Weighting procedure and variance estimation for the 2004 Living Standards Survey

Alistair Gray *

23 September 2004

Abstract

This paper describes in detail the weighting for the 2004 Living Standards Survey (LSS) and the method of estimating the sampling errors.

The early versions of this paper were written as proposals for weighting. In particular, how to account for the different sampling schemes for the primary sampling units; how to estimate the Economic Family Units; what auxiliary data to use for calibration.

As in usual in a complex survey, the choice of estimation and variance estimation was an iterative process. The final decision was made by the author and some of the LSS team: John Jensen, Sathi Sathiyandra and Matt Spittal.

This version of the paper documents the final estimation method, but also gives the intermediate steps leading to the final method.

Keywords: Multistage sample design, Nonresponse, Calibration, Delete-a-group Jackknife.

*Statistics Research Associates Ltd, PO Box 12 649, Thorndon, Wellington, New Zealand. email: `alistair@statsresearch.co.nz`

Contents

List of Tables

1 Design Weights

The approach to calculating the sample design weights for the LSS is as follows.

The major issue was how to deal with nonresponse and the initial approach required knowing which households belonged to the second stage cluster (half AU map). The adopted approach does not need that information.

The sample design weights are (ignoring nonresponse) the inverse of the probabilities of selection. These weights are often called simple expansion weights or rate-up weights. Since this is a multistage design, it is usual to calculate the conditional probabilities of selection at each stage.

The population was stratified into 54 TNS regions and the sample size was allocated proportional to the population in the TNS region. The number of AUs sampled in each region was calculated using the expected number of interviews.

1.1 1st stage: selecting Area Units

The primary sampling units (psus) in this design are Area Units (AUs). The initial selection of AUs were by mistake selected systematic probability proportional to size without replacement (sppswor). Interviewing had taken place in 77 of these before this was realized. Interviewing continued in these and they were considered to be self-representing psus: i.e. selected with probability 1. This is some what contentious since it is usual to make psus self-representing based on their special characteristics. These AUs were in the main confined to major urban areas, where many AUs would be selected. So the impact on the overall design was not thought to be great.[1]

A further 445 AUs were selected from the remaining AUs simple random sample without replacement (srswor). Note that the overall sampling fraction of AUs (521 from 1678) is very high.

Let n_h be the number of AUs selected srswor in the TNS region h. Let N_h be the total number of AUs in region h less the AUs selected by systematic ppswor. Probability of selecting the ith AU is:

$$\pi_{hi} = \frac{n_h}{N_h}.$$

1.2 2nd stage: selecting cluster of dwellings

The secondary sampling units (ssus) are implicitly defined clusters of dwellings inside the AUs. It would be more usual to choose meshblocks as ssus. However, TNS did not have

[1] Discussions with Robert Templeton of MSD, who has been investigating the weighting of this and the last survey, suggests that in retrospect it may have been better to treat them as if they had been selected srswor, as it seems that the mean square error of the estimates under this assumption are a little less than the mean square error for estimates finally produced.

meshblock maps so split the AU in half and then took a random start point in each half and then selected every kth dwelling. This procedure ensures that the two implicitly defined clusters are geographically separated and in particular do not overlap.

Suppose there are M_{hi} dwellings in the ith sampled AU. Suppose at the random start point in each half AU a sample of m_{hi} dwellings is taken systematically using a sampling interval of 1 in k. Then at each start point this defines an implicit cluster of size $k \times m_{hi}$, of which 1 in k are sampled. So there are $M_{hi}/(k \times m_{hi})$ such clusters in the AU, and 2 of them are taken. So the probability of selecting the jth cluster is:

$$\pi_{j|hi} = \frac{2}{M_{hi}/(k \times m_{hi})}.$$

Note that in the AUs selected sppswor $k = 5$ and $m_{hi} = 10$ and in the AUs selected srswor $k = 3$ and $m_{hi} = 7$.

1.3 3rd stage: selecting dwellings

Clearly the probability of selecting the kth dwelling is the sampling interval 1 in k, i.e.

$$\pi_{k|jhi} = \frac{1}{k}.$$

It is at this point that the first nonresponse adjustment is carried out (of course the calibration will be another nonresponse adjustment). Suppose instead of 7 responding dwellings there are 6. Then the 1 in 3 sampling interval is adjusted by the fraction 6/7. It is clear that the dwellings belonging to this cluster need to be known.

This is a common approach (particularly if nonresponse is low and no calibration to population totals is being done) to ensure that the sample design weights estimate the the population total closely. (If nonresponse is ignored then the sum of the sample design weights will be considerably less than the population total.)

However, doing the adjustment are the 2nd stage cluster level has pros and cons. The pro is principally that this adjustment to the weight is (implicitly) like substituting the mean of the other responses in the cluster for the nonrespondent and hence if the clusters are reasonably homogeneous, substituting like for like. The con is that the sample size at this level is small so that estimate of the mean is not so accurate and more variability is introduced to the estimator. It might be better to average over a larger number of observations. For example, SNZ does this sort of adjustment in most of its household surveys, but the smoothing is often done at the stratum or higher level.

So, the obvious modification to make is to do this adjustment not at the 2nd stage cluster level but at the 1st stage cluster level (AU). For that only the response rate only needs to be known at the AU level. This information TNS has supplied.

There is an additional small technical point. TNS selected in some AUs additional dwellings when the nonresponse was high. It could be assumed that they were selecting off a substitute list of dwellings which was predetermined before the main list was interviewed. (This is so because the main list was chosen as a 1 in 3 sample and the substitute list was another 1 in 3 sample). Hence the sampling for the dwelling is still a 1 in 3 sample and if the dwelling nonresponds, a substitute household is taken which carries the original households selection probability.

1.4 4th stage: selecting adult or Economic Family Unit

Sampling unit individual who defines Economic Family Unit (EFU).

