School of Mathematical
and Physical Sciences

Department of Mathematics and Statistics

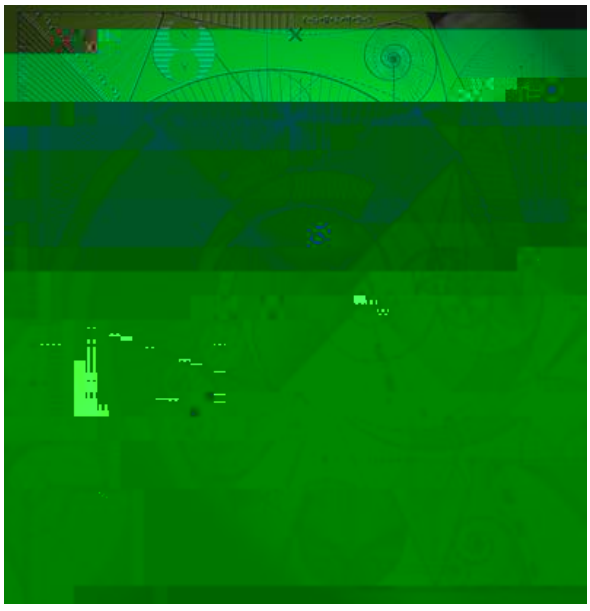
Preprint MPS_2011-01

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

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		

$p_0 = \exp(-\lambda)$ 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 B

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

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

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

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

$$\hat{r}_x$$

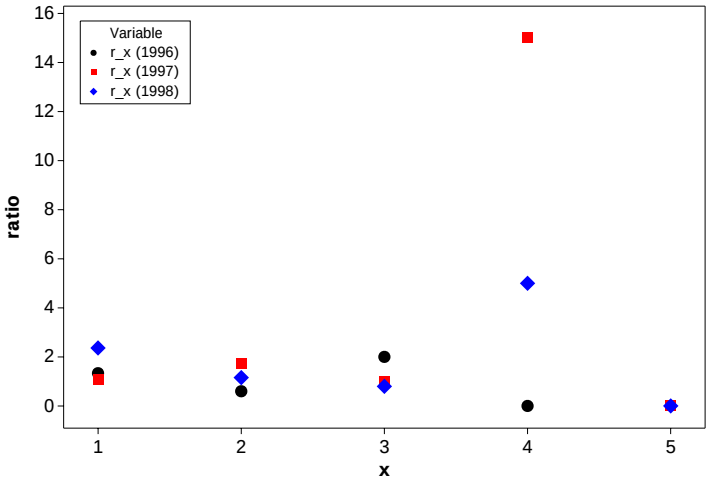


Figure 1: Ratio plot of observed Grizzly bears in the Yellowstone ecosystem for the period 1996-1998

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

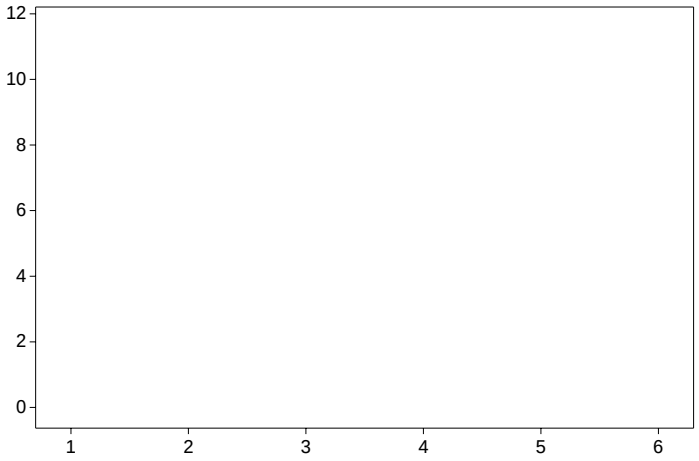


Figure 2: Ratio plot of observed scrapie infected herds in Great Britain based upon the Scrapie Notifications Database (SND) for the period 2002-2004

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?
-

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 $k = 1$

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

In addition, for the Turing estimator we have

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

$$\to [1$$

$\alpha Po(x;\mu)$ for $x = 0, 1, 2, \ldots$, $\alpha = 0.5$ 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

