School of Mathematical
and Physical Sciences

Department of Mathematics and Statistics

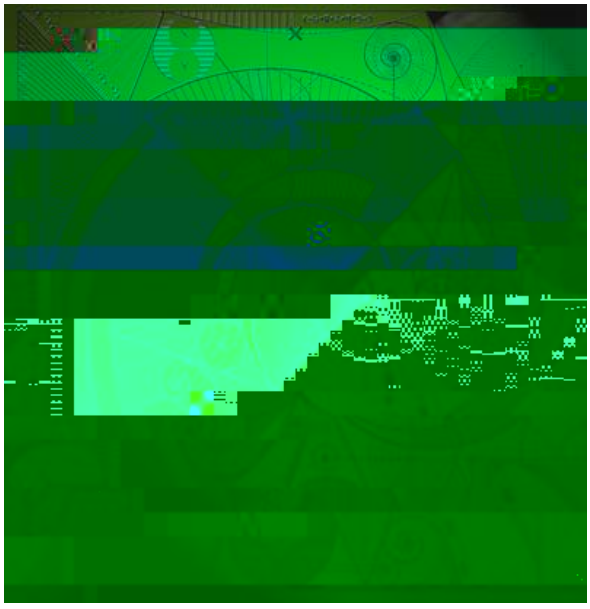
Preprint MPS_2010-25

12 June 2010

Population Size Estimation Based upon Ratios of Recapture Probabilities

by

Irene Rocchetti, John Bunge and
Dankmar Böhning



Population Size Estimation Based upon Ratios of Recapture Probabilities

Irene Rocchetti

Department of Demography,

Sapienza University of Rome, Rome, Italy

John Bunge

Department of Statistical Science, Cornell University,

Ithaca (NY), USA

Dankmar Böhning

Applied Statistics, School of Biological Sciences,

University of Reading, Reading, UK

June 12, 2010

Abstract

Estimating the size of an elusive target population is of prominent interest in many areas in the life and social sciences. Our aim is to provide an accurate and workable method to estimate the unknown population size, given the frequency distribution of counts of repeated identifications of units of the population of interest. This counting variable is necessarily zero-truncated, since units that have never been identified are not in the sample. We consider several applications: clinical medicine, where interest is in estimating patients with adenomatous polyps which have been overlooked by the diagnostic procedure; drug user studies, where interest is in estimating the number of hidden methamphetamine users; veterinary surveillance of sheep in Great Britain, where interest is in estimating the hidden amount of scrapie; and entomology and microbial ecology, where interest is in estimating the number of unobserved species of organisms. In all these examples, simple models such as the homogenous Poisson are not appropriate since they do not account for present and latent heterogeneity. The Poisson-gamma (negative binomial) model provides a flexible alternative and often leads to well-fitting models. It has a long history and was recently used in the development of the Chao-Bunge estimator. Here we use a different property of the Poisson-gamma model: ratios of neighboring Poisson-gamma probabilities are linearly related to the counts of repeated iden-

components. In criminology the number of people with illegal behavior is of high interest (Van der Heijden, Cruy□, and Houwelingen 2003), and in ecology we wish to estimate the number of rare species of organisms (Chao *et al.* 2001). All of these situations fall under the following setting. We assume that there are N units in the population, which is closed (no birth, death or migration), and that there is an endogenous mechanism such as a register, a diagnostic device, a set of reviewers, or a trapping system, which identifies n distinct units from the population. A given unit may be identified exactly once, or it may be observed twice, three times, or more. We denote the number of units observed i times by f_i, so that $n = f_1 + f_2 + f_3 + \ldots$; the number of unobserved or missing units is f_0, so $N = f_0 + n$

Table 1: *Methamphetamine data — frequency distribution of treatment episodes per drug user*

f_1	f_2	f

Table 3: *Scrapie data — frequency distribution of the scrapie count within each holding for Great Britain in 2005*

