

Sampling Subpopulations in Multi-Stage Surveys

Robert Clark, Angela Forbes, Robert
Templeton



This research was funded by the Statistics NZ Official Statistics Research Fund 2007/2008, and builds on the NZ Health Survey 2006/2007 sample design conducted for the NZ Ministry of Health.

Outline



- ❑ Surveying Rare Populations
- ❑ Snowball Sampling and Intercept Point Surveys
- ❑ Screening:
 - Proxy screening of households
 - Accuracy of proxy screening
- ❑ Disproportionate Sampling
 - Optimal one-stage and two-stage allocations
 - Intercensal mobility
- ❑ Dual Frame using the Maori Electoral Roll
- ❑ ABS Findings on Sampling Indigenous Australians
- ❑ Conclusions

Surveying Subpopulations



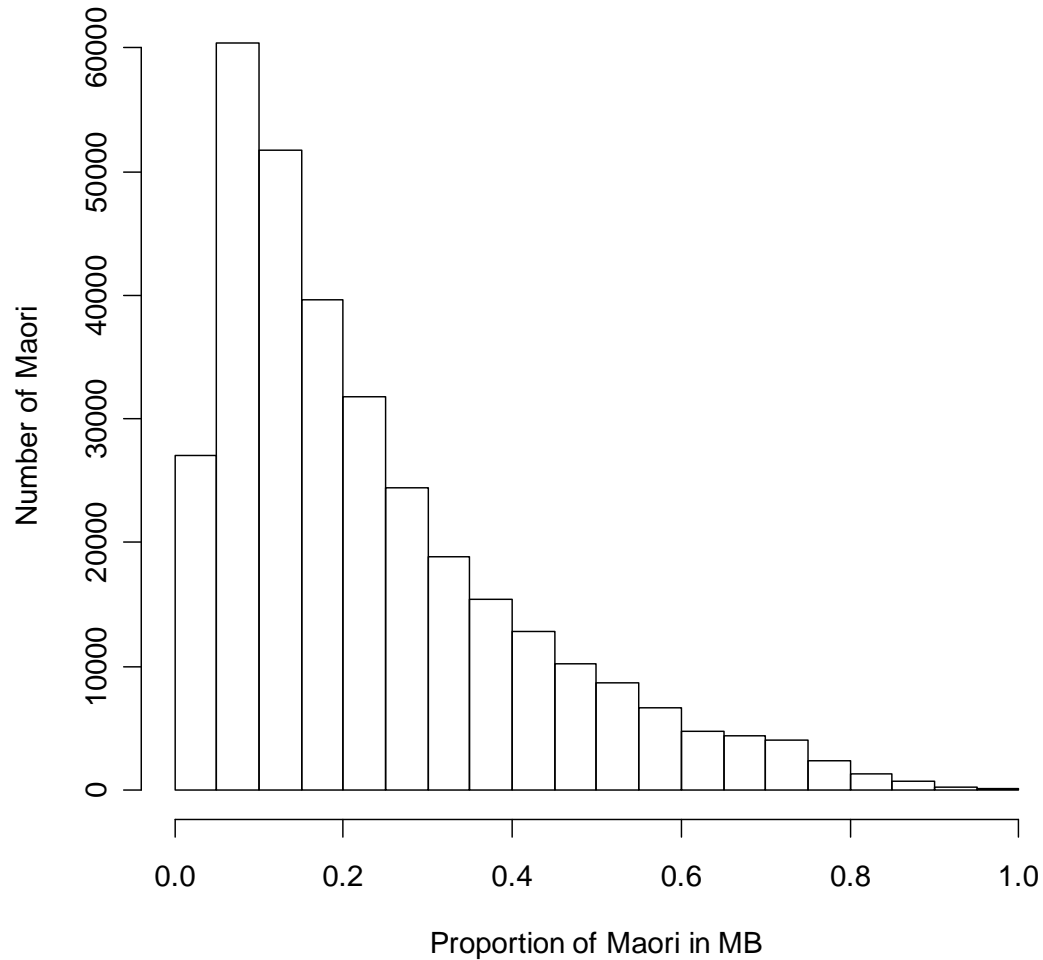
- Group of interest is a relatively small subset of the population.
- No reliable list of the subpopulation.
- Common problems:
 - not highly geographically clustered;
 - over-surveyed?
 - mobile population
 - frequent identification errors / variability

