21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Can we use the approaches of ecological inference to learn about the potential for dependence bias in dual-system estimation?

An application to cancer registration data

J. J. Brown[a], E. J. Beh[b] and I. L. Hudson[b]

[a] *School of Mathematical and Physical Sciences, University of Technology Sydney, New South Wales*
[b] *School of Mathematical and Physical Sciences, University of Newcastle, New South Wales*
Email: james.brown@uts.edu.au

Abstract: The dual-system estimator, or estimators with a similar underlying set of assumptions and structure, is a widely used approach to estimate the unknown size of a population. Within official statistics its use is linked with population census, while in health applications it is often used to estimate true levels of incidence from imperfect reporting systems; the classic example being work by Sekar and Deming exploring the estimation of births in India in the 1940s. Critical to the implementation of dual-system estimation are the assumptions that the probability of being counted in a source is homogeneous and that the event of being counted in each source is independent. When either of these assumptions fails, the two by two table will have an odds ratio different to one and the dual-system estimator will be biased.

Inferential frameworks such as the aggregate association index (AAI) have been developed to allow the researcher to assess the plausibility of independence between two variables in a two by two table, when only the margins are observed. Given any appropriate measure of relationship, this strategy relies on determining the AAI, which provides an indication of the likely association structure between the variables given only the marginal information. Further advances of the AAI have also been established including its link with the odds ratio and its relationship with the size of the study being undertaken. Determining the population size from a two by two table given limited information is an alternative variation of the framework on which the AAI is built. Therefore the underlying theoretical properties of the two by two table are identical in both scenarios – it is only the nature of the unknown information that differs.

In this paper we make the first steps to exploring the use of an AAI type framework (and its relatives) to assess the plausibility of an independence assumption in applications of population size estimation. We use alternative data set-ups based on real data relating to historical cancer registration (with three sources of registration) to demonstrate that the chi-square statistic behaves differently over a range of values for the missing data for differing true relationships between the two variables. We then apply the approach to the cancer registration from two of the registration systems to show that we can see evidence of potential dependence from the observed but incomplete data.

The first results in this paper demonstrate the possibility of exploring the independence assumption when estimating the unknown population size from two lists. As with the AAI framework, the aim is not to directly estimate the level of the association but rather alert the analyst to the potential for an association and its direction allowing them to assess the likelihood of a biased estimate for the population size. This has important implications within a health setting where it is potentially useful to understand if the true population size, of say cancer patients, is likely to be higher or lower than the estimate constructed assuming independence. Within the official statistics setting, it can alert us to situations where it is advantageous to explore whether external data exist that would allow an adjustment for dependence in our two lists.

Keywords: *Aggregate association index, dual-system estimation, dependence, ecological inference, measures of association*

Brown *et al*., Can we use the approaches of ecological inference to learn about the potential for dependence bias in dual-system estimation: application to cancer registration data?

1. INTRODUCTION

Capture re-capture (CR) has a long history being used to measure the size of human populations going back to Sekar and Deming (1949). In recent years, there has been an explosion of the use of the method (Böhning and van der Heijden, 2009) with many applications to count hidden or hard-to-reach human populations in the areas of epidemiology and public health (see for example Hook *et al*., 2012 and Laure and Stéphane, 2008). In this paper we revisit the data presented in Robles *et al*. (1988) by examining the completeness of cancer registration data. In parallel, the approach is widely used in the estimation of wild-life populations (Seber, 1982; Pollock, 1991) and the specific area of assessing census coverage (Wolter, 1986; Brown *et al*., 1999; United Nations, 2010). The use of dual lists gives rise to two main issues in the CR framework, namely, list dependence and heterogeneity; the latter can be further classified as observed and unobserved heterogeneity. The heterogeneity issue is typically tackled through modelling capture rates extending the post-stratification approach outlined in Sekar and Deming (1949) with further extensions such as use of mixture models (Böhning *et al*., 2005). However, adjusting for dependence requires additional information. This can be in the form of additional lists as in Robles *et al*. (1988); see Baffour *et al*. (2013) for an exploration of triple-system estimators, see Thandrayen and Wang (2009, 2010) for further extensions with multiple sources, or by the introduction of external information as in Wolter (1990) or Brown *et al*. (2006). In fact, the application to UK Census coverage outlined in Brown *et al*. (2006) is one of the few attempts to produce population size estimates accounting for dependence within the dual-system framework. The US Census Bureau apply the approach in Wolter (1990) to explore potential bias in their dual-system estimates of census coverage, but do not make a direct adjustment.

