]$
PGF[a,b]	$\sigma + (1-\sigma)e^{\mu(t-1)/(1-\sigma)}$
pf[a]	$\begin{cases} \sigma + (1-\sigma)e^{-\left(\frac{\mu}{1-\sigma}\right)}, & \text{if} \quad y = 0 \\ \frac{\mu^y}{y!(1-\sigma)^{y-1}} e^{-\left(\frac{\mu}{1-\sigma}\right)}, & \text{if} \quad y = 1, 2, 3, \ldots \end{cases}$
cdf[b]	$\sigma + \frac{(1-\sigma)\Gamma(y+1, \frac{\mu}{1-\sigma})}{\Gamma(y)}$
Reference	reparameterize μ to $\mu/(1-\sigma)$ in ZIP(μ, σ)

```
disc2("ZIP2", mu=c(.5,1,5), sigma=c(.1, .5), miny=0, maxy=20)
```

Figure 18.16



R code on page 339

Figure 18.16: The zero inflated Poisson, ZIP2(μ, σ), distribution type 2 with $\mu = .5, 1, 5$ and $\sigma = .1, .5$.

18.3 Count data three parameters distributions

18.3.1 Beta negative binomial distribution, BNB(μ, σ, ν)

The probability function of the beta negative binomial distribution, denoted by BNB(μ, σ, ν), is given by

$$P(Y = y|\mu, \sigma, \nu) = \frac{\Gamma(y + \nu^{-1})}{\Gamma(y + 1)} \frac{B(y + \mu\sigma^{-1}\nu, \sigma^{-1} + \nu^{-1} + 1)}{\Gamma(\nu^{-1}) \; B(\mu\sigma^{-1}\nu, \sigma^{-1} + 1)} \qquad (18.17)$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$, $\sigma > 0$ and $\nu > 0$. Note that the parameterization of BNB(μ, σ, ν) only includes beta negative binomial distributions with a finite mean μ.

The beta negative binomial distribution is also called the beta Pascal distribution or the generalized Waring distribution.

The BNB(μ, σ, ν) distribution has mean μ and is an overdispersed negarive binomial distribution. The Waring, WARING(μ, σ), distribution (which is an over dispersed geometric distrbution) is a special case of BNB(μ, σ, ν) where $\nu = 1$. The negative binomial distribution is a limiting case of the beta negative binomial since BNB$(\mu, \sigma, \nu) \to$ NBI(μ, ν) as $\sigma \to 0$ (for fixed μ and ν).

The BNB(μ, σ, ν) distribution is a re-parametrization of the distribution given in Wimmer and Altmann (1999) p19, where $m = \sigma^{-1} + 1$ and $n = \mu\sigma^{-1}\nu$ and $k = \nu^{-1}$. Hence $\mu = kn(m-1)^{-1}$, $\sigma = (m-1)^{-1}$ and $\nu = 1/k$.

The BNB(μ, σ, ν) distribution can be derived as a beta mixture of negative binomial distributions, by assuming $Y|\pi \sim NBo(k, \pi)$ where $\pi \sim BEo(m, n)$. Hence the BNB(μ, σ, ν) can be considered as an overdispersed negative binomial distribution.

For large y, $P(Y = y|\mu, \sigma, \nu) \sim q y^{-(\sigma^{-1}+2)}$ where

$$q = \Gamma\left([\mu\nu + 1]\sigma^{-1} + 1\right) / [B(\sigma^{-1} + 1, \nu)\Gamma(\mu\sigma^{-1}\nu)]$$

and hence the BNB(μ, σ, ν) distribution has a heavy right tail, especially for large σ.

The probability function (18.17) and the mean, variance, skewness and kurtosis in Table 18.19 of $Y \sim$ BNB(μ, σ, ν) can be obtained from Johnson et al. [2005] p 259 and p 263 setting $\alpha = -\mu\nu\sigma^{-1}$, $b = (\mu\nu + 1)\sigma^{-1}$ and $n = -\nu^{-1}$. The equivalence of their probability function to (18.17) is shown using their equation (1.24).

To interpret the parameters of BNB(μ, σ, ν), μ is the mean, σ is a right tail heaviness parameter (increasing the variance for $\sigma < 1$) and ν increases the variance (for $\nu^2 > \sigma/\mu$ and $\sigma < 1$), while the variance is infinite for $\sigma \geq 1$.

Table 18.19: Beta negative binomial distribution

BNB(μ,σ,ν)	
Ranges	
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$, mean
σ	$0<\sigma<\infty$
ν	$0<\nu<\infty$
Distribution measures	
mean	μ
variance	$\begin{cases}\mu(1+\mu\nu)(1+\sigma\nu^{-1})(1-\sigma)^{-1}, & \text{if } \sigma<1\\ \infty, & \text{if } \sigma\geq 1\end{cases}$
skewness	$\begin{cases}\frac{(2\mu\nu+1)(1+2\sigma\nu^{-1})(1-\sigma)^{1/2}}{\mu^{1/2}(\mu\nu+1)^{1/2}(1+\sigma\nu^{-1})^{1/2}(1-2\sigma)}, & \text{if } \sigma<1/2\\ \infty, & \text{if } \sigma\geq 1/2\end{cases}$
excess kurtosis	β_2-3 where $\beta_2=\begin{cases}\frac{(1-\sigma)\left\{1+\sigma+6\sigma\nu^{-1}+6\sigma^2\nu^{-2}+3\mu(\mu\nu+1)[1+2\nu+\sigma\nu^{-1}+6\sigma+6\sigma^2\nu^{-1}]\right\}}{\mu(\mu\nu+1)(1+\sigma\nu^{-1})(1-2\sigma)(1-3\sigma)}, & \text{if } \sigma<1/3\\ \infty, & \text{if } \sigma\geq 1/3\end{cases}$
PGF[a]	$\frac{{}_2F_1(\nu^{-1},\mu\sigma^{-1}\nu;\mu\sigma^{-1}\nu+\sigma^{-1}+\nu^{-1}+1;t)}{{}_2F_1(\nu^{-1},\mu\sigma^{-1}\nu;\mu\sigma^{-1}\nu+\sigma^{-1}+\nu^{-1}+1;1)}$
pf[a]	$\frac{\Gamma(y+\nu^{-1})B(y+\mu\sigma^{-1}\nu,\sigma^{-1}+\nu^{-1}+1)}{\Gamma(y+1)\Gamma(\nu^{-1})B(\mu\sigma^{-1}\nu,\sigma^{-1}+1)}$
Reference	[a] Wimmer and Altmann [1999] p. 19, reparameterized by $m=\sigma^{-1}+1$, $n=\mu\sigma^{-1}\nu$ and $k=\nu^{-1}$ and hence $\mu=kn(m-1)^{-1}$, $\sigma=(m-1)^{-1}$ and $\nu=1/k$

Figure 18.17
```
disc3("BNB", mu=c(1,2,5), sigma=c(0.1,0.5), nu=c(.5, 1) , miny=0, maxy=20)
```

R code on page 342

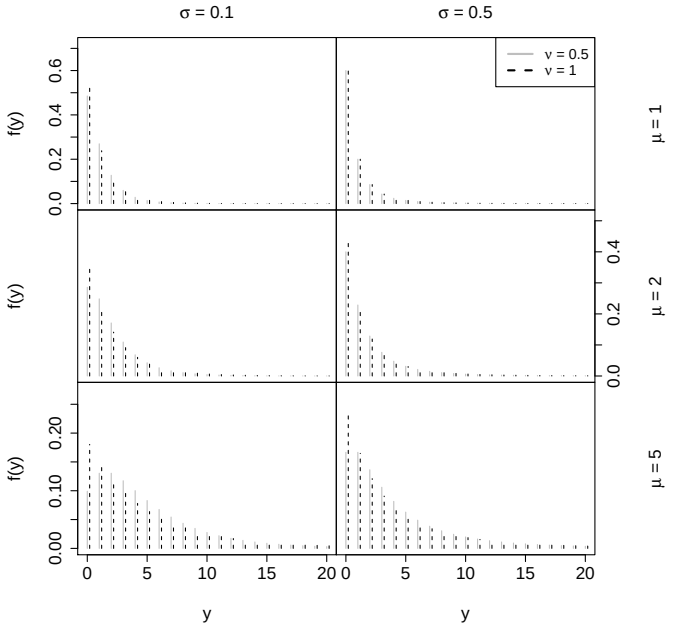


Figure 18.17: The beta negative binomial, $\texttt{BNB}(\mu, \sigma, \nu)$, distribution with $\mu = 1, 2, 5$, $\sigma = 0.1, 0.5$ and $\nu = 0.5, 1$.

18.3.2 Delaporte distribution, DEL(μ, σ, ν)

The probability function of the Delaporte distribution, denoted by DEL(μ, σ, ν), is given by

$$P(Y = y|\mu, \sigma, \nu) = \frac{e^{-\mu\nu}}{\Gamma(1/\sigma)} \left[1 + \mu\sigma(1-\nu)\right]^{-1/\sigma} S \tag{18.18}$$

where

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

for $y = 0, 1, 2, 3, \ldots$, 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 $\rho = [1 + \mu\sigma(1-\nu)]^{-1}$. This parametrization was used by Rigby et al. [2008]. For large y, $\log P(Y = y \,|\, \mu, \sigma, \nu) \sim \log\left[1 + \frac{1}{\mu\sigma(1-\nu)}\right]$

Table 18.20: The Delaport distribution

DEL(μ, σ, ν)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$, mean
σ	$0 < \sigma < \infty$
ν	$0 < \nu < 1$
Distribution measures	
mean	μ
variance	$\mu + \mu^2\sigma(1-\nu)^2$
skewness	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = \mu[1 + 3\mu\sigma(1-\nu)^2 + 2\mu^2\sigma^2(1-\nu)^3]$
excess kurtosis	$\frac{k_4}{[Var(Y)]^2}$ where $k_4 = \mu[1 + 7\mu\sigma(1-\nu)^2 + 12\mu^2\sigma^2(1-\nu)^3 + 6\mu^3\sigma^3(1-\nu)^4]$
PGF	$e^{\mu\nu(t-1)}[1 + \mu\sigma(1-\nu)(1-t)]^{-1/\sigma}$
pf	$\frac{e^{-\mu\nu}}{\Gamma(1/\sigma)} \left[1 + \mu\sigma(1-\nu)\right]^{-1/\sigma} S$ where $S = \sum_{j=0}^{y} \binom{y}{j} \frac{\mu^y \nu^{y-j}}{y!} \left[\mu + \frac{1}{\sigma(1-\nu)}\right]^{-j} \Gamma\left(\frac{1}{\sigma} + j\right)$
Reference	Rigby et al. [2008]

