

ANALYSIS OF HIDDEN SCRAPIE IN GREAT BRITAIN: CAPTURE-RECAPTURE ESTIMATION WITH A GEOGRAPHICAL APPLICATION

ISDS, 03/12/2009

Victor J del Rio Vilas, Department for Environment, Food
and Rural Affairs (Defra), UK

Dankmar Bohning, University of Reading, UK

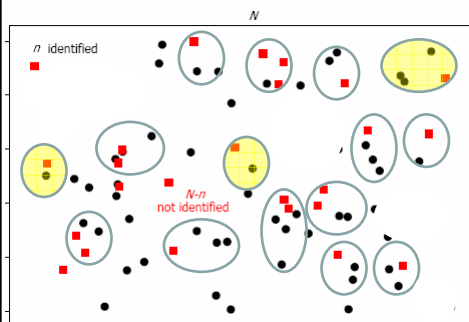
Ronny Kuhnert, Robert Koch Institute Berlin, Germany

1

Perfect estimators, why?

- Providing unbiased parameters for
 - Transmission models
 - Models to measure interventions' impact (surveillance Se, disease control programmes)
- Relative resource allocation (between competing alternatives).
- Hook and Regal (2004) found little impact in PH arena
- Because?
 - A drop in an ocean of uncertainty (too little too late)
 - Conservative control strategies that go beyond minimum
 - National level estimates (i.e. too big a problem: unmanageable)

1. To apply it at geographical level to show surveillance inequalities
2. To develop a capture-recapture estimator (CRC) that overcomes limitations of traditional estimators
3. To apply a degree of smoothness to inform under-reporting affecting clinical notifications of scrapie.



- One list with
 - A count of repeated identifications within cluster (e.g. households with cholera, animal holdings with scrapie)
 - Objective: estimate affected clusters not recorded by identifying mechanism (p_0).
 - Prob. of identifying a cluster is $(1-p_0)$
 - And $N = (1-p_0)N + p_0 N - n + p_0 N$
 - So that, Horvitz-Thompson estimator follows

$$\hat{N} = \frac{n}{1 - p_0}$$

- McKendrick (1926)
- Drakos (2007) on estimating hidden transnational terrorist activity with f_0 "countries with unrecorded terrorist activities".

| f_0 | f_1 | f_2 | f_3 | f_4 | f_5 | f_6 | f_7 | f_8 | f_9 | ... | f_{136} | n |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----|-----------|-----|
| - | 286 | 114 | 101 | 59 | 33 | 21 | 20 | 19 | 11 | ... | 1 | 785 |

- Del Rio Vilas et al (2008) on estimating hidden scrapie affected holdings in GB.

| f_0 | f_1 | f_2 | f_3 | f_4 | f_5 | f_6 | f_7 | f_8 | n |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----|
| - | 84 | 15 | 7 | 5 | 2 | 1 | 2 | 2 | 118 |

- "Everything is related to anything else, but near things are more related than distant things" (Tobler's law).

- Fatal neurological disease affecting small ruminants
- Of the family of BSE, CJD and other TSE's
- Endemic, classical form, in GB and notifiable since 1993.
- Previous estimates of under-reporting by multiple methods (anonymous surveys, CRC). All national.



- Presence of heterogeneity in capture probabilities (every cluster may be different) so simpler models (Poisson λ) lead to underestimation.
- Multiple estimators (e.g. Chao's lower bound, Zelterman)

$$\hat{N}_Z = \frac{n}{1 - \exp(-\frac{2f_2}{f_1})}$$

- Problems associated with them (lower bound, large variance for Zelterman).
- New estimator suggested here:
 - A smoothed generalisation of Zelterman's that overcomes the over-estimation

$$\hat{N}_{\text{NPMLE},i} = \sum_{\ell=1}^m \frac{f_{\ell,i}}{1 - \exp(-(\ell + 1) \frac{\hat{p}_{\ell+1}}{\hat{p}_{\ell}})}$$

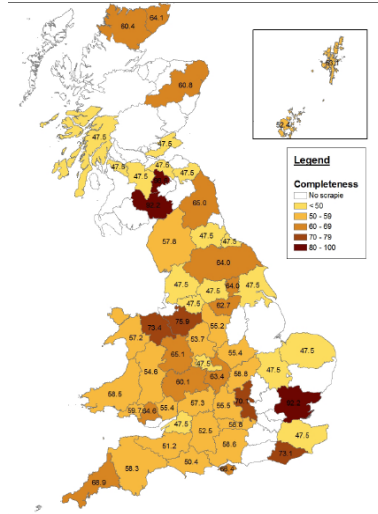
- Uses an empirical Bayes approach with different choices (parametric (gamma) and non-parametric) for priors

