Dual system estimation for under-count adjustment

Li-Chun Zhang[1,2]

[1] ***University of Southampton (L.Zhang@soton.ac.uk)***
[2] ***Statistisk sentralbyrå, Norway***
[3] ***Universitetet i Oslo***

How many fish in the pond?

	Second fishing	
First fishing	Caught (List B)	Not caught
Caught (List A)	4 (n_{11})	8 (n_{12})
Not caught	6 (n_{21})	? (n_{22})

1st occasion: capture A, marked and released

2nd occasion: recaptures (AB) of the marked ones

Capture-Recapture methodology: numerous applications in wild-life, medical, social studies (Böhning, D. et al., 2017)

Dual system estimator (DSE)

Let $N = n_{11} + n_{12} + n_{21} + n_{22}$. Unknown n_{22} and N.

DSE, also known as the Lincoln-Petersen estimator:

$$\hat{n}_{22} = \frac{n_{12}n_{21}}{n_{11}} \qquad \text{and} \qquad \hat{N} = \frac{n_{1+}n_{+1}}{n_{11}}$$

Chapman correction if n_{11} is small or zero:

$$\hat{N}_C = \frac{(n_{1+} + 1)(n_{+1} + 1)}{n_{11} + 1} - 1$$

NB. General case with K incomplete lists (Fienberg, 1972)

NB. An evaluation of 10 TSEs with 3 lists (Griffin, 2014)

Dual system estimator (DSE)

Odds ratio $R_{r_1 r_2; c_1 c_2}$ in two-way contingency table:

$$R_{r_1 r_2; c_1 c_2} = \frac{E(n_{r_1 c_1}) E(n_{r_2 c_2})}{E(n_{r_1 c_2}) E(n_{r_2 c_1})} \equiv 1$$

i.e. constant odds ratio, provided the row-classification is independent of the column-classification. For a 2×2-table, setting $r_1 = c_1 = 1$ and $r_2 = c_2 = 2$ yields

$$\frac{E(n_{11}) E(n_{22})}{E(n_{12}) E(n_{21})} = 1 \quad \Leftrightarrow \quad E(n_{22}) = \frac{E(n_{12}) E(n_{21})}{E(n_{11})}$$

Replacing $E(\cdot)$ by observed values gives the DSE of $E(n_{22})$.

- "Causal independence" assumption of Wolter (1986)

Dual system estimator (DSE)

Other assumptions for DSE (Wolter, 1986):

- "Closure" of the target population, denoted by U
- "Multinomial" distribution of $(\delta_{iA}, \delta_{iB})$, for $i \in U$
- "Spurious Events": no duplicated or erroneous count
- "Nonresponse": complete keys available for matching
- "Matching": subset AB identified without error
- "Autonomous Independence": δ_{iL} independent of δ_{jL}, for $i \neq j \in U$, where $L = A, B$
- Homogenous catch probability: $\Pr(\delta_{iL} = 1) = p_L > 0$, for any $i \in U$, where $L = A, B$

Dual system estimator (DSE)

Provided the 8 assumptions above,

- $\hat{N}$ is consistent for N, asymptotically as $n_{11} \to \infty$
- Variance estimators for $\hat{N}$:

$$\widehat{V}(\hat{N}) \approx n_{11}^{-3} n_{1+} n_{+1} n_{12} n_{21}$$

- $100(1-\alpha)\%$ confidence interval of N:

$$n_{11} + n_{12} + n_{21} + \frac{(n_{12}+0.5)(n_{21}+0.5)}{n_{11}+0.5} \exp(\pm z_{0.5\alpha} \hat{\tau})$$

$$\hat{\tau}^2 = \frac{1}{n_{11}+0.5} + \frac{1}{n_{12}+0.5} + \frac{1}{n_{21}+0.5} + \frac{n_{11}+0.5}{(n_{12}+0.5)(n_{12}+0.5)}$$

Application: A = Census, B = CCS

- "Closure": CCS asks about presence on Census day; CCS fieldwork close to Census date [NB. tension below?]

- "Causal Independence": completely different head office teams; CCS sample areas unknown to Census; interviewers work in different areas in Census and CCS

- "Autonomous Independence" [NB. household cluster effects?]