Example: Maori Population



- ❑ Maori comprise about 12% of the adult population
- ❑ 60% of Maori live in Meshblocks (primary sampling units containing about 50 dwellings) where the proportion of Maori is 20% or less.
- ❑ New Zealand Health Survey 07-08: equal probability would give approx 1500 Maori in sample, more like 3000 are needed
- ❑ best possible outcome for Maori sample
 - Disproportionate allocation according to MB density → simple random sampling (SRS) + 15.9%

Distribution of Proportion of Maori in Meshblocks



Snowball Sampling



- ❑ Select sample of people;
- ❑ Ask subpopulation members to identify others among their acquaintance;
- ❑ Advantage: don't need to contact as many people to achieve the same number of subpopulation members in sample.
- ❑ Disadvantages:
 - Can be biased towards people with more friends?
 - unbiased estimation possible provided that everyone is linked to others;
 - Subpopulation members need to know each other;
 - Image problem for government?

Intercept Point Survey



- Example: sample homeless population by selecting individuals visiting selected soup kitchens (*"aggregation points"*) at selected times.
- At each location, individuals are asked how often they visit this and other aggregation points.
- Very cost-efficient, but biased, possibly extremely so.

- ❑ McKenzie and Mistaien, "Surveying Migrant Populations: A Comparison of Census-Based, Snowball and Intercept Point Surveys" (2008; Journal of the Royal Statistical Society Series A, cond.accepted)
- ❑ Group of interest: Japanese-Brazilian families (about 0.9% of population)
- ❑ Snowball Survey: poor response rate; most respondents did not want to provide referrals
- ❑ Snowball and intercept point surveys selected individuals more tied to the Nikkei community.
- ❑ Intercept point can be useful for exploratory investigation but snowball not much cheaper than probability sampling.

Screening



- ❑ Not so much a method as the absence of a method.
- ❑ Select a large sample of people and identify whether they belong to the subpopulation.
- ❑ Conduct the survey on all identified members.
- ❑ Important to make screening as cheap as possible per household or person!
- ❑ If the initial identification is subject to error, take a subsample of the (apparent) non-members (two-phase):
 - initial screen needs to be much cheaper than the second phase costs (6:1 or better) and
 - screening needs to be quite accurate (at least 75% of the subpopulation classified to stratum a).

(Kalton and Anderson, 1986, quoting Deming, 1977).

Proxy Screening



- A number of NZ surveys, including the NZ Health Survey, have improved the efficiency of screening by:
 - Each PSU has a main sample and an oversample.
 - Collect household information from any contacted adult in selected households, including ethnicity and age.
 - In the main sample, one adult is selected at random. In the oversample, one (apparently) eligible adult is selected at random.

Table 1: Survey Eligibility by Screener Eligibility (Core Only)

| | | Screener Proxy Report | | |
|-----------------------|--------------|-----------------------|----------|--------|
| | | Not Eligible | Eligible | Total |
| <i>frequency</i> | | | | |
| <i>row percent</i> | | | | |
| <i>column percent</i> | | | | |
| Survey Direct Report | Not Eligible | 6830 | 111 | 6941 |
| | | 98.40 | 1.60 | 76.45 |
| | | 95.11 | 5.85 | |
| | Eligible | 351 | 1787 | 2138 |
| | | 16.42 | 83.58 | 23.55 |
| | | 4.89 | 94.15 | |
| | Total | 7181 | 1898 | 9079 |
| | | 79.09 | 20.91 | 100.00 |

| | Misclassification Rate (%) | |
|---|----------------------------|----------------------------|
| Ethnicity | Single Person Households | Multiple Person Households |
| Maori | 18.0 | 21.5 |
| Pacific | 15.4 | 7.6 |
| Asian | 16.5 | 10.6 |
| Maori, Pacific or Asian | 17.2 | 15.3 |
| <i>Two or More of Maori, Pacific or Asian</i> | <i>21.6</i> | <i>8.1</i> |

