

# VGAM Family Functions for Positive, Zero-altered and Zero-Inflated Discrete Distributions

T. W. Yee

November 28, 2008

Beta Version 0.7-8

© Thomas W. Yee

Department of Statistics,  
University of Auckland,  
New Zealand

[yee@stat.auckland.ac.nz](mailto:yee@stat.auckland.ac.nz)

<http://www.stat.auckland.ac.nz/~yee>

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Positive distributions</b>	<b>2</b>
2.1	The positive binomial distribution . . . . .	2
2.1.1	Example: albinotic children data . . . . .	5
2.2	The positive Poisson distribution . . . . .	6
2.2.1	Example 1: Oregon group data . . . . .	7
2.2.2	Example 2: immunogold data . . . . .	8
2.3	The positive normal distribution . . . . .	9
2.3.1	Example: approximating the Poisson distribution . . . . .	9
2.4	Other functions . . . . .	10

<b>3</b>	<b>Zero-inflated distributions</b>	<b>10</b>
3.1	The zero-inflated Poisson distribution . . . . .	11
3.2	The zero-inflated binomial distribution . . . . .	11
3.3	The zero-inflated negative binomial distribution . . . . .	11
3.4	Other functions . . . . .	13
<b>4</b>	<b>Zero-altered distributions</b>	<b>13</b>
4.1	The zero-altered Poisson distribution . . . . .	13
4.2	The zero-altered negative binomial distribution . . . . .	13
<b>5</b>	<b>Conclusion</b>	<b>13</b>
	<b>Exercises</b>	<b>13</b>
	<b>References</b>	<b>15</b>

[Important note: This document and code is not yet finished, but should be completed one day ...]

## 1 Introduction

This document describes in detail VGAM family functions for *zero-truncated (positive)*, *zero-inflated* and *zero-altered* discrete distributions. They are summarized in Tables 1, 6, and 7.

Many of the following features of VGAM come from `glm()` and `gam()` so that readers unfamiliar with these functions are referred to Chambers and Hastie (1993).

## 2 Positive distributions

In this document, positive discrete distributions are distributions that are usually defined with a positive probability for  $P(Y = 0)$  but now have  $P(Y = 0) = 0$  so that the other probabilities are scaled up by a factor of  $1 - P(Y = 0)$ . Positive continuous distributions have  $P(Y \leq 0)$  set to zero, and  $P(Y > 0)$  scaled upwards. Positive distributions are also known as zero-truncated distributions.

### 2.1 The positive binomial distribution

The truncated binomial distribution can be classified into four types: left, right, double and multiple truncation. The most common form of left truncation is the omission of the zero class. The binomial truncated below one is usually called positive binomial distribution. Suppose that

$t$

Distribution	Density function $f(y; \boldsymbol{\theta})$	Range of $y$	Range of $\boldsymbol{\theta}^a$	Mean	VGAM family function
Positive binomial	$\frac{1}{(1-p)^N} \binom{N}{Ny} p^{Ny} (1-p)^{N(1-y)}$	$\frac{1}{N} (\frac{1}{N}) 1$	$0 < p < 1$	$\frac{p}{1-(1-p)^N}$	posbinomial()
Positive Poisson	$\frac{1}{1-e^{-\lambda}} \frac{e^{-\lambda} \lambda^y}{y!}$	$1(1)\infty$	$\lambda > 0$	$\frac{\lambda}{1-e^{-\lambda}}$	pospoisson()
Positive negative binomial	$\frac{1}{1-(k/(k+\mu))^k} \binom{y+k-1}{y} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{k+\mu}\right)^k$	$1(1)\infty$	$\sigma > 0$	$\frac{\mu}{1-(k/(k+\mu))^k}$	posnegbinomial()
Positive normal	$\frac{1}{\sigma} \frac{\phi((y-\mu)/\sigma)}{[1-\Phi(-\mu/\sigma)]}$	$(0, \infty)$	$\sigma > 0$	$\mu + \sigma \frac{\phi(-\mu/\sigma)}{1-\Phi(-\mu/\sigma)}$	posnormal1()

Table 1: Summary of positive distributions currently supported by the VGAM package.

