Submitted to Statistical Science
arXiv: math.PR/0000000

Ratio plot and ratio regression with applications to social and medical sciences

Dankmar Böhning

Southampton Statistical Sciences Research Institute & Mathematical Sciences, University of Southampton

Abstract. We consider count data modeling, in particular, the zero-truncated case as it arises naturally in capture-recapture modeling as the marginal distribution of the count of identifications of the members of a target population. Whereas in wildlife ecology these distributions are often of a well-defined type, this is less the case for social and medical science applications since study types are often entirely observational. Hence, in these applications, violations of the assumptions underlying closed capture-recapture are more likely to occur than in carefully designed capture-recapture experiments. In consequence, the marginal count distribution might be rather complex. The purpose of this note is to sketch some of the major ideas in the recent developments in ratio plotting and ratio regression designed to explore the pattern of the distribution underlying the capture process. Ratio plotting and ratio regression are based upon considering the ratios of neighboring probabilities which can be estimated by ratios of observed frequencies. Frequently, these ratios show patterns which can be easily modeled by a regression model. The fitted regression model is then used to predict the frequency of hidden zero counts. Particular attention is given to regression models corresponding to the negative-binomial, multiplicative binomial and the Conway-Maxwell-Poisson distribution.

Key words and phrases: closed capture-recapture, Conway-Maxwell-Poisson, mixtures, multiplicative-binomial, negative-binomial, zero-truncated count distributions.

1. INTRODUCTION

We are interested in zero-truncated count distributional modelling which arises naturally in capture-recapture experiments or studies. The size N of a target population needs to be determined. For this purpose a trapping experiment or study is done where members of the target population are identified at T occasions where T might be known or not. Furthermore, the sampling occasions

Dankmar Böhning is Professor in Medical Statistics, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, UK (e-mail: d.a.bohning@soton.ac.uk)

*The author is grateful to all editors and reviewers for the many helpful comments.

might be specified prior to the study or they might occur randomly during the observational period. For each member i the count of identifications X_i is returned where X_i takes values in $\{0, 1, 2, \cdots\}$ for $i = 1, \cdots, N$. However, zero-identifications are not observed; they remain hidden in the study. Hence, a zero–truncated sample $X_1, \cdots, X_n$ is observed, where we have assumed w.l.o.g. that $X_{n+1} = \cdots = X_N = 0$. So, n is the number of recorded individuals. The associated untruncated and zero-truncated densities will be denoted as $p_x(\theta)$ and $p_x^+(\theta) = p_x(\theta)/[1 - p_0(\theta)]$, respectively. The setting above has been developed primarily for wildlife populations (Bunge and Fitzpatrick 1993, Borchers *et al.* 2004, Chao 2001, Sanathanan 1977, Wilson and Collins 1992). We are interested here to apply the framework to social and medical scenarios as we will illustrate in the following three examples.

1.1 Homeless population of the city of Utrecht

As illustration of the problem, we consider the question of estimating the homeless population of Utrecht (NL). The city of Utrecht runs a shelter where homeless people can stay overnight. Data are available for a period of 14 nights in 2013 and are shown in Table 1. It can be assumed that the shelter covers only the city of Utrecht. The table contains information on how often homeless people stayed in the shelter within this 14-nights period. For example, $f_1 = 36$ people stayed exactly one night, whereas $f_2 = 11$ people stayed exactly two nights, and so forth. In total, 222 different homeless people stayed in the shelter, spending a total of $S = \sum_{x=1}^{14} x f_x = 2,009$ nights there. For more details see van der Heijden *et al.* (2014a). In this case, the number of occasions is known with $T = 14$ and also the occasions are specified in the observational period. Whereas some homeless people use the shelter frequently, others use it only occasionally or very rarely. Hence the register for homeless people based on the shelter is incomplete. The city of Utrecht is interested in the total size of its homeless population. Hence, we are interested to find an estimate of N, or, equivalently, of f_0, the size of the hidden homeless population.

TABLE 1
Frequency distribution of the number of nights x stayed in the shelter per homeless person for the city of Utrecht for a period of 14 nights in 2013

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	n
f_x	36	11	6	11	5	7	6	11	3	8	7	12	22	77	222

1.2 Domestic violence in NL

In a study of domestic violence, van der Heijden *et al.* (2014b) reports perpetrator offense data in the Netherlands for the year 2009. The data represent the Netherlands excluding the police region for The Hague. Here the perpetrator study is reported in Table 2. In this case T is unknown and there are no prespecified sampling occasions as domestic violence incidents occurred at unplanned time points in the observational period 2009.

TABLE 2
Frequencies of the number of times perpetrators have been identified in a domestic violence incident in the Netherlands in the year 2009

x	1	2	3	4	5	6	7	8	9	n
f_x	15,169	1,957	393	99	28	8	6	1	1	17,662

There were 15,169 perpetrators identified as being involved in a domestic violence incident exactly once, 1,957 exactly twice, and so forth. In total, there were 17,662 different perpetrators identified by the police in the Netherlands for 2009. As not every case of domestic violence is reported to the police, an unknown number of perpetrators remain hidden. Hence, here the target population of interest consists of the perpetrators in the Netherlands (excluding The Hague) in the year 2009, whether they have been identified by the police or not.

1.3 Size of forced labour worldwide

The International Labour Office (ILO) undertook a study to estimate the size of forced labour worldwide (ILO 2012). Here *forced labour* is characterized by provision of some form of work or service which is done under threat of penalty and undertaken involuntarily. Frequently the term *slave labour* is used instead (Bales 2012). Due to its hidden nature forced labour is hard to measure. For this reason, the ILO launched a capture-recapture study to estimate the size of forced labour worldwide. Teams were established and searched for reports on forced labour. Sources of information included media, government reports, academic and trade union reports and many more. In total about 2500 different sources have been used. The period that was covered was the years 2002 – 2011. Reports were collected from anywhere in the world and therefore considerable heterogeneity should be expected. Table 3 shows the zero-truncated frequency distribution of the count x, the number of times a case of forced labour has been identified in any of the sources. There were 4069 cases of forced labour that were exactly identified by 1 report, 1,186 cases that were identified by 2 different reports etc. Each case will have a certain number of persons involved. From this an estimate of the size of forced labour (the number of people involved) can be derived. Here we are interested in estimating the number of cases f_0 that were identified by $x = 0$ reports.

TABLE 3
Frequency distribution of forced labour report counts

x	1	2	3	4	5	6	7	8	9	10	11	n
f_x	4,069	1,186	167	46	10	7	3	1	0	1	1	5,491

1.4 Screening for bowel cancer

Bowel cancer can develop without any early warning signs. The Faecal Occult Blood Test (FOBT) can detect small amounts of blood in the bowel motion. This might be indicative of a problem such as cancer but also something else such as polyps or nothing at all. Lloyd and Frommer (2008, 2004a, 2004b) present results of a screening study for bowel cancer in Sydney (Australia). From 1984 onwards about 50000 subjects were screened for bowel cancer using the FOBT. Self-administered testing took place on $T = 6$ successive days and at each of the

6 occasions absence or presence of blood in faeces was recorded. If at least one of the T tests is positive a gold standard evaluation took place and results could be healthy, polyps, or cancer. A person that tested negatively on all T tests is not further assessed. Out of exactly 49,927 persons, 46,553 tested negatively on all six tests (and these were not further investigated). Out of the other 3374 subjects who tested positively at least once, 3106 were examined and their true disease status determined. The other 268 subjects who tested positively were lost to the study. In Table 4 we see the frequency distribution of the 228 persons with cancer where x is the count of positive tests in the 6-days period. As 46,553 remained without further assessment the question arises of how much hidden cancer is present among this unassessed population.

