School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

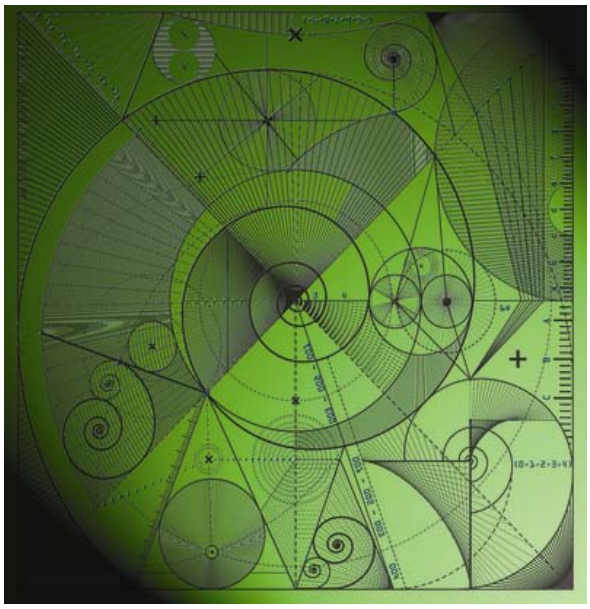
Preprint MPS_2011-01

25 January 2011

Use of the Ratio Plot in Capture-Recapture Estimation

by

Dankmar Böhning, M. Fazil Baksh, Rattana Lerdsuwansri and James Gallagher



Use of the Ratio Plot in Capture-Recapture Estimation

Dankmar Böhning, M. Fazil Baksh, Rattana Lerdsuwansri, and James Gallagher

January 25, 2011

Dankmar Böhning
Department of Mathematics and Statistics
School of Mathematical and Physical Sciences
University of Reading, Reading RG6 6BX, England
E-mail : d.a.w.bohning@reading.ac.uk
Tel: +44 (0)118 378 6211 Fax: +44 (0)118 378 8032

M. Fazil Baksh (address as above)
E-mail : m.f.baksh@reading.ac.uk

Rattana Lerdsuwansri (address as above)
E-mail : r.lerdsuwansri@reading.ac.uk

James Gallagher
Statistical Services Centre
School of Mathematical and Physical Sciences
University of Reading, Reading RG6 6FN, England
E-mail : j.gallagher@reading.ac.uk

Abstract

In this paper we present and assess a graphical device for choosing a method for estimating population size in capture-recapture studies. The basic concept is derived from a homogeneous Poisson distribution, where the ratios of neighboring Poisson probabilities multiplied by the value of the larger neighbor count are constant. This property extends to the zero-truncated Poisson distribution which is of fundamental importance in capture–recapture studies. The ratio plot can be used for assessing specific departures from a Poisson distributions. For example, simple contaminations of an otherwise homogeneous Poisson model can be easily detected and a robust estimator for the population size can be suggested. Several robust estimators are developed and a simulation study is provided to give some guidance on which should be used in practice. More systematic departures can also easily be detected using the ratio plot. In this paper focus is on Gamma-mixtures of the Poisson distribution which leads to a linear pattern in the ratio plot. More generally, the paper shows that the ratio plot is monotone for arbitrary mixtures of power series densities.

Keywords: Capture-recapture; Chao and robust and generalized Chao estimator; Turing estimator; robust Turing estimator; generalized Turing estimator; Poisson-

gamma model, ratio plot, structured heterogeneity.

1 Introduction

Capture-recapture studies, concerned with estimating the size of populations that are hidden or difficult to enumerate, make use of some "capture" mechanism (e.g. live trapping, register, surveillance system) capable of repeatedly identifying observational units in time, or in clusters (Bunge and Fitzpatrick 1993; Chao *et al.* 2001). Capture–recapture methods are now widely used in a variety of application areas, including public health and epidemiology, clinical medicine, bioinformatics (estimating biodiversity), criminology and terroristic research, systems engineering (estimating the number of unknown errors in a software) as well as investigating forms of deviating behavior in social sciences, in addition to the traditional field of wildlife biology/ecology. As a consequence, the statistical community has developed a major interest in the use of capture–recapture methods.

For studies based on repeated sampling in time there is an observational period in which each member (unit) of the target population can be potentially detected on several occasions. An example of sampling in time taken from Chao and Huggins (2005) is reproduced in Table 1. Here, the number of detections of female grizzly bears with cubs-of-the-year for three different observational periods were

recorded in a study of the bear population in Yellowstone from 1996 to 1998. For instance, in 1996 a total of 15 female bears were observed exactly once, 10 exactly twice and so on, leading to a total of 45 detections of 28 bears.

Table 1: *Female Grizzly Bears in the Yellowstone ecosystem*

	Frequency of detection							Number of observed bears	Number of detections
Year	f_1	f_2	f_3	f_4	f_5	f_6	f_7	n	S
1996	15	10	2	1	0	0	0	28	45
1997	13	7	4	1	3	0	1	29	65
1998	11	13	5	1	1	0	2	33	75

For a study of a population of size N units, let X_i denote the number of times unit i is detected in the observational period, $i = 1, 2, \ldots, N$ and let $p_x = P(X_i = x)$. Also, let f_x denote the frequency of units detected exactly x times, $x = 0, \ldots, m$ and let n denote the number of observed units. As $X_i = 0$ is *not* observed, the corresponding frequency $f_0 = N - n$ is unknown and, in order to obtain an estimate for N, may be replaced by its expected value Np_0. When p_0 is known, this leads to the familiar Horvitz-Thompson estimator of population size

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

In most problems, p_0 is unknown and itself has to be estimated. Under the assumption that each X_i follows a Poisson distribution with parameter λ, we obtain

$p_0 = \exp(-\lambda)$ and consequently $\hat{N} = n/[1 - \exp(-\hat{\lambda})]$ where $\hat{\lambda}$ is an estimate of λ. In the well-known Turing or Good-Turing estimator $\hat{N} = n/(1 - f_1/S)$ (Good 1953), the estimate of p_0 is $\hat{p}_0 = f_1/S$ where $S = f_1 + 2f_2 + ... + mf_m$. Another approach uses the maximum likelihood estimate of λ. It should be emphasized that both these estimates of population size are only appropriate under the homogeneous Poisson model.

