

# Discrete Distributions

Mikis Stasinopoulos<sup>1</sup>   Bob Rigby<sup>1</sup>

<sup>1</sup>STORM, London Metropolitan University

GAMLSS: Short course Cordoba, Argentina, Oct 2012

- 1 Count data
  - Different (overdispersed) count data approaches
  - Random effect at the observation level
  - Explicit continuous mixture distribution
  - Comparison of the marginal distributions
  - Families modelling the variance-mean relationship
  - Examples
    - A stylometric application
    - The fish species data
  
- 2 Binomial response variables
  - Examples
    - The alveolar-bronchiolar adenomas data
    - The hospital stay data
  
- 3 End

# Different (overdispersed) count data approaches

(i) *Ad-hoc* solutions

- (a) quasi-likelihood (QL), Extended QL
- (b) Efron's Double Exponential
- (c) pseudo-likelihood (PL)

(ii) Discretized continuous distributions

for example if  $F_W(w)$  is the cdf a continuous random variable  $W$  defined in  $\mathfrak{R}^+$  then  $f_Y(y) = F_W(y + 1) - F_W(y)$

(iii) Random effect at the observation level solutions.

$$f_Y(y) = \int f(y|\gamma)f_\gamma(\gamma)d\gamma.$$

## Random effect at the observation level

- (a) when an explicit continuous mixture distribution,  $f_Y(y)$ , exists.
- (b) when a continuous mixture distribution,  $f_Y(y)$ , is not explicit but is approximated by integrating out the random effect using approximations, e.g. Gaussian quadrature or Laplace approximation.
- (c) when a 'non-parametric' mixture (effectively a finite mixture) is assumed for the response variable.

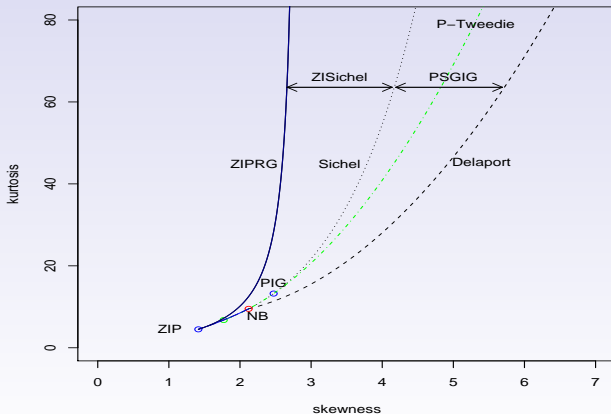
# Explicit continuous mixture distribution

Distributions	R Name	mixing distribution for $\gamma$
Poisson	$PO(\mu)$	-
Neg. bin. I	$NBI(\mu, \sigma)$	$GA(1, \sigma^{\frac{1}{2}})$
Neg. bin. II	$NBII(\mu, \sigma)$	$GA(1, \sigma^{\frac{1}{2}}/\mu)$
Poisson IG	$PIG(\mu, \sigma)$	$IG(1, \sigma^{\frac{1}{2}})$
Sichel	$SICHEL(\mu, \sigma, \nu)$	$GIG(1, \sigma^{\frac{1}{2}}, \nu)$
Delaporte	$DEL(\mu, \sigma, \nu)$	$SG(1, \sigma^{\frac{1}{2}}, \nu)$
Zero inflated Poisson	$ZIP(\mu, \sigma)$	$BI(1, 1 - \sigma)$
Zero inflated Poisson 2	$ZIP2(\mu, \sigma)$	$(1 - \sigma)^{-1}BI(1, 1 - \sigma)$
Zero inflated neg. bin.	$ZINBI(\mu, \sigma, \nu)$	zero inflated gamma
Poisson-Tweedie	-	Tweedie family

Table: Discrete gamlss family distributions for count data

R Name	params	mean	variance
$PO(\mu)$	1	$\mu$	$\mu$
$NBI(\mu, \sigma)$	2	$\mu$	$\mu + \sigma\mu^2$
$NBII(\mu, \sigma)$	2	$\mu$	$\mu + \sigma\mu$
$PIG(\mu, \sigma)$	2	$\mu$	$\mu + \sigma\mu^2$
$SICHEL(\mu, \sigma, \nu)$	3	$\mu$	$\mu + h(\sigma, \nu)\mu^2$
$DEL(\mu, \sigma, \nu)$	3	$\mu$	$\mu + \sigma(1 - \nu)^2\mu^2$
$ZIP(\mu, \sigma)$	2	$(1 - \sigma)\mu$	$(1 - \sigma)\mu + \sigma(1 - \sigma)\mu^2$
$ZIP2(\mu, \sigma)$	2	$\mu$	$\mu + \frac{\sigma}{(1 - \sigma)}\mu^2$