$a$   $-\infty < \theta_j < \infty$  and  $\theta_j$  real-valued is assumed unless otherwise stated.

$nY$  has the positive binomial distribution where  $n$  is a constant number and  $Y$  is a random variable. The probability function of  $Y$  is

$$f(y; p) = \begin{cases} \frac{\binom{n}{ny} p^{ny} (1-p)^{n-ny}}{1 - (1-p)^n} & y = \frac{1}{n}, \frac{2}{n}, \dots, 1 \text{ and } 0 < p < 1, \\ 0 & y = 0. \end{cases}$$

The mean and variance are

$$\begin{aligned} E(Y) &= \frac{p}{1 - (1-p)^n}, \\ \text{Var}(Y) &= \frac{p(1 + p(n-1))}{n(1 - (1-p)^n)} - \frac{p^2}{(1 - (1-p)^n)^2}. \end{aligned}$$

The log-likelihood for the parameter  $p$  is

$$\ell_i(p_i; y_i) = \log \binom{n_i}{n_i y_i} + n_i y_i \log(p_i) + (n_i - n_i y_i) \log(1 - p_i) - \log(1 - (1 - p_i)^{n_i}),$$

the score and Hessian functions are

$$\begin{aligned} \frac{\partial \ell_i}{\partial p_i} &= \frac{n_i y_i}{p_i} - \frac{n_i - n_i y_i}{1 - p_i} - \frac{n_i (1 - p_i)^{n_i - 1}}{1 - (1 - p_i)^{n_i}}, \\ \frac{\partial^2 \ell_i}{\partial p_i^2} &= -\frac{n_i y_i}{p_i^2} - \frac{n_i - n_i y_i}{(1 - p_i)^2} + \frac{n_i (n_i - 1) (1 - p_i)^{n_i - 2}}{1 - (1 - p_i)^{n_i}} \\ &\quad + \left( \frac{n_i (1 - p_i)^{n_i - 1}}{1 - (1 - p_i)^{n_i}} \right)^2. \end{aligned}$$

and  $E \left( \frac{\partial^2 \ell_i}{\partial p_i^2} \right) = -\frac{n_i}{p_i (1 - (1 - p_i)^{n_i})} - \frac{n_i}{(1 - p_i)^2} + \frac{n_i p_i}{(1 - p_i)^2 (1 - (1 - p_i)^{n_i})}$   
 $+ \frac{n_i (n_i - 1) (1 - p_i)^{n_i - 2}}{1 - (1 - p_i)^{n_i}} + \left( \frac{n_i (1 - p_i)^{n_i - 1}}{1 - (1 - p_i)^{n_i}} \right)^2.$

The maximum likelihood estimator  $\hat{p}_i$  of  $p$  is given by setting  $\frac{\partial \ell_i}{\partial p_i} = 0$  and satisfies the equation

$$y_i = \frac{\hat{p}_i}{1 - (1 - \hat{p}_i)^{n_i}}.$$

The fitted values are  $\hat{\mu}_i = \frac{\hat{p}_i}{1 - (1 - \hat{p}_i)^{n_i}}$ . We cannot solve this explicitly, therefore VGAM maximizes  $\ell$  rather than minimize the deviance. As well as a logit link (default), others link functions are available, e.g., probit link and complementary log-log link, which are

$$\eta = \log(p/(1-p)), \eta = \Phi^{-1}(p) \text{ or } \eta = \log(-\log(1-p)).$$

The truncated binomial is implemented in the family function `posbinomial()`, which operates in a similar way as `binomialff()`.

### 2.1.1 Example: albinotic children data

The data in Table 2 are reproduced from Patil (1962), who attributes them to Karl Pearson in his studies on albinism. For a sample of 60 families, each with  $n = 5$  children, the table gives the number  $f_k$  of families with exactly  $k$  albinotic children, for  $k = 1, \dots, 5$ .

