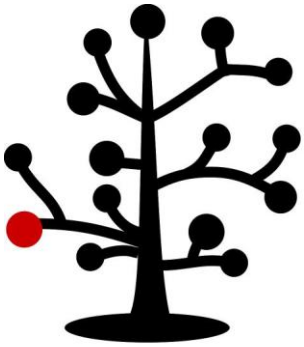

N.i.D.S.

NATIONAL INCOME DYNAMICS STUDY

Weights: Report on NIDS Wave 1

Technical Paper no. 2

Martin Wittenberg
Department of Economics, University of Cape Town
Martin.Wittenberg@uct.ac.za

July 2009

The NIDS weights were derived in two stages. In the first, the design weights were calculated as the inverse of the inclusion probability. In the second, the weights were calibrated to the 2008 midyear estimates. In practice the process was a bit more complicated so there are not just two sets of weights available. In this document we set out how the weights were calculated and what choices were necessary in the process.

1 Calculating the design weights

The theoretical formula for calculating design weights is straightforward: the Horvitz-Thompson estimator (1952) is given by

$$w_i = \frac{1}{\pi_i}$$

where w_i is the weight and π_i is the inclusion probability of the i-th unit. We therefore need to be able to calculate π_i. This means taking account of the two-stage sampling design: first drawing a sample of PSUs and then drawing a sample of dwellings within them.

1.1 The probability of including a PSU

In principle the calculation of the probability of a PSU being selected is quite simple, it is the probabability of the PSU appearing in the master sample (supplied by Statistics South Africa) times the probability of being drawn from that master sample. The problem is that nine PSUs that were drawn were not visited at all and were replaced. This implies that the probability of a PSU appearing in the NIDS sample conditional on it being in the master sample is now given by:

$$\begin{aligned}\pi_{PSU_j} \quad = \quad & \Pr(\text{PSU } j \text{ is selected}) * \Pr(\text{fieldwork is possible}) + \\ & \sum_{k \neq j} \Pr(\text{PSU } k \text{ is selected}) * \Pr(\text{fieldwork is not possible in } k) * \\ & \Pr(\text{PSU } j \text{ is selected as replacement for } k) + \ldots\end{aligned}$$

(The dots signal that there should be a third and higher-order replacement terms also, given that there is a non-zero probability of fieldwork not happening in a replacement PSU.) There are a number of imponderables in this. Firstly what is the probability of fieldwork being possible in a particular area? If we treat this as a stochastic variable then we need to recalculate the inclusion probabilities even for PSUs that were not replaced, since presumably the *ex post* discovery that fieldwork was possible does not mean that the *ex ante* probability of this event was one. So how might we get estimates of $\Pr(\text{fieldwork is possible})$? One approach is to estimate this given the available information: geography type, location, average household size, predominant "race" group. It turns out that this is not quite so straightforward, because within certain categories (e.g. urban informal, several of the provinces) no replacement ever took place, so the best estimate of the probability is, in fact, one. Within the areas where the outcome did vary the available information does not predict the outcome very well (the coefficients in a probit turn out to be all insignificant). The simplest procedure might therefore be to assign fieldwork difficulties entirely to "geography", e.g. calculate the fraction of PSUs that needed to be replaced within a district council and use that as the probability that fieldwork is not possible.

The second difficulty is that we have no information about $\Pr(\text{PSU } j \text{ is selected as replacement for } k)$. The most tractable assumption is to assume that the second draw is made randomly within the district council. If we also assume that $\Pr(\text{fieldwork is not possible in } k)$ is constant within district,

then the formula for selection becomes

$$\begin{aligned}\pi_{PSU_j} &= \Pr(\text{PSU } j \text{ is selected}) * \Pr(\text{fieldwork is possible}) + \\ &\left(\sum_{k \neq j} \Pr(\text{PSU } k \text{ is selected})\right) * \Pr(\text{fieldwork is not possible}) * \\ &\Pr(\text{PSU } j \text{ is selected as replacement}) + \ldots\end{aligned}$$

Indeed since draws within the district council are made with equal probability, it turns out that these probabilities must all be equal within district councils, so we could effectively ignore the replacement procedure and proceed as though all our PSUs were selected at the beginning. Of course this is valid only if the replacement PSUSs were really drawn at random and if the probability of fieldwork being possible is constant within district councils.

