

# Capture-Recapture Methods for Human and Animal Populations Based Upon Nonparametric Maximum Likelihood

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

*Invited Presentation at  
KMITL International Conference  
on Science and Applied Science 2006  
8-10 March 2006*

Joint work with:

**Ronny Kuhnert, Ekkehart Dietz**

Institute for Social Medicine, Epidemiology,  
and Health Economics, Charité Medical School Berlin, Germany

**Dieter Schön**

Robert Koch Institute, Berlin, Germany

**Valentin Patilea**

CREST-ENSAI, France

**Chukiat Viwatwongkasem**

Mahidol University, Bangkok, Thailand

**Busaba Supawattanabodee**

Cinical Epidemiology Unit, Bangkok Metropolitan  
Administration Medical College and Vajira Hospital  
Bangkok, Thailand

## *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., Suppawattanabodee, B., Kusolvisitkul, W, and Viwatwongkasem, C. (2004). [Estimating the Number of Drug Users in Bangkok 2001: A Capture-Recapture Approach Using Repeated Entries in One List.](#) *European Journal of Epidemiology* **19**, 1075-1083.

*Key-Reference*

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\*\*](http://www.reading.ac.uk/~sns05dab)

# Overview

- Motivation and Background (15 min)
- Truncated Mixtures or Mixtures of Truncated Distributions? (5 min)
- Some Equivalence Results (15 min)
  - Model Spaces and Likelihood Surfaces (5 min)
  - Model Transformations (5 min)
  - Population Size Estimates (5 min)
- Epilogue (2 min)

# Capture-Recapture Procedures based upon Counting Distributions

- Basic objective of CR: estimate population size
- In particular of interest in areas where direct counting is difficult such as
  - a wildlife population (historic genesis)
  - how many people drive a car without license?
  - how many practicing physicians are alcohol dep.?
  - how many cases of a disease remain undetected?
- Adjustment for undercount



# How many cases **N** in a population?

- Some mechanism identifies  $n$  cases
- $p_0$  probability of being **not** identified by the mechanism
- **Then:**

$$\begin{aligned} N &= N p_0 + (1 - p_0) N \\ &= \text{unobserved} + \text{observed cases} \\ &= N p_0 + n \end{aligned}$$



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

(Horwitz-Thompson)

# Horwitz-Thompson-Approach seems easy, but ...

inclusion probability often unknown

approaches differ in the way they  
estimate the inclusion probability,

or in other words, how they

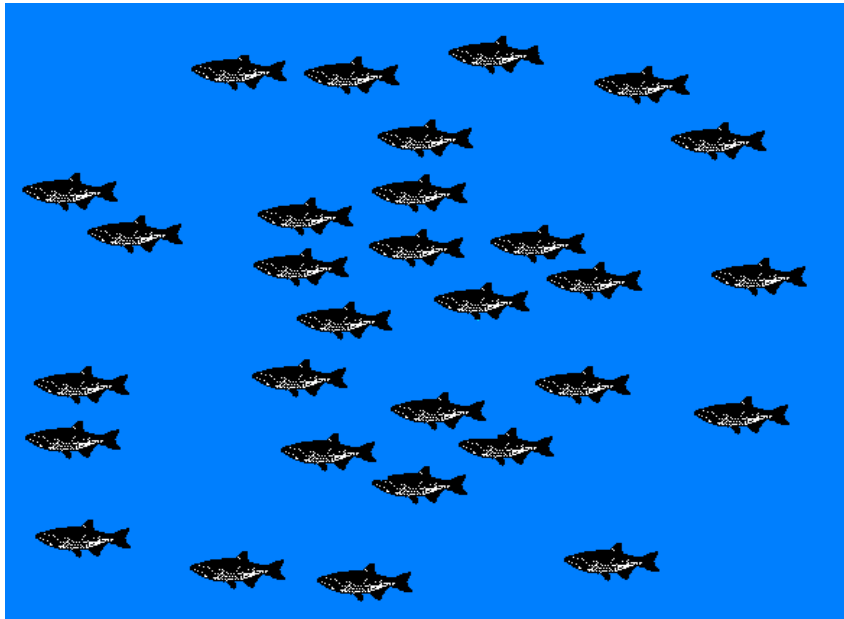
model  $p_0$  



# Two sample capture-recapture method (historic interest)

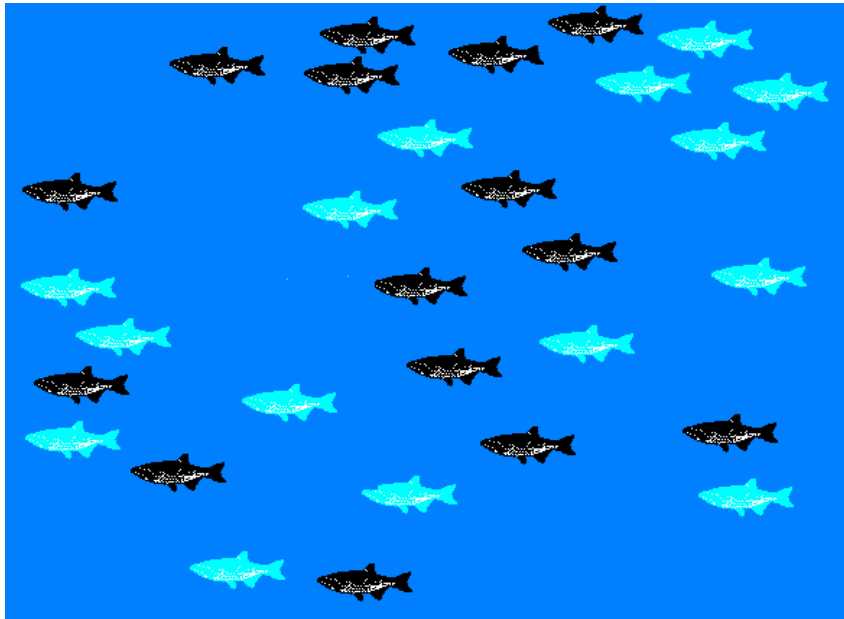
- Animal populations
  - Capture a sample of fish
  - Mark them
  - Release them
  - Recapture a sample at a later date
  - Look for marks
  - Estimate population size

# Example - fish



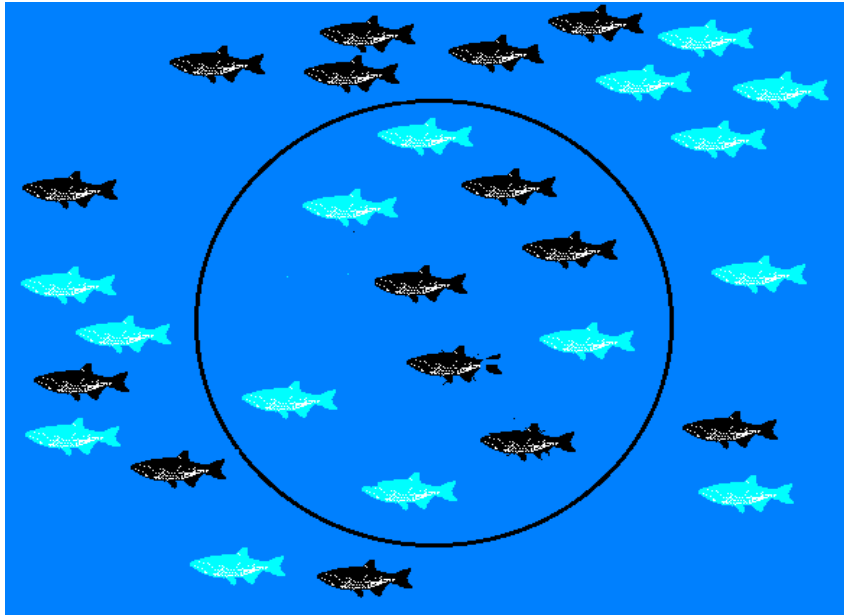
- Unknown number of fish in a lake

# Example - fish



- Unknown number of fish in a lake
- Catch a sample and mark them
- Let them loose

# Example - fish



- Unknown number of fish in a lake
- Catch a sample and mark them
- Let them loose
- Recapture a sample and look for marks

# Estimate population size

$n_{10}$  = number in first sample, but not in second

$n_{01}$  = number in second sample, but not in first

$n_{11}$  = number in both samples

$N$  = total population size

	Sample 2		total
Sam- ple 1	$n_{11}$	$n_{10}$	$n_{1+}$
	$n_{01}$	$n_{00}$	$n_{0+}$
total	$n_{+1}$	$n_{+1}$	$N$