TABLE 4
Frequency distribution of number of positive tests of those with cancer and testing positive at least once (Lloyd and Frommer 2008)

x	1	2	3	4	5	6	n
f_x	46	27	26	33	39	57	228

1.5 Assumptions involved in the ratio-regression approach

The ratio-regression approach to be presented is not assumption-free. We assume that the target population is closed, e. g. that there is no migration, no deaths, and no births. Specifically, the no-migration assumption can be questionable in some of the examples above such as the homeless study in the city of Utrecht. Here, the size of the time window is a steering element in satisfying the closed-population assumption. The larger the observational period the more likely is the occurrence of migration. The smaller the period the less homeless people are observed. In this case, it was found that 14 days established a reasonable compromise as increasing the period by one week did not add substantially more homeless people to the observed part of the homeless population. An alternative way to proceed would be to use open population modeling such as the Cormack-Jolly-Seber model (McCrea and Morgan 2015; see also Cormack 1964, Jolly 1965, Seber 1965 and this special issue) in which the time-specific dependency of the data is incorporated.

In some cases it might be unclear how the target population is defined. Whereas in the case of the bowel cancer study 1.4 the target population is the *disease-free* screened population, this is less clear in the domestic violence study 1.2 of the Netherlands. Here, we define the target population to be all perpetrators that actually performed acts of domestic violence whether this has been identified by the authorities or not.

2. RATIO PLOT

We aim to estimate the population size N. As $N = Np_0 + N(1-p_0)$ where p_0 is the probability of a zero count or missing an observation, we can get an estimate of N by using the moment estimate n for $N(1-p_0)$ and solving $\hat{N} = \hat{N}p_0 + n$ for $\hat{N} = n/(1-p_0)$, a Horvitz-Thompson estimate of N. As p_0 is unknown in most applications (and certainly in those of section 1) we need to come up with some estimate for p_0. A natural way to proceed is to use a parametric model $p_x = p_x(\theta)$ for $x = 0, 1, \cdots$, derive some estimate $\hat{\theta}$ for θ on the basis of

$p_x^+(\theta)$ for $x = 1, 2, \cdots$, and use $\hat{\theta}$ in $p_0(\hat{\theta})$ to estimate N. See also Sanathanan (1977). McCrea and Morgan (2015) call $\hat{N} = n/[1 - p_0(\hat{\theta})]$ a Horvitz-Thompson-like estimate to distinguish it from the conventional Horvitz-Thompson estimate $\hat{N} = n/(1 - p_0)$.

A natural starting point for searching for an appropriate count distribution is the power series density

$$p_x(\theta) = a_x \theta^x / \eta(\theta) \tag{1}$$

where a_x is a known, nonnegative coefficient, θ a positive parameter and $x = 0, 1, \cdots$ ranges over the set of nonnegative integers. Also, $\eta(\theta) = \sum_{x=0}^{\infty} a_x \theta^x$ is the normalizing constant. The power series distribution contains the Poisson ($a_x = 1/x!$), the binomial ($a_x = \binom{T}{x}$ for $x = 0, \cdots, T$ with positive integer T and $a_x = 0$ for $x > T$) or the geometric ($a_x = 1$).

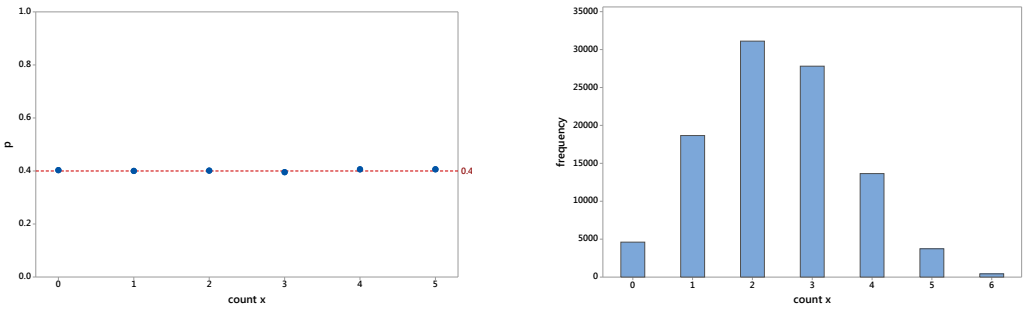

FIG 1. *Ratio plot for a sample of 100,000 counts from a binomial with $T = 6$ and event parameter $p = 0.4$ (left panel) and frequency distribution (right panel). The vertical axis in the left panel shows $\hat{p} = \hat{r}_x/(1 + \hat{r}_x)$.*

It is a fundamental property of the power series distribution that

$$r_x = \frac{p_{x+1}(\theta)/a_{x+1}}{p_x(\theta)/a_x} = \frac{p_{x+1}^+(\theta)/a_{x+1}}{p_x^+(\theta)/a_x} = \theta, \tag{2}$$

the ratio of neighboring probabilities multiplied by the inverse of their respective coefficients is a constant, independent of x, in fact it is the parameter θ itself. This property occurs for the untruncated as well as for the zero-truncated distribution as the normalizing constant $(1 - p_0(\theta))$ cancels out. The quantity r_x can be used to develop a *diagnostic device* for the presence of a particular distribution. As p_x is an unknown quantity we replace it by its nonparametric estimate f_x/N so that we obtain an empirical ratio

$$\hat{r}_x = \frac{a_x}{a_{x+1}} \frac{f_{x+1}}{f_x}, \tag{3}$$

as the unknown quantity N cancels out. Plotting $\hat{r}_x$ against x provides the *empirical ratio plot* or simply the *ratio plot.* If the ratio plot shows a horizontal line pattern we can take this as supportive evidence for the presence of the distribution of interest. The determining quantity in (3) is the ratio a_x/a_{x+1} of

the coefficients of the power series family member. For the Poisson this ratio is $a_x/a_{x+1} = x + 1$, for the binomial it is $a_x/a_{x+1} = (x + 1)/(T - x)$, and for the geometric it is simply $a_x/a_{x+1} = 1$. As the ratio plot construction depends on the coefficient a_x we emphasize this by mentioning the family member. For example, if we use the concept for the binomial we speak of the *binomial* ratio plot, if we use it for the geometric we speak of the *geometric* ratio plot. If there is no doubt of which family member is used we simply speak about the ratio plot. The ratio plot has been developed in its basic form in Böhning *et al.* (2013). We illustrate the concept for the binomial in Figure 1. 100,000 counts have been sampled from a binomial with size parameter $T = 6$ and event parameter $p = 0.4$ corresponding to the parameterization in the power series of $\theta = p/(1 - p) = 2/3$. In the left panel of Figure 1 we see the ratio plot for the binomial on the event parameter scale. There is clear evidence of a horizontal line pattern supporting the binomial distribution. The benefit of the diagnostic device becomes clear when comparing it to the bar chart provided in the right panel of Figure 1 where the binomial distribution is more difficult to recognize.

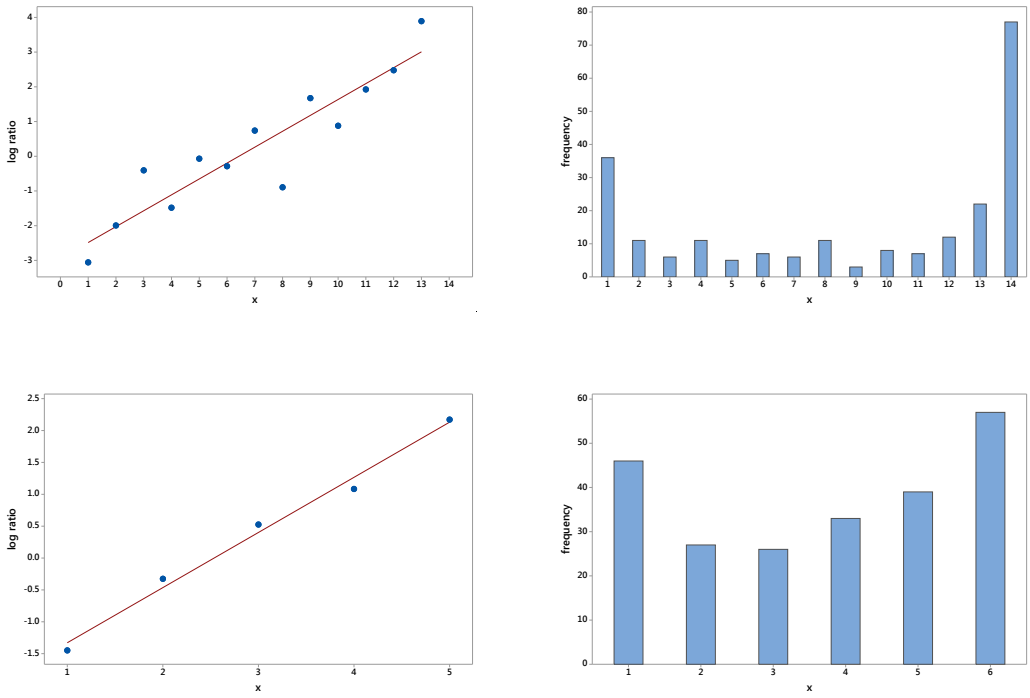

FIG 2. *Binomial ratio plot (on the log-scale) for homeless data of section 1.1 (upper left panel) and bowel cancer data of section 1.4 (lower left panel) with associated frequency distributions (upper and lower respective right panel)*

