Capture-Recapture Methodology in the Biological and Health Sciences – an Approach Based upon Generalized Chao Bounds

Presentation at Maejo University Chiang Mai, 22. August 2007

Dankmar Böhning

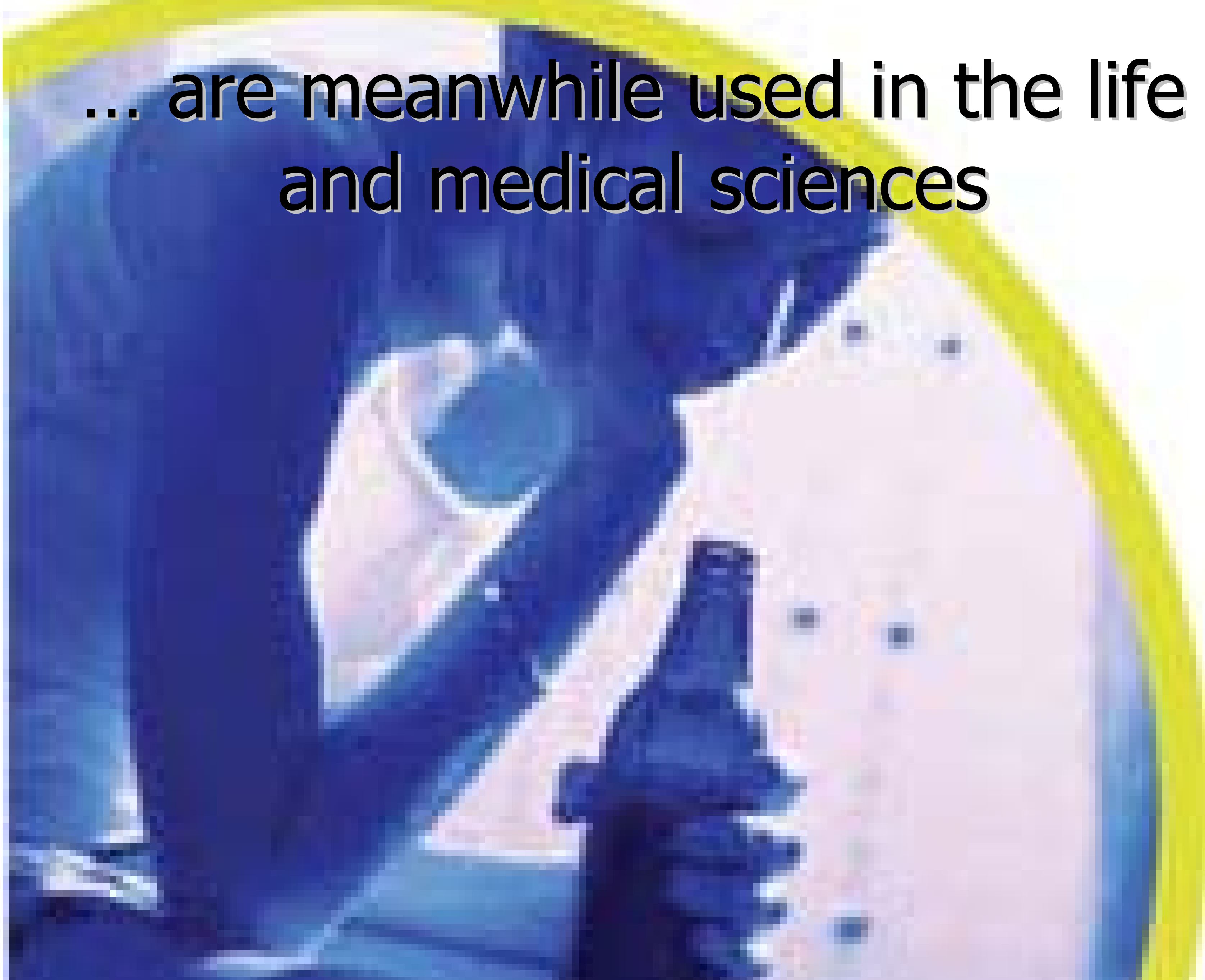
Applied Statistics

School of Biological Sciences

The University of Reading

Capture-Recapture experiments come from Biological Sciences

… are meanwhile used in the life and medical sciences



… as well as in the social sciences

Objective

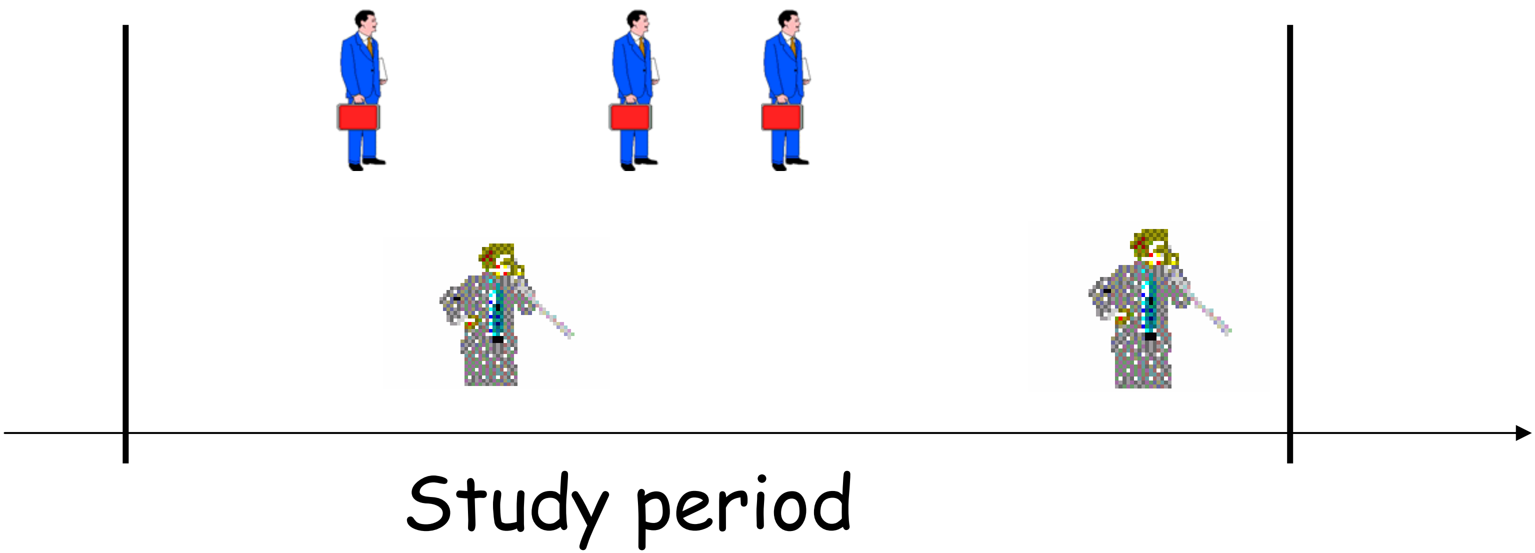
- develop a population size estimator using capture-recaputre techniques
- interest in population size estimator which is valid under a wider range of scenarios

Overview

- Introduction
- Chao,s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II
- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman`s Estimator as an Upper Bound
  - Generalising Zelterman
- A Simulation Study

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, ..., 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 on illicit drug use in Bangkok 2001 (4th Quarter)

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

frequencies of drug users with 1, 2, 3, ..., m contacts to treatment institutions (hospitals):

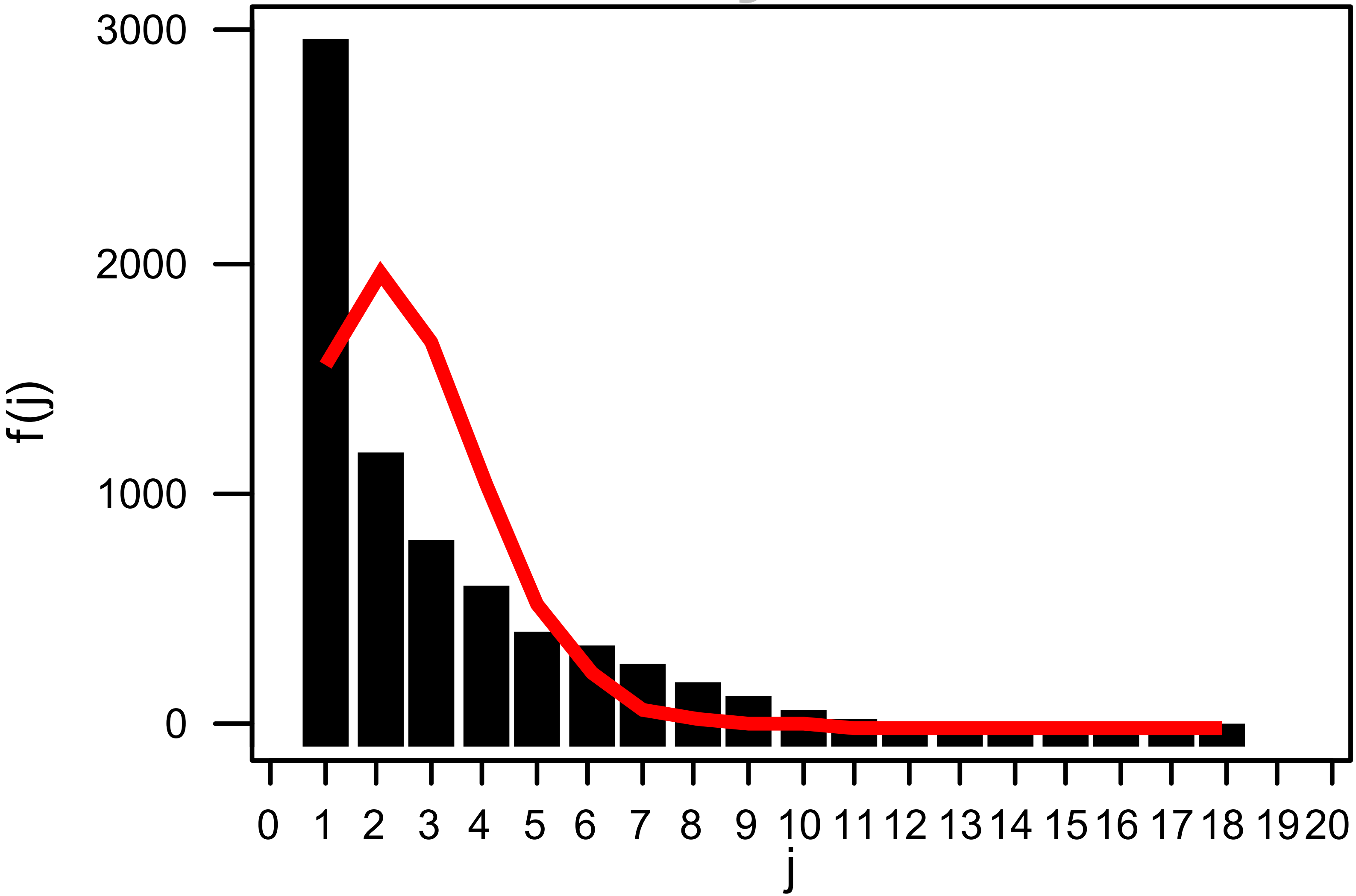
$$f_1 = 2955, f_2 = 1186, f_3 = 803, f_4 = 611, \ldots$$

f_0 is number of hidden (unseen) drug users

adjusted size of drug user population:

$$N = f_0 + n = f_0 + 6966$$

Frequency Distribution of BKK-Drug Users with j Contacts



Idea of Modelling

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

look at associated probabilities:

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

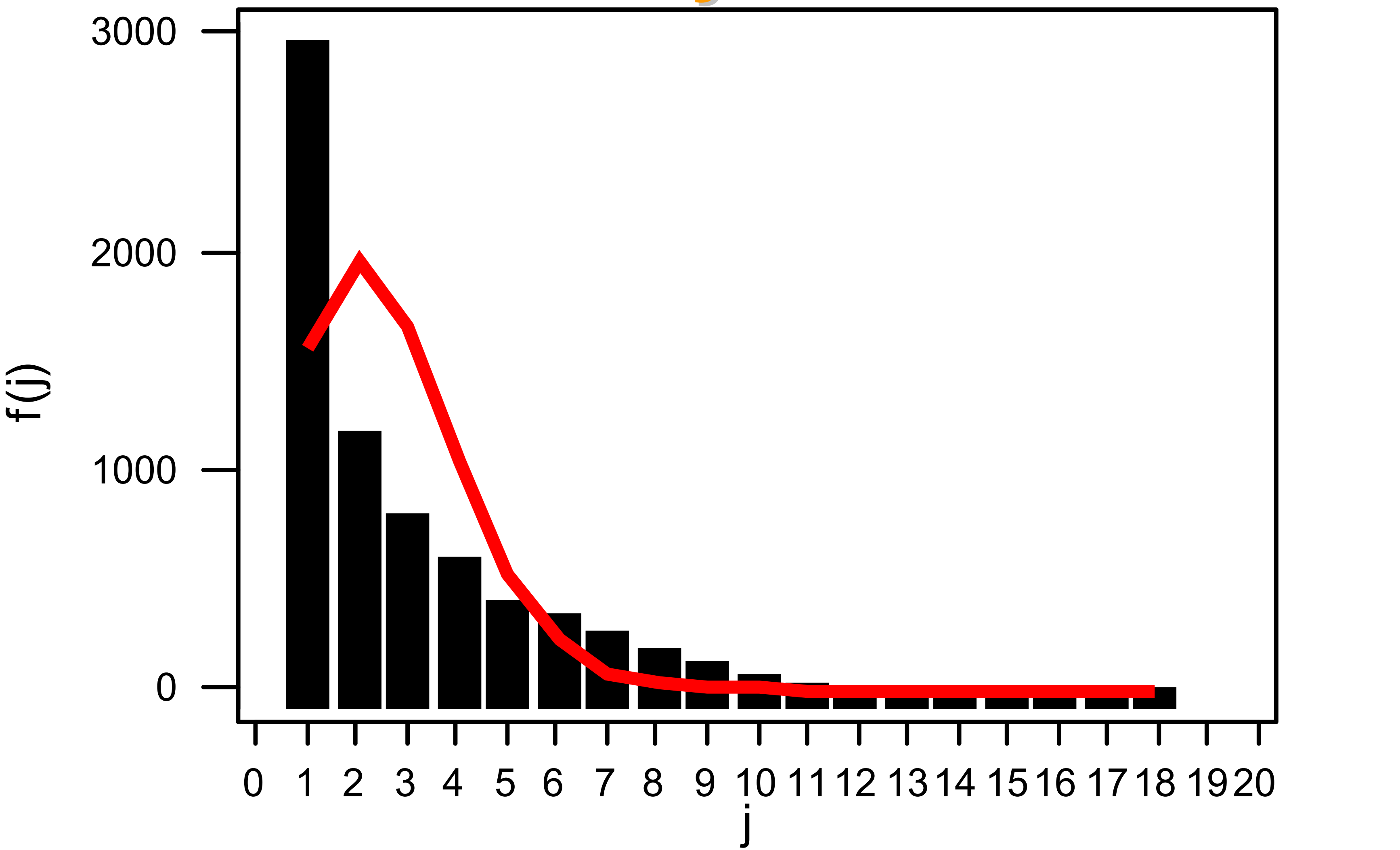
and choose a model (Poisson)

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

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

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

Frequency Distribution of BKK-Drug Users with j Contacts



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 Chao

ook at mixed Poisson:

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

Cauchy-Schwartz: $[E(XY)]^2 \leq E(X^2)E(Y^2)$

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

with $x = \sqrt{e^{-\theta}}$ and $y = \sqrt{e^{-\theta}}\theta$

Idea of Chao

ook at mixed Poisson:

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

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

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

Chao's lower bound estimate

Extending the idea of Chao: way I

ook at mixed Poisson:

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

Cauchy-Schwartz: $[E(XY)]^2 \le E(X^2)E(Y^2)$

$$\left(\int_0^\infty e^{-\theta}\theta^j f(\theta)d\theta\right)^2 \le \int_0^\infty e^{-\theta}\theta^{j-1} f(\theta)d\theta \int_0^\infty e^{-\theta}\theta^{j+1} f(\theta)d$$

with $x = \sqrt{e^{-\theta}\theta^{j-1}}$ and $y = \sqrt{e^{-\theta}\theta^{j+1}}$

Extending the idea of Chao: way I

ook at mixed Poisson:

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

$$\left(\int_0^\infty e^{-\theta}\theta^j f(\theta)d\theta\right)^2 \leq \int_0^\infty e^{-\theta}\theta^{j-1} f(\theta)d\theta \int_0^\infty e^{-\theta}\theta^{j+1} f(\theta)d\theta$$

$$(j! \times p_j)^2 \leq (j-1)!p_{j-1} \times (j+1)!p_{j+1}$$

$$\frac{j \times p_j}{p_{j-1}} \leq \frac{(j+1)p_{j+1}}{p_j}$$

Extending the idea of Chao: way I

$$\frac{j\, p_j}{p_{j-1}} \leq \frac{(j+1)\, p_{j+1}}{p_j}$$

so ... ratios of mixed Poissons are monotone non-decreasing with increasing j

Extending the idea of Chao: way I- a new diagnostic device

monotone pattern should be visible

$$\frac{j \times p_j}{p_{j-1}} \leq \frac{(j+1)p_{j+1}}{p_j}$$

when replacing p_j by f_j :

$$\frac{j \times f_j}{f_{j-1}} \leq \frac{(j+1)f_{j+1}}{f_j}$$

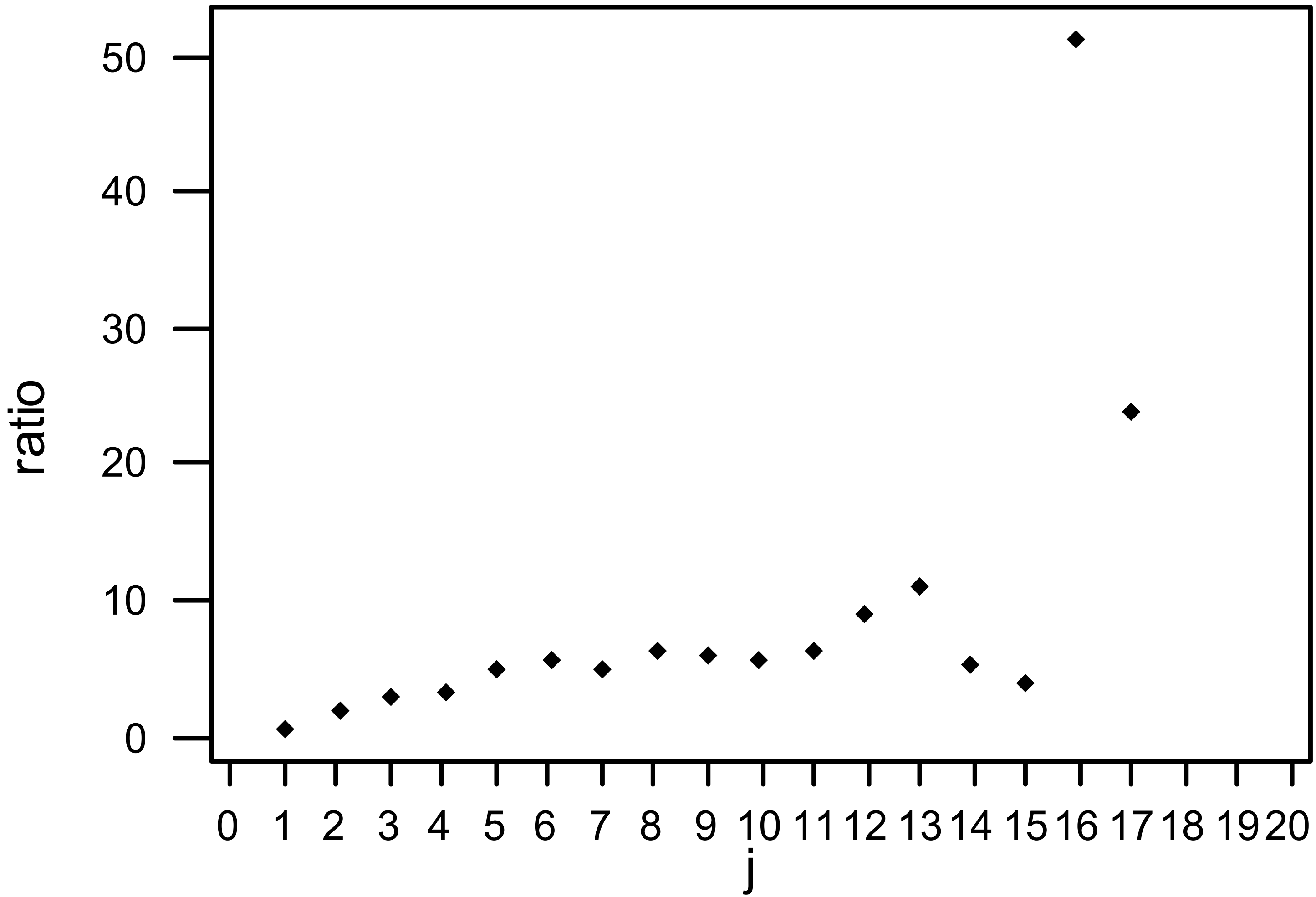
monotone non-decreasing with increasing j