The above notation can also be used for studies based on multiple detections within a cluster (e.g. herd, village, household). Here N is the total number of clusters, X_i is the number of units detected in cluster i, $i = 1, 2, \ldots, N$ and f_x is the frequency of clusters with exactly x units detected, $x = 0, \ldots, m$. An example of repeated identifications in clusters (herds) is provided by Böhning and Del Rio Vilas (2008) who examined scrapie occurrence in Great Britain based upon the Scrapie Notifications Database (SND). The frequency distribution of the case count per herd is shown in Table 2.

Table 2: *Scrapie surveillance in Great Britain based upon the Scrapie Notifications Database (SND)*

Year	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_{8+}	n
2002	74	23	15	6	8	3	3	12	144
2003	66	29	12	2	3	2	3	17	134
2004	83	29	14	6	5	6	0	8	151

The probability of the inclusion of an individual or unit in a capture-recapture study frequently depends on measured covariates such as age, gender and size, as well as on unobserved factors. This heterogeneity often invalidates the assumption that the X_i's are identically distributed. If this heterogeneity is ignored the estimators of population size can be severely negatively biased (Böhning and Schön 2005, van der Heijden *et al.* 2003). Heterogeneity is closely connected to the occurrence of over-dispersion. Recently (Baksh *et al.* 2011) a distribution-free test procedure to detect over-dispersion has been suggested which modifies a previously developed over-dispersion test for zero-truncated data. A method to account for heterogeneity in the estimation of population size (Chao 1987) models the Poisson parameter as a random variable with a latent *heterogeneity distribution* $\lambda(t)$. This gives

$$p_x(\lambda) = \int_0^\infty \frac{\exp(-t)t^x}{x!}\, \lambda(t)dt\ . \tag{2}$$

Here, we exploit the above model for p_x to develop a graphical method for identifying heterogeneity in capture-recapture data. In particular, we provide a tool for assessing if the homogeneous Poisson model, with and without contaminations, is appropriate, or whether or not there is structured heterogeneity in the observed data. The contaminated Poisson model and structured heterogeneity will be discussed in the next section. In addition, we develop further a number of common

estimators, and evaluate their performance under different heterogeneity assumptions.

2 The Ratio Plot

Define $r_x = \frac{(x+1)p_{x+1}}{p_x}$ where $x = 0, 1, 2, \ldots$. For the Poisson random variable X with mean λ, it is clear that $r_x = \lambda$ and hence the plot of r_x against x is a horizontal line. The concept of using a graphical device for deciding about the suitability of the Poisson model has been proposed in the literature at various points in time. Hoaglin (1980) suggested using that $\log p_x + \log x! = -\lambda + x \log \lambda$ is linear in x under the Poisson assumption. The associated plot of $\log f_x + \log x!$ against x has been called the *Poissonness Plot*. Gart (1970) mentions the possibility of plotting an estimate of r_x against x. In practice, plotting the ratio has not been widely used. Occasionally, it occurs in textbooks as marginal notes, such as in Pawitan (2001; p.110) in an exercise. In capture-recapture studies, zero-counts are truncated. Hence, observed sample frequencies $f_1, f_2, ...$ arise from the zero-truncated distribution $p_x/(1 - p_0)$. However, the ratio plots are for both cases identical since

$$r_x = \frac{(x+1)p_{x+1}}{p_x} = \frac{(x+1)p_{x+1}/(1-p_0)}{p_x/(1-p_0)}.$$

This is an important result, making the ratio plot applicable to the capture-recapture scenario (with zero-truncated count distributions). In practice the ratio r_x is estimated by

$$\hat{r}_x = \frac{(x+1)f_{x+1}}{f_x}$$

while the graph of $\hat{r}_x$ against x is called the ratio plot. A horizontal line is consistent with homogeneous Poisson observations; conversely, departures from a horizontal line provide evidence for violation of Poisson homogeneity. For example, the plot in Figure 1 of the grizzly bears data for 1997 (Table 1) clearly shows that the frequency of 5 sightings is larger than expected under a homogeneous Poisson model, but there is no evidence to suggest that the other observed frequencies violate the homogeneity assumption. In studies with such contaminated Poisson data $S = f_1 + 2f_2 + ... + mf_m$ will be too large and consequently the Good-Turing estimate will be biased downward.

The ratio plot of the Scrapie data in Table 2 is given in Figure 2. Again, for all three years, there is a clear suggestion that a homogeneous Poisson model is inappropriate. However, it is unclear whether the cause is due to contaminations or whether there is some latent structure to the data causing the ratios to increase with x.

The general importance of the concept of the ratio plot stems from the following

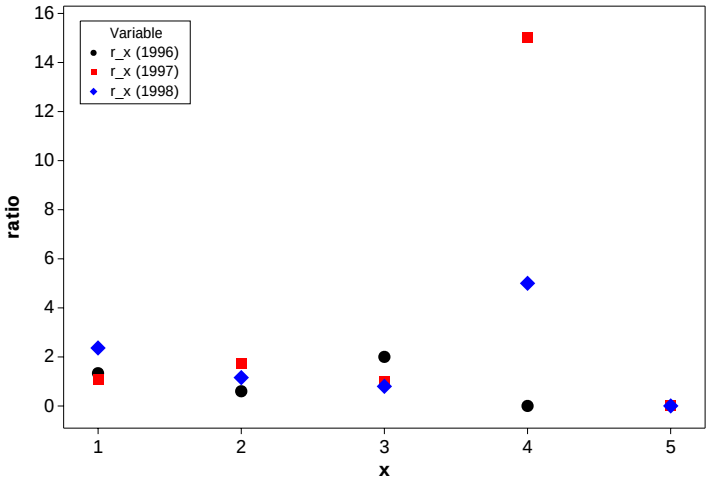


Figure 1: Ratio plot of observed Grizzly bears in the Yellowstone ecosystem for the period 1996-1998

result which, in essence, says that under arbitrary mixing on the Poisson parameter the ratio plot should show a monotone increasing pattern.