We now apply the binomial ratio plot to the homeless study data of section 1.1 and the bowel cancer data of section 1.4. Figure 2 shows the ratio plot for the homeless data of Utrecht (upper left panel) and for the bowel cancer data (lower left panel). Note that the ratio $\hat{r}_x = (x + 1)/(T - x) \times f_{x+1}/f_x$ is plotted on the log-scale. The associated frequency distributions are provided in the right panels of Figure 2. There is clear evidence that a horizontal line pattern does not

hold. There could be various reasons why a horizontal line pattern is violated in the ratio plots present in Figure 2. It could be that the repeated visits to the homeless shelter (upper left panel) are not independent or that homeless people have different tendencies to visit the shelter. Similar issues might occur in the bowel cancer data (lower left panel) where repeated testing might not be independent or different patients might have different risk for a positive test. An alternative approach to deal with dependencies between occasions is the approach using log-linear models as suggested by Fienberg (1972) and Cormack (1989). See also Chao (2001). This approach requires availability of the data in form of complete capture histories x_{ij} where $x_{ij} = 1$ if unit i is identified at occasion j and 0 otherwise. In certain applications such as the homeless or bowel cancer data occasion-specific data might be available (although we did not have access to these for the present work), in other applications such as the worldwide forced labour study only $x_i = \sum_{j=1}^{T} x_{ij}$ is available and log-linear modelling is not possible in these cases.

Let us now turn to the domestic violence data of section 1.2. The Poisson ratio plot in Figure 3 (left panel) provides evidence for a violation of the Poisson assumption in this case. There is a clear positive trend visible in the ratio plot. However, there is no reason why we can expect domestic violence counts to follow a Poisson distribution. We might as well consider the geometric distribution and its associated ratio plot implying plotting $x \to f_{x+1}/f_x$ as provided in the right panel of Figure 3. Apparently, there is also a positive trend visible although this appears more diminished in the geometric ratio plot than in the Poisson ratio plot. We denote by T_0 the largest count considered, in this case $T_0 = 5$. Note that $T_0 \leq T$ if the number of sampling occasions is known. An inspection of the chi-square goodness-of-fit statistic $\chi^2 = \sum_{x=1}^{T_0-1}(\log \hat{r}_x - \log \bar{r}_x)^2/\widehat{var}(\log \hat{r}_x)$ confirms this impression. Here, $\widehat{var}(\log \hat{r}_x) = 1/f_{x+1} + 1/f_x$ (Rocchetti *et al.* (2011), Böhning *et al.* (2013)), and for estimating the parameter θ we use the consistent and asymptotically unbiased estimates $\bar{r}_x = \sum_{x=1}^{T_0-1}(x+1)f_{x+1}/f_x$ for the Poisson ratio plot and $\bar{r}_x = \sum_{x=1}^{T_0-1} f_{x+1}/f_x$ for the geometric Poisson ratio plot. We find $\chi^2 = 382.54$ for the Poisson and $\chi^2 = 94.86$ for the geometric ratio plot, both with $T_0 - 1 = 4$ df. Thus, it is clear that even in the case of the geometric the fit is not yet acceptable, and we will turn to ratio regression in the next section to extend the modelling framework considerably.

3. RATIO REGRESSION

The basic idea is to extend the ratio plot to a full regression approach. Consider $\hat{r}_x = \frac{a_x}{a_{x+1}}\frac{f_{x+1}}{f_x}$ and the regression model in count x

$$\hat{r}_x = \sum_{j=1}^{p} \beta_j g_j(x) \tag{4}$$

where $g_j(x)$ is a known regression function of count x. In most applications we have in mind, $p = 2$ or $p = 3$, and $g_j(x)$ is of simple structure such as $g_1(x) = 1$ and $g_2(x) = x$ or $g_2(x) = \log(x+1)$. After estimating the coefficients $\beta_1, \cdots, \beta_p$ we can estimate f_0 as

$$\hat{f}_0 = \frac{a_0}{a_1}\frac{f_1}{\sum_{j=1}^{p}\hat{\beta}_j g_j(0)}. \tag{5}$$

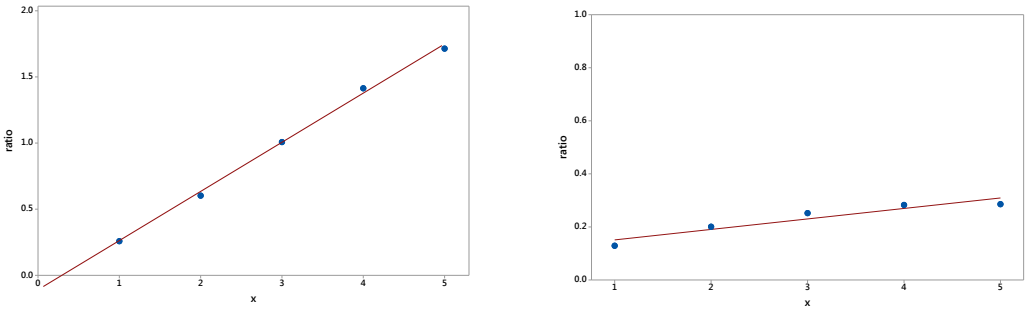

FIG 3. *Poisson ratio plot (left panel) and geometric ratio plot (right panel) for the domestic violence data of the Netherlands ignoring low frequency data* $x \geq 6$

3.1 Ratio regression and mixtures

We are interested in connecting the presence of unobserved heterogeneity (which could be described by a latent variable) with the concept of the ratio plot and ratio regression. If the target population consists of subpopulations and subpopulation membership is not observed we speak of the occurrence of unobserved heterogeneity. For example, in the case study on forced labour 1.3, reports were collected from anywhere in the world and therefore considerable heterogeneity should be expected. Assuming that in each subpopulation a power series distribution is valid then unobserved heterogeneity leads to a mixture of power series distributions $m_x = \int_\theta p_x(\theta) f(\theta) d\theta$, where $f(\theta)$ represents the mixing distribution, the distribution of the subpopulation parameter θ in the population. Hence mixtures of parametric count distributions have attracted some attention in capture-recapture modelling (Dorazio and Royle 2005, Pledger 2005, Norris and Pollock 1996, Wang and Lindsay 2005, 2008, Mao and You 2009, Böhning and Kuhnert 2006). We can likewise consider the ratio plot for mixtures

$$r_x = \frac{a_x}{a_{x+1}} \frac{m_{x+1}}{m_x} \tag{6}$$

where we use again the known coefficients a_x associated with the mixture kernel, for example, in the case of a Poisson kernel $a_x = 1/x!$ or the case of a geometric kernel $a_x = 1$. The estimate of r_x will not change; however, the interpretation of the observed pattern in the ratio plot will. This is mainly due to the following result (Böhning and Del Rio Vilas 2008):

THEOREM 1. *Let* $m_x = \int_\theta p_x(\theta) f(\theta) d\theta$ *where* $p_x(\theta)$ *is a member of the power series family and* $f(\theta)$ *an arbitrary density. Then, for* $r_x = \frac{a_x}{a_{x+1}} \frac{m_{x+1}}{m_x}$ *we have the following monotonicity:*

$$r_x \leq r_{x+1}$$

for all $x = 0, 1, \cdots$.

