Biometrics 72, 697–706
September 2016

DOI: 10.1111/biom.12485

A Flexible Ratio Regression Approach for Zero-Truncated Capture–Recapture Counts

Dankmar Böhning,[1,*] Irene Rocchetti,[2,] Marco Alfó,[3,***] and Heinz Holling[4,****]**

[1]Department of Mathematical Sciences and Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton, SO17 1BJ, U.K.
[2]Istituto Nazionale di Statistica, Rome, Italy
[3]Dipartimento di Scienze Statistiche, Sapienza Universitá di Roma, Rome, Italy
[4]Department of Psychology and Sport Science, University of Münster, Münster, Germany
*email: d.a.bohning@soton.ac.uk
**email: irocchetti@istat.it
***email: marco.alfo@uniroma1.it
****email: holling@uni-muenster.de

Summary. Capture–recapture methods are used to estimate the size of a population of interest which is only partially observed. In such studies, each member of the population carries a count of the number of times it has been identified during the observational period. In real-life applications, only positive counts are recorded, and we get a truncated at zero-observed distribution. We need to use the truncated count distribution to estimate the number of unobserved units. We consider ratios of neighboring count probabilities, estimated by ratios of observed frequencies, regardless of whether we have a zero-truncated or an untruncated distribution. Rocchetti et al. (2011) have shown that, for densities in the Katz family, these ratios can be modeled by a regression approach, and Rocchetti et al. (2014) have specialized the approach to the beta-binomial distribution. Once the regression model has been estimated, the unobserved frequency of zero counts can be simply derived. The guiding principle is that it is often easier to find an appropriate regression model than a proper model for the count distribution. However, a full analysis of the connection between the regression model and the associated count distribution has been missing. In this manuscript, we fill the gap and show that the regression model approach leads, under general conditions, to a valid count distribution; we also consider a wider class of regression models, based on fractional polynomials. The proposed approach is illustrated by analyzing various empirical applications, and by means of a simulation study.

Key words: Capture–recapture; Mixed binomial distributions; Ratio regression estimator; Zero-truncated model.

1. Introduction and Background

Let us consider a target population with size N, and assume we are interested in estimating its global size. Often, for this purpose, an identification mechanism is repeatedly used to register units from the population. Only a portion of the population is registered, and we need to estimate the number of unobserved units. Let us consider the binary indicator variable $x_{it},\ i = 1, \ldots, N,\ t = 1, \ldots, T$, where $x_{it} = 1$ if the i-th unit has been identified at the t-th measurement occasion. It is assumed that $x_i = \sum_{t=1}^{T} x_{it}$ is available only if $x_i > 0$, i.e., $\exists t = 1, \ldots, T : x_{it} > 0$. When $x_i = 0$, the i-th unit remains *unobserved*. The quantity T may be known a priori, or it may denote the maximum observed count, e.g., if we look at the number of lesions of a given type in a sample of patients. By simply arranging units indices, we may distinguish between the untruncated population $X_1, X_2, \ldots, X_N$ and the truncated sample $X_1, X_2, \ldots, X_n$ where, without lack of generality, $X_{n+1} = \cdots = X_N = 0$. The target population can be described by a probability density function (x, p_x), where $x = 0, 1, \ldots$, and p_x denotes the probability of exactly x identifications for a generic unit, $p_x \geq 0$ and $\sum_{x=0}^{\infty} p_x = 1$. If we denote by f_x the empirical frequency of units with count x, we notice that partial observation leads to a zero-truncated sample with size $n = \sum_{x>0} f_x$. The empirical relative frequency f_x/N (which cannot be computed since N is unknown) gives an estimate of p_x, while the observed empirical relative frequency f_x/n provides an estimate of the zero-truncated probability $p_x/(1 - p_0)$. As a result of the study design, f_0 and $N = \sum_{x=0}^{T} f_x$ remain unknown. Starting from the observed zero-truncated distribution, our purpose is to find an estimate of the population size N, a special form of the capture–recapture problem (see Wilson and Collins, 1992; Bunge and Fitzpatrick, 1993; Chao, 2001).

A popular choice is to model the distribution of X using an appropriate counting distribution, e.g., the binomial for fixed T or the Poisson when T is not known in advance. We will focus on the case with T fixed in advance. Since the observed counts derive from repeated measurements of the same unit, and due to potential individual-specific unobserved heterogeneity, mixed binomial distributions are an obvious choice. It is worth noticing that, using a mixed binomial distribution, we account for between-individual variation but not for within-individual variation. In the capture–recapture framework, the choice of a mixed binomial model suffers from the lack of

identifiability of the mixing distribution (see Link, 2003), and from boundary problems in ML estimation, see, e.g., Dorazio and Royle (2005). Starting from the beta-binomial case, Rocchetti et al. (2014) propose a simple regression model to describe ratios of successive probabilities, which can be fitted using observed frequencies. Once fitted, the model is projected backward onto $x = 0$ to estimate the frequency f_0 of unobserved units. One may wonder whether a clear connection does exist between the regression model and a proper counting distribution. That is, whether any regression model corresponds to a proper counting distribution. In this manuscript, we fill the gap and show that, under general conditions, each regression model corresponds to a proper counting distribution. In this respect, we may also consider a wider and more flexible family of regression models, based on the use of fractional polynomials, to cover a wide variety of empirical situations.

The article is structured as follows: in Section 2, we introduce some benchmark data examples, to be used throughout the article. In Section 3, the probability ratio plot is proposed as a screening tool to detect potential departures from homogeneity. Starting from the properties of the ratio plot, Section 4 discusses the estimation of the global size for the target population. In Section 5, the proposed estimators are applied to the benchmark data examples introduced in Section 2; we also discuss a model-averaged estimator. Identifiability of the mixing distribution in zero-truncated binomial mixtures is discussed in Section 6. Section 7 entails the analysis of a further, historic, example, where a mixed Poisson distribution represents a sensible choice. The last section contains discussion and concluding remarks. The results from a simulation study are available as Supplementary Material.

2. Examples

Let us consider some real-life benchmark data examples. In two cases, f_0 is known, and the studies will be used to illustrate how well any estimator can recover f_0.