Figure 18.18

```
disc3("DEL", mu=c(1,2, 5), sigma=c(.5,1), nu=c(.1,.8) , miny=0, maxy=20)
```

R code on page 343

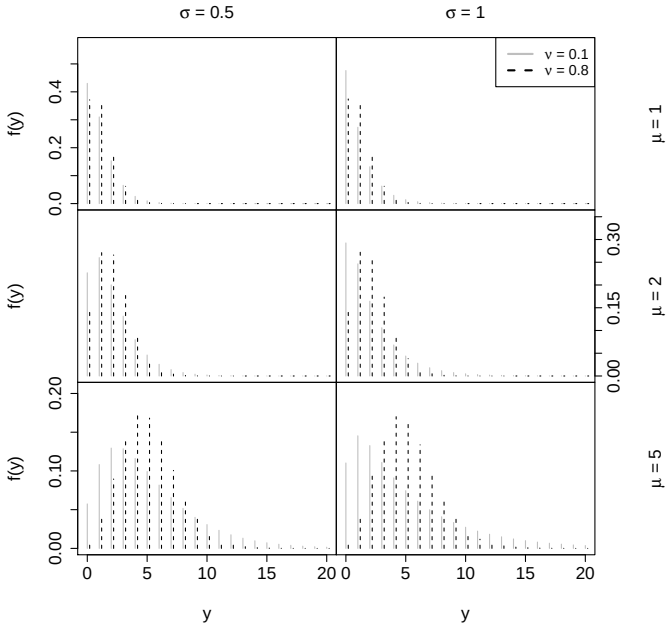


Figure 18.18: The Delaport, DEL(μ, σ, ν), distribution with $\mu = 1, 2, 5$, $\sigma = .5, 1$ and $\nu = .1, .8$.

18.3.3 Negative Binomial Family, NBF(μ, σ, ν).

The probability function of the negative binomial family distribution, denoted NBF(μ, σ, ν), is given by

$$P(Y = y|\mu, \sigma, \nu) = \frac{\Gamma\left(y + \sigma^{-1}\mu^{2-\nu}\right)\sigma^{y}\mu^{y(\nu-1)}}{\Gamma\left(\sigma^{-1}\mu^{2-\nu}\right)\Gamma\left(y+1\right)\left(1+\sigma\mu^{\nu-1}\right)^{\sigma^{-1}\mu^{2-\nu}}} \tag{18.19}$$

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

This family of reparameterazations of the negative binomial distribution is obtained by reparameterizing σ to $\sigma\mu^{\nu-2}$ in NBI(μ, σ). The variance of $Y \sim$ NBF(μ, σ, ν) is $\text{Var}(Y) = \mu + \sigma\mu^{\nu}$. Hence ν is the power in the variance-mean relationship.

Table 18.21: Negative binomial family distribution

NBF(μ, σ, ν)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$ mean
σ	$0 < \sigma < \infty$ dispersion
ν	$0 < \nu < \infty$ power parameter in variance-mean relationship
Distribution measures	
mean	μ
variance	$\mu + \sigma\mu^{\nu}$
skewness	$(1 + 2\sigma\mu^{\nu-1})(\mu + \sigma\mu^{\nu})^{-0.5}$
excess kurtosis	$6\sigma\mu^{\nu-2} + (\mu + \sigma\nu^{\nu})^{-1}$
PGF	$\left[1 + \sigma\mu^{\nu-1}(1-t)\right]^{-(\sigma\mu^{\nu-2})}$
pf	$\frac{\Gamma\left(y+\sigma^{-1}\mu^{2-\nu}\right)\sigma^{y}\mu^{y(\nu-1)}}{\Gamma(\sigma^{-1}\mu^{2-\nu})\Gamma(y+1)(1+\sigma\mu^{\nu-1})^{\sigma^{-1}\mu^{2-\nu}}}$
cdf	$1 - \frac{B\left(y+1,\sigma^{-1}\mu^{2-\nu},\sigma\mu^{\nu-1}\left(1+\sigma\mu^{\nu-1}\right)^{-1}\right)}{B(y+1,\sigma^{-1}\mu^{2-\nu})}$
Reference	Reparameterized σ to $\sigma\mu^{\nu-2}$ in NBI(μ, σ)

18.3.4 Sichel distribution, SICHEL(μ, σ, ν) and SI(μ, σ, ν).

First parametrization, SICHEL(μ, σ, ν)

This parametrization of the Sichel distribution, Rigby et al. [2008], denoted by SICHEL(μ, σ, ν), has probability function given by

$$P(Y = y|\mu, \sigma, \nu) = \frac{(\mu/b)^y K_{y+\nu}(\alpha)}{y!\,(\alpha\sigma)^{y+\nu}\, K_\nu\left(\frac{1}{\sigma}\right)} \tag{18.20}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, and $\alpha^2 = \sigma^{-2} + 2\mu(b\sigma)^{-1}$, $b = K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$ and $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp[-\frac{1}{2}t(x + x^{-1})]dx$ is the modified Bessel function of the third kind. Note

$$\sigma = \left[\left(\frac{\mu^2}{b^2} + \alpha^2\right)^{0.5} - \frac{\mu}{b}\right]^{-1}.$$

The SICHEL(μ, σ, ν) distribution is a reparameterization of the SI(μ, σ, ν) distribution given by setting μ to μ/b, so that the mean of the SICHEL(μ, σ, ν) distribution is its parameter μ.

As $\sigma \to 0$, SICHEL$(\mu, \sigma, \nu) \to$ PO(μ). For $\nu > 0$, as $\sigma \to \infty$, SICHEL$(\mu, \sigma, \nu) \to$ NBI(μ, ν^{-1}). For $\nu = 0$ then as $\sigma \to \infty$, SICHEL$(\mu, \sigma, \nu) \to$ ZALG(μ_1, σ_1) where $\mu_1 = (2\mu \log \sigma)/(1 + 2\mu \log \sigma)$ and $\sigma_1 = 1 - [\log(1 + 2\mu \log \sigma)] / [2 \log \sigma]$, which tends to a degenerate point probability 1 at $y = 0$ in the limit as $\sigma \to \infty$.

For large y, $P(Y = y \,|\, \mu, \sigma, \nu) \sim q \exp\left[-y \log\left(1 + \frac{b}{2\mu\sigma}\right) - (1 - \nu) \log y\right]$ where q does not depend on y. i.e. essentially an exponential tail. The tail becomes heavier as $2\mu\sigma/b$ increases.

Figure 18.19

```
disc3("SICHEL", mu=c(1,2, 5), sigma=c(.5,1), nu=c(-5,0),
      miny=0, maxy=20)
```

Second parametrization, SI(μ, σ, ν)

The probability function of the first parametrization of the Sichel distribution, denoted by SI(μ, σ, ν), is given by

$$P(Y = y|\mu, \sigma, \nu) = \frac{\mu^y K_{y+\nu}(\alpha)}{y!(\alpha\sigma)^{y+\nu} K_\nu(\frac{1}{\sigma})} \tag{18.21}$$

Table 18.22: Sichel distribution, SICHEL

SICHEL(μ,σ,ν)	
Ranges	
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$, mean
σ	$0<\sigma<\infty$
ν	$-\infty<\nu<\infty$
Distribution measures	
mean	μ
variance	$\mu+\mu^2 g_1$ where $g_1=\frac{2\sigma(\nu+1)}{b}+\frac{1}{b^2}-1$ and $b=\frac{K_{\nu+1}(1/\sigma)}{K_\nu(1/\sigma)}$
skewness	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3=\mu+3\mu^2 g_1+\mu^3(g_2-3g_1)$ where $g_2=\frac{2\sigma(\nu+2)}{b^3}+\frac{[4\sigma^2(\nu+1)(\nu+2)+1]}{b^2}-1$
excess kurtosis	$\frac{k_4}{[Var(Y)]^2}$ where $k_4=\mu+7\mu^2 g_1+6\mu^3(g_2-3g_1)+\mu^4(g_3-4g_2+6g_1-3g_1^2)$ and where $g_3=\frac{[1+4\sigma^2(\nu+2)(\nu+3)]}{b^4}+\frac{[8\sigma^3(\nu+1)(\nu+2)(\nu+3)+4\sigma(\nu+2)]}{b^3}-1$
PGF	$\frac{K_\nu(q)}{(q\sigma)^\nu K_\nu(1/\sigma)}$ where $q^2=\sigma^{-2}+2\mu(1-t)(b\sigma)^{-1}$
pf	$\frac{(\mu/b)^y K_{y+\nu}(\alpha)}{y!(\alpha\sigma)^{y+\nu}K_\nu\left(\frac{1}{\sigma}\right)}$ where $\alpha^2=\sigma^{-2}+2\mu(b\sigma)^{-1}$, $b=K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$ and $K_\lambda(t)=\frac{1}{2}\int_0^\infty x^{\lambda-1}\exp[-\frac{1}{2}t(x+x^{-1})]dx$
Reference	Rigby et al. [2008]

R code on page 346

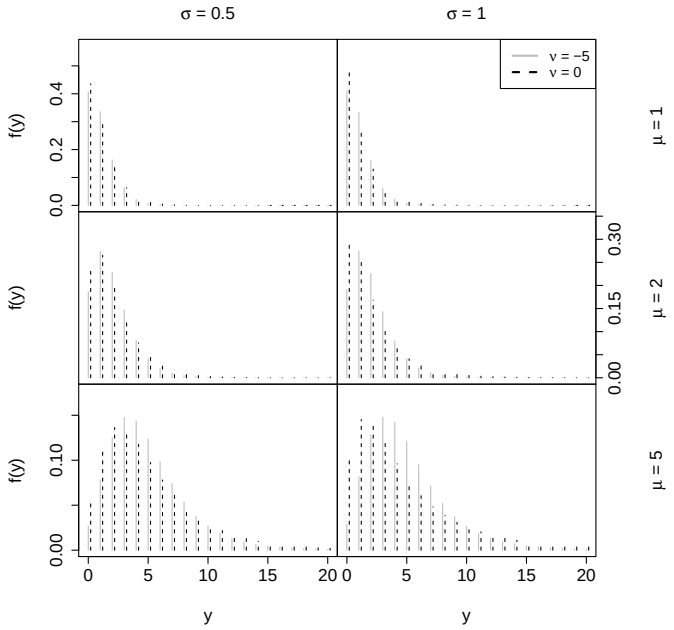


Figure 18.19: The Sichel, `SICHEL`(μ, σ), distribution with $\mu = 1, 2, 5$, $\sigma = .5, 1$ and $\nu = -5, 0$.