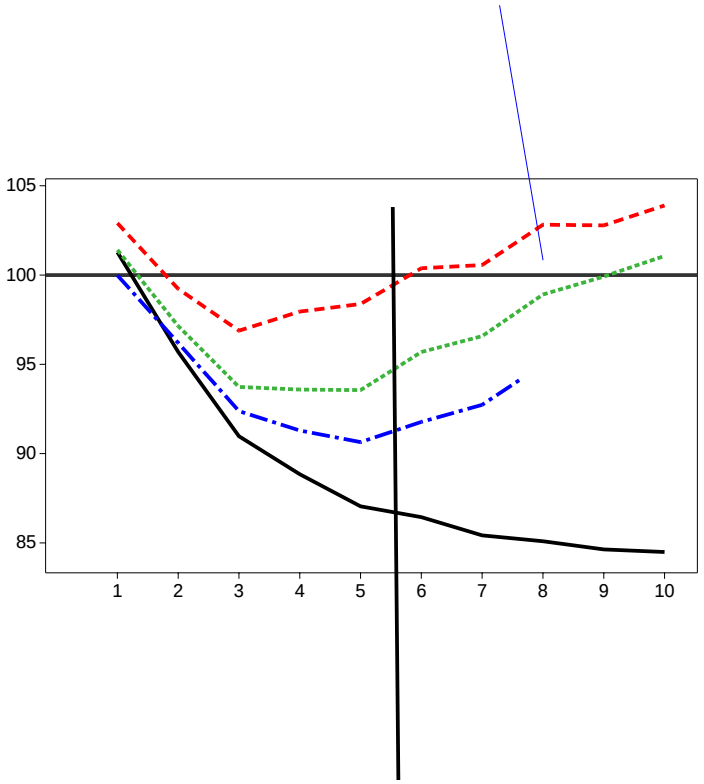


Figure 3: Mean population size estimator in contamination model $p_x = (1 - \alpha)Po(x;\lambda) + \alpha Po(x;\mu)$ for $N = 100$; N(k) denotes the robust Turing estimator $\hat{N}_k$ and 'Chao' corresponds to $\hat{N}_1$

$bias^2(\hat{N}_1) \leq bias^2(\hat{N}_2) \leq bias^2(\hat{N}_3)$ bi8991482889863d

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

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

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

where $\hat{\lambda}_k$ is given by equation (4), $Po_+(x;$

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

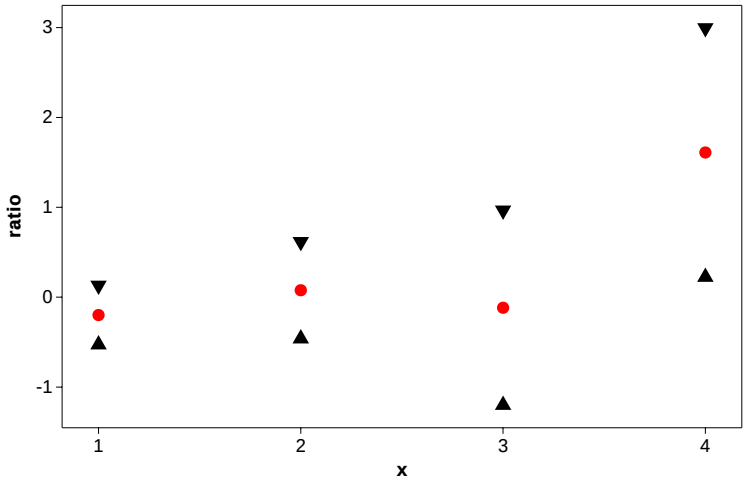


Figure 5: Ratio plot (on log-scale) for the dystrophin data (bullets) with approximate 95% confidence limits (upper and lower triangle)

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

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* $\beta = 0$

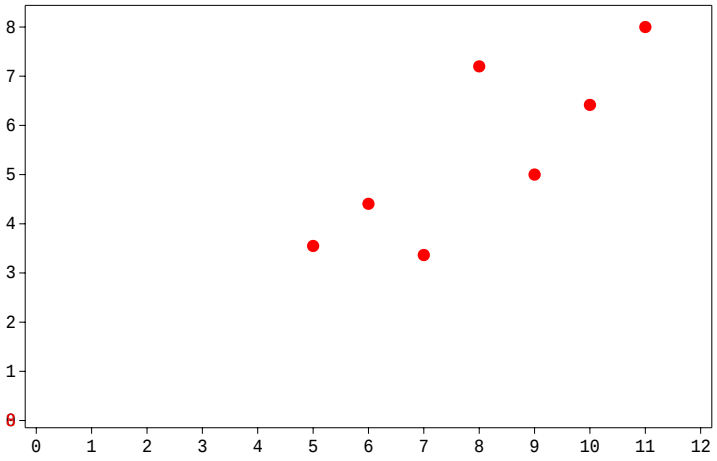


Figure 6: Ratio plot of episode count per drug user in Los Angeles in 1989

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

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 $\pi_1/E(X) = \pi^{\kappa+1}$ 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* $p_x =$ $\Gamma(\kappa+x_1$

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-*

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.

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*

$$[E(\quad$$

result holds:

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

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 $g_2(T) = \sqrt{T^{x+1}\mu(T)}$.

[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]

Handbook of capture-recapture analysis

[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.

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.