2.1. *Utrecht Homeless Data*

The city of Utrecht (NL) runs a shelter where homeless people can stay overnight. The shelter is assumed to cover the city of Utrecht only. The distribution of the number of nights homeless people stayed in the shelter within a 14-nights period in 2013 is reported in the Supplementary Material. In this case $T = 14$, $f_1 = 36$ people stayed one night, $f_2 = 11$ people stayed two nights, and so forth. In total, 222 different homeless people stayed in the shelter during the period, spending a total of $S = \sum_{x=1}^{14} x f_x = 2009$ nights there. For more details, see van der Heijden et al. (2014). It can be argued that not all the homeless people spent at least one night at the shelter during the analyzed period. With the aim at improving social policy interventions, the city of Utrecht is interested in estimating the total size of its homeless population, N, or, equivalently, the size f_0 of the hidden homeless population.

2.2. *Golf Tees Data*

The data entail a well-known field experiment: 250 groups of golf tees, of two colors, have been placed in groups with different sizes in a survey region of $1680\,\mathrm{m}^2$, either exposed above the surrounding grass, or partly hidden by it. They have been surveyed by the 1999 statistics honors class at the University of St. Andrews (Scotland). Borchers et al. (2004) give details. A total of 162 groups of tees were found and $f_0 =$ 88 group of tees were missed. The observed distribution refers to the count of times each group of tees has been found by eight independent observers, see the Supplementary Material.

2.3. *Bowel Cancer Data*

Over several years, from 1984 onwards, about 50,000 subjects were screened for bowel cancer at St Vincent's Hospital in Sydney (Australia), see Lloyd and Frommer (2004a, 2004b, 2008). The screening procedure was based on a sequence of binary diagnostic tests, self-administered on $T = 6$ successive days. Since no screening test is 100% accurate, replications of the diagnostic test over a number of days may help identify most cases. On each of the six occasions, the presence $x_{it} = 1$ of blood in feces has been recorded. People with six negative tests were not further assessed and it remains unknown which disease status they have, while people with at least one positive test had their true disease status determined by physical examination, sigmoidoscopy, and colonoscopy. The aim is to estimate how many (say f_0) cancer patients have been *missed* by adopting this screening procedure. The frequency distribution of the number of positive tests X is provided in the Supplementary Material, see the row *cancer (primary)*. Lloyd and Frommer (2004b) mention that 122 patients with confirmed cancer status were screened again using the identical screening procedure. The frequency distribution is provided in the Supplementary Material, see the row *cancer (secondary)*. We will focus on this secondary distribution as f_0 is known there.

3. The Probability Ratio

Approaches to estimating the population size, N, or the number of unregistered units, f_0, from the observed, zero-truncated, count distribution follow a general scheme; p_x is modeled by using some known distribution $p_x(\theta)$, indexed by the parameter θ. Based on the observed data, and using the zero-truncated distribution $p_x(\theta)/[1 - p_0(\theta)]$, an estimate $\hat{\theta}$ is used to derive an estimate of N by means of the Horvitz–Thompson estimator:

$$\hat{N} = n/[1 - p_0(\hat{\theta})], \quad \hat{f}_0 = n\frac{p_0(\hat{\theta})}{1 - p_0(\hat{\theta})}.$$

To illustrate this procedure, let us consider the binomial probability distribution

$$p_x(\theta) = P(X = x \mid \theta) = \binom{T}{x}\theta^x(1-\theta)^{T-x}, \tag{1}$$

$x = 0, \ldots, T$ and $p_x(\theta) = 0$ for $x > T$. In this case, $p_0(\theta) = (1 - \theta)^T$ and the Horvitz–Thompson estimator is defined by

$$\hat{N} = n/[1 - (1 - \hat{\theta})^T].$$

Usually, θ is estimated fitting a zero-truncated distribution to the observed data, e.g., through an EM-type algorithm. The major problem with homogeneous binomial models is that they are often not flexible enough to produce good fit to the

observed (zero-truncated) distribution. In fact, unobserved heterogeneity may play a role in determining variability in the probability to be registered; so, it is important to have a screening tool for *binomiality*. This tool may be built on an interesting property of the binomial distribution, see Hoaglin (1980):

$$\frac{p_{x+1}(\theta)}{p_x(\theta)} = \frac{\binom{T}{x+1}\theta^{x+1}(1-\theta)^{T-x-1}}{\binom{T}{x}\theta^{x}(1-\theta)^{T-x}} = \frac{T-x}{x+1}\frac{\theta}{1-\theta},$$

that is

$$R_x = \frac{x+1}{T-x}\frac{p_{x+1}(\theta)}{p_x(\theta)} = a_x\frac{p_{x+1}(\theta)}{p_x(\theta)} = \frac{\theta}{1-\theta}, \tag{2}$$

where $a_x = \binom{T}{x}/\binom{T}{x+1} = (x+1)/(T-x)$. Therefore, in the binomial distribution, the ratio R_x is constant with respect to x. It is straightforward to estimate $R_x = a_x\frac{p_{x+1}}{p_x}$ by

$$r_x = a_x\frac{f_{x+1}/N}{f_x/N} = a_x\frac{f_{x+1}}{f_x},$$

where f_x denotes the number of units that have been identified x times; this estimate does not change whether we consider the truncated or the untruncated distributions. The graph $x \to r_x = a_x\frac{f_{x+1}}{f_x}$ is called the *ratio plot* and was developed as a *diagnostic device* for the binomial by Böhning et al. (2013). In a ratio plot, the pattern of a horizontal line can be taken as supporting evidence for a binomial distribution. This is shown in Figure 1a in the Supplementary Material, where 50,000 simulated data from a binomial distribution with index $T = 6$ and parameter $\theta = 0.4$ are reported on the ratio scale (left panel). The ratio plot shows clear evidence for a binomial distribution, while this feature is more difficult to recognize in the frequency plot (right panel). Despite the (almost) absence of any random error, the nature of the distribution is not easily recognized, whereas the binomial structure can easily be evinced from the ratio plot. Hence, the motivation for the use of the ratio plot is in that it clearly shows whether substantial departures from the homogeneous binomial distribution are observed; in the presence of a high sample size and number of trials, it may help detecting a discrete mixing. For a smaller sample size, random error comes in and the ratio plot could be supplemented by error bars to account for uncertainty. If we apply the ratio plot concept to homeless people data, there is no evidence of a horizontal line, and the same is true for the golf tees data. Instead, we observe a monotone pattern which might be used as supporting evidence for population heterogeneity; a similar increasing pattern can be observed for the bowel cancer data, see Figure 1b–d in the Supplementary Material. As a consequence, we will consider models where a mixing distribution $h(\theta)$ describes population heterogeneity in the identification rates. The marginal distribution is as follows:

$$p_x = \int_0^1 \binom{T}{x}\theta^x(1-\theta)^{T-x}h(\theta)d\theta. \tag{3}$$

The shape of the marginal distribution may vary substantially as a function of the mixing distribution, as this term controls the departure from the homogeneous binomial model.

