School of Mathematical
and Physical Sciences

# Department of Mathematics and Statistics

**Preprint MPS_2010-30**

**17 August 2010**

# Capture-Recapture Estimation by Means of Empirical Bayesian Smoothing with an Application to the Geographical Distribution of Hidden Scrapie in Great Britain

Dankmar Böhning
Department of Mathematics and Statistics
School of Mathematical and Physical Sciences
University of Reading, Whiteknights
Reading, RG6 6BX, UK
email: d.a.w.bohning@reading.ac.uk

Ronny Kuhnert
Robert Koch-Institute, FG 23, Berlin, Germany
email: KuhnertR@rki.de

Victor Del Rio Vilas
Department for Environment, Food and Rural Affairs (Defra)
London, SW1P 3JR, UK
email: victor.delriovilas@defra.gsi.gov.uk

August 17, 2010

# 1 Introduction

For integer $N$, we consider a sample of counts $x_1, x_2, ..., x_N \in \{0, 1, 2, ...,\}$ arising from a count random variable $X$ having a mixture probability density function

$$p_x = \int_0^\infty p(x|\lambda)q(\lambda)d\lambda \tag{1}$$

with unspecified mixing density $q(\lambda)$ and a mixture kernel $p(x|\lambda)$ which needs to be specified. In this paper, a typical choice for the mixture kernel is the Poisson $p(x|\lambda) = Po(x|\lambda) = \exp(-\lambda)\lambda^x/x!$ though other choices are possible as well. Whenever $x_i = 0$ unit $i$ remains unobserved, so that only a zero-truncated sample of size $n = \sum_{j=1}^{m} f_j$ is observed, where $f_j$ is the frequency of counts with value $x = j$ and $m$ is the largest observed count. Hence, $f_0$ and consequently $N$ are unknown. The purpose is to find an estimate of the size $N$. Since frequently the count variable $X$ represents repeated identifications of an individual in an observational period, the problem at hand is a special form of the capture-recapture problem (see Bunge and Fitzpatrick (1993) for a review on the topic).

The sample of counts $x_1, x_2, ..., x_N$ can occur in several ways. A target population which might be difficult to count consists out of $N$ units. This population might be a wildlife population, a population of homeless people, drug addicts, software errors or animals with a specific disease. Furthermore, let an identification device (a trap, a register, a screening test) be available that identifies unit $i$ at occasion $t$ where $t = 1, .., T$. Let the binary result be $x_{it}$ where $x_{it} = 1$ means that unit $i$ has been identified at occasion $t$ and $x_{it} = 0$ means that unit $i$ has not been identified at occasion $t$. The indicators $x_{it}$ might be observed or not, but it is assumed that $x_i = \sum_{t=1}^{T} x_{it}$ is observed if at least one $x_{it} > 0$ for $t = 1, ..., T$. Only if $x_{i1} = x_{i2} = ... = x_{iT} = 0$ and, consequently $x_i = 0$,

occurs by repeated identifications of the same unit.

In another setting, which is also the basis for this work, the clustering occurs by means of a grouping variable such as herds, holdings, households, or villages. In this case, $x_{it}$ denotes if the $t-$th element in cluster $i$ is identified ($x_{it} = 1$) or not ($x_{it} = 0$). In the example given in the next section the clusters are holdings of sheep and $x_{it}$ informs about the disease status of the $t-$th animal in holding $i$. Note that $x_i = \sum_t x_{it}$ is observed only if it is positive. In other examples the cluster corresponds to villages or households, one of the earliest applications of this kind is the cholera-outbreak in a community in India studied by McKendrick (1926) in which the cluster corresponds to households in a village. A more recent example involves cholera occurrence in rural East Pakistan where the cluster structure consists of villages (see also Mosley *et al.* (1972)).

The paper is organized as follows. The next section 2 introduces the data on scrapie in Great Britain. In section 3 we review some of the existing approaches in the capture-recapture methodology for the setting of interest. Section 4 describes the development of a new set of empirical Bayes estimators which are then further evaluated by means of a simulation study. The application of the empirical Bayes estimator to the spatial data on scrapie in Great Britain, including the development of maps at county level of completeness and observed–hidden ratio, ends the paper in section 5.