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$, and where $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$, and $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\{-\frac{1}{2}t(x + x^{-1})\}dx$ is the modified Bessel function of the third kind. Note that the above parametrization is different from Stein et al. [1987] section 2.1, who use the above probability function but treat μ, α and ν as the parameters. Note that $\sigma = [(\mu^2 + \alpha^2)^{0.5} - \mu]^{-1}$.

As $\sigma \to 0$, $\texttt{SI}(\mu, \sigma, \nu) \to \texttt{PO}(\mu)$. Following Stein et al. [1987], for $\nu = 0$ then as $\sigma \to \infty$, $\texttt{SI}(\mu, \sigma, \nu) \to \texttt{ZALG}(\mu_1, \sigma_1)$ where $\mu_1 = 2\mu\sigma/(1 + 2\mu\sigma)$ and $\sigma_1 = 1 - [\log(1 + 2\mu\sigma)]/[2\log\sigma]$ which tends to a degenerate $\texttt{ZALG}(1, 0.5)$ in the limit as $\sigma \to \infty$.

For large y, $P(Y = y \,|\, \mu, \sigma, \nu) \sim q \exp\left[-y \log(1 + \frac{1}{2\mu\sigma}) - (1 - \nu)\log y\right]$, where q does not depends on y. i.e. essentially an exponential tail. The tail becomes heavier as $2\mu\sigma$ increases.

Table 18.23: Sichel distribution. SI.

$\texttt{SI}(\mu, \sigma, \nu)$	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$
σ	$0 < \sigma < \infty$
ν	$-\infty < \nu < \infty$
Distribution measures	
mean	$b\mu$ where $b = \frac{K_{\nu+1}(1/\sigma)}{K_\nu(1/\sigma)}$
variance	$b\mu + b^2\mu^2 g_1$
skewness	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = b\mu + 3b^2\mu^2 g_1 + b^3\mu^3(g_2 - 3g_1)$
excess kurtosis	$\frac{k_4}{[Var(Y)]^2}$ where $k_4 = b\mu + 7b^2\mu^2 g_1 + 6b^3\mu^3(g_2 - 3g_1) + b^4\mu^4(g_3 - 4g_2 + 6g_1 - 3g_1^2)$
PGF	$\frac{K_\nu(q)}{(q\sigma)^\nu K_\nu(1/\sigma)}$ where $q^2 = \sigma^{-2} + 2\mu(1 - t)\sigma^{-1}$
pf	$\frac{\mu^y K_{y+\nu}(\alpha)}{y!(\alpha\sigma)^{y+\nu} K_\nu(\frac{1}{\sigma})}$ where $\alpha^2 = \sigma^{-2} + 2\mu\sigma^{-1}$, and $K_\lambda(t) = \frac{1}{2}\int_0^\infty x^{\lambda-1} \exp\{-\frac{1}{2}t(x + x^{-1})\}dx$
Reference	Reparameterized μ to $b\mu$ in $\texttt{SICHEL}(\mu, \sigma, \nu)$
Note	Formulae for g_1, g_2 and g_3 are given in $\texttt{SICHEL}(\mu, \sigma, \nu)$

```
disc3("SI", mu=c(1,2, 5), sigma=c(.5,1), nu=c(-1,1), miny=0,
      maxy=20)
```

Figure 18.20

R code on page 349

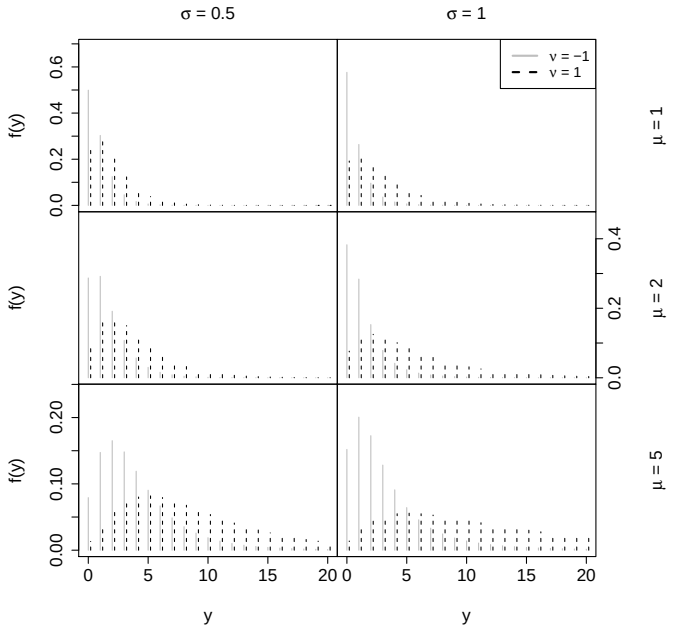


Figure 18.20: The Sichel, $\texttt{SI}(\mu, \sigma, \nu)$, distribution with $\mu = 1, 2, 5$, $\sigma = .5, 1$ and $\nu = -1, 1$.

18.3.5 Zero adjusted (or altered) negative binomial distribution, ZANBI(μ, σ, ν)

Let $Y = 0$ with probability ν and $Y = Y_0$ with probability $(1 - \nu)$, where $Y_0 \sim$ NBItr(μ, σ) and where NBItr(μ, σ) is a negative binomial, NBI(μ, σ), truncated at zero. Then Y has a zero adjusted (or altered) negative binomial distribution, denoted by ZANBI(μ, σ, ν), with probability function given by

$$P(Y = y|\mu, \sigma, \nu) = \begin{cases} \nu, & \text{if } y = 0 \\ cP(Y_1 = y|\mu, \sigma), & \text{if } y = 1, 2, 3, \ldots \end{cases} \tag{18.22}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$, where $Y_1 \sim$ NBI(μ, σ) so

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

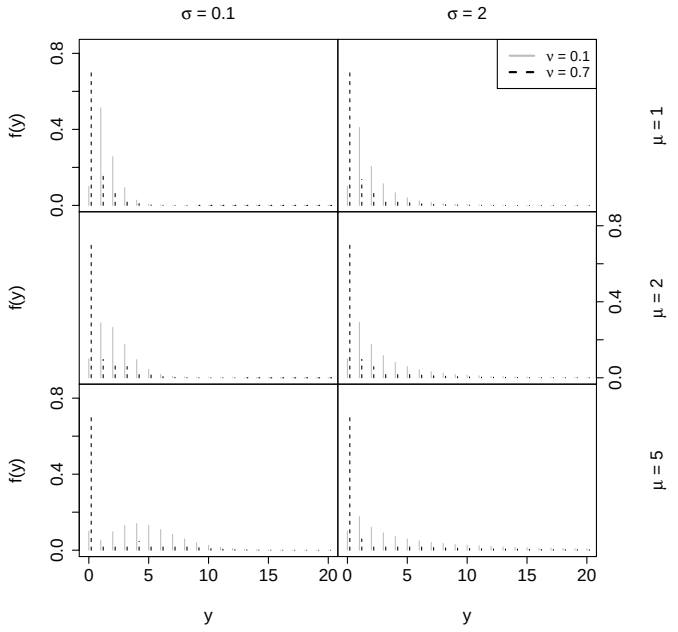
for $y = 0, 1, 2, 3, \ldots$ and $c = (1 - \nu)/(1 - p_0)$ where $p_0 = P(Y_1 = 0|\mu, \sigma) = (1 + \mu\sigma)^{-1/\sigma}$.

```
disc3("ZANBI", mu=c(1,2, 5), sigma=c(.1,2), nu=c(.1,.7) , miny=0,
      maxy=20)
```

Figure 18.21

Table 18.24: Zero adjusted negative binomial distribution

ZANBI(μ,σ,ν)	
Ranges	
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$, mean of negative binomial component before truncation at 0
σ	$0<\sigma<\infty$ dispersion of negative binomial component before truncation at 0
ν	$0<\nu<1$, exact probability that $Y=0$
Distribution measures	
mean [a,a2]	$c\mu$ where $c=\frac{(1-\nu)}{[1-(1+\mu\sigma)^{-1/\sigma}]}$
variance [a,a2]	$c\mu+c\mu^2(1+\sigma-c)$
skewness[a,a2]	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3=c\mu\left[1+3\mu\left(1+\sigma-c\right)+\mu^2\left(1+3\sigma+2\sigma^2-3c-3c\sigma+2c^2\right)\right]$
excess kurtosis [a,a2]	$\frac{\mu_4}{[Var(Y)]^2}-3$ where $\mu_4=c\mu\left[1+\mu\left(7+7\sigma-4c\right)+6\mu^2\left(1+3\sigma+2\sigma^2-2c-2c\sigma+c^2\right)+\mu^3\left(1-4c+6c^2-3c^3+6\sigma+11\sigma^2+6\sigma^3-12c\sigma+6c^2\sigma-8c\sigma^2\right)\right]$
PGF [a]	$(1-c)+c[1+\mu\sigma(1-t)]^{-1/\sigma}$
pf [a]	$\begin{cases}\nu, & \text{if } y=0\\ cP(Y_1=y\,\vert\,\mu,\sigma), & \text{if } y=1,2,3,\ldots\end{cases}$ where $Y_1\sim$ NBI(μ,σ).
cdf [a]	$\nu+c\left[1-\frac{B(y+1,\sigma^{-1},\mu\sigma(1+\mu\sigma)^{-1})}{B(y+1,\sigma^{-1})}-(1+\mu\sigma)^{-1/\sigma}\right]$
Reference	[a] obtained from equations (5.14), (5.15), (5.16), and (5.17), where $Y_1\sim$ NBI(μ,σ). [a2]for moments set $\nu=(1-c)$ in moments of ZINBI(μ,σ,ν)



R code on page 351

Figure 18.21: The zero adjusted negative binomial, `ZANBI`(μ, σ, ν), distribution with $\mu = 1, 2, 5$, $\sigma = .1, 2$ and $\nu = 0.1, 0.7$.