1.4.1 Selecting an adult

Clearly if there are p_{hijkl} adults in the lth household then the probability of selection for an adult is

$$\pi_{l|kjhi} = \frac{1}{p_{hijkl}}.$$

1.4.2 Selecting an EFU

If there are e_{hijkl} estimated EFUs in the lth household (estimated because not all of them can necessarily be determined e.g. under 18 year olds with independent income; however, there will be few such EFUs so this bias is ignorable.)

$$\pi_{l|kjhi} = \frac{1}{e_{hijkl}}.$$

2 Controlling for size of Area Units

It is usual when the psus (AUs in this case) are selected srswor and a constant sample size is selected in each psu to use ratio estimation to control for the size of the psu and the variability in the overall rate-up weights. See for example Cochran (1977) Chapter 11 and page 303 in particular.

This approach is necessary in the LSS design because as Table 1 indicates, the estimate of number of occupied dwellings in each of the TNS regions based on the sample of AUs is typically very different from the Census counts. In particular in the major urban areas the estimated number of dwellings is on the low side, whereas in the remaining areas it is one the high side. The situation is probably worse than that depicted in Table 1 since there has been a substantial population increase in major urban areas, especially Auckland, since the 2001 Census.

Recall that 445 AUs were sampled srswor so that it is to be expected that the average size of AUs sampled to be very close to the average size of all AUs excluding the ones

Table 1: Comparison of estimated number of occupied dwellings and actual number on Census night 2001 in TNS regions for Area Units excluding those selected sppswor.

	sample estimate	Census
Ashburton	15807.00	9327
Auckland Rural	55616.40	39234
Blenheim	22452.75	15639
Cambridge Zone	7035.00	5181
Central Auckland Zone	126263.00	122325
Christchurch	115357.57	120576
Christchurch Rural	28266.86	23682
Dunedin	37716.67	38343
Dunedin Rural	29513.25	21318
Fielding	8662.50	5061
Gisborne	2895.00	8376
Gisborne Rural	3990.00	7293
Gore	3834.00	4860
Greymouth	22011.00	9945
Hamilton Rural	59567.14	47373
Hamilton Zone	39019.89	41847
Hastings Zone	23214.60	18648
Hawera	5701.50	4131
Invercargill	17316.00	16719
Invercargill Rural	4560.00	11463
Kapiti	24876.00	15363
Levin	5121.00	7380
Lower Hutt Zone	37674.00	33846
Masterton	11103.43	10905
Napier Hastings Rural	2132.00	5142
Napier Zone	15600.00	13284
Nelson	12789.00	15939
Nelson Rural	19786.00	15405
New Plymouth	10143.00	16227
New Plymouth Rural	26160.00	16317
Northern Auckland Zone	61776.82	64302
Oamaru	10417.50	8664
Palmerston North	24183.00	24366
Palmerston North Rural	38994.33	26937
Porirua Zone	12474.00	12906
Pukekohe	12024.00	6387
Rotorua	15574.00	17412
Rotorua Rural	2250.00	3249
Southern Auckland Zone	92413.79	92880
Taupo	19668.00	14952
Tauranga	33462.00	31551
Tauranga Rural	10398.00	11307
Te Awamutu Zone	3945.00	5013
Timaru	44370.00	18831
Tokoroa	5475.00	4872
Upper Hutt Zone	13623.00	11742
Wanganui	22250.00	14172
Wanganui Rural	4452.00	3096
Wellington Rural	1386.00	2436
Wellington Zone	53387.84	59649
Western Auckland Zone	47909.05	48627
Whakatane	32786.00	17079
Whangarei	20773.67	16464
Whangarei Rural	39081.82	34599

selected sppswor. However as the summary statistics below indicate, the sampled AUs are consistently larger than expected. Note that the standard error on the mean for the sampled AUs is about 20.

```
Number of occupied dwellings

Sampled srswor AUs
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    51.0   516.0   963.0   975.8  1356.0  3261.0

All AUs excluding those selected sppswor
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     0.0   219.0   663.0   735.7  1148.0  3261.0
```

The ratio estimation at the first stage means replacing the rate-up estimator

$$\hat{Y}_h = \sum_i \frac{\hat{Y}_{hi}}{\pi_{hi}},$$

where $\hat{Y}_{hi}$ is the estimate of the total for psu i in TNS region h, and $\pi_{hi} = \frac{n_h}{N_h}$, with the estimator

$$\hat{Y}_h^R = \frac{\sum_{N_h} M_{hi}}{\sum_i \frac{M_{hi}}{\pi_{hi}}} \sum_i \frac{\hat{Y}_{hi}}{\pi_{hi}},$$

or since is $\pi_{hi} = N_h/n_h$ for all psus in region h,

$$\hat{Y}_h^R = \frac{\sum_{N_h} M_{hi}}{\frac{N_h}{n_h}\sum_i M_{hi}} \frac{N_h}{n_h} \sum_i \hat{Y}_{hi} = \frac{\sum_{N_h} M_{hi}}{\sum_i M_{hi}} \sum_i \hat{Y}_{hi}.$$

I.e. the inclusion probabilities are constant within stratum so they cancel. In effect, the sampling estimates of the total number of dwellings $\sum_i \frac{M_{hi}}{\pi_{hi}}$, are being compared with the actual number $\sum_{N_h} M_{hi}$ and adjusting for any miscount.

Note that $\hat{Y}_h^R$ uses the rate-up estimator at all stages subsequent to the first, i.e. $\hat{Y}_{hi}$ is just the rate-up estimator using the conditional selection probabilities at the 2nd and subsequent stages.

Note also that in the language of theory of ratio estimators for stratified samples this is a separate ratio estimator (one for each stratum) rather than a combined one (summed over all strata). This is justifiable because the variability in AU size is considerable across TNS Regions.