A new diagnostic device for heterogeneity: some examples

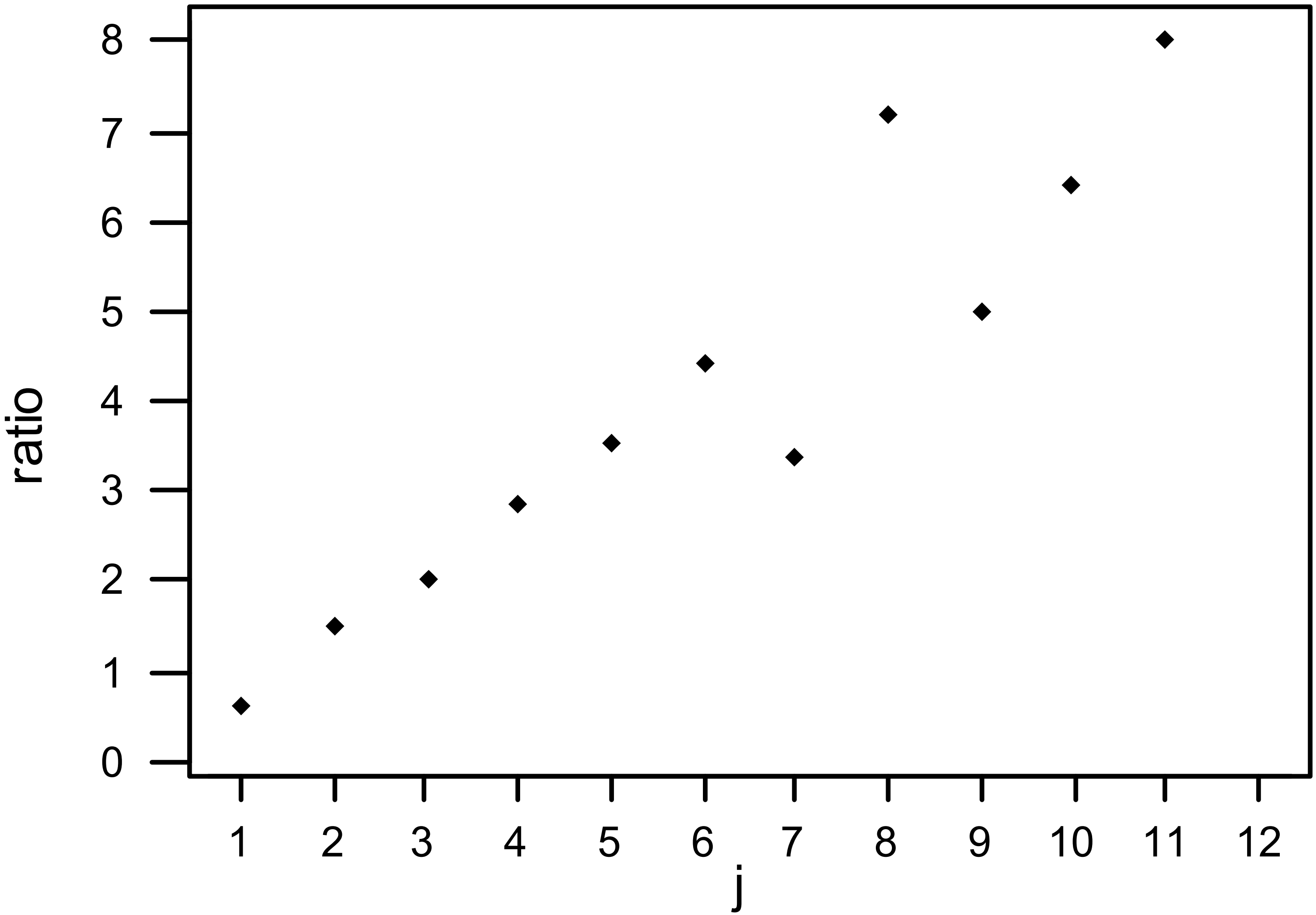
$$\text{graph: } j \ \rightarrow \text{ratio} = \frac{(j+1) f_{j+1}}{f_j}$$

- Drug user data Bangkok (1/4 year)
- Drug user data L.A. (Hser 1992)
- Drug user data Scotland (Hay and Smit 2003)

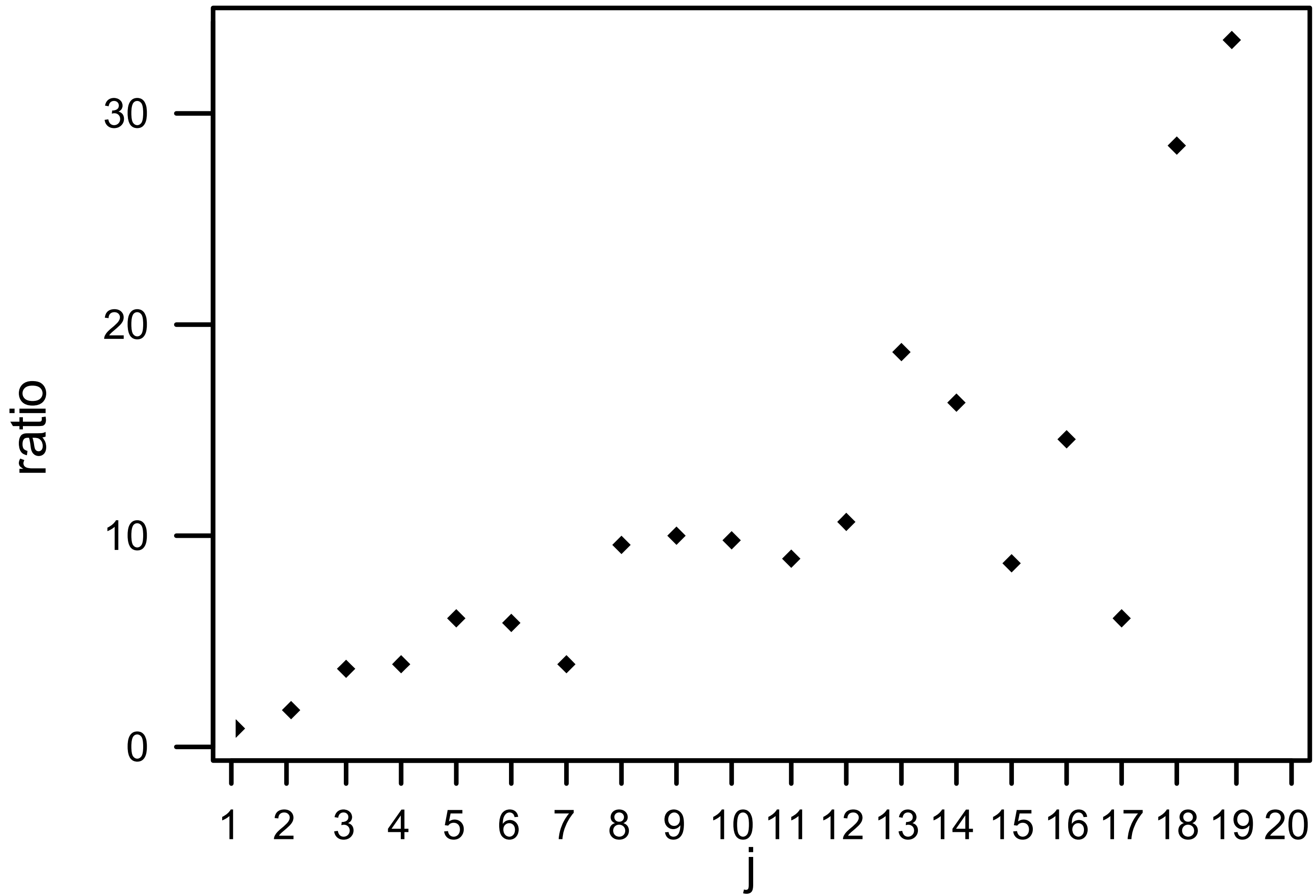
Ratio for BKK Drug User Data



Ratio for L.A. Drug User Data



Ratio for Scottish Drug User Data



Conclusion

Ratio plot seems to work as a diagnostic device for presence of a mixed Poisson

	f_1	f_2	n	$\hat{f}_0$	$\hat{N} = \hat{f}_0 + n$	$n / \hat{N}$
BKK:	2955	1186	6966	3681	10647	0.65
LA:	11982	3893	20198	18439	38637	0.52
Scotl.:	175	85	647	180	827	0.78

Extending the idea of Chao: way II

from mixed Poisson to mixed Power series distribution:

$$p_j = \int_0^\infty e^{-\theta}\theta^j / j! f(\theta)d\theta \rightarrow \quad p_j = \int_0^\infty \mu(\theta)\theta^j a_j f(\theta)d\theta$$

Similar Results!