18.3.6 Zero adjusted (or altered) Poisson inverse Gaussian distribution, ZAPIG(μ, σ, ν)

Let $Y = 0$ with probability ν and $Y = Y_0$ with probability $(1 - \nu)$, where $Y_0 \sim$ PIGtr(μ, σ) and where PIGtr(μ, σ) is a Poisson inversed Gaussian, PIG(μ, σ), truncated at zero. Then Y has a zero adjusted (or altered) Poisson inversed Gaussian distribution, denoted by ZAPIG(μ, σ, ν), with probability function given by:

$$P(Y = y|\mu, \sigma, \nu) = \begin{cases} \nu, & \text{if } y = 0 \\ cP(Y_1 = y|\mu, \sigma), & \text{if } y = 1, 2, 3, \ldots \end{cases} \tag{18.23}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$, where $Y_1 \sim$ PIG(μ, σ), so

$$P(Y_1 = y|\mu, \sigma) = \left(\frac{2\alpha}{\pi}\right)^{1/2} \frac{\mu^y e^{1/\sigma} K_{y-0.5}(\alpha)}{y!\,(\alpha\sigma)^y}$$

for $y = 0, 1, 2, 3, \ldots$ and $c = (1-\nu)/(1-p_0)$ where $p_0 = P(Y_1 = 0|\mu, \sigma) = e^{1/\sigma - \alpha}$.

Figure 18.22

```
disc3("ZAPIG", mu=c(1,2, 5), sigma=c(.1,2), nu=c(.1,.7), miny=0,
      maxy=20)
```

Table 18.25: Zero adjusted Poisson inverse Gaussian

	$\texttt{ZAPIG}(\mu,\sigma,\nu)$
	Ranges
Y	$0,1,2,3\dots$
μ	$0<\mu<\infty$, mean of PIG component before truncation at 0
σ	$0<\sigma<\infty$ dispersion of PIG component before truncation at 0
ν	$0<\nu<1$, exact probability that $Y=0$
	Distribution measures
mean [a,a2]	$c\mu$ where $c=\frac{(1-\nu)}{\left(1-e^{1/\sigma-\alpha}\right)}$
variance [a,a2]	$c\mu+c\mu^2(1+\sigma-c)$
skewness [a,a2]	$\mu_3/[Var(Y)]^{1.5}$ where $\mu_3=c\mu\left[1+3\mu(1+\sigma-c)+\mu^2\left(1+3\sigma+3\sigma^2-3c-3c\sigma+2c^2\right)\right]$
excess kurtosis [a,a2]	$\{k_4/[Var(Y)]^2\}$ where $k_4=c\mu\left[1+7\mu(1+\sigma-c)+6\mu^2\left(1+3\sigma+3\sigma^2-3c-3c\sigma+2c^2\right)+\mu^3\left(1-7c+12c^2-6c^3+6\sigma+15\sigma^2+15\sigma^3-18c\sigma+12c^2\sigma-15c\sigma^2\right)\right]$
PGF [a]	$(1-c)+ce^{1/\sigma-q}$ where $q^2=\sigma^{-2}+2\mu(1-t)\sigma^{-1}$
pf [a]	$\begin{cases}\nu, & \text{if } y=0\\ cP(Y_1=y\mid\mu,\sigma), & \text{if } y=1,2,3,\dots\end{cases}$ where $Y_1\sim\texttt{PIG}(\mu,\sigma)$.
Reference	[a] obtained from equations (5.14), (5.15), and (5.17), where $Y_1\sim\texttt{PIG}(\mu,\sigma)$. [a2] for moments set $\nu=(1-c)$ in moments of $\texttt{ZIPIG}(\mu,\sigma,\nu)$

R code on page 354

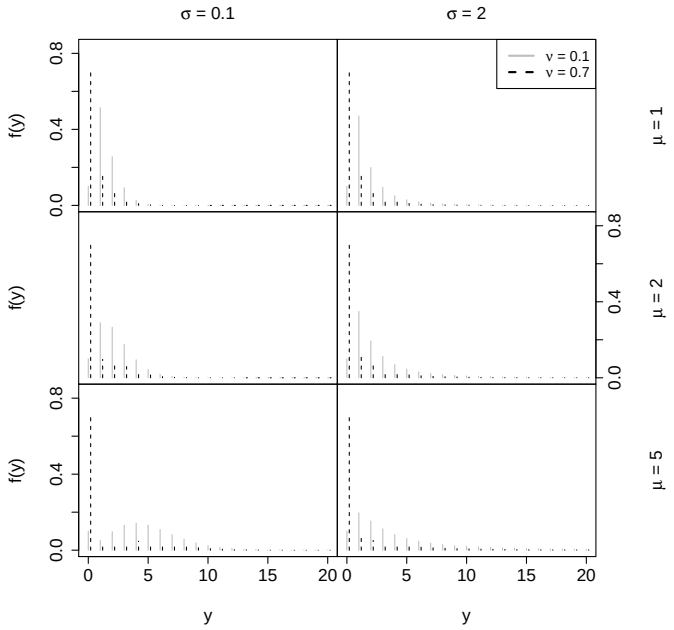


Figure 18.22: The zero adjusted Poisson inverse Gaussian, $\texttt{ZAPIG}(\mu, \sigma, \nu)$, distribution with $\mu = 1, 2, 5$, $\sigma = .1, 2$ and $\nu = 0.1, 0.7$.

18.3.7 Zero inflated negative binomial distribution, ZINBI(μ,σ,ν)

Let $Y=0$ with probability ν and $Y=Y_1$ with probability $(1-\nu)$, where $Y_1 \sim$ NBI(μ,σ). Then Y has a zero inflated negative binomial distribution, denoted by ZINBI(μ,σ,ν), with probability function given by

$$P(Y=y|\mu,\sigma,\nu)=\begin{cases}\nu+(1-\nu)\,P(Y_1=0|\mu,\sigma), & \text{if} \quad y=0\\ (1-\nu)P(Y_1=y|\mu,\sigma), & \text{if} \quad y=1,2,3,\ldots\end{cases} \tag{18.24}$$

for $\mu>0$, $\sigma>0$ and $0<\nu<1$, where $Y_1 \sim NBI(\mu,\sigma)$ and so $P(Y_1=0\,|\,\mu,\sigma)=(1+\mu\sigma)^{-1/\sigma}$. The mean of Y is given by $\mathrm{E}(Y)=(1-\nu)\,\mu$ and the variance by $\mathrm{Var}(Y)=\mu\,(1-\nu)+\mu^2\,(1-\nu)\,(\sigma+\nu)$. Hence $\mathrm{Var}(Y)=\mathrm{E}(Y)+[\mathrm{E}(Y)]^2\,(\sigma+\nu)/(1-\nu)$.

Table 18.26: Zero inflated negative binomial distribution

ZINBI(μ,σ,ν)	
Ranges	
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$ mean of the negative binomial component
σ	$0<\sigma<\infty$ dispersion of the negative binomial component
ν	$0<\nu<1$, inflated (i.e. extra) probability that $Y=0$
Distribution measures	
mean [a]	$(1-\nu)\mu$
variance [a]	$\mu(1-\nu)+\mu^2(1-\nu)(\sigma+\nu)$
skewness [a]	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3=\mu(1-\nu)\left[1+3\mu(\sigma+\nu)+\mu^2\left(2\sigma^2+3\sigma\nu+2\nu^2-\nu\right)\right]$
excess kurtosis [a]	$\frac{\mu_4}{[Var(Y)]^2}-3$ where $\mu_4=\mu(1-\nu)\left[1+\mu(3+7\sigma+4\nu)+6\mu^2\left(\sigma+2\sigma^2+2\sigma\nu+\nu^2\right)+\mu^3\left(3\sigma^2+6\sigma^3+6\sigma\nu^2+8\sigma^2\nu+\nu-3\nu^2+3\nu^3\right)\right]$
PGF [a]	$\nu+(1-\nu)[1+\mu\sigma(1-t)]^{-1/\sigma}$
pf [a]	$\begin{cases}\nu+(1-\nu)\,P(Y_1=0\|\mu,\sigma), & \text{if} \quad y=0\\ (1-\nu)P(Y_1=y\|\mu,\sigma), & \text{if} \quad y=1,2,3,\ldots\end{cases}$ where $Y_1 \sim$ NBI(μ,σ)
cdf [a]	$1-\frac{(1-\nu)B(y+1,\sigma^{-1},\mu\sigma(1+\mu\sigma)^{-1})}{B(y+1,\sigma^{-1})}$
Reference	[a]obtained from equations (5.10), (5.11), (5.12) and (5.13), where $Y_1 \sim$ NBI(μ,σ)

18.3.8 Zero inflated Poisson inverse Gaussian distribution, ZIPIG(μ, σ, ν)

Let $Y = 0$ with probability ν and $Y = Y_1$ with probability $(1 - \nu)$, where $Y_1 \sim$ PIG(μ, σ). Then Y has a zero inflated Poisson inverse Gaussian distribution, denoted by ZIPIG(μ, σ, ν), with probability function given by

$$P(Y = y|\mu, \sigma, \nu) = \begin{cases} \nu + (1 - \nu)\, P(Y_1 = 0|\mu, \sigma), & \text{if } y = 0 \\ (1 - \nu)P(Y_1 = y|\mu, \sigma), & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

for $\mu > 0$, $\sigma > 0$ and $0 < \nu < 1$, where $Y_1 \sim PIG(\mu, \sigma)$. The mean of Y is given by $\mathrm{E}(Y) = (1 - \nu)\,\mu$ and the variance by $\mathrm{Var}(Y) = \mu(1-\nu) + \mu^2(1-\nu)(\sigma+\nu)$. Hence $\mathrm{Var}(Y) = \mathrm{E}(Y) + [\mathrm{E}(Y)]^2\,(\sigma + \nu)/(1 - \nu)$.

