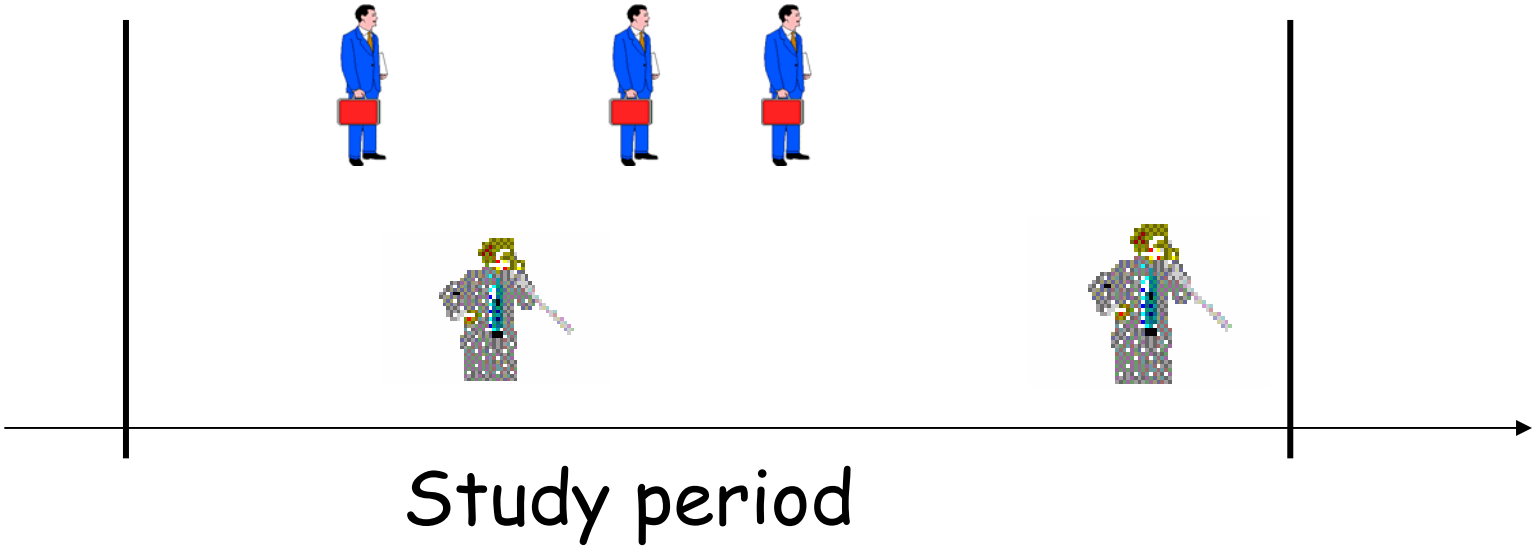
Nonparametric Mixture Maximum Likelihood Truncated Count Distributions

Dankmar Böhning
Professor and Chair in Applied Statistics
School of Biological Sciences
University of Reading, UK

Round Table: Microbial Biodiversity Research
11th International Symposium on Microbial Ecology
20-25 August 2006

Counts of capture-recaptures as outcome of continous time CR-experiment

- CR of Wildlife Populations
- CR in Public Health and Surveillance



Situation in Continuous CR Experiment

$f_1, f_2, f_3, \ldots, f_m$

frequencies of units identified $1, 2, 3, \ldots, m$ times

f_0 is unobserved

population size: $N = f_0 + f_1 + \ldots + f_m = f_0 + n$

if probability p_0 for zero-count known:

$$N = Np_0 + n \Rightarrow \hat{N} = n/(1 - p_0)$$

Illustration: Project with DEFRA on Scrapie in the UK 2002

$$f_1, f_2, f_3, \ldots, f_m$$

frequencies of holdings reporting $1, 2, 3, \ldots, m$ cases:

$$f_1 = 74,\ f_2 = 23,\ f_3 = 15, \ldots,\ f_{11} = 3,\ f_{12} + = 7$$

f_0 is number of hidden holdings with scrapies

adjusted size of scrapie: $N = f_0 + n = f_0 + 177$

Idea of Modelling

$$f_0, f_1, f_2, f_3, \ldots, f_m$$

look at associated probabilities:

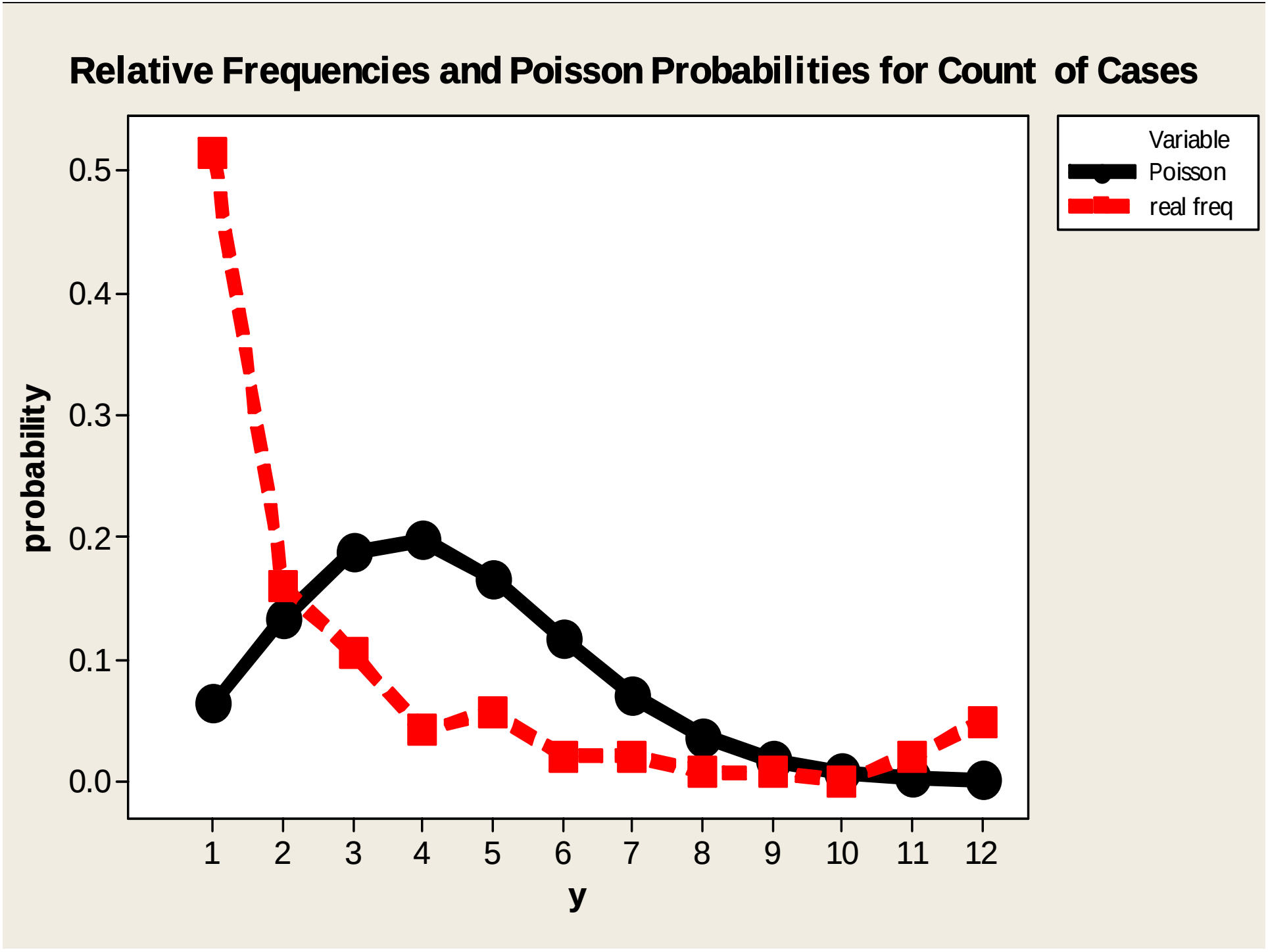
$$p_0, p_1, p_2, p_3, \ldots, p_m$$

and choose a model (Poisson)

$$p_0 = e^{-\theta}, p_1 = e^{-\theta}\theta, p_2 = e^{-\theta}\theta^2 / 2 , \ldots,$$

estimate θ with $\hat{\theta}$, get $\hat{p}_0 = e^{-\hat{\theta}}$

$$\hat{N} = n/(1-\hat{p}_0)$$


Relative Frequencies and Poisson Probabilities for Count of Cases
probability
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10
11
12
y
Variable
Poisson
real freq