# Estimate population size

assume that samples are independent:

$$\begin{aligned}n_{11}/N &= (n_{11} + n_{10})/N \times (n_{11} + n_{01})/N \\ &= (n_{1+}/N) (n_{+1}/N)\end{aligned}$$



$$\hat{N} = (n_{1+} n_{+1})/n_{11}$$

Lincoln (1896) – Petersen (1930)

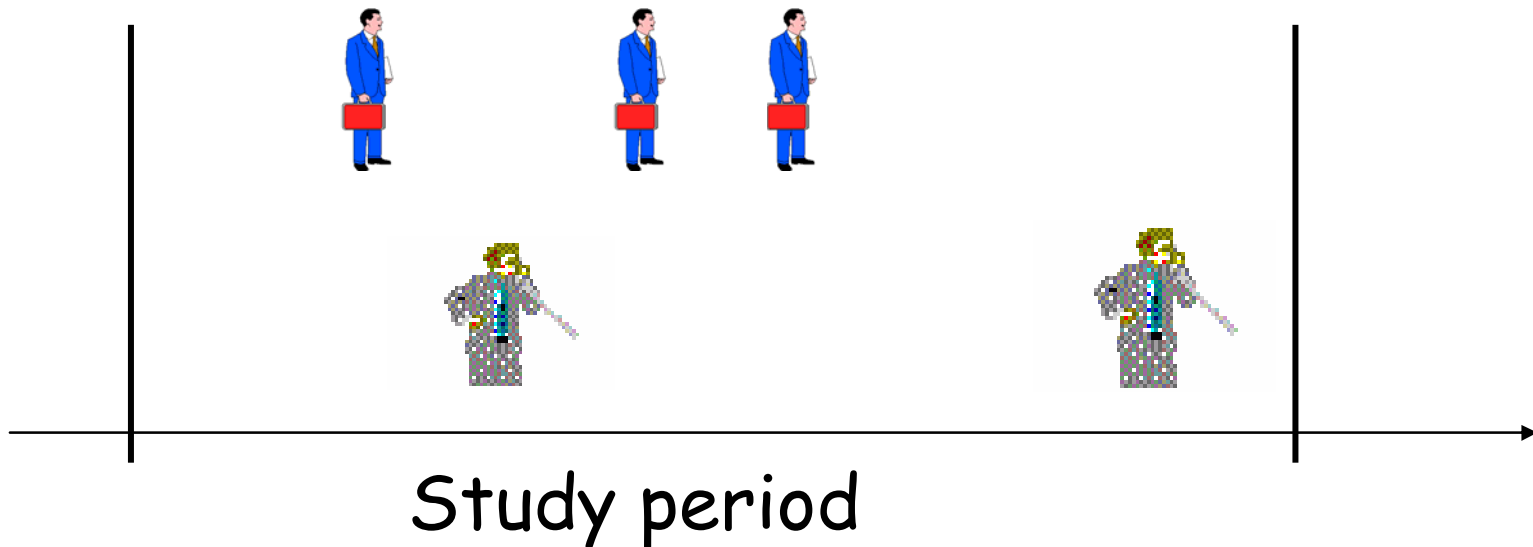
# More samples (traps, sources)

<b>ID</b>	<b>Sample</b>	<b>Sample</b>	<b>Sample</b>		<b>Counting</b>
	<b>1</b>	<b>2</b>	<b>3</b>		<b>captures</b>
001	1	0	0	...	1
002	0	1	1	...	2
003	0	1	0	...	1
004	1	0	1	...	2
005	1	1	1	...	3
...	...	...	...	...	...

Could use log-linear modelling of multi-way frequency table (Chapter 6, Bishop, Holland, and Fienberg 1975)

# Counts of capture-recaptures as outcome of continuous time CR-experiments

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



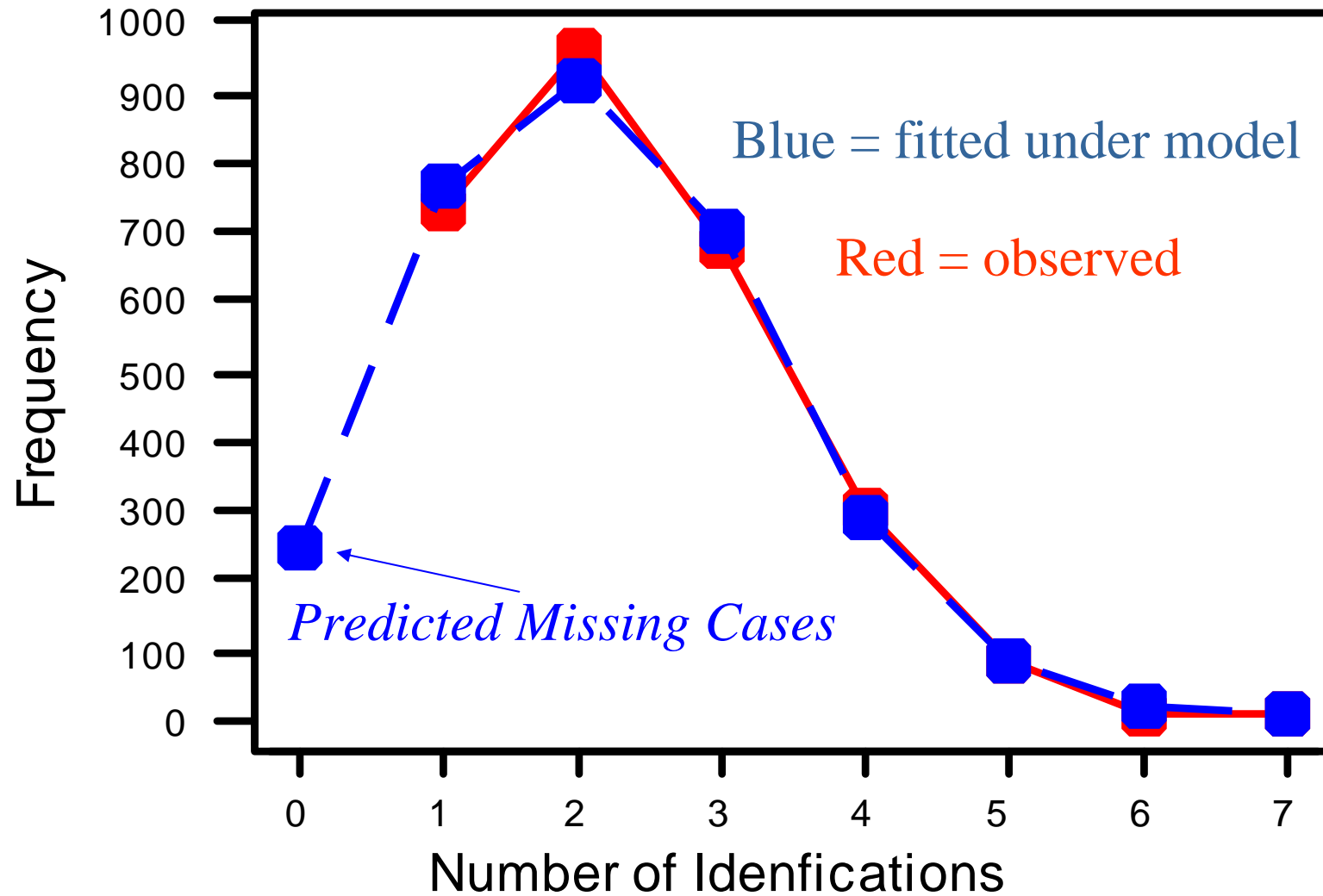


# The Counting Distribution

... occurs when the mechanism can catch multiple identifications (s.a. police identifies and expells an illegal immigrant several times)

Count of identifications $i$	Frequency of counts with $i$ identifications	observed
0	$n_0$	no
1	$n_1$	yes
2	$n_2$	yes
3	$n_3$	yes
4	$n_4$	yes
...	...	...

# Distribution of Observed and Predicted Counts of Sources *for fictional data of multiple identifications*






# The Counting Distribution: A historic Example

- McKendrick's cholera data
- Village in India had households with cholera cases  $n_1=32$ ,  $n_2=16$ ,  $n_3=6$ ,  $n_4=1$
- McKendrick ignored the houses with no cases
- Constructed an estimate (moment) based upon a Poisson assumption for the counts

Cholera Epidemic in an Indian Village (1915-1920)



-  House not affected, no cases
-  House affected, no cases
-  House affected,  $m$  cases

# Simple Distributional Count Models

Poisson (for unobservable counts)

$$f(y, \theta) = e^{-\theta} \theta^y / y! , y = 0, 1, 2 \dots$$

truncated Poisson (for observable counts)

$$f(y, \theta) = \frac{1}{1 - e^{-\theta}} e^{-\theta} \theta^y / y! , y = 1, 2 \dots$$

Predicted Probability of a Zero:

$$p_0 = f(y, \theta) = e^{-\theta}$$

# Simple Distributional Count Models

after  $\theta$  is identified ...