Table 18.27: Zero adjusted Poisson inverse Gaussian distribution

ZIPIG(μ, σ, ν)	
Ranges	
Y	$0, 1, 2, 3 \ldots$
μ	$0 < \mu < \infty$ mean of the PIG component
σ	$0 < \sigma < \infty$ dispersion of the PIG component
ν	$0 < \nu < 1$, inflated (i.e. extra) probability that $Y = 0$
Distribution measures	
mean	$(1 - \nu)\mu$
variance	$\mu(1 - \nu) + \mu^2(1 - \nu)(\sigma + \nu)$
skewness	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3 = \mu(1-\nu)[1 + 3\mu(\sigma+\nu) + \mu^2(3\sigma^2 + 3\sigma\nu + 2\nu^2 - \nu)]$
excess kurtosis	$\{k_4/[Var(Y)]^2\}$ where $k_4 = \mu(1-\nu)\left[1 + 7\mu(\sigma+\nu) + 6\mu^2\left(3\sigma^2 + 3\sigma\nu + 2\nu^2 - \nu\right) + \mu^3\left(\nu - 6\nu^2 + 6\nu^3 + 15\sigma^3 - 6\sigma\nu + 12\sigma\nu^2 + 15\sigma^2\nu\right)\right]$
PGF	$\nu + (1 - \nu)e^{(1/\sigma)-q}$ where $q^2 = \sigma^{-2} + 2\mu(1 - t)\sigma^{-1}$
pf	$\begin{cases} \nu + (1 - \nu)\, P(Y_1 = 0|\mu, \sigma), & \text{if } y = 0 \\ (1 - \nu)P(Y_1 = y|\mu, \sigma), & \text{if } y = 1, 2, 3, \ldots \end{cases}$ where $Y_1 \sim PIG(\mu, \sigma)$
Reference	obtained from equations (5.10), (5.11) and (5.13), where $Y_1 \sim$ PIG(μ, σ)

18.4 Count data four parameters distributions

18.4.1 Poisson shifted generalized inverse Gaussian distribution PSGIG(μ, σ, ν, τ)

The Poisson shifted generalized inverse Gaussian distribution, Rigby et al. [2008], denoted by PSGIG(μ, σ, ν, τ), has probability function given by

$$P(Y = y|\mu, \sigma, \nu, \tau) = \frac{e^{-\mu\tau} T}{K_\nu(1/\sigma)} \tag{18.25}$$

where $y = 0, 1, 2, 3, \ldots$ where $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ and $0 < \tau < 1$ and where

$$T = \sum_{j=0}^{y} \binom{y}{j} \frac{\mu^y \tau^{y-j} K_{\nu+j}(\delta)}{y! d^j (\delta\sigma)^{\nu+j}} \tag{18.26}$$

and where $d = b/(1-\tau)$, $b = K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$ ans $\delta^2 = \sigma^{-1} + 2\mu(d\sigma)^{-1}$. Note

$$\sigma = \left[\left(\frac{\mu^2}{d^2} + \delta^2 \right)^{1/2} - \frac{\mu}{d} \right]^{-1}.$$

For large y, $\log P(Y = y \,|\, \mu, \sigma, \nu, \tau) \sim -y \log \left[1 + \frac{b}{2\mu\sigma(1-\tau)} \right]$

Table 18.28: Poisson shifted generalised inverse Gaussian distribution

PSGIG(μ,σ,ν,τ)	
Ranges	
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$, mean
σ	$0<\sigma<\infty$
ν	$-\infty<\nu<\infty$
τ	$0<\tau<1$
Distribution measures	
mean	μ
variance	$\mu+(1-\tau)^2\mu^2 g_1$ where $g_1=\frac{2\sigma(\nu+1)}{b}+\frac{1}{b^2}-1$ and $b=\frac{K_{\nu+1}(1/\sigma)}{K_\nu(1/\sigma)}$
skewness	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3=\mu+3(1-\tau)^2\mu^2 g_1+(1-\tau)^3\mu^3(g_2-3g_1)$ where $g_2=\frac{2\sigma(\nu+2)}{b^3}+\frac{[4\sigma^2(\nu+1)(\nu+2)+1]}{b^2}-1$
excess kurtosis	$\frac{k_4}{[Var(Y)]^2}$ where $k_4=\mu+7(1-\tau)^2\mu^2 g_1+6(1-\tau)^3\mu^3(g_2-3g_1)+(1-\tau)^4\mu^4(g_3-4g_2+6g_1-3g_1^2)$ and where $g_3=\frac{[1+4\sigma^2(\nu+2)(\nu+3)]}{b^4}+\frac{[8\sigma^3(\nu+1)(\nu+2)(\nu+3)+4\sigma(\nu+2)]}{b^3}-1$
PGF	$\frac{e^{\mu\tau(t-1)}K_\nu(r)}{(r\sigma)^\nu K_\nu(1/\sigma)}$ where $\tau^2=\sigma^{-2}+2\mu(1-t)(d\sigma)^{-1}$ and $d=b/(1-\tau)$ and $b=K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$
pf	$\frac{e^{\mu\tau}T}{K_\nu(1/\sigma)}$ where $T=\sum_{j=0}^{y}\binom{y}{j}\frac{\mu^y\tau^{y-j}K_{\nu+j}(\delta)}{y!d^j(\delta\sigma)^{\nu+j}}$
Reference	Rigby et al. [2008]

18.4.2 Zero adjusted (or altered) beta negative binomial, ZABNB(μ, σ, ν, τ)

Let $Y = 0$ with probability τ and $Y = Y_0$ with probability $(1 - \tau)$, where $Y_0 \sim$ BNBtr(μ, σ, ν), has a beta negative binomial, BNB(μ, σ, τ), truncated at zero. Then Y has a zero adjusted (or altered) beta negative binomial distribution, denoted by ZABNB(μ, σ, ν, τ), with probability function given by

$$P(Y = y|\mu, \sigma, \nu, \tau) = \begin{cases} \tau, & \text{if } y = 0 \\ cP(Y_1 = y|\mu, \sigma, \nu), & \text{if } y = 1, 2, 3, \dots \end{cases} \quad (18.27)$$

for $y = 0, 1, 2, 3, \dots$, where $\mu > 0$, $\sigma > 0$, $\nu > 0$ and $0 < \tau < 1$, where $Y_1 \sim$ BNB(μ, σ, ν) and $c = (1 - \tau)/(1 - p_0)$ where $p_0 = P(Y_1 = 0|\mu, \sigma, \nu) = B(\mu\sigma^{-1}\nu, \sigma^{-1} + \nu^{-1} + 1)/B(\mu\sigma^{-1}\nu, \sigma^{-1} + 1)$.

The moments of Y can be obtained from the moments of Y_1 using equation (5.15) from which the mean, variance, skewness and excess kurtosis of Y can be found. In particular the mean of Y is $\mathrm{E}(Y) = c\mu$. The variance of Y is $\mathrm{Var}(Y) = c\mu(1 + \mu\nu)(1 + \sigma\nu^{-1})(1 - \sigma)^{-1} + c(1 - c)\mu^2$ for $\sigma < 1$, while the variance is infinity for $\sigma \geq 1$.

Table 18.29: Zero adjusted beta negative binomial

ZABNB(μ, σ, ν)	
Ranges	
Y	$0, 1, 2, 3 \dots$
μ	$0 < \mu < \infty$, mean of BNB component before truncation at 0
σ	$0 < \sigma < \infty$
ν	$0 < \nu < \infty$
τ	$0 < \tau < 1$, exact probability that $Y = 0$
Distribution measures	
mean ^{a,a_2}	$c\mu$ where $c = \frac{(1-\tau)}{(1-p_0)}$ and $p_0 = P(Y_1 = 0\|\mu, \sigma, \nu)$ where $Y_1 \sim$ BNB(μ, σ, ν)
variance ^{a,a_2}	$\begin{cases} c\mu(1 + \mu\nu)(1 + \sigma\nu^{-1})(1 - \sigma)^{-1} + c(1 - c)\mu^2, & \text{for } \sigma < 1 \\ \infty, & \text{for } \sigma \geq 1 \end{cases}$
pf ^{a}	$\begin{cases} \tau, & \text{if } y = 0 \\ cP(Y_1 = y \mid \mu, \sigma, \nu), & \text{if } y = 1, 2, 3, \dots \end{cases}$ where $Y_1 \sim$ BNB(μ, σ, ν).
Reference	^{a} obtained from equations (5.14) and (5.15) where $Y_1 \sim$ BNB(μ, σ, ν). ^{a_2}for moments set $\nu = (1 - c)$ in moments of ZIBNB(μ, σ, ν)

Plots for 4 parameters is needed

18.4.3 Zero adjusted (or altered) Sichel, ZASICHEL(μ, σ, ν, τ)

Let $Y = 0$ with probability τ and $Y = Y_0$ with probability $(1 - \tau)$, where $Y_0 \sim$ SICHELtr(μ, σ, ν), has a Sichel, SICHEL(μ, σ, τ), truncated at zero. Then Y has a zero adjusted (or altered) Sichel distribution, denoted by ZASICHEL(μ, σ, ν, τ), with probability function given by

$$P(Y = y|\mu, \sigma, \nu) = \begin{cases} \tau, & \text{if } y = 0 \\ cP(Y_1 = y|\mu, \sigma, \nu), & \text{if } y = 1, 2, 3, \ldots \end{cases} \tag{18.28}$$

for $y = 0, 1, 2, 3, \ldots$, where $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$ and $0 < \tau < 1$, where $Y_1 \sim$ SICHEL(μ, σ, ν) and $c = (1 - \tau)/(1 - p_0)$ where $p_0 = P(Y_1 = 0\,|\,\mu, \sigma, \nu) = K_\nu(\alpha)/\left[(\alpha\sigma)^\nu K_\nu(1/\sigma)\right]$ where $\alpha^2 = \sigma^{-2} + 2\mu(b\sigma)^{-1}$ and $b = K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$.