Extending the idea of Chao: way II

mixed Power series $p_j = \int_0^\infty \mu(\theta)\theta^j a_j f(\theta)d\theta$:

$$\frac{p_j / a_j}{p_{j-1} / a_{j-1}} \leq \frac{p_{j+1} / a_{j+1}}{p_j / a_j}$$

replace again

$$\frac{f_j / a_j}{f_{j-1} / a_{j-1}} \leq \frac{f_{j+1} / a_{j+1}}{f_j / a_j}$$

Extending the idea of Chao: way II: a diagnostic device for the Power series distribution

plot

$$j \rightarrow \frac{f_{j+1} / a_{j+1}}{f_j / a_j}$$

and see if pattern monotone

… by the way: generalised Chao bound

$$\frac{p_1 / a_1}{p_0 / a_0} \leq \frac{p_2 / a_2}{p_1 / a_1}$$

$$\frac{(p_1 / a_1)^2 a_0}{p_2 / a_2} \leq p_0$$

replace again by observed frequemcies

$$\hat{f}_0 = \frac{(f_1 / a_1)^2 a_0}{f_2 / a_2}$$

An example: mixed binomial

Binomial with size parameter m :

$$\binom{m}{j}\theta^{j}(1+\theta)^{-m} = \binom{m}{j}p^{j}(1-p)^{m-j}$$

so that $a_j = \binom{m}{j}$ and $\mu(\theta) = (1+\theta)^{-m}$

… by the way: generalised Chao bound

$$\hat{f}_0 = \frac{(f_1 / a_1)^2 a_0}{f_2 / a_2}$$

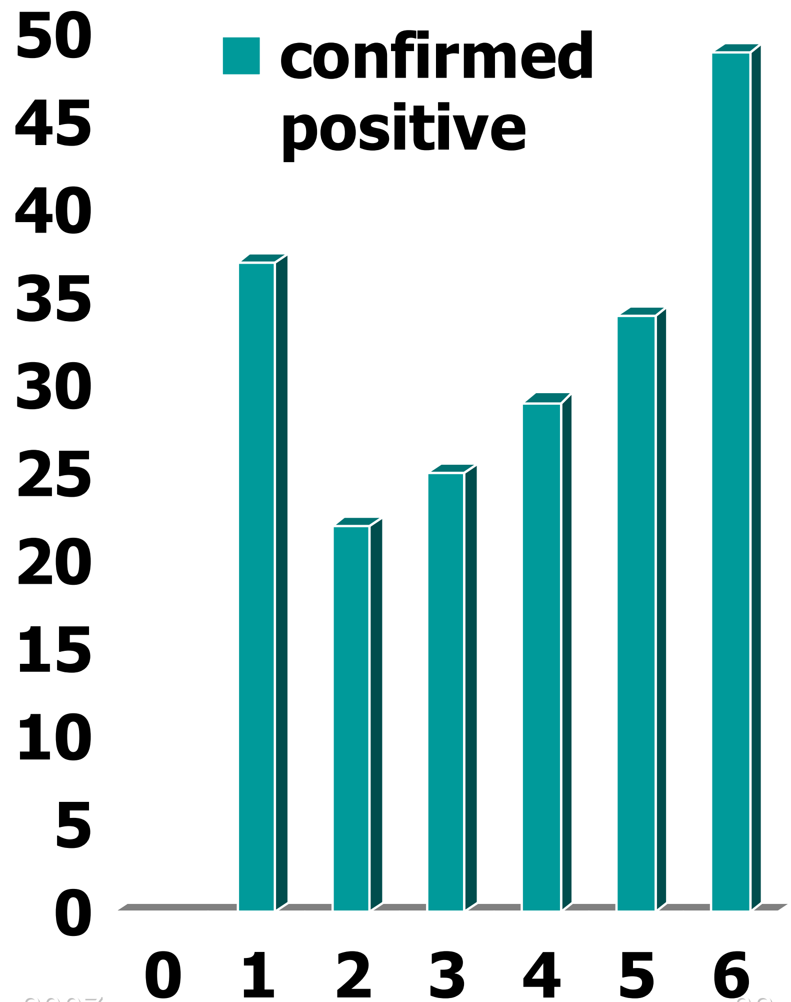
$$= \frac{f_1^{\;2}}{2 f_2} \frac{(m-1)}{m}$$

Exemplified at a recent example from screening

- Lloyd & Frommer (2004, Applied Statistics) screening for bowel cancer
- 38,000 men screened in Sidney at 6 consecutive days by means of self-tesing for blood in stools
- 3,000 tested positively a least once and cancer status evaluated
- 196 were confirmed positive to have bowel cancer
- How many of 35,000 unconfirmed negative have bowel cancer?

The counting distribution: a recent example from screening

- frequency f_0 of those tested negative at all 6 times with bowel cancer is unknown
- an estimate of f_0 might be constructed from the distribution $f_1, f_2, f_3 \ldots$ of counts

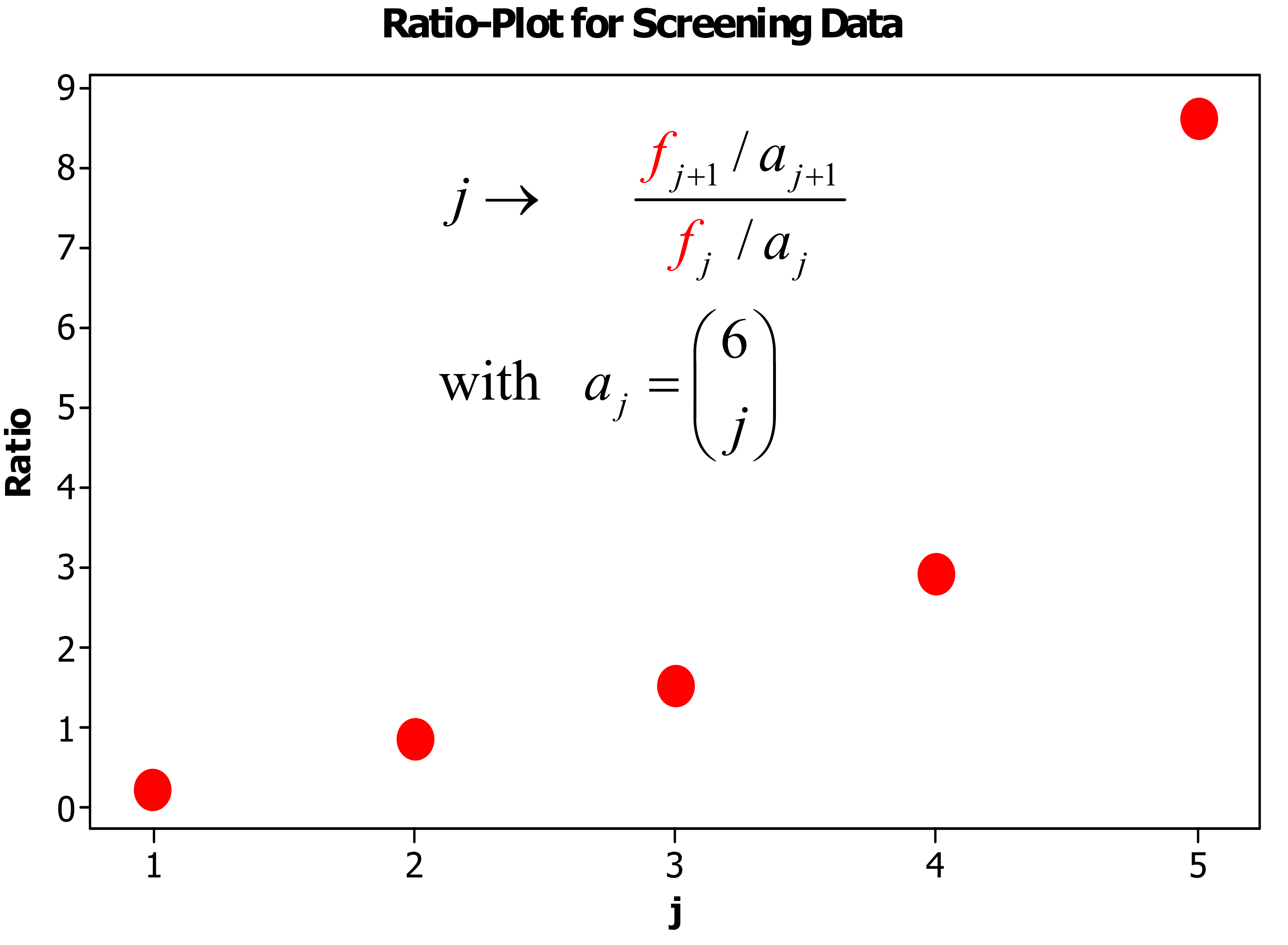


mixed binomial

binomial with size parameter 6 :

$$\binom{6}{j}\theta^{j}(1+\theta)^{-6}$$

so that $a_j = \binom{6}{j}$ and $\mu(\theta) = (1+\theta)^{-6}$



Ratio-Plot for Screening Data
$j \rightarrow \frac{f_{j+1} / a_{j+1}}{f_j / a_j}$
with $a_j = \binom{6}{j}$
Ratio
j
0
1
2
3
4
5
6
7
8
9
1
2
3
4
5

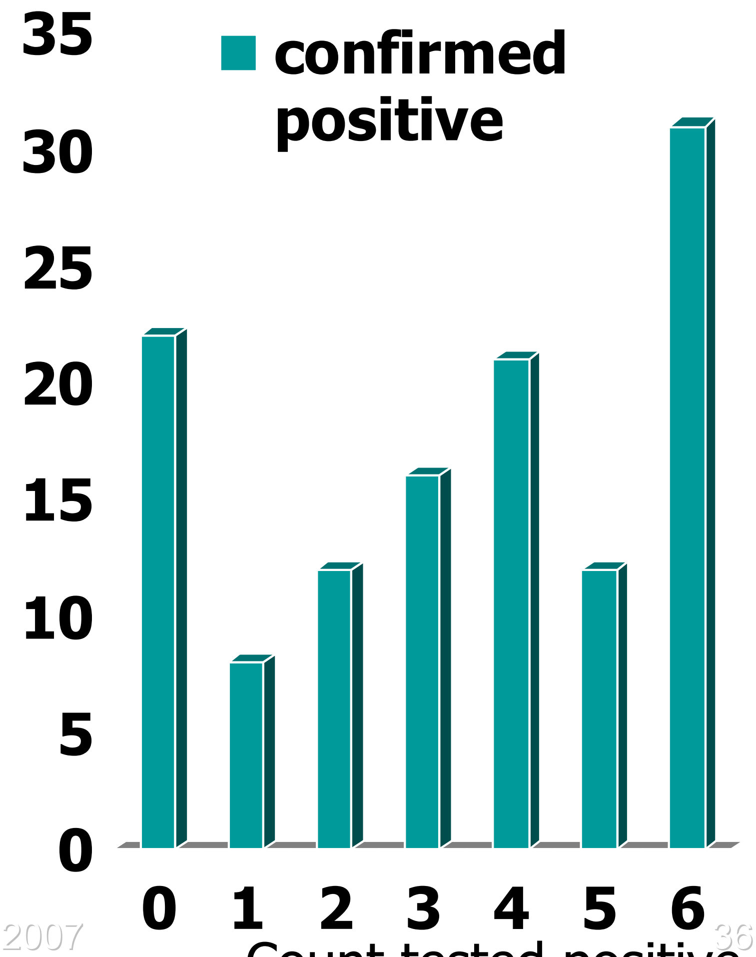
Conclusion

Ratio plot seems to work also as a diagnostic device for heterogeneity for the Power series distribution

f_1	f_2	n	$\hat{f}_0$	$\hat{N} = \hat{f}_0 + n$	$\hat{f}_0 / \hat{N}$
37	22	196	26	222	0.12