Theorem 1 *Let p_x be given according to (2). Then, the following monotonicity result holds:*

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

A proof of this theorem is provided in the appendix. Note the special case of

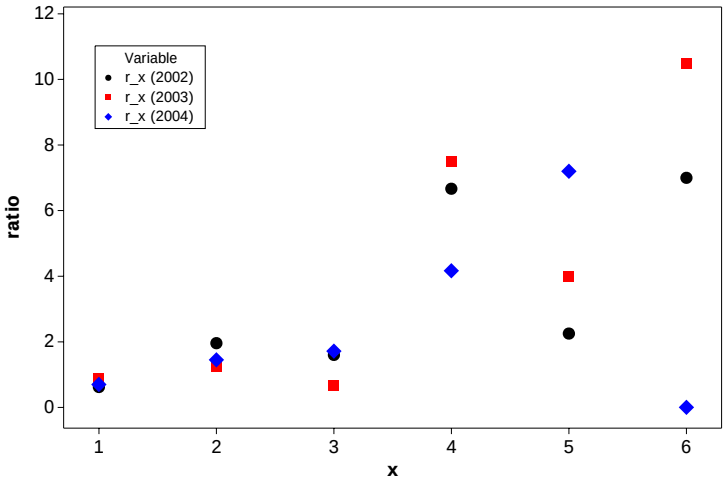


Figure 2: Ratio plot of observed scrapie infected herds in Great Britain based upon the Scrapie Notifications Database (SND) for the period 2002-2004

Poisson homogeneity is included as all inequalities become equalities. In the remainder of this paper we examine specific departures from Poisson homogeneity as well as specific forms of monotonicity. For example:

- is there a contamination of an underlying, but otherwise, homogeneous Poisson model?

- or, is there a form of monotonicity which can be described by a simple

monotone structure such as a straight line with positive slope (structured heterogeneity)?

- or, is there no recognizable form of monotone pattern (unstructured heterogeneity)?

The next two sections consider population size estimation for the first two of the above departures from Poisson homogeneity.

3 The Robust Turing Estimator

For $k = 1, 2, 3, \ldots, m-1$ define the robust estimator for λ by

$$\hat{\lambda}_k = \frac{\sum_{x=1}^{k}(x+1)f_{x+1}}{\sum_{x=1}^{k} f_x} \,. \tag{4}$$

The form of $\hat{\lambda}_k$ is very similar to an estimator for λ suggested by Moore (1952) in the case of a Poisson distribution with larger counts truncated. Estimating p_0 in equation (1) by $(f_1/N)/\hat{\lambda}_k$ (as under Poisson homogeneity $p_0 = p_1/\lambda$) we obtain a modification of the Turing estimator of population size for contaminated Poisson data as a solution of the equation $N = n/[1 - (f_1/N)/\hat{\lambda}_k]$ for N as

$$\hat{N}_k = n + \frac{f_1}{\hat{\lambda}_k} \,. \tag{5}$$

We will call $\hat{N}_k$ the *robust Turing* estimator since it will be less influenced by large, contaminating observations than the original Turing estimator, in particular, if k is

chosen small. In the case where $k = 1$ we have $\hat{\lambda}_1 = 2f_2/f_1$ which is identical to the Zelterman (1988) estimator of the Poisson parameter. Zelterman showed that this estimator was more robust against mis-specification of the Poisson model than the estimator based on the maximum likelihood estimate of λ. This is intuitively clear since the estimator remains unchanged for distributional changes associated with counts larger than 2. The corresponding estimator for the population size $\hat{N}_1$ becomes $n + f_1^2/(2f_2)$ which is the lower bound estimator of Chao (1987, 1989).

3.1 An optimality property of the Robust Turing Estimator

The beneficial behavior of $\hat{N}_k$ can be seen in the following result for a simple contamination model in which the Poisson distribution is contaminated by a second Poisson component with weight α.

Theorem 2 *Let X be a discrete random variable with probability mass function* $p_x = (1-\alpha)Po(x;\lambda)+\alpha Po(x;\mu)$*, where* $Po(x;\nu) = e^{-\nu}\nu^x/x!$ *for* $x = 0, 1, 2, \ldots$ *and* $0 \leq \alpha \leq 1$*; Also, let* $f_1, \cdots, f_m$ *be the observed frequencies in sample of size* n *from* p_x *with largest observed count* m*. If* $\hat{N}_k = n + f_1/\hat{\lambda}_k$ *denotes the robust Turing estimator for* $1 \leq k \leq m-1$*, then*

$$lim_{N\to\infty} E(\hat{N}_k)/N = (1 - p_0) + p_1 \frac{p_1 + ... + p_k}{2p_2 + ... + kp_{k+1}}$$

$$\to 1 \text{ as } \mu \to \infty.$$

In addition, for the Turing estimator we have

$$lim_{N\to\infty}E(\hat{N})/N = lim_{N\to\infty}E(\frac{n}{1-f_1/S})/N = \frac{(1-p_0)}{1-p_1/E(X)}$$

$$\to [1-(1-\alpha)\exp(-\lambda)] \leq 1 \textit{ as } \mu \to \infty.$$

Proof:

For sufficiently large N we can replace f_x by its expected value Np_x to get

$$\frac{E(N_k)}{N} \to (1-p_0) + p_1\frac{p_1+\ldots+p_k}{2p_2+\ldots+(k+1)p_{k+1}} \text{ as } N \to \infty \ .$$

The result follows from the fact that

$$p_x \to (1-\alpha)\frac{e^{-\lambda}\lambda^x}{x!} \ ,$$

as $\mu \to \infty$, since then

$$p_1\frac{p_1+\ldots+p_k}{2p_2+\ldots+(k+1)p_{k+1}}$$
$$= (1-\alpha)e^{-\lambda}\lambda\frac{(1-\alpha)e^{-\lambda}\lambda+\ldots+(1-\alpha)e^{-\lambda}\lambda^k/k!}{(1-\alpha)e^{-\lambda}\lambda^2+\ldots+(1-\alpha)e^{-\lambda}\lambda^{k+1}/k!} = (1-\alpha)e^{-\lambda} = p_0$$

For the Turing estimator, note that $E(X) = (1-\alpha)\lambda+\alpha\mu$ goes to infinity if μ becomes large. So $p_1/E(X)$ goes to zero and a persistent bias of $-(1-\alpha)exp(-\lambda)$ remains. □