Table 5: *Protistan diversity in the Gotland Deep — frequency counts of observed species*

f_1	f_2	f_3	f_4	f_6	f_8	f_9	f_{10}	f_{11}	
48	9	6	2	2	2	1	2	1	
f_{12}	f_{13}	f_{16}	f_{17}	f_{18}	f_{20}	f_{29}	f_{42}	f_{53}	n
1	1	2	1	1	1	1	1	1	84

in Table 5 stem from a recent work by Stock *et al.* (2009). Microbial ecologists are interested in estimating the number of species N in particular environments. Unlike butterflies, microbial species membership is not clear from visual inspection, so individuals are defined to be members of the same species (or more general taxonomic group) if their DNA sequences (derived from a certain gene) are identical up to some given percentage, 95% in this case. Here the study concerned protistan diversity in the Gotland Deep, a basin in the central Baltic Sea. The sample was collected in May 2005. The maximum observed frequency was 53.

The classical approach to estimation of N is to assume that each population unit enters the sample independently with probability p (dealing with heterogeneous capture probabilities by modeling and averaging). Given p, the unbiased Horvitz-Thompson estimator of N is n/p, and the maximum likelihood estimator is its integer part $\lfloor n/p \rfloor$. One then estimates p using any of several methods, and the final estimate of N is $n/\hat{p}$ or $\lfloor n/\hat{p} \rfloor$ (Lindsay and Roeder 1987, Böhning *et al.* 2005, Böhning and van der Heijden 2009, Wilson and Collins 1992, Bunge and Barger 2008, Chao 1987, 1989, Zelterman 1988).

Here we take a new approach: we consider *ratios of successive frequency counts*,

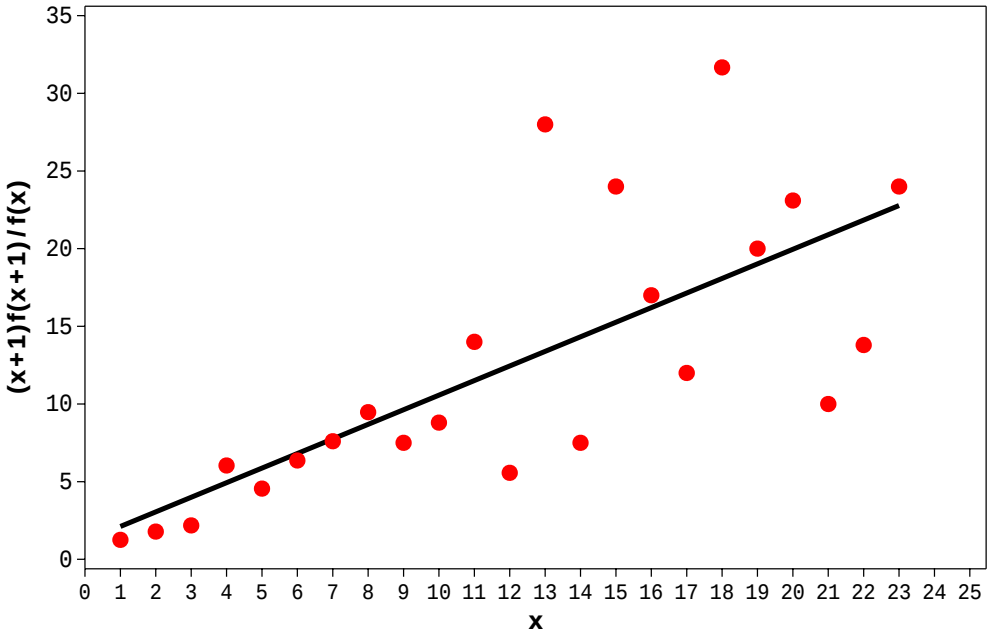


Figure 2: *Scatterplot with regression line of* $(x+1)f_{(x+1)}/f_x$ *vs.* x *for the butterfly data*