## 2 The data of scrapie in Great Britain

We now consider as a specific case study the spatial distribution of scrapie in Great Britain. Classical scrapie, a neurological fatal disease of small ruminants is endemic in Great Britain (see Del Rio Vilas *et al.* (2006) for more details). There is ample evidence to support the occurrence of under-reporting affecting the clinical notification of scrapie cases (Hoinville *et al.* (2000), Del Rio Vilas

*et al.* (2005), Böhning *et al.* (2008)). Although not established to date, there is reason to believe that, reflecting population and surveillance related heterogeneities, under–reporting presents an uneven distribution across Great Britain. The spatial analysis presented in the following uses county-specific disease data from the Scrapie Notifications Database (SND) (see Vilas *et al.* (2006) for more details), more specifically the number of confirmed clinical cases. Table 1 shows the frequency distribution $f_x$ of the count of confirmed clinical cases $X$ for $x = 1, 2, 3, \ldots$ by county. Evidently, there is a considerable range in the number

(1988) arises. In the approach of Zelterman the homogeneous Poisson serves only as a working model and it was suggested by Zelterman that the estimate $\hat{}$

of $p$ and $\lambda$ the factor can be considerably larger than 1. For example if $q = 0.5$ and $\lambda \leq 0.4$ the factor is larger than 2, so that the Zelterman estimate would overestimate severely. The question arises as to what is the source of this overestimation bias. We approach this question in the next theorem which states that the Zelterman estimator uses the *wrong* expected value in predicting $f_0$.

Theorem 2 *i) Let* $\log L(\lambda) = f_1 \log(p_1) + f_2 \log(p_2)$ *with* $p_1 = e^{-\lambda}\lambda/(e^{-\lambda}\lambda + e^{-\lambda}\lambda^2/2) = 2/(\lambda + 2)$ *and* $p_2 = e^{-\lambda}\lambda^2/2/(e^{-\lambda}\lambda + e^{-\lambda}\lambda^2/2) = \lambda/(\lambda + 2)$ *being the Poisson probabilities truncated to counts of ones and twos. Then* $\log L(\lambda)$ *is maximized for*

$$\hat{\lambda} = 2f_2/f_1.$$

*ii)*

$$E(f_0|f_1, f_2; \hat{\lambda}) = f_1^2/(2f_2), \text{ for } \hat{\lambda} = 2f_2/f_1.$$

*Proof.* For the first part, it is clear that $f_1 \log(p_1) + f_2 \log(p_2)$ is maximal for $\hat{p}_1 = f_1/(f_1 + f_2)$, which is attained for $\hat{\lambda} = 2f_2/f_1$. For the second part, we see that with $e_x = E(f_x|f_1, f_2; \lambda) = Po(x|\lambda)N$:

$$e_x = Po(x|\lambda)N = Po(x|\lambda)N = Po(x|\lambda)(e_0 + f_1 + f_2 + \sum_{j=3}^{\infty} e_j)$$

so that

$$e_0 + e_3^+ = [1 - Po(1|\lambda) - Po(2|\lambda)](e_0 + e_3^+) + [1 - Po(1|\lambda) - Po(2|\lambda)](f_1 + f_2)$$

with $e_3^+ = \sum_{j=3}^{\infty} e_x$. Hence

$$e_0 + e_3^+ = \frac{1 - Po(1|\lambda) - Po(2|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2)$$

and

$$e_0 = Po(0|\lambda)(f_1 + f_2 + e_0 + e_3^+) = Po(0|\lambda)(f_1 + f_2) + Po(0|\lambda)$$
$$= \frac{1 - Po(1|\lambda) - Po(2|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2)$$

$$= \frac{Po(0|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2) = \frac{f_1 + f_2}{\lambda + \lambda^2/2}.$$

Plugging in the maximum likelihood estimate $\hat{\lambda} = 2f_2/f_1$ for $\lambda$ yields the desired result. □

Theorem 2 establishes a close connection between the approach by Zelterman and Chao's estimator. It shows that Zelterman's estimator of the Poisson parameter $\lambda$ arises when all counts are truncated except counts of ones and twos and when the resulting likelihood is maximized. If the correct prediction for $f_0$ is used, namely the conditional expectation for the truncated Poisson model, the Chao estimator arises. Hence the strong overestimation of the original Zelterman estimator stems from using a *wrong* conditional expectation.