The theorem shows that the Turing estimator is sensitive to large contaminating counts whereas the robust Turing estimator is less affected. This is further explored using simulations, here with $N = 100$ and $p_x = (1-\alpha)Po(x;\lambda) +$

$\alpha Po(x;\mu)$ for $x = 0, 1, 2, ...,$ $\alpha = 0.5$ and $\lambda = 0.5$. The results with respect to bias and mean squared error (MSE) are given in Figure 3 and Figure 4, respectively. Figure 3 shows the expected ordering of bias in the sense

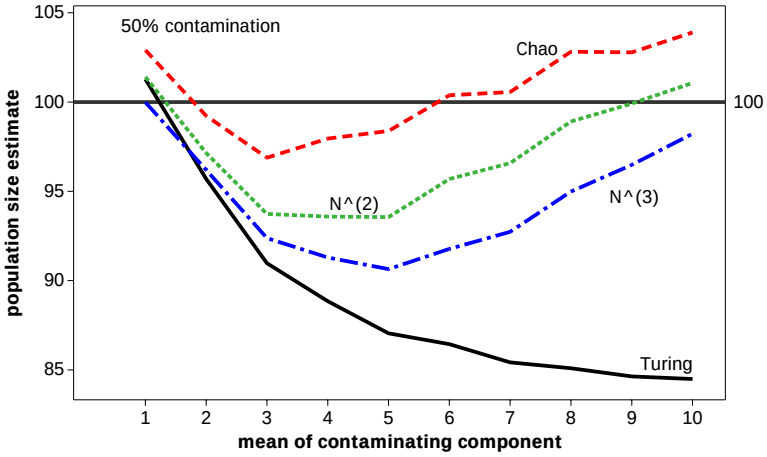


Figure 3: Mean population size estimator in contamination model $p_x = (1 - \alpha)Po(x;\lambda) + \alpha Po(x;\mu)$ for $N = 100$; $\hat{\text{N}}$(k) denotes the robust Turing estimator $\hat{N}_k$ and 'Chao' corresponds to $\hat{N}_1$

$bias^2(\hat{N}_1) \leq bias^2(\hat{N}_2) \leq bias^2(\hat{N}_3)$. The figure also illustrates the undesirable behavior of the Turing estimator when contaminations increase. Hence it is interesting to look at the mean squared error as a summary measure of bias and

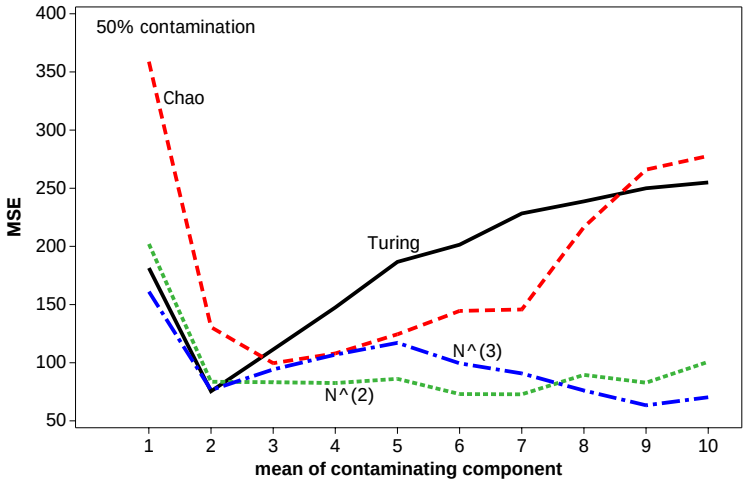


Figure 4: Mean squared error of population size estimator in contamination model $p_x = (1-\alpha)Po(x;\lambda) + \alpha Po(x;\mu)$ for $N = 100$; N$\hat{}$(k) denotes the robust Turing estimator $\hat{N}_k$ and 'Chao' corresponds to $\hat{N}_1$

variance in Figure 4. Once again, the Turing estimator does not perform well. There is evidence that Chao's estimator can be improved upon by using the robust Turing estimator $\hat{N}_2$ or $\hat{N}_3$.

3.2 Finding the upper truncation point k

In practice, evidently a value for k is required in order to calculate $\hat{N}_k$. Again the ratio plot can be used. For example, from the plot in Figure 1 of the grizzly

bears data for 1997 we deduced that the ratio r_4 is larger than expected under a homogeneous Poisson model. We suggest that this is formally tested using the following χ^2 test based upon the truncated distribution

$$\chi^2(k) = \sum_{x=1}^{k+1} \frac{[f_x - n_k Po_+(x; \hat{\lambda}_k)]^2}{n_k Po_+(x; \hat{\lambda}_k)} \tag{6}$$

where $\hat{\lambda}_k$ is given by equation (4), $Po_+(x;\lambda) = Po(x;\lambda)/[Po(1;\lambda)+...+Po(k+1;\lambda)]$ and $n_k = f_1 + ... + f_{k+1}$. Under the null hypothesis of a homogeneous Poisson model this statistic approximately follows a χ^2_{k-1} distribution. Note that for $k = 1$ a perfect fit is achieved, resulting in no degrees of freedom. Table 3 shows the significance tests for the grizzly bear data for 1997 along with the modified Turing estimates of population size for different values of k. These results support the findings from the ratio plot; for $k = 4$ we have borderline significance at the 5% level. Hence, we conclude that the robust estimate for the mean parameter is $\hat{\lambda}_3 = 1.25$ (using $k = 3$) and $\hat{N}_3 = 39.4$. As expected, the Turing estimate (36.25) is smaller.

Another illustration of applying the ratio plot uses data from a study of Cullen *at al.* (1990) on dystrophin density in human muscle (see also Matthews and Appleton 1993). Dystrophin, a gene product of possible importance in muscular dystrophies, can be located within muscle fibers using an electron microscope . Units (epitops) of dystrophin cannot be detected until they have been labelled

Table 3: Robust Turing estimates of the number of Female Grizzly Bears in the Yellowstone ecosystem for 1997

k	$\chi^2(k)$	p-value	$\hat{\lambda}_k$	$\hat{N}_k$
1	0.000	1.000	1.08	41.1
2	0.241	0.623	1.30	39.0
3	0.264	0.876	1.25	39.4
4	7.627	0.054	1.80	36.2
5	10.473	0.033	1.61	37.1

by a suitable electron-dense substance such as gold-conjugated antibodies which adhere to the dystrophin. Not all units can be labelled and more than one anti-body molecule may attach to a dystrophin unit. To achieve an unbiased estimate of the dystrophin density, it is important to account for all labelled and unlabelled units. Table 4 shows the observed count of the number of antibody molecules on each dystrophin unit within the muscle fibres of biopsy specimens taken from normal patients. Interest is in f_0, the number of unobserved (unlabelled) dystrophin units.