The result in Theorem 1 can be interpreted as saying that the presence of unobserved heterogeneity will force a monotone increasing pattern in the ratio plot. In some special cases for the mixing distribution stronger results are possible.

Suppose that $X_{|\Theta=\theta}$ is Poisson with density $p_x(\theta)$ and suppose further that the density $f(\theta)$ of Θ is a gamma with parameters k and β. Then, using standard knowledge,

$$\begin{aligned} m_x = & \int_\theta p_x(\theta) f(\theta) d\theta = \tfrac{1}{\Gamma(k)\beta^k} \int_0^\infty \tfrac{\exp(-\theta)\theta^x}{x!} \times \theta^{k-1} \exp(-\theta/\beta) d\theta \\ = & \tfrac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)} \beta^{-k} \left(\tfrac{\beta}{\beta+1}\right)^{k+x}, \end{aligned} \tag{7}$$

which corresponds to a negative-binomial with parameter $p = 1/(\beta+1)$ so that

$$m_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)} (1-p)^x p^k. \tag{8}$$

It is easy to work out that $r_x = \frac{a_x}{a_{x+1}} \frac{m_{x+1}}{m_x} = (x+1)\frac{m_{x+1}}{m_x} = (1-p)(x+k)$ in this case, so that the monotone pattern in the ratio plot becomes a straight line with intercept $(1-p)k$ and positive slope $(1-p)$.

3.2 Ratio regression and Chao estimation

Another question is how the result in Theorem 1 connects to established estimators such as Chao's estimator (Chao 1987, 1989). Chao's estimator of f_0 has been developed as a lower bound estimator under $m_x = \int_\theta p_x(\theta) f(\theta) d\theta$ where $p_x(\theta)$ is the Poisson density and $f(\theta)$ an arbitrary mixing distribution. The original estimator takes the form $\hat{f}_0 = f_1^2/(2f_2)$ and is one of the most frequently used estimators in capture-recapture modeling.

We let $p_x(\theta)$ be any member of the power series now. Then Theorem 1 implies

$$\frac{a_0}{a_1}\frac{m_1}{m_0} \le \frac{a_x}{a_{x+1}}\frac{m_{x+1}}{m_x} \text{ for } x = 0, 1, \cdots. \tag{9}$$

For $x = 1$ it follows that $m_0 \ge (a_0 a_2/a_1^2)(m_1^2/m_2)$, and replacing m_x by f_x/N leads to Chao's lower bound estimator $(a_0 a_2/a_1^2)(f_1^2/f_2)$ for f_0 in the case of the power series family, and in particular to $f_1^2/(2f_2)$ in the Poisson case. The lower bound estimator becomes asymptotically unbiased if there is no heterogeneity (the mixing distribution becomes a one mass point distribution). Note that the lower bound estimator is valid for any mixing distribution on θ including a discrete mixing distribution with point mass at zero (leading to a zero-inflated distribution) as this is a special case of a discrete mixing distribution. However, its bias will depend on the choice of the mixture kernel. For example, in the case of the domestic violence data of section 1.2 we can expect, by inspecting the ratio plot in Figure 3, that the geometric lower bound will have a smaller bias than the Poisson lower bound as the bias-determining difference

$$\frac{a_0}{a_1}\frac{f_1}{f_0} - \frac{a_1}{a_2}\frac{f_2}{f_1} \tag{10}$$

can be expected to be smaller for the geometric than for the Poisson.

The result of Theorem 1 allows many lower bound estimators since $m_0 \ge (a_0 a_{x+1}/a_x)\ (m_1 m_x/m_{x+1})$ for $x = 1, 2, \cdots$. For example, $(a_0 a_3/a_2)(f_1 f_2/f_3)$ provides a lower bound estimator for f_0 if we choose $x = 2$. However, none will be as sharp as Chao's lower bound, the one we obtain for $x = 1$. Nevertheless, considering the ratios r_x for $x > 1$ can be helpful and ratio regression can be viewed as a way of projecting to a best lower bound.

3.3 Ratio regression and empirical Bayes

The ratio $r_x = \frac{a_x}{a_{x+1}}\frac{m_{x+1}}{m_x}$ has an interesting connection to Bayesian inference. In fact,

$$\begin{aligned} r_x = & \quad \frac{a_x}{a_{x+1}}\frac{m_{x+1}}{m_x} = \frac{a_x}{a_{x+1}}\frac{\int_\theta a_{x+1}\theta^{x+1}/\eta(\theta)f(\theta)d\theta}{\int_\theta a_x\theta^x/\eta(\theta)f(\theta)d\theta} \\ = & \quad \int_\theta \theta \times \frac{a_x\theta^x/\eta(\theta)f(\theta)}{\int_\theta a_x\theta^x/\eta(\theta)f(\theta)d\theta}d\theta = \int_\theta \theta f(\theta|x)d\theta \end{aligned} \tag{11}$$

is the *posterior mean* w.r.t. the prior distribution $f(\theta)$ on θ. Here $f(\theta|x) = \frac{a_x\theta^x/\eta(\theta)f(\theta)}{\int_\theta a_x\theta^x/\eta(\theta)f(\theta)d\theta}$ is the posterior distribution. Hence $\hat{r}_x = \frac{a_x}{a_{x+1}}\frac{f_{x+1}}{f_x}$ provides an estimate of the posterior mean *without assuming any knowledge of the prior distribution nor is there any requirement for estimating the prior distribution*, an idea which goes back to Robbins (1955) and is considered the origin of empirical Bayes. For more details see Carlin and Louis (2011). In conclusion, when modelling r_x we are modelling the posterior mean.

3.4 Ratio regression and count distribution modeling

We return to the ratio regression approach (4). To ensure positive fitted values we need to incorporate a link-function leading to the ratio-regression model

$$\log \hat{r}_x = \sum_{j=1}^{p} \beta_j g_j(x) \tag{12}$$

where $g_j(x)$ is a known regression function of count x. Indeed, fitting a simple straight line to $\hat{r}_x$ in the domestic violence data of section 1.2 would lead to a negative intercept estimate (see left panel of Figure 3) and, hence, to a non-feasible estimate of f_0. This is not a specific problem of the least-squares estimation technique used here, but a more general deficiency of the negative-binomial as also the maximum likelihood estimate of the shape parameter lies on the boundary of the parameter space. Invoking an appropriate link-function such as the log-link avoids this non-feasibility, but we are also losing the interpretation of the straight line ratio regression as the negative-binomial model. Instead of working with the negative-binomial we can try the Conway-Maxwell-Poisson distribution given by

$$m_x = \frac{1}{C}\frac{\theta^x}{(x!)^\nu}, \tag{13}$$

for $x = 0, 1, \cdots$ and positive θ and ν. The normalizing constant $C = \sum_{x=0}^{\infty}\theta^x/(x!)^\nu$ is not available in closed form. For more details see Sellers and Shmueli (2010). It is easy to see that $r_x = (x+1)m_{x+1}/m_x = \theta(x+1)^{1-\nu}$ which suggests the ratio regression approach with log-link

$$\log r_x = \beta_1 + \beta_2 \log(x+1) \tag{14}$$

where $\beta_1 = \log\theta$ and $\beta_2 = (1-\nu)$ and the restriction $\beta_2 \leq 1$. Hence working with the Conway-Maxwell-Poisson distribution is equivalent to working with a straight line model on the log-scale for the ratio regression. We see the $\log[(x+1)f_{x+1}/f_x]$ and the model fit for $\beta_1 + \beta_2\log(x+1)$ in Figure 4 for the domestic violence data of section 1.2. While the fit of the Conway-Maxwell-Poisson distribution is

