WP. 14
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (ii): Disclosure risk, information loss and usability of data

A 'MICRODATA FOR RESEARCH' SAMPLE FROM A NEW ZEALAND CENSUS

Supporting Paper

Submitted by Statistics New Zealand[1]

[1] Prepared by Mike Camden (mike.camden@stats.govt.nz).

A 'Microdata for Research' sample from a New Zealand census

Mike Camden*

* Statistical Methods, Statistics New Zealand, PO Box 2922, Wellington, New Zealand, mike.camden@stats.govt.nz.

Abstract: Statistics New Zealand has provided researcher access to many unit record datasets since 1995 in its three internal data laboratories. It began a programme to produce licensed Confidentialised Unit Record Files (CURFs) from social surveys and censuses in 2004, and it is currently investigating remote access, synthetic datasets and other ways for enabling access to microdata for research.

We will focus on our new 2% census sample CURF, which has just been pilot-tested by a set of New Zealand researchers. We will assume that cell count has a hypergeometric distribution, and use this to quantify sampling error, (a measure of both information loss and usability), justify the sample size and assess disclosure risk from the sample. Our sampling method preserves almost exact census proportions for a few census variables, and decreases sampling error for others. We will outline this method and its effects on cell counts and proportions.

We will summarise Statistics New Zealand's future plans for meeting the growing demand for microdata for research, in a country of four million people.

1. Introduction.

In September 2005, Statistics New Zealand released a 2% sample from its 2001 Census to a set of ten researchers. The researchers have been asked to comment on the two core issues in microdata for research: usability and disclosure risk. A few of the researchers were specifically asked to test the disclosure risk by behaving as intruders. This pilot is part of a larger program of licenced Confidentialised Unit Record Files (CURFs).

The 2001 New Zealand census dataset contains 3.8 M records. This population size raises disclosure risk issues of uniqueness and confidentiality. It also raises usability issues for a small sample from it. To help us with decisions involving these issues, we considered cell counts (or proportions) in tables, and used the hypergeometric distribution to predict the behaviour of these counts. This helped us to calculate sampling errors, and hence to decide the sample size. In fact, the sample selection method produces sampling errors, for some tables, that are smaller than the

hypergeometric ones. We are in the process of analysing how well the model fits the real behaviour of the sample.

The population is diverse in ethnicity, origin, income and employment. This raises the question of how to lessen uniqueness and minimise damage to usability. Considerable effort has been put into the concepts and practice of confidentiality for census data, including the application of appropriate rules for micro- and aggregate data (Statistics New Zealand, 2005).

An examination of the issues (Dunlop, 2004) recommended that we should proceed with extreme caution to produce a pilot CURF.

We will use the researchers' feedback in deciding the future for licenced CURFs from censuses. It will help us decide which variables to include, what sample sizes to use, what sampling methods to use and which of our five-yearly censuses to use.

2. The provision of microdata for research in New Zealand.

In 2004, the New Zealand government completed a review of the official statistics system. This gave Statistics New Zealand a leadership role in collection and storage of datasets, and dissemination of information from them. The Annual Report 2004 (Statistics New Zealand, 2004) states: 'users will be able to use a variety of standard methods' to access information.

The Statistics Act (New Zealand, 1975) obliges Statistics New Zealand to protect confidential information from people and businesses, hold it securely, limit use to statistical purposes and prevent disclosure of identifiable information. The need to preserve respondent trust implies similar obligations.

In response to these two needs, Statistics New Zealand currently is actively extending its provision of microdata for research. Three methods of provision are detailed below. In each, it seeks to provide access to microdata, within its legal and contractual constraints.

2.1 The Data Laboratory.

Researchers (academic, government or private) make an application, stating data needs, proposed outcomes and how the research will contribute to the improvement of official statistics. If the application is approved, Statistics New Zealand prepares a customised dataset with identifiers removed, arranges access in the data laboratory room, and checks all output. The three data laboratory rooms are in the Statistics New Zealand offices in Auckland, Wellington and Christchurch. This process has been running since 1995, with census datasets often being used. Sensitive variables,

like geographic and household ones, are supplied where the need is clear and the disclosure risk allows.