When the mixing distribution is not described by a one-point mass (leading to the binomial distribution), it can be shown that the ratios R_x are increasing in x. The ratio plots we have seen for the benchmark data examples seem to suggest the presence of unobserved population heterogeneity. Parametric choices for $h(\theta)$ such as the beta distribution have been considered which often improve the fit considerably when compared to the binomial model. Discrete mixture models have also been suggested, see, e.g., Norris and Pollock (1996), Pledger (2000), and Böhning and Kuhnert (2006). However, boundary problems may arise when the parameter approaches the borders of the segment $(0, 1)$, see Wang and Lindsay (2005, 2008), and identifiability is an issue of great concern, see Link (2003). Given that we only observe the zero-truncated distribution, we are left with the unsolved problem of choosing which mixing is the best, not in terms of the observed fit, but rather in terms of estimating the unknown f_0. While a general solution to the problem does not exist, a sub-optimal solution is to restrict the attention to identifiable families of distributions. The question is how do we achieve alternative families? Could the ratio plot be used to determine the family of interest? In this article, we propose a general approach which produces identifiable families of distributions that can be used to estimate the population size.

4. The Regression Approach

Let us start by the mixture model in equation (3). If we assume that the parameter θ is distributed according to an arbitrary density $h(\theta)$, the marginal distribution is

$$p_x = \int_0^1 \binom{T}{x}\theta^x(1-\theta)^{T-x}h(\theta)d\theta. \tag{4}$$

As remarked before, it can be easily proven that the marginal distribution satisfies the following monotonicity property (Böhning et al., 2013)

$$a_0\frac{p_1}{p_0} \le a_1\frac{p_2}{p_1} \le a_2\frac{p_3}{p_2} \le \ldots,$$

where $a_x = \binom{T}{x}/\binom{T}{x+1} = \frac{x+1}{T-x}$. In other words, the ratio plot for binomial mixtures is monotone nondecreasing. In the context of species richness estimation, Hwang and Shen (2010) consider the reciprocals of the elements in the ratio plot, and prove they define a monotone nonincreasing sequence. Rivest and Baillargeon (2014) consider terms $\log\left(p_x/\binom{T}{x}\right)$ and show that, in the presence of individual-specific heterogeneity, they define a convex sequence in $x = 1, \ldots, T$, that is, the ratio plot is nondecreasing. These results and the analyzed examples suggest to model explicitly R_x as a nondecreasing function of x. This *ratio regression* approach can be used to identify an appropriate distributional form without the need to parametrically specify the mixing density $h(\theta)$. Let us assume that there exists an unknown probability distribution $p_1, \ldots, p_T$ with all probabilities positive, i.e., $p_x > 0$, $\forall x = 0, \ldots, T$, and

let us consider the ratios:

$$R_x = a_x p_{x+1}/p_x, \tag{5}$$

$x = 0, \ldots, T-1$. The coefficients a_x are known constants, determined by the choice of the *reference* distribution we would like to include. The reference distribution defines the homogeneous distribution we get when no unobserved heterogeneity is present; that is, the conditional distribution we use in (4). To give an example, if the upper limit T is known and fixed, we may choose the binomial as reference distribution, with $a_x = (x+1)/(T-x)$. If the range of the counts has no upper limit, we may would like to include the Poisson as the reference distribution with $a_x = (x+1)$. The point is that, if the observed count data follow the reference distribution, the associated ratios $R_x = a_x p_{x+1}/p_x$, $x = 0, \ldots, T-1$ are constant over X. This implies that any regression model for R_x (or a suitable transformation of it) with only the intercept term represents the reference distribution and, for this reason, a non-null slope implies some unobserved heterogeneity. Let us assume that R_x can be linked to a known set of predictor functions $z_0(x), \ldots, z_p(x)$, so that the following model is defined:

$$g(R_x) = \beta' \mathbf{z}(x), \tag{6}$$

where $x = 0, \ldots, T-1$, and $g(\cdot)$ is a monotone link-function, e.g., $\log(R_x) = \beta_0 + \beta_1 x$ with $z_0(x) = 1$, $z_1(x) = x$, that is $R_x = \exp(\beta_0 + \beta_1 x)$. A general result can be proven.

THEOREM 1. *Let $R_x > 0$ be given for $x = 0, \ldots, T-1$, and let a_x, $x = 0, \ldots, T-1$, be known positive coefficients. Then, there exists a unique probability distribution $p_0, \ldots, p_T > 0$ such that:*

$$p_{x+1} = R_x p_x / a_x, \quad x = 0, \ldots, T-1.$$

Furthermore, we have that

$$p_0 = \left[1 + R_0/a_0 + (R_0/a_0)(R_1/a_1) + \cdots + \prod_{x=0}^{T-1} R_x/a_x \right]^{-1}.$$

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

$$\begin{aligned} 1 = p_0 + \cdots + p_T &= p_0 + p_0 R_0/a_0 + p_0 R_0/a_0 R_1/a_1 \\ &+ \cdots + p_0 \prod_{x=0}^{T-1} R_x/a_x = p_0 \left(1 + R_0/a_0 + (R_0/a_0)(R_1/a_1) \right. \\ &\left. + \cdots + \prod_{x=0}^{T-1} R_x/a_x \right). \end{aligned} \tag{7}$$

Hence, it follows that

$$p_0 = 1 \Big/ \left[1 + R_0/a_0 + (R_0/a_0)(R_1/a_1) + \cdots + \prod_{x=0}^{T-1} R_x/a_x \right],$$

necessarily, and $0 < p_0 < 1$. The remaining probabilities follow from the recurrence formula, and $p_{x+1} = R_x p_x / a_x$ implies that $0 < p_{x+1} < 1$, $x = 0, \ldots, T-1$. This ends the proof. □