The CR situation is not the only framework for analysing incomplete two-by-two contingency tables. Another seemingly unrelated area has been the development of approaches for ecological inference (EI) of two-by-two tables. This technique involves estimating the cells (or some function of the cells) of stratified two-by-two contingency tables when only the marginal, or aggregate data is available. One may refer to, for example, Goodman (1959), Freedman *et al*. (1991), King (1997), Wakefield (2004), Steel, Beh and Chambers (2004) and Wakefield, Haneuse, Dobra and Teeple (2011) for detailed strategies for making such inferences. Hudson, Moore, Beh and Steel (2012) demonstrated the effectiveness of a variety of techniques for performing EI by considering early historical New Zealand gendered election data covering the period 1893 – 1919. However, EI techniques are subject to two key limitations. Firstly, they require the imposition of assumptions of the unknown cell frequencies that are either restrictive or untestable; see, for example, Hudson *et al*. (2012). Secondly, despite the common use of data summarised in a single two-by-two table, all EI techniques are only applicable in studies involving multiple (stratified) two-by-two tables. In other words, similar assumptions are required to make progress with EI as in CR.

Therefore, in the case of EI, rather than trying to estimate the cell frequencies, an alternative strategy is to focus on the association structure between the variables. Analysing aggregate data using this strategy can be undertaken using the aggregate association index, or AAI, proposed by Beh (2008, 2010) and discussed in terms of occupational epidemiology by Tran, Beh and Smith (2012). This index quantifies the extent of association that may exist between two dichotomous variables at the α level of significance, given only the aggregate data from a single two-by-two contingency table. Underlying the theory of the AAI is Pearson's chi-squared statistic. Therefore, rather than estimating the cells of multiple two-by-two contingency tables, the purpose of the AAI is to quantify the likelihood that a statistically significant association exists between the two dichotomous variables. Unlike the numerous EI techniques that are now available, the AAI is applicable to the analysis of a single (as well as multiple) table and has been studied further recently. For example, Beh, Cheema, Tran and Hudson (2014) examine the role of the sample size on the magnitude of the AAI and Beh, Tran and Hudson (2013) consider the derivation of the AAI in terms of the odds ratio of the two-by-two contingency table.

Table 1. Demonstrating the link between the EI Problem and Capture Re-capture for Estimating Population Size

Ecological Inference

	Counted	Missed	
Counted	~~N_{11}~~		$\mathbf{N_{1+}}$
Missed			
	$\mathbf{N_{+1}}$		$\mathbf{N_{++}}$

Capture Recapture

	Counted	Missed	
Counted	$\mathbf{N_{11}}$		$\mathbf{N_{1+}}$
Missed			
	$\mathbf{N_{+1}}$		~~N_{++}~~

Brown *et al*., Can we use the approaches of ecological inference to learn about the potential for dependence bias in dual-system estimation: application to cancer registration data?

In this paper we start to bring together these two approaches, by taking the same approach as AAI but within the context of CR, to tackle the issue of assessing dependence between two lists within traditional capture re-capture. Table 1 shows the available information in the two situations. When faced with EI, we observe the margins of the table but know nothing about the cell frequencies that lie within the table, while in CR we see observe a single joint cell frequency but not the total size. Therefore, in both cases we see three out of the four needed cells to define the entire two-by-two table; and the underlying association structure between the two lists. In fact, in the EI situation we do have slightly more information as the margins define an upper and lower bound for the unknown (1,1) cell (Duncan and Davis, 1953; Beh, 2008); while in the CR situation the unknown population size just has a lower bound defined by setting the unobserved count N_{00}, the unknown count for those missing on both lists, to zero.