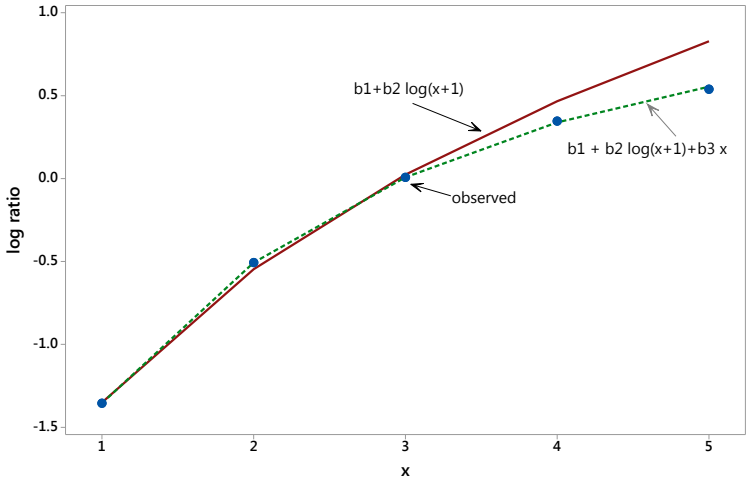

FIG 4. *Poisson ratio regression using* $\beta_1+\beta_2\log(x+1)$ *(solid curve) and* $\beta_1+\beta_2\log(x+1)+\beta_3 x$ *(dashed curve) for the domestic violence data of section 1.2*

good for $x = 1, 2, 3$, it is deteriorating for values $x = 4, 5$. The model $\log r_x = \beta_1 + \beta_2 \log(x + 1) + \beta_3 x$ provides an excellent fit for all $x-$values as Figure 4 shows. An estimate $\hat{f}_0$ is simply found from the estimated regression coefficients as $\hat{f}_0 = f_1 \exp(-\hat{\beta}_1)$. Here as well as for the general case of the model $E(\mathbf{Y}) = \mathbf{X}\beta$, we use the weighted least-squares estimate

$$\hat{\beta} = (\mathbf{X^T W X})^{-1}\mathbf{X^T W Y}, \tag{15}$$

where $Y = (\log r_1, \cdots, \log r_{T_0-1})^T$, $\mathbf{X}$ is the design matrix containing the regression functions of the model, $\mathbf{W}$ is a diagonal matrix containing the estimated inverse variances of $Y_1, \cdots, Y_{T_0-1}$, more precisely, $w_i = (1/f_i + 1/f_{i+1})^{-1}$. Here T_0 is the largest count considered. Note that the estimated covariance matrix of (15) is readily available as

$$\widehat{\mathbf{cov}}(\hat{\beta}) = (\mathbf{X^T W X})^{-1}. \tag{16}$$

The ratio regression approach opens the door to a huge arena of techniques. However, whatever we choose as regression model we would like to make sure to include an intercept term as this guarantees that the power series family is included as a special case. For example, in the Poisson ratio regression case $\hat{r}_x = (x + 1)p_{x+1}/p_x = \sum_{j=1}^{p} \beta_j g_j(x)$ we always choose $g_1(x) = 1$ as this will include the Poisson model as a special case ($\beta_2 = \cdots \beta_p = 0$).

In the search for better fitting ratio regression models we are also moving away from known corresponding probability models. In fact, the question arises does

a model such as $\log r_x = \beta_1 + \beta_2 \log(x+1) + \beta_3 x$ correspond to a probability distribution at all? The answer to this question is given by Theorem 2 and is basically a *yes* under the mild assumption that $r_x > 0$ for all $x = 0, \cdots, T_0 - 1$ and underlines the importance of an appropriate link function. [We think of r_x as arising from some regression model $r_x = \exp[\sum_{j=1}^{p} \beta_j g_j(x)]$.]

THEOREM 2. *Let $r_x > 0$ be given for $x = 0, 1, ..., T_0 - 1$. Then there exists a unique count distribution p_x for $x = 0, ..., T_0$ with the properties*

1.
$$p_{x+1} = p_x r_x a_{x+1}/a_x$$
for $x = 0, 1, ..., T_0 - 1$

2.
$$p_0 = \frac{1}{1 + r_0 a_1/a_0 + (r_0 a_1/a_0)(r_1 a_2/a_1) + ... + \prod_{x=0}^{T_0-1} r_x a_{x+1}/a_x}$$

A proof of Theorem 2 is given in the appendix. The value of Theorem 2 lies in the fact that it guarantees the existence of a proper probability distribution for any valid ratio regression model. It will also allow the construction of an estimator for p_0 by means of

$$\hat{p}_0 = \frac{1}{1 + \hat{\hat{r}}_0 a_1/a_0 + (\hat{\hat{r}}_0 a_1/a_0)(\hat{\hat{r}}_1 a_2/a_1) + ... + \prod_{x=0}^{T_0-1} \hat{\hat{r}}_x a_{x+1}/a_x}, \tag{17}$$

where $\hat{\hat{r}}_x = \exp[\sum_{j=1}^{p} \hat{\beta}_j g_j(x)]$ is the fitted regression model for $x = 0, 1, \cdots, T_0 - 1$, ultimately leading to the Horvitz-Thompson-like estimator $n/(1 - \hat{p}_0)$. Note that we are using here $\hat{\hat{r}}_x$ for the fitted value to distinguish it from the empirical observed ratios $\hat{r}_x$, and the theoretical model ratios r_x. However, in the following applications we will only use the simpler estimator for f_0, namely $\hat{f}_0 = \exp[-\sum_{j=1}^{p} \hat{\beta}_j g_j(0)] \times f_1 = \hat{\hat{r}}_0 \times f_1$.

3.5 Ratio regression and variance estimation

Another benefit of the ratio regression approach is that variance estimators for $\hat{f}_0$ can easily be developed as variance estimators for the estimated regression coefficients are easily available. We will demonstrate this for the binomial straight line ratio regression model. In this case, $\hat{f}_0 = f_1 \exp(-\hat{\beta}_1)/T$. Using conditioning moment techniques (Böhning 2008)

$$Var(\hat{f}_0) = E[Var(\hat{f}_0|f_1)] + Var[E(\hat{f}_0|f_1)] \tag{18}$$

$$\approx \frac{1}{T^2}\left(f_1^2 \exp(-\hat{\beta}_1)^2 Var(\hat{\beta}_1) + f_1 \exp(-\hat{\beta}_1)^2 (1 - \tfrac{f_1}{n+\hat{f}_0})\right) \tag{19}$$

$$= \frac{1}{T^2} f_1 \exp(-\hat{\beta}_1)^2 \left(f_1 Var(\hat{\beta}_1) + 1 - f_1/(n + \hat{f}_0)\right), \tag{20}$$

where we have used the $\delta-$method for the first term on the RHS of (18). Note that an estimate of $Var(\hat{\beta}_1)$ is readily available from (16). Hence a *prediction interval* for f_0 can be constructed as $\hat{f}_0 \pm 1.96\sqrt{Var(\hat{f}_0)}$ and for N as $n + \hat{f}_0 \pm 1.96\sqrt{Var(\hat{f}_0)}$.

4. APPLICATIONS

We start with the data on the homeless population of Utrecht discussed in section 1.1. We have seen in the binomial ratio plot (upper left panel of Figure 2) that the model

$$\log\left(\frac{x+1}{T-x}\frac{p_{x+1}}{p_x}\right) = \beta_1 + \beta_2 x \tag{21}$$

provides a good approximation of the observed log-ratio $\log\left(\frac{x+1}{T-x}\frac{f_{x+1}}{f_x}\right)$. Hence we use this model to predict $\hat{f}_0 = \exp(-\hat{\beta}_1)f_1/T = 66$, leading to a population size estimate of $\hat{N} = n + \hat{f}_0 = 288$.