Table 4: Distribution of antibody counts attached to dystrophin units

f_0	f_1	f_2	f_3	f_4	f_5	n
-	122	50	18	4	4	198

Figure 5 shows the ratio plot (on log-scale) for the dystrophin data. Also shown are 95% confidence limits using $\log(\hat{r}_x) \pm 1.96 * \sqrt{\text{Var}[\log(\hat{r}_x)]}$ where $\text{Var}[\log(\hat{r}_x)] = 1/f_{x+1} + 1/f_x$ (Böhning 2008). Although there is progressively less reliability in

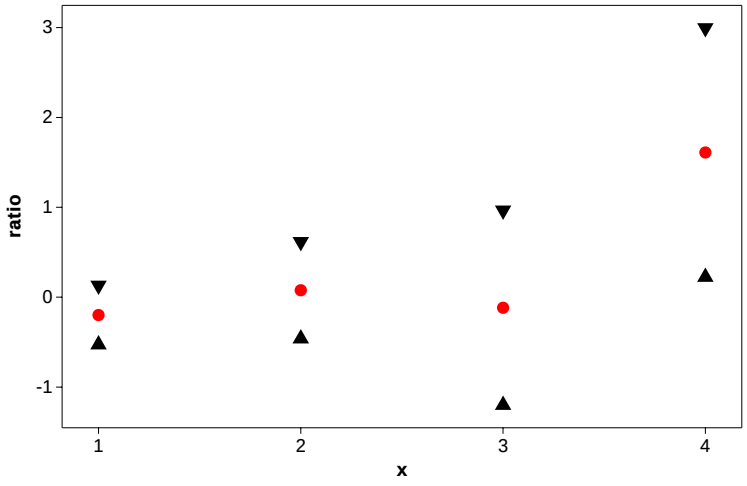


Figure 5: Ratio plot (on log-scale) for the dystrophin data (bullets) with approximate 95% confidence limits (upper and lower triangle)

the estimated ratios, nonetheless there is evidence that frequency f_5 is contaminated. This assertion is supported by the χ^2-test (see Table 5). It is interesting to note that the robust estimate (334; $k = 3$) is larger than the Plackett, MLE and MVUE estimates of 321, 315 and 313, respectively (see also Matthews and Appleton 1993 where also the values for the estimators Plackett, MLE and MVUE are taken from).

Table 5: Robust Turing estimates of the number of dystrophin units

k	$\chi^2(k)$	p-value	$\hat{\lambda}_k$	$\hat{N}_k$
1	0	1	0.82	347
2	0.53	0.47	0.90	334
3	0.58	0.75	0.89	334
4	12.00	0.01	0.98	323

4 The Ratio Plot under Structured Heterogeneity and the Generalised Turing Estimator

4.1 Structured heterogeneity

To illustrate the situation of *structured heterogeneity* we begin with an example from illicit drug user research. The data set comes from a study concerned with estimating hidden intravenous drug users in Los Angeles (Hser 1993). Intravenous drug users in Los Angeles county were entered into the California Drug Abuse Data System (CAL-DADS): Table 6 shows the frequency distribution of

the episode (contact with treatment center) count per drug user in the year 1989, and the ratio plot is in Figure 6. The most interesting feature of this plot is the apparent linear trend with positive slope. As suggested earlier, this is evidence in support of violation of Poisson homogeneity. Furthermore, as shown below, this is indicative of structured heterogeneity due to a latent Gamma distribution of the mean parameter.

Definition 1 *The ratio plot exhibits structured heterogeneity if*

$$r_x = \alpha + \beta x$$

with $\beta > 0$. *The case* $\beta = 0$ *exhibits Poisson homogeneity.*

Table 6: The frequency distribution of the episode count per drug user in Los Angeles for 1989 as obtained from the California Drug Abuse Data System ($n = 20,198$)

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
-	11,982	3,893	1,959	1,002	575	340	214	90	72	36	21	14

The question arises for which mixing distribution $\lambda(t)$ does *structured heterogeneity* arise. This is now partly answered. Using equation (2), suppose that $\lambda(t)$ is the Γ density with parameters π and κ. Then $p_x = \frac{\Gamma(\kappa+x)}{\Gamma(x+1)\Gamma(\kappa)}\pi^\kappa(1-\pi)^x$ is the negative binomial density with event parameter π and shape parameter κ, and

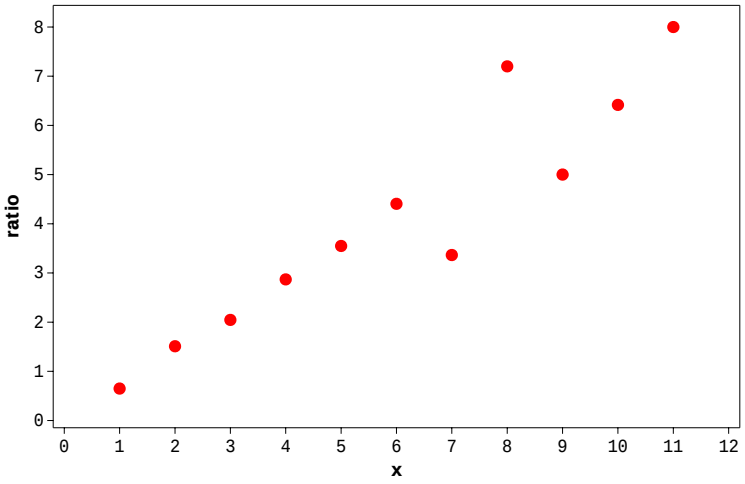


Figure 6: Ratio plot of episode count per drug user in Los Angeles in 1989

consequently $r_x = (x + \kappa)(1 - \pi)$. It follows that the ratio plot is expected to be a straight line with slope $1 - \pi$ and intercept $\kappa(1 - \pi)$. Hence, structured heterogeneity in the ratio plot relates to a prominent class of mixing distributions, the Gamma-distribution or in its marginal form, the negative-binomial. These forms of structured heterogeneity arise frequently in capture-recapture data (Dorazio and Royle 2005; Pledger 2005; Chao and Bunge 2002).