This simple and powerful method applies exactly when the frequency counts emanate from the Katz family of distributions, namely the binomial, Poisson, and gamma-mixed Poisson or negative binomial, and it applies approximately to extensions of the Katz family and to general Poisson mixtures. It can be implemented using any statistical software package that performs weighted least squares regression, and it is superior to existing methods for the negative binomial model (including maximum likelihood) in several ways. In addition, it substantially mitigates the effect of truncating large counts (recaptures or replicates), which is an issue with almost every existing method, parametric or nonparametric. In section 2 we discuss the method and its scope of applicability; in section 3 we describe weighting schemes; in

section 4 we look at goodness of fit of the linear model; and in section 5 we compare our method with existing techniques, analyze the five datasets, and discuss the implications of our findings. An appendix covers aspects of the approximation used for reaching the linear model as well as a comparative simulation study, a discussion of standard error approximations, and an assessment of the effect of deleting large "outlying" frequencies.

2 Linear regression and the Katz distributions

Let $p_0, p_1, p_2, \ldots$ denote a probability distribution on the non-negative integers. The condition

$$r(x) := (x + 1)p_{x+1}$$

lower bound for the population size, since the homogeneity assumption leads to downwardly biased estimation in the presence of heterogeneity. In this case the frequency count data $f_1, f_2, \ldots$ summarizes the nonzero values of $m_1, \ldots, m_N$.

- Now suppose that population unit i appears a random number of times m_i in the sample, but now $m_1, \ldots, m_N$ are i.i.d. Poisson random variables with (homogeneous) mean λ. This model arises naturally in *species abundance sampling* where each species contributes some number of representatives to the sample; it also appears as an approximation to the binomial model with $\lambda \approx kq$, for large k and small q

for estimating N in a variety of situations.

We make two further comments on distribution theory. First, it may be readily shown using the Cauchy-Schwartz inequality that the ratio on the left-hand side of (1) is non-decreasing for *any* mixed-Poisson distribution. This means that the linear relation, and hence our weighted linear regression procedure below, can be regarded as a first-order linear approximation for any Poisson mixture (not just gamma), thus justifying a degree of robustness of our method across a wide range of heterogeneity models. Second, there are extended versions of relation (1) which give rise to distributional extensions of the Katz family that need not be mixed-Poisson (Johnson *et al.*, 2005). Such extensions may be parameterized and we conjecture that our method below will be robust to small perturbations along these parameters.

Condition (1) suggests linear regression of the left-hand side upon the right, in some form. Observe that the natural estimate of p_x would be $\hat{p}_x(N) := f_x/N$, if N were known. But

(/N, if N e9939Tf18.-1tfl%h6F1817.965255 2T]TJF1811.95526004f5.77-1.051 :=

and we fit the model

$$\log\left[\frac{(x+1)f_{x+1}}{f_x}\right] = \log \hat{r}(x) = \gamma + \delta x + \varepsilon_x. \qquad (2)$$

We consider this in terms of linear regression in the next section. To obtain a simple

so that we have $\log(x + k) + \log(1 - p) \approx \log(1 - p) + \log(k) + x/k$. Note that this approximation is exact for $x = 0$ (the point where we predict) and good for $x = 1$ (corresponding to the informative "singleton" frequency count). In the Appendix we discuss this approximation further, as well as alternatives. With reference to model (2) we have $\gamma = \log(1 - p) + \log(k)$ and $\delta = 1/k$. We focus on this model in the discussion below.

Note also that due to the simple structure of the estimator $\hat{f}_0 = f$

3 Heteroscedasticity and weighted least squares

nomial with cell probabilities $\pi = (\pi_1, \ldots, \pi_m)^T$. Then it is well-known that $\mathbf{f} = (f_1, \ldots, f_m)^T$ has covariance matrix $\Sigma = n[\Lambda(\pi) - \pi\pi^T]$, where $\Lambda(\pi)$ is a diagonal matrix with elements π on the diagonal, and $n = f_1 + \ldots + f_m$. Writing

$$\Sigma = n[\Lambda(\pi) - \pi\pi^T] = \Lambda(n\pi) - \frac{1}{n} n\pi \; n\pi^T,$$