- Homogenous ("Multinomial") catch probability: post-stratification $U_1, ..., U_h, ..., U_H$, e.g. by Age, area, etc. [NB. bias-var. trade-off; modelling of factors $n_{11}^{(h)}/n_{1+}^{(h)}$]

Application: $A =$ Census, $B =$ CCS

- "Matching" & "Nonresponse": quality of record linkage

$$\hat{N}_{RL} = \frac{n_{1+}n_{+1}}{\tilde{n}_{11}}$$

where $\tilde{n}_{11}$ is the size of linked set [$\neq$ match set], and

$$\tilde{n}_{11} = n_{11} - m_{11} + u_{11}$$

where m_{11} is the no. missing matches, and u_{11} the no. false links. Thus

$$\begin{cases} \hat{N}_{RL} > \hat{N} & \text{if } m_{11} > u_{11} \\ \hat{N}_{RL} < \hat{N} & \text{if } m_{11} < u_{11} \end{cases}$$

Application: A = Census, B = CCS

- "Spurious Events": no duplicated or erroneous count

- RL for de-duplication of either Census or CCS
- Use Census follow-up survey (Nirel and Glickman, 2009):

$$\hat{N}_O = \hat{\beta} n_{1+} \Big(\frac{n_{+1}}{n_{11}}\Big)$$

where $1 - \hat{\beta}$ is the estimator of Census over-coverage rate, and assuming negligible spurious events in CCS

NB. See e.g. Hogan (1993) for an account in US, Renaud (2007) for Swiss Census 2000, and Abbott (2009) for 2011 UK Census.

Admin register replacing census: A = SPD, S = PCS

Statistical Population Dataset (SPD) based on admin data

- Patient Register, Tax Records, School Census in UK
- direct tabulation from processed SPD unlikely suffices
- considerable "spurious records" in SPD — more later on

Population Coverage Survey (PCS) probability design

NB. Reverse Record Check (RRC): S = Census...

NB. Zhang and Dunne (2017): an Irish application based on admin registers *entirely*, S = Driver License Renewal

$A =$ SPD, $S =$ PCS: Assumptions reconsidered

SPD under-coverage unlike Census; may be systematic

Treat SPD as **fixed**, PCS as the only **random** source:

(i) No duplicated records in A or S, $A \subset U$ and $S \subset U$

(ii) Matches between A and S identified without errors

(iii) Homogenous capture in S: for any $i \in U$,

$$\pi_i = \pi \quad \text{and} \quad 0 < \pi < 1$$

(iv) Uncorrelated captures in S: for any $i \neq j \in U$,

$$Cov(\delta_i, \delta_j | U) = 0$$

NB. See Zhang (2018) for details

A = SPD, S = PCS: Assumptions reconsidered

- "Closure" unnecessary: population ref. date in PCS

SPD	PCS		Total
	Caught (S)	Not caught	
Caught (A)	4 (n_{11})	8 (n_{12})	12 (n_{1+})
Not caught	6 (n_{21})	? (n_{22})	
Total	10 (n_{+1})		? (N)

Provided (i), all the 12 SPD-enumerations belong to U

Provided (ii) & (iii), $\hat{\pi} = \frac{4}{12}$ = PCS-catch rate estimate

$$E(n_{+1}) = N\pi \quad \Rightarrow \quad \hat{N} = \frac{n_{+1}}{\hat{\pi}} = \frac{n_{+1}n_{1+}}{n_{11}}$$

Consistency of $\hat{N}$, as $N \to \infty$, provided (iv) in addition

A = SPD, S = PCS: Assumptions reconsidered

- "Causal Independence" / "Multinomial" assumptions: unnecessary / no longer applicable with fixed SPD

  - Independence only defined for two random variables

    Let k be a constant and X a random variable:

$$Cov(k, X) \equiv 0$$

- Homogeneous catch by assumption (iii) and "Autonomous Independence" among PCS-captures by assumption (iv)
- "Spurious Events" by assumption (i)
- "Matching" & "Nonresponse" by assumption (ii)

$A = \text{SPD}$, $S = \text{PCS}$: Assumptions reconsidered

Let $\hat{N} = xn/m$, with $x = |A|$, $n = |S|$, $m = |A \cap S|$.