Distribution of counting the number of days testing positive for 122 men with confirmed colon cancer

- Now frequency f_0 of those tested negative at all 6 times with bowel cancer is known to be 22
- validation sample



Conclusion

f_1	f_2	n	$\hat{f}_0$	$\hat{N} = \hat{f}_0 + n$	$\hat{f}_0 / \hat{N}$
37	22	196	26	222	0.12

from validation sample: $f_0 = 22, f_0 / N = 22/122 = 0.18$

Overview

- Introduction
- Chao‘s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II
- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman‘s Estimator as an Upper Bound
  - Generalising Zelterman
- A Simulation Study

Idea of Mixed Modelling

look at mixed Poisson:

$$p_j = \int_0^\infty e^{-\theta}\theta^j / j! \, f(\theta) d\theta \approx \sum_{i=1}^{k} e^{-\theta_i}\theta_i^j / j! \, q_i$$

(to capture heterogeneity in θ)

reasonable: since NPMLE is always discrete

Idea of Mixed Modelling

now let $\theta_{\min} = \min\{\theta_1, \ldots, \theta_k\}$ then:

$$p_0 = \sum_{i=1}^{k} e^{-\theta_i} q_i \le e^{-\theta_{\min}} \sum_{i=1}^{k} q_i = e^{-\theta_{\min}}$$

$$\hat{N} = \frac{n}{1 - e^{-\theta_{\min}}} \ge \frac{n}{1 - \sum_{i=1}^{k} e^{-\theta_i} q_i} = \frac{n}{1 - p_0}$$

Idea of Mixed Modelling

since for a mixed Poisson:

$$\frac{p_1}{p_0} \le \frac{2p_2}{p_1} \le \frac{3p_3}{p_2} \le \frac{4p_4}{p_3} \ldots$$

reasonable

$$\theta_{\min} \approx \frac{2p_2}{p_1}$$

$$\frac{2p_2}{p_1} = \frac{2\sum_j q_j Po(2,\theta_j)}{\sum_j q_j Po(1,\theta_j)} \approx \frac{2q_1 Po(2,\theta_1)}{q_1 Po(1,\theta_1)} = \theta_1 = \theta_{\min}$$

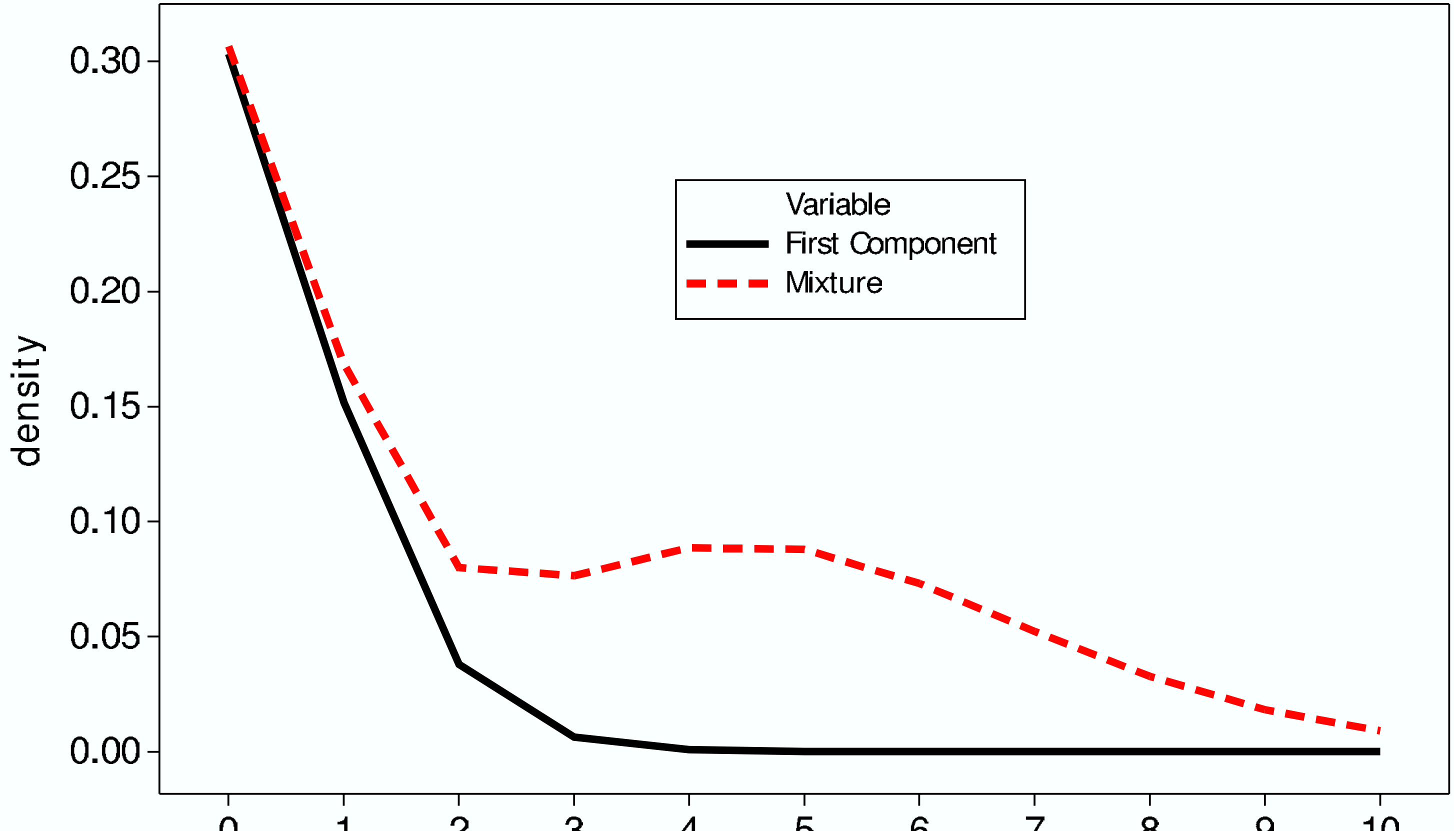


Illustration of approximation

$$p_j = \int_0^\infty e^{-\theta}\theta^j / j!\, f(\theta)d\theta \approx \sum_{i=1}^k e^{-\theta_i}\theta_i^j / j!\, q_i$$

large

1: $f(\theta) = Po(0.5)0.5 + 0.5Po(5)$ $\quad \frac{2p_2}{p_1} = 0.9499$

2: $f(\theta) = Po(0.5)0.9 + 0.1Po(5)$ $\quad \frac{2p_2}{p_1} = 0.5549$

3: $f(\theta) = Po(0.5)0.5 + 0.5Po(1)$ $\quad \frac{2p_2}{p_1} = 0.7741$

4: $f(\theta) = Po(0.5)0.9 + 0.1Po(1)$ $\quad \frac{2p_2}{p_1} = 0.5594$

small

Estimation

estimating

$$\theta_{\min} \approx \frac{2p_2}{p_1}$$

leads to

$$\hat{\theta}_{\min} = \frac{2\hat{p}_2}{\hat{p}_1} = \frac{2f_2}{f_1}$$

and Zelterman estimator arises:

$$\hat{N}_Z = \frac{n}{1-\exp(-\hat{\theta}_{\min})} = \frac{n}{1-\exp(-\frac{2f_2}{f_1})}$$

Zelterman's as truncated estimator

write (truncated Poisson likelihood for count 1 or 2)

$$p_1 = \frac{e^{-\theta}\theta}{e^{-\theta}\theta + e^{-\theta}\theta^2/2} = \frac{1}{1+\theta/2}$$

$$p_2 = \frac{e^{-\theta}\theta^2/2}{e^{-\theta}\theta + e^{-\theta}\theta^2/2} = \frac{\theta/2}{1+\theta/2}$$

so that binomial likelihood

$$f_1 \log(p_1) + f_2 \log(p_2)$$

occurs which is maximized at

$$\hat{\theta} = \frac{2f_2}{f_1}$$

Benefits of the truncated likelihood

binomial likelihood

$$f_1 \log(p_1) + f_2 \log(p_2)$$

is well studied:

1) $\mathrm{var}(\hat{p}_2) = \mathrm{var}(\frac{f_2}{f_1 + f_2}) = p_2(1 - p_2)/(f_1 + f_2)$

2) covariates might be easily included with logistic regression

Overview

- Introduction
- Chao‘s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II
- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman‘s Estimator as an Upper Bound
  - Generalising Zelterman
- A Simulation Study

Extending Zelterman`s estimator to the Power Series

write (truncated Poisson likelihood for count 1 or 2)

$$p_1 = \frac{\mu(\theta)\theta a_1}{\mu(\theta)\theta a_1 + \mu(\theta)\theta^2 a_2} = \frac{a_1}{a_1 + \theta a_2}$$

$$p_2 = \frac{\mu(\theta)\theta^2 a_2}{\mu(\theta)\theta a_1 + \mu(\theta)\theta^2 a_2} = \frac{\theta a_2}{a_1 + \theta a_2}$$

so that binomial likelihood occurs:

$$f_1 \log(p_1) + f_2 \log(p_2)$$

with

$$p = p_2 = \frac{\theta a_2}{a_1 + \theta a_2} \quad \text{or} \quad \theta = \frac{p}{1-p} \frac{a_1}{a_2}$$

since $\frac{\hat{p}}{1-\hat{p}} = \frac{f_2}{f_1}$, $\hat{\theta} = \frac{f_2}{f_1} \frac{a_1}{a_2}$