## 3.3 Comparing some conventional estimators in a simulation

Before we continue developing the generalized, adjusted version of the Zelterman estimator, we consider the performance of Chao and Zelterman estimators in a small simulation study. In the case of a homogeneous Poisson the maximum likelihood estimate is found by maximizing the likelihood of zero-truncated Poisson observations in $\lambda$:

$$\prod_{j=1}^{m} \frac{p_{\lambda}}{}$$

where $Q$ is the discrete mixing distribution giving $k$ weights $q$

# 4 A new empirical Bayes estimator of population size

Although it is clear that $2f_2/f_1$ estimates the Poisson parameter in the case that $p_x = Po(x|\lambda)$, it is not clear what it estimates when there is a mixing distribution present instead of Poisson homogeneity. Here, a Bayesian perspective is helpful. We think of the mixing distribution $q(\lambda)$ as a prior distribution on $\lambda$ so that

$$E(\lambda|x) = \int_0^\infty \lambda \frac{Po(x|\lambda)q(\lambda)}{\int_0^\infty Po(x|\theta)q(\theta)d\theta} d\lambda \tag{4}$$

is the *posterior mean* w.r.t the prior $q(\lambda)$ and Poisson likelihood for observation $x$. Note that (4) can be further simplified to

$$\lambda_x = E(\lambda|x) = \frac{\int_0^\infty \lambda Po(x|\lambda)q(\lambda)d\lambda}{\int_0^\infty Po(x|\lambda)q(\lambda)d\lambda}$$

$$= (x+1)\frac{\int_0^\infty Po(x+1|\lambda)q(\lambda)d\lambda}{\int_0^\infty Po(x|\lambda)q(\lambda)d\lambda} = (x+1)\frac{p_{x+1}}{p_x},$$

where $p_x$ is the marginal density (1). Before we continue on the ways to estimate the ratio of marginals we point out an important monotonicity property.

Theorem 3

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m.$$

*Proof.* Consider

$$p_j = \int_0^\infty \exp(-\lambda)\lambda^j/j!q(\lambda)d\lambda$$

with unknown $q(\lambda)$ for $\lambda > 0$. Then, by means of the *Cauchy-Schwarz inequality* for random variables $X$ and $Y$:

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

we have that

$$\left[\int_0^\infty \left[\frac{X}{\exp(-\lambda)\lambda^{(j-1)/2}}\right]\left[\frac{Y}{\exp(-\lambda)\lambda^{(j+1)/2}}\right]d\lambda\right]^2$$

$$\leq \left[\int_0^{\infty} \exp(-\lambda)\lambda^{(j-1)}\,d\lambda\right]^{X^2} \left[\int_0^{\infty} \exp(-\lambda)\lambda^{(j+1)}\,d\lambda\right]^{Y^2}$$

or,

$$(j!\,p_j)^2 \leq (j-1)!\,p_{j-1}(j+1)!\,p_{j+1},$$

or, finally $\frac{jp_j}{p_{j-1}} \leq \frac{(j+1)p_{j+1}}{p_j}$. □

Theorem 3 has an important application. Since under heterogeneity we have that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$, we expect that the graph $x \rightarrow \hat{\lambda}_x = (x+1)f_{x+1}/f_x$ shows a monotone increasing pattern if heterogeneity is present. Hence we

appropriately since

$$\sum_{x=2}^{m} \frac{f_x}{1-\exp(-\lambda_x)} \geq \sum_{x=2}^{m} f_x.$$

We are now considering ways of doing so.

The marginal density $p_x$ can be estimated by the relative, empirical frequency $f_x/N$ so that

$$E(\lambda|x) = \hat{\lambda}_x = (x+1)\frac{f_{x+1}}{f_x}$$

provides an estimate of the posterior mean $E(\lambda|x) = \lambda_x$ using the fact that the unknown denominators $N$ cancel out. Hence, the Zelterman estimate occurs as a special case of the nonparametric, empirical Bayes estimator for observation $x$ (Robbins (1955), Carlin and Louis (1997)).

The understanding of Zelterman's original estimator of $\lambda$ as $\hat{\lambda}_1 = 2f_2/f_1$ as empirical Bayes estimator for observation $x = 1$ is useful, since it helps to find ways to eliminate the overestimation bias. We need to define a Horvitz-Thompson estimator that takes into account the different counts $x = 1, 2, ..$ separately. This can be accomplished by defining

$$\hat{N} = \frac{f_1}{1-\exp(-\hat{\lambda}_1)} + \frac{f_2}{1-\exp(-\hat{\lambda}_2)} + ... + \frac{f_m}{1-\exp(-\hat{\lambda}_m)}. \qquad (6)$$