If  $p$  is the probability that a child is albinotic in one of the families, then the distribution of the number of albinos in one family is a positive binomial. The proportion of albinotic children in  $i$ th family is  $y_i = k_i/n$  for  $i = 1, \dots, 60$ . Then

```
> y <- c(rep(1, 25), rep(2, 23), rep(3, 10), 4, 5)
> n <- rep(5, 60)
> fit <- vglm(cbind(y, n - y) ~ 1, posbinomial)
```

or equivalently,

```
> yprop <- y/n
> fit <- vglm(yprop ~ 1, posbinomial, weight = n)
```

The output is

```
> print(summary(fit))
```

Call:

```
vglm(formula = yprop ~ 1, family = posbinomial, weights = n)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
logit(p)	-0.97046	-0.97046	0.19409	0.19409	3.6877

Coefficients:

	Value	Std. Error	t value
(Intercept)	-0.80559	0.15036	-5.3576

Number of linear predictors: 1

Name of linear predictor: logit(p)

Dispersion Parameter for posbinomial family: 1

Log-likelihood: -356.4802 on 59 degrees of freedom

Number of Iterations: 3

The maximum likelihood estimate  $\hat{p} = e^{\hat{\eta}}/(1 + e^{\hat{\eta}})$  is

```
> Coef(fit)
```

Table 2: Number of albinotic children in families with five children.

number of albinos ( $k$ )	1	2	3	4	5
frequency ( $f_k$ )	25	23	10	1	1

0.3088317

or

```
> exp(coef(fit))/(1 + exp(coef(fit)))
```

```
(Intercept)
```

```
0.3088317
```

The estimated mean  $\hat{\mu}$  is

```
> fitted(fit)[1]
```

```
[1] 0.3666667
```

That is, the maximum likelihood estimate  $\hat{p}$  is 0.309 and the estimated mean  $\hat{\mu}$  of  $y$  is 0.367.

## 2.2 The positive Poisson distribution

The truncated Poisson distribution can be classified into four types, left, right, double and multiple truncation. The most common form of left truncation is the omission of the zero class. The Poisson truncated below one is also known as the positive Poisson. Its probability function is

$$f(y; \lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^y}{y! (1 - e^{-\lambda})} & y = 1, 2, \dots \text{ and } \lambda > 0, \\ 0, & y = 0. \end{cases}$$

where  $\lambda$  is the rate parameter. The mean and variance are

$$E(Y) = \frac{\lambda}{1 - e^{-\lambda}},$$
$$\text{Var}(Y) = \frac{\lambda(1 + \lambda)}{1 - e^{-\lambda}} - \frac{\lambda^2}{(1 - e^{-\lambda})^2} = \frac{\lambda e^{\lambda}(-1 + e^{\lambda} - \lambda)}{(e^{\lambda} - 1)^2}.$$

The log-likelihood is

$$\ell_i(\lambda; y_i) = -\lambda_i + y_i \log \lambda_i - \log(1 - e^{-\lambda_i}) - \log y_i!,$$

the score and Hessian functions are

$$\frac{\partial \ell_i}{\partial \lambda_i} = -1 + \frac{y_i}{\lambda_i} - \frac{1}{e^{\lambda_i} - 1},$$
$$\frac{\partial^2 \ell_i}{\partial \lambda_i^2} = -\frac{y_i}{\lambda_i^2} + \frac{e^{\lambda_i}}{(e^{\lambda_i} - 1)^2},$$
$$\text{and } E\left(\frac{\partial^2 \ell_i}{\partial \lambda_i^2}\right) = -\frac{e^{\lambda_i}}{(e^{\lambda_i} - 1)} \left(\frac{1}{\lambda_i} - \frac{1}{e^{\lambda_i} - 1}\right).$$

The maximum likelihood estimator  $\hat{\lambda}_i$  of  $\lambda$  is given by setting  $\frac{\partial \ell_i}{\partial \lambda_i} = 0$  satisfies the equation

$$y_i = \frac{\hat{\lambda}_i}{1 - e^{-\hat{\lambda}_i}}.$$

The fitted values are  $\hat{\mu}_i = \frac{\hat{\lambda}_i}{1 - e^{-\hat{\lambda}_i}}$ . A log link (default) or identity link can be used; these are

$$\eta = \log \lambda \text{ or } \eta = \lambda.$$

Since an expression for  $\ell_{max}$  cannot be found easily, the VGAM will maximize a log-likelihood  $\ell(\lambda)$  rather than minimize a deviance.

### 2.2.1 Example 1: Oregon group data

The data from Coleman and James (1961) are in Table 3. A possible model for these data is a truncated zero Poisson distribution. We can treat  $y$  as group size and the frequency as weights. The model is fitted by VGAM as follows:

```
> y <- 1:6
> w <- c(1486, 694, 195, 37, 10, 1)
> fit <- vglm(y ~ 1, pospoisson, weights = w)
```

The output is

```
> print(summary(fit), presid = FALSE)
```

Call:

```
vglm(formula = y ~ 1, family = pospoisson, weights = w)
```

Coefficients:

```
                Value Std. Error t value
(Intercept) -0.11373    0.026777 -4.2473
```

Number of linear predictors: 1

Name of linear predictor: log(lambda)

Dispersion Parameter for pospoisson family: 1

Log-likelihood: -2304.659 on 5 degrees of freedom

Number of Iterations: 5

The maximum likelihood estimate  $\hat{\lambda} = e^{\hat{\eta}}$ , is

```
> exp(coef(fit))
```

```
(Intercept)
  0.892496
```

or

```
> Coef(fit)
```

```
lambda
0.892496
```

Table 3: On a spring afternoon in Portland, Oregon, data on the sizes of different groups observed in public places were collected by Coleman and James (1961).

Group size	1	2	3	4	5	6
Frequency	1486	694	195	37	10	1

The estimated mean is

```
> fitted(fit)[1]
[1] 1.511762
```

That is, the mean group size is estimated to be 1.512 persons for the observed data and 0.892 persons for the expected Poisson distribution.

### 2.2.2 Example 2: immunogold data

The data from Cullen et al. (1990) are reproduced in Table 4. The total number of labelled sites is 198 and the total number of gold particles counted is 312. The number of unlabelled sites is unknown. Let  $y$  be the number of gold particles counted in a labelled site. Therefore  $y$  is truncated below one. The observed distribution for this data is possibly a positive Poisson. The maximum likelihood estimate  $\hat{\lambda} = 0.9906$  and the estimated standard error = 0.0860 are obtained from Matthews and Appleton (1993). We can fit the model using VGAM with weights as the example in Section 2.2.1 or as follows:

```
> y <- c(rep(1, 122), rep(2, 50), rep(3, 18), rep(4, 4),
+       rep(5, 4))
> fit <- vglm(y ~ 1, pospoisson(link = "identity"))
```

We get

```
> print(summary(fit))
```

Call:

```
vglm(formula = y ~ 1, family = pospoisson(link = "identity"))
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
lambda	-0.71213	-0.71213	-0.71213	0.52473	4.2353

Coefficients:

	Value	Std. Error	t value
(Intercept)	0.99059	0.087073	11.377

Number of linear predictors: 1

Name of linear predictor: lambda

Dispersion Parameter for pospoisson family: 1

Log-likelihood: -205.9477 on 197 degrees of freedom

Number of Iterations: 4

Table 4: Immunogold assay data were from Cullen et al. (1990).

Number of particles	1	2	3	4	5
Observed number of sites	122	50	18	4	4

The maximum likelihood estimate  $\hat{\lambda}$  is

```
> Coef(fit)
```

```
lambda  
0.9905855
```

The estimated mean is

```
> fitted(fit)[1]
```

```
[1] 1.575758
```

We obtain the same  $\hat{\lambda}$  as Cullen et al. (1990). The rate parameter or the expected mean number of particles is estimated to be 0.9906. The observed mean number of particles is estimated to be 1.576. Our standard error is slightly different from their paper.