Idea of Mixed Modelling

instead of simple Poisson

$$p_j = e^{-\theta}\theta^j / j!$$

look at mixed Poisson:

$$p_j = \int_0^\infty e^{-\theta}\theta^j / j!\, f(\theta)d\theta$$

(to capture heterogeneity in θ)

Idea of Mixed Modelling

possible to estimate $f(\theta)$ in mixed Poisson:

$$\hat{p}_j = \int_0^\infty e^{-\theta}\theta^j / j!\, \hat{f}(\theta)d\theta$$

by means of nonparametric mixtures

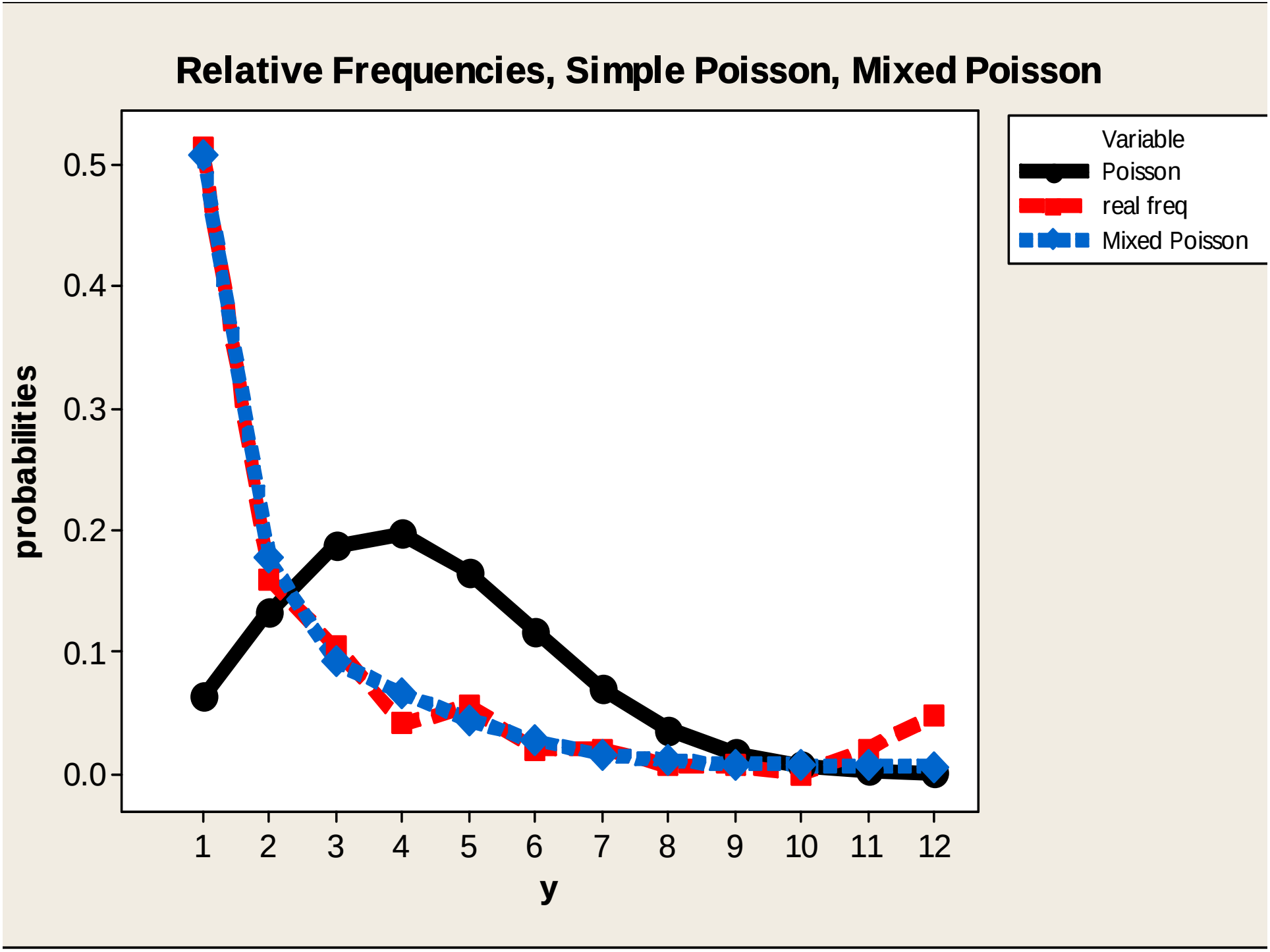
Estimate of $f(\theta)$ for Scrapie 2002