.... probability of a zero count:

$$p_0 = f(y = 0, \theta) = e^{-\theta}$$

$$\Rightarrow \hat{N} = \frac{n}{1 - p_0} = \frac{n}{1 - e^{-\theta}}$$

# ML-Estimation in Zero-Truncated Poisson Models

Step 1: suppose  $\hat{n}_0$  would be available

$$\hat{\theta} = \frac{1}{n + \hat{n}_0} \sum_{i=1}^m i n_i$$

Step 2: suppose  $\hat{\theta}$  would be available

$$\hat{N} = \frac{n}{1 - p_0} = \frac{n}{1 - e^{-\hat{\theta}}} \Rightarrow \hat{n}_0 = \hat{N} - n = n \frac{e^{-\hat{\theta}}}{1 - e^{-\hat{\theta}}}$$

# EM-Algorithm

Step 1 (M-Step): suppose  $\hat{n}_0$  would be available

$$\hat{\theta} = \frac{1}{n + \hat{n}_0} \sum_{i=1}^m i n_i$$

Step 2 (E-Step): suppose  $\hat{\theta}$  would be available

$$\hat{n}_0 = E(n_0 \mid \hat{\theta}; n_1, n_2, \dots) = n \frac{p_0}{1 - p_0} = n \frac{e^{-\hat{\theta}}}{1 - e^{-\hat{\theta}}}$$

# The counting distribution: 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 at least once and cancer status evaluated
- 196 were confirmed positive to have bowel cancer
- How many of 35,000 **unconfirmed** negative have bowel cancer?



38,000 men screened

3,000 tested positive  
at least once

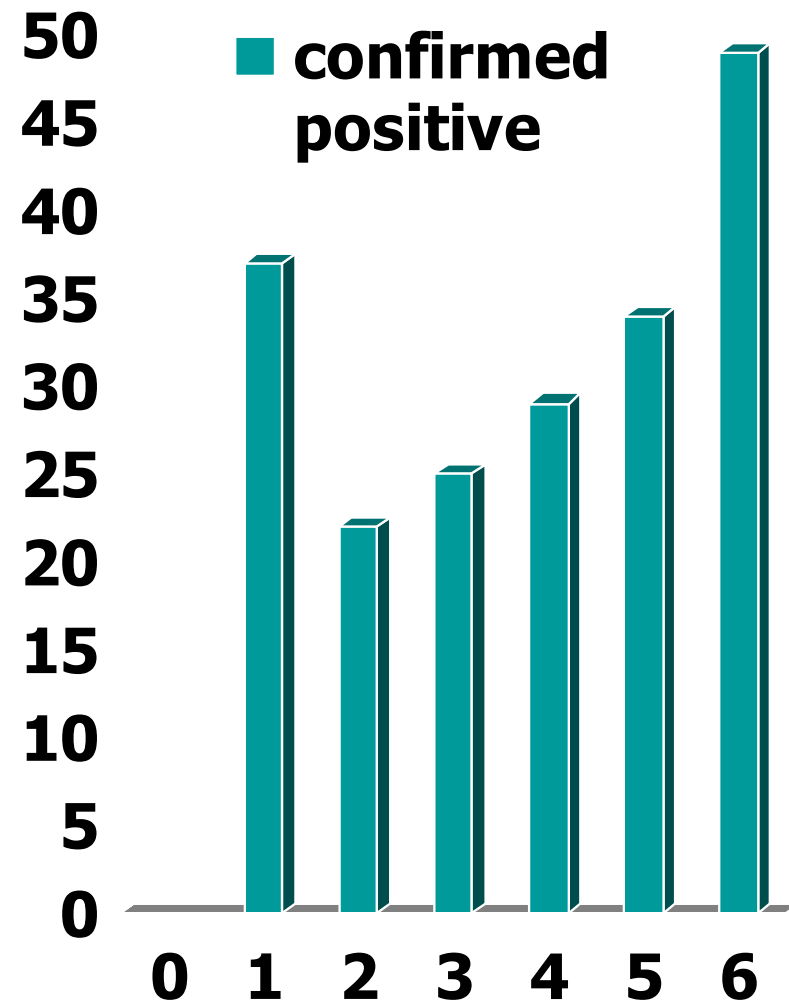
196 confirmed  
positive



Men with cancer, but  
tested always negative

# The counting distribution: a recent example from screening

- frequency  $n_0$  of those tested negative at all 6 times with bowel cancer is unknown
- an estimate of  $n_0$  might be constructed from the distribution  $n_1, n_2, n_3, \dots$  of counts



# Second Example: Surveillance Study on Drug Use in Thailand

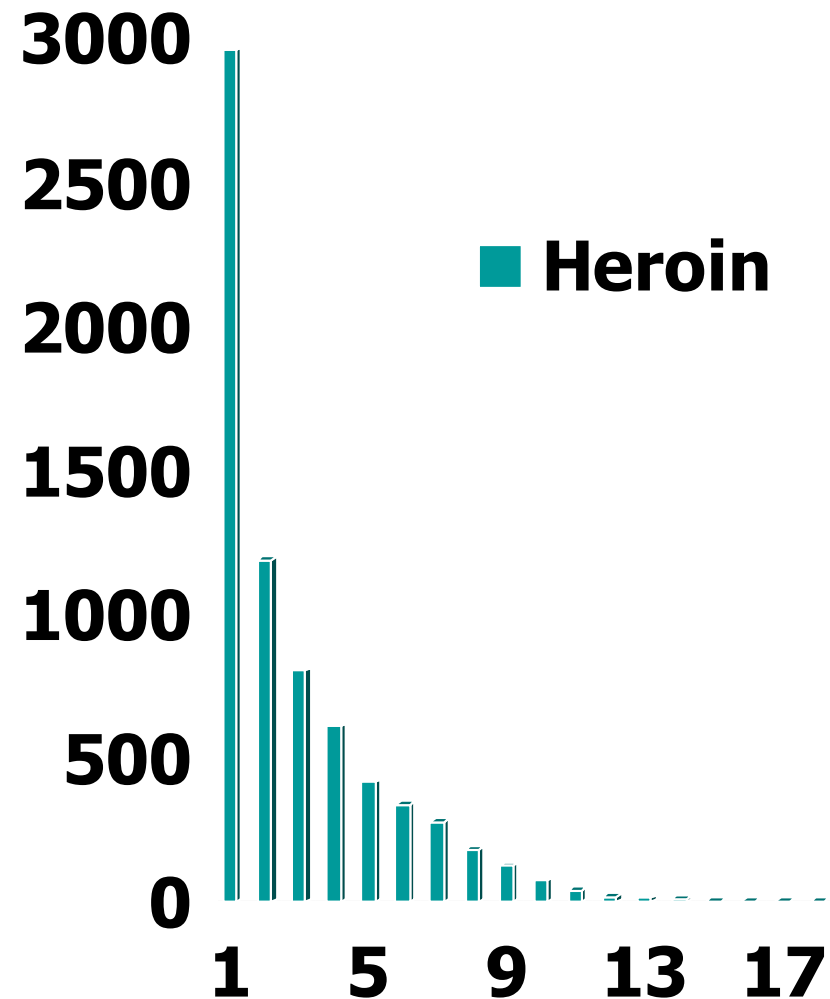
- Ministry of Public Health (Th) collects routinely data on drug use via the ONCB on drug users visiting treatment institutions
- In a pilot study (Böhning, Busaba, Chukiat et al. 2004 *EUJE*) CR-Poisson mixture model applied to data from 2001 (last quarter)
- Major emphasis on heroin and metamphetamin users

Welcome all participants to a special lecture  
"Capture-Recapture Procedures in Public Health"  
Speaker: Prof. Dr. Dankmar Böhning  
Organized by Department of Biostatistics, Faculty of Public Health,  
Mahidol University. March 15 - April 16, 2004



# Application: surveillance study on drug use in Thailand

- Count distribution (counting number of visits) for heroin users
- $n = 7,048$  observed heroin users (2001, 4)