Table 18.30: Zero adjusted Sichel distribution

ZASICHEL(μ,σ,ν,τ)	
Ranges	
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$, mean of SICHEL component before truncation at 0
σ	$0<\sigma<\infty$
ν	$-\infty<\nu<\infty$
τ	$0<\tau<1$, exact probability that $Y=0$
Distribution measures	
mean ^{a,a_2}	$c\mu$ where $c=\frac{(1-\tau)}{(1-p_0)}$ and $p_0=P(Y_1=0)$ where $Y_1\sim$ SICHEL(μ,σ,ν)
variance ^{a,a_2}	$c\mu+c^2\mu^2h_1$ where $h_1=\frac{1}{c}\left[\frac{2\sigma(\nu+1)}{b}+\frac{1}{b^2}\right]-1$
skewness ^{a,a_2}	$\frac{\mu_3}{[Var(Y)]^{1.5}}$ where $\mu_3=c\mu+3c^2\mu^2h_1+c^3\mu^3(h_2-3h_1-1)$ where $h_2=\frac{1}{c^2}\left\{\frac{2\sigma(\nu+2)}{b^3}+\frac{1}{b^2}\left[4\sigma^2(\nu+1)(\nu+1)+1\right]\right\}$
excess kurtosis ^{a,a_2}	$\frac{k_4}{[Var(Y)]^2}$ where $k_4=c\mu+7c^2\mu^2h_1+6c^3\mu^3(h_2-3h_1-1)+c^4\mu^4\left(h_3-4h_2+6h_1-3h_1^2+3\right)$ where $h_3=\frac{1}{c^3}\left\{\frac{1}{b^4}\left[1+4\sigma^2(\nu+2)(\nu+3)\right]+\frac{1}{b^3}\left[8\sigma^3(\nu+1)(\nu+2)(\nu+3)+4\sigma(\nu+2)\right]\right\}$
PGF ^{a}	$(1-c)+\frac{cK_\nu(q)}{(q\sigma)^\nu K_\nu(1/\sigma)}$ where $q^2=\sigma^{-1}+2\mu(1-\tau)(b\sigma)^{-1}$
pf ^{a}	$\begin{cases}\tau, & \text{if } y=0\\ cP(Y_1=y\,\vert\,\mu,\sigma,\nu), & \text{if } y=1,2,3,\ldots\end{cases}$ where $Y_1\sim$ SICHEL(μ,σ,ν).
Reference	^{a} obtained from equations(5.14), (5.15) and (5.17) where $Y_1\sim$ SICHEL(μ,σ,ν). ^{a_2}for moments set $\tau=(1-c)$ in moments of ZISICHEL(μ,σ,ν,τ)

18.4.4 Zero inflated beta negative binomial, ZIBNB(μ, σ, ν, τ)

Let $Y = 0$ with probability τ and $Y = Y_1$, with probability $(1 - \tau)$ where $Y_1 \sim$ BNB(μ, σ, ν). Then Y has a zero inflated beta negative binomial distribution, denoted by ZIBNB(μ, σ, ν, τ), with probability function given by

$$P(Y = y|\mu, \sigma, \nu, \tau) = \begin{cases} \tau + (1 - \tau)\, P(Y_1 = 0|\mu, \sigma, \nu), & \text{if } y = 0 \\ (1 - \tau)P(Y_1 = y|\mu, \sigma, \nu), & \text{if } y = 1, 2, 3, \dots \end{cases}$$

for $\mu > 0$, $\sigma > 0$, $\nu > 0$ and $0 < \tau < 1$ and $P(Y_1 = 0 \,|\, \mu, \sigma, \tau) = B(\mu\sigma^{-1}\nu, \sigma^{-1} + \nu^{-1} + 1)/B(\mu\sigma^{-1}\nu, \sigma^{-1} + 1)$.

The moments of Y can be obtained from the moments of Y_1, using equations (5.11), from which the mean, variance, skewness and excess kurtosis of Y can be found. In particular the mean of Y is $\mathrm{E}(Y) = (1 - \tau)\mu$. The variance is $\mathrm{Var}(Y) = (1 - \tau)\mu(1 + \mu\nu)(1 + \sigma\nu^{-1})(1 - \sigma)^{-1} + \tau(1 - \tau)\mu^2$ for $\sigma < 1$, while the variance is infinite for $\sigma \geq 1$.

Table 18.31: Zero inflated Sichel distribution

ZIBNB(μ, σ, ν, τ)	
Ranges	
Y	$0, 1, 2, 3 \dots$
μ	$0 < \mu < \infty$ mean of the BNB component
σ	$0 < \sigma < \infty$
ν	$0 < \nu < \infty$
τ	$0 < \tau < 1$, inflated (i.e. extra) probability that $Y = 0$
Distribution measures	
mean	$(1 - \tau)\,\mu$
variance	$(1 - \tau)\mu(1 + \mu\nu)(1 + \sigma\nu^{-1})(1 - \sigma)^{-1} + \tau(1 - \tau)\mu^2$
pf	$\begin{cases} \tau + (1 - \tau)\, P(Y_1 = 0\|\mu, \sigma, \nu), & \text{if } y = 0 \\ (1 - \tau)P(Y_1 = y\|\mu, \sigma, \nu), & \text{if } y = 1, 2, 3, \dots \end{cases}$ where $Y_1 \sim$ BNB(μ, σ, ν)
Reference	Obtained from equations (5.10) and (5.11), where $Y_1 \sim$ BNB(μ, σ, ν)

18.4.5 Zero inflated Sichel, ZISICHEL$(\mu, \sigma, \nu.\tau)$

Let $Y = 0$ with probability τ and $Y = Y_1$, with probability $(1 - \tau)$ where $Y_1 \sim$ SICHEL(μ, σ, ν). Then Y has a zero inflated Sichel distribution, denoted by ZISICHEL(μ, σ, ν, τ), with probability function given by

$$P(Y = y|\mu, \sigma, \nu, \tau) = \begin{cases} \tau + (1 - \tau)\, P(Y_1 = 0|\mu, \sigma, \nu), & \text{if } y = 0 \\ (1 - \tau)P(Y_1 = y|\mu, \sigma, \nu), & \text{if } y = 1, 2, 3, \ldots \end{cases}$$

for $\mu > 0$, $\sigma > 0$ $-\infty < \nu < \infty$ and $0 < \tau < 1$, where $P(Y_1 = 0|\mu, \sigma, \nu) = K_\nu(\alpha)/\left[(\alpha\sigma)^\nu K_\nu(1/\sigma)\right]$ where $\alpha^2 = \sigma^{-2} + 2\mu(b\sigma)^{-1}$ and $b = K_{\nu+1}(1/\sigma)/K_\nu(1/\sigma)$.

Table 18.32: Zero inflated Sichel distribution

ZISICHEL(μ,σ,ν,τ)	
	Ranges
Y	$0,1,2,3\ldots$
μ	$0<\mu<\infty$, mean of the Sichel component
σ	$0<\sigma<\infty$
ν	$-\infty<\nu<\infty$
τ	$0<\tau<1$, inflated (i.e. extra) probability that $Y=0$
	Distribution measures
mean	$(1-\tau)\,\mu$
variance	$(1-\tau)\,\mu+(1-\tau)^2\,\mu^2 h_1$ where $h_1=\frac{1}{(1-\tau)}\left[\frac{2\sigma(\nu+1)}{b}+\frac{1}{b^2}\right]-1$ and $b=\frac{k_{\nu+1}(1/\sigma)}{k_\nu(1/\sigma)}$
skewness	$\mu_3/\left[Var(Y)\right]^{1.5}$ where $\mu_3=(1-\tau)\,\mu+3\,(1-\tau)^2\,\mu^2 h_1+(1-\tau)^3\,\mu^3\,(h_2-3h_1-1)$, where $h_2=\frac{1}{(1-\tau)^2}\left\{\frac{2\sigma(\nu+2)}{b^3}+\frac{1}{b^2}\left[4\sigma^2\,(\nu+1)\,(\nu+2)+1\right]\right\}$
excess kurtosis	$k_4/\left[Var(Y)\right]^2$ where $k_4=(1-\tau)\,\mu+7\,(1-\tau)^2\,\mu^2 h_1+6\,(1-\tau)^3\,\mu^3\,(h_2-3h_1-1)+$ $(1-\tau)^4\,\mu^4\left(h_3-4h_2+6h_1-3h_1^2+3\right)$ where $h_3=\frac{1}{(1-\tau)^3}\left\{\frac{1}{b^4}\left[1+4\sigma\,(\nu+2)\,(\nu+3)\right]+\frac{1}{b^3}\left[8\sigma^3\,(\nu+1)\,(\nu+2)\,(\nu+3)\right.\right.$ $\left.\left.+4\sigma\,(\nu+2)\right]\right\}$
PGF	$\tau+(1-\tau)\frac{K_\nu(q)}{(q\sigma)^\nu K_\nu(1/\sigma)}$ where $q^2=\sigma^{-2}+2\mu(1-t)(b\sigma)^{-1}$
pf	$\begin{cases}\tau+(1-\tau)\,P(Y_1=0\|\mu,\sigma,\nu), & \text{if } y=0\\ (1-\tau)P(Y_1=y\|\mu,\sigma,\nu), & \text{if } y=1,2,3,\ldots\end{cases}$ where $Y_1\sim SICHEL(\mu,\sigma,\nu)$
Reference	Obtained from equations (5.10), (5.11) and (5.13), where $Y_1\sim$ SICHEL(μ,σ,ν)

Chapter 19

Binomial type data distributions

19.1 Binomial type data one parameter distributions

19.1.1 The Binomial distribution BI(n, μ)

The probability function of the binomial distribution, denoted here as BI(n, μ), is given by

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

for $y = 0, 1, 2, ..., n$, where $0 < \mu < 1$, (and n is a known positive integer), with $\mathrm{E}(Y) = n\mu$ and $\mathrm{Var}(Y) = n\mu(1-\mu)$. See Johnson *et al.* (1993), p 105 where $\mu = p$.

19.2 Binomial type data two parameters distributions

19.2.1 Beta Binomial distribution BB(n, μ, σ)

The probability function of the beta binomial distribution denoted here as BB(n, μ, σ) is given by

$$p_Y(y|\mu,\sigma) = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(\frac{1}{\sigma})\Gamma(y+\frac{\mu}{\sigma})\Gamma[n+\frac{(1-\mu)}{\sigma}-y]}{\Gamma(n+\frac{1}{\sigma})\Gamma(\frac{\mu}{\sigma})\Gamma(\frac{1-\mu}{\sigma})} \qquad (19.1)$$

for $y = 0, 1, 2, \ldots, n$, where $0 < \mu < 1$ and $\sigma > 0$ (and n is a known positive integer). Note that $\mathrm{E}(Y) = n\mu$ and $\mathrm{Var}(Y) = n\mu(1-\mu)\left[1 + \frac{\sigma}{1+\sigma}(n-1)\right]$.