The impact of using this ratio estimator is beneficial. Firstly the estimate of Census night dwellings is much closer to the actual value. Using the rate-up estimator at the first stage gives an estimate of 1438456 dwellings compared with an actual of 1367661. That is a 5% overestimate. Using the ratio estimator gives an estimate of 1364287. That is a 0.2% underestimate. Secondly summary statistics for the dwelling weights (i.e. the weights associated with selecting dwellings but not yet people) for the rate-up and ratio estimators are given below and it can be seen that the ratio estimator has less variables weights. The impact on the design effect (deff) will be significant. Using the measure considered by Kish, $n \sum_i w_i^2 / (\sum_i w_i)^2$, where w_i is the weight for the ith dwelling and n is the total number of dwelling in the sample, the component of the deff for sampling dwellings for the rate-up estimator is 1.96 and for the ratio estimator 1.73: i.e. a reduction of about 12%.

```
Dwelling weights

Using rate-up weights at 1st stage
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.83  109.60  201.60  288.30  375.00 2315.00

Using ratio weights at 1st stage
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.17  110.70  210.20  273.50  368.00 1771.00
```

3 Final sample design weights

Summary statistics for final sample designs weights for an individual and an EFU are given below.

```
Individual weight pfinwgt
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.17  184.70  393.30  543.70  729.80 5313.00

EFU weight efinwgt
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.17  123.90  254.60  365.20  481.00 5313.00
```

The large `pfinwgts` and `efinwgts` arise from Area Units which have very large number of dwellings, which is why ppswor schemes are usually chosen in these situations. The small `pfinwgts` and `efinwgts` arise from Area Units which have very small number of dwellings and were in the original sample selected ppswor, so have a first stage selection weight of 1.

4 Calibration weights

4.1 Potential areas of unrepresentativeness due to differential nonresponse.

In the 2000 LSS for the working age population, to account for differential nonresponse calibration was done to 2001 Census totals adjusted to the population count at March

2000.

The calibration cells used were:

- age × sex (18-24, then 5 year age groups)
- age × ethnicity (Māori / NonMāori)
- sex × ethnicity
- location × sex (Auckland, Wellington, Other Major Urban areas, Secondary/Minor urban areas, rest)
- location × ethnicity
- age × location

There was also some comparison of other variables against the 1996 Census as a result of which home ownership and salary or wages income variables were added. Again the benchmarks came from the 2001 Census.

The 2004 LSS also had differential response rates with an overall response rate of 62.2%.

Comparison of the sample estimates with those from the 2001 Census showed that there were substantial differences in proportions for several socio-demographic variables.

Table 2 compares sample estimates and 2001 Census usually resident percentages for Age × Sex. The sample consistently overestimates the percentage of women except in the ≥ 65 age group.

Table 3 compares sample estimates and 2001 Census usually resident percentages for Sex × Prioritized Ethnicity. Respondents were allowed to give up to three ethnic groups, so the calibration cells could have some multiple ethnicities e.g. Māori-European, Māori-Pasifika. However, given the overall sample size of 5000 and the need for largish cell sizes to keep the variance of the estimator small, there is a need to collapse these multiple ethnicities. Since there is an interest in the Māori population, there is a case to use the prioritized ethnic groups, where anyone reporting Māori is classified as Māori, then anyone reporting a Pasifika ethnic group is classified as Pasifika, etc.

The sample overestimates the percentage of Māori and European women and underestimates the percentage of Asian women. The Other group is too small to make a strong inference.

Table 4 compares sample estimates and 2001 Census usually resident percentages for Age × Prioritized Ethnicity. The sample varies considerably from the Census data in many cells but particularly too many young Māori, too few young Europeans and possibly too many young Asians.

Table 5 shows that there has been higher response rates in single Economic Family Unit households and in multiple EFU households where the respondent is a couple with

Table 2: Comparison of sample estimates and 2001 Census usually resident percentages for Age $\times$ Sex.

	18-24	25-34	35-44	45-54	55-64	≥ 65
Census						
Male	6.33	9.22	10.36	8.86	6.12	7.11
Female	6.36	10.17	11.12	9.12	6.29	8.93
Sample						
Male	4.63	6.18	7.85	7.38	5.83	8.28
Female	7.5	10.9	13.83	11.16	8.04	8.41

Table 3: Comparison of sample estimates and 2001 Census usually resident percentages for Sex $\times$ Prioritized Ethnicity.

	Male	Female
Census		
Māori	5.18	5.77
Pasifika	2.1	2.3
Asian	2.73	3.21
Other	0.31	0.27
European	35.51	38.5
Missing	2.03	2.11
Sample		
Māori	3.67	8.7
Pasifika	2.23	2.67
Asian	3.07	2.68
Other	1.6	2.13
European	29.42	43.55
Missing	0.11	0.17

no children. Particularly under-represented are single people with no children in multiple EFU households. The estimates of Labour Force Status in the sample closely matched the 2001 Census.

However, given the high proportion of missing responses in the 2001 Census for Wages as Income Source and Home Ownership, little can be said about under or over representation in these variables of the sample estimates.

It seems likely that the ratio estimation at 1st stage has corrected regional (location) imbalance.

Since the 2001 Census the New Zealand population has increased by around 5%, principally due to migration. So calibration to 2001 census totals, would require adjustment to the totals to account for this increase in population.

However, looking at the June 2004 provisional estimates of the population by age

Table 4: Comparison of sample estimates and 2001 Census usually resident percentages for Age × Prioritized Ethnicity.