Government departments can apply to access data in their offices if they can demonstrate a secure environment similar to that of the data laboratories. Statistics New Zealand audits these secure environments regularly.

2.2 The Remote Access system.

This was trialled in 2004, on a modified version of the New Zealand Income Survey dataset, and decisions are pending. A modified version of a dataset is prepared, and kept secure. Researchers send in SAS code and receive outputs. Both are checked automatically, and are audited by our staff. Software written by the Australian Bureau of Statistics is used. The trial was with a social survey dataset, but census datasets could be used.

2.3 The CURF programme.

The programme began in 2004. CURFs are issued to researchers, who sign a licence agreement. Datasets are modified carefully, and sent out on a CD. So far, we have issued CURFs for the New Zealand Income Survey 2002 and 2003. We are working towards issuing a joint CURF for the Income Survey and Household Labour Force Survey 2004. The samples for these CURFs contain about 28,000 records each. CURFs for further socio-economic surveys are in preparation or planning. The CURFs do not contain sample design variables, like stratum, and hence we attach datasets with 100 replicate weight variables.

3. Size and contents of this census CURF.

The pilot CURF has 33 variables, of the approximately 100 output variables available, and 76,415 (2%) of the 3,820,749 records available. All the CURF variables are categorical, and most of them have categories collapsed from their original versions. The variables are about demographics, residence, ethnicity, origin, income and employment. Variables dealing with geographic location and household structure are omitted from the CURF.

The two drivers of CURF design are the needs to maximise usability and minimise disclosure risk.

4. Selection of the 2% sample size.

Several issues influenced our choice of sample size. They are outlined below. Given these considerations, for New Zealand 1% is too small and 3% is too large. The

sampling method means that a whole-number percentage is more convenient. So the conclusion, for this pilot, is clear!

4.1 Overseas practice

Many other countries produce the equivalent of CURFs, some licensed and some more freely available. We considered several countries, all with populations much larger than ours, and observed small sample proportions: 1% to 5%. We decided to be conservative, and like them, to aim for a small proportion.

4.2 The relationship of disclosure risk to sample size.

Most types of disclosure risk depend on the number of records. It is reasonable then to assume that much disclosure risk increases linearly with sample proportion.

4.3 The relationship of usability to sample size.

Many types of output have a sampling error which decreases with the square root of sample size. These types include cell counts and proportions, and regression coefficients (this assumes simple random sampling (SRS)). Sampling error is a measure of one form of information loss, and an inverse measure of usability. As sample proportion increases, usability increases, but with a square root law of diminishing returned.

4.4 Existing sample surveys and CURFs.

We already have two large ongoing surveys, with nearly 1% of the population in each. The first is the Income Survey/Household Labour Force Survey, and CURFS are being produced from this. The second is the more recent Survey of Family, Income & Employment (SoFIE). Both contain much more socio-economic data per person or household than the census. So we needed a sample bigger than these.

5. Disclosure risk and uniqueness.

5.1 Types of disclosure risk.

The CURF will go to licensed researchers, some of whom may be research students. It should not, but could, fall into the hands of other persons. We'll assume that some of these recipients may behave as 'intruders'.

These events are possible:

- A researcher spontaneously recognises someone who is unique in the CURF and in the population, on a small number (3-5) of identifying demographic variables.

- An intruder hunts for a person or type, on a larger number of demographic variables.

- A researcher or intruder finds what they think is their own (or a neighbour's) record, because it appears to be unique on all or most of the 33 variables, and these values match their own values.

- A rogue researcher or intruder links this dataset with another one using software.

The risk from any of these events increases with sample size, as well as with level of detail in the variables.

5.2 Quantifying uniqueness.

We carried out all the processing of variables on the entire census dataset, and drew the sample at the end. Hence we are able to look for population and sample uniqueness, using our 33 variables. Children (22.2%) and visitors from overseas (2.2%) have many structural missing values, so we will examine adult New Zealand residents.

Using these 2.9 M people, and using all our 33 variables, a high percentage of us (74.4%) are population uniques. For the 2% sample, the conditional probability that a person is a population unique, given that the person is a sample unique, is 81.3%. The level of uniqueness, and the size and behaviour of the conditional probability, are further reasons to keep the sample proportion small.