The binomial $\texttt{BI}(n, \mu)$ distribution is the limiting distribution of $\texttt{BB}(n, \mu, \sigma)$ as $\sigma \to 0$. For $\mu = 0.5$ and $\sigma = 0.5$, $\texttt{BB}(n, \mu, \sigma)$ is a uniform distribution.

19.2.2 Zero altered (or adjusted) binomial $\texttt{ZABI}(n, \mu, \sigma)$

Let $Y = 0$ with probability σ and $Y \sim \texttt{BItr}(n, \mu)$ with probability $(1 - \sigma)$, where $\texttt{BItr}(n, \mu)$ is a Binomial truncated at zero distribution, then Y has a zero altered (or adjusted) binomial distribution, denoted by $\texttt{ZABI}(n, \mu, \sigma)$, given by

$$p_Y(y|n, \mu, \sigma) = \begin{cases} \sigma, & \text{if} \quad y = 0 \\ \frac{(1-\sigma)n!\mu^y(1-\mu)^{n-y}}{[1-(1-\mu)^n]y!(n-y)!}, & \text{if} \quad y = 1, 2, 3, \ldots \end{cases} \tag{19.2}$$

For $0 < \mu < 1$, and $0 < \sigma < 1$. The mean and variance of Y are given by

$$\mathrm{E}(Y) = \frac{(1-\sigma)\, n\mu}{[1-(1-\mu)^n]}$$

and

$$\mathrm{Var}(Y) = \frac{n\mu\,(1-\sigma)\,(1-\mu+n\mu)}{[1-(1-\mu)^n]} - [\mathrm{E}\,(Y)]^2$$

respectively.

19.2.3 Zero inflated binomial $\texttt{ZIBI}(n, \mu, \sigma)$

Let $Y = 0$ with probability σ and $Y \sim \texttt{BI}(n, \mu)$ with probability $(1 - \sigma)$, then Y has a zero inflated binomial distribution, denoted by $\texttt{ZIBI}(n, \mu, \sigma)$, given by

$$p_Y(y|n, \mu, \sigma) = \begin{cases} \sigma + (1-\sigma)\,(1-\mu)^n, & \text{if} \quad y = 0 \\ \frac{(1-\sigma)n!\mu^y(1-\mu)^{n-y}}{y!(n-y)!}, & \text{if} \quad y = 1, 2, 3, \ldots \end{cases} \tag{19.3}$$

For $0 < \mu < 1$, and $0 < \sigma < 1$. The mean and variance of Y are given by

$$\mathrm{E}(Y) = (1-\sigma)\, n\mu$$

and

$$\mathrm{Var}(Y) = n\mu\,(1-\sigma)\,[1-\mu+n\mu\sigma]$$

respectively.

19.3 Binomial type data three parameters distributions

19.3.1 Zero altered (or adjusted) beta binomial ZABB(n, μ, σ, ν)

Let $Y = 0$ with probability ν and $Y \sim \texttt{BBtr}(n, \mu, \sigma)$ with probability $(1 - \nu)$, where $\texttt{BBtr}(n, \mu, \sigma)$ is a beta binomial truncated at zero distribution, then Y has a zero altered (or adjusted) beta binomial distribution, denoted by $\texttt{ZABB}(n, \mu, \sigma, \nu)$, given by

$$p_Y(y|n, \mu, \sigma, \nu) = \begin{cases} \nu, & \text{if} \quad y = 0 \\ \frac{(1-\nu)p_{Y'}(y|n,\mu,\sigma)}{[1-p_{Y'}(0|n,\mu,\sigma)]}, & \text{if} \quad y = 1, 2, 3, \ldots \end{cases} \tag{19.4}$$

where $Y' \sim \texttt{BB}(n, \mu, \sigma)$. For $0 < \mu < 1$, $\sigma > 0$ and $0 < \nu < 1$. The mean and variance of Y are given by

$$\mathrm{E}(Y) = \frac{(1 - \nu)\, n\mu}{[1 - p_{Y'}(0|n, \mu, \sigma)]}$$

and

$$\mathrm{Var}(Y) = \frac{(1 - \nu)\left\{n\mu\,(1 = \mu)\left[1 + \frac{\sigma}{1+\sigma}\,(n - 1)\right] - n^2\mu^2\right\}}{[1 - p_{Y'}(0|n, \mu, \sigma)]} - [E(Y)]^2$$

respectively.

19.4 Binomial type data three parameters distributions

19.4.1 Zero inflated beta binomial ZIBB(n, μ, σ, ν)

Let $Y = 0$ with probability ν and $Y \sim \texttt{BB}(n, \mu, \sigma)$ with probability $(1-\nu)$, then Y has a zero inflated beta binomial distribution, denoted by $\texttt{ZIBB}(n, \mu, \sigma, \nu)$, given by

$$p_Y(y|n, \mu, \sigma, \nu) = \begin{cases} \nu + (1 - \nu)\, p_{Y'}(0|n, \mu, \sigma), & \text{if} \quad y = 0 \\ (1 - \nu)\, p_{Y'}(y|n, \mu, \sigma), & \text{if} \quad y = 1, 2, 3, \ldots \end{cases} \tag{19.5}$$

For $0 < \mu < 1$, $\sigma > 0$ and $0 < \nu < 1$ where $Y' \sim \texttt{BB}(n, \mu, \sigma)$. The mean and variance of Y are given by

$$\mathrm{E}(Y) = (1 - \nu)\, n\mu$$

and

$$\mathrm{Var}(Y) = (1-\nu)\, n\mu\,(1-\mu)\left[1 + \frac{\sigma}{1+\sigma}\,(n-1)\right] + \nu\,(1-\nu)\, n^2\mu^2$$

respectively.

Bibliography

M. A. Aitkin. Statistical inference: an integrated bayesian/likelihood approach. Chapman & Hall/CRC, 2010.

H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

H. Akaike. Information measures and model selection. Bulletin of the International Statistical Institute, 50:277–290, 1983.

DF Andrews, PJ Bickel, FR Hampel, PJ Huber, WH Rogers, and JW Tukey. Robust estimation of location: Survey and advances. Technical report, Princeton University Press, Princeton, NJ, 1972.

F. J. Anscombe. Sampling theory of the negative binomial and logarithmic series approximations. Biometrika, 37:358–382, 1950.

A. Azzalini. A class of distributions which includes the normal ones. Scand. J. Statist., 12:171–178, 1985.

A. Azzalini. Further results on a class of distributions which includes the normal ones. Statistica, 46:199:208, 1986.

A. Azzalini and A. Capitanio. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J. R. Statist. Soc. B, 65:367–389, 2003.

K. P. Balanda and H. L. MacGillivray. Kurtosis: A critical review. The American Statistician, 42:111–119, 1988.

Vic Barnett. Comparative statistical inference, volume 522. John Wiley & Sons, 1999.

E. Bowman, A. Crawford, G. Alexander, and R. W. Bowman. rpanel: Simple interactive controls for r functions using the tcltk package. Journal of Statistical Software, 17(9):1–18, 2007.

G. E. P. Box and D. R. Cox. An analysis of transformations (with discussion). J. R. Statist. Soc. B., 26:211–252, 1964.

G. E. P. Box and G. C. Tiao. Bayesian Inference in Statistical Analysis. Wiley, New York, 1973.

R.J. Butler, J. B. McDonald, R.D. Nelson, and S.B. White. Robust and partially adaptive estimation of regression models. The review of Economics and Statistics, 72(2):321–327, 1990.

R. J. Carroll and D. Ruppert. Robust estimation in heteroscedastic linear models. Ann. Statist., 10:429–441, 1982.

R. J. Carroll and D. Ruppert. Transformations and Weighting in Regression. Chapman and Hall, London, 1988.

John M. Chambers. Software for Data Analysis, Programming with R. Springer, 2008.

T. J. Cole and P. J. Green. Smoothing reference centile curves: The LMS method and penalized likelihood. Statistics in Medicine., 11:1305–1319, 1992.

P. C. Consul. Generalized Poisson Distributions. Marcel Dekker, New York, 1989.

Prem C Consul and Gaurav C Jain. A generalization of the poisson distribution. Technometrics, 15(4):791–799, 1973.

D. R. Cox and D. V. Hinkley. Theoretical Statistics. Chapman & Hall/CRC, 1979.

Crowder, M. J., Kimber, A. C., Smith R. L. and T. J. Sweeting. Statistical Analysis of Reliability Data. Chapman and Hall, London, 1991.

M. Davidian and R. J. Carroll. A note on extended quasi-likelihood. J. R. Statist. Soc., 50:74–82, 1988.

G. C. Jr. Davis and M. H. Kutner. The lagged normal family of probability density functions applied to indicator-dilution curves. Biometrics, 32:669–675, 1976.

C. Dean, J. F. Lawless, and G. E. Willmot. A mixed Poisson-inverse-Gaussian regression model. Canadian Journal of Statistics, 17(2):171–181, 1989.

T. J. DiCiccio and A. C. Monti. Inferential aspects of the skew exponential power distribution. Journal of the American Statistical Association, 99:439–450, 2004.

S. Dossou-Gbete and D. Misere. An overview of probability models for statistical modelling of count data. Monografias del Seminario Matematico Garcia de Galdeano, 33:237–244, 2006.

P. K. Dunn and G. K. Smyth. Randomized quantile residuals. Journal of Computational and Graphical Statistics, 5:236–244, 1996.

A.W. Edwards. Likelihood: an account of the statistical concept of likelihood and its application to scientific inference. Cambridge University Press, London, 1972.

B. Efron. Double exponential families and their use in generalized linear regression. J. Am. Statist. Ass., 81:709–721, 1986.

B. Efron and R.J. Tibshirani. An introduction to the bootstrap. Chapman & Hall, New York, 1993.

D. A. Evans. Experimental evidence concerning contagious distributions in ecology. Biometrika, 40:186–211, 1953.

C. Fernandez and M. F. J. Steel. On Bayesian modelling of fat tails and skewness. Journal of the American Statistical Association, 93:359–371, 1998.

C. Fernandez, J. Osiewalski, and M. J. F. Steel. Modeling and inference with v-spherical distributions. Journal of the American Statistical Association, 90: 1331–1340, 1995.