# Comparison of the marginal distributions using a (ratio moment) diagram of their skewness and kurtosis



# Families modelling the variance-mean relationship

$V[Y] = \mu + \mu^2 V[\gamma]$  where  $V[\gamma] = v(\sigma, \nu, \tau)$  is a function of the parameters of the mixing distribution  $f_\gamma(\gamma)$ .

Alternative variance-mean relationship can be obtained by reparametrization.

i.e NB type I  $V[Y] = \mu + \sigma\mu^2$ .

If  $\sigma = \sigma_1/\mu$  then

$V[Y] = (1 + \sigma_1)\mu$  (negative binomial type II)

$\sigma = \sigma_1\mu$  then  $V[Y] = \mu + \sigma_1\mu^3$ .

More generally  $\sigma = \sigma_1\mu^{2-\nu}$  giving  $V(Y) = \mu + \sigma_1\mu^\nu$



# A stylometric application

## Data summary:

R data file: `stylo` in package **gamlss.data** of dimensions  $64 \times 2$

source: Dr Mario Corina-Borja

### variables

`word` : is the number of times a word appears in a single text

`freq` : the frequency of the number of times a word appears in a text

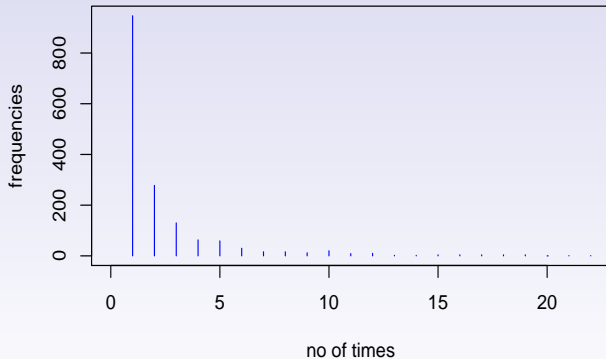
`purpose`: to demonstrate the fitting of a truncated discrete dist.

`conclusion` the truncated SICHEL distributions fits best

# A stylometric application

```
library(gamlss.tr)
data(stylo)
plot(freq ~ word, data = stylo, type = "h", xlim =
+ c(0, 22), xlab = "no of times", ylab =
+ "frequencies", col = "blue")
```

# The stylometric data



## A stylometric application

```
> library(gamlss.tr)
```

```
> gen.trun(par = 0, family = PO, type = "left")
```

A truncated family of distributions from PO has been generated and saved under the names:

```
dP0tr pP0tr qP0tr rP0tr P0tr
```

The type of truncation is left and the truncation parameter

```
> gen.trun(par = 0, family = NBII, type = "left")
```

```
...
```

```
> gen.trun(par = 0, family = DEL, type = "left")
```

```
...
```

```
> gen.trun(par = 0, family = SICHEL, type = "left",
```

```
+   delta = 0.001)
```

```
...
```

# A stylometric application

```

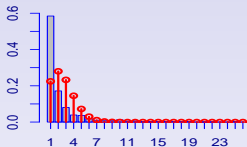
> mPO <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = P0tr, trace = FALSE)
> mNBII <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = NBIItr, n.cyc = 50, trace = FALSE)
> mDEL <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = DELtr, n.cyc = 50, trace = FALSE)
> mSI <- gamlss(word ~ 1, weights = freq, data = stylo,
+ family = SICHELtr, n.cyc = 50, trace = FALSE)
> GAIC(mPO, mNBII, mDEL, mSI)

```

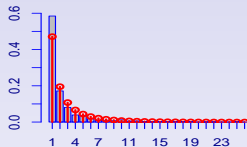
	df	AIC
mSI	3	5148.454
mDEL	3	5160.581
mNBII	2	5311.627
mPO	1	9207.459