	18-24	25-34	35-44	45-54	55-64	≥ 65
Census						
Māori	2.21	2.91	2.61	1.63	0.93	0.65
Pasifika	0.9	1.22	0.99	0.64	0.36	0.27
Asian	1.28	1.33	1.52	0.97	0.48	0.35
Other	0.1	0.16	0.15	0.09	0.04	0.03
European	7.59	12.82	15.21	13.8	10	14.44
Missing	0.54	0.86	0.88	0.75	0.53	0.77
Sample						
Māori	2.91	2.52	3.18	2.02	1.11	0.63
Pasifika	0.77	1.5	1.22	0.77	0.29	0.36
Asian	1.49	1.46	1.42	0.83	0.36	0.2
Other	0.49	0.58	0.97	0.79	0.47	0.43
European	6.45	10.94	14.88	14.04	11.64	15.02
Missing	0.03	0.07	0.02	0.1	0.01	0.05

and sex indicated some large changes in the proportions from the 2001 Census. Table 6 compares Provisional June 2004 population estimates and 2001 Census usually resident percentages for Age × Sex. There have been some important shifts in the age sex structure. The 55-64 age group has increased by about 0.5%, 25-34 old males have decreased by about 0.4% and 25-34 females by 0.8%. 18-24 year old males have increased by about 0.6%. It might be thought that the use of current data would reduce the need for calibration but as Table 7 shows, there is still considerable imbalance in the sample estimates.

Hence it was decided to use the provisional June 2004 estimates for calibration. Here ethnicity is rather limited and what is available is Māori and Other. In any case because ethnicity is not recorded on migration cards, the estimates for the Māori and Other populations are not as reliable as for the total population. At the time calibration was being carried out, the estimates were available for the Total population but not the ethnic breakdown. A comparison was made between the June 2003 Māori estimates with the year ending December 2003. The proportions in the 5 year age groups seemed stable. Hence the June 2003 Māori proportions on the June 2004 data was used to produce Age × Sex × Ethnicity calibration cells.

The initial age groups for calibration were the standard 5 year age groups, except for the 15-19 age group where this group was split into 15-17 and 18-19, and the old, where the 85 and older were formed into one group.

The reason for considering calibrating on children (those under 18) is that analysis will be done on EFUs. There is some imbalance in the household structure, in particular too many children, so there is need to calibrate.

Table 5: Comparison of sample estimates and 2001 Census usually resident percentages for estimated Economic Family Units.

	Single EFU	Multiple EFUs
Census		
Couple no children	14.7	6.9
Couple with children	15.9	5.2
Single no children	9.9	38.8
Single with children	4.5	4.2
Sample		
Couple no children	20.1	11.2
Couple with children	17.5	6.6
Single no children	14.1	20.6
Single with children	5.3	4.2

Table 6: Comparison of Provisional June 2004 population estimates and 2001 Census usually resident percentages for Age × Sex.

	June 2004		Census 2001	
	Males	Females	Males	Females
18-24	6.90	6.61	6.33	6.36
25-34	8.81	9.37	9.22	10.17
35-44	10.05	10.74	10.36	11.12
45-54	8.84	9.06	8.86	9.12
55-64	6.64	6.75	6.12	6.29
65+	7.17	9.04	7.11	8.93

4.2 Integrated weighting

Because one of the analysis units is the EFU, and because as argued above, there is a need to calibrate both respondents and members of the respondent's EFU (Family Group 1) and because members in Family Group 1 will typically have different ages, sexes or ethnicities, they will end up with different weights, whereas they had the same sample design weight.

In such situations, a common approach is to use integrated weighting, where some adjustments are made to the calibrated weights so that all members in Family Group 1 have the same calibrated weight.

This is not straightforward for the method of calibration used last time (see below) but turns out to be straightforward for the method used this time (see below).

Given that the respondent in Family Group 1 has a calibrated (and integrated) weight, it would be natural to construct their weight (for when a respondent analysis is needed) by taking the Family Group 1 calibrated weight and multiplying by the number of adults

Table 7: Percentage differences in age groups between survey and provisional June 2003 population estimates.

	male	female
0-4	13.3	23.1
5-9	11.9	13.9
10-14	3.0	15.1
15-19	-9.0	-14.0
20-24	-30.3	-6.2
25-29	-25.7	-18.6
30-34	-14.1	5.3
35-39	-11.7	-2.6
40-44	-5.8	8.7
45-49	-21.1	-0.5
50-54	26.0	15.1
55-59	-6.1	13.9
60-64	2.6	10.2
65-69	2.5	1.9
70-74	8.8	9.6
75-79	13.9	-5.9
80-84	6.2	-32.0
85-89	-14.7	-59.0
90+	-65.2	-81.8

in Family Group 1. But as the results below show, this gave more extreme values to the weights. This would impact on the overall design effect. Furthermore there were still some imbalances, so it was decided to calibrate the respondents weight to the population totals (of course for the 18+ population).

Finally, and again for reasons of minimizing the variability in weights (the sample sizes were also too small in some groups), the age groups were collapsed as follows:

```
[0,10)  [10,18)  [18,25)  [25,35)  [35,45)  [45,55)  [55,65)  [65,75)  [75,100)
```

4.3 Calibration methodology

Some imputation for sex (1 missing value in Family Group 1 data) and age (36 missing values in Family Group 1 and 15 missing values in respondent data) was necessary so that everyone in the sample could get a calibrated weight. The approach taken with imputation was either to construct a plausible value using other household members data: e.g. the age of a female partner is usually less than that of a male partner) or the hot deck approach where the information is borrowed form the next nearest person in similar household structure.

The basic idea behind calibration is an adjustment of the weights derived from the inverse of the inclusion probabilities. Call these the design weights $d_k = 1/\pi_k$. The adjustment is made so that the new weights, call these w_k, match known population totals of certain auxiliary variables, e.g. age group or sex counts. But also the w_k's need to be as close as possible to the d_k's. In effect the w_k's can be expressed in terms of what are called g-factors as:

$$w_k = g_k d_k \qquad \text{or} \qquad w_k = \frac{g_k}{\pi_k}.$$

It is sensible to consider making the g-factors close to 1 by minimizing an appropriate distance between 1 and the g-factors.

An example of a distance function is

$$\sum_{k=1}^{N} h_k (g_k - 1)^2, \tag{1}$$

where h_k is a known set of constants.