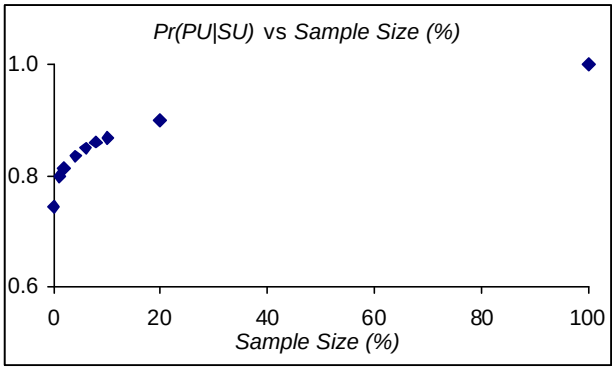

Fig 1 The conditional probability Pr(Population Unique | Sample Unique) rises with sample size.

6. Methods for limiting risk, and their impacts on usability.

To lessen all the types of risk listed above, we used the methods below. There is further protection in the licence agreement that limits use and distribution.

6.1 Omission of household and location variables

We omitted all variables dealing with family and household structure, and with geographic location. Both these sets of variables are very useful to researchers, and their non-selection is an important form of information loss. Responses to the pilot may suggest that we provide a location variable in future CURFs. We may need to balance this with further collapsing or omissions.

6.2 Collapsing of categories.

We aimed to minimise risk and preserve usability. For each variable included, we examined the univariate distribution. Where categories had about 1% or less of the dataset, they were combined with others. When possible, new categorisations followed existing classifications. Most variables have similar proportions in each category. We aimed to use categories that would be useful to researchers. For example, people who cycled to work on Census Day 2001 (1.10%) were put with others to form "bicycle, walked or jogged" (3.60%). Careful design of the classification is a way of both limiting uniqueness and minimising the loss of usable information.

AgeGroup	Percent (population)	Percent (CURF)
0-4 Years	7.13	7.13
5-14 Years	15.16	15.16
15-19 Years	7.08	7.08
20-24 Years	6.53	6.52
25-34 Years	14.22	14.22
35-44 Years	15.50	15.50
45-64 Years	22.24	22.24
65 Years and Over	12.14	12.14

Table 1 In the original dataset, age is in years. The CURF variable *AgeGroup* has eight categories. They match life stages, and are all above 5%. This categorisation aims to preserve useful detail, and remove other detail. The population and CURF percentages are almost the same, due to the sampling method.

We regard *IncomeGroup* as the target variable that an intruder might want to find for a person they have recognised. This has 8 categories, with the top one being $70,001 (€40,000)/year upwards and having 3.43% of the population. The categorisation substantially lessens the variable's usefulness to an intruder, while it minimally lessens the usefulness to a researcher.

6.3 Special Uniques analysis.

This process applied the ideas of Elliot et al (2002). We selected a subset of 14 variables that were considered very identifying: ie likely to be known about a neighbour or colleague. One of the variables we used was the five binary ethnicity variables combined into one variable with 32 categories. We took *Sex*, *AgeGroup* and every combination of three of the remaining 12 variables, and marked the records that showed up as uniques.

This process adds a variable (number of occurrences as a unique), and also shows which variables produce them. We decided to treat the 15,000 records (of the 3.8 M) that had two or more occurrences. We set up rules for modifying the value of the "worst" variable for each of these, ran seven iterations of this process and reduced the number of these 'special unique' records from 15,000 to 2,800.

This process reduces risk, but changes a tiny proportion (0.014%) of the values in the dataset. These values are replaced by neutral categories (like NEI), and not by wrong values. The process therefore produces minimal information loss.

7. The Sampling Method and its Consequences.

Our sampling method was controlled on three variables. New Zealand is divided into 1,860 'Area Units', which contain on average 2,100 people, but vary widely in size. We added a random-number variable, and sorted the census dataset by *Sex*, *AgeGroup*, *AreaUnit* and the random-number variable. We divided this sorted dataset into groups of 100 records, and sampled two records from each group. *AreaUnit* is not included in the CURF.