# More General Zero-Truncated Count Distributional Models

general count distribution

$$f(y, \theta), y = 0, 1, 2, \dots$$

assoc. zero-truncated distribution

$$\frac{1}{1 - f(0, \theta)} f(y, \theta), y = 1, 2, \dots$$

# Overview

- Motivation and Background
- Truncated Mixtures or Mixtures of Truncated Distributions?
- Some Equivalence Results
  - Model Spaces and Likelihood Surfaces
  - Model Transformations
  - Population Size Estimates

# More flexible and robust approach through mixtures

- Simple counting sources distributions such as Binomial and Poisson require assumptions such as homogeneity of identification probabilities that are seldom met in reality
- allowing the identification probability to vary in unobserved sub-populations will be more realistic

# More flexible and robust approach through mixtures

G.A.F. Seber (2001, JABES):

*However, heterogeneity of capture is an ever present problem, and a natural way of modeling heterogeneity is to use a mixture distribution for the probability of capture. This involves assuming that there are  $G$  groups in the population, for which the probability of capture is constant within each group.*

Norris and Pollock (1996, 1998)

Pledger (2000), Link (2003)



# The mixture approach in a nutshell

**mixture density:** (for  $y = 0, 1, 2, 3, \dots$ )

$$f(y, \theta) = f(y, \lambda_1)q_1 + \dots + f(y, \lambda_k)q_k$$

$f(y, \lambda)$  is **component density**

Example:  $f(y, \lambda) = Po(y, \lambda) = e^{-\lambda} \lambda^y / y!$

$\theta = \begin{pmatrix} \lambda_1 & \dots & \lambda_k \\ q_1 & \dots & q_k \end{pmatrix}$  is **mixing distribution**

# two ways of setting up the mixture for the zero-truncated counts

- truncated mixture of Poisson distributions  
(primal modal)
- or ...
- mixture of truncated Poisson distributions  
(dual model)

# truncated Poisson mixture (primal model)

$$\sum_{j=1}^k q'_j \text{Po}(y, \lambda'_j)$$

---

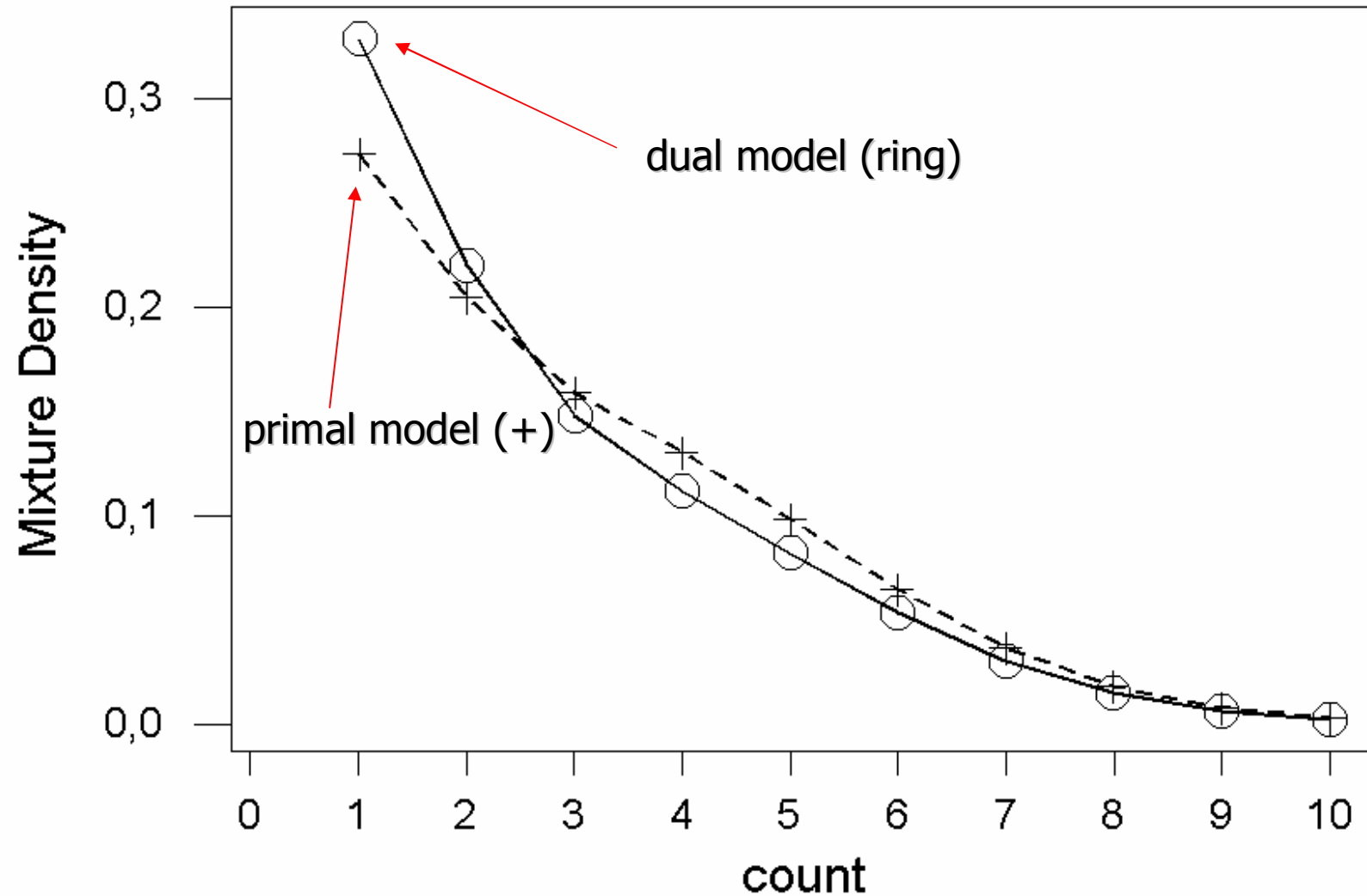
$$1 - \sum_{j=1}^k q'_j \text{Po}(0, \lambda'_j)$$

mixture of truncated Poissons  
(dual model)

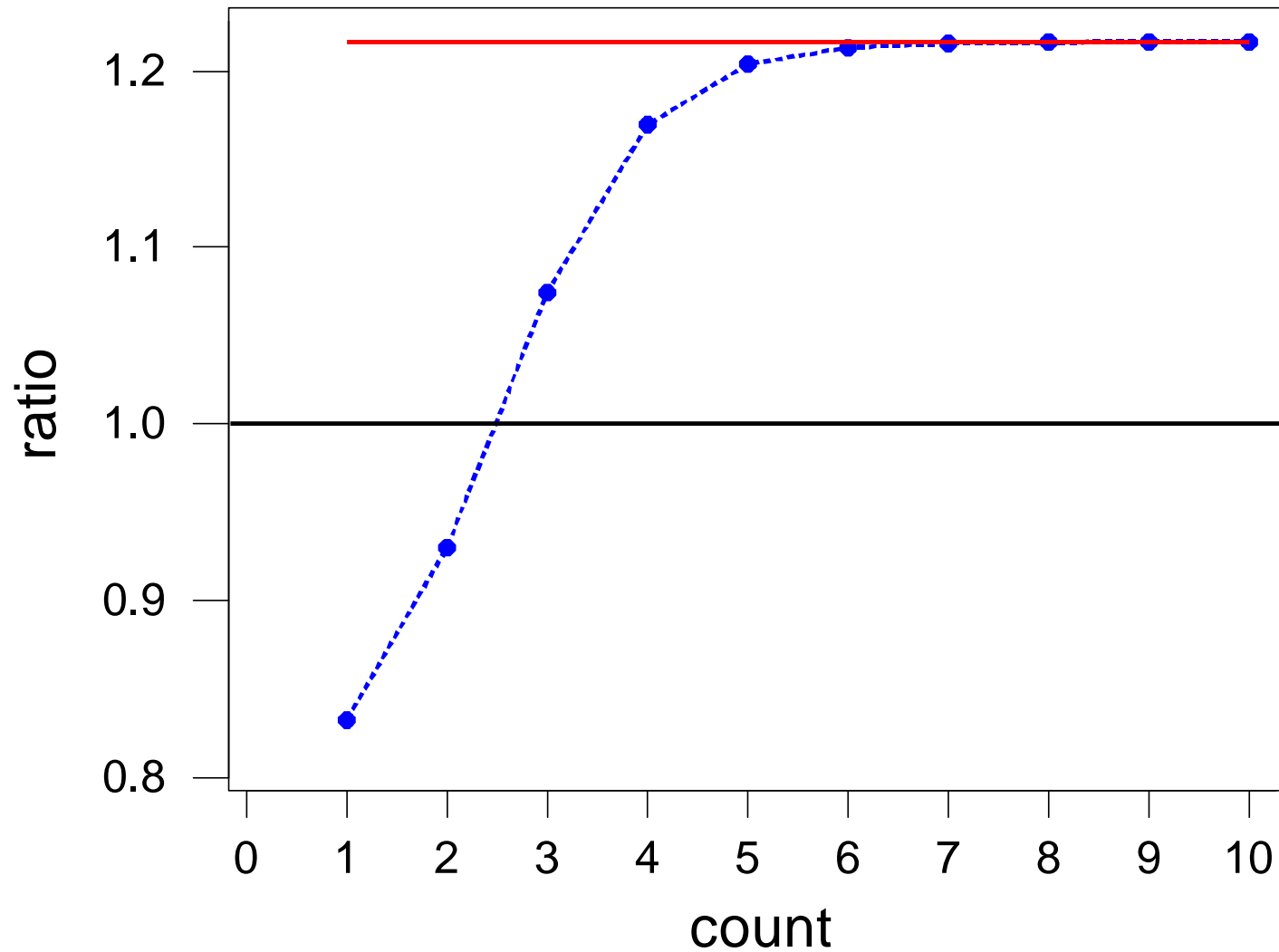
$$\sum_{j=1}^k q_j \frac{Po(y, \lambda_j)}{1 - Po(0, \lambda_j)}$$