## 2.3 The positive normal distribution

The VGAM family function `posnormal1()` implements Fisher scoring. It is not too difficult to show

$$\begin{aligned} -E \left[ \frac{\partial^2 \ell}{\partial \mu^2} \right] &= \frac{1}{\sigma^2} \left\{ 1 - \frac{\mu\phi}{\sigma(1-\Phi)} - \frac{\phi^2}{(1-\Phi)^2} \right\}, \\ -E \left[ \frac{\partial^2 \ell}{\partial \sigma^2} \right] &= \frac{2}{\sigma^2} - \frac{\mu\phi}{\sigma^3(1-\Phi)} \left\{ 1 + \frac{\mu^2}{\sigma^2} + \frac{\phi\mu}{\sigma(1-\Phi)} \right\}, \\ -E \left[ \frac{\partial^2 \ell}{\partial \mu \partial \sigma} \right] &= \frac{\phi}{\sigma^2(1-\Phi)} \left\{ 1 + \frac{\mu^2}{\sigma^2} + \frac{\phi\mu}{\sigma(1-\Phi)} \right\} \end{aligned}$$

where  $\phi = \phi(-\mu/\sigma)$  and  $\Phi = \Phi(-\mu/\sigma)$ .

### 2.3.1 Example: approximating the Poisson distribution

It is well known that the normal distribution approximates the Poisson distribution well when  $\mu$  is large, e.g.,  $\mu > 10$ . But for  $\mu$  small, one possibility is to try a positive-normal approximation. Here is an example.

```
> lambda = 1  
> NN = 10  
> y = 0:NN  
> wts = dpois(y, lambda = lambda)  
> y = 0.5 + y  
> yfine = seq(0, max(y), len = 200)  
> fit = vglm(y ~ 1, fam = posnormal1(imean = lambda, isd = 1),  
+ weight = wts)  
> coef(fit, matrix = TRUE)  
  
mean log(sd)  
(Intercept) 0.9161409 0.3145143  
  
> Cfit = Coef(fit)  
> mygrid = seq(0, NN, len = 200)
```

Note  $\frac{1}{2}$  is added to the response so that a continuity correction can be performed. Then Figure 1 was produced by

```
> plot(y, wts, type = "h", col = "blue", ylim = range(wts) *  
+ 1.1, ylab = "Probabilities", main = paste("mu =", lambda),  
+ xlim = c(0, max(y) + 0.5))  
> lines(yfine, dnorm(yfine, lambda + 0.5, sqrt(lambda)),  
+ col = "darkgreen")  
> lines(mygrid, dposnorm(mygrid, m = Cfit[1], sd = Cfit[2]),  
+ col = "darkred", lty = "dashed")
```

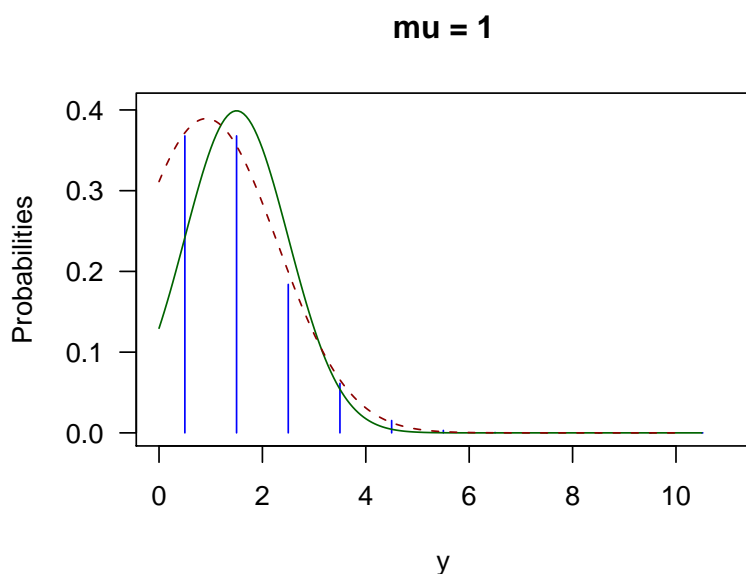


Figure 1: Positive-normal approximation to the Poisson for small  $\mu$ . The green curve is the normal approximation to the Poisson, and the dashed red curve is the positive-normal approximation to the Poisson.