Expanding $\hat{N}$ with respect to (n, m) around (μ_n, μ_m) yields

$$\begin{aligned}
\hat{N} &= N + \frac{x}{\mu_m}(n - \mu_n) - \frac{N}{\mu_m}(m - \mu_m) \\
&\quad - \frac{x}{\mu_m^2}(n - \mu_n)(m - \mu_m) + \frac{N}{\mu_m^2}(m - \mu_m)^2 + R_3 \\
E(\hat{N}|A)/N - 1 &= (1 - \frac{x}{N})\mu_m^{-2} V(m|A) + E(R_3)/N \\
V(\hat{N}) &\approx \frac{(N - x)^2}{\mu_m^2} V(m|A) + \frac{x^2}{\mu_m^2} V(n - m|A)
\end{aligned}$$

where $V(m|A) = x\pi(1 - \pi)$ and $V(n - m|A) = (N - x)\pi(1 - \pi)$.

Now that $\frac{x}{N} = O(1)$ asymptotically, as $N \to \infty$, and R_3 is the lower-order remainder, $\hat{N}$ is consistent for N.

A = SPD, S = PCS: Relaxing assumptions

- Can allow intra-cluster correlation in PCS instead of (iv)

(iv.c) $Cov(\delta_i, \delta_j) = 0$ for $i \in U_k$ and $j \in U_l$, for $1 \leq k \neq l \leq K$, and $U = \bigcup_{k=1}^{K} U_k$ partitioned into K clusters. The variance is then

$$V(\hat{N}) = \frac{(N-x)^2}{\mu_m^2} V(m|A) + \frac{x^2}{\mu_m^2} V(n-m|A) - 2\frac{x(N-x)}{\mu_m^2} Cov(n-m, m|A)$$

- Assumption (iii) can be relaxed in various ways:

(iii.h) $\pi_i = \pi_h$ and $0 < \pi_h < 1$, for $i \in U_h$, where $U_1, ..., U_H$ form a post-stratification of the target population U.

(iii.a) $\bar{\pi}_A = \bar{\pi}_A^c$, with $\bar{\pi}_A = \sum_{i \in A} \pi_i / x$ and $\bar{\pi}_A^c = \sum_{i \in U \setminus A} \pi_i / (N - x)$ as the *average* capture probabilities in and out of A, respectively.

(iii.ha) $\bar{\pi}_{A_h} = \bar{\pi}_{A_h}^c$, $\bar{\pi}_{A_h} = \sum_{i \in A \cap U_h} \pi_i / x_h$, $\bar{\pi}_{A_h}^c = \sum_{i \in U_h \setminus A} \frac{\pi_i}{N_h - x_h}$

A = SPD, S = PCS: Alternative assumption (iii)

Rewrite the DSE in a *prediction* form:

$$\hat{N} = n + (x - m)\frac{n}{m} = \sum_{i \in S} \delta_i + \sum_{i \in A \setminus S} \frac{n}{m} = \sum_{i \in S} (\delta_i - \frac{n}{m}) + \sum_{i \in A} \frac{n}{m}$$

where $\sum_{i \in S} \delta_i = n$ is the no. counts in the PCS, and n/m is a factor adjusting the under-counting of A. Under (iii), the factor is a constant over U. To allow for *heterogenous* factors, let

$$\hat{N} = \sum_{i \in S} (\delta_i - \mathbf{a}_i^\top \mathbf{b}) + \sum_{i \in A} \mathbf{a}_i^\top \mathbf{b} = \sum_{i \in S} (\delta_i - \xi_i) + \sum_{i \in A} \xi_i$$

For instance, under (iii.h), we can let

$$\xi_i = \mathbf{a}_i^\top \mathbf{b} = \frac{n_h}{m_h} \quad \text{for } i \in U_h$$

where $\mathbf{a}_i$ = dummy stratum vector, and $\mathbf{b} = (\frac{n_1}{m_1}, ..., \frac{n_H}{m_H})^\top$

$A = \text{SPD}$, $S = \text{PCS}$: Alternative assumption (iii)

NB. Can use $\sum_{i \in S}(\delta_i - \xi_i)/\pi_i$ to account for out-of-PCS areas

Two concerns:

- to be applicable to $A \setminus S$, the values $\mathbf{a}_i$ need to be known for $i \in A$

- however, heterogeneity may depend on values <u>only</u> observed in S

  [NB. $\mathbf{a}_i$ known for $i \in A$ may be subject to measurement error]

Let $\mathbf{z}_i =$ the q-vector of heterogeneity factors observed for $i \in S$

Let $\mathbf{a}_i =$ the known p-vector of choice for all $i \in A$

Let $d_i = 1$ if $i \in U_g$ and 0 otherwise, for partition $U = \cup_{g=1}^{G} U_g$

Need to model $E(d_i|a_i)$ <u>without</u> the assumption

$$E(d_i|a_i) = E(d_i|a_i, i \in A)$$

$A = \text{SPD}$, $S = \text{PCS}$: Alternative assumption (iii)

Let $\tilde{\mathbf{a}}_i = \mathbf{a}_i$ if $i \in A$, and $\tilde{\mathbf{a}}_i = \mathbf{0}$ if $i \in U \setminus A$.