$$= \sum_{j=1}^k q_j Po_+(y, \lambda_j)$$

Illustration: use mixture with equal weights and component means 1 and 4 in both models



# Ratio of truncated mixture (primal) to mixture of truncated Poissons (dual)



# truncated Poisson mixture (primal model)

$$\frac{\sum_{j=1}^k q'_j \text{Po}(y, \lambda'_j)}{1 - \sum_{j=1}^k q'_j \text{Po}(0, \lambda'_j)}$$

- close to the original problem, easy to understand and to communicate
- used in the CR-literature: Dahiya & Gross (73), Blumenthal et al. (79), Scollnik (97), van der Heijden et al. (03), Grogger & Carson (91) Cameron & Trivedi (98), Winkelmann (03)
- But technical difficult, because of **non-concavity**

# mixture of truncated Poissons (dual model)

$$\sum_{j=1}^k q_j \frac{Po(y, \lambda_j)}{1 - Po(0, \lambda_j)} = \sum_{j=1}^k q_j Po_+(y, \lambda_j)$$

- less close to the original problem
- but convex problem with strong results available on NPMLE and global ML estimation



# Benefit in using the dual model

$$\sum_{j=1}^k q_j \frac{Po(y, \lambda_j)}{1 - Po(0, \lambda_j)} = \sum_{j=1}^k q_j f_+(y, \lambda_j) = f_+(y, Q)$$

let  $l(Q) = \sum_{i=1}^m n_i \log f_+(i, Q)$  be the log-likelihood

discrete mixing distribution  $\hat{Q}$  such that

$$l(\hat{Q}) \geq l(Q)$$

for all (discrete) mixing distributions is called the

*nonparametric maximum likelihood estimator* (NPMLE)

# Benefit in using the dual model

Equivalence Theorem for the NPMLE;  
(Böhning 82, Lindsay 83):

$$l(\hat{Q}) \geq l(Q) \text{ for all discrete } Q$$

$$\Leftrightarrow d(\lambda, \hat{Q}) \leq 1 \text{ for all } \lambda$$

where  $d(\lambda, Q) = \frac{1}{n} \sum_{i=1}^m n_i \frac{f_+(i, \lambda)}{f_+(i, Q)}$  gradient function

McKendrick's cholera data: Village in India had households with cholera cases  $n_1=32$ ,  $n_2=16$ ,  $n_3=6$ ,  $n_4=1$

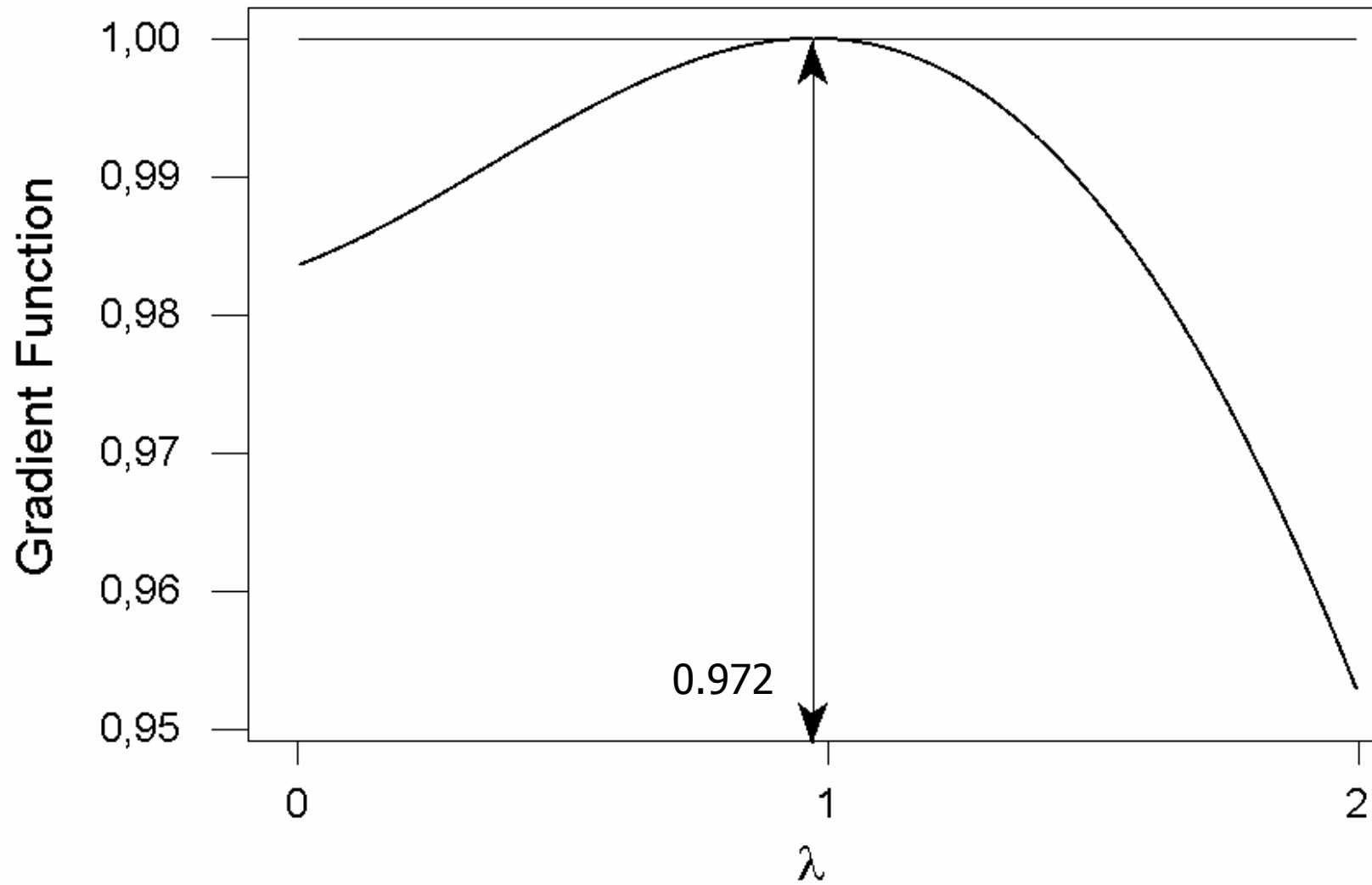
homogenous Poisson: one component mixture

$$d(\lambda, \hat{Q}) = \frac{1}{n} \sum_{i=1}^m n_i \frac{f_+(i, \lambda)}{f_+(i, \hat{Q})}, \text{ where } f_+(i, \lambda) = \frac{e^{-\lambda} \lambda^i}{1 - e^{-\hat{\lambda}}}$$

where  $\hat{Q}$  puts all mass at  $\hat{\lambda} = 0.972$

*e.g.*

$$d(\lambda, \hat{Q}) = d(\lambda, 0.972) = \frac{1}{n} \sum_{i=1}^m n_i \frac{f_+(i, \lambda)}{f_+(i, 0.972)}$$



# Benefit in using the dual model

- Algorithms exist finding the globally the NPMLE
- VDM, VEM, ISDM
- EM, EMGFU
- Others

# Some results

- n=7,048 (observed)
- N=17,278
- N-n=10,230 (hidden)
- Ratio:  
observed/hidden=0.69