# The stylometric data

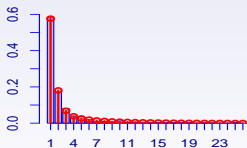
(b) Poisson



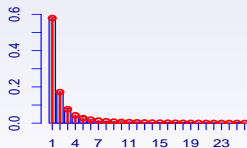
(c) negative binomial II



(c) Delaporte



(d) Sichel



# The fish species data

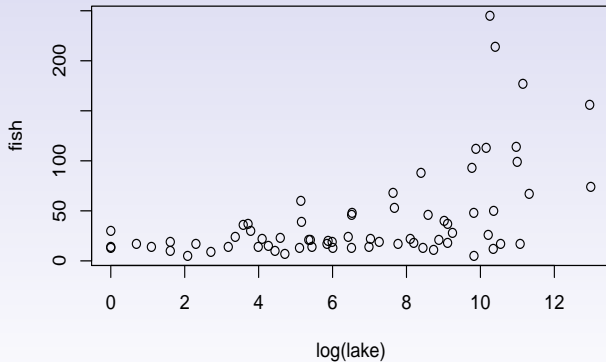
**Data summary:** the fish species data

**R data file:** species in package gamlss.data of dimensions  $70 \times 2$   
**variables**

**fish** : the number of different species in 70  
lakes in the world

**lake** : the lake area

# The fish species data





# The fish species data

There are several questions that need to be answered.

- How does the mean of  $y$  depend on  $x$ ?
- Is  $y$  overdispersed Poisson?
- How does the variance  $y$  depend on its mean?
- What is the distribution of  $y$  given  $x$ ?
- Do the scale and shape parameters of the distribution of  $y$  depend on  $x$ ?

# Overdispersed count data approaches

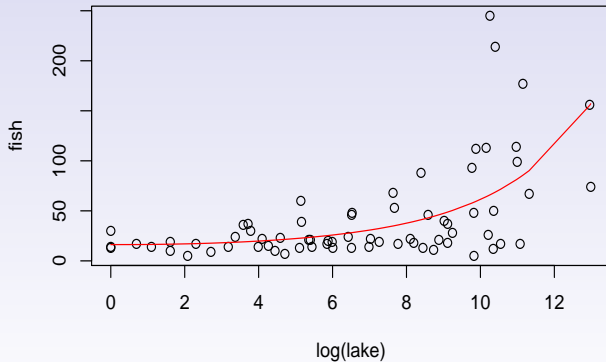
Table: Comparison of models for the fish species data

Model	$f_Y(y)$	$\mu$	$\sigma$	$\nu$	$DEV$	$df$	$AIC$	$SBC$
1	PO	$x < 2 >$	-	-	1849.3	3	1855.3	1862.0
2	NBI	$x$	1	-	619.8	3	625.8	632.6
3	NBI	$x < 2 >$	1	-	614.3	4	622.3	631.3
4	NBI	$cs(x, 3)$	1	-	611.9	6	623.9	637.4
5	NBI	$x < 2 >$	$x$	-	605.0	5	615.0	626.2
6	NBI-fam	$x < 2 >$	1	1	606.0	5	616.0	627.3
7	NBI-fam	$x < 2 >$	$x$	1	604.9	6	616.9	630.4

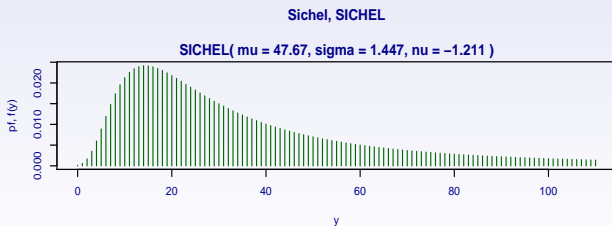
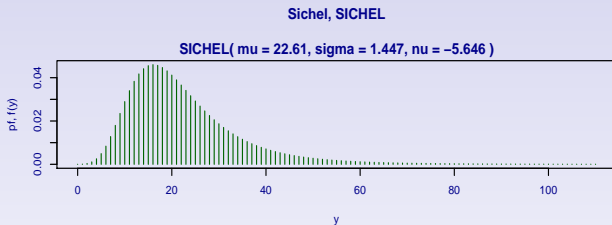
## Overdispersed count data approaches

