# Regression estimators for capure-recapture frequency data

Dankmar Böhning, Marco Alfò, Irene Rochetti

**Abstract** Mixed binomial models are often used to provide estimates for the unknown size of a partially observed population when the number of identification (sampling) sources is finite and known. By using a simple recursive relation, we develop a regression based estimator which is always well defined and does not suffer from weak identifiability, which is a characteristic of ML estimator.

**Key words:** Capture recapture, count data, ratio plot, regression model.

## 1 Introduction

Capture-recapture methods are used to estimate the unknown size $N$ of a *partially observed* population, based on samples observed through one or more identification mechanisms. Established in the wildlife setting, these methods have been extended to epidemiology, see eg Chao (1989), and repeated diagnostic testing, see eg Böhning and Patilea (2008). We focus on those empirical cases where the number of sampling occasions, $m$, is known and fixed, $n$ is the number of units identified by the mechanism at least once, where $n = n_1 + n_2 + \cdots + n_m$, and $n_x$, $x = 1,\ldots,m$ represent the number of units identified exactly $x$ times. If we denote by $p_x$, $x = 1,\ldots,m$ the probability of exactly $x$ identifications, the ML estimator of $N$ is known to be the integer part of the Horvitz-Thompson estimator

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

Dankmar Böhning
Southampton Statistical Sciences Research Institute, e-mail: D.A.Bohning@soton.ac.uk

Marco Alfò
Dip.to di Scienze Statisiche, Sapienza Univ. di Roma, e-mail: marco.alfo@uniroma1.it

Irene Rocchetti
Dip. Censimenti ed Archivi Amministrativi e Statistici, ISTAT e-mail: irocchetti@istat.it

where $p_0$ represents the probability that an individual is not registered. To obtain an estimate $\hat{N}$, we can proceed by estimating $p_0$ or, directly, $n_0$ given that $N = n_0 + n$.

## 2 The ratio plot

According to the hypothesis of a fixed, known, number of sampling occasions $m$, the number of times the $i$-th individual is identified can be modeled by a Binomial distribution with probability $\pi$ and index $m$:

$$p_x = \Pr(X = x) = \binom{m}{x} \pi^x (1-\pi)^{m-x}$$

where $p_0 = (1-\pi)^m$; in this case, the Horvitz-Thompson estimator is given by $\hat{N} = \frac{n}{1-(1-\hat{\pi})^m}$.

If we look at ratios of probabilities corresponding to successive counts, we may observe that the following property holds:

$$\frac{p_{x+1}}{p_x} = \frac{\binom{m}{x+1}\pi^{x+1}(1-\pi)^{m-x-1}}{\binom{m}{x}\pi^x(1-\pi)^{m-x}} = \frac{m-x}{x+1}\frac{\pi}{1-\pi} = \frac{1}{\alpha_x}\frac{\pi}{1-\pi} \rightarrow \alpha_x \frac{p_{x+1}}{p_x} = \frac{\pi}{1-\pi}$$

$x = 0, \ldots, m-1$. Using an empirical Bayes procedure, this ratio can be estimated by

$$\alpha_x \frac{n_{x+1}/N}{n_x/N} = \alpha_x \frac{n_{x+1}}{n_x}$$

Plotting $\alpha_x \frac{n_{x+1}}{n_x}$ versus $x$ defines a graphical tool to measure the potential departure of the observed frequency counts from a homogeneous Binomial data generating process. This plot, called *the ratio plot* has been introduced by Hoaglin (1980) in the Poisson case, for a review see Böhning et al. (2011).

The need for such a graphical device stems from the empirical evidence that, in real applications, the homogeneous Binomial may not be appropriate because it does not account for individual- or group-specific heterogeneity. Heterogeneity may be observed (and stored in a covariate vector) and/or (partially) unobserved; in the latter case, individual-specific variability in detection probabilities is modeled through a mixing distribution on the Binomial parameter and mixed binomial models are used to provide estimates for the unknown size of the population of interest. A major problem is, in this context, the lack of identifiability of the mixing distribution which may lead to inconsistent estimates of the population size, see e.g. Sanathanan (1977) and Link (2003). To explain, let us suppose to observe the realizations of a truncated mixed binomial distribution $P_Q(x)$ describing the number of times that an individual is observed, with *true* mixing distribution $Q$. It can be proved that different mixing distributions, say $Q_1 \neq Q_2$ may lead to identical *truncated* marginal distributions,

i.e. $P_{Q_1}(x)$=$P_{Q_2}(x)$, $\forall x = 1,\ldots,m$. Therefore, $P_Q$ is not identifiable and the same argument applies to the probability an individual is unseen.