The value of this theorem lies in the fact that *any* regression model fulfilling the regularity condition $R_x > 0$, $x = 0, \ldots, T-1$ leads to a proper probability distribution, which is obtained by mixing the reference distribution, specified by the coefficients a_x. The link function defines a one-to-one mapping from the positive axis into the real line, and guarantees the regularity conditions $R_x > 0$, $x = 0, \ldots, T-1$ hold. Estimation may be based on the likelihood function

$$L(\beta) = \prod_{x=1}^{T} \left(\frac{p_x}{1 - p_0} \right)^{f_x},$$

where p_x is a function of $R_x = g^{-1}(\beta' \mathbf{z}(x))$, and hence of β, via Theorem 1. We suggest to use the following procedure for practical purposes. We estimate R_x by its empirical counterpart, $r_x = a_x \frac{f_{x+1}}{f_x}$, and study its dependence on x. This process could help generate ideas on how to develop an appropriate regression model. Once we have chosen the link function $g(\cdot)$, we fit the model

$$g(r_x) = \beta' \mathbf{z}(x) + \boldsymbol{\epsilon}_x, \tag{8}$$

where ϵ_x is such that $\mathrm{E}(\boldsymbol{\epsilon}_x) = \mathbf{0}$ and $\mathrm{cov}(\boldsymbol{\epsilon}_x) = \Sigma$. Here, $\beta = (\beta_0, \ldots, \beta_p)'$ represents a $(p+1)$-dimensional vector of unknown fixed parameters, associated to the regression functions $\mathbf{z}(x) = (z_0(x), \ldots, z_p(x))'$. If an estimate $\hat{\Sigma}$ is available, the generalized least squares estimate of β is known to be

$$\hat{\beta} = (\mathbf{X}' \hat{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}' \hat{\Sigma}^{-1} \mathbf{Y},$$

where $\mathbf{Y}$ has elements $g(r_x)$ and $\mathbf{X}$ has rows $z_0(x), \ldots, z_p(x)$, $x = 1, \ldots, T-1$, since no observation is available for $x = 0$. Details on how to estimate Σ are discussed in Rocchetti et al. (2011, 2014). One of the peculiar features of the ratio regression approach is that the model remains invariant whether the untruncated or the zero-truncated distribution is considered. In fact, we may observe that:

$$R_x = a_x \frac{p_{x+1}}{p_x} = a_x \frac{p_{x+1}/(1-p_0)}{p_x/(1-p_0)},$$

$x = 0, \ldots, T-1$. Clearly, R_0 is defined for the untruncated distribution only. For the zero-count frequency, a regression-based estimator can be derived as follows:

$$\widehat{g(r_0)} = \hat{\beta}' \mathbf{z}(0) \Rightarrow \hat{r}_0 = g^{-1}\left(\hat{\beta}' \mathbf{z}(0) \right).$$

Two estimators can be defined on the basis of the estimated regression model for r_x. First, we can use the fitted values

$\hat{r}_x = g^{-1}(\hat{\beta}'\mathbf{z}(x))$, $x = 0, \ldots, T-1$ to estimate the corresponding probability mass at 0 according to Theorem 1:

$$\hat{p}_0 = \left[1 + \hat{r}_0/a_0 + (\hat{r}_0/a_0)(\hat{r}_1/a_1) + \cdots + \prod_{x=0}^{T-1} \hat{r}_x/a_x\right]^{-1}. \quad (9)$$

Given this probability mass, the Horvitz–Thompson estimator is as follows:

$$\hat{N}_{\mathrm{HT}} = \frac{n}{1-\hat{p}_0} = n + \hat{f}_{0,HT}.$$

Second, once a given regression model has been fitted and corresponding parameters estimated, we may use the recurrence relation $r_x = a_x f_{x+1}/f_x$ and project it onto $x = 0$, to get an estimate of f_0:

$$\hat{f}_{0,\mathrm{reg}} = a_0 f_1/\hat{r}_0 = a_0 f_1/g^{-1}(\hat{\beta}'\mathbf{z}(0)).$$

The associated population size estimator follows:

$$\hat{N}_{\mathrm{reg}} = n + \hat{f}_{0,\mathrm{reg}}.$$

4.1. *Using Fractional Polynomials*

The ratio regression approach allows a wide range of regression models to be considered to fit empirical ratios. The only restriction is that the model should have an intercept β_0 included and the link function should be such that $\hat{r}_x = g^{-1}(\hat{\beta}'\mathbf{z}(0)) > 0$, $x = 0, \ldots, T-1$. The former guarantees that the associated reference distribution is included, the latter ensures that the regression model corresponds to a proper probability distribution. For example, if $a_x = x+1$, then $\log(R_x) = \beta_0$ implies that the associated distribution, according to Theorem 1, is the Poisson. In principle, there are no further restrictions on the side of possible regression models. In the following, we will use a log link, that is $g(r_x) = \log(r_x)$, but other choices are possible as well. While a single choice for the link function may be considered as a restriction, we propose to widen the family by considering a more flexible regression predictor. Since any regression model will ultimately be used for prediction, it should be simple to estimate and robust to departures due to sampling variability; that is, it should perform stable. We found that fractional polynomials may be appropriate in this context because they are simple, stable, and may approximate a wide range of continuous functions, see Royston and Sauerbrei (2008). A fractional polynomial of order k for r_x is as follows:

$$g(r_x) = \beta_0 + \sum_{j=1}^{k} \beta_j (x+1)^{\alpha_j}, \quad (10)$$

where α_j is chosen from a standard set of powers, say $\mathcal{S} = \{-2, -1, -0.5, 0., 0.5, 1, 2, 3\}$, with the convention that when $\alpha_j = 0$ $\log(x+1)$ is used. According to Royston and Sauerbrei (2008), $\mathcal{S}$ includes the most commonly used power transformations. They also point out that " (...) $k \leq 2$ *provides enough flexibility for modeling many types of continuous functions we encounter in the health sciences and elsewhere*." We consider fractional polynomials of at most order 2, since higher order polynomials would lead to overfitting. This restriction would not imply the class of fractional polynomial regressions is not wide enough to provide a good fit; if one looks at the 8×8 matrix of all $\alpha_1 \times \alpha_2$ combinations of powers, it can be noticed that the lower (upper) triangular matrix identifies 28 different fractional polynomials of order 2 (with $\alpha_1 \neq \alpha_2$), plus the 8 fractional polynomials of order 1 on the diagonal. A similar approach is discussed by Hwang and Shen (2010); by rewording their proposal using the current notation, we get the nonlinear regression model:

$$\frac{1}{r_x} = \gamma_0 \exp(\gamma_1 x^{\gamma_2}) + \varepsilon_x,$$

which, however, if not properly constrained, may lead to negative estimates at $x = 0$. Considering the linear predictor only and adopting a log scale, the model can be equivalently written as follows:

$$\log r_x = -\log(\gamma_0) - \gamma_1 x^{\gamma_2},$$

which is a particular case of the regression model we are discussing, but with a power not fixed a priori. Rocchetti et al. (2011) describe a regression estimator based on the ratio plot for distributions in the Katz family, proving linearity of the same. Willis and Bunge (2015) generalize this work and consider Kemp-type family of distributions; they adopt a nonlinear regression model for the ratios of successive probabilities, based on ratios of polynomials. The family of Kemp-type distributions is shown to include mixed Poisson distributions, but it is wider and allows to handle departures from the mixed Poisson. Model parameters are estimated by nonlinear least squares.