The question arises as to which way the estimator $\hat{\lambda}_x$ should be constructed. A naive estimator would follow the Robbins-type estimation to arrive at

$$\hat{N}_R = \frac{f_1}{1-\exp(-2f_2/f_1)} + \frac{f_2}{1-\exp(-3f_3/f_2)} + ... + \frac{f_{m-1}}{1-\exp(-mf_m/f_{m-1})} + f_m, \qquad (7)$$

where we define

$$\frac{f_j}{1-\exp(-(j+1)f_{j+1}/f_j)} = \begin{cases} 0, & \text{if } f_j = 0; \\ f_j, & \text{if } f_{j+1} = 0. \end{cases}$$

Although the estimator (7) is intuitively attractive, it has some considerable difficulties. Not only is it unclear what to do with the largest count $m$ (in (7) it is not up-weighted), but also various counts could have frequencies zero which

would leave some of the frequencies $f_x$ unweighted. More importantly, most of the observed count data will lie on the lower counts resulting in highly unstable estimates for larger counts.

It is more attractive to consider a *smoothed* version of the Bayes estimator. This can be accomplished by constructing an estimate of the marginal distribution $p_x = \int_0^\infty p(x|\lambda)q(\lambda)d\lambda$ using a discrete, finite mixture

$$p_x = \sum_{j=1}^{k} Po(x|\lambda_j)q_j,$$

where $\lambda_j > 0$ and the non-negative weights $q_j$ sum up to 1. Estimates can be constructed by means of the EM algorithm or using some gradient-type algorithm. For details see Böhning and Kuhnert (2006). Some attention needs to be given to the question of the number of components $k$. Two strategies will be looked at:

- The number of components is determined by the nonparametric maximum likelihood estimator (NPMLE).
- The mixture model is chosen on the basis of the Bayesian Information Criterion (BIC) defined as $-\log L(Q) + (2k-1)\log(n)$.

In both cases we arrive at some estimate of the marginal distribution

$$\hat{p}_x = \sum_{j=1}^{k} Po(x|\hat{\lambda}_j)\hat{q}_j \qquad (8)$$

leading to smoothed estimates of the population size

$$\hat{f}$$

We will also consider two further ways of estimating the mixing distribution $q(\lambda)$ in $\int_0^\infty Po(x|\lambda)q(\lambda)d\lambda$. The first estimator is based upon the idea of using the empirical distribution itself as an estimator of the mixing distribution. To accomplish this task we have to consider the appropriate transformation of the observed frequencies. Let $\tilde{q}_i = f_i/n$ denote the relative frequencies of the observed, zero-truncated sample. According to Böhning and Kuhnert [1] the associated relative proportions of the zero-truncated mixture are given as

$$\hat{q}_i = \frac{\tilde{q}_i/[1 - Po(0|x_i)]}{\sum_{=1}^{n} \tilde{q}_i/[1 - Po(0|x\ )]},$$

so that $\hat{p}_x = \sum_{j=1}^{m} Po(x|x_j)\hat{q}_j$ and

$$\hat{N}_{\text{EDF}} = \sum_{=1}^{m} \frac{f}{1 - \exp(-(\ +1)\frac{\hat{p}_{\ +1)}}{\hat{p}})},$$

where the index EDF associates with the empirical distribution function. The benefit of this approach is that the estimate of the mixing distribution is readily available without any computational e□ort. The second additional estimator is building upon the Γ-distribution for $q(\lambda)$ in $p_x = \int_0^\infty Po(x|\lambda)q(\lambda)d\lambda$

counts $X_1, \ldots X_N$ were drawn from a two-component mixture of Poisson densities: $X \quad 0.5Po(1) + 0.5Po(\lambda)$, evidently with equal weights $q_1 = q_2 = 0.5$. The population size was set to $N = 100$ and $1,000$ replications used. Here, we will concentrate on the main findings. More details are available in the supplementary material Böhning *et al.* (2010). We see from Table 2 that both empirical Bayes estimators perform better with respect to their standard error and root mean square error than the other estimators adjusting for heterogeneity. If we compare the two empirical Bayes estimators it appears that the one based upon the nonparametric mixture model as smaller variance which is reflected also in a better mean squared error.