It is interesting to note the connection to the multiplicative-binomial distribution (Altham 1978) defined as

$$p_x = \binom{T}{x}\theta^x(1-\theta)^{T-x}\eta^{x(T-x)}/C \tag{22}$$

where $\eta > 0$ is an additional positive parameter and $C = \sum_{x=0}^{T}\binom{T}{x}\theta^x(1-\theta)^{T-x}\eta^{x(T-x)}$. Clearly, if $\eta = 1$ the multiplicative-binomial reduces to the standard binomial. The parameter η catches over- as well as underdispersion although there are no simple ranges for η representing the two forms of non-equidispersion. For more details see Lovison (1998). The ratio regression approach for the multiplicative-binomial yields

$$\log r_x = \log\left(\frac{x+1}{T-x}\frac{p_{x+1}}{p_x}\right) = \beta_1 + \beta_2 x \tag{23}$$

with no restrictions on $\beta_1 = \log[\theta/(1-\theta)] + (m-1)\log(\eta)$ and $\beta_2 = -2\log\eta$. Hence the straight line model for the binomial ratio regression is identical to the multiplicative-binomial.

In Table 5 we have given two additional estimators. One is Chao's estimator provided as $\hat{f}_0 = (T-1)/T f_1^2/(2f_2) = 55$ for the binomial as developed in section 3.2, corresponding to a population size estimate of $\hat{N} = 277$. We can see that the ratio regression approach corrects the Chao estimator upwards. The other is Turing's estimator under homogeneity. Here the idea is to express p_0 as a function of p_1 and the mean. As it turns out for the binomial, $p_0 = (p_1/E(X))^{T/(T-1)}$ where X is binomial with size parameter T. The Turing estimate $\hat{N} = n/(1-\hat{p}_0)$ with $\hat{p}_0 = (f_1/S)^{T/(T-1)}$ follows. Note that $S = f_1 + 2f_2 + \cdots + Tf_T$. One can also view the Turing estimator as a form of coverage estimator as $1 - f_1/S$ represents the sample coverage. For more details on Good-Turing estimation see Good (1953), Bunge and Fitzpatrick (1993), and Chao and Bunge (2002). In the case of the homeless data we find the Turing estimate of the population size of the homeless population to be 225, considerably smaller than the other two estimates which is as expected.

Here we look at the bowel cancer screening data of section 1.4. As the lower left panel of Figure 2 suggests we can use the straight line regression model in this case. Besides the 228 cancer cases detected by the screening programme we estimate 71 additional undetected cancer cases in contrast to Chao's estimator with 33 additional cases. The Turing estimator provides only 7 additional cases, clearly too low. For the details see Table 5.

We have already discussed in section 3.4 the modelling for the Poisson ratio regression of the domestic violence data of section 1.2 where it was found that the model $\log[(x+1)f_{x+1}/f_x] = \beta_1+\beta_2\log(x+1)+\beta_3 x+\epsilon_x$ provided an excellent fit. Note again that the term *Poisson* solely refers to the construction of the response $\log[(x+1)f_{x+1}/f_x]$. Using the model $\log[(x+1)f_{x+1}/f_x] = \beta_1+\beta_3 x+\epsilon_x$ in the ratio regression, we find an estimate of the total number of domestic violence perpetrators in the Netherlands of 131,668. For comparison Chao's estimator $n+f_1^2/(2f_2)$ provides an estimate of 76,451 perpetrators and Turing's estimator $n/(1-f_1/S)$ yields 64,370 persons, only half the size of the ratio regression estimator. Using the better fitting model $\log[(x+1)f_{x+1}/f_x] = \beta_1+\beta_2\log(x+1)+\beta_3 x+\epsilon_x$, we find an estimate of 328,224 perpetrators. The AIC for this model is -47.7 which compares well with the AIC of -6.6 of the former model. However, we have seen in section 2 that there is evidence that the geometric distribution provides a better fit to the domestic violence data than the Poisson distribution. Note that a geometric ratio regression can be viewed as a Poisson ratio regression with an offset term $\log(x+1)$. In our case, the geometric ratio regression model $\log[p_{x+1}/p_x] = \beta_1+\beta_3 x$ is equivalent with the Poisson ratio regression $\log[(x+1)p_{x+1}/p_x] = \log(x+1)+\beta_1+\beta_3 x$. Hence this appears to be a reasonable alternative model to use. Fitting a geometric ratio regression model $\log[f_{x+1}/f_x] = \beta_1+\beta_3 x+\epsilon_x$ leads to an estimate of the total number of domestic violence perpetrators in the Netherlands of 179,979. Based on this estimate, the sample coverage is very low at about 10%, hence the police data base provides only a small peak of the domestic violence iceberg in the Netherlands. This is as expected since dark number research[1] estimates the number of reported domestic crimes between 10% and 20% (Summers and Hoffman 2002). The details are found in Table 5.

Let us now look at the data on the magnitude of worldwide forced labour. We see in Figure 5 that the Poisson ratio regression model $\log[(x+1)f_{x+1}/f_x] = \beta_1+\beta_2 x+\epsilon_x$ provides a reasonable approximation of the pattern visible in the ratio plot. The ratio regression estimate for worldwide number of reports on forced labour is 14,096, almost three times as much as has been found in the sources ($n =$ 5,491). The estimators of Chao and Turing are 12,471 and 12,475, respectively. The details are again in Table 5. Note that Turing and Chao are very close here, despite the fact that there is considerable heterogeneity, illustrating that Chao's estimator is not always able to adjust for heterogeneity satisfactorily. The ratio regression model used here does not correspond to a known probability density although it can be thought of as an approximation of the Conway-Maxwell-Poisson distribution as $\log(x+1) \approx x$ in the vicinity of 1.

[1]Dark number research is a social sciences term for research focussing on elusive target populations such as populations untertaking illegal activities or behaviors.

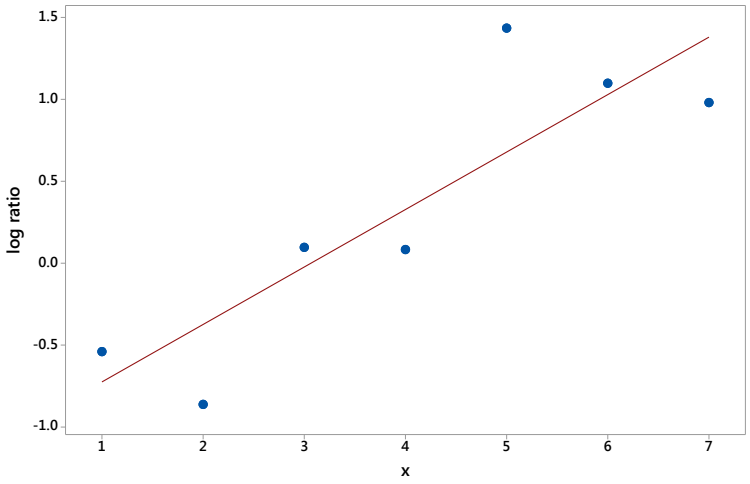

FIG 5. *Poisson ratio regression using $\beta_1 + \beta_2 x$ for the forced labour data of section 1.3*

TABLE 5
Estimates of the population size N for the various applications; RR denotes the ratio regression approach; 'P' stands for Poisson, 'G' for geometric and 'B' for Binomial

				estimates of N with 95% prediction intervals		
application	n	a_x	model	RR	Chao	Turing
1.1	222	B	$\beta_1 + \beta_2 x$	288 (233 - 342)	277 (229 - 324)	225 (224 - 226)
1.2	17,662	P	$\beta_1 + \beta_3 x$	131,668 (106,583 - 156,753)	76,451 (73,363 - 79,538)	64,370 (62,302 - 66,438)
1.2	17,662	P	$\beta_1 + \beta_3 x$ $+\beta_2 \log(x+1)$	328,224 (320,586 - 335,862)		
1.2	17,662	G	$\beta_1 + \beta_2 x$	179,979 (156,718 - 203,240)		
1.3	5,491	P	$\beta_1 + \beta_2 x$	14,096 (10,749 - 17,443)	12,471 (11,916 - 13,026)	12,475 (12,016 - 12,934)
1.4	228	B	$\beta_1 + \beta_2 x$	299 (269 - 329)	261 (238 - 283)	235 (232 - 238)
golf tees ($N = 250$)	162	B	$\beta_1 + \beta_2 x$	218 (173 - 263)	195 (173 - 217)	172 (168 - 176)
taxicabs ($N = 420$)	283	B	$\beta_1 + \beta_2 x$	411 (310 - 512)	395 (353 - 437)	376 (350 - 402)