we see that Σ can be estimated as

$$\hat{\Sigma} = \Lambda(\mathbf{f}) - \frac{1}{n} \mathbf{f} \; \mathbf{f}^T.$$

An application of the multivariate delta-method then shows that an estimate of $\mathrm{cov}(Y)$ is

$$\mathbf{J}_{\mathbf{f}}(Y(\mathbf{f}))\, \hat{\Sigma} \left[\mathbf{J}_{\mathbf{f}}(Y(\mathbf{f})) \right]^T = \begin{bmatrix} \frac{1}{f_1} + \frac{1}{f_2} & \frac{-1}{f_2} & 0 & \ldots & 0 & \ldots & 0 \\ \frac{-1}{f_2} & \frac{1}{f_2} + \frac{1}{f_3} & \frac{-1}{f_3} & 0 & & \ldots & 0 \\ 0 & & \ddots & & & & \\ \vdots & & & \ddots & & & \\ 0 & \ldots & 0 & \frac{-1}{f_i} & 1 & & \end{bmatrix}$$

terms in $\mathrm{cov}(Y)$) with little loss of precision for our purposes. This corresponds to our intuition that covariances between adjacent log-ratios may not play a large role in reducing MSE. Let

$$\Lambda(\mathbf{f}) = \begin{bmatrix} \frac{1}{f_1} + \frac{1}{f_2} & 0 & 0 & \dots & 0 & 0 \\ 0 & \frac{1}{f_2} + \frac{1}{f_3} & 0 & \dots & 0 & 0 \\ \vdots & & \ddots & & & \\ \vdots & & & \ddots & & \\ 0 & 0 & 0 & \frac{1}{f_i} + \frac{1}{f_{i+1}} & 0 & 0 \\ \vdots & & & & \ddots & \\ \dots & & & & 0 & \frac{1}{f_{m-1}} + \frac{1}{f_m} \end{bmatrix} \tag{5}$$

be the diagonal part of (4); we then suggest using (5) in our weighted regression model. This is computationally simpler, especially when dealing with a high number of recaptures. A small simulation study confirms the precision of this simplification, at least within the domain of the simulation. We computed the bias of $\hat{N}$ using the weighted regression model under three scenarios: with weights according to (4), according to (5) and according to $\mathbf{W} = I_m$ (the m-dimensional identity matrix, i.e., unweighted). Frequency data were drawn from a negative binomial distribution with parameters $p = 0.8$ and $k = 7$, and replicated $1,000$ times. Table 6 shows results for $N = 100$ and $N = 1,000$. It is clear that weighting is important in fitting the model: the unweighted regression model leads to potentially heavily biased estimators of the population size, whereas the effect of ignoring the covariance between $\log(xf_x/f_{x-1})$ and $\log((x+1)f_{x+1}/f_x)$ is negligible. Finally we note that weighted least squares can introduce numerical problems, especially in sparse-data situations (Björck, 1996,

Table 6: *The effect of different weight matrices according to (4), (5) and* $W = I_m$ *for frequency data from the Negative Binomial distribution with parameters* $k = 7$, $p = 0.8$

	Bias of $\hat{N}$		
N	(4)	(5)	unweighted
100	3.05	3.40	8.81
1,000	2.70	0.36	45.86
	Standard error of		

is the "truncation point" or maximum frequency used in the analysis (we return to this issue below). We make the further approximation $\exp(\hat{\gamma} + \hat{\delta}x) \approx (x+1)\hat{f}_{x+1}/\hat{f}_x$, leading to the recursive relation $\hat{f}_{x+1} = \hat{f}_x \exp(\hat{\gamma} + \hat{\delta}x)/(x+1)$, $x = 1, 2, \ldots, m-1$. Since $\hat{f}_0$ is given, this defines the sequence $\{\hat{f}_x, x = 0, 1, \ldots, m\}$. We then define our χ^2 statistic as