To make progress in either situation, we typically assume independence between the two lists. In other words we assume the odds ratio (OR) is one and therefore the $\log_e$ of the OR, which we will define as γ, is zero. Under independence this then leads to estimates of the missing count given by

$$\hat{N}_{11} = \frac{N_{1+} \times N_{+1}}{N_{++}} \text{ (EI) and } \hat{N}_{++} = \frac{N_{1+} \times N_{+1}}{N_{11}} \text{ (CR with Dual-System Estimation).}$$

In the EI framework there has been extensive work that explores the plausibility of this independence assumption using the information available in the observed margins, primarily through the use of the AAI. However, in the CR framework there has to date been no such development. Therefore, as far as the authors are aware, this paper represents a first attempt to borrow from the ideas behind measures such as AAI in the EI framework to explore the sensitivity of a dual-system estimate to potential dependence between the two lists.

2. THE EXAMPLE DATA

To illustrate the ideas in this paper we consider the data described in Robles *et al*. (1988) relating to the estimation of cancer registrations. In their situation Robles *et al*. (1988) observed three lists allowing us to explore the relationship between two lists knowing the correct answer for the sub-population defined by the third list. Figure 1, reproduced from Robles *et al*. (1988), shows the observed information. From this we can see that hospital discharges provide most of the new registrations with considerable overlap to both pathology reports and death certificates. We therefore start by working with death certificates and pathology reports as our two lists for the 'population' defined by hospital discharges. In the context of Table 1 this implies N_{11} is 5152 while the 'unobserved' count N_{00} is 7406.

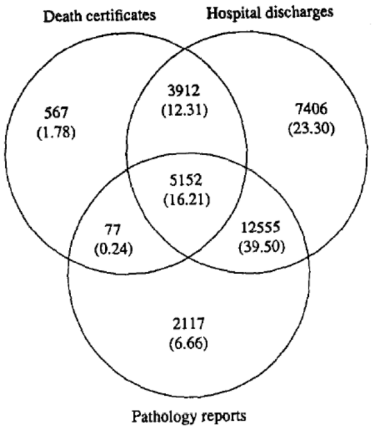

Figure 1. Distribution of sources of registration for the Ontario Cancer Registry, 1982. Percent of the total number of new registrations (n = 31786) is shown under the number for each combination of sources.

See Figure 1, Robles *et al*. (1988)

3. EXPLORING THE INFORMATION IN THE DATA

The work behind the development of the AAI is built on exploring the behaviour of the chi-squared statistic when we 'impute' different values for the missing count N_{11}, or equivalently when we vary the assumed value for γ away from zero. As a start to exploring within the CR framework we take a similar approach. There is a simple link between γ and the unobserved count of the missing, N_{00}, given by

$$\hat{N}_{00} = \frac{(N_{1+} - N_{11}) \times (N_{+1} - N_{11}) \times \exp(\gamma)}{N_{11}} = \frac{3912 \times 12555 \times \exp(\gamma)}{5152} \quad (1)$$

where the numbers are the appropriate counts from Figure 1 treating the hospital discharges data as the population of interest; with the observed data as the margins and (1,1) cell coming from the two other lists for this population defined by hospital discharges. Using the relationship given in (1), we can vary the value of γ, and therefore the value of the missing count N_{00}. For each value we can then calculate the classic chi-squared statistic, χ^2, and plot this against γ. Figure 2a shows this for the observed data. Figure 2a also demonstrates the relationship assuming the margins of the two lists (pathology and deaths) are fixed, along with the unknown

size of the population defined by hospital discharges, but the underlying association between the two lists captured by the OR is varied. This is achieved by suitably adjusting the value of the count N_{11}.

a) b)

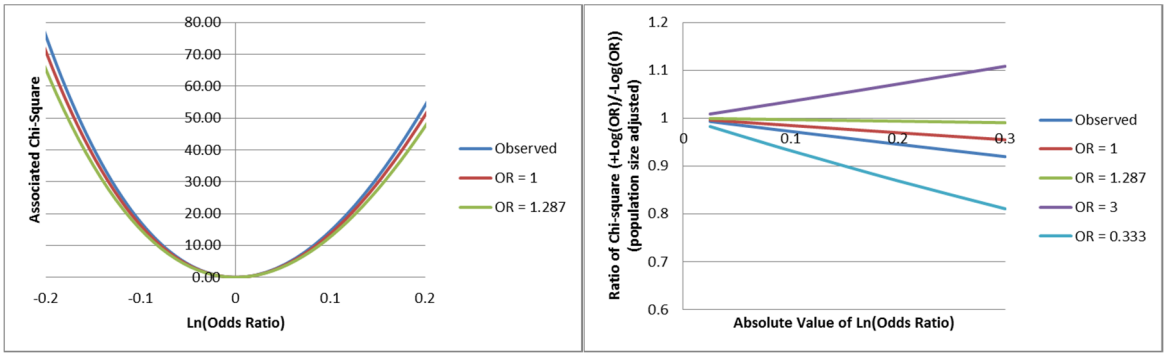

Figure 2. Exploring the relationship between χ and γ for different values of the true under-lying association as captured by the OR