We conclude this section by applying the method to two data sets for which the true population size is known. The first one is reported in Borchers *et al.* (2004) and goes back to a capture-recapture experiment. Golf tees were placed

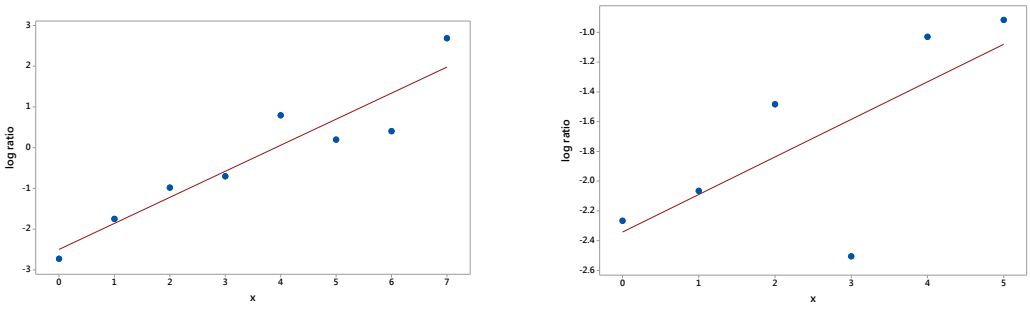

FIG 6. *Binomial ratio plot with $T = 8$ for golf tees data (left panel) and binomial ratio plot with $T = 10$ for taxicab data (right panel)*

in 250 clusters in a specific area on grounds of the University of St. Andrews (Scotland) and 8 surveyors were used to recover them. Of the total of 250 golf tee clusters 162 could be recovered successfully. Details are provided in Table 6. Note that here $f_0 = 88$ is known but we will not use this information in the estimation process. The binomial ratio plot for these data is shown in Figure 6 (left panel). Based on this graph we think that a straight line regression $\log\left([(x+1)/(T-x)f_{x+1}/f_x]\right) = \beta_1 + \beta_2 x + \epsilon_x$ is not inappropriate. The estimators for these data are presented in Table 5. For the binomial ratio regression we find an estimate of 218 which improves upon Chao's (195) and Turing's (172) estimate and compares favorably with the true size of 250. In fact, the prediction interval for N (see also section 3.5) based on the binomial ratio regression is $(173-263)$ with the upper interval end covering the true $N = 250$. The prediction interval for N based upon Chao's estimator is instead $(173 - 217)$, not including the true $N = 250$. Similarly, the prediction interval for N based upon Turing's estimator is $(168 - 176)$, clearly not including the true N.

TABLE 6
Frequency distribution of recovery count per golf tee cluster (Borchers et al. *2004)*

x	0	1	2	3	4	5	6	7	8	n
f_x	(88)	46	28	21	13	23	14	6	11	162

TABLE 7
Frequency distribution of count of identifications per taxicab (Carothers 1973); there were no counts larger than $T_0 = 6$

x	0	1	2	3	4	5	6	n
f_x	(137)	142	81	49	7	3	1	283

The second data set is reported in McCrea and Morgan (2015) and goes back to Carothers (1973). The number of registered taxicabs in Edinburgh (Scotland) is known to be $N = 420$ at the time of the experiment. On $T = 10$ sighting occasions the passing taxicabs are identified and the count of re-sightings per taxicab determined. The details are provided in Table 7. $n = 283$ different taxicabs could

be identified of which 142 were seen only once, 81 twice, and so forth. No taxicab had been identified more than 6 times. For the binomial ratio regression we also use a straight line model as is motivated by the right panel in Figure 6. The associated binomial ratio regression estimate of the population size is 411 which is close to the known number of taxicabs of 420. Chao's and Turing's estimates are 395 and 376, respectively (see also Table 6). In this data set there is more variation as the prediction intervals for N based on the binomial ratio regression and Chao's estimator are both wide: $(310 - 512)$ for the ratio regression and $(353 - 437)$ for Chao's estimator. Both easily cover the true $N = 420$. Turing's estimator underestimates with a prediction interval of $(350 - 402)$, not including the true $N = 420$.

These examples and applications show that the ratio regression approach can be a valuable tool in estimating population size.

5. EXTENSIONS AND DISCUSSION

The question arises of what happens if part of the target population remains undetectable. For example, in the case study of homeless people in Utrecht 1.1, some homeless people might never visit a shelter to stay overnight. As for any other method the ratio regression approach assumes that there is a positive detection probability. If this is not the case, then, even if the observational period is chosen to be large, some homeless people remain undetected and the ratio regression approach will provide only a lower bound. Hence it is crucial to discuss with practitioners responsible for the well-being of homeless people how realistic the assumption is that every homeless person is likely to visit the shelter at some time.

Some guidance for the practical use of the ratio regression model might be appropriate. The first important choice is the *base* family as this leads to the coefficients a_x, for $x = 0, 1, \cdots$. For example, if there is a finite number of trapping occasions such as in applications 1.1 and 1.4 the natural base family is the binomial and every regression model considered should include an intercept term so that the binomial is included as a special case. In data examples such as 1.2 or 1.3, the base family is less clear as at least the Poisson or the geometric distributions could be considered. Here ratio plotting might help and the distribution with least positive trend might be chosen as base family (and hence determine the coefficients a_x). The choice of link function is usually not a problem as the log-link is typically suitable. Choosing the regression model is clearly important and guidance can be received again from the ratio plot. However, several models might appear equally suitable and model selection criteria such as the Akaike information criterion might be used to select models. Finally, goodness-of-fit analysis could be provided as already mentioned in section 2. The ratio regression approach can be widely applied, clearly also to ecological data. However, it should be mentioned that sample sizes should be at least moderate as the ratios f_{x+1}/f_x need to be constructed on the basis of frequency distribution of the count of captures X. Depending on the spread of the distribution, our experience is that $n > 50$ is desirable.

The approach can be extended in several ways. A very interesting extension is that validation information can be easily incorporated into the ratio regression modeling. To demonstrate, we again consider the bowel cancer application of

section 1.4. For some reason a subsample of the screened persons with confirmed bowel cancer took the diagnostic test a second time for six consecutive times. The results for $n = 122$ persons with confirmed cancer are found in Table 8. As we know the cancer status of all 122 persons participating in this secondary application we do observe zero counts. For 22 persons with bowel cancer the diagnostic test was negative at all times. We call this a *validation sample* as there are no hidden cases here.

TABLE 8
Frequency distribution of number of positive secondary tests of those with confirmed cancer (Lloyd and Frommer 2004b)

x	0	1	2	3	4	5	6	n
f_x	22	8	12	16	21	12	31	122

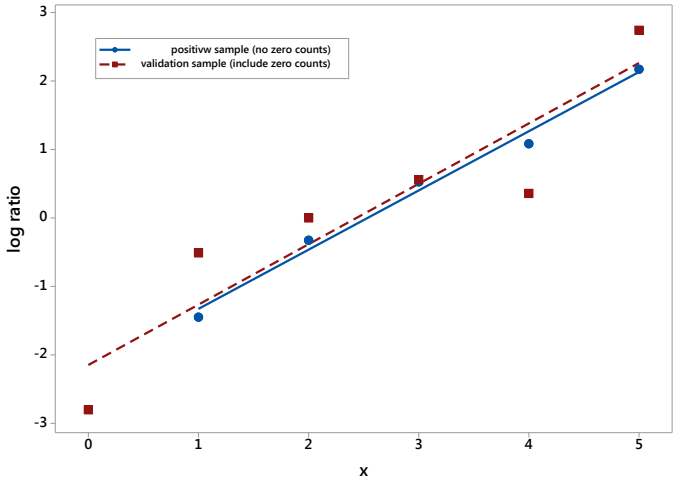

FIG 7. *Binomial ratio regression with additional validation sample using* $\log[(x+1)/(T-x)f_{x+1}/f_x] = \beta_1 + \beta_2 x + \beta_3 \texttt{set} + \beta_4 x \times \texttt{set} + \epsilon_x$ *for the bowel cancer data of section 1.4 with additional validation information*