However, because when doing calibration a sample only is available, it is possible to minimize only a sample version of this:

$$\sum_{k \in s} \frac{h_k}{\pi_k} (g_k - 1)^2,$$

so the g-factors are sample dependent. This could be rewritten as

$$\sum_{k \in s} \frac{h_k}{d_k} (w_k - d_k)^2.$$

This quantity is minimized subject to say

$$\sum_{k \in s} w_k x_{jk} = t_{x_j},$$

where x_j is some auxiliary variable (e.g. a dummy variable which is 1 if the a respondent is say a male aged 35-44, but 0 otherwise), and t_{x_j} is the exact total (e.g. population count for males aged 35-44).

Using this distance measure, $\sum_{k \in s} (h_k/d_k)(w_k - d_k)^2$, there is an analytic solution to the minimization problem, namely

$$w_k = \left[1 + \left[\left(t_X - \hat{t}_X \right)' \left(X'WX \right)^{-1} X'W \right]_k \right] d_k, \tag{2}$$

where t_X is the vector of population totals for the auxiliary variables, $\hat{t}_X$ is the estimate of these totals from the sample, X is the matrix of auxiliary variables, W is a diagonal (weight) matrix whose k, k element is d_k/h_k. Clearly the g-factor is the term in the outer brackets on the right hand side.

It turns out that Equation 2 can be simplified if a certain technical condition holds[2]. Roughly speaking the condition is that if weighting the auxiliary variables by the inverse

[2]Mary Thompson in her book Thompson (1997) calls this the $X\Sigma$ condition.

of h_k there is a linear combination of these weighted variables which is constant on all observations. This condition holds here.

With this condition holding Equation 2 simplifies to

$$w_k = \left[t'_X \left(X'WX\right)^{-1} X'W\right]_k d_k, \tag{3}$$

This distance measure was used with $h_k = 1$ for all k.

Last time, partly because incomplete multiway tables were used (e.g. Age × Sex and Age × Ethnicity instead of Age × Sex × Ethnicity), and partly to ensure negative weights did not occur, the method of calibration was the generalized raking ratio. The distance measure corresponding to this is

$$\sum_{k \in s} w_k \log\left(\frac{w_k}{d_k}\right) - w_k + d_k. \tag{4}$$

The minimization problem with this distance measure does not have an analytic solution but e.g. the iterative proportional fitting algorithm provides a rapid numerical answer.

See Deville and Särndal (1992), Deville et al. (1993), Lundström and Särndal (1999), and Renssen and Martinus (2002) for details.

4.3.1 Integrated weighting

If dummy variables x_g are constructed corresponding to the Age × Sex × Ethnicity cells, (e.g.which is 1 if you are a Māori male aged 35-44), then for integrated weighting a new variable, z_g, is constructed as follows. For the kth member of the Family Group 1 sample

$$z_{g,k} = \frac{\text{number in cell } g \text{ in Family Group 1}}{\text{number in Family Group 1}}. \tag{5}$$

Summing $z_{g,k}$ over a Family Group 1 gives the number of people in that Family Group 1 in calibration cell g, and summing $z_{g,k}$ over all Family Group 1's gives the population count for that cell.

See Lemaître and Dufour (1987) for details.

4.3.2 Summary of calibrated and integrated weights

Fortunately there are no negative weights which as mentioned above can be a potential side effect of using the distance measure used. The table below gives the summary statistics for the original design weights for respondent and EFU as well as the calibrated and integrated weights summary statistics

```
pfinwgt original respondent weight
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.17  184.70  393.30  543.70  729.80 5313.00

cpfinwgt calibrated respondent weight
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.26  194.20  417.30  600.10  803.40 6818.00

efinwgt original efu weight
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.17  123.90  254.60  365.20  481.00 5313.00

cefinwgt calibrated and integrated efu weight
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.05  127.90  261.10  396.70  499.50 6913.00

cefinwgt * # adults in efu
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 14.05   194.00   423.90   600.10   804.70 10910.00
```

Note as mentioned above using the EFU calibrated weight multiplied by the number of adults in the EFU produces more extreme weights than the calibrated respondent weight.

The impact of the calibration and integration is small the g_k factors vary from around .86 to 2.7 for respondent weight and .38 to 2.7 for EFU weight.

The large and small weights are typically due to the factors mentioned in Section 3 above, as well as differential response rates giving large or small calibration factors.

Note that for the respondent weight if an srswor of 4989 people 18+ was taken from 2993800 then the weight would be 600.08 which is, as expected, the same as the mean of the calibrated respondent weight.

For the EFU weight the mean weight would be 396.71 if an srswor of 4989 EFUs was taken from (an estimated) 1979200.

So the variability in weights is fairly high (for respondent the ratio of the 2.5% quantile to 97.5% quantile is about 43, and for the EFU about 48.

However, the impact of the Kish deff measure (see Section 2 above) which looks at variability in weights is not so pronounced. The calibration has a small effect as indicated in the table below compared with the variability of the design weights.

```
deff efinwgt   deff cefinwgt
2.03           2.26

deff pfinwgt   deff cpfinwgt  deff cefinwgt * # adults in efu
1.89           2.00           2.03
```

A quick look suggests that household composition, regional estimates, etc look plausible with the calibrated weights. The differences may well be changing demographics.

For example at Census 2001 estimated EFUs were 1.627 million, now the estimate is 1.979 million. This means in 3 years a reduction in EFU size from 2.32 people per EFU to 2.05.

4.3.3 Controlling for EFU type

As MSD began producing tables for the report, it became clear that the proportions of EFU type for the current survey were quite different to those from the previous survey. Clearly factors such as the aging population, migration, and steady move over the last decade to more single person EFUs and Couples without children would mean the proportions would not be expected to be the exactly the same. Nevertheless, on balance these factors did not seem likely to explain the observed differences.

So it was decided to calibrate on EFU type. The calibration totals were constructed in a mixed fashion because there are no current populations counts for the population as a whole.