Table 5: Alternative Designs (Equal-Sized Main Sample and Oversample)

| | | (1a) | (1b) | (1c) | Design (2a) | (2b) | (2c) | (2d) |
|-----------------------|--------|-------|-------|-------|----------------|-------|-------|-------|
| Households Approached | | 13333 | 17670 | 17892 | 13333 | 17392 | 17821 | 18147 |
| People Surveyed | Total | 13333 | 11165 | 11054 | 13333 | 11304 | 11089 | 10927 |
| | Subpop | 3516 | 4660 | 4355 | 3056 | 4602 | 4072 | 3821 |
| | Maori | 1746 | 2314 | 2096 | 1642 | 2549 | 2173 | 2002 |
| Design Effect | Total | 1.00 | 1.10 | 1.09 | 1.21 | 1.30 | 1.28 | 1.25 |
| | Subpop | 1.00 | 1.00 | 1.07 | 1.23 | 1.27 | 1.33 | 1.29 |
| | Maori | 1.00 | 1.00 | 1.08 | 1.22 | 1.23 | 1.36 | 1.32 |
| Effective Sample Size | Total | 13333 | 10177 | 10141 | 11041 | 8687 | 8669 | 8730 |
| | Subpop | 3516 | 4660 | 4088 | 2483 | 3624 | 3055 | 2962 |
| | Maori | 1746 | 2314 | 1935 | 1351 | 2077 | 1594 | 1521 |

(1a): Main Sample Only, SRS of People

(1b): Main Sample & Oversample both SRS of People assuming Perfect Screening;

(1c): Main Sample & Oversample both SRS of People with Proxy Screener;

(2a): Main Sample only, SRS of Households, One/Household;

(2b): Main Sample & Oversample both SRS of Households (Screen then Subsample) with Perfect Screening;

(2c): Main Sample & Oversample both SRS of Households (Screen then Subsample) with Proxy Screener;

(2d): Main Sample & Oversample both SRS of Households (Subsample then Screen) with Proxy Screener;

Could we ever just omit the main sample?



Table 7: Bias for Just Māori

| Variable | Mean of Variable | | | | p-value |
|-------------------------|------------------|-----------|--------------------------------|--|---------|
| | Core | Screeener | Eligible according to Screener | Ineligible according to screener (272 cases) | |
| Obesity | 0.382 | 0.445 | 0.450 | 0.291 | 0.000* |
| Smoker | 0.413 | 0.431 | 0.434 | 0.380 | 0.1446 |
| Visited GP in Past Year | 0.787 | 0.794 | 0.793 | 0.781 | 0.6929 |
| Diabetes | 0.056 | 0.060 | 0.064 | 0.039 | 0.0520 |
| Asthma | 0.150 | 0.158 | 0.152 | 0.162 | 0.7328 |

* CI of difference: 0.094-0.224

K&A(1986): Optimal One-Stage Allocation (Subpopulation Mean)



- Let N_k be population in stratum k
- Let φ_k be proportion of stratum k who are in subpopulation
- Let π_k be probability of selection for people in stratum k . Then:

$$E[n(\text{subpop})] = \sum_k \pi_k N_k \varphi_k$$

- BUT, there is a penalty from using unequal π_k .
This leads to the variance being multiplied by:

$$\begin{aligned} deff &= 1 + RV\left(\pi_i^{-1} : i \in \text{sample} \cap \text{subpop}\right) \\ &\approx \left(\sum_k N_k \varphi_k \pi_k^{-1} \right) \left(\sum_k N_k \varphi_k \pi_k^{-1} \right) / \left(\sum_k N_k \varphi_k \right)^2 \end{aligned}$$

- Cost = $C_1 n + C_2 n_{\text{sub}}$
where C_1 =cost per screen, C_2 =cost per interview
- Variance proportional to $n_{\text{sub}}^{-1} / \text{deff}$
- Minimize variance for fixed cost:

$$\pi_k \propto \sqrt{\varphi_k / (C_1 + C_2 \varphi_k)}$$

Comparison of Targeting Strategies

All Designs Cost Equivalent with $C_1/C_2=0.4$

| π_k prop to: | # screened | Sample size (eligible) | Deff | Effective sample size (eligible) |
|---|------------|---------------------------|-------|---|
| Constant | 14,514 | 1,695 | 1.00 | 1,695 |
| Sqrt(φ_k) | 13,187 | 2,225 | 1.19 | 1,867 |
| $\sqrt{\varphi_k / (C_1 + C_2\varphi_k)}$ | 13,566 | 2,073 | 1.09 | 1,895 |
| φ_k | 11,848 | 2,761 | 2.00 | 1,383 |
| φ_k^2 | 9,710 | 3,616 | 13.78 | 262 |