5. Examples (Continued)

5.1. *Golf Tees*

We recall the fractional polynomial of order 1:

$$\log r_x = \beta_0 + \beta_1 (x+1)^{\alpha} + \epsilon_x. \quad (11)$$

Let us now consider the fitted values obtained by estimating the fractional polynomial above for the golf tees data. As a first step, we evaluated the likelihood obtained by fitting the regression model with varying $\alpha \in \mathcal{S}$, to the complete distribution, considering f_0 as known. The best fit corresponds to the power $\alpha = 0.5$.

The fit with $\alpha = 0.5$ is good with observed $\chi^2 = \sum_{x=0}^{T} \frac{(f_x - \hat{f}_x)^2}{\hat{f}_x} = 9.98$, on $\nu = (T-1) - p = 9 - 1 - 2 = 6$ degrees of freedom. For comparison purposes, we also include the fitted frequencies obtained by the homogeneous binomial model which is entirely unsatisfactory; an improved fit can be found by applying the beta-binomial model. The fit of the beta-binomial is evidently and considerably better ($\chi^2 = 9.11$) than the fit for the binomial, and it is found to be comparable to the fractional polynomial (11) with $\alpha = 0.5$. Our interest is, however, in predicting f_0 when it is unobserved. Hence, as a second step, we choose

Table 1
Golf tees data: untruncated distribution, observed and fitted frequencies. Estimates for binomial, beta-binomial, and regression models based on fractional polynomial of order 1, power $\alpha = 0.5$.

Count	Observed	Binomial	Beta-binomial	$\alpha = 0.5$
0	88	19.51	91.05	86.01
1	46	58.60	40.33	44.26
2	28	77.02	28.47	26.63
3	21	57.85	22.51	19.48
4	13	27.15	18.65	16.82
5	23	8.16	15.79	16.28
6	14	1.53	13.41	16.35
7	6	0.16	11.19	14.95
8	11	0.01	8.61	9.22
χ^2		16,306.15	9.11	9.98

the order of the fractional polynomial by considering the truncated distribution; after the order has been chosen, we predict $\log(\hat{r}_0) = \hat{\beta}_0 + \hat{\beta}_1(x+1)^{\alpha}$ at $x = 0$, leading to $\hat{f}_{0,\text{reg}} = a_0 f_1 \exp(-\hat{\beta}_0 - \hat{\beta}_1)$. The optimal value of α was found to be $\alpha = 1$. The estimates of f_0 using the proposed regression approach, also for $\alpha = 0.5$, the best value for the untruncated distribution, and some competing estimators are reported in Table 2. For the binomial and the beta-binomial models, we have used an EM algorithm, see Web Appendix B. As standard comparative estimators, we provide the Chao lower bound estimator (Chao, 1987) and the Turing estimator (Good, 1953). Details on these estimators are given in Web Appendix B. In all cases, however, the proposed approach clearly outperforms the others.

5.2. *Bowel Cancer Data*

As a first step, we evaluated the likelihood associated to varying $\alpha \in \mathcal{S}$, and used the complete distribution (f_0 observed and known) to estimate model parameters. The best fit corresponds to powers $\alpha = 0.5$ and $\alpha = 0$.

Table 3 reports the observed and the fitted frequencies obtained through the binomial, beta-binomial, and the fractional polynomial models with $\alpha = 0$ and $\alpha = 0.5$. As a second step of the analysis, we have fitted the regression model for varying values of $\alpha \in \mathcal{S}$ to the observed (truncated) distribution, with f_0 unknown, as it would be in real-life cases.

The estimates of f_0 for $\alpha = 0$, $\alpha = 0.5$ (best-fitting powers for the untruncated data) and $\alpha = 1$ (best-fitting power for the truncated data) are provided in Table 4 with $\hat{f}_0 = 11$ for $\alpha = 0.5$ and $\hat{f}_0 = 6$ for $\alpha = 1$. The ratio regression approach provides the best estimate though there seems to be space for improvements. By using the model with $\alpha = 0$, that is

Table 3
Secondary bowel cancer data: untruncated distribution, observed and fitted frequencies. Estimates for binomial, beta-binomial, and regression models based on fractional polynomial of order 1, power $\alpha = 0, 0.5$.

Count	Observed	Binomial	Beta-binomial	$\alpha = 0$	$\alpha = 0.5$
0	22	0.88	18.22	22.37	24.74
1	8	6.75	14.19	8.23	11.89
2	12	21.5	13.37	7.55	8.89
3	16	36.5	13.61	10.55	9.7
4	21	34.85	14.8	17.42	13.95
5	12	17.75	17.84	27.3	22.65
6	31	3.77	29.98	28.6	30.18
χ^2		726.87	8.59	14.96	15.35

$\log r_x = \beta_0 + \beta_1 \log(x+1) + \epsilon_x$, we get $\hat{f}_0 = 21$; it is worth noticing that this is actually supported as the best fractional polynomial model when we consider the untruncated data (see Table 3) and the Chi-square as index of model fit. This power is, however, not the best choice when the truncated distribution is considered. From this discussion, it is clear that the model that best fits the truncated distribution is not necessarily the model that best fits the untruncated distribution, and therefore it may not result in the best estimate of f_0. So, it could be interesting to consider not only the *best* model but rather a range of good models, which could be averaged to get a more reliable estimate for f_0. To this aim, we consider the best three models with respect to the maximized log-likelihood or the Akaike information criterion (AIC) if model complexity varies. For the general case, let $\text{AIC}_{(i)}$ denote the value of the AIC for the best ($i = 0$), the second best ($i = 1$), and the third best ($i = 2$) model, so that $\text{AIC}_{(0)} \leq \text{AIC}_{(1)} \leq \text{AIC}_{(2)}$. The AIC-based weights are as follows:

$$w_i = \exp(\text{AIC}_{(0)} - \text{AIC}_{(i)}),$$

$i = 0, 1, 2$. Burnham and Anderson (2002) discuss AIC-based model averaging in details. In the case of equally parameterized models, the previous expression reduces to the (exponentiated) difference in the maximized log-likelihood values, that is the (log-) ratio of the maximized likelihood value to the likelihood value for the best fitting model. We use these weights to produce a *model-averaged estimate* for f_0:

$$\hat{f}_0^{ma} = \sum_i w_i \hat{f}_{0,i} \Big/ \sum_i w_i,$$

where $\hat{f}_{0,i}$ denotes the estimate obtained according to model i, $i = 0, 1, 2$. The weighted estimate is $f_0^a = 94$ for the golf tees data and $f_0^a = 11$ for the secondary bowel cancer data. These

Table 2
Golf tees data: observed (f_0) and estimated frequency ($\hat{f}_0$) for binomial, beta-binomial, and regression models with first-order fractional polynomial and power $\alpha = 0.5, 1$, Chao and Turing estimators.