# 5 Application to spatial analysis of scrapie in Great Britain

Following the results of the previous section we will concentrate on using the

## 5.1 Determining the NPMLE for the SND data

We have seen in section 4 using the ratio plot that there is strong evidence for heterogeneity captured by a mixing distribution. We consider the marginal distribution over all counties as available from Table 1: $f_1$ = 298, $f_2$ = 89, $f_3$ = 42,..., $f_{29}$ = 2. We are using this (truncated) count distribution to determine the maximum likelihood estimators for the various possible mixture models. The results are provided in Table 3. For each number of components $k$, starting with the homogeneous case $k$ = 1, the estimated mixture model $\hat{Q}$ is provided, the Poisson parameters $\hat{\lambda}_j$ and associated component weights $\hat{q}_j$. This is followed by the log-likelihood $\log L(\hat{Q})$ and the BIC-value $-2\log L(\hat{Q}) + (2k-1)\log(n)$. Note that there are two estimates of the population size of scrapie-a□ected holdings given. One is based upon the direct computation using the mixture model estimated as provided in (3), the other is the empirical Bayes estimate using the estimated mixture as prior distribution (10). It is evident from columns 6 and and 7 in Table 3, that the empirical Bayes estimate of the population size is less sensitive to the choice of the number of components. Furthermore, the empirical Bayes estimates is not prone to spurious estimates as is the conventional mixture model based estimator. We have already mentioned that Figure 1 supports that there is considerable evidence for a monotone increasing pattern. In addition, the estimate of the posterior mean based upon the estimated mixture model with 4 components (this is what the BIC suggests) shows that this monotone pattern is met. Note that columns 6 and and 7 in Table 3 contain also (in brackets) an estimate of the standard error of the repsective population size estimate. This was achieved by applying the nonparametric bootstrap as adapted to capture–recapture situations by van der Heijden *et al.* (2003) and Böhning (2008). It is evident from columns 7 in Table 3 that the conventional mixture model based estimator is prone to extreme variance inflation when the number of components

become large.

## 5.2 Estimating the number of hidden scrapie-a□ected holdings per county

We now apply these results to the individual counties using (11). Note that we are using the same mixture distribution in (11) estimated from the entire SND data. This is necessary since the county specific case distributions are frequently very sparse. Take for example county 1 in Table 1: we find $f_{1,1} = 2$, $f_{2,1} = 1$, $f_{3,1} = 1$, so $n_1 = 4$. It is clear that a reliable estimation of a mixing distribution is not possible from this count distribution. Hence we use the mixing distribution estimated from the entire data set and assume that the heterogeneity found for the entire data set is also valid in each county. Then we compute the *predicted* number of scrapie-a□ected holdings by applying the weight $(1 - \exp[-( + 1)\frac{\hat{\rho}_{+1)}}{\hat{\rho}}])^{-1}$ to the frequency $f_{,i}$ of count in the $i-$th county and summing up over all observed frequencies $f_{,i}$ leading to

$$\hat{N}_i = \sum_{=1}^{\square} \frac{f_{,i}}{1 - \exp(-( + 1)\frac{\hat{\rho}_{+1)}}{\hat{\rho}})}.$$

This process is very similar to indirect standardization used in epidemiologic methodology (see Waller and Gotway (2004, p. 17). The results are provided in Table 4. In addition, two further measures are computed. The *observed–hidden ratio* defined as $n_i/(\hat{N}_i - n_i)$ and the *completeness* measure defined as $n_i/\hat{N}_i$, provided as columns 4 and 5 in Table 4. The completeness ranges betwe8re5u.r81.448% and 99%re5u. Figure 2 shows a scatterplot of the completeness against the observed count (on log-scale) of scrapie-a□ected holdings. There is no evidence for a specific pattern, though the variation of completeness seems to decrease with increasing observed count of scrapie-a□ected holdings. Median observed–hidden ratio is 1.29 with 95% nonparametric CI (1.11, 1.43) and completeness is 56.36 with 95% nonparametric CI (52.62%, 58.83%).

Figure 3 shows the geographical distribution of county–specific completeness and observed–hidden ratios. Completeness is fairly stable with most counties in the 50-59% category and fewer counties in the upper completeness categories. Note that as well as providing completeness and observed/hidden ratios, we can also estimate adjusted measures of disease occurrence for each county. However, for our particular case, this would not have a clear biological interpretation as annual data was pooled to increase the power of our analyses.

# 6 Discussion

As described in section four and five, providing theoretical evidence and empirical support respectively, $\hat{N}_i = \hat{N}_{BIC,i}$ represents a lower bound of the population size in each county $i$. Hence, the estimated completeness $n_i/\hat{N}_i$ in county $i$ will be an upper bound for $n_i/N_i$, so that the estimated values for completeness will be too large on average. Consequently, since the observed values already have an upper limit of almost 100%, it is expected that only the observed minimum for completeness of 48% will be in fact a bit lower. Similarly, we expect that the observed-hidden ratios are overestimated. Typically, we have seen in the simulation study that $N$ is underestimated by 5 – 10%, never more than 20%.