4.2 The Generalised Turing Estimator

Furthermore, since $p_0 = \pi^\kappa$, $p_1 = \kappa\pi^\kappa(1-\pi)$, $E(X) = \kappa(1-\pi)/\pi$, we have that $\pi_1/E(X) = \pi^{k+1}$ and $p_0 = [\pi^{\kappa+1}]^{\kappa/(\kappa+1)} = [p_1/E(X)]^{\kappa/(\kappa+1)}$. This leads to the generalised Turing estimator

$$\hat{N}_{GT} = \frac{n}{1-(\frac{f_1}{S})^{\kappa/(\kappa+1)}} \; . \tag{7}$$

Theorem 3 *Let* $p_x = \frac{\Gamma(\kappa+x)}{\Gamma(x+1)\Gamma(\kappa)}\pi^\kappa(1-\pi)^x$. *Then, we have the following property for the generalised Turing estimator:*

$$E(\hat{N}_{GT})/N \to (1-\pi^k)/[1-(\pi^{\kappa+1})^{\kappa/(\kappa+1)}] = 1,$$

for $N \to \infty$.

Proof: For sufficiently large N, $E(f_1/N)/E(S/N) \to \pi^{\kappa+1}$ and $E(n/N) \to (1-\pi^\kappa)$ so that

$$E(\hat{N}_{GT})/N \to \frac{N(1-\pi^\kappa)}{N[1-(\pi^{\kappa+1})^{\kappa/(\kappa+1)}]} = 1,$$

with N becoming large. □

The form of the generalised Turing estimator is interesting. It uses the frequency f_1 of units detected only once which is usually a large quantity. And it also uses S which makes use of all the information in the sample. This is in contrast, for

example, to Chao's estimator $n + f_1^2/(2f_2)$ which uses only the frequencies of counts equal to one and two. Clearly, to make the generalised Turing estimator work practically, we need to have an estimate for κ. This can be accomplished by utilizing the ratio plot and constructing a weighted regression estimator for the regression coefficients in $r_x = \alpha + \beta x$ with a diagonal weight matrix containing the inverse variances of $\hat{r}_x = (x+1)f_{x+1}/f_x$ as entries (Böhning 2008). An estimate for κ can then be given as $\hat{\alpha}/\hat{\beta}$.

We demonstrate the application of these methods with a further case study, again from illicit drug user research. Hay and Smit (2003) collated data on individuals who have visited a Scottish needle exchange during 1997. Hay and Smit (2003) preferred not to explicitly state the needle exchange from which they obtained these data. The authors stated however, that *"the data were collated during a programme of drug misuse prevalence research in Scotland and was the only one operating in that area at that time. The needle exchange assigns a unique identifier number to each individual accessing the service, thus enabling it to produce statistics on the number of people who had contacted the service over a given period."* We show these data in Table 7. For these data (as it is the case also with many other data sets) it should be noted that the ratio plot shows strong indication of *exponential* mixing. That is the ratio plot is consistent with a (truncated) *geo-*

Table 7: The frequency distribution of the episode count per participant in a needle exchange program in Scotland for 1997 (Hay and Smit 2003)

f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11+}	n
-	175	85	50	47	37	38	32	16	17	17	133	647

metric distribution, since the Poisson-exponential mixture is a geometric density. The latter is a special case of the negative binomial distribution and occurs for $\kappa = 1$. To identify this case most easily from the ratio plot we define a different but essentially equivalent form:

$$r'_x = xp_x/p_{x-1} = [x + (\kappa - 1)](1 - \pi) \text{ for } x = 1, ..., m,$$

so that the case of exponential mixing ($\kappa = 1$) results in a ratio plot which is a line passing through the origin. Figure 7 shows the associated empirical ratio plot for the participants of the Scottish needle exchange programme with fitted regression line. Clearly, the line passes very near to the origin. Hence we take $\kappa = 1$ and the generalised Turing estimator to be $\hat{N}_{GT} = \frac{n}{1-\sqrt{f_1/S}}$. In our case, we have $n = 647$, $f_1 = 175$ and $S = 3596$ leading to $\hat{N}_{GT} = 830$ which clearly adjusts for a likely underestimation bias of the conventional Turing estimator $\hat{N}_T = 680$.

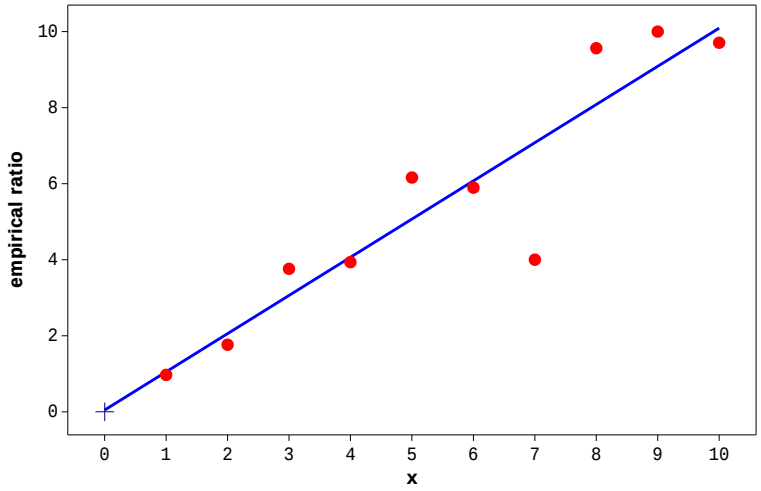


Figure 7: Ratio plot of episode count per participant in for 1989 with fitted regression line