## 2.4 Other functions

The R functions in Table 5 summarize those currently available in VGAM. These functions calculate the density, cumulative distribution function, quantiles and random variates. See the online help for further details.

## 3 Zero-inflated distributions

A (discrete) random variable  $Y_i$  is said to have a zero-inflated distribution if it has value 0 with probability  $\phi$ , otherwise it has some other distribution with  $P(Y = 0) > 0$ . Hence  $P(Y = 0)$  comes from two sources, and the  $\phi$  source can sometimes be thought of as a structural zero.

Distribution	Random variates functions
Zero-altered negative binomial	[dpqr]zanegbin()
Zero-altered Poisson	[dpqr]zapois()
Zero-inflated binomial	[dpqr]zibinom()
Zero-inflated negative binomial	[dpqr]zinegbin()
Zero-inflated Poisson	[dpqr]zipois()
Positive binomial	[dpqr]posbinom()
Positive negative binomial	[dpqr]posnegbinom()
Positive normal	[dpqr]posnorm()
Positive Poisson	[dpqr]pospois()

Table 5: Summary of VGAM functions for generating random variates etc. The prefix “d” means the density, “p” means the distribution function, “q” means the quantile function and “r” means random deviates.

### 3.1 The zero-inflated Poisson distribution

The most common example of a zero-inflated distribution is the zero-inflated Poisson. It has value 0 with probability  $\phi$  else is Poisson( $\lambda$ ) distributed. Thus its probability function is

$$\begin{aligned} f(y) &= (1 - \phi)e^{-\lambda}\lambda^y/y!, & y = 1, 2, \dots, \\ f(0) &= \phi + (1 - \phi)e^{-\lambda}. \end{aligned} \quad (1)$$

A possible example is the distribution of the number of children a person has given birth to from a population where there is a proportion  $1 - \phi$  of females.

The VGAM family function zipoisson() implements (1). It is based on the expected information matrix given in Thas and Rayner (2005), and is recommended over the VGAM family function yip88() which estimates a *conditional likelihood*. The theory presented in Yip (1988) did not mention the use of explanatory variables other than an intercept term, however, zipoisson() should work reasonably in general. It is a good idea to set zero=1 if there are numerical problems.

### 3.2 The zero-inflated binomial distribution

The VGAM family function zibinomial() implements

$$\begin{aligned} f(y) &= (1 - \phi) \binom{N}{Ny} \mu^{Ny} (1 - \mu)^{N(1-y)}, & y = \frac{1}{N}, \frac{2}{N}, \dots, 1, \\ f(0) &= \phi + (1 - \phi) (1 - \mu)^N. \end{aligned}$$

The value  $N$  here is the number of trials and corresponds to the argument size in the software.

### 3.3 The zero-inflated negative binomial distribution

The VGAM family function zinegbinomial() implements

$$f(y) = (1 - \phi) \binom{y+k-1}{y} \left( \frac{\mu}{\mu+k} \right)^y \left( \frac{k}{k+\mu} \right)^k, \quad y = 1, 2, \dots \quad (2)$$

$$f(0) = \phi + (1 - \phi) (k/(k + \mu))^k. \quad (3)$$

Zero-inflated distribution	Probability function $f(y; \boldsymbol{\theta})$	Range of $y$	Range of $\boldsymbol{\theta}^a$	Mean	VGAM family function
ZI negative binomial	$I(y=0)\phi + (1-\phi) \times \binom{y+k-1}{y} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{k+\mu}\right)^k$	$0(1)\infty$	$0 < \phi < 1, \mu > 0, k > 0$	$(1-\phi)\mu$	<code>zinegbinomial()</code>
ZI binomial	$I(y=0)\phi + (1-\phi) \times \binom{N}{Ny} p^{Ny} (1-p)^{N(1-y)}$	$0(1/N)1$	$0 < \phi < 1, 0 < p < 1$	$(1-\phi)p$	<code>zibinomial()</code>
ZI Poisson	$I(y=0)\phi + (1-\phi) \frac{e^{-\lambda} \lambda^y}{y!}$	$0(1)\infty$	$0 < \phi < 1, \lambda > 0$	$(1-\phi)\lambda$	<code>zipoisson()</code>

Table 6: Summary of zero-inflated discrete distributions currently supported by VGAM. “ZI” stands for “zero-inflated”.