It is now possible to incorporate the information coming from the validation sample into the modeling as is done in (24):

$$\log[(x+1)/(T-x)f_{x+1}/f_x] = \beta_1 + \beta_2 x + \beta_3 \texttt{set} + \beta_4 x \times \texttt{set} + \epsilon_x. \tag{24}$$

Here `set` represents a dummy variable which takes on the value 1 if f_x is from the validation sample and 0 otherwise. The model (24) allows two completely separate lines for the positive sample (where no zero counts are observed) and the validation sample (where zero counts are observed), respectively. The associated

graph is given in Figure 7. If $\beta_4 = 0$ the parallel line model occurs and if $\beta_3 = 0$, in addition, the two lines become identical. Tests for these hypotheses can be done in standard ways and, in our case, there is no evidence to reject the identical lines model as also Figure 7 indicates. The resulting estimate is 298 persons with cancer which is not much different from the estimate of 299 achieved by the positive sample only. Using a validation sample does not only lead to an increased efficiency it also reassures that the model, used for the positive sample to predict the frequency of hidden zero counts, is also a reasonable model for the prediction. In the parallel line model, the prediction would still partly use the validation sample whereas in the separate line model the validation sample is not used at all in predicting f_0.

Ratio plotting has been proposed in Böhning *et al.* (2013) and connected work has been referenced therein. See also McCrea and Morgan (2015). Ratio regression for the Poisson case has been suggested in Rocchetti *et al.* (2011) and a special fractional polynomial model for the binomial ratio case by Hwang and Shen (2010). This paper develops the most general form of the ratio regression approach as it allows any member of the power series distribution as base distribution, a basically unlimited choice of regression model which is connected to the ratio of neighboring frequencies by a feasible link function.

REFERENCES

[1] Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Journal of the Royal Statistical Society Series C (Applied Statistics)* **27** 162-167.

[2] Bales, K. (2012). *Disposable People: New Slavery in the Global Economy* Oakland, University of California Press.

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

[5] Böhning, D. and Del Rio Vilas, V. (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.

[6] Böhning, D., Baksh, M.F., Lerdsuwansri, R. and Gallagher, J. (2013). The use of the ratio-plot in capture-recapture estimation. *Journal of Computational and Graphical Statistics* **22** 135–155.

[7] Borchers, D.L., Buckland, S.T. and Zucchini, W. (2004). *Estimating Animal Abundance. Closed Populations.* London, Springer.

[8] Bunge, J. and Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association* **88** 364–373.

[9] Carlin, B. P. and Louis, T. A. (2011). *Bayesian methods for data analysis, 3rd edition.* Monographs on Statistics and Applied Probability, London, Chapman & Hall/CRC.

[10] Carothers, A.D. (1973). Capture-Recapture methods applied to a population with known parameters. *The Journal of Animal Ecology* **42** 125–146.

[11] Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43** 783–791.

[12] Chao, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics* **45** 427–438.

[13] Chao, A. (2001). An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics* **6** 158–175.

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

[15] Cormack, R.M. (1964). Estimates of survival from sightings of marked animals. *Biometrika* **51** 429–438.

[16] Cormack, R.M. (1989). Log-linear models for capture-recapture. *Biometrics* **45** 395–413.

[17] 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.
[18] Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59** 591–603.
[19] Good, I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264.
[20] Hwang, W.-H. and Shen, T.-J. (2010). Small-sample estimation of species richness applied to forest communities. *Biometrics* **66** 1052–1060.
[21] ILO (2012). *ILO Global Estimate of Forced Labour. Results and Methodology.* International Labour Organization, Geneva.
[22] Jolly, G.M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model . *Biometrika* **52** 225–247.
[23] Lloyd, C.J. and Frommer (2004a). Estimating the false negative fraction for a multiple screening test for bowel cancer when negatives are not verified. *Australian and New Zealand Journal of Statistics* **46** 531–542.
[24] Lloyd, C.J. and Frommer (2004b). Regression based estimation of the false negative fraction when multiple negatives are unverified. *Journal of the Royal Statistical Society Series C* **53** 619–631.
[25] Lloyd, C.J. and Frommer (2008). An application of multinomial logistic regression to estimating performance of a multiple-screening test with incomplete verification. *Journal of the Royal Statistical Society Series C* **57** 89–102.
[26] Lovison, G. (1998). An alternative representation of Altham's multiplicative-binomial distribution. *Statistics & Probability Letters* **36** 415–420.
[28] Mao, C.-X. and You, N. (2009). On comparison of mixture models for closed population capture-recapture studies. *Biometrics* **65** 547-553.
[28] Mao, C.-X. (2008). On the nonidentifiability of population sizes. *Biometrics* **64** 977-979.
[29] McCrea, R.S. and Morgan, B.J.T. (2015). *Analysis of Capture-Recapture Data*. Chapman & Hall/CRC, Boca Raton.
[30] Norris, J.L. and Pollock, K.H. (1996). Nonparametric MLE under two closed capture-recapture models with heterogeneity. *Biometrics* **52** 639–649.
[31] Pledger, S. A. (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics* **61** 868–876.
[32] Robbins, H. (1955). An empirical Bayes approach to statistics. *In Proceedings of the 3rd Berkeley Symposium on Mathematical Statistitcs and Probability 1.* Berkeley, CA: University of California Press, 157–164.
[33] Rocchetti, I., Bunge, J. and Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics*, **5** 1512–1533.
[34] Sanathanan, L (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association* **72** 669–672.
[35] Seber, G.A.F. (1965). A note on the the multiple-recapture census. *Biometrika* **52** 249–259.
[36] Sellers, K.F. and Shmueli, G. (2010). A flexible regression model for count data. *The Annals of Applied Statistics* **4** 943–961.
[37] Summers, R.W. and Hoffman, A.M. (2002). *Domestic Violence. A Global View*. Greenwood Press, Westport.
[38] van der Heijden, P. G. M., Cruyff, M. and 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.
[39] Van der Heijden, P. G. M., Cruyff, M. and Böhning, D. (2014a). *Analyses daklozen Utrecht 2013.* Universiteit Utrecht en University of Southampton. Utrecht, 28 januari 2014.
[40] van der Heijden, P.G.M., Cruyff, M. and Böhning, D. (2014b). Capture-recapture to estimate crime populations. In: G.J.N. Bruinsma and D.L. Weisburd (eds.). *Encyclepedia of Criminology and Criminal Justice.* Berlin: Springer, 267 278.
[41] 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.
[42] 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.

[43] Wilson, R.M. and Collins, M.F. (1992). Capture-recapture estimation with samples of size one using frequency data. *Biometrika* **79** 543–553.

APPENDIX

Proof. Let $r_x > 0$ be given for $x = 0, \cdots, T_0 - 1$. Any probability distribution $p_0, \ldots, p_{T_0} > 0$ will meet the constraint $p_0 + \cdots + p_{T_0} = 1$. Since the probability distribution needs also to fulfill the recurrence relation $p_{x+1} = r_x p_x a_{x+1}/a_x$, we have that

$$1 = p_0 + \cdots + p_{T_0} = p_0 + p_0 r_0 a_1/a_0 + (p_0 r_0 a_1/a_0)(r_1 a_2/a_1) + \cdots + p_0 \prod_{x=0}^{T_0-1} r_x a_{x+1}/a_x$$

$$= p_0[1 + r_0 a_1/a_0 + (r_0 a_1/a_0)(r_1 a_2/a_1) + \cdots + \prod_{x=0}^{T_0-1} r_x a_{x+1}/a_x].$$

Hence, it follows that

$$p_0 = 1/[1 + r_0 a_1/a_0 + (r_0 a_1/a_0)(r_1 a_2/a_1) + \cdots + \prod_{x=0}^{T_0-1} r_x a_{x+1}/a_x]$$

necessarily, and $0 < p_0 < 1$. The remaining probabilities follow uniquely from the recurrence formula. According to the latter, $p_{x+1} = r_x p_x/a_x$ implies that $0 < p_{x+1} < 1$, $x = 0, \ldots, T_0 - 1$. □