Observed	Binomial	Turing	Beta-binomial	Chao	$\alpha = 0.5$	$\alpha = 1$
88	2	10	126	33	93	56

Table 4

Secondary bowel cancer data: observed (f_0) and estimated frequency ($\hat{f}_0$) for binomial, beta-binomial, and regression models with first-order fractional polynomial, power $\alpha = \{0, 0.5, 1\}$, Turing and Chao estimators.

Observed	Binomial	Turing	Beta-binomial	Chao	$\alpha = 0$	$\alpha = 0.5$	$\alpha = 1$
22	0	1	6	2	21	11	6

estimates seem to give corrections in the right direction to the estimates provided by the *best* model for both datasets. This point will be further evaluated in a simulation study available as Supplementary Material.

5.3. *Utrecht Homeless Data*

We now turn to the homeless data for the city of Utrecht. Table 5 reports the fitted frequencies obtained via the homogeneous binomial, the beta-binomial models, and the ratio regression model with $\alpha = 1$ the best fitting first-order fractional polynomial. In this case, f_0 is unknown and, therefore, we are not able to compare the models that best fit the untruncated and the truncated distribution. Further, given that f_0 is unknown, we have no benchmark to compare with.

The estimates of f_0 are provided in Table 6 with $\hat{f}_0 = 66$ for the model with $\alpha = 1$; the ratio regression approach seems to provide a realistic estimate by adjusting the lower bound estimate of Chao $\hat{f}_0 = 55$ to the above. The binomial estimate $\hat{f}_0 = 0$ is clearly too low as is the Turing estimate $\hat{f}_0 = 3$.

The results from the beta-binomial need some comments. Evidently, the beta-binomial model reaches a good fit to the truncated distribution for the homeless data, but the estimate of 881 for f_0 seems unrealistically high. How can this be explained? Again, the ratio regression approach may be of help. Let us consider the beta-binomial distribution

$$p_x = \binom{T}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(x+a)\Gamma(T-x+b)}{\Gamma(T+a+b)}, \qquad (12)$$

Table 5

Utrecht homeless data: observed (truncated) and fitted frequencies: binomial, beta-binomial, and regression models, first-order fractional polynomial, $\alpha = 1$.

Count	Observed	Binomial	Beta-binomial	$\alpha = 1$
1	36	0.00	27.80	16.07
2	11	0.03	15.25	6.84
3	6	0.24	11.06	3.00
4	11	1.19	9.03	1.51
5	5	4.33	7.91	0.92
6	7	11.89	7.28	0.70
7	6	24.83	6.97	0.68
8	11	39.72	6.91	0.85
9	3	48.41	7.12	1.35
10	8	44.24	7.65	2.68
11	7	29.41	8.72	6.46
12	12	13.44	10.93	17.88
13	22	3.78	16.86	50.86
14	77	0.49	78.51	112.21
χ^2		493,376.33	14.42	368.65

leading to the ratio $R_x = a_x \frac{p_{x+1}}{p_x} = \frac{x+a}{T-x-1+b}$, $a_x = (x+1)/(T-x)$, or, equivalently:

$$\log(R_x) = \log(x+a) - \log(T-x-1+b).$$

The beta-binomial requires $a, b > 0$, but the ratio plot with fitted beta-binomial shows that the best parameter estimate for a is negative. The value of $a = -0.36$ does not create difficulties for the range of observed counts $x = 1, \ldots, T$ but leads to an infeasible value for $x = 0$ producing an infinite value estimate for f_0. Restricting the parameter space to $a > 0$ does not avoid this problem since the maximum-likelihood estimate occurs at the boundary. For a thorough discussion about boundary and identifiability-related problems, see Mao and You (2009). The value of 881 has likely occurred at a stage where the computational algorithm has reached a lack-of-progress stopping rule. For some data constellations, potentially with a misleading excellent fit to the zero-truncated data, these boundary problems occur and clearly pose some questions about the beta-binomial as a feasible model. In our perspective, the ratio regression approach may help recognize these situations.

6. Identifiability

The issue of identifiability within the general class of mixtures of zero-truncated binomial distributions has been brought to a general audience by Link (2003). The key argument is best explained by one of his examples. Let us consider the mixed binomial distribution:

$$p_x = \int_0^1 \binom{4}{j} \theta^x (1-\theta)^{4-x}\, h(\theta) d\theta, \qquad (13)$$

$x = 0, \ldots, 4$. Link (2003) considers two choices for $h(\theta)$:

- $h(\theta) \sim U(a, b)$ with $a = 0.026$ and $b = 0.80$
- $h(\theta) \sim 0.576421 \times \delta_{0.286245} + 0.423579 \times \delta_{0.676474}$,

where δ_θ is the one-point distribution putting a unit mass at θ. The untruncated binomial mixtures we can derive by

Table 6

Utrecht homeless data: estimated frequency ($\hat{f}_0$) for binomial, beta-binomial, and regression models with first-order fractional polynomial, power $\alpha = 1$, Turing and Chao estimators.

Observed	Binomial	Turing	Beta-binomial	Chao	$\alpha = 1$
?	0	3	881	55	66

Table 7
Untruncated and truncated count distributions according to model (13)

Mixture model	Probability	Count x				
		0	1	2	3	4
Uniform	p_x	0.227	0.255	0.243	0.190	0.085
	$p_x/(1-p_0)$	-	0.329	0.315	0.246	0.110
Discrete	p_x	0.154	0.279	0.266	0.208	0.093
	$p_x/(1-p_0)$	-	0.329	0.315	0.246	0.110

using (13) are different, but the associated zero-truncated mixtures are identical, see Table 7. Should we have observed only the zero-truncated distribution, it would be impossible to distinguish between the uniform and the two-component distribution when looking for the best fitting model. Hence, unless we refer to a specific form (say, e.g., continuous) for the mixing distribution, it would be impossible to derive a unique estimate for p_0, and consequently, for f_0 and N. This leads to the fact that N (which is a function of p_0) is non-identifiable in the general class of zero-truncated binomial mixture models. A key point in the example of Link (2003) consists in the fact that the class of mixing distributions on the zero-truncated binomial kernel is not identifiable itself as two different mixing distributions may give identical mixture distributions. Working with an identifiable class of discrete zero-truncated mixtures (in the sense that identical mixtures invoke identical mixing distributions) will avoid the problem illuminated by Link (2003) as it cannot happen that identical mixtures lead to different mixing distributions. To make valid inference for p_0, however, it is necessary to assume that the class of mixture models is also valid when we look at the frequency of units with $x = 0$. It is one of the benefits of the ratio regression approach that identifiability for the zero-truncated part is easy to check as we outline below.