Gelman, A. Carlin, J. B. Stern, H. S. and D. B. Rubin. Bayesian Data Analysis, 2nd ed. Chapman and Hall/CRC, London, 2004.

J. F. Gibbons and S. Mylroie. Estimation of impurity profiles in ion-implanted amorphous targets using joined half-gaussian distributions. Appl. Phys. Lett., 22:568–569, 1973.

R. Gilchrist. Regression models for data with a non-zero probability of a zero response. Commun. Statist. Theory Meth., 29:1987–2003, 2000.

Gilks, W.R. Richardson, S. and Spiegelhalter D.J. Markov Chain Monte Carlo in Practice. Chapman and Hall/CRC, 1996.

J. Gill. Bayesian methods: a social and behavioral sciences approach (Second ed.). Chapman and Hall/CRC, London, 2006.

B. Hansen. Autoregressive conditional density estimation. International Ecomomic Review, 35:705–730, 1994.

H. L. Harter. Maximum-likelihood estimation of the parameters of a four parameter generalized gamma population from complete and censored samples. Technometrics, 9:159–165, 1967.

B. M. Hill. A simple general approach to inference about the tail of a distribution. Ann. Statist., 3:1163–1174, 1975.

Hougaard, P., Lee, M-L. T. and G. A. Whitmore. Analysis of overdispersed count data by mixtures of poisson variables and poisson processes. Biometrics, 53:1225–1238, 1997.

N. L. Johnson. Systems of frequency curves generated by methods of translation. Biometrika, 36:149–176, 1949.

N. L. Johnson, S. Kotz, and N. Balakrishnan. Continuous univariate distributions, volume 1. Wiley, New York, 2nd edition, 1994.

N. L. Johnson, S. Kotz, and N. Balakrishnan. Continuous univariate distributions, volume 2. Wiley, New York, 2nd edition, 1995.

N. L. Johnson, A. W. Kemp, and S. Kotz. Univariate discrete distributions. Wiley, New York, 3rd edition, 2005.

M. C. Jones. In discussion of Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape,. Appl. Statist., 54: 507–554, 2005.

M. C. Jones and M. J. Faddy. A skew extension of the t distribution, with applications. J. Roy. Statist. Soc B, 65:159–174, 2003.

M. C. Jones and A. Pewsey. Sinh-arcsinh distributions. Biometrika, 96:761–780, 2009.

B. Jørgensen. Statistical properties of the generalized inverse Gaussian distribution. Lecture Notes in Statistics No. 9. Springer-Verlag, New York, 1982.

B. Jørgensen. The Theory of Dispersion Models. Chapman and Hall: London, 1997.

D. Lambert. Zero-inflated poisson regression with an application to defects in manufacturing. Technometrics, 34:1–14, 1992.

P. Lambert and J.K. Lindsey. Analysing financial returns using regression models based on non-symmetric stable distributions. Appl. Statist., 48:409–424, 1999.

K. L. Lange, R. J. A. Little, and J. M. G. Taylor. Robust statistical modelling using the t distribution. Journal of the American Statistical Association, 84: 881–896, 1989.

J. K. Lindsey. Parametric Statistical Inference. Clarendon Press, Oxford, 1996.

James Lindsey. Modelling frequency and count data. Oxford University Press, 1995.

J.K. Lindsey. On the use of correction for overdispersion. Appl. Statist., 48: 553–561, 1999.

A. Lopatatzidis and P. J. Green. Nonparametric quantile regression using the gamma distribution. Private Communication, 2000.

G. Lovison and C. Schindler. Separate regression modelling of the gaussian and exponential components of an EMG response from respiratory physiology. In T. Kneib, F. Sobotka, J. Fahrenholz, and H. Irmer, editors, Proceedings of the 29th International Workshop on Statistical Modelling, 2014.

H.L. MacGillivray. Skewness and asymmetry: measures and orderings. Annals of Statistics, 14:994–1011, 1986.

B.B. Mandelbrot. Fractals and scaling in finance: discontinuity, concentration, risk: selecta volume E. Springer Verlag, 1997.

P. McCullagh and J. A. Nelder. Generalized linear lodels. Chapman & Hall, London, 2nd edition, 1989.

J. B. McDonald. Some generalized functions for the size distributions of income. Econometrica, 52:647–663, 1984.

J. B. McDonald. Parametric models for partially adaptive estimation with skewed and leptokurtic residuals. Economic Letters, 37:273–278, 1991.

J. B. McDonald. Probability distributions for financial models. In G. S. Maddala and C. R. Rao, editors, Handbook of Statistics, Vol. 14, pages 427–460. Elsevier Science, 1996.

J. B. McDonald and W. K. Newey. Partially adaptive estimation of regression models via the generalized t distribution. Econometric Theory, 4:428–457, 1988.

J. B. McDonald and Y. J. Xu. A generalisation of the beta distribution with applications. Journal of Econometrics, 66:133–152, 1995.

A. K. Nandi and D. Mämpel. An expension of the generalized gaussian distribution to include asymmetry. J. Franklin Inst., 332:67–75, 1995.

J. A. Nelder and Y. Lee. Likelihood, quasi-likelihood and psuedolikelihood: Some comparisons. Journal of the Royal Statistical Society, Series B, 54: 273–284, 1992.

J. A. Nelder and D. Pregibon. An extended quasi-likelihood function. Biometrika, 74:221–232, 1987.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. Journal of the Royal Statistical Society, Series A, 135:370–384, 1972.

D. B. Nelson. Conditional heteroskedasticity in asset returns: a new approach. Econometrica, 59:347–370, 1991.

J. P. Nolan. Stable Distributions - Models for Heavy Tailed Data. Birkhauser, Boston, 2012. In progress, Chapter 1 online at academic2.american.edu/∼jpnolan.

W. F. Perks. On some experiments in the graduation of mortality statistics. Journal of the Institute of Actuaries, 58:12–57, 1932.

T. Pham-Gia and Q. P. Duong. The generalized beta and f distributions in statistical modelling. Mathematical and Computer Modelling, 13:1613–1625, 1989.

M. Pokorny and J. Sedgwick. Profitability trends in Hollywood: 1929 to 1999: somebody must know something. Economic History Review, 63:56–84, 2010.

Klaas Poortema. On modelling overdispersion of counts. Statistica Neerlandica, 53(1):5–20, 1999.

Frank Proschan. Theoretical explanation of observed decreasing failure rate. Technometrics, 5(3):375–383, 1963.

A. E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. Biometrika, 83:251–266, 1996.

A. E. Raftery. Bayes Factors and BIC, comment on: A critique of the Bayesian Information Criterion for Model Selection. Sociological Methods & Research, 27:411–427, 1999.

R. A. Rigby and D. M. Stasinopoulos. Robust fitting of an additive model for variance heterogeneity. In R. Dutter and W. Grossmann, editors, COMPSTAT : Proceedings in Computational Statistics, pages 263–268. Physica, Heidelberg, 1994.

R. A. Rigby and D. M. Stasinopoulos. Construction of reference centiles using mean and dispersion additive models. Statistician, 49:41–50, 2000.

R. A. Rigby and D. M. Stasinopoulos. Smooth centile curves for skew and kurtotic data modelled using the Box Cox power exponential distribution. Statistics in Medicine, 23:3053–3076, 2004.

R. A. Rigby and D. M. Stasinopoulos. Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. Statistical Modelling, 6(3):209, 2006. ISSN 1471-082X.

R. A. Rigby, D. M. Stasinopoulos, and C. Akantziliotou. A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. Computational Statistics & Data Analysis, 53 (2):381–393, 2008. ISSN 0167-9473.

B. D. Ripley. Pattern recognition and neural networks. Cambridge University Press, Cambridge, 1996.

JL Rosenberger and M Gasko. Comparing location estimators: Trimmed means, medians and trimean. In DC Hoaglin, F Mosteller, and JW Tukey, editors, Understanding Robust and Exploratory Data Analysis, pages 297–338. John Wiley, New York, 1983.

G. E. Schwarz. Estimating the dimension of a model. Annals of Statistics, 6(2): 461–464, 1978.

H. S. Sichel. Anatomy of a generalized inverse gaussian-poisson distribution with special applications to bibliometric studies. Information Processing and Management, 28:5–17, 1992.

B. W. Silverman. Density Estimation for Statitsics and Data Analysis. Chapman & Hall, 1988.

D. M. Stasinopoulos. Contribution to the discussion of the paper by leee and nelder, double hierarchical generalized linear models. Appl. Statist., 55:171–172, 2006.

D. M. Stasinopoulos and R. A. Rigby. Generalized additive models for location scale and shape (GAMLSS) in R. Journal of Statistical Software, 23(7):1–46, 2007.

D. M. Stasinopoulos, R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani. Flexible Regression and Smoothing: Using GAMLSS in R. Chapman and Hall, Boca Raton, 2017.

G. Z. Stein, W. Zucchini, and J. M. Juritz. Parameter estimation of the sichel distribution and its multivariate extension. Journal of American Statistical Association, 82:938–944, 1987.

M. T. Subbotin. On the law of frequency of errors. Mathematicheskii Sbornik, 31:296–301, 1923.

P. Theodossiou. Financial data and the skewed generalized t distribution. Management Science, 44:1650–1661, 1998.

W. N. Venables and B. D. Ripley. S Programming. Springer, 2000. URL `http://www.stats.ox.ac.uk/pub/MASS3/Sprog/`. ISBN 0-387-98966-8.

V. Voudouris, R. Gilchrist, R. Rigby, J. Sedgwick, and D. Stasinopoulos. Modelling skewness and kurtosis with the BCPE density in GAMLSS. Journal of Applied Statistics, 39(6):1279–1293, 2012.

M. P. Wand and M. C. Jones. Kernel smoothing. Chapman & Hall, Essen, Germany, 1999.

R. W. M. Wedderburn. Quasi-likelihood functions, generalised linear models and the GaussmNewton method. Biometrika, 61:439–447, 1974.

G. Wimmer and G. Altmann. Thesaurus of univariate discrete probability distributions. Stamm Verlag, Essen, Germany, 1999.

D. Wurtz, Y. Chalabi, and L. Luksan. Parameter estimation of arma models with garch/aparch errors. an r and splus software implementation. Journal of Statistical Software, 2006.

Andreas Ziegler. Generalized estimating equations, volume 204. Springer Science & Business Media, 2011.