For the 65 and over population, there is reliable administrative data using Guaranteed Retirement income beneficiaries. From this data the following is known: the number of couples, one or both of whom are over 65; the number of singles; and the number of couples one of whom is in a non-private dwelling, e.g. resthome or hospital.

For the 18-64 population it decided to use the proportions at last Census applied to the estimated number of EFUs from this survey.

The following table gives the calibration totals.

EFU type	number	percentage
single person resp < 65	163774	8.27
couple only resp < 65	242153	12.23
couple with children resp < 65	262112	13.24
single person with children < 65	74007	3.74
single person < 65 multiple efu	641331	32.40
couple only resp < 65 mutilple efu	113182	5.72
couple with children resp < 65 multiple efu	85837	4.34
single person with children < 65 multiple efu	68591	3.47
couple both >= 65	114178	5.77
couple resp >= 65	11019	0.56
single >= 65	203053	10.26

Here resp means respondent.

Because the EFU weights are estimating EFUs and respondent weights are estimating the 18 and over population, it is not possible to calibrate in the one step using say Equation 3. Instead, an iterative procedure similar to that used last time was used.

First the EFU type was calibrated using iterative proportional fitting (that is using the distance measure Equation 4). In fact because the calibration cells are in fact a complete one-way table there is no iteration. This amounts to the usual poststratification estimator.

Second the resulting EFU weight was applied to Family Group 1 members and then the integrated weighting procedure used previously was carried out.

This procedure was iterated until a reasonably close fit was obtained between the EFU type, age, sex, and ethnicity calibration totals. After 10 iterations there was a very close fit. Specifically the EFU type was exact and the age, sex, and ethnicity percentage differences less than 0.03%: see next table.

```
Percentage differences in age, sex, and ethnicty calibration cells for the
integrated EFU weight

maori                       other

              m f                          m     f
[18,25)    0.00 0           [18,25)    0.01  0.01
[25,35)    0.00 0           [25,35)    0.01  0.00
[35,45)    0.00 0           [35,45)    0.01  0.01
[45,55)    0.00 0           [45,55)    0.00  0.01
[55,65)    0.00 0           [55,65)    0.00  0.00
[65,75)    0.00 0           [65,75)   -0.02 -0.02
[75,100)  -0.01 0           [75,100)  -0.01 -0.02
```

Finally, the respondent weight for the Family Group 1 respondent was constructed by multiplying the integrated EFU weight by the number of adults in Family Group 1. There is a reasonably close fit to the age sex ethnic calibration totals. The differences in percentages are:

```
maori                       other

              m     f                      m     f
[18,25)   -0.04  0.04       [18,25)   -0.16  0.05
[25,35)   -0.08  0.07       [25,35)   -0.33  0.09
[35,45)   -0.06  0.13       [35,45)    0.12  0.13
[45,55)    0.02  0.06       [45,55)   -0.13  0.20
[55,65)    0.02  0.03       [55,65)    0.33 -0.15
[65,75)    0.07 -0.11       [65,75)    0.15 -0.35
[75,100)   0.00  0.00       [75,100)  -0.16  0.07
```

Recall that in the previous calibration constructing the respondent weight this way produced large percentage differences and hence this respondent weight was recalibrated, thus destroying the simple relationship between the EFU weight for Family Group 1 and the respondent weight.

Given the reasonably close fit, certainly well within sampling error, and the desire to maintain that simple relationship it was decided not to recalibrate this weight.

In summary, the EFU weight is produced by an iterative procedure calibrating first on EFU type then integrating the Family Group 1 members EFU weight to age, sex and ethnicity totals. The respondent weight is the EFU weight times the number of adults in the EFU.

The table below gives the summary statistics for the new calibrated and integrated weights. It is clear that improving the estimates of the EFU type has come at the cost of increased variability in the weights. Although the interquartile ranges are similar to the previous weights, the tails, particularly the right hand one are longer.

```
calibrated and integrated EFU weight
    Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
   6.699   120.000   248.500   421.300   497.600 11590.000

calibrated and integrated respondent weight
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  12.41   186.10   394.50   600.10   784.50 11590.00
```

5 Proposed method of variance estimation

Because for this survey:

- there is effectively a self-representing stratum of those AUs initially selected sppswor;
- because for the AUs selected srswor ratio estimation was used at the first stage selection;
- calibration is used as a nonresponse adjustment method

it does not seem feasible nor realistic to use an analytic approach to calculate variances for the variables estimated in this survey.

With such a design it is more natural to consider replicated methods such as the jackknife. This method will produce a set of *jackknife weights*. An additional benefit is that the parameters estimated by modelling of the data can have standard errors calculated taking account for the complex survey design.

5.1 Jackknife variance estimation

Suppose there is an estimator $\hat{\theta}$ of some population parameter θ based on the full sample. Suppose for the moment that the sample is not stratified. Then the jackknife technique has the following steps.

1. Partition the sample of size n into K random groups of equal size m.

   Assume that, for any given sample s each group is a simple random sample from the sample s even if the sample s is not a simple random sample.

2. For each group $k \in K$, calculate $\hat{\theta}_{[-k]}$, an estimator of the same functional form as $\hat{\theta}$ but based on the data omitting the kth group.

3. Define for each $k \in K$ the kth pseudovalue:

$$\hat{\theta}_k = K\hat{\theta} - (K-1)\hat{\theta}_{[-k]}$$

   This is motivated by the case of the usual sample mean estimator where the sample value X_i can be written as

$$X_i = n\bar{X} - (n-1)\bar{X}_{[-i]}$$

   where $\bar{X}$ is the sample mean for the full sample, and $\bar{X}_{[-i]}$ is the sample mean for the sample with unit i omitted.

4. Form the jackknife estimator of θ (an alternative to $\hat{\theta}$)

$$\hat{\theta}_{JK} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$$

   Note that the difference between $\hat{\theta}_{JK}$ and $\hat{\theta}$ is called the jackknife bias.