$$\chi^2 = \sum_{x=1}^{m} \frac{(f_x - \hat{f}_x)^2}{\hat{f}_x}$$

and simulations support that this has a χ^2 distribution with $m-2$ degrees of freedom if the regression model holds. Note that we have m unconstrained frequencies, since $n = \sum_{x=1}^{m} f_x$ is random, and we lose 2 degrees of freedom due to estimating the intercept and slope parameters. Note also that the estimate of the intercept parameter fixes $\hat{f}_1 = f_1$, so that the degrees of freedom are indeed only reduced by 2. This approach has the benefit of gaining one degree of freedom when compared to a goodness-of-fit measure based solely on the regression model which works with the $m-1$ values $\hat{y}_x$, $x = 1, \ldots, m-1$.

estimate *ex post facto*. Bunge and Barger (2008) propose a goodness-of-fit criterion for selecting m, while the coverage-based nonparametric methods of Chao and co-authors fix m heuristically at 10 (see Chao and Bunge, 2002). Our weighted linear regression approach also has the potential for loss of fit as m increases, depending on the realized structure of the data, and again we can fix m and collapse all frequencies greater than this threshold to one value. Sensitivity of the various methods to the choice of m is a complex topic (Bunge and Barger (2008) compute all estimates at all possible values of m); however, our data analyses below show that the the weighted linear regression model is considerably less sensitive to m than its chief competitors in the negative binomial case, namely ML and the Chao-Bunge estimator.

Finally we note that in the ML approach, if the negative binomial fit is less than ideal (although perhaps still acceptable), numerical maximum likelihood algorithms often do not converge, or converge to the edges of the parameter space, which in turn distorts the apparent fit. The regression-based method described here o□ers a more robust approach to parameter estimation, and appears not to be prone to the numerical problems which arise for maximum likelihood estimation under the negative binomial model. In fact, the negative binomial parameter estimates $(\hat{p}, \hat{k})$ derived from the regression model could be used as starting values for a numerical search for the ML estimates. This is a topic for further research.

5 Alternative estimators, data analyses, and discussion

5.1 Alternative estimators

We first consider certain other options for the negative binomial model.

- *Maximum likelihood.* This approach is well-studied and has a long history (see Bunge and Barger (2008)), but as noted above, good numerical solutions for the model parameters (p, k) seem to be remarkably difficult to obtain, even using reasonably sophisticated search algorithms with high-precision settings. In our experience we get good numerical convergence only when the frequency data is smooth and fits the negative binomial well, or when the right-hand tail is fairly severely truncated. The latter issue causes the additional computational burden of investigating many truncation points, each involving numerical optimization. Nonetheless we can obtain ML results for the negative binomial in some cases. The ML estimator $\hat{N}_{ML}$ is consistent for N given that the model is correct.

- *Chao-Bunge.* Let τ denote the probability of observing a unit at least twice, i.e., $\tau = 1 - p_0 - p_1$. Chao and Bunge (2002) developed a nonparametric estimator $\hat{\tau}$ for τ, and on this basis proposed the estimator

$$\hat{N}_{CB} := \sum_{j=2}^{m} \frac{f_j}{\hat{\tau}}$$

  for N. They showed that $\hat{N}_{CB}$ is consistent for N under the negative binomial model. However, in applied data analysis $\hat{\tau}$ may be very small or even negative,

leading to very large or negative values of $\hat{N}_{CB}$. This is one reason why Chao and Bunge set $m = 10$ (as noted above). In fact $\hat{N}_{CB}$ fails roughly as often as $\hat{N}_{ML}$, although not necessarily in the same situations.