As alternative to assumption (iii), suppose

$$E(d_i|\mathbf{z}_i, \tilde{\mathbf{a}}_i) = E(d_i|\mathbf{z}_i, \tilde{\mathbf{a}}_i, i \in S) = E(d_i|\mathbf{z}_i, i \in S) = \mathbf{z}_i^\top \boldsymbol{\theta}_{q\times 1}$$

$$E(\tilde{\mathbf{a}}_i^\top|\mathbf{z}_i) = E(\tilde{\mathbf{a}}_i^\top|\mathbf{z}_i, i \in S) = \mathbf{z}_i^\top \boldsymbol{\gamma}_{q\times p}$$

We have then,

$$E(d_i|\tilde{\mathbf{a}}_i) = E\big(E(d_i|\mathbf{z}_i, \tilde{\mathbf{a}}_i)|\tilde{\mathbf{a}}_i\big) = E(\mathbf{z}_i^\top \boldsymbol{\theta}|\tilde{\mathbf{a}}_i) = E(\mathbf{z}_i^\top|\tilde{\mathbf{a}}_i)\boldsymbol{\theta}$$

$$\mathbf{z}_i^\top = E(\tilde{\mathbf{a}}_i^\top|\mathbf{z}_i)\boldsymbol{\gamma}^- \qquad \text{[NB. generalised inverse } \boldsymbol{\gamma}^-]$$

Chipperfield et al (2017) propose the empirical PREG

$$\xi_i = \widehat{E(d_i|\tilde{\mathbf{a}}_i)} = \tilde{\mathbf{a}}_i^\top \mathbf{b} \qquad \mathbf{b} = (\sum_{i\in S} \mathbf{z}_i \tilde{\mathbf{a}}_i^\top)^-(\sum_{i\in S} \mathbf{z}_i d_i)$$

where $\hat{\boldsymbol{\gamma}} = (\sum_{i\in S} \mathbf{z}_i \mathbf{z}_i^\top)^{-1}(\sum_{i\in S} \mathbf{z}_i \tilde{\mathbf{a}}_i^\top)$, $\hat{\boldsymbol{\theta}} = (\sum_{i\in S} \mathbf{z}_i \mathbf{z}_i^\top)^{-1}(\sum_{i\in S} \mathbf{z}_i d_i)$

NB. $\sum_{i\in U} \xi_i = \sum_{i\in U} \tilde{\mathbf{a}}_i^\top \mathbf{b} = \sum_{i\in A} \mathbf{a}_i^\top \mathbf{b}$ since $\tilde{\mathbf{a}}_i = \mathbf{0}$ for $i \notin A$

A = SPD, S = PCS: Alternative assumption (iii)

NB. PREG uses $\hat{\mathbf{z}}_i^\top = \tilde{a}_i^\top \hat{\boldsymbol{\gamma}}^-$ instead of the observed $\mathbf{z}_i$.

As another (untried!) possibility, suppose

$$E(d_i|\mathbf{z}_i, \tilde{\mathbf{a}}_i) = E(d_i|\mathbf{z}_i, \tilde{\mathbf{a}}_i, i \in S) = E(d_i|\mathbf{z}_i, i \in S) = \mathbf{z}_i^\top \boldsymbol{\theta}_{q\times 1}$$

$$E(\mathbf{z}_i^\top|\tilde{\mathbf{a}}_i) = E(\mathbf{z}_i^\top|\tilde{\mathbf{a}}_i, i \in S) = \tilde{\mathbf{a}}_i^\top \boldsymbol{\beta}_{p\times q}$$

We have then,

$$E(d_i|\tilde{\mathbf{a}}_i) = E\big(E(d_i|\mathbf{z}_i, \tilde{\mathbf{a}}_i)|\tilde{\mathbf{a}}_i\big) = E(\mathbf{z}_i^\top \boldsymbol{\theta}|\tilde{\mathbf{a}}_i) = E(\mathbf{z}_i^\top|\tilde{\mathbf{a}}_i)\boldsymbol{\theta}$$

$$= \tilde{\mathbf{a}}_i^\top \boldsymbol{\alpha} \qquad\qquad [\boldsymbol{\alpha} = \boldsymbol{\beta}\boldsymbol{\theta}]$$