5. Form the jackknife variance estimator:

$$\hat{V}_{JK1} = \frac{1}{K(K-1)}\sum_{k=1}^{K}(\hat{\theta}_k - \hat{\theta}_{JK})^2$$

   Note that given the definition of $\hat{\theta}_k$ this can be rewritten as

$$\hat{V}_{JK1} = \frac{K-1}{K}\sum_{k=1}^{K}(\hat{\theta}_{[-k]} - \bar{\theta}_{JK})^2.$$

   Essentially, this is the sum of squares of the jackknife estimators about their mean $\bar{\theta}_{JK}$.

The estimator $\hat{V}_{JK1}$ is used to estimate $V(\hat{\theta})$ as well as $V(\hat{\theta}_{JK1})$. So $\sqrt{\hat{V}_{JK1}}$ can be used to estimate the sampling error for the complex estimator under the complex design.

If the $\hat{\theta}_k$'s were uncorrelated then $\hat{V}_{JK1}$ would be unbiased for V_{JK}. But in general the $\hat{\theta}_k$'s are correlated so unbiasedness doesn't hold. There are no exact results for the properties (bias variance, asymptotic distribution, etc) of the jackknife estimator and the jackknife variance estimator for complex estimators, but empirical evidence suggest that it gives good estimates of sampling errors for many complex statistics.

A more conservative estimator is

$$\hat{V}_{JK2} = \frac{K-1}{K} \sum_{k=1}^{K} (\hat{\theta}_{[-k]} - \hat{\theta})^2, \tag{6}$$

essentially the sum of squares of the jackknife estimators about the estimator $\hat{\theta}$, i.e. the mean square error.

Finally, for generalized regression estimators (such as the calibrated estimator used here) the jackknife estimator can be reduced to a set of jackknife weights which can be calculated once and then applied to any variable.

For multistage sampling the random groups for the jackknife technique are usually the primary sampling units (psus) (in this case Area Units) and typically the estimators are nonlinear functions of the (unbiased) psu totals. However, practical sample designs for household typically have many psus (in this case 522 in total) and so the number of jackknife replicates becomes large. Whilst this is generally not a problem now that computers are powerful, it is usual to use groups of psus.

It is well known that one has to be careful when applying the jackknife to stratified designs (see e.g. Wolter (1985) page 175).

A common method (see Rust (1985)) is to apply the jackknife technique at the stratum level and then combine the estimates using independence of stratum samples as in the usual Horvitz-Thompson estimator. Where a design is highly stratified such as this (There are 54 TNS regions plus the self representing stratum), even with grouping of psus, this lead to a large number of replicates.

More typically now the method is to delete a psu or groups of psus in one stratum at a time while keeping the remaining strata fixed. See Rust and Rao (1996).

5.2 Delete-a-group-jackknife

An alternative approach is to form the groups of psus across strata. Specifically, the psus are formed into a list ordered by stratum and randomly within strata. Groups are then formed by systematically selected psus down the list with a constant sampling interval.

With a small number of strata this can reduce the number of replicates considerably whilst not increasing the bias of the variance estimate This method typically has a bias bounded by

$$\frac{K-1}{K} \min_h \frac{1}{n_h - 1}$$

where K is the number of groups. With K reasonably large, this term is dominated by $\min_h \frac{1}{n_h-1}$. Since the largest srswor sample comes from the TNS region Auckland Central Zone and is 45, the upper bound on the bias in the sampling error would be about 15%. Given, that there are other biases unaccounted for in the sampling errors, having a conservative estimator like this is probably sensible so that real differences are correctly reported.

See Kott (1998a), Kott (1998b) and Kott (2001) for details.

Where the original sample has been selected by systematic probability proportional to size scheme, this approach is similar to the random group estimator proposed for such designs (see Wolter (1985) page 288.)

Harry Smith (see Smith (2001)) investigated this approach for the Household Labour Force Survey which is highly stratified with 120 strata and found that even a delete-a-group-jackknife with as few as 40 groups produced acceptable variance estimates.

This method was used for SNZ's Disability survey run after the 2001 Census of Population and Dwellings. This has also been used in the 2003 Physical Health Survey.

Therefore, it was decided that this delete-a-group-jackknife be used to estimate the sampling errors and deffs for the LSS.[3]

5.3 Delete-a-group-jackknife for the LSS

The choice for the numbers of groups was as follows.

- There are 522 AUs and this factors into 18×29, so if unequal sized groups and possibly *boundary effects* in the systematic sampling are to be avoided, 18 or 29 are the choices for the group size.
- In the major urban areas the TNS regions typically have at least 18 AUs selected in the srswor sample, so choosing 18 would mean each group has at least one major urban area AU deleted. This should provide good balance.

The conservative estimator

$$\hat{V}_{JK2} = \frac{K-1}{K} \sum_{k=1}^{K} (\hat{\theta}_{[-k]} - \hat{\theta})^2,$$

was chosen.

[3]Subsequent discussions with Robert Templeton of MSD, who implemented the jackknife estimation and carried out a quality check on its performance suggest that there is a reduction in variance for some subpopulations when more replicates are used. This has to be traded off against ease of analysis and modelling.

5.3.1 Algorithm for producing jackknife weights

This algorithm is designed for the SAS datasets which were available at the time sampling errors were to be produced. Clearly it could be modified for other software.