- Chao (1987, 1989) proposed the nonparametric statistic

$$\hat{N}_{Ch} = n + \frac{f_1^2}{2f_2},$$

which is valid as a (nonparametricmetri8.0ricmetri8.0rich is v2

Table 7: *Data analyses.* $\hat{N}$ *= weighted linear regression model;* $\hat{N}_{ML}$ *= negative binomial maximum likelihood estimate;* $\hat{N}_{CB}$ *= Chao-Bunge estimator;* $\hat{N}_{Ch}$ *= Chao lower bound; SE = standard error; p = p-value from* χ^2 *goodness-of-fit test; * = estimation failed.*

study	$\hat{N}$	SE	*p*	$\hat{N}_{ML}$	SE	*p*	$\hat{N}_{CB}$	SE	$\hat{N}_{Ch}$
Meth.	61,133	17,088.8	0.000	*	*	*	*	*	33,090
Polyps – low	495	37.15	0.340	892	342.3	0.619	668	141.4	458
Polyps – high	513	52.0	0.001	587	77.2	0.010	584	72.0	511

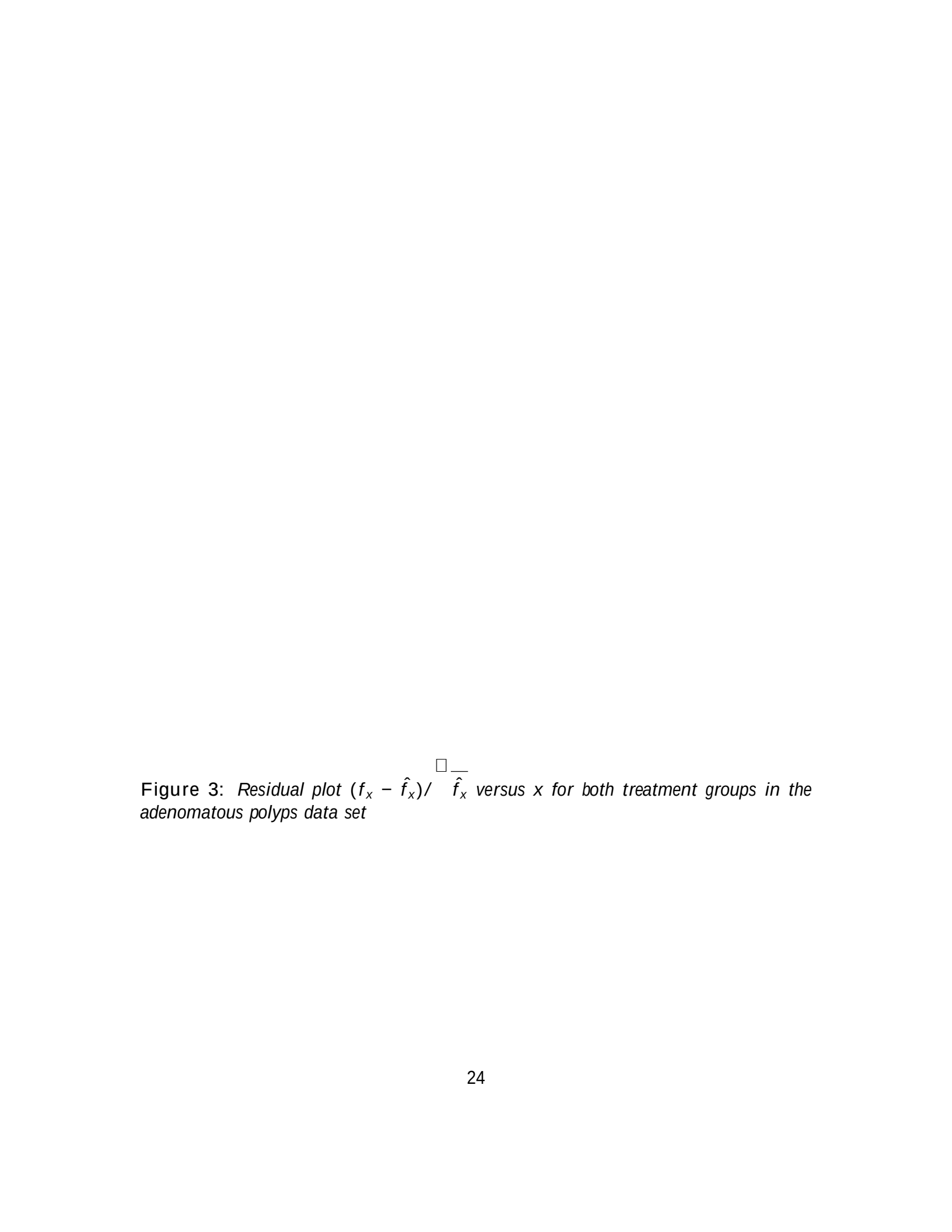


Figure 3: *Residual plot* $(f_x - \hat{f}_x)/\sqrt{\hat{f}_x}$ *versus* x *for both treatment groups in the adenomatous polyps data set*

way or risk the severe downward bias of procedures based on the assumption of homogeneity, that is, on "pure" binomial or Poisson models. Since the time of Fisher *et al.* (1943) considerable success has been achieved using mixed-Poisson models with

model to data, and that it gives us a view of a new and little-known territory for exploring the robustness and extensions of that model.

6 Appendix

6.1 Comparative simulation study

We begin with one further extension. The suggested weighted linear regression estimator $\hat{N}$ depends on a first-order Taylor approximation which might not be good for larger values of x. One might consider a second-order approximation, but this leads to an estimator with large variance due to the functional relationship of x and x^2. An alternative linear approximation is possible by developing $\log(k + x) = \log((k - 1) + (x + 1))$ linearly around $x + 1$ leading to the approximation

$$\log(x + 1) + (k - 1)/(x + 1)$$

and the regression model

$$\log\left[\frac{(x + 1)f_{x+1}}{f_x}\right] - \log(x + 1) = \gamma + \delta/(x + 1) + \varepsilon_x. \qquad (6)$$

We call this the *hyperbolic model* (HM). The hyperbolic model is also of very simple structure and prediction is possible since the model is defined for $x = 0$ leading to $\hat{f}_0 = f_1/\exp(\hat{\gamma} + \hat{\delta})$. We denote the estimator based on this model by $\hat{N}_{HM}$.