An example: mixed binomial

Binomial with size m: $\binom{m}{j}\theta^j(1+\theta)^{-m}$

so that $a_j = \binom{m}{j}$ and $\mu(\theta) = (1+\theta)^{-m}$

$$\hat{\theta} = \frac{f_2}{f_1}\frac{a_1}{a_2} = \frac{f_2}{f_1}\frac{m}{m(m-1)/2} = \frac{f_2}{f_1}\frac{2}{(m-1)}$$

$$\hat{N}_Z = \frac{n}{1-\hat{p}_0}, \hat{p}_0 = 1/(1+\hat{\theta})^m$$

Example: Screening for Bowel Cancer by taking Stool Samples at 6 Consecutive Days

	f_1	f_2	n	$\hat{f}_0$	$\hat{N} = \hat{f}_0 + n$	$\hat{f}_0 / \hat{N}$
Chao	37	22	196	26	222	0.12
Zelterman	37	22	196	75	271	0.26

from validation sample: $f_0 = 22, f_0 / N = 22/122 = 0.18$ (true)

Critical appraisal of Zelterman`s conventional estimator

- Collins and Wilson (1992 Biometrika):

…*For although it often does have a smaller bias than the other estimators, it does so at the cost of having a larger standard deviation which overwhelms the reduced bias* …

Generalising Zelterman

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

frequencies are concentrated on f_1, f_2, f_3

frequencies of drug users with $1, 2, 3, \ldots, m$ contacts to treatment institutions (hospitals) $(n = 6966)$:

$$f_1 = 2955, f_2 = 1186, f_3 = 803, f_4 = 611, \ldots$$

Generalising Zelterman

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

frequencies are concentrated on f_1, f_2, f_3

frequencies of drug users with $1, 2, 3, \ldots, m$ contacts to treatment institutions (hospitals) $(n = 6966)$:

$$f_1 = 2955, f_2 = 1186, f_3 = 803, f_4 = 611, \ldots$$

Zelterman's as triple truncated estimator

write (truncated Poisson likelihood for count 1,2 or 3)

$$p_1 = \frac{e^{-\theta}\theta}{e^{-\theta}\theta + e^{-\theta}\theta^2/2 + e^{-\theta}\theta^3/6} = \frac{1}{1+\theta/2+\theta^2/6}$$

$$p_2 = \frac{e^{-\theta}\theta^2/2}{e^{-\theta}\theta + e^{-\theta}\theta^2/2 + e^{-\theta}\theta^3/6} = \frac{\theta/2}{1+\theta/2+\theta^2/6}$$

$$p_3 = \frac{e^{-\theta}\theta^3/6}{e^{-\theta}\theta + e^{-\theta}\theta^2/2 + e^{-\theta}\theta^3/6} = \frac{\theta^2/6}{1+\theta/2+\theta^2/6}$$

so that multinomial likelihood in θ

$$f_1 \log(p_1) + f_2 \log(p_2) + f_3 \log(p_3)$$

occurs which is maximized at

$$\hat{\theta} = -\frac{3}{2}\frac{f_1 - f_3}{f_2 + 2f_1} + \sqrt{\frac{6(f_2+2f_3)}{f_2+2f_1} + \frac{9}{4}\frac{(f_1-f_3)^2}{(f_2+2f_1)^2}} \geq 0$$

Overview

- Introduction
- Chao,s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II
- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman`s Estimator as an Upper Bound
  - Generalising Zelterman
- A Simulation Study

Overview

- Introduction
- Chao,s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II
- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman`s Estimator as an Upper Bound
  - Generalising Zelterman
- A Simulation Study: Estimators considered

The (upper bound) estimators Z1

Zelterman's conventional estimator

$$\hat{\theta} = \frac{2\hat{p}_2}{\hat{p}_1} = \frac{2f_2}{f_1}$$

and

$$\hat{N}_{Z1} = \frac{n}{1-\exp(-\hat{\theta})} = \frac{n}{1-\exp(-\frac{2f_2}{f_1})}$$

The (upper bound) estimators Z2

Zelterman's generalized estimator

$$\hat{\theta} = -\frac{3}{2}\frac{f_1 - f_3}{f_2 + 2f_1} + \sqrt{\frac{6(f_2 + 2f_3)}{f_2 + 2f_1} + \frac{9}{4}\frac{(f_1 - f_3)^2}{(f_2 + 2f_1)^2}}$$

and

$$\hat{N}_{Z2} = \frac{n}{1 - \exp(-\hat{\theta})} = \frac{n}{1 - \exp(-\hat{\theta})}$$

The (upper bound) estimators Z3

not only $2p_2 / p_1 = \frac{2e^{-\theta}\theta^2 / 2}{e^{-\theta}\theta} = \theta,$ but also

$$\frac{2p_2 + 3p_3}{p_1 + p_2} = \frac{2e^{-\theta}\theta^2 / 2 + 3e^{-\theta}\theta^3 / 6}{e^{-\theta}\theta + e^{-\theta}\theta^2 / 2} = \theta \frac{e^{-\theta}\theta + e^{-\theta}\theta^2 / 2}{e^{-\theta}\theta + e^{-\theta}\theta^2 / 2} = \theta$$

motivates

$$\hat{\theta} = \frac{2\hat{p}_2 + 3\hat{p}_3}{\hat{p}_1 + \hat{p}_2} = \frac{2f_2 + 3f_3}{f_1 + f_2}$$

$$\hat{N}_{Z3} = \frac{n}{1 - \exp(-\hat{\theta})} = \frac{n}{1 - \exp(-\hat{\theta})}$$

The (lower bound) estimators C1

under mixed Poisson sampling

$$\frac{p_1}{p_0} \leq \frac{2p_2}{p_1} \leq \frac{3p_3}{p_2} \leq \dots$$

$\Rightarrow$ C1 (original Chao estimator):

$$\frac{p_1 p_1}{2p_2} \leq p_0 \quad \text{replacing with estimates}$$

$$\hat{f}_0 = \frac{f_1^2}{2f_2}, \; N_{C1} = n + \hat{f}_0$$

The (lower bound) estimators
C2

under mixed Poisson sampling

$$\frac{p_1}{p_0} \leq \frac{2p_2}{p_1} \leq \frac{3p_3}{p_2} \leq \ldots$$

$\Rightarrow$ C2 (generalized Chao estimator):

$$\frac{p_1 p_2}{3p_3} \leq p_0 \quad \text{replacing with estimates}$$

$$\hat{f}_0 = \frac{f_1 f_2}{3f_3}, N_{C2} = n + \hat{f}_0$$

Classical estimator under Poisson homogeneity M

under Poisson sampling

$$\theta = \frac{p_1}{p_0} = \frac{2p_2}{p_1} = \frac{3p_3}{p_2} = \ldots$$

$$\Rightarrow \theta = \frac{2p_2 + 3p_3 + 4p_4 \ldots}{p_1 + p_2 + p_3 + \ldots} \quad \Rightarrow \hat{\theta} = \frac{2f_2 + 3f_3 + 4f_4 \ldots}{f_1 + f_2 + f_3 + \ldots}$$

$$\hat{N}_M = \frac{n}{1 - \exp(-\hat{\theta})} = \frac{n}{1 - \exp(-\hat{\theta})}$$

Overview

- Introduction
- Chao‚s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II
- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman‵s Estimator as an Upper Bound
  - Generalising Zelterman
- A Simulation Study: Design

Six Experiments
N=100, replication=1,000

0: $f(\theta) = Po(0.5)$

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

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

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

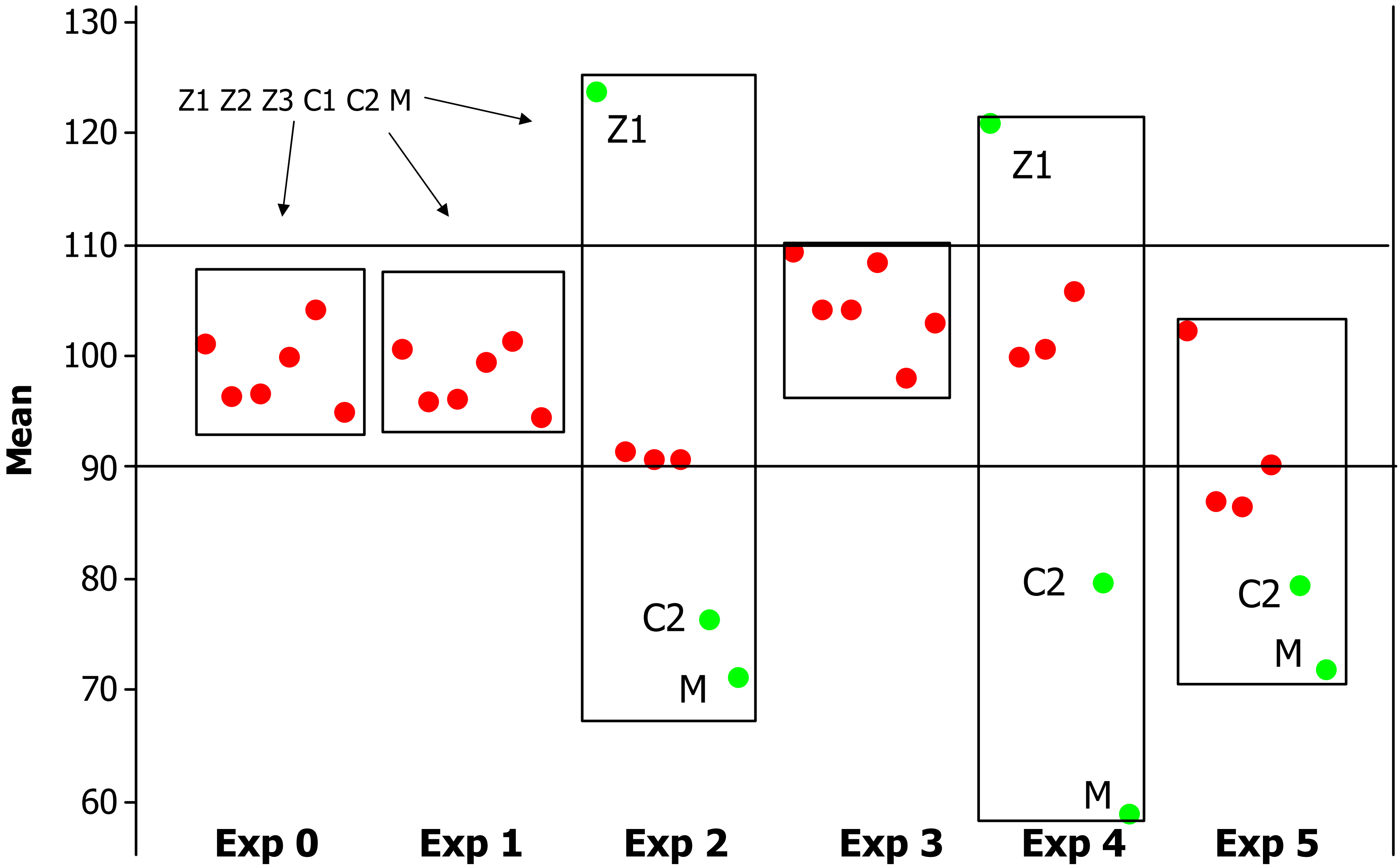
4: $f(\theta) = Po(0.5)0.9 + 0.1Po(5)$

5: $$f(\theta) = Po(0.5)0.5 + 0.1Po(1) + 0.1Po(2) + 0.1Po(3) + 0.1Po(4) + 0.1Po(5)$$