## Estimating the Number of Heroin Users:

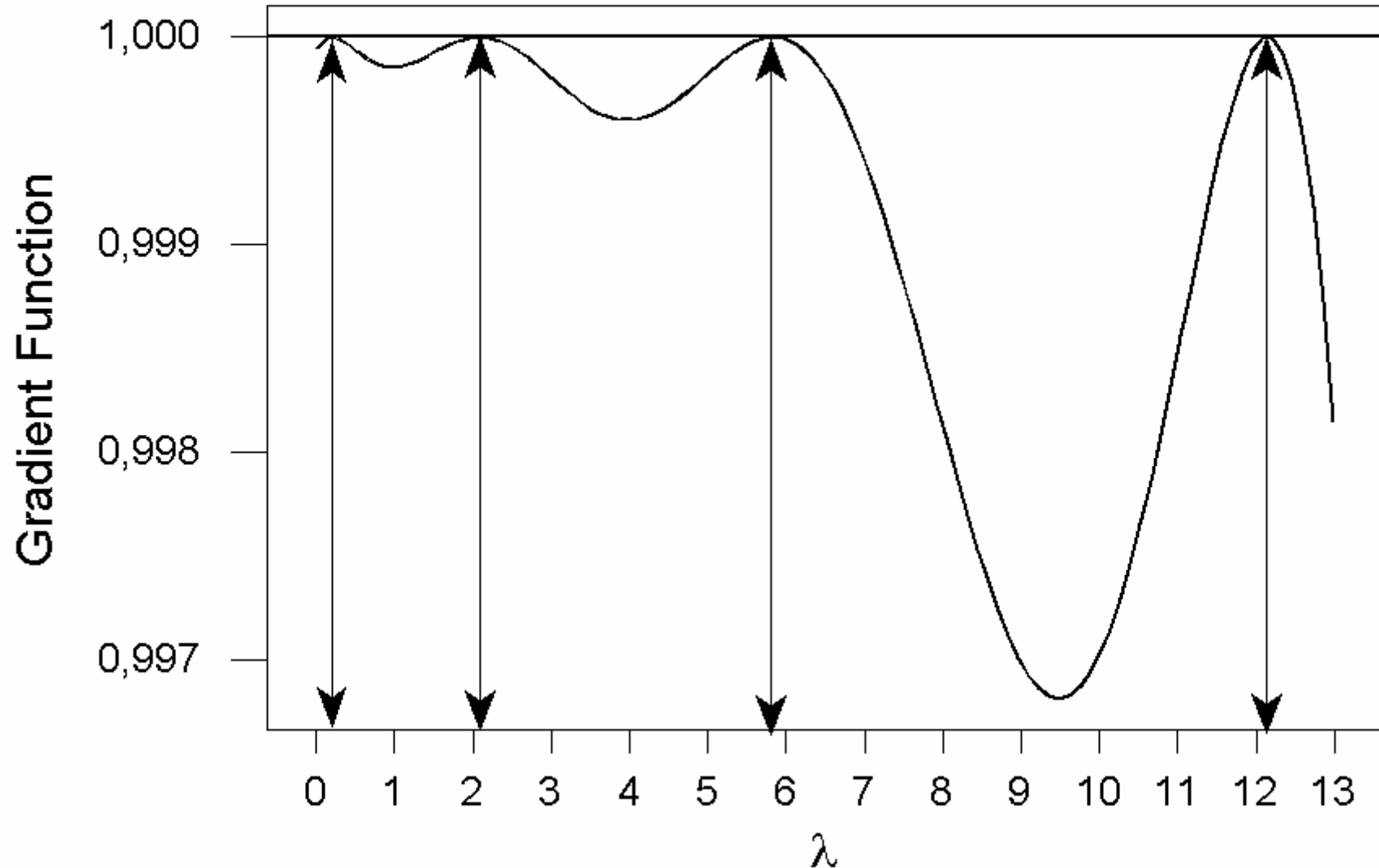
$k$	$\hat{\lambda}_j$	$\hat{q}_j$	log-likelih.	$AIC$	$BIC$	$\hat{N}$
1	2.75	1.00	-15462	-30927	-30934	7543
2	0.88 5.40	0.75 0.25	-13214	-26434	-26455	10226
3	0.41 2.97 6.80	0.69 0.22 0.09	-13134	-26279	-26313	13350
4	0.21 2.13 5.84 12.20	0.70 0.19 0.10 0.01	-13120	<b>-26255</b>	<b>-26303</b>	17278

$$AIC = 2 \times \log\text{-likelihood} - (2k - 1)2$$

$$BIC = 2 \times \log\text{-likelihood} - (2k - 1) \log(n)$$

# Gradient Function Graph

for Heroin Users in BKK Drug User Study



# Overview

- Motivation and Background
- Truncated Mixtures or Mixtures of Truncated Distributions?
- **Some Equivalence Results**
  - Model Spaces and Likelihood Surfaces
  - Model Transformations
  - Population Size Estimates



# Some equivalence results: How are dual and primal model related?

- Both share same model spaces!
- Both share the same likelihood surfaces!
- MLEs can be explicitly transformed into each other
- $\hat{N} = \hat{N}'$

# Model Spaces

Primal:

$$M' = \left\{ (m'_1, m'_2, m'_3, \dots)^T \mid m'_i = \frac{\sum_{j=1}^k q'_j Po(i, \lambda'_j)}{1 - \sum_{j=1}^k q'_j Po(0, \lambda'_j)} \right\}$$

Dual:

$$M = \left\{ (m_1, m_2, m_3, \dots)^T \mid m_i = \sum_{j=1}^k q_j \frac{Po(i, \lambda_j)}{1 - Po(0, \lambda_j)} \right\}$$

# Model Spaces

$$M' = M$$

# Proof (a): $M' \subseteq M$

$$m' \in M' \text{ with } m'_i = \frac{\sum_{j=1}^k q'_j Po(i, \lambda'_j)}{1 - \sum_{j=1}^k q'_j Po(0, \lambda'_j)}$$

$$\text{define } q_j = \frac{q'_j (1 - Po(0, \lambda'_j))}{\sum_{j=1}^k q'_j (1 - Po(0, \lambda'_j))}$$

$$\Rightarrow \sum_{j=1}^k q_j \frac{Po(i, \lambda'_j)}{1 - Po(0, \lambda'_j)} = m'_i \Rightarrow m' \in M$$

# Proof (b): $M \subseteq M'$

$$m \in M \text{ with } m_i = \sum_{j=1}^k q_j \frac{Po(i, \lambda_j)}{1 - Po(0, \lambda_j)}$$

$$\text{define } q'_j = \frac{q_j / (1 - Po(0, \lambda_j))}{\sum_{j=1}^k q_j / (1 - Po(0, \lambda_j))}$$

$$\Rightarrow \frac{\sum_{j=1}^k q'_j Po(i, \lambda_j)}{1 - \sum_{j=1}^k q'_j Po(0, \lambda_j)} = m_i \Rightarrow m \in M'$$

# Model Spaces

$$M' = M$$

$$\Rightarrow \{L(m') \mid m' \in M'\} = \{L(m) \mid m \in M\}$$

$$\text{with } L(m') := \sum_i n_i \log(m'_i)$$

$\Rightarrow$  NPMLs agree for both models

$$\Rightarrow \hat{N}' = \frac{n}{1 - \sum_{j=1}^k q'_j e^{-\lambda'_j}} = \hat{N} = n \sum_{j=1}^k \frac{q_j}{1 - e^{-\lambda_j}}$$

# Epilogue

- Can we estimate something which is hidden or unobserved?
- And if, how valid is such an estimate?

# Australian Screening Study for Colon Cancer

- 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 at least once and cancer status evaluated
- 196 were confirmed positive to have bowel cancer
- How many of 35,000 **unconfirmed** negative have bowel cancer?