If $\Pr(\text{PSU } j \text{ is selected as replacement})$ is not random, e.g. if there is any matching on the characteristics of the area, then we would need to to know for how many other PSUs a given one is the best match and we would need to know the probability of selection and probability of fieldwork within those. Calculating these probabilties is not possible given the information to hand.

A different approach would be to "undo" the replacement procedure. If replacement hadn't happened, then the probability of inclusion is just

$$\pi_{PSU_j} = \Pr(\text{PSU } j \text{ is selected}) * \Pr(\text{fieldwork is possible})$$

Assuming again that $\Pr(\text{fieldwork is possible})$ is constant within district council, we can estimate this probability. The resulting weights essentially "weight up" the observed PSUs within the district council to compensate for the missing ones.

In summary there are two approaches:

- weight PSUs as though $\Pr(\text{fieldwork is possible})$ is constant within district council and replacement happened at random

  The resulting weights can be thought of as weights ignoring the problem of replacement

- weight PSUs as though $\Pr(\text{fieldwork is possible})$ is constant within district council and replacement never occurred, i.e. weight replaced PSUs at zero

1.2 The probability of interviewing a household

Within PSUs twenty-four (or forty-eight) dwellings where extracted from a listing of dwellings compiled in 2006. This means, in effect, that people living in dwellings constructed on greenfields sites since then have a zero probability of appearing in the sample. Dwellings that have become vacant introduce a further complication. Obviously no interview is possible from such a site. Once an occupied dwelling has been visited, the household(s) present there have to consent to being interviewed. The probability of a household in a sampled PSU being interviewed is therefore

$$\begin{aligned}\pi_h &= \Pr(\text{participation}|\text{hh selected}) * \\ &\Pr(\text{hh selected}|\text{occupied dwelling selected}) * \\ &\Pr(\text{occupied dwelling selected}|\text{PSU selected})\end{aligned} \tag{1}$$

The final term in this expression would be easy to calculate if we knew how many occupied dwellings there were in the PSU, but we don't. One approach would be to rewrite this expression as

$$\Pr(\text{occupied}|\text{dwelling selected}) \Pr(\text{dwelling selected})$$

To make this more concrete assume that there are N_c dwellings listed within the cluster, that we extracted a sample of size 24 and that there were n_o occupied dwellings. This formula would suggest that

$$\Pr\left(\text{occupied dwelling selected}|\text{PSU selected}\right) = \frac{n_o}{24} * \frac{24}{N_c} = \frac{n_o}{N_c}$$

This approach assumes that the universe within which we are operating is that of all dwellings, instead of that of occupied dwellings. Implicitly we would be assuming that the occupied dwellings would have to proxy for the vacant ones. That might be a valid approach if the households in the vacant plots had built new structures somewhere else, i.e. the vacant plot is a type of missing household.

A different approach is to assume that there really isn't a household corresponding to the vacant plot somewhere outside the coverage of the survey. In that case we need an estimate of the number of occupied dwellings within the cluster. The most straightforward estimate is

$$N_{oc} = N_c * \left(\frac{n_o}{24}\right)$$

Then the probability of a particular occupied dwelling being selected from all the occupied dwellings in the cluster is

$$\Pr\left(\text{occupied dwelling selected}|\text{PSU selected}\right) = \frac{n_o}{N_{oc}} = \frac{24}{N_c}$$

This is, in fact, the approach that was adopted.

The middle term in equation 1 is supposedly one, so this creates no problems. The first term again creates difficulties. There is no information on households that refuse to participate, so the easiest procedure is to assume that this probability is constant across all households within the PSU. We can estimate this probability as the fraction of participating households over the total number of households within the sample of occupied dwellings. In the available data it is the number of households observed in the sample divided by "countDUs".

1.3 Trimming the weights

In a final step the weights were "trimmed" to reduce the influence of a few households with very large weights. These arose in PSUs in which only one or two households were interviewed. The weights were trimmed to the 95th percentile of the weights.