In the following simulation comparison, then, we compare $\hat{N}$, $\hat{N}_{HM}$, $\hat{N}_{CB}$ and $\hat{N}_{Ch}$. We generated counts from a negative binomial distribution with dispersion

parameters equal to 1, 2, 4, 6, and 10 and event probability parameter such that the associated mean matches 1. The population sizes to be estimated were $N = 100$ and $N = 1,000$. For each simulated data set $f_0, f_1, \ldots, f_m$ were generated; then f_0 was ignored and $f_1, \ldots, f_m$ were used to compute the various estimators. This process was repeated 1,000 times and bias, variance and MSE were calculated from the resulting values. The results are shown in Table 10. Clearly $\hat{N}$ performs better than $\hat{N}_{HM}$ since the former always has smaller MSE than the latter. In fact, there is only once case in which $\hat{N}_{HM}$ had smaller bias than $\hat{N}$, namely $N = 1000$ and $k = 1, 2$ and the smaller bias here was balanced by the smaller variance of $\hat{N}$. Hence, we do not consider $\hat{N}_{HM}$ any further. We see in addition that $\hat{N}$ and $\hat{N}_{CB}$ overestimate the true size $N = 100$ whereas $\hat{N}_{Ch}$ tends to underestimate. We need to point out that $\hat{N}_{CB}$ produced many negative values so its bias and RMSE were evaluated on the basis of the positive values. The bias of $\hat{N}$ is smaller than that of $\hat{N}_{CB}$, and the same size that of $\hat{N}_{Ch}$. Also, the RMSE of $\hat{N}_{CB}$ is a lot larger than that of $\hat{N}$. The situation changes for $N = 1,000$. In this case both the bias and MSE for $\hat{N}$ are lower than those from $\hat{N}_{Ch}$ for every value k of the dispersion parameter. We notice, however, that $\hat{N}_{CB}$ shows a reduced bias, but the RMSE of the $\hat{N}$ is still smaller. Overall, we find that $\hat{N}$ and $\hat{N}_{CB}$ are behaving somewhat similarly for larger population sizes; however, a major benefit of $\hat{N}$ is that it is well-defined in the many situations where $\hat{N}_{CB}$ fails.