5 Residual Heterogeneity

Let us now assume that a Poisson-Gamma mixture

$$p_x(\pi,\kappa) = \int_0^\infty \frac{\exp(-t)t^x}{x!}\,\lambda(t)dt = \frac{\Gamma(\kappa+x)}{\Gamma(x+1)\Gamma(\kappa)}\pi^\kappa(1-\pi)^x \qquad (8)$$

has been successfully identified. Clearly, (8) incorporates all available structured heterogeneity. The question arises whether there is any remaining *residual, unstructured heterogeneity* in the data. Note that, conditional upon κ, the negative binomial density is part of the power series family $p_x = a_x t^x \mu(t)$ with

$a_x = \frac{\Gamma(\kappa+x)}{\Gamma(x+1)\Gamma(\kappa)}$ and $\mu(t) = (1-t)^\kappa$. Hence, we can consider mixing the negative binomial $\frac{\Gamma(\kappa+x)}{\Gamma(x+1)\Gamma(\kappa)}\pi^\kappa(1-\pi)^x$ together with some arbitrary mixing density $\lambda(t)$:

$$g_x(\lambda|\kappa) = \int_0^1 \frac{\Gamma(\kappa+x)}{\Gamma(x+1)\Gamma(\kappa)}(1-t)^\kappa t^x \ \lambda(t)dt = \int_0^1 a_x\mu(t)t^x \ \lambda(t)dt, \qquad (9)$$

and we can apply the general monotonicity result of the appendix, showing that the *generalised ratio plot* $r_x = \frac{g_{x+1}/a_{x+1}}{g_x/a_x}$ vs. x should show a monotone increasing pattern if heterogeneity is still present. If there is residual homogeneity the generalised ratio plot reduces to a horizontal line.

This property of the Poisson, namely mixing a Poisson with a Gamma resulting in a negative binomial which, if again, mixed with an arbitrary mixing distribution resulting in a monotone ratio, allows the construction of a *generalized Chao* estimator which might provide an additional correction for *unstructured, residual heterogeneity*. Since

$$\frac{g_1/a_1}{g_0/a_0} \leq \frac{g_2/a_2}{g_1/a_1},$$

we can write the generalized Chao estimator as

$$\hat{N}_{GC} = n + \frac{(f_1/a_1)^2}{f_2/a_2} = n + \frac{\kappa+1}{\kappa}\frac{f_1^2}{2f_2}.$$

To illustrate these findings, we use the Scottish needle exchange data. In section 4, we have found evidence for a geometric density ($\kappa = 1$). However, the question arises whether there is any residual heterogeneity in this data set. The ratio plot

associated with a geometric is

$$r_x = \frac{g_{x+1}/a_{x+1}}{g_x/a_x} = g_{x+1}/g_x,$$

which can be simply estimated as $\hat{r}_x = f_{x+1}/f_x$. Figure 8 shows the empirical generalized ratio plot, from which there appears to be little evidence for residual heterogeneity. The generalized Chao estimator is $\hat{N}_{GC} = n + \frac{\kappa+1}{\kappa} f_1^2/(2f_2) = n + f_1^2/f_2 = 1007$ (since $\kappa = 1$) supporting the impression of little evidence for residual heterogeneity.

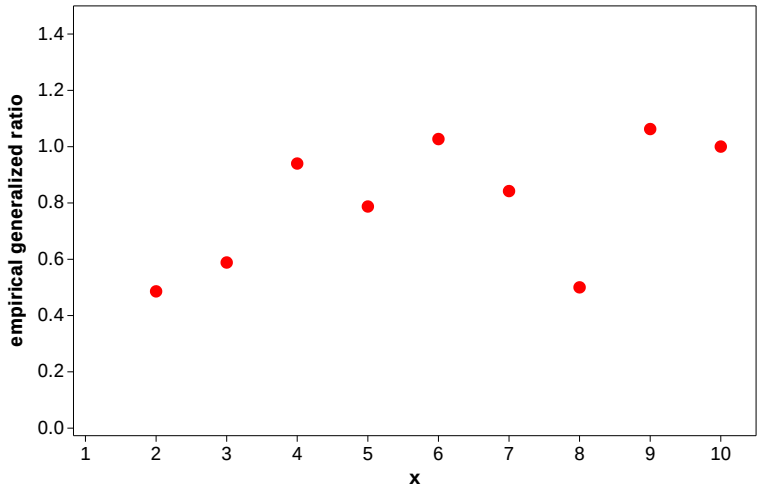


Figure 8: Generalised ratio plot of episode count per participant in a needle exchange programme in Scotland in 1997 (Hay and Smit 2003)

6 Concluding Remarks

The occurrence of Poisson homogeneity is rare in practice. This results in the need for identifying and allowing for heterogeneity (Böhning and Kuhnert 2006). However, a general approach allowing for arbitrary mixing distributions is problematic because of the identifiability problem (Link 2003; Holzmann *et al.* 2006; Link 2006) and the boundary problem (Wang & Lindsay 2005, 2008). The latter report an overestimation bias for the nonparametric mixture model for zero-truncated Poisson distributions. In practice this leads to the occurrence of spurious population size estimates as illustrated in Kuhnert *et al.* (2008). Consequently, to achieve identifiability and avoid spurious solutions it is reasonable to constrain the feasible class of mixing distributions to parametric mixing distributions with a small number of parameters or to rely on lower bounds (Chao 1987; Mao 2006; Mao 2007; Mao and Lindsay 2007).

To help avoid the aforementioned difficulties we have suggested utilizing a graphical device, the ratio plot, to identify structured heterogeneity, characterized by a parametric mixing distribution. An appropriately modified Chao-lower bound may be used to correct for potential residual heterogeneity. We also note that the methodology evolving from the ratio plot can also be used with kernels other than the Poisson. In particular, the binomial distribution where the size parameter

might correspond to the number of trapping occasions, if this is known, in the capture–recapture study.

Appendix: Monotonicity of the Ratio Plot for Mixtures of Power Series Densities

Let us consider the mixed power series family

$$p_x(\lambda) = \int_0^\infty a_x t^x \mu(t)\ \lambda(t) dt\ , \tag{10}$$

where a_x are known non-negative coefficients and $\mu(t)$ is the normalizing function in the power series satisfying $1/\mu(t) = \sum_{x=0}^{\infty} a_x t^x$. Note that the Power Series includes the Poisson ($a_x = 1/x!$, $\mu(t) = \exp(-t)$), the binomial, the geometric or, more generally, the negative binomial with known shape parameter κ.