It is interesting to view this example from the perspective of the ratio regression approach. Clearly, R_0 and its empirical counterpart, r_0 are only defined for the untruncated count distribution; Figure 1 shows the ratio plot for the two mixtures. The ratios are identical for the zero-truncated and for the untruncated count mixtures at $x = 1, 2, 3$. Obviously, what makes the difference is the observed value for R_0.

The figure makes it clear that it is impossible to say which of the two models is correct if only the untruncated part has been observed. By adopting the argument of Sanathanan (1977), we could say that R_0 cannot be identified by looking at the truncated distribution only. The ratio regression approach cannot change this situation, but it could be interesting to see which way the ratio regression approach would take in this case. Just from the zero-truncated part, it seems reasonable to use a straight line model (the dashed line in Figure 1). The regression parameters are definitely identifiable as long as the design matrix $\mathbf{X}$ is of full rank. The straight line (based only on the zero-truncated distribution) seems to be a reasonable (and intermediate) guess/approximation for the *true* model, either the uniform or the discrete one. It is interesting to see that the quadratic model (the solid line), estimated on the zero-truncated distribution clearly favors the uniform mixture. In this particular situation, however, we may not be interested in using the quadratic model since it is too complex for the situation at hand and it would not allow any goodness-of-fit evaluation since it is a saturated model.

One conclusion from this analysis is that one has to be careful in allowing the size of the class of models under consideration becoming too large; we feel that one way to achieve flexible and well fitting models (keeping identifiability) is via the ratio regression approach. This class can be enriched by looking at suitable classes of functions other than fractional polynomials, even if we suggest to choose simple/robust models, especially for small T.

7. A Historic Example: Shakespeare Word Frequency Data

In a historic article, Efron and Thisted (1976) used the work of Shakespeare to illustrate the number of species problem. According to Spevack (1968), Shakespeare's known works comprise 884,647 total words, of which 14,376 are types appearing just once, 4343 are types appearing twice, etc. In our notation, X_i denotes the number of times the i-th word appears in Shakespeare's work, so that f_x is the number of words appearing exactly x times. In this situation, it seems reasonable to work with a Poisson mixture:

$$p_x = \int_0^\infty \exp(-\theta)\theta^x/x!\; h(\theta)d\theta, \tag{14}$$

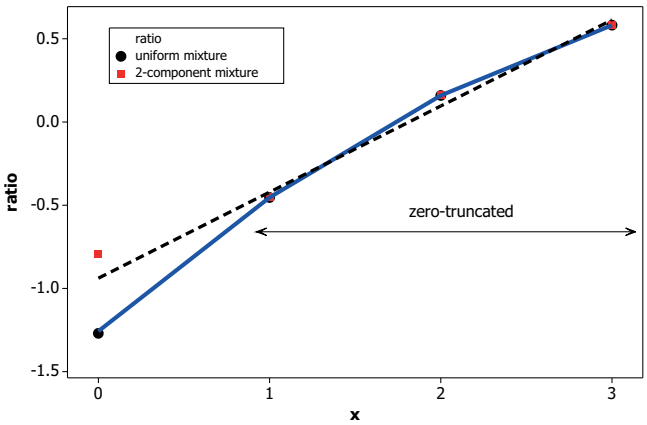

Figure 1. Ratio plot for the untruncated and the zero-truncated uniform (bullet) and discrete (square) count mixture (identical for the zero-truncated part). Straight line (dashed) and parabola (solid) based on the zero-truncated ratios.

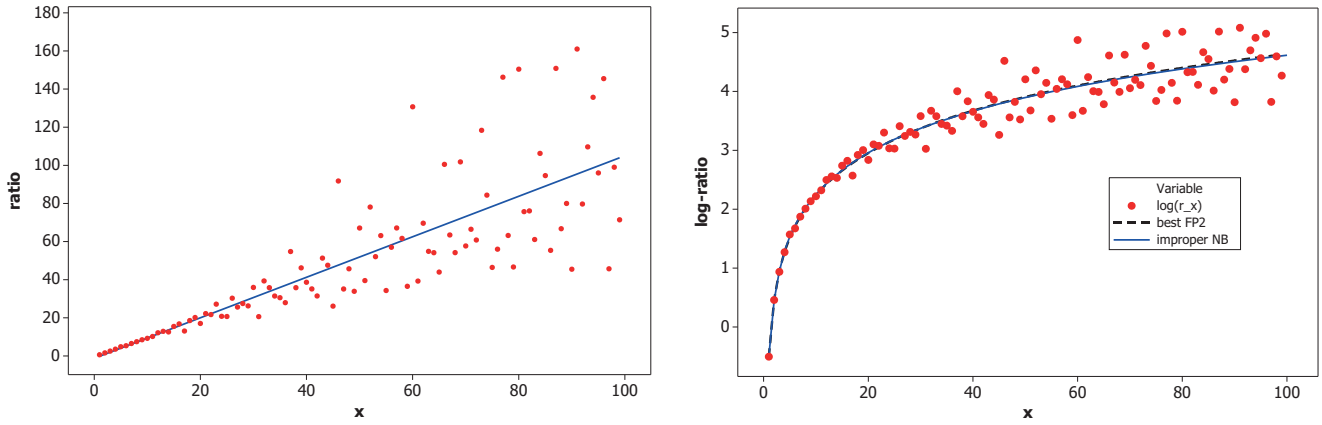

Figure 2. Word frequency count of Shakespeare's works: ratio plot r_x against x, (left panel), improper negative binomial distribution (solid), and best fractional polynomial of order 2 (dashed), (right panel).

where $h(\theta)$ is some mixing density coping with potential heterogeneity in the Poisson rate parameter. Note that in this case, we set $a_x = x + 1$ to include the Poisson as the base distribution, which corresponds to a constant (log) ratio R_x over the range of x. Using a gamma density for $h(\theta)$, we obtain a marginal negative binomial distribution:

$$p_x = \frac{\Gamma(x+k)}{\Gamma(x+1)\Gamma(k)}(1-p)^x p^k,$$

with event parameter $p \in (0, 1)$, and shape parameter $k > 0$. The negative binomial is one of the models discussed by Efron and Thisted (1976). In this case, $a_x = (x + 1)$ and

$$R_x = (1 - p)(x + k).$$

It defines a straight line in x with intercept term $\beta_0 = k(1 - p)$ and slope $\beta_1 = (1 - p)$. This model is seemingly supported when looking at the ratio plot of $r_x = (x + 1)f_{x+1}/f_x$ against x; in fact, it seems to give some evidence of a straight line pattern (see Figure 2, left panel).