6.2 Standard errors

In Tables 9 and 10 we compare the standard error calculated from (3) with the true standard error. This was done by taking 10,000 replications of $\hat{N}$, say $\hat{N}_i, i =$

Table 8: *RMSE and Bias for estimators based upon the WLRM, the HM, the Chao-Bunge estimator and the lower bound estimator of Chao,* $N = 100$ *and* $N = 1000$*,* $k = 1, 2, 4, 6, 10$*, where* k

1, ..., 10, 000. Then the mean $(1/10,000) \sum_i \widehat{Var}(\hat{N}_i)$ was computed and the root of it forms column 2 in the tables. The third column was constructed by simply computing the empirical standard deviation of $\hat{N}_i$, $i = 1, ..., 10,000$. We see that the approximation is good for larger values of N and reasonable for smaller values of N.

Table 9: *Estimated (using (3)) and true standard error for WLRM estimator* $\hat{N}$*;* $N = 100$ *and* $N = 1,000$, $k = 1, 2, 4, 6, 10$,

Table 10: *Estimated (using (3)) and true standard error for WLRM estimator* $\hat{N}$*;* $N = 100$ *and* $N = 1,000$*,* $k = 7, 8, 9, 11$*,* $p = 0.8$*; Results are based on* $10,000$ *replications*

k	$S.E.(\hat{N})$	true $S.E.(\hat{N})$
N=100		
7	12.20	11.80
8	9.29	8.96
9	7.45	7.35
11	5.03	4.99
N=1000		
7	30.52	31.67
8	24.43	25.46
9	20.05	20.71
11	14.02	14.59

6.3 Dependence of estimators on the truncation point

Table 11: *Dependence of the weighted least-squares $\hat{N}$ and the Chao-Bunge estimator on the truncation point, compared for all datasets*

	polyps – low		polyps – hi		butterflies		microbial	
m	WLRM	C-B	WLRM	C-B	WLRM	C-B	WLRM	C-B
3	609	411	881	446	754	682	767	266
4	525	440	620	459	744	696	364	492
5	509	471	542	472	776	715	364	492
6	523	524	513	482	759	727	364	-240
7	519	596	512	497				

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

[8]

[15]

[23] Van der Heijden, P. G. M., Van Putten, W., Van Rongen, R. (2006). A comparison of Zelterman's and Chao's estimators for the size of an unknown population by capture-recapture frequency data. Personnel Communication with P.v.d. Heijden.

[24] Holzmann, H., Munk, A., and Zucchini, W. (2003). On identifiability in capture-recapture models. *Biometrics* 62, 934–939.

[25] Hsu, Chiu-Hsien (2007). A weighted zero-inflated Poisson model for estimation of recurrence of adenomas. *Statistical Method in Medical Research* 16, 155–166.

[26] Johnson, N.L., Kemp, A.W., Kotz, S. (2005). *Univariate Discrete Distributions*.

[32] Pledger, S. A. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics* **56**, 434–442.

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

[34] Quince, C., Curtis, T. P., and Sloan, W. T. (2008). The rational exploration of microbial diversity. *The ISME Journal* **2**, 997–1006.

[35] Roberts, J.M. & Brewer, D.D. (2006). Estimating the Prevalence of male clients of prostitute women in Vancouver with a simple capture-recapture method. *Journal of the Royal Statistical Society (Series A)* **169**, 745–756.

[36] Stock, A., Jürgens, K., Bunge, J., and Stoeck, T. (2009). Protistan diversity in the suboxic and anoxic waters of the Gotland Deep (Baltic Sea) as revealed by 18S rRNA clone libraries. *Aquatic Microbial Ecology* **55**, 267-284.

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

[38] Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Berlin-Heidelberg-New York: Springer.

[39]

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