The maps are based upon an estimated size of the scrapie population in county $i$, given as

$$\hat{N}_i = \sum_{=1}^{m} \frac{f_{,i}}{1 - \exp(-( \ + 1)\frac{\hat{\rho}_{\ +1)}}{\hat{\rho}})} = \sum_{=1}^{m} \hat{w} \ f_{,i},$$

where $\hat{\rho}$ is found from (10) with an estimated BIC-selected nonparametric mixing distribution. Since the estimated weights $\hat{w} \ = 1/[1 - \exp(-( \ + 1)\frac{\hat{\rho}_{\ +1)}}{\hat{\rho}})]$ do not depend on the county index $i$ we have that

$$\sum_{i} \hat{N}_i = \sum_{i} \sum_{=1}^{m} \hat{w} \ f_{,i} = \sum_{=1}^{m} \hat{w} \ \sum_{i} \hat{f}_{,i} = \sum_{=1}^{m} \hat{w} \ f \ = \hat{N},$$

where $f = \sum_i f_{,i}$, so that the margin (over counties) of the county-specific estimates of the size of the scrapie population and the estimate of the size of the scrapie population, unstratified by county, coincide.

Finally, note that it is also possible tlsoe4660-1(.)smates alfo(P)73(also)-462(p)-28(ossossibi)]TJ, coincal

[17] 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.

[18] Hoinville, L.J., Hoek, A.R., Gravenor, M.B., McLean, A.R. (2000). Descriptive epidemiology of scrapie in Great Britain: results of a postal survey. *Veterinary Record* 146, 455–461.

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

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

[21] Link, W.A. (2003). Response to a paper by Holzmann, Munk and Zucchini. *Biometrics* 62, 936–939.

[22] McKendrick, A.G. (1926): Application of Mathematics to Medical Problems. *Proceedings of the Edinburgh Mathematical Society* 44, 98-130.

[23] Mosley, W.H., Bart, K.J., and Sommer, A. (1972). An epidemiological assessment of cholera control programs in rural East Pakistan. *International Journal of Epidemiology* 1, 5–11.

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

[25] Roberts, J.M. and 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.

[26] Rocchetti, I., Bunge, J. and Böhning, D. (2010). Population size estimation based upon ratios of recapture probabilities. *submitted for publication*.

[27] Robbins, H. (1955). An empirical Bayes approach to statistics. *In Proc.3rd Berkeley Symp. on Math Statist. and Prob.,1, Berkeley, CA: University of California Press*, 157–164.

[28] Waller, L.A., Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data.* Hoboken, NJ, Wiley.

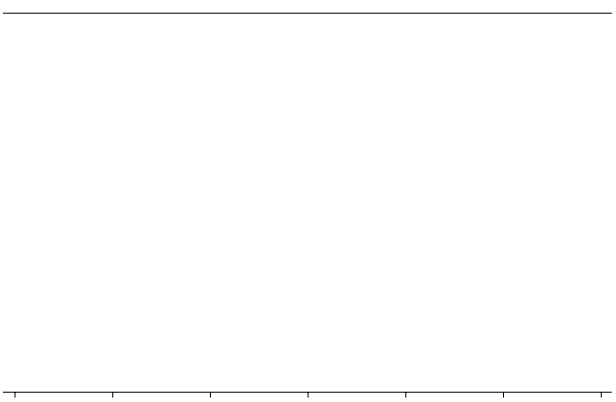

Figure 1: *Ratio plot for SND data 2002-2006, unstratified by county, for Robbins estimate of posterior mean as well as the discrete mixture (4 components) based empirical Bayes estimate of the posterior mean*

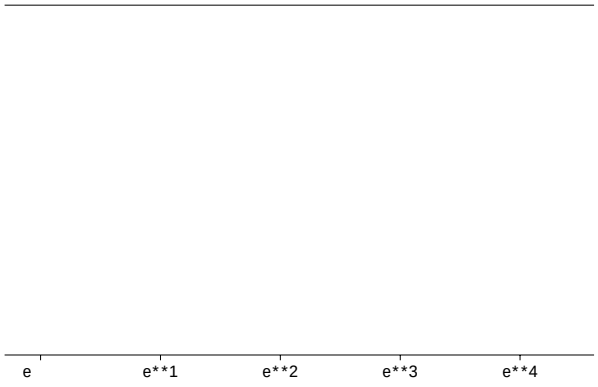


Figure 2: *Scatterplot of completeness of surveillance stream per county against observed count of scrapie a ected holdings per county*