Model	$f_Y(y)$	$\mu$	$\sigma$	$\nu$	$DEV$	$df$	$AIC$	$SBC$
8	PIG	$x < 2 >$	1	-	613.3	4	621.3	630.3
9	SI	$x < 2 >$	1	$x$	597.7	6	609.7	623.2
10	DEL	$x < 2 >$	1	$x$	600.6	6	612.6	626.1
11	DEL	$x < 2 >$	-	$x$	600.6	5	610.6	621.9
12	PO-Normal	$x < 2 >$	1	-	615.2	4	623.2	632.2
13	NBI-Normal	$x < 2 >$	$x$	1	603.7	6	615.7	629.2
14	PO-NPFM(5)	$x < 2 >$	-	-	601.9	13	627.9	657.2
15	NB-NPFM(2)	$x < 2 >$	1	-	611.9	6	623.9	637.4
16	doublePO	$x < 2 >$	$x$	-	616.4	5	626.4	637.6
17	IGdisc	$x < 2 >$	1	-	603.3	4	611.3	620.3

# Fitted mean of the Sichel distribution



# Fitted Sichel distributions for observations (a) 40 and (b) 67



# Binomial response variables

There are only two distributions here

- binomial
- beta binomial

# The alveolar-bronchiolar adenomas data

## Data summary:

**R data file:** alveolar in package **gamlss.data** of dimensions  $23 \times 2$

**source:** Tamura and Young (1987), and Hand *et al.* (1994)

### variables

**r** : number of mice having alveolar-bronchiolar

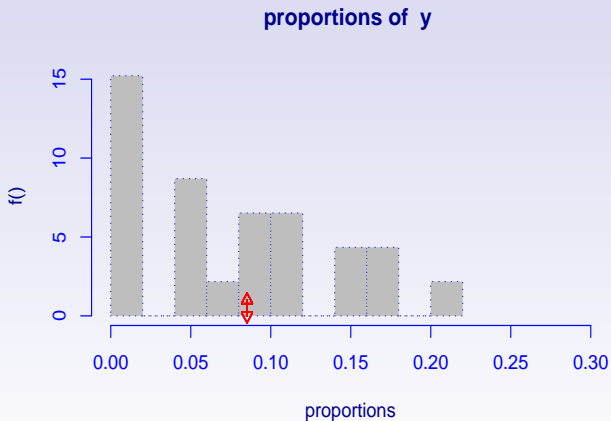
**n** : total number of mice

## The alveolar-bronchiolar adenomas data

```
> data(alveolar)
> alveolar$y <- with(alveolar, cbind(r, n - r))
> m1 <- gamlss(y ~ 1, data = alveolar, family = BI)
> m2 <- gamlss(y ~ 1, data = alveolar, family = BB)
> GAIC(m1, m2)
df AIC
m1 1 73.1292
m2 2 75.0665
> m3 <- with(alveolar, histDist(y, "BI", xlim = c(0, 0.3)))
```



# The alveolar-bronchiolar adenomas data



# The hospital stay data

## Data summary:

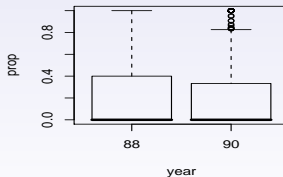
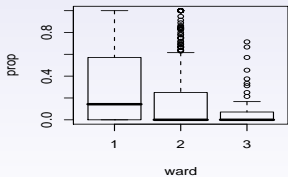
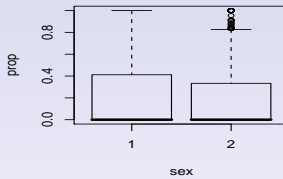
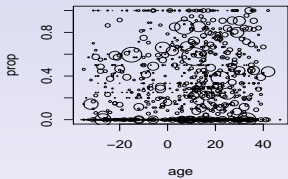
R data file: aep in package **gamlss** of dimensions  $1383 \times 8$

source: Gange *et al.* (1996)

### variables

- `los` : total number of days
- `noinap` : number of inappropriate days patient stay in hospital
- `loglos` : the log of `los/10`
- `sex` : the gender of patient
- `ward` : type of ward in the hospital (medical, surgical)
- `year` : 1988 or 1990
- `age` : age of the patient subtracted from 55
- `y` : the response variable, a matrix with columns `(noinap, los-noinap)`