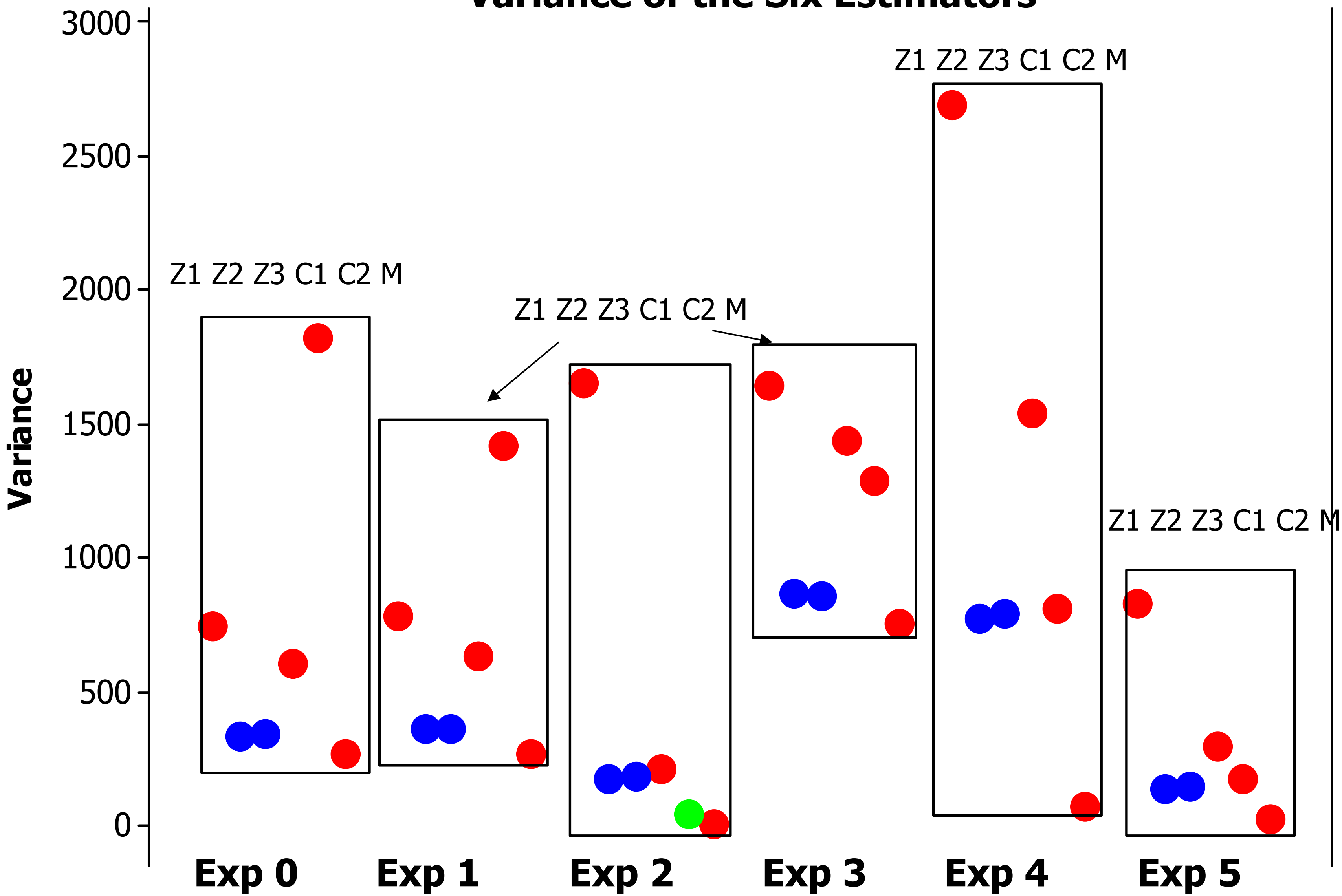
Overview

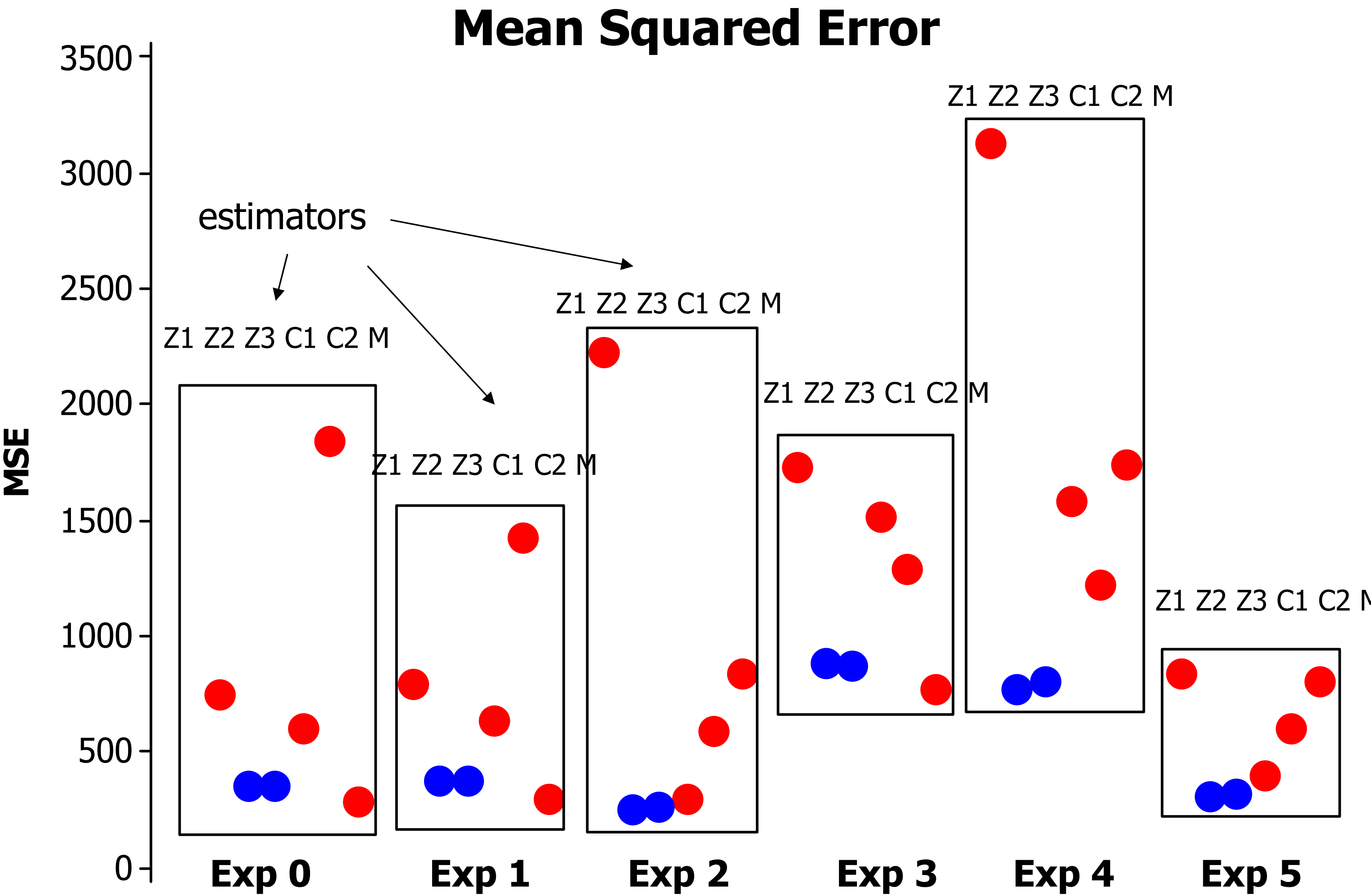
- Introduction
- Chao,s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II
- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman`s Estimator as an Upper Bound
  - Generalising Zelterman
- A Simulation Study: Results

Mean for the Six Estimators (N=100 is true)



Variance of the Six Estimators





Mean Squared Error
MSE
3500
3000
2500
2000
1500
1000
500
0
estimators
Z1 Z2 Z3 C1 C2 M
Z1 Z2 Z3 C1 C2 M
Z1 Z2 Z3 C1 C2 M
Z1 Z2 Z3 C1 C2 M
Z1 Z2 Z3 C1 C2 M
Z1 Z2 Z3 C1 C2 M
Exp 0
Exp 1
Exp 2
Exp 3
Exp 4
Exp 5

Illustration: Project on illicit drug use in Bangkok 2001 (4th Quarter)

frequencies of drug users with $1, 2, 3, \ldots, m$ contacts to treatment institutions (hospitals):

$$f_1 = 2955, f_2 = 1186, f_3 = 803, f_4 = 611, \ldots$$

$$n = f_1 + f_2 + \ldots + f_m = 6,966$$

$$\hat{N}_{Z1} = 12,622 \qquad \hat{N}_{C1} = 10,647$$

$$\hat{N}_{Z2} = 7,987 \qquad \hat{N}_{C2} = 8,421$$

$$\hat{N}_{Z3} = 10,172$$

Overview

- Introduction

- Chao‘s Idea and Lower Bounds
  - Extending Chao: Way I
  - Extending Chao Way II