Figure 2a shows the quadratic relationship between the chi-squared statistic and the log-odds ratio. Such a curve is akin to the AAI curves of Beh, Tran and Hudson (2013) that are used to graphically explore the association between the variables of a two-by-two contingency table using only the aggregate data. From the AAI curve, the AAI is quantified. Another technique that is commonly used to graphically explore association in contingency tables is correspondence analysis (Beh and Lombardo, 2014). While the focus of this paper is not to quantify the AAI we do borrow from the features that underly the index. Figure 2a does show that the association is likely to be very strong between the two variables since the area under the curve, but above the line defined by the critical value of χ^2 with one degree of freedom is very large. Figure 2a also allows us to see the possibility to spot the difference as the relationship is not symetric about zero for γ, and the nature of the asymmetry appears to depend on the underlying association. To capture this differing asymmetry, we first scale the values for χ^2 by dividing by the implied population size associated each value of γ to get values we define as χ_N. This removes the impact of an increasing population size on the chi-squared statistic. Define $\chi_N(+\gamma)$ and $\chi_N(-\gamma)$ to be those values of χ_N when the log-odds ratio is positive and negative, respectively. Then in Figure 2b we plot the relationship between the absolute value of the log-odds ratio and $\frac{\chi_N(+\gamma)}{\chi_N(-\gamma)}$, to capture the asymmetry when the underlying OR varies, against the absolute value of γ.

Figure 2b reveals a clear differential in the behaviour of this ratio when the true OR moves well away from independence. A positive slope indicates that the underlying OR is greater than one, and a negative slope occurs when the OR is less than one. For the observed relationship, the slope is marginally less than one reflecting the fact that in this case the true OR is just less than 0.8.

4. APPLICATION TO DIFFERENT SUB-POPULATIONS

In Section 3 we have borrowed from the idea behind the development of the AAI to propose a measure, in this case our ratio of the population size adjusted chi-square statistic, to explore whether there is evidence that the underlying association captured by the unobserved OR is moving away from independence. To demonstrate whether there is further potential with this approach, we now apply the method to the two other populations defined by death registrations and pathology reports.

The results presented in Figure 3 suggest that in both cases the underlying OR is greater than one, and hence if we estimate the unknown population size with the classic dual-system estimator for CR we will under-estimate the number of cancer registrations in each population. This is indeed **true in both cases**. For the population of death registrations, the association between hospital and pathology is captured by an OR of 9.7. The dual-system estimator would be around 9200 while the true sub-population size (see Figure 1) is over 9700. For the population of pathology reports, the association between hospital and deaths is captured by an OR of 11.3. The dual-system estimator would be just under 18000 while the true population size (see Figure 1) is just over 19900. Therefore, while using the approach in Figure 3 we cannot directly adjust the dual-system estimator; we can alert those using the estimates to the likely direction of any bias. In both of these cases we

would be warning that the dual-system estimate of the unknown population size is likely to be greatly under-estimated. In that sense it mirrors the role of the AAI in that we assess for the presence of an association without directly estimating it.