2 Calibrating the weights

2.1 Why post-stratify?

Post-stratification involves adjusting the weights of a survey so that the application of those weights makes the sample look like the population, e.g. in terms of its distribution across provinces and demographic characteristics. There is no intrinsic reason why a random sample should look representative of the population. Sampling theory allows us to place bounds on that variability. The attraction of post-stratification is that it can reduce that variability further **provided** that the attributes that we are trying to measure are correlated with characteristics we are using to post-stratify on. So, for instance, if the probability of employment is strongly related to geographical factors, then down-weighting over-represented provinces and weighting up under-represented ones should give us a more accurate measure of employment than simply ignoring the lopsided nature of the sample. Adding in additional information from outside the survey can therefore improve the accuracy of measures provided that this auxiliary information is itself accurate.

The particular reason for wanting to post-stratify the NIDS survey becomes apparent in Table 1. The last line of that table shows that the NIDS sample has 35% too few Indians and 17% too few Whites. The entries within the table show that there are certain age groups that are prone to be overrepresented (in particular the elderly), while young adults (25 to 29 year olds in particular) are significantly underrepresented.

Table 1: Percentage difference between NIDS sample and Midyear population estimates

Age band	AM	AF	CM	CF	IM	IF	WM	WF
0-4	10.0%	3.0%	12.4%	1.8%	-27.8%	-10.6%	-9.8%	-13.7%
5-9	-6.9%	-11.3%	16.6%	-1.6%	1.7%	-25.2%	-3.5%	8.4%
10-14	3.8%	-8.5%	20.8%	23.3%	30.9%	40.6%	-9.1%	6.1%
15-19	0.1%	1.5%	-4.0%	22.5%	62.0%	18.6%	20.0%	-29.8%
20-24	-9.6%	11.3%	-2.2%	-1.0%	7.7%	25.7%	-25.1%	11.4%
25-29	-10.2%	-16.6%	-1.1%	-12.2%	-76.1%	-47.9%	-27.2%	-44.0%
30-34	-17.5%	-23.6%	-36.7%	-38.6%	-28.9%	-38.2%	32.8%	4.2%
35-39	-17.9%	-3.9%	-31.5%	1.2%	49.8%	-31.2%	0.3%	9.5%
40-44	26.1%	19.0%	0.5%	-2.5%	17.2%	100.3%	-15.0%	13.4%
45-49	7.7%	10.2%	4.4%	28.3%	-50.8%	10.4%	21.6%	13.3%
50-54	26.2%	10.0%	-17.6%	-3.1%	0.2%	-3.0%	-0.6%	29.3%
55-59	21.0%	16.0%	5.3%	-15.7%	64.3%	10.5%	-6.4%	-12.3%
60-64	-0.1%	-0.4%	14.8%	-21.0%	-15.4%	28.1%	2.2%	7.8%
65-69	26.5%	25.4%	-6.0%	-12.9%	-30.2%	-60.7%	-14.6%	-1.2%
70-74	-1.3%	21.8%	11.6%	4.2%	-11.5%	-73.0%	19.5%	46.8%
75-79	39.0%	78.2%	316.6%	55.5%	-92.3%	60.3%	40.0%	-17.6%
80+	169.3%	122.4%	-13.3%	-40.3%	191.9%	-94.8%	53.4%	-64.4%
Total	-1.3%	5.8%	3.2%	9.1%	-37.9%	-31.9%	-20.4%	-13.8%

NIDS estimates calculated using the design weights.
A: African C: Coloured I: Indian W: White; M: Male F: Female

2.2 How the calibration was done

2.2.1 The constraints

The sample weights were adjusted so that the resident NIDS population conformed to the age-sex-race distribution of the 2008 midyear population estimates released by Statistics South Africa. A separate constraint was that the distribution by provinces should correspond to that released in those

population estimates and that the total weights should add up to the estimated total population of 48,687,000. A further constraint imposed was that the weights should be constant within households. This is based on the assumption that the mismatch is due to the fact that we disproportionately missed certain types of households, rather than that we disproportionately underenumerated particular age groups within the households that we found.

In order to implement these constraints we needed to decide how to deal with individuals where the age, sex or race was missing. These were all allocated to a residual category. We imposed the condition that the proportionate weight of these individuals (around 1.4% of the sample) should not change due to the reweighting.

2.2.2 The technique

Weights adjustments are often done by a "raking" procedure, i.e. the ex-ante weights (in this case the design weights) are scaled up or down to make the weights sum up to one of the constraints (e.g. the age-sex-race counts), then rescaled to fit the second constraint (provincial totals), then back to the first until all constraints are met.