1. Order the strata as follows:
   (a) Self-representing.
   (b) Major Cities: Northern Auckland Zone, ..., Hamilton, ..., Porirua Zone, ..., Christchurch, Dunedin.
   (c) Minor Cities: Whangarei, Tauranga, Rotorua, Gisborne, Napier, Hastings, New Plymouth, Palmerston North, Wanganui, Nelson, Timaru.
   (d) remaining Upper North Island.
   (e) remaining Lower North Island.
   (f) remaining South Island.
2. Randomly order the AUs within strata.
3. Identify the 18 (or number of replicates) systematic samples through the strata. For each sample, work out the number of AUs deleted in each TNS region.
4. For the AU sample of self representing strata, `rrpps`, if the AU is in the kth systematic sample delete it. Then work out the AU weight:

$$w = \frac{77}{77 - n_{[-k]}},$$

where $n_{[-k]}$ is the number of AUs deleted in the kth systematic sample. Now work out the dwelling weight which accounts for nonresponse at the AU level.

$$dwgt = w \times \frac{M_i}{m_i},$$

where M_i is the number of occupied dwellings at 2001 Census in the ith AU and m_i is the number of responding dwellings.

For the AU samples in TNS-region strata, `rrsrs`, if the AU is in the kth systematic sample the delete it. Then work out the AU inverse of selection probability weight:

$$w_{h[-k]} = \frac{N_h}{n_h - n_{[-k]}},$$

where N_h is the total number of AUs in stratum h, and n_h is the number TNS sampled. As before $n_{h[-k]}$ is the number of AUs deleted in the kth systematic sample, in stratum h. Note in the file `rrsrs` N_h and n_h are called `Nh` and `nh` respectively.

Then work out the design weight which is

$$d_h = \frac{\sum_{N_h} M_{hi}}{\sum_{i \notin [-k]} M_{hi}}.$$

So the key here is to work out for each replicate sample (the complement of the systematic sample, denoted by $i \notin [-k]$ in the formula above) the sum of occupied dwellings.

Now work out the dwelling weight which accounts for nonresponse at the AU level.

$$dwgt = d_h \times \frac{M_{hi}}{m_{hi}},$$

where M_{hi} is the number of occupied dwellings at 2001 Census in the ith AU in the hth stratum, and m_{hi} is the number of responding dwellings.

5. Using a subset of the respondent file which has variables:
   - identifier for respondent,
   - any variable used for calibration,
   - number of adults in household
   - number of EFUs in household

   merge with the two AU files and work out the `pfinwgt` and `efinwgt`.

   $$\texttt{pfinwgt} = dwgt \times n^a_{hij}$$

   where n^e_{hij} is the number of adults in respondent j's household. Similarly,

   $$\texttt{efinwgt} = dwgt \times n^e_{hij}$$

   where n^e_{hij} is the number of EFUs in respondent j's household.

6. Create a family Group 1 file which has all the members in Family Group 1, with the variables
   - identifier for respondent assigned to each member of Family Group 1,
   - any variable used for calibration,
   - number of adults in Family Group 1
   - number of children in Family Group 1
   - `efinwgt`

7. Do calibration on the Family Group 1 file including integration using `efinwgt`. This is an iterative procedure.

   (a) Adjust `efinwgt` so that calibrates to the EFU type totals. Simply construct the appropriate 1-way table using `efinwgt` and for each cell i, work out the factor $f_i = t_i / \sum_{j \in i} \texttt{efinwgt}_j$. Apply this factor to all `efinwgt`'s belonging to cell i.

   (b) Using this new `efinwgt` do the integrated weighting on Family Group 1.

   This involves constructing a design matrix X from dummy variables corresponding to the calibration cells and totals as outlined in Section 4.3 and in particular Equation 2 and Equation 5. Note that is SAS the weight matrix W is implemented in regression procedures e.g. `PROC REG` by using the `WEIGHT` statement.

Iterate until the age sex ethnic totals are close to the calibration totals. Some judgement is required here. Probably if there is less than 0.1% difference the iteration can stop.

Call this weight `cefinwgtk` where `k` indicates the kth systematic sample which was deleted. This is then, the kth jackknife or replicated weight. Merge this weight back onto the main respondent file. For the respondents in the deleted AUs `cefinwgtk` will be missing in a SAS merge. These can be set to zero so that other software which does not handle missing values can use these weights.

8. Construct the respondent weight by multiplying `cefinwgtk` by the number of adults in Family Group 1. Call this weight `cpfinwgtk` where `k` indicates the kth systematic sample which was deleted. Merge this weight back onto the main respondent file.

5.3.2 Producing sampling error estimates

Given the $K = 18$ jackknife or replicate weights the sampling errors for any variable is estimated using the method outlined in Section 5.1 and in particular Equation 6. The sampling error is the square root of this which then can be multiplied by whatever z-value corresponds to the desired confidence level to produce a half width confidence interval.

References

Bethlehem, J. G. and Keller, W. J. (1987). Linear Weighting of Sample Survey Data. *Journal of Official Statistics*, 3(2):141–153.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley and Sons, Inc., New York, 3rd edition.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized Raking Procedures in Survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11(1):55–77.

Kott, P. S. (1998a). Using the Delete-A-Group Jackknife Variance Estimator in NASS surveys. NASS Research Report RD-98-01, Research Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC. Revised July 2001.

Kott, P. S. (1998b). Using the delete-a-group jackknife variance estimator in practice. In *ASA Proceedings of the Survey Research Methods Section*, pages 763–768.

Kott, P. S. (2001). The Delete-A-Group Jackknife. *Journal of Official Statistics*, 17(4):521–526.

Lemaître, G. and Dufour, J. (1987). An Integrated method for Weighting Persons and Families. *Survey Methodology*, 13(2):199–207.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15(2):305–327.

Renssen, R. H. and Martinus, G. H. (2002). On the uses of Generalized Inverse Matrices in Sampling Theory. *Survey Methodology*, 28(2):209–212.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1(4):381–397.

Rust, K. F. and Rao, J. N. K. (1996). Variance estimation techniques for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5:283–310.

Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Smith, H. (2001). Investigation of the delete-a-group jackknife variance estimator. Research and Analytical Report 13, Statistics New Zealand.

Thompson, M. E. (1997). *Theory of Sample Surveys*. Chapman and Hall, London.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York.