However, we cannot estimate $\boldsymbol{\alpha}$ based on A directly, since

$$E(d_i|\tilde{\mathbf{a}}_i) \neq E(d_i|\tilde{\mathbf{a}}_i, i \in A)$$

But one can use

$$\xi_i = \tilde{\mathbf{a}}_i^\top \mathbf{b} \qquad \mathbf{b} = \Big(\sum_{i\in S} \tilde{\mathbf{a}}_i \tilde{\mathbf{a}}_i^\top\Big)^{-1}\Big(\sum_{i\in S} \tilde{\mathbf{a}}_i \mathbf{z}_i^\top\Big)\Big(\sum_{i\in S} \mathbf{z}_i \mathbf{z}_i^\top\Big)^{-1}\Big(\sum_{i\in S} \mathbf{z}_i d_i\Big)$$

$A = \text{SPD}$, $S = \text{PCS}$: Violation of assumption (ii)

Record linkage methods:

- Deterministic: uniques matches on chosen key variables
- Probabilistic (Fellegi & Sunter, 1969; Herzog et al., 2007):
  $A \times B = M \cup U$, $M =$ Matched pairs, $U =$ Unmatched pairs
  Likelihood Ratio Test $H_0 : (a, b) \in M$ vs. $H_1 : (a, b) \in U$
- Bayesian 'latent entity' formulation (e.g. Stoerts et al., 2016)

Not a major issue in BNU-network; otherwise violation the assumption (ii), if no. false links $\neq$ no. missing links

- Ding and Fienberg (1994): one-direction linkage
- Di Consiglio and Tuoto (2015): both-direction linkage
- See Tuoto et al. (2018) for a recent, comprehensive discussion

[1] Abbott, O. (2009) *2011 UK Census Coverage Assessment and Adjustment Methodology.* ONS.

[2] Böhning, D., Van der Heijden, P.G.M. and Bunge, J. (2017). *Capture-Recapture Methods for Social and Medical Sciences*. Chapman and Hall/CRC.

[3] Chipperfield, J., Brown, J. and Bell, P. (2014). Estimating the count error in the Australian Census. *Journal of Official Statistics*, **33**, 43-59.

[4] Di Consiglio, L. and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, **31**, 415-429.

[5] Ding, Y. and Fienberg, S.E. (1994). Dual system estimation of Census undercount in the presence of matching error. *Survey Methodology*, **20**, 149-158.

[6] Fellegi, I.P. and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183–1210.

[7] Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, **59**, 409–439.

[8] Griffin, R.A. (2014). Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020. *Journal of Official Statistics*, **30**, 177-189.

[9] Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). *Data Quality and Record Linkage Techniques*. Springer.

[10] Hogan, H. (1993). The Post-Enumeration Survey: Operations and results. *Journal of the American Statistical Association*, **88**, 1047–1060.

[11] Nirel, R. and Glickman, H. (2009). Sample surveys and censuses. In *Sample Surveys: Design, Methods and Applications, Vol 29A (eds. D. Pfeffermann and C.R. Rao)*, Chapter 21, pp. 539-565.

[12] ONS-M8 (2013). *Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory*. `https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme/reportsandpublications`

[13] Renaud, A. (2007). Estimation of the coverage of the 2000 census of population in Switzerland: Methods and results. *Survey Methodology*, **33**, 199–210.

[14] Tuoto, T., Di Consiglio, L. and Zhang, L.-C. (2018). Capture-recapture methods in the presence of linkage errors. In *Analysis of Integrated Data, eds. L.-C. Zhang and R-L Chambers*. Chapman & Hall/CRC. *To appear.*

[15] Stoerts, R., Hall, R. and Fienberg, S. (2016). A Bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, **111**, 1660-1672.

[16] Wolter, K. (1986). Some coverage error models for census data. *Journal of the American Statistical Association*, vol. **81**, pp. 338-346.

[17] Zhang, L.-C. (2018). A note on dual system population size estimator. *Journal of Official Statistics, to appear.*

[18] Zhang, L.-C. and Dunne, J. (2017). Trimmed Dual System Estimation. In *Capture-Recapture Methods for the Social and Medical Sciences, eds. D. Böhning, J. Bunge and P. v. d. Heijden*, Chapter 17, pp. 239-259. Chapman & Hall/CRC.