## 3 Regression estimator

The ratio plot could helps detect substantial departures from the homogeneous binomial model, but its scope could be wider, due to the following property of *general* mixed binomial models. A mixed binomial distribution with mixing $Q(\pi)$ is

$$p_x = \int_0^1 \binom{n}{x} \pi^x (1-\pi)^{n-x} Q(\pi) d\pi$$

Following the arguments given in Chao (1989) for $m = 2$, it is possible to prove that, whatever the mixing distribution, the following propert holds:

$$\alpha_0 \frac{p_1}{p_0} \leq \alpha_1 \frac{p_2}{p_1} \leq \alpha_2 \frac{p_3}{p_2} \leq \cdots \leq \alpha_{m-1} \frac{p_m}{p_{m-1}}$$

That is, the ratio plot is nondecreasing in $x$ with equality holding only when homogeneity (ie standard binomial model) holds, see Böhning et al. (2011). Thus, we may consider a regression model, a quasi-generalized linear model, where the function $g(\cdot)$ links the estimated probability ratios to the number of identifications, $x$:

$$g\left(\alpha_x \frac{p_{x+1}}{p_x}\right) = \beta' \mathbf{z}(x)$$

Here $\mathbf{z}$ represents a vector containing a unit term together with one or more trasfomations of $x$. Thus, every mixed binomial distribution could be represented by a regression model via the ratio plot; a simple binomial distribution would lead to a model with only the intercept term ($\alpha_x \frac{p_{x+1}}{p_x} = \frac{p}{1-p} = \beta_0$), while in the Beta Binomial case we would obtain:

$$\alpha_x \frac{p_{x+1}}{p_x} = \frac{x+\alpha}{m-x-1+\beta}$$

The question is whether every regression model may also lead to a proper count distribution. It can be proved that given $g(r_x) = \beta' \mathbf{z}(x)/\alpha_x > 0$, $x = 0,1,\ldots,m-1$ there exists a unique distribution $p_x$, $x = 0,\ldots,m$ satisfying the following properties

- $p_{x+1} = r_x p_x \qquad x = 0,\ldots,m-1$
- $\sum_{x=1}^m p_x = 1 \rightarrow p_0 = \frac{1}{1+r_0+r_0 r_1+\cdots+\prod_{x=0}^{m-1} r_x}$

Thus, we may start by looking at observed frequencies of identification, define the probability ratio estimates and the empirical model :

$$g\left(\alpha_x \frac{n_{x+1}}{n_x}\right) = \beta' \mathbf{z}(x)$$

where obseved ratios may have a non-diagonal covariance matrix $\Sigma$. Fitting the model trough weighted least squares and finding the fitted values, we get

$$\hat{r}_x = g^{-1}\left[\hat{\beta}'\mathbf{z}(x)\right]/\alpha_x$$

By exploiting the *recursive* property linking $p_0$ to probability ratios, we achieve finally the following estimates for $p_0$ and $N$:

$$\hat{p}_0 = \frac{1}{1+\hat{r}_0+\hat{r}_0\hat{r}_1+\cdots+\prod_{x=0}^{m-1}\hat{r}_x}$$
$$\hat{N}_0 = \frac{n}{1-\hat{p}_0} = n\frac{1+\hat{r}_0+\hat{r}_0\hat{r}_1+\cdots+\prod_{x=0}^{m-1}\hat{r}_x}{\hat{r}_0+\hat{r}_0\hat{r}_1+\cdots+\prod_{x=0}^{m-1}\hat{r}_x}$$

We get an estimate for $p_0$, resp. $N$, derived from the observed frequency distribution only, where we do not need to choose the *true* (if any) mixing distribution, but rather the function of $x$ that best approximates the observed count distribution. The main result of a simulation study performed by considering the Sydney screening data, see Lloyd and Frommer (2008), is that the *final* estimate of the population size is quite robust wrt the choice of the function of $x$ used in the linear predictor.

## References


BÖHNING, D. AND BAKSH, M. F. AND LERDSUWANSRI, R. AND GALLAGHER, J. (2011). Use of the ratio plot in capture-recapture estimation. *Journal of of Computational and Graphical Statistics*, to appear.

BÖHNING, D. AND PATILEA, V. (2008). A capture-recapture approach for screening using two diagnostic tests with availability of disease status for the test-positives only. *Journal of the Amer. Stat. Ass.* **103**, 212–221.

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

CHAO, A., TSAY, P. K., LIN, S.-H., SHAU, W.-Y. AND CHAO, D.-Y. (2001). Tutorial in Biostatistics. The applications of capture–recapture models to epidemiological data. *Statistics in Medicine* **20**, 3123–3157.

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

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

LLOYD, C.J., FROMMER, D.J. (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.

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