- Upper Bounds and Zelterman approach
  - Motivation
  - Zelterman‘s Estimator as an Upper Bound
  - Generalising Zelterman

- A Simulation Study: improve upon Z3?

improve upon Z3 ?

$$\hat{N}_Z = \frac{n}{1-\exp(-\hat{\theta})}$$

not only $2p_2/p_1 = \frac{2e^{-\theta}\theta^2/2}{e^{-\theta}\theta} = \theta$, but also

$$\frac{2p_2+3p_3}{p_1+p_2} = \frac{2e^{-\theta}\theta^2/2+3e^{-\theta}\theta^3/6}{e^{-\theta}\theta+e^{-\theta}\theta^2/2} = \theta\frac{e^{-\theta}\theta+e^{-\theta}\theta^2/2}{e^{-\theta}\theta+e^{-\theta}\theta^2/2} = \theta$$

motivates

$$\hat{\theta}_3 = \frac{2\hat{p}_2+3\hat{p}_3}{\hat{p}_1+\hat{p}_2} = \frac{2f_2+3f_3}{f_1+f_2}$$

$$\frac{2p_2+3p_3+4p_4}{p_1+p_2+p_3} = \frac{2e^{-\theta}\theta^2/2+3e^{-\theta}\theta^3/6+4e^{-\theta}\theta^4/24}{e^{-\theta}\theta+e^{-\theta}\theta^2/2+e^{-\theta}\theta^3/6} = \theta$$

motivates

$$\hat{\theta}_4 = \frac{2\hat{p}_2+3\hat{p}_3+4\hat{p}_4}{\hat{p}_1+\hat{p}_2+\hat{p}_3} = \frac{2f_2+3f_3+4f_4}{f_1+f_2+f_3}$$

improve upon Z3 ?

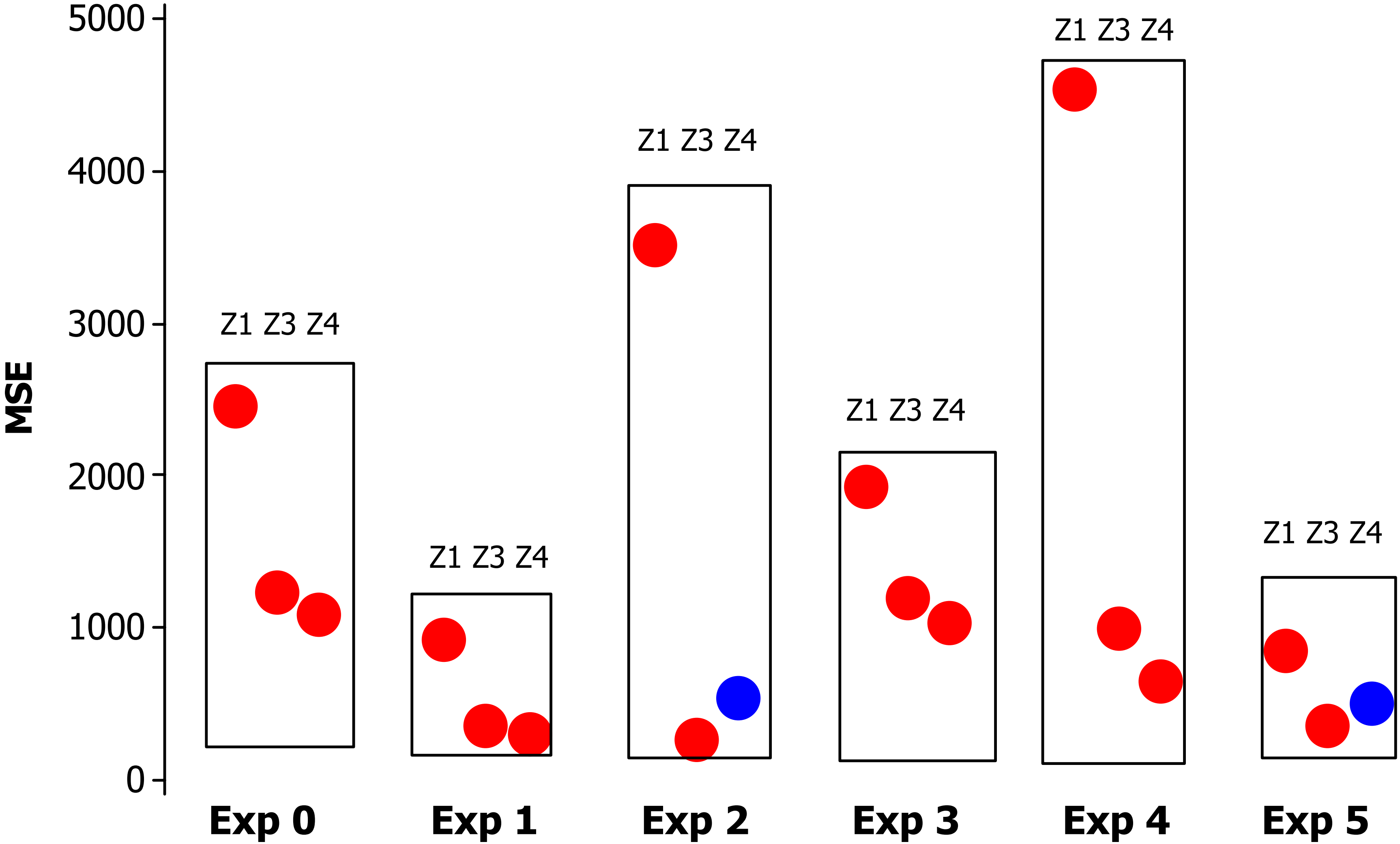
Three Estimators: $\hat{N}_Z = \dfrac{n}{1-\exp(-\hat{\theta})}$

Z1: $\hat{\theta}_1 = \dfrac{2f_2}{f_1}$

Z3: $\hat{\theta}_3 = \dfrac{2f_2 + 3f_3}{f_1 + f_2}$

Z4: $\hat{\theta}_4 = \dfrac{2f_2 + 3f_3 + 4f_4}{f_1 + f_2 + f_3}$

MSE for Three Generalized Zelterman Estimators



Key-References

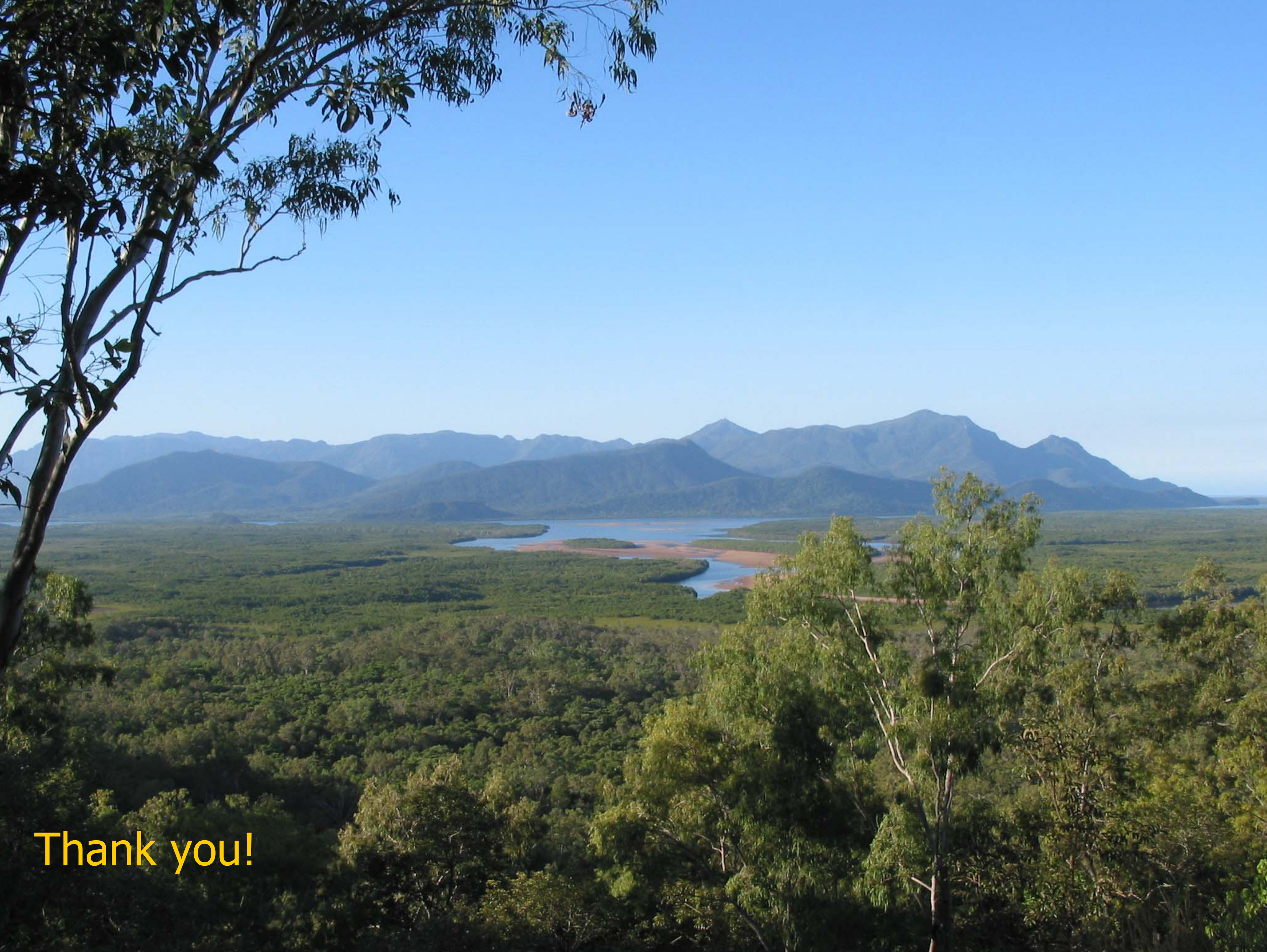
Böhning, D. and Kuhnert, R. (2006). The Equivalence of Truncated Count Mixture Distributions and Mixtures of Truncated Count Distributions. *Biometrics* ***62***, 1207-1215.

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


Thank you!