We will prove the monotonicity result (11) in Theorem 4 for which we use the following version of the Cauchy-Schwarz inequality.

Lemma 1 *For any random variable Z with density $f(z)$ let $g_1(z)$ and $g_2(z)$ be arbitrary functions with existing first and second moments. Then*

$$[E(g_1(Z)g_2(Z)]^2 \leq E[g_1(Z)]^2 E[g_2(Z)]^2.$$

Theorem 4 *Let g_x be given according to (9). Then, the following monotonicity*

result holds:

$$\frac{g_1/a_1}{g_0/a_0} \leq \frac{g_2/a_2}{g_1/a_1} \leq \frac{g_3/a_3}{g_2/a_2} \leq ... \tag{11}$$

Proof.

We show

$$\left[\int_0^\infty t^x \mu(t)\ \lambda(t) dt\right]^2 \leq \int_0^\infty t^{x-1} \mu(t)\ \lambda(t) dt \int_0^\infty t^{x+1} \mu(t)\ \lambda(t) dt.$$

But this follows from Lemma 1 by choosing $T = Z$, $g_1(T) = \sqrt{T^{x-1}\mu(T)}$ and $g_2(T) = \sqrt{T^{x+1}\mu(T)}$. □

References

[1] Baksh, M. F., Böhning, D. and Lerdsuwansri, R. (2011). An Extension of an over-dispersion test for count data. *Computational Statistics and Data Analysis* **55**, 466–474.

[2] Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statistical Methodology* **5**, 410–423.

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

[4] Böhning, D. and Del Rio Vilas, V. J. (2008). Estimating the hidden number of scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological and Environmental Statistics* **13**, 1–22.

[5] Böhning, D. and Schön, D. (2005). Nonparametric maximum likelihood estimation of the population size based upon the counting distribution. *Journal of the Royal Statistial Society, Series C* **54**, 721–737.

[6] Bunge, J. and Fitzpatrick, M. (1993). Estimating the Number of Species: a Review. *Journal of the American Statistical Association* **88**, 364–373.

[7] Chao, A. (1987). Estimating the Population Size for Capture-Recapture data with Unequal Catchability. *Biometrics* **43**, 783–791.

[8] Chao, A. (1989). Estimating Population Size for Sparse Data in Capture-Recapture Experiments. *Biometrics* **45**, 427–438.

[9] Chao, A. and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* **58**, 531–539.

[10] Chao, A. and Huggins, R. M. (2005). Modern closed-population capture-recapture models. In: Amstrup, S.C., McDonald, T.L., Manly, B.F.J. (Eds.),

Handbook of capture-recapture analysis. Princeton University Press, Princeton, pp. 58–87.

[11] Chao, A., Tsay, P. K., Lin, S. H., Shau, W. Y. and Chao, D. Y. (2001). Tutorial in Biostatistics: The Applications of capture-recapture models to epidemiological data. *Statistics in Medicine* **20**, 3123–3157.

[12] Cullen, M. J., Walsh, J., Nicholson, L. V. B., and Harris, J. B. (1990). Ultrastructural localization of dystrophin in human muscle by using gold immunolabelling. *Proceedings of the Royal Society of London, Series B* **20**, 197–210.

[13] Dorazio, R. M. and Royle, J. A. (2005). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351–364.

[14] Gart, J. J. (1970). Some Simple Graphically Oriented Statistical Methods for Discrete Data. in G.P. Patil (ed.), Random Counts in Scientific Work, Vol.1 *Random Counts in Models and Structures*, University Park, Pa: The Pennsylvania State University Press.

[15] Good, I. J. (1953). On the Population Frequencies of species and the Estimation of Population Parameters. *Biometrika* **40**, 237–264.

[16] Hay, G. and Smit, F. (2003). Estimating the number of drug injectors from needle exchange data. *Addiction Research and Theory* **11**, 235–243.

[17] Hser, Y. I. (1993). Population estimates of intravenous drug users and HIV infection in Los Angeles county. *The International Journal of Addictions* **28**, 695–709.

[18] Hoaglin, D. C. (1980). A Poissonness Plot. *The American Statistician* **34**, 146–149.

[19] Holzmann, H., Munk, A., and Zucchini, W. (2006). On identifiability in capture-recapture models. *Biometrics* **62**, 934–939.

[20] Kuhnert, R., Del Rio Vilas, V. J., Gallagher, J. and Böhning, D. (2008). A bagging–based correction for the mixture model estimator of population size. *Biometrical Journal* **50**, 993–1005.

[21] Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123–1130.

[22] Link, W. A. (2006). Response to a paper by Holzmann, Munk and Zucchini. *Biometrics* **62**, 936–939.

[23] Matthews, J. N. S. and Appleton, D. R. (1993). An Application of the Truncated Poisson Distribution to Immunogold Assay. *Biometrics* **49**, 617–621.

[24] Mao, C. X. (2006). Inference on the number of species through geometric lower bounds. *Journal of the American Statistical Association* **101**, 1663–1670.

[25] Mao, C. X. (2007). Estimating Population sizes for capture–recapture sampling with binomial mixtures. *Computational Statistics and Data Analysis* **51**, 5211–5219.

[26] Mao, C. X. and Lindsay, B. G. (2007). Estimating the number of classes. *Annals of Statistics* **35**, 917–930.

[27] Moore, P. G. (1952). The estimation of the Poisson parameter from a truncated distribution. *Biometrika* **39**, 247–251.

[28] Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford, Oxford University Press.

[29] Pledger, S. A. (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics* **61**, 868–876.

[30] Van der Heijden, P. G. M., Cruyff, M., van Houwelingen, H. C. (2003).

Estimating the Size of a Criminal Population from Police Records Using the Truncated Poisson Regression Model. *Statistica Neerlandica* **57**, 1–16.

[31] Wang, J.-P. and Lindsay, B. G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100**, 942–959.

[32] Wang, J.-P. and Lindsay, B. G. (2008). An exponential partial prior for improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology* **5**, 30–45.

[33] Zelterman, D. (1988). Robust estimation in truncated discrete distributions with applications to capture-recapture experiments. *Journal of Statistical Planning and Inference* **18**, 225–237.