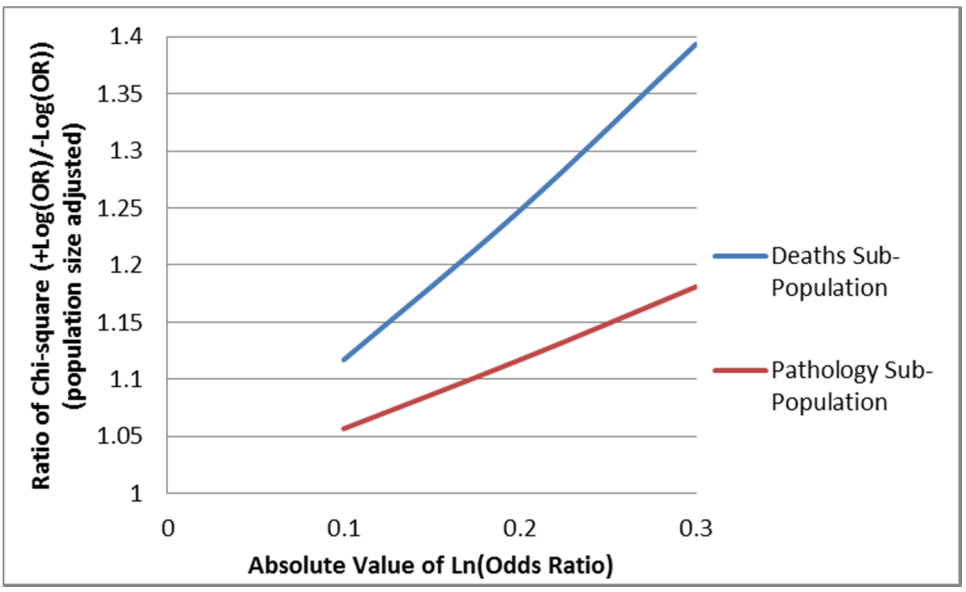

Figure 3. Exploring evidence for the under-lying OR moving away from independence in the sub-populations defined by death registrations and pathology reports.

5. DISCUSSION

There are well developed approaches within EI to explore the potential association between two lists when the joint count N_{11} is not observed. This allows us to explore the likely bias in an estimate of N_{11} based on an independence assumption. In CR, the aim is to estimate the unknown population total and classic dual-system estimation achieves this by also making the independence assumption. However, the ability to explore the plausibility of this assumption from the observed data on the two lists is not developed, even though the approach is widely used to estimate the size of hidden or hard-to-reach populations in both health and social science applications.

In this paper, we have borrowed ideas from EI to explore the underlying association in the CR system where the assumption of independence between the lists is made. The initial work presented here shows there is potential to develop a more formal measure along the lines of the AAI that has been developed for EI. This measure can be developed in terms of the odds ratio of the incomplete two-by-two contingency table. Being able to inform a policy-maker of the likely direction of any underlying association has a real benefit because it alerts them to the presence and direction of bias in the dual-system estimate of population size. In the application used in this paper, if we apply dual-system estimation to the full data for the pathology and hospital lists, we would estimate the population of new cancer cases to be around 32620. The original paper using all three lists estimated it around 33500. The approach developed in Section 3 would have suggested the log odds ratio was likely positive (odds ratio greater than one) for the two lists. This would imply that our dual-system estimator is likely to be negatively biased, thereby alerting the policy-maker to be cautious with respect to any future planning using the estimate.

In this application the population size and overlap are large. However, CR is more sensitive to the dependence assumptions when the overlap between the two lists is small, as demonstrated in Gerritse, van der Heijden and Bakker (2015). Moving forward, the use of simulations will help demonstrate the usefulness of this approach with smaller sample sizes, and smaller overlap, when detecting movement from independence is more crucial for CR.

REFERENCES

Baffour, B**.,** Brown, J.J, and Smith, P.W. (2013) An investigation of triple system estimators in censuses. *Journal of the International Association of Official Statistics (JIOAS), **24**, 53-58.*

Beh, E. J. (2008), Correspondence analysis of aggregate data: The 2x2 table, *Journal of Statistical Planning and Inference*, **138**, 2941-2952.

Beh, E. J. (2010), The aggregate association index, *Computational Statistics & Data Analysis*, **54**, 1570 – 1580.

Beh. E. J., Tran, D. and Hudson, I. L. (2013), A reformulation of the aggregate association index using the odds ratio, *Computational Statistics & Data Analysis*, **68**, 52 – 65.

Beh, E. J., Cheema, S. A., Tran, D. and Hudson, I. L. (2014), Adjustment to the aggregate association index to minimize the impact of large samples, *Advances in Latent Variables* (eds Carpita, M., Brentari, E. and Qannari, E. M.), pp. 241 – 251, Springer; Switzerland.

Beh, E. J. and Lombardo, R. (2014), *Correspondence Analysis: Theory, Practice and New Strategies*, Chichester: Wiley.