38,000 men screened

3,000 tested positive  
at least once

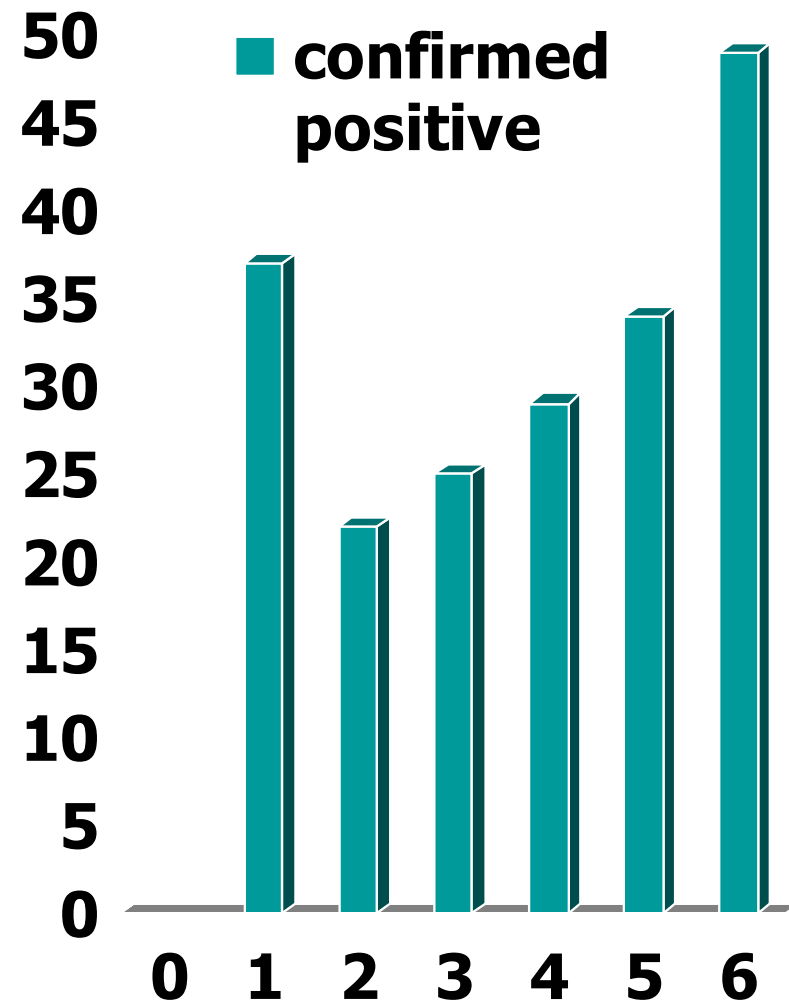
196 confirmed  
positive



Men with cancer, but  
tested always negative

# The counting distribution: a recent example from screening

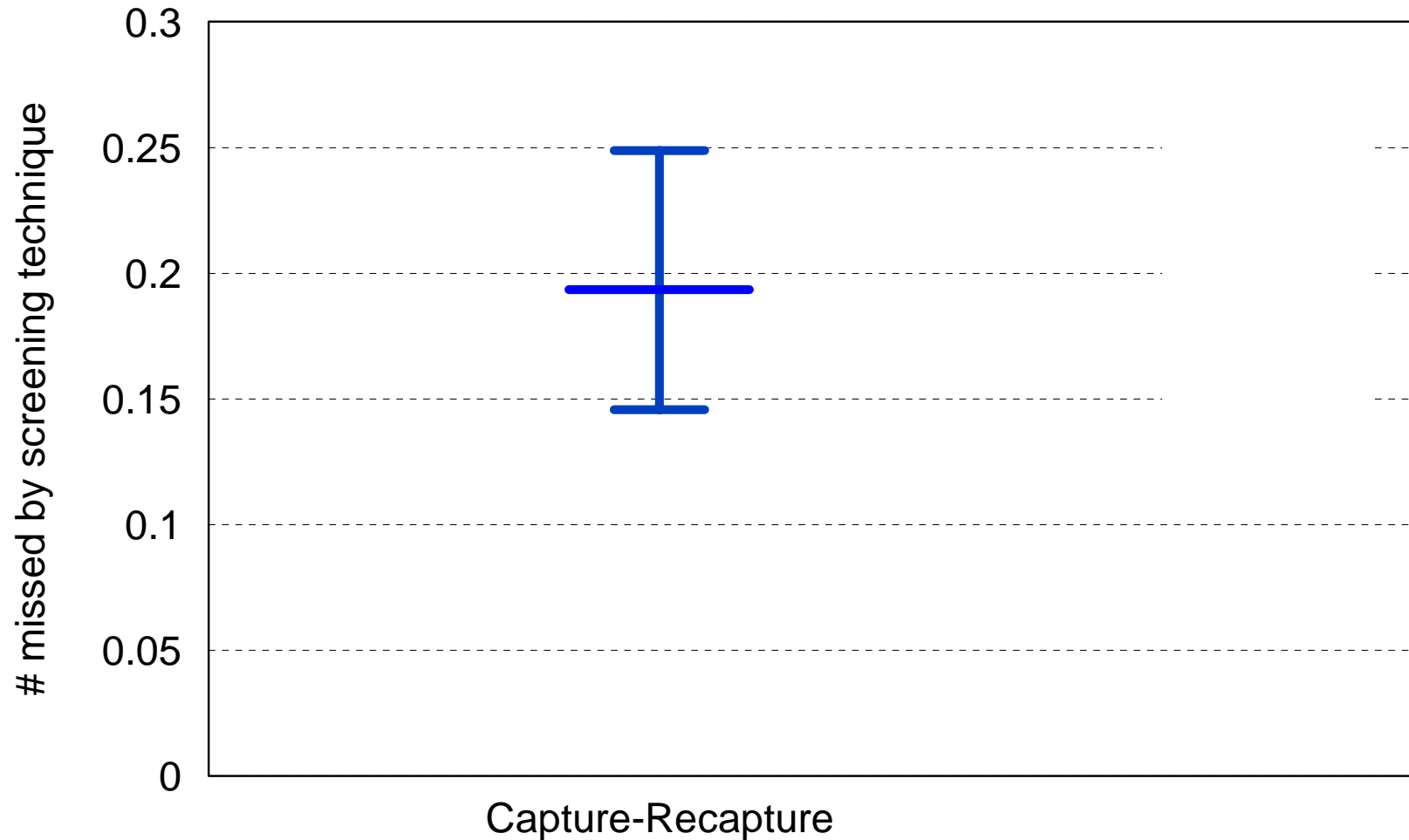
- frequency  $n_0$  of those tested negative at all 6 times with bowel cancer is unknown
- an estimate of  $n_0$  might be constructed from the distribution  $n_1, n_2, n_3, \dots$  of counts