The sort on the three named variables divides the census dataset into about 29,000 cells, with an average of about 130 people in each. Some cells are much smaller. The cells are homogeneous on at least the three named variables. Neighbouring cells are usually similar, and hence most of the groups of 100 records are homogeneous. We plan to investigate the effect of small cells on the value of this sampling method.

We can distinguish three types of variable: C: Controlled: *Sex*, *AgeGroup* (and *AreaUnit*); D: Dependent on these Controlled variables: ranging from highly dependent to slightly dependent; I: Independent of these Controlled variables. There are probably no completely independent variables, but they would form a worst case, and hence their properties need to be examined.

This distinction has some use, as it affects sampling error and hence information loss and usability. A sampling method that limits sampling error for some variables is of value, as it increases usability.

For the Type C variables, cell counts are extremely close to 2% of the population cell count; they are about ± 1 person away. The sample is an extremely close image of the population, by sex, age and geographical location. Other types, and combinations of them, are discussed below.

8. The distribution of cell counts under independence

We assume that some researchers will make frequency tables using one, two or more of the variables. Each cell in these tables will have a population count k, which remains unknown to CURF users. It will have a sample count x from the CURF, and a sample proportion p. We assume here that some variables are independent from the three controlled variables. This gives us a 'worst case'; other cases will usually have less variation.

If we assume that the CURF behaves like a simple random sample (without replacement) of n people from N people, then x is the number of people who are both in this cell for the population and in the sample. We will treat x as having a hypergeometric distribution, with parameters (N, k, n). (In fact, if a table has c cells, then only c-1 of the x-values can be independent, but c is large for most tables.)

There are two convenient approximations. If n/N is small, then x will be approximately binomial, with parameters (n, k/N). If k/N is small too, then x will be approximately Poisson, with parameter nk/N. These give simple expressions for the standard deviation of sampling error, for counts (sx = v(kn/N) and proportions (sp = v($k/(nN)$). In fact, n is small (2%), and k/N) is small for most cells of interest to researchers.

All three models give the law of diminishing returns: sampling error decreases with the square root of sample size.

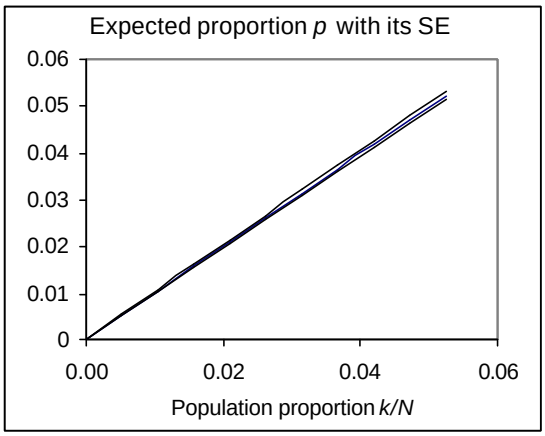

Fig 2: The graph relates size of sampling error for p, for a 2% sample, to the quantity being estimated. If a cell has a population proportion k/N, we locate this value on the x-axis, and move upwards to see its CURF proportion and sampling error (shown as ±1 standard deviation). Even in this this 'worst case', the 2% sample gives relatively small errors. If k/N is say 5%, then the error is ±0.08%.

9. Sampling error and the controlled variables.

With three types of variables (C, D, I), there are 2^3 (= 8) types of combinations. The sampling error for D and combinations of two or more types will usually be between the best case (C) and the SRS case (I).

For Type D variables, and for combinations of C and I variables, standard error will usually be smaller than for SRS. Unfortunately, for cells of practical usefulness, the improvement is small. The expected behaviour can be studied analytically, but we will instead graph examples from the CURF.