Böhning, D. Dietz, E., Kuhnert R, D. Schön, D. (2005) Mixture models for capture-recapture count data. *Stat. Methods Appl.*, **14**, 29-43.

Böhning, D. and van der Heijden, P. (2009) Recent developments in life and social science applications of capture-recapture methods. *Advances in Statistical Analysis*, **93**, 1-3.

Brown, J.J., Diamond, I.D., Chambers, R.L., Buckner, L.J. and Teague, A.D. (1999) A methodological strategy for a one-number census in the UK. *J. R. Statist. Soc. A*, **162**, 247-267.

Brown, J., Abbott, O. & Diamond, I. (2006) Dependence in the 2001 one-number census project. *J. R. Statist. Soc.* A, **169**, 883–902.

Freedman, D. A., Klein, S. P., Sacks, J., Smyth, C. A. and Everett, C. G. (1991) Ecological regression and voting rights. *Evaluation Review*, **15**, 673–711.

Gerritse, S., van der Heijden, P. and Bakker, B. (2015) Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models. *Journal of Official Statistics*, **31**, 357–379.

Goodman, L.A. (1959) Some alternatives to ecological correlation. *American Journal of Sociology*, **64**, 610 – 625.

Hook, E. B., Hsia, M. S., and Regal, R. R. (2012) Accuracy of Capture-Recapture Estimates of Prevalence. *Epidemiologic Methods*, **1**, 1-11.

Hudson, I.L., Moore, L., Beh, E.J., Steel, D.G. (2010) Ecological inference techniques: an empirical evaluation using data describing gender and voter turnout at New Zealand elections, 1893–1919. *J. R. Statist. Soc. A*, **173**, 185 – 213.

King, G. (1997) *A Solution to the Ecological Inference Problem*. Princeton: Princeton University Press.

Laure, V. and Stéphane, L. (2008) Capture-recapture estimates of the local prevalence of problem drug use in six French cities. *European Journal of Public Health*, **19**, 32 – 37

Pollock K. (1991) Modeling capture, recapture and removal statistics for estimation of demographic parameters for fish and wildlife populations. *J Amer Stat Assn*, **86,** 225–238.

Robles, S. C., Marrett, L. D., Clarke, E. A. & Risch, H. A. (1988) An application of capture-recapture methods to the estimation of completeness of cancer registration. *J. Clin. Epidemiol.*, **41**, 495-501.

Seber, G. A. F. (1982). *The estimation of animal abundance and related parameters*. Second edition published by Charles Griffin & Company Ltd, London.

Sekar, C. C. and Deming, W. E. (1949) On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, **44**, 101-115.

Steel, D. G., Beh, E. J. and Chambers, R. L. (2004) The information in aggregate data. In *Ecological Inference: New Methodological Strategies* (eds G. King, O. Rosen and M. Tanner), pp. 51 – 68. Cambridge: Cambridge University Press.

Tran, D., Beh, E. J. and Smith, D. R. (2012) Real-word occupational epidemiology, part 3: An aggregate data analysis of Selikoff's "20-year rule". *Archives of Environmental & Occupational Health*, 67, 243 – 248.

Thandrayen, J. and Wang, Y. (2009) A latent variable regression model for capture-recapture data. *Comput. Statist. Data Anal.*, **53**, 2740-2746.

Thandrayen, J. and Wang, Y. (2010). Capture–recapture analysis with a latent class model allowing for local dependence and observed heterogeneity. *Biometrical Journal*, **52**, 552-561.

United Nations (2010) *Post Enumeration Surveys; Operational guidelines*. Technical Report by Department of Economic and Social Affairs, Statistics Division, New York, USA. (http://unstats.un.org/unsd/demographic/standmeth/handbooks/Manual_PESen.pdf)

Wakefield, J. (2004) Ecological inference for 2 × 2 tables (with discussion). *J. R. Statist. Soc. A*, **167**, 385–445.

Wakefield, J., Haneuse, S., Dobra, A. and Teeple, E. (2011) Bayes computation for ecological inference. *Statistics in Medicine*, **30**, 1381 – 1396.

Wolter, K. M. (1986) Some Coverage Error Models for Census Data. *J Amer Stat Assn*, **81**, 338-346.

Wolter, K. (1990) Capture-Recapture Estimation in the Presence of a Known Sex Ratio. *Biometrics*, **50**, 1219-1221.