Optimal Two-Stage Allocation



- Select a sample of primary sampling units with some probabilities;
- Select a sample of people from PSUs and screen them;
- Select a subsample of eligibles and a subsample of ineligibles.
- $\text{Cost} = C1 \#PSUs + C2 \#approached + C3 \#interviewed$
- Trade-off between cost and variance

- If screen perfectly accurate, and subpopulation means are the only objective, then:
 - Select PSUs with probability proportional to density times population;
 - Sampling fraction within PSU for screening proportional to

$$1 / \sqrt{\varphi_g (C_2 + C_3 \varphi_g)}$$

- i.e. over-target high-concentration PSUs, but then under-sample within them!

Intercensal Mobility



- The optimal designs assume that the concentration of subpopulation members is known exactly for every PSU. In practice, out of date census data is used →
 - Designs less efficient than they appear;
 - A less targeted design would be appropriate: use $E[\text{density}|\text{census data}]$ rather than census-density.
- Over 50% of New Zealanders change addresses over a five year period.
- Correlations between 2001 and 2006 densities:
 - Meshblocks: 0.911
 - PSUs: 0.939
 - Territorial Authority: 0.997

Comparison of Designs based on 2001 Census Data
Cost Fixed at 12500, C1=2, C2=0.3, C3=1
rho=0.05

| Design | SE(%) in 2001 | SE(%) in 2006 <i>(simulated)</i> | Undercoverage in 2006 (%) <i>(simulated)</i> |
|--|------------------|--|--|
| Using 2001 MB densities unadjusted | 1.022 | 1.046 | 1.77(?) |
| Assume ≥ 1 Maori per MB | 1.046 | 1.087 | 0.00 |
| Shrinkage estimate of MB density | 1.040 | 1.091 | 0.00 |
| Shrinkage estimates of MB density and total population | 1.042 | 1.058 | 0.00 |

Dual Frame using Maori Electoral Roll?



- Available addresses from the NZ Health Survey sample were matched to the Maori electoral roll
- Thus we had a sample of addresses, and for each address:
 - Did a Maori adult live at the address (Y/N) (as measured by NZHS)
 - Did a Maori adult live at the address according to the Maori electoral roll?

□ Results:

- In urban areas, approximately 85% of Maori in the matched sample lived in an address found on the electoral roll. Of addresses on the roll, 77% would be found to have a Maori resident by the survey.
- Results were less good for rural areas, partly due to more ambiguous addresses.

Sampling the Australian Indigenous Population: Some ABS Findings



- ❑ From Working Paper, “Sample Design Issues for National Surveys of Aboriginal and Torres St Islander Populations” (Alistair Rogers and Geoffrey Brent), www.abs.gov.au
- ❑ Indigenous Australians about 2.3% of the total population of Australia (less at household level). 24% live in remote areas (vs 3% general population).
- ❑ “At regional levels, many indigenous populations can be summarised as either
 - geographically clustered and relatively inaccessible, or
 - relatively accessible but geographically diverse.”

- Many split-meshblocks (SMBs) had zero Indigenous population in census. Some options:
 1. Exclude them. Leads to unacceptable undercoverage due to intercensal changes;
 2. Give them a reduced probability of selection;
 3. Make use of SMB and CD census numbers, to exclude some size-0 SMBs, such that a conservative estimate of undercoverage was less than 5% in each region. This led to very substantial savings (>30% reduction in screening in some areas).
- A combination of (2) and (3) was used.

Conclusions



- ❑ Multiplicity sampling and Intercept Point surveys seem tempting but generally only good for indicative information.
- ❑ A rough cost-variance approach can lead to improved efficiency of the order for Maori sampling. New two-stage allocations can yield modest further gains.
- ❑ Proxy screening gives some gains but under-identification can quickly degrade these.
- ❑ Intercensal mobility can be ignored to some extent.
- ❑ Maori electoral roll shows promise.
- ❑ For rarer populations such as Indigenous Australians, an ABS study suggests a combination of (roughly) optimal allocation and limited intentional undercoverage (<5%).
- ❑ The main thing is to avoid over-targeting.

- www.cssm.uow.edu.au
- www.uow.edu.au/~rclark/talks.html