# The hospital stay data



## The hospital stay data

```

> mI <- gamlss(y ~ ward + year + loglos, sigma.fo = ~year,
+ family = BB, data = aep)
> mII <- gamlss(y ~ ward + year + loglos, sigma.fo = ~year
+ ward, family = BB, data = aep)
> mIII <- gamlss(y ~ ward + year + cs(loglos, 1),
+ sigma.fo = ~year+ward, family = BB, data = aep)
> mIV <- gamlss(y ~ ward + year + cs(loglos, 1) + cs(age, 1)
+ sigma.fo = ~year + ward, family = BB, data = aep)
> GAIC(mI, mII, mIII, mIV, k = 0)

```

	df	AIC
mIV	12.00010	4454.362
mIII	10.00045	4459.427
mII	9.00000	4483.020
mI	7.00000	4519.441

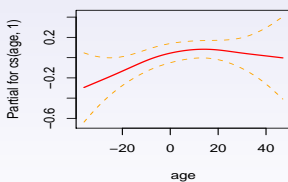
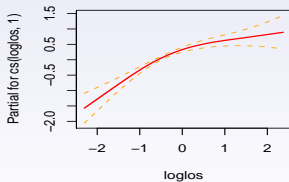
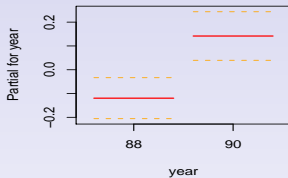
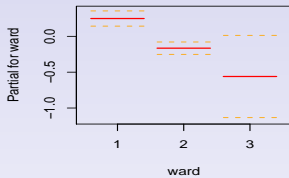
# The hospital stay data

Models	Links	Terms	GD (AIC) [SBC]
I	logit( $\mu$ ) log( $\sigma$ )	1+ward+loglos+year 1+year	4519.4 (4533.4) [4570.1]
II	logit( $\mu$ ) log( $\sigma$ )	1+ward+loglos+year 1+year+ward	4483.0 (4501.0) [4548.1]
III	logit( $\mu$ ) log( $\sigma$ )	1+ward+cs(loglos,1)+year 1+year+ward	4459.4 (4479.4) [4531.8]
IV	logit( $\mu$ ) log( $\sigma$ )	1+ward+cs(loglos,1)+year+cs(age,1) 1+year+ward	4454.4 (4478.4) [4541.2]

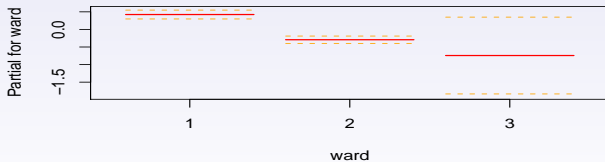
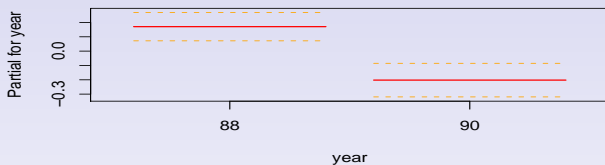
## The hospital stay data

```
> op <- par(mfrow = c(2, 2))
> term.plot(mIV, se = T)
> par(op)
> op <- par(mfrow = c(2, 1))
> term.plot(mIV, "sigma", se = T)
> par(op)
> rqres.plot(mIV)
```

# The hospital stay data

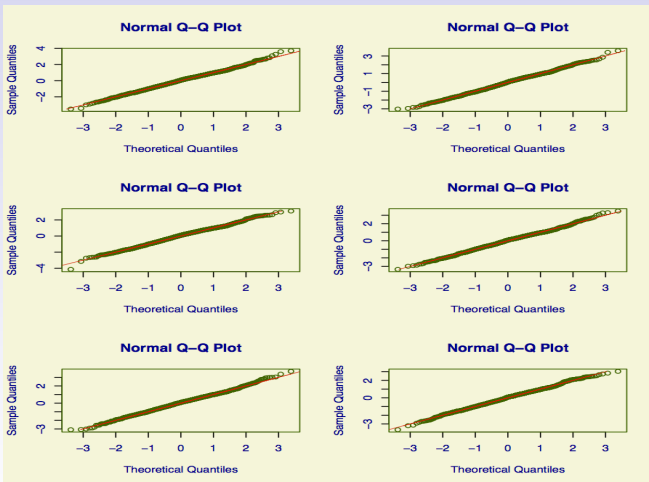


# The hospital stay data





# The hospital stay data



# END

for more information see

[www.gamlss.org](http://www.gamlss.org)