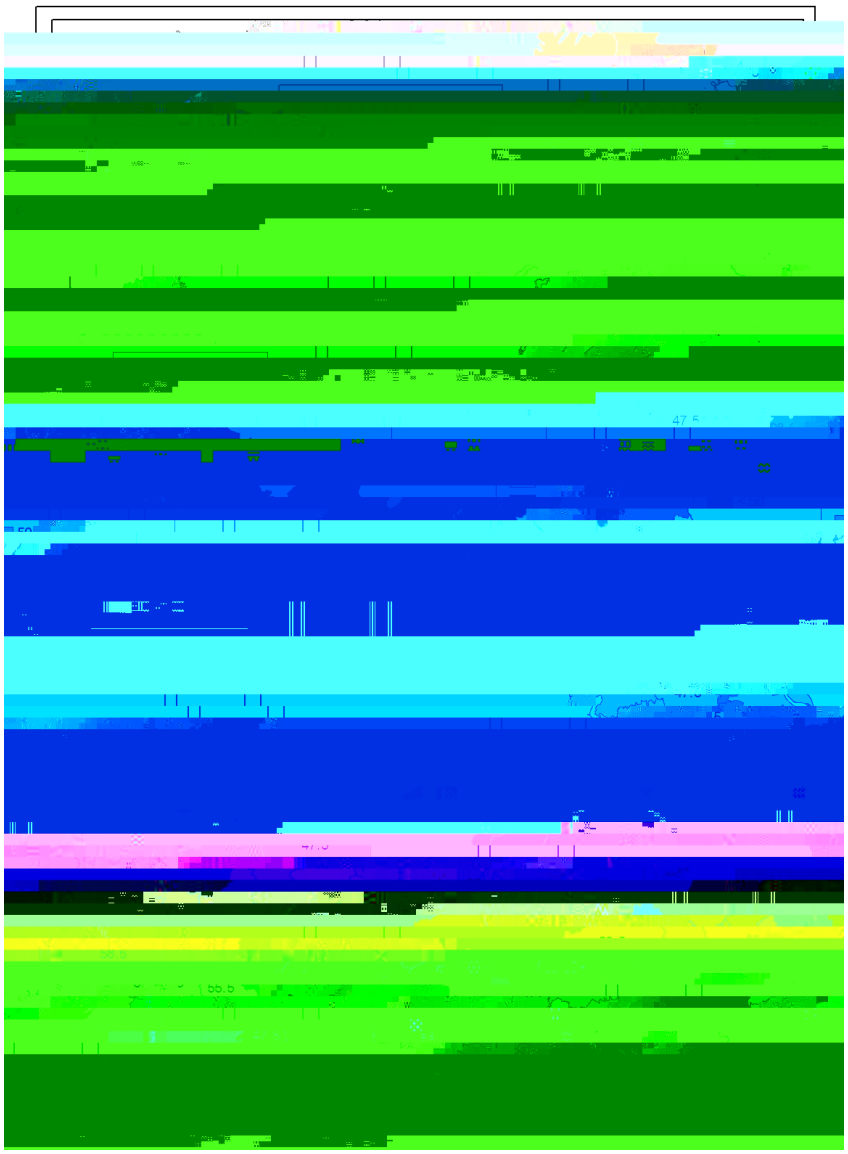

Figure 3: *Map of estimated completeness on county level for SND data 2002-2006*

Table 1: *Distribution of confirmed scrapie–a ected holdings from the SND database 2002–2006 by county*

| county | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $f_7$ | $f_8$ | $f_9$ | $f_{10+}$ | $n$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 2 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 |
| 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 3 | 7 |
| 6 | 4 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 11 |
| 7 | 12 | 1 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 1 | 19 |
| 8 | 7 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 11 |
| 9 | 25 | 8 | 5 | 1 | 1 | 1 | 2 | 0 | 0 | 2 | 45 |
| 10 | 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 11 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 12 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 13 | 2 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 14 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 15 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | 5 | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 10 |
| 17 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 18 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| 19 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 20 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 21 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 22 | 3 | 3 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 9 |
| 23 | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 7 |
| 24 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 25 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 3 |
| 26 | 6 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 8 |
| 27 | 5 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 7 |
| 28 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 3 |
| 29 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 |
| 30 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 31 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| 32 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 33 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 34 | 14 | 10 | 3 | 1 | 3 | 0 | 2 | 1 | 0 | 3 | 37 |
| 35 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

... continued on next page ...