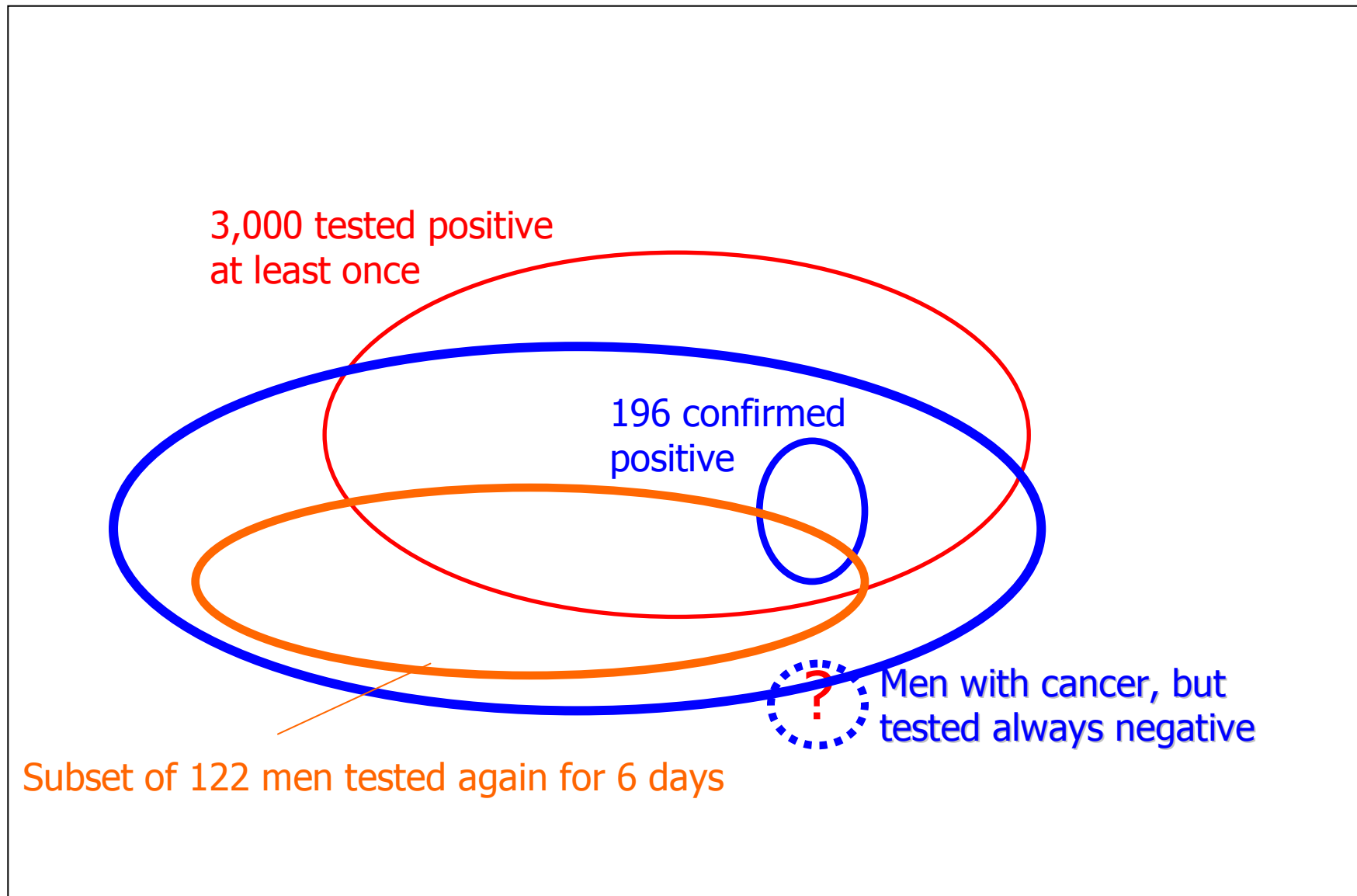
# Results from ML

k	$\lambda_j$	$q_j$	L	$n_0$	N
1	0.6241	1	-436.72	1	197
2	0.8548	0.5664	-349.09	13	209
	0.2821	0.4336			
3	0.9352	0.3452	-344.18	47	243
(NPMLE)	0.5971	0.4199			
	0.1088	0.2349			

# Number missed by screening technique

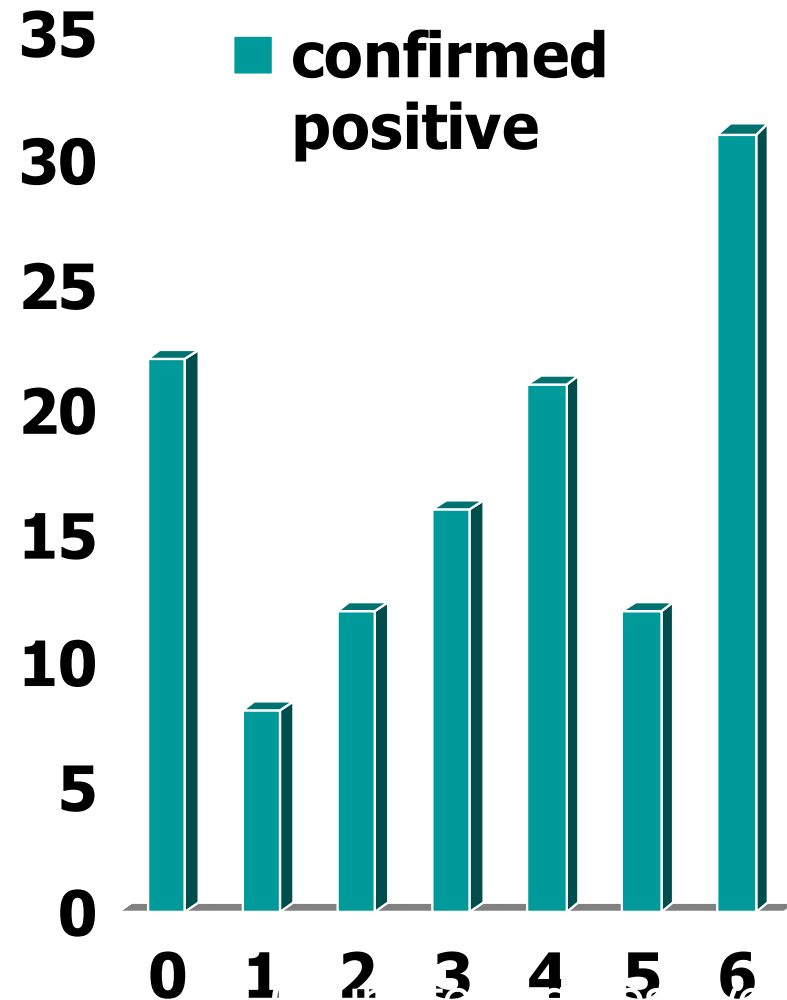


38,000 men screened

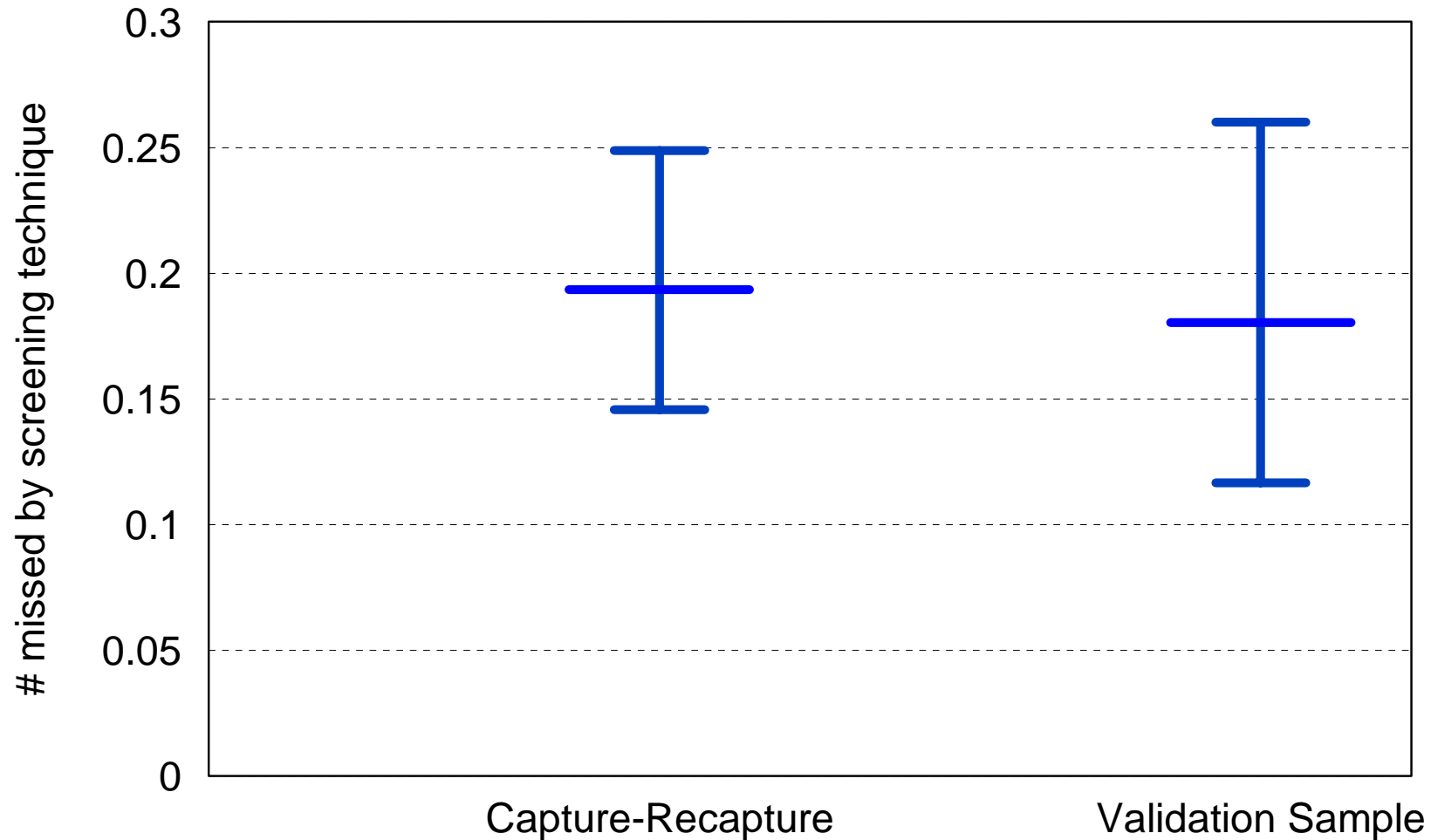


# Distribution of counting the number of days testing positive for 122 men with confirmed colon cancer

- Now frequency  $n_0$  of those tested negative at all 6 times with bowel cancer is known
- validation sample



# Relative number missed by screening technique







# In Summary

- Zero-truncated Count Mixtures Historical important and more useful for applications
- Dual Model offers better properties (global NMLE, Convergences of algorithms, ...)
- Need to be transported to practitioners