Poisson mixture

$$\hat{p}_j = \int_0^{\infty} e^{-\theta}\theta^j / j!\, \hat{f}(\theta)d\theta = \sum_{l=1}^{k} e^{-\theta_l}\hat{\theta}_l^{\,j} / j!\hat{f}_l$$

we find

$$\hat{f}(\theta) = \begin{array}{c} \begin{array}{cc} \hat{\theta}_l & \hat{f}_l \end{array} \\ \begin{pmatrix} .475 & 0.7955 \\ 3.406 & 0.1594 \\ 10.362 & 0.0277 \\ 22.367 & 0.0173 \end{pmatrix} \end{array}$$


Relative Frequencies, Simple Poisson, Mixed Poisson
probabilities
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10
11
12
y
Variable
Poisson
real freq
Mixed Poisson

Estimates of hidden Scrapie 2002

Simple Poisson:

$$\hat{N} = n/(1-e^{-\hat{\theta}}) = 144 + 2$$

Mixed Poisson :

$$\hat{N} = n/(1-\sum_{l=1}^{4} e^{-\theta_l}\ \hat{f}_l) = 144 + 143$$

Two other nonparametric ideas

- Zelterman‘s robustness idea
- Chao‘s lower bound estimate

Idea of Zelterman

look at count probabilities:

$$p_0, p_1, p_2, p_3, \ldots, p_m$$

and modelled by Poisson

$$p_0 = e^{-\theta}, p_1 = e^{-\theta}\theta, p_2 = e^{-\theta}\theta^2/2, \ldots,$$

ratios of consecutive Poissons

$$p_1/p_0 = \theta,\ p_2/p_1 = \theta/2,\ p_3/p_2 = \theta/3 \ldots,$$

$$\Rightarrow \hat{\theta} = 2f_2/f_1$$

$$\hat{N}_Z = n/(1-\exp(-2f_2/f_1))$$

Idea of Chao

look at mixed Poisson:

$$p_j = \int_0^\infty e^{-\theta}\theta^j / j!\, f(\theta)d\theta$$

Cauchy-Schwartz:

$$\left(\int_0^\infty e^{-\theta}\theta f(\theta)d\theta\right)^2 \le \int_0^\infty e^{-\theta} f(\theta)d\theta \int_0^\infty e^{-\theta}\theta^2 f(\theta)d\theta$$

$$p_1^2 \le p_0 p_2 \Rightarrow f_0 \ge f_1^{\,2}/(2 f_2)$$

Chao's lower bound estimate

Estimates of hidden Scrapie 2002

Simple Poisson:

$$\hat{N} = n/(1-e^{-\hat{\theta}}) = 144 + 2$$

Mixed Poisson :

$$\hat{N} = n/(1-\sum_{l=1}^{4} e^{-\theta_l}\ \hat{f}_l) = 144 + 143$$

Zelterman:

$$\hat{N}_Z = n/(1-e^{-2\,f_2/\,f_1}) = 144 + 167$$

Chao:

$$\hat{N}_C = n + f_1^{\,2}/(2\,f_2) = 144 + 119$$

Comparing Zelterman and Chao Simulation: 4 Experiments with Poisson Model Violation