| ... continued from previous page ... | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| county | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ | $f_7$ | $f_8$ | $f_9$ | $f_{10+}$ | $n$ |
| 36 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

Table 2: *Simulation using* $X \quad 0.5Po(1) + 0.5Po(\lambda)$ *and* $N = 100$*; provided are estimates of* $E(\hat{N})$, $Var(\hat{N})^{1/2}$ *and* $[E(\hat{N} - N)^2]^{1/2}$ *as mean, SD and RMSE*

| $\lambda$ | estimator | mean | SD | RMSE |
|---|---|---|---|---|
| 1 | MLE-hom | 102 | 13 | 13 |
| | NPMLE | 484 | 12,098 | 20,028 |
| | Chao | 104 | 19 | 19 |
| | Zelterman | 105 | 21 | 22 |
| | EB-NPMLE | 105 | 15 | 15 |
| | EB-Robbins | 108 | 21 | 22 |
| 2 | MLE-hom | 94 | 7 | 9 |
| | NPMLE | 4599 | 35 | 21,328 |
| | Chao | 99 | 12 | 12 |
| | Zelterman | 101 | 16 | 16 |
| | EB-NPMLE | 98 | 8 | 9 |
| | EB-Robbins | 102 | 12 | 12 |
| 3 | MLE-hom | 88 | 5 | 13 |
| | NPMLE | 12,517 | 52,425 | 23,955 |
| | Chao | 97 | 10 | 11 |
| | Zelterman | 102 | 15 | 16 |
| | EB-NPMLE | 93 | 7 | 10 |
| | EB-Robbins | 96 | 9 | 10 |
| 4 | MLE-hom | 85 | 4 | 16 |
| | NPMLE | 11,715 | 54,501 | 23,114 |
| | Chao | 97 | 10 | 10 |
| | Zelterman | 108 | 20 | 20 |

Table 3: *Estimated mixture models for 1, 2, 3 , 4 and 5 (NPMLE) number of components with associated estimator of the size of the scrapie–a ected population of holding from the unstratified SND database 2002–2006*

| | | | | | discrete mixture model based | |
|---|---|---|---|---|---|---|
| $k$ | $\hat{\lambda}_j$ | $\hat{q}_j$ | $\log L(\hat{Q})$ | BIC | $\hat{N}_{\text{NPMLE}}$ (10), $(SE)$ | $\hat{N}_{\text{NPMLE}}$ (3), $(SE)$ |
| | | | | | | |
| 1 | 2.33 | 1.00 | -1,279.0 | 2,561.4 | 572 (9.4) | 572 (9.4) |
| 2 | 0.97 | 0.88 | -865.4 | 1,740.8 | 776 (32.4) | 793 (34.6) |
| | 9.80 | 0.12 | | | | |
| 3 | 0.67 | 0.80 | -807.8 | 1,632.4 | 869 (44.8) | 946 (65.8) |
| | 5.46 | 0.17 | | | | |
| | 19.10 | 0.03 | | | | |
| 4 | 0.56 | 0.75 | -802.3 | 1,628.2 | 896 (48.0) | 1,036 (60,102) |
| | 4.03 | 0.19 | | | | |
| | 10.35 | 0.05 | | | | |
| | 23.58 | 0.01 | | | | |
| 5 | 0.01 | 0.27 | -801.2 | 1,632.7 | 916 (25.5) | 528,694 (419,663) |
| | 1.08 | 0.54 | | | | |
| | 5.13 | 0.14 | | | | |
| | 11.76 | 0.03 | | | | |
| | 23.98 | 0.01 | | | | |

Table 4: *Observed and hidden scrapie-a ected counts of holdings by county, observed–hidden ratio and completeness of surveillance stream*

| county | $n$ | $\hat{N}$ | $o/h$ | completeness |
|---|---|---|---|---|
| 1 | 4 | 7 | 1.4 | 59 |
| 2 | 4 | 6 | 2.3 | 70 |
| 3 | 1 | 2 | 0.9 | 48 |
| 4 | 1 | 2 | 0.9 | 48 |
| 5 | 7 | 9 | 3.1 | 76 |
| 6 | 11 | 16 | 2.2 | 69 |
| 7 | 19 | 33 | | |

| county | $n$ | $\hat{N}$ | $o/h$ | completeness |
|---|---|---|---|---|
| ... continued from previous page ... | | | | |
| 37 | 75 | 137 | 1.2 | 55 |
| 38 | 11 | 19 | 1.3 | 57 |
| 39 | | | | |