The corresponding nonlinear model $\log(r_x) = \beta_0 + \beta_1 \log(x + k)$ experiences a negative estimate $\hat{k} = -0.3890$, which is indeed very close to the value given in Efron and Thisted (1976), equal to -0.3954. Although the fit is excellent, as Figure 2 (right panel) shows, the negative binomial distribution becomes *improper* since $\hat{r}_0$ is negative, a useless model for predicting f_0. Implementing a boundary condition $k > 0$ diminishes the fit considerably. Alternatively, we can consider the ratio regression approach; the best second-order fractional polynomial model is provided here by the following specification:

$$\log r_x = \beta_0 + \beta_1(x + 1)^{-2} + \beta_2 \log(x + 1). \qquad (15)$$

The corresponding fit is illustrated in Figure 2, right panel, with virtually no visible difference to the fit of the *improper* negative binomial. The benefit of the ratio regression approach is that a valid count distribution can be derived from model (15) via the result of Theorem 1. The result is that, when the conditions of Theorem 1 are valid, an estimate for f_0 can be easily derived, and this may help solve the boundary problems.

8. Discussion

In this article, starting from the work of Rocchetti et al. (2011, 2014), we introduce a regression approach to estimate the unknown size of a potentially elusive population. The approach is based on modeling ratios of successive probabilities, and can be readily applied to arbitrary mixtures of count distributions. The idea of using a regression approach to develop estimators for the population size may be fruitfully linked to the ratio plot developed by Böhning et al. (2013) as a diagnostic tool for homogeneity. A regression approach to estimate the size of a population has been also investigated by Hwang and Shen (2010), Rivest and Baillargeon (2014), and Willis and Bunge (2015). The empirical behavior of the regression estimator has been investigated in the context of the Katz family of distributions by Rocchetti et al. (2011) and for beta-binomial distributions by Rocchetti et al. (2014). However, all these proposals still lack a general perspective and do not discuss the conditions for the regression model to lead to a proper counting distribution. In the present article, we proved that, under simple conditions, any regression model for the ratios provides a feasible count distribution. This means that a regression model may lead to a proper marginal distribution, but the latter does not necessarily correspond to any known or standard-mixed count distribution. This is a relevant finding of the proposed approach. Furthermore, the approach is based on finding the most appropriate regression model with respect to fitting the available (truncated) data, considering a wide range of fractional polynomial functions. These functions are often flexible enough to cope with various and general forms of nonlinearity. We have shown how a modified Horvitz–Thompson estimator can be defined, where the probability of missing a unit is estimated through the proposed regression model.

As it is well known, a major problem with mixed binomial distributions is that we can not identify the mixing distribution if no limitations on the class of mixing distributions are introduced. In this context, identifiability still is a concern, but the effort is moved to choosing the model of a given order

that best fits the observed, truncated, distribution, and this is essentially unique. Another aspect of the proposed approach is the introduction of model-averaged estimators. This idea seems to mitigate the potential bias of the *best* AIC estimator which may overfit the truncated part of the distribution. By using an AIC-weighted estimator we also have the chance, within a specified class of models, to recover potential problems related to a single model, as shown in the simulation study.

9. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2.1–2.3 and 5.1 are available with the article at the *Biometrics* website on Wiley Online Library. The web-based supplementary materials also include a simulation study to analyze the proposed model behavior under various conditions.

Acknowledgements

We thank the Editor, an Associate Editor, and two referees for helpful comments on a previous version of the manuscript. Marco Alf'øacknowledges the financial support from the grant RBFR12SHVV of the Italian Government (FIRB project Mixture and latent variable models for causal inference and analysis of socio-economic data).

References

Böhning, D. and Kuhnert, R. (2006). The equivalence of truncated count mixture distributions and mixtures of truncated count distributions. *Biometrics* **62**, 1207–1215.

Böhning, D., 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.

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

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

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information–Theoretic Approach*. New York: Springer.

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

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

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.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63**, 435–447.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

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.

Van der Heijden, P. G. M., Cruyff, M., and Böhning, D. (2014). *Analyses daklozen Utrecht 2013.* Universiteit Utrecht en University of Southampton. Utrecht, 28 januari 2014.

Hoaglin, D. C. (1980). A Poissonness plot. *The American Statistician*, **34**, 146–149.

Hwang, W.-H. and Shen, T.-J. (2010). Small-sample estimation of species richness applied to forest communities. *Biometrics* **66**, 1052–1060.

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

Lloyd, C. J. and Frommer, D. (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.

Lloyd, C. J. and Frommer, D. (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.

Lloyd, C. J. and Frommer, D. (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.

Mao, C. X. and You, N. (2009). On comparison of mixture models for closed population capture–recapture studies. *Biometrics*, **65**, 547–553.

Norris, J. L. and Pollock, K. H. (1996). Nonparametric MLE under two closed capture–recapture models with heterogeneity. *Biometrics* **52**, 639–649.

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

Rivest, L.-P. and Baillargeon, S. (2014). Capture–recapture methods for estimating the size of a population: Dealing with variable capture probabilities. In *Statistics in Action: A Canadian Outlook*, J. F. Lawless (ed.). Boca Raton (FL): Chapman and Hall/CRC.

Rocchetti, I., Bunge, J., and Böhning, D. (2011). Population size estimation based upon ratios of recapture probabilities. *Annals of Applied Statistics*, **5**, 1512–1533.

Rocchetti, I., Alfó, M., and Böhning, D. (2014). A regression estimator for mixed binomial capture–recapture data. *Journal of Statistical Planning and Inference* **145**, 165–178.

Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-Building. A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables.* Hoboken: Wiley.

Sanathanan, L. (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association* **72**, 669–672.

Spevack, M. (1968). *A Complete and Systematic Concordance to the Works of Shakespeare*. Vols. 1–6. Hildesheim: George Olms.

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.

Willis, A. and Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics* **71**, 1042–1049.

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

Received April 2015. *Revised December* 2015.
Accepted December 2015.