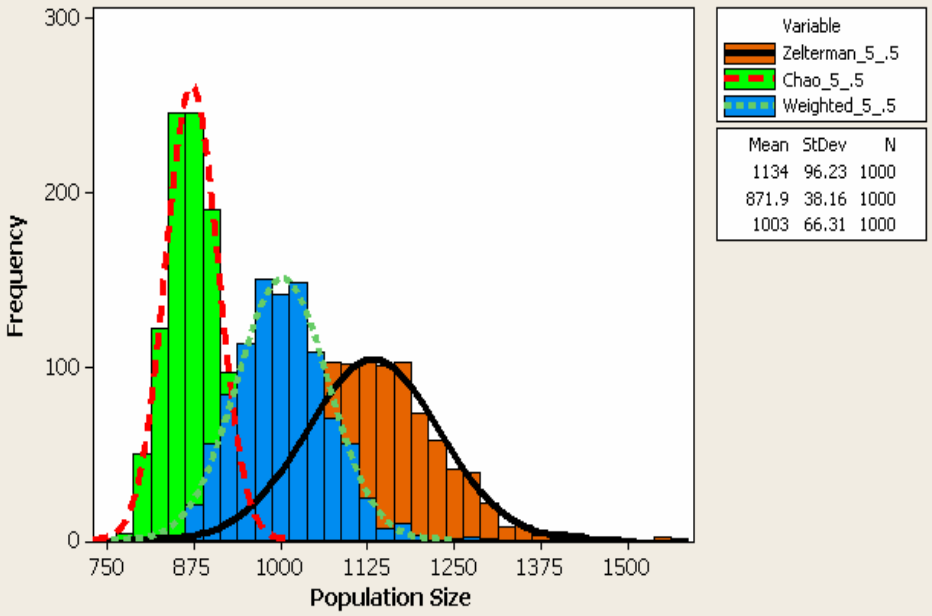
1: $f(\theta) = Po(0.5)0.5 + 0.5Po(5)$

2: $f(\theta) = Po(0.5)0.9 + 0.1Po(5)$

3: $f(\theta) = Po(0.5)0.5 + 0.5Po(1)$

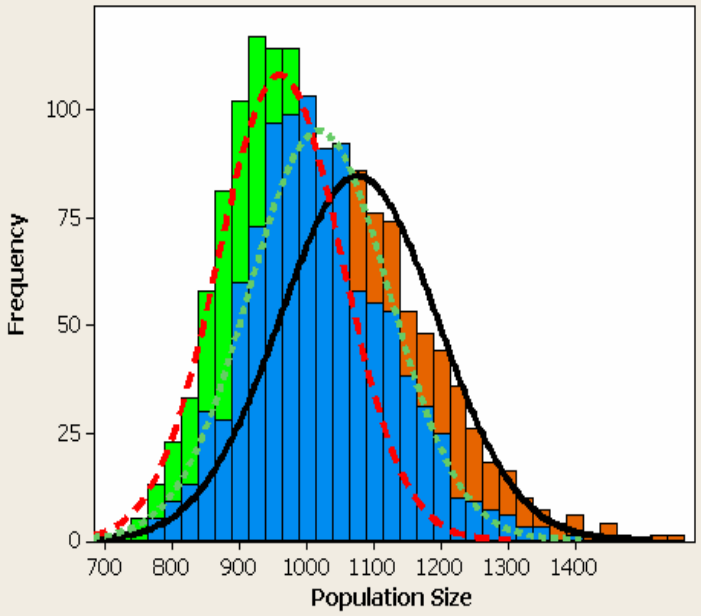
4: $f(\theta) = Po(0.5)0.9 + 0.1Po(1)$

Experiment 1



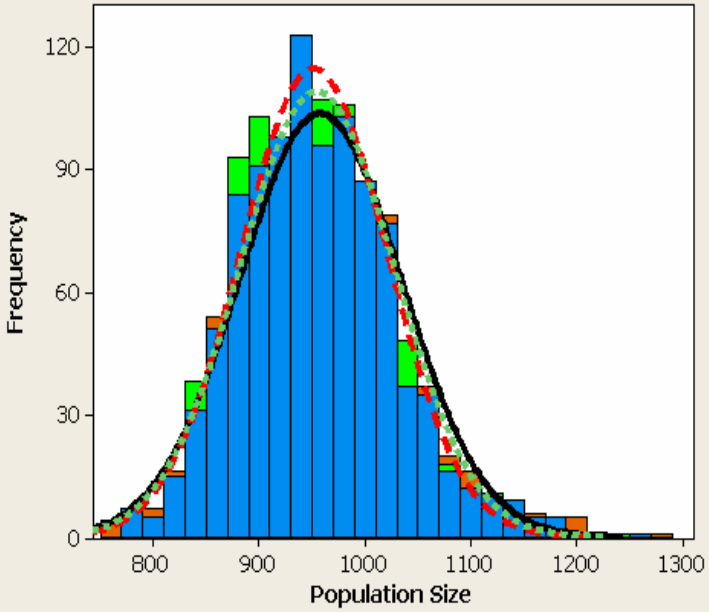
Variable	Mean	StDev	N
Zelterman_5_.5	1134	96.23	1000
Chao_5_.5	871.9	38.16	1000
Weighted_5_.5	1003	66.31	1000

Experiment 2



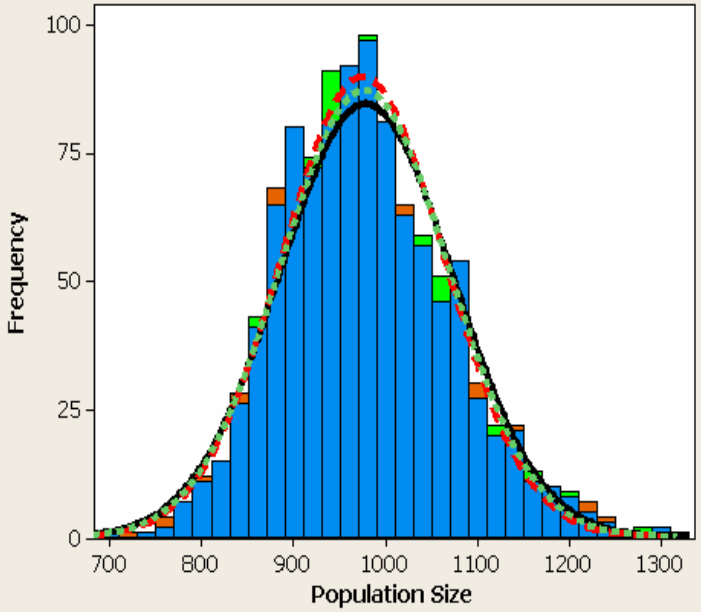
Variable	Mean	StDev	N
Zelterman_5_.1	1077	118.1	1000
Chao_5_.1	960.0	92.38	1000
Weighted_5_.1	1018	105.1	1000

Experiment 3



Variable	Mean	StDev	N
Zelterman_1_.5	957.9	76.97	1000
Chao_1_.5	952.0	69.56	1000
Weighted_1_.5	954.9	73.18	1000

Experiment 4



Variable	Mean	StDev	N
Zelterman_1_.1	978.6	94.37	1000
Chao_1_.1	975.6	88.83	1000
Weighted_1_.1	977.1	91.53	1000

Conclusions

- Chao‘s estimator keeps the lower bound
- New Weighted estimator: average between Zelterman and Chao?

- For more valid inference, mixed modelling can‘t be avoided

Key-References

Böhning, D. and Kuhnert, R. (2006). The Equivalence of Truncated Count Mixture Distributions and Mixtures of Truncated Count Distributions. *Biometrics* (to appear).

Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of the population size based upon the counting distribution. *Journal of the Royal Statistical Society, Series C, Applied Statistics* **54**, 721-737.

Böhning, D. and Patilea, V. (2005). Asymptotic Normality in Mixtures of Power Series Distributions. *Scandinavian Journal of Statistics* **32**, 115-132.

Papers download at (also copy of this talk):

www.reading.ac.uk/~sns05dab