Instead we calculated the post-stratified weights by the "cross-entropy" estimation procedure (Golan, Judge and Miller 1996, p.29). The idea is to minimise the cross-entropy measure

$$\sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}$$

where p_i is the set of weights to be chosen (one for each individual) and q_i is the set of ex-ante weights (rescaled to sum to one). We used the original design weights and not the trimmed ones, since these weights would get rescaled through the procedure anyhow. The minimisation is done subject to the set of constraints imposed on the problem, i.e.

$$\begin{aligned} \sum_{i=1}^{n} p_i &= 1 \\ y_j &= \sum_{i=1}^{n} x_{ij} p_i \end{aligned}$$

In our case y_j is a particular population proportion (e.g. the proportion of people in the Western Cape is 0.10803) and x_{ij} is a dummy variable indicating whether the i-th individual in the data set is in the Western Cape or not. Altogether there are 146 of these constraints: 9 provincial proportions, 136 age-sex-race proportions plus the proportion "missing". Two of these constraints are redundant, since the province proportions add up to unity, as do the age-sex-race plus "missing" proportions.

It is relatively straightforward to show that the cross-entropy solution is equivalent to the solution that would be obtained by rescaling the proportions iteratively until convergence is achieved (Wittenberg 2009b). In a sense the weights p_i are those as close to the original weights q_i as possible, while obeying all the constraints.

The set of weights p_i obtained through the cross-entropy estimation were converted to "raising weights" by multiplying them by the population total 48,687,000 as given in the mid-year population estimates. The distribution of the NIDS sample when reweighted with these weights is shown in Table 2. The total in that table is 48,024,624 which is 1.36% below the figure of 48,687,000 due to individuals with missing age, gender or race information. The program used to calculate the weights is available (Wittenberg 2009a).

Table 2: Counts in each Age-Sex-Race cell when applying the post-stratified weights

Age band	AM	AF	CM	CF	IM	IF	WM	WF
0-4	2,184,175	2,127,655	208,425	205,861	47,939	46,755	126,554	122,510
5-9	2,232,213	2,177,862	209,017	206,748	45,670	44,486	135,333	131,289
10-14	2,220,770	2,175,396	207,340	205,170	50,898	49,813	151,017	146,677
15-19	2,142,746	2,109,505	201,323	200,337	53,956	53,167	163,150	158,415
20-24	1,935,012	1,980,287	185,935	190,670	58,789	56,718	157,330	153,878
25-29	1,711,001	1,825,818	180,116	192,051	61,354	57,901	140,660	139,082
30-34	1,496,756	1,610,685	184,160	198,463	52,673	50,997	133,952	132,571
35-39	1,080,695	1,292,967	167,983	185,146	43,697	43,993	145,691	144,211
40-44	722,337	927,113	138,884	156,541	39,456	40,738	166,898	165,418
45-49	669,368	867,929	119,847	136,320	36,990	38,371	169,068	172,126
50-54	566,783	743,742	93,609	108,997	33,242	35,116	164,432	170,449
55-59	442,300	583,946	68,160	83,153	28,606	31,269	151,116	158,218
60-64	337,742	458,279	47,051	61,058	21,898	25,153	134,051	146,381
65-69	241,075	351,453	30,578	43,895	14,500	18,248	98,541	114,619
70-74	154,568	252,714	19,235	32,946	8,779	12,724	63,425	84,140
75-79	85,816	158,316	9,864	20,320	4,833	8,187	35,609	59,973
80+	52,575	120,833	5,622	14,796	3,058	6,609	29,592	71,612
Total	18,275,931	19,764,500	2,077,151	2,242,471	606,337	620,246	2,166,420	2,271,569

References

Golan, Amos, George Judge, and Douglas Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, Chichester: Wiley, 1996.

Horvitz, D.G. and D.J. Thompson, "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 1952, *47* (260), 663–685.

Wittenberg, Martin, "An introduction to maximum entropy and minimum cross-entropy estimation using Stata," School of Economics and SALDRU, University of Cape Town 2009.

___ , "Sample Survey Calibration: An Information-theoretic perspective," School of Economics and SALDRU, University of Cape Town 2009.