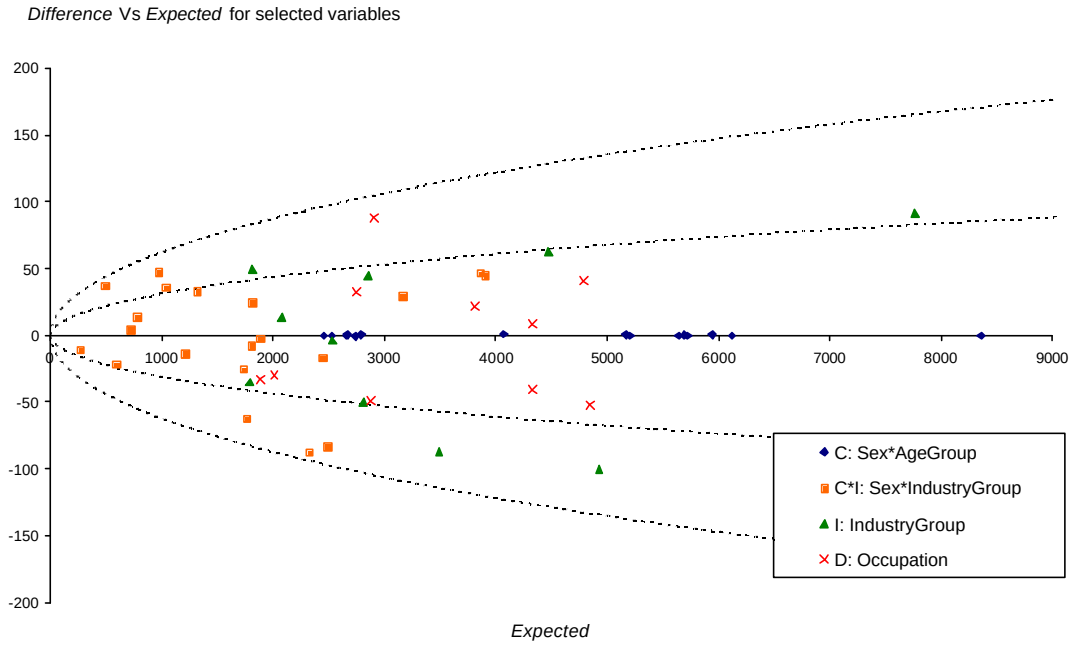

Fig 3 *Expected* is 2% of the population cell count (= *kn/N*), and *Difference* is CURF cell count minus *Expected*. The curves show ±1 and ±2 standard errors, assuming independence and hypergeometric behaviour. The relationship is shown for the cells of four tables that exemplify types C, I, D and C with I. We used *IndustryGroup* as an example of Type I, as it is weakly related to Type C variables. The type C table shows very small values for *Difference*. The other tables have behaviour consistent with, or slightly less variable than, the hypergeometric distribution.

10. Acknowledgements

I wish to acknowledge the work and insights of these Statistic New Zealand Colleagues: Lisa Corscadden, Jamie Pou, Tiri Sullivan, Zoe Wood, Jamas Enright and Jason O'Sullivan (all, at the time, in Statistical Methods), and Adele Quinn, Adam Bedford and Peter Lafferty (Population and Census).

11. Conclusions

Statistics New Zealand already makes information from the 2001 Census freely available via its website, in the form of tables, with counts random rounded to base three. Tables can contain geographic and household variables. This pilot CURF enables a constrained group of people to produce a range of tables that is more limited in some ways and wider in other ways. Counts in these tables, if weighed up, would resemble counts random rounded to base 50. These researchers could perform any other analyses applicable to categorical variables. The CURF is a new form of access, with its own limitations and advantages.

We will use the feedback from the researchers who have trialled the pilot to consider new balances between disclosure control and usability, in possible future census CURFs.

References

Dunlop, A. (2004). "Census CURFs in New Zealand: an examination of the issues". Internal paper. Statistics New Zealand.

Elliot, M.J., Manning, A.M. & Ford, R.W. (2002). "A computational algorithm for handling the special uniques problem", *International Journal of Uncertain.ty, Fuzziness and Knowledge-Based Systems*, 5:10, 493-509.

New Zealand. (1975). Statistics Act 1975

Statistics New Zealand (2004). *Annual Report of the Government Statistician for the year ended 30 June 2004*, Statistics New Zealand, Wellington.

Statistics New Zealand (2005). "2006 census confidentiality rules", http://www.stats.govt.nz/census/2006-census/methodology-papers/confidentiality-rules.htm (3 Oct 2005)