Table 1: Distribution of confirmed scrapie cases 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 |
|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-----------|-----|
| 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| 3 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 |
| 4 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 3 | 7 |
| 7 | 4 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 3 | 11 |
| 8 | 12 | 1 | 0 | 2 | 3 | 0 | 0 | 0 | 0 | 1 | 19 |

- **Outputs:**
 - Biological interest: maps of completeness and adjusted prevalence
 - Methodological interest: from simulation comparing performance of conventional traditional estimators.

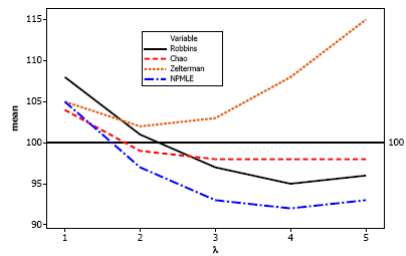
Results (biological)



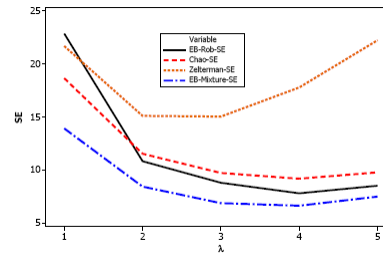
12/10/2009

9

Results (methodological)



Expected values of two Empirical Bayes estimators vs. Chao and Zelterman



Standard errors of two Empirical Bayes estimators vs. Chao and Zelterman

12/10/2009

10

- Need to appreciate other forms of smoothing (more related to neighbouring effects/informing the “blanks”?)
 - Borrowing information from 1st/2nd neighbours?
 - Need to incorporate systematic variability (e.g. Holdings number/area, surveillance related variables to ascertain coverage inequalities so interventions are possible, etc).
 - Only host, environment and identifying mechanisms are relevant
- Other biological problems/applications?

Thanks

- More details in **Bohning, Kuhnert and Del Rio Vilas**. “*Capture-recapture estimation by means of empirical Bayesian smoothing with an application to spatial analysis of hidden scrapie in Great Britain*” To appear in Journal of Agricultural, Biological and Environmental Statistics (JABES)

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

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

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

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 λ so that

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

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

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

Now, the problem we had was with the Zelterman estimator

Now we need ways to estimate the ratio of marginals

12/10/2009

13

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

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

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 [24], Carlin and Louis [5]).

That is, we observed that Zelterman estimator of λ , $\lambda_1 = 2f_2/f_1$ is the empirical Bayes estimator for $x=1$.

We need to define a Horvitz-Thompson estimator that takes into account the different counts $x = 1, 2, \dots$ separately. This can be accomplished by defining

$$\hat{N}^* = \frac{f_1}{1 - \exp(-\hat{\lambda}_1)} + \frac{f_2}{1 - \exp(-\hat{\lambda}_2)} + \dots + \frac{f_m}{1 - \exp(-\hat{\lambda}_m)} \quad (6)$$

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

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

where we define

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

(7) appears attractive but some counts may have frequencies zero which would leave some frequencies unweighted. Also, most of the observed count data will lie on the lower counts. This leads to highly unstable estimates for larger counts. 14

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

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

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

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

leading to smoothed estimates of the population size

NPMLE of the mixing distribution as the smoothed empirical Bayes estimate of the prior distribution for further analysis, in particular

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

$$\hat{p}_x = \sum_{j=1}^k Po(x|\hat{\lambda}_j)\hat{q}_j, \quad (10)$$

as derived in (8). In a first step, this will be done using the entire SND data, unstratified by county. Once an estimate for the mixing distribution has been achieved, a smoothed *county-specific* estimate of the population size can be developed as follows:

$$\hat{N}_{\text{NPMLE},i} = \sum_{\ell=1}^m \frac{f_{\ell,i}}{1 - \exp(-(\ell + 1) \frac{\hat{p}_{\ell+1}}{\hat{p}_\ell})}, \quad (11)$$

where $f_{\ell,i}$ is the frequency of holdings with ℓ cases in the i -th county and \hat{p}_ℓ is taken from (10).