<sup>a</sup> $-\infty < \theta_j < \infty$  and  $\theta_j$  real-valued is assumed unless otherwise stated.

### 3.4 Other functions

All VGAM family functions for zero-inflated models come with dpqr-type functions, e.g., for the zero-inflated Poisson they are `dzipois()`, `pzipois()`, `qzipois()`, `rzipois()`. These functions calculate the density, cumulative distribution function, quantiles and random variates. See Table 5 and the online help for further details.

## 4 Zero-altered distributions

One problem with the zero-inflated distributions is that the probability that  $Y = 0$  is no smaller than the nominal probability. This can cause numerical difficulties when there are fewer 0s in the response than the ordinary reference distribution. One way around this problem is to consider *zero-altered* version of the distribution.

A zero-altered distribution is where  $P(Y = 0)$  is modelled separately from  $P(Y > 0)$ . For example,  $P(Y = 0)$  could be a logistic regression and the distribution of  $Y$ , given that  $Y > 0$ , is modelled using a positive distribution. We say the response  $Y$  is zero with probability  $\phi$  else  $Y$  has a positive distribution. For example, the zero-altered negative binomial distribution differs from the zero-inflated negative binomial distribution in that the former has zeros coming from one source, whereas the latter has zeros coming from the negative binomial distribution too. Some people call the zero-altered negative binomial a *hurdle* model. Table 7 lists zero-altered discrete distributions currently supported by VGAM.

Other packages such as `pscl` and `MASS` can fit some zero-inflated and zero-altered count models. Note that `pscl` calls `optim()` rather than using Fisher scoring as in here.

### 4.1 The zero-altered Poisson distribution

The details are given in Table 7. Note this is a mixture of  $Y = 0$  and a positive Poisson distribution.

### 4.2 The zero-altered negative binomial distribution

The details are given in Table 7. Note this is a mixture of  $Y = 0$  and a positive negative binomial distribution.

## 5 Conclusion

This document has described some simple zero-truncated, zero-altered and zero-inflated univariate distributions. There are many more distributions that could be done, as well as truncation of other types, e.g., the upper tail rather than at zero. Users are encouraged to submit VGAM family functions that they have written so that others may use them.

## Acknowledgements

The work on positive distributions was motivated by a project by Alison Yu.

## Exercises

$t$

Zero-altered distribution	Probability function $f(y; \theta)$	Range of $y$	Range of $\theta^a$	Mean	VGAM family function
ZA negative binomial	$I(y=0)\phi + I(y>0)\frac{(1-\phi)}{1-(k/(k+\mu))^k} \times$ $\binom{y+k-1}{y} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{k+\mu}\right)^k$	$0(1)\infty$	$0 < \phi < 1, \mu > 0, k > 0$	$\frac{(1-\phi)\mu}{1-(k/(k+\mu))^k}$	zanegbinomial()
ZA Poisson	$I(y=0)\phi + I(y>0)(1-\phi)\frac{e^{-\lambda}\lambda^y}{(1-e^{-\lambda})y!}$	$0(1)\infty$	$0 < \phi < 1, \lambda > 0$	$\frac{(1-\phi)\lambda}{1-e^{-\lambda}}$	zapoisson()

Table 7: Summary of zero-altered discrete distributions currently supported by VGAM. “ZA” stands for “zero-altered”.

---

$a$ — $\infty < \theta_j < \infty$  and  $\theta_j$  real-valued is assumed unless otherwise stated.

1. Show that the variance of the zero-inflated Poisson (1) is  $\lambda(1 - \phi)(1 + \phi\lambda)$ . Derive the variance of the zero-inflated negative binomial distribution ((2)–(3)).
2. Similarly, show that the variance of the zero-altered Poisson is  $\lambda(1 - \phi)^2$ , and derive the variance of the zero-altered negative binomial distribution.
3. Consider an intercept-only zero-inflated Poisson model. Present a heuristic argument explaining why  $\hat{P}(Y = 0)$  is equal to the sample proportion of zeros.
4. Consider the *zero-altered Poisson* (ZAP) model

$$P(Y = y) = \begin{cases} \phi, & y = 0; \\ (1 - \phi) e^{-\theta} \theta^y / \{(1 - e^{-\theta})y!\}, & y = 1, 2, \dots; \end{cases} \quad (4)$$

which is similar to the zero-inflated Poisson distribution. Given data  $(y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , show that the MLE of  $\phi$  is the number of observations with  $y_i = 0$  divided by  $n$ . Explain how the other parameter  $\theta$  can be estimated by fitting a positive Poisson distribution. Note: the VGAM family function `zapoisson()` fits the ZAP model.

5. If  $\boldsymbol{\theta} = (\phi, \lambda)^T$  in the zero-inflated Poisson distribution, determine the score vector and show that the expected information matrix (EIM) is given by

$$\begin{pmatrix} \frac{1 - e^{-\lambda}}{(1 - \phi)(\phi + (1 - \phi)e^{-\lambda})} & \frac{-e^{-\lambda}}{\phi + (1 - \phi)e^{-\lambda}} \\ \frac{-e^{-\lambda}}{\phi + (1 - \phi)e^{-\lambda}} & \frac{1 - \phi}{\lambda} - \frac{\phi(1 - \phi)e^{-\lambda}}{\phi + (1 - \phi)e^{-\lambda}} \end{pmatrix}.$$

This is used by the VGAM family function `zipoisson()`.

6. Following the same argument as above with the `zipoisson()`, derive the score vector and the EIM for the zero-inflated binomial distribution.
7. The reciprocal of a positive normal random variable has an *alpha distribution*: it has

$$F(y; \alpha, \beta) = \Phi(\alpha - \beta/y) / \Phi(\alpha)$$

for  $\alpha > 0$  and  $\beta > 0$ . The resulting pdf is

$$f(y; \alpha, \beta) = \frac{\beta}{\sqrt{2\pi} y^2 \Phi(\alpha)} \cdot \exp \left\{ -\frac{1}{2} (\alpha - \beta/y)^2 \right\}.$$

Verify the above formulae.

8. Let  $Y$  be defined on a population which contains a subset of elements on which  $Y = 0$ . Denote the mean and variance of the conditional distribution of  $Y$  for  $Y \neq 0$  by  $\mu$  and  $\sigma^2$ , respectively. Then if  $P(Y \neq 0) = p$ , the mean  $\alpha$  and variance  $\beta$ , of  $Y$  are given by

$$\alpha = p\mu \quad (5)$$

$$\beta = p(1 - p)\mu^2 + p\sigma^2. \quad (6)$$

Prove these results.

## References

- Chambers, J. M., Hastie, T. J. (Eds.), 1993. *Statistical Models in S*. Chapman & Hall, New York.
- Coleman, J. S., James, J., 1961. The equilibrium size distribution of freely-forming groups. *Sociometry* 24, 36–45.
- Cullen, M. J., Walsh, J., Nicholson, L. V. B., Harris, J. B., 1990. Ultrastructural localisation of dystrophin in human muscle by using gold immunolabelling. *Proceedings of the Royal Society of London, Series B* 210, 197–210.
- Matthews, J. N. S., Appleton, D. R., 1993. An application of the truncated Poisson distribution to immunogold assay. *Biometrics* 49, 617–621.
- Patil, G. P., 1962. Maximum likelihood estimation for generalised power series distributions and its application to a truncated binomial distribution. *Biometrika* 49, 227–237.
- Thas, O., Rayner, J. C. W., 2005. Smooth tests for the zero-inflated Poisson distribution. *Biometrics* 61 (3), 808–815.
- Yip, P., 1988. Inference about the mean of a Poisson distribution in the presence of a nuisance parameter. *The Australian Journal of Statistics* 30, 299–306.