B

WAVE 1

National Income Dynamics Study (NIDS) – Coronavirus Rapid Mobile Survey (CRAM)

Sample design and weighting in the NIDS-CRAM survey

Andrew Kerr - University of Cape Town
Cally Ardington - University of Cape Town
Rulof Burger - University of Stellenbosch

15 July 2020

N.i.D.S.
NATIONAL INCOME DYNAMICS STUDY

CRAM
CORONAVIRUS RAPID MOBILE SURVEY 2020

Sample Design and Weighting in the NIDS-CRAM Survey

Andrew Kerr, Cally Ardington and Rulof Burger [1]

15 July 2020

Data Access

The NIDS-CRAM data is publicly available for download and use from the Data First portal www.datafirst.uct.ac.za

[1] We thank Martin Wittenberg, Nicola Branson, Tim Brophy, Reza Daniels and Kim Ingle for very helpful inputs on the sample design and usage of the NIDS wave 5 data as the sample frame.

The NIDS-CRAM Sample Design

The sample frame for NIDS-CRAM is NIDS wave 5, which was conducted in 2017. Continuing sample members (CSMs) and temporary sample members (TSMs) who were 18 years or older at the time of the NIDS-CRAM wave 1 fieldwork in April 2020 were re-interviewed. The NIDS-CRAM sample is drawn using a stratified sampling design but with “batch sampling”.

This batch sampling method was designed to allow flexibility to adjust the sampling rate in each stratum as information about stratum response rates became available as the “fieldwork” progressed. The motivation is that there was substantial uncertainty about both the level and drivers of non-response to a telephone survey and about the sample size that was possible given the response rates and budget and time constraints. Batch sampling means that sampled individuals were sent to the fieldwork team in batches of 2500 individuals. The individuals are randomly drawn within each of 99 strata, which are a combination of household per capita income decile, race, age and urban/rural.

As information about the response rates from the initial batches of 2500 respondents each was obtained, the batch sampling method allowed for changes in the number of individuals sampled in each stratum in the subsequent batches. Specifically, more individuals from the strata where response rates were lower were sampled, and fewer individuals were sampled from strata where response rates were higher. This process continued until the target number of successfully interviewed respondents in each stratum was obtained, or the individuals in the stratum have been exhausted. The final sample size was 17568 individuals, of which approximately 40% responded, giving a realised sample size of 7074 completed interviews.

The sample design and sampling process followed thus incorporates a non-response adjustment, albeit in a different manner to the usual type of adjustments that are made ***after*** the fieldwork is complete. The adjustment is that the sampling rate is modified ***during the survey*** to oversample strata with low response rates. Individuals who did not provide NIDS with any primary or alternate contact telephone numbers in wave 5 and who could therefore not be contacted are also included as non-respondents in the stratum they were in. Further non-response adjustments were possible after the fieldwork was complete, as we explain below.

Weighting

The NIDS-CRAM survey is drawn from NIDS wave 5, a broadly representative sample of individuals resident in South Africa in 2017, who are then re-interviewed in 2020. The probability of selection of an individual i in stratum s in NIDS-CRAM conditional on being in the NIDS wave 5 sample is the final sampling rate in the individual’s stratum. This means that the design weight for NIDS-CRAM is:

$$NIDS\ CRAM\ \ design\ weight_{is} = weight\ NIDS\ wave\ 5_{\ i} * \ 1/(\ \text{sampling rate}_s)$$

Non-response adjustment

The response rate in NIDS-CRAM was approximately 40%. 48% of the sample was classified as a non-contact, ie the fieldwork team was not able to speak to the individual, either on the telephone number provided or on a number of alternative contact numbers (including the phone numbers of co-resident household members in 2017). This is a concern because respondents may be different to those who could not be contacted. A further 8% of the selected respondents were classified as a refusal- they were contacted but refused to be interviewed.

As discussed above, the sampling process incorporated a non-response adjustment by oversampling strata where strata response rates in the initial batches was low. Further non-response adjustments were made because non-response was not random within strata. This further non-response adjustment is similar to the NIDS wave 1-5 panel weight adjustments for attrition (Branson and Wittenberg, 2019). We multiply the design weight by the inverse of the conditional probability of being interviewed. This conditional probability was estimated from a probit regression with a dummy dependent variable where 1 is a response and 0 is non-response. The explanatory variables in the regression are the NIDS-CRAM stratum, the individual's race, gender, language, log of household per capita income in wave 5, an urban dummy, the individual's province, their wave 5 employment status, wave 5 household size and whether or not an individual was successfully interviewed in wave 5.

Thus, the NIDS weight for individual ***i*** in stratum ***s*** adjusted for non-response is:

$$\begin{aligned} &NIDS\ CRAM\ \ panel\ weight_{is} \\ &\qquad = post\ stratified\ weight\ NIDS\ wave\ 5_{\,i} * \ 1/(\text{ sampling rate}_s \\ &\qquad * conditional\ probability\ of\ response_s) \end{aligned}$$

A final adjustment is made to the weights by trimming. Weights below the 1st percentile of all weight values were set to the 1st percentile and those weights above the 99th percentile were set to the 99th percentile. This is the same process that was undertaken for the panel weights in NIDS waves 1-5 (see Branson and Wittenberg, 2019).

Representativity of the Weighted Sample

Because NIDS-CRAM survey respondents are selected and re-interviewed from NIDS wave 5, a broadly representative sample of individuals resident in South Africa in 2017, NIDS-CRAM is a panel. **There is therefore no post-stratification adjustment to the weights so the weighted sample does not match the South African population in 2020. Instead, one should think of the weighted NIDS-CRAM survey data as reflecting the outcomes for a broadly representative sample of those 15 years and older in 2017 who were followed up 3 years later.**

In interpreting or discussing results from the NIDS-CRAM survey researchers should note this. For example, in discussing employment losses, a researcher should note "For a broadly representative sample of South African adults from 2017, who were re-interviewed in 2020 for NIDS-CRAM, the estimated employment loss between February and April 2020 is 2.7 million." There is also statistical uncertainty about any estimate due to NIDS and NIDS-CRAM being survey, and we discuss variance estimation below.

The extent to which NIDS wave 5 was indeed representative of South Africa in 2017 depends on the extent to which the NIDS wave 5 weights, adjusted for attrition and non-response and post-stratified on gender, race, age and province, compensated for attrition, migration and non-response that had occurred since NIDS wave 1 in 2008. If the weighted NIDS wave 5 respondents did not match the South African population in 2017, due to attrition on unobservable characteristics between 2008 and 2017 for example, and the migration of individuals into South Africa between 2008 and 2017, then this will also mean that the 2020 NIDS-CRAM weighted data will not be representative of the changes in the life circumstances of individuals living in South Africa in 2017. This potential negative is far outweighed by three important benefits. The first is that NIDS collected the respondents' contact numbers, allowing for a telephone survey while in person fieldwork was impossible due to the lock down. The second is that researchers have potentially 5 waves of NIDS data on the respondents stretching back to 2008. The third is that it allows for a rich set of data to model non-response in CRAM, described above. This is particularly advantageous given the low response rates and unknown sample frames typical of other telephone surveys.

Obtaining correct variance estimates

The original NIDS sample design was a two-stage cluster sample with stratification. Researchers using NIDS-CRAM must take this original sample design of NIDS into account when estimating variances. To do this, researchers should use the original cluster an individual's household was drawn from in NIDS, the NIDS-CRAM weight, discussed above, and the original NIDS stratum variable, which was the district council. These variables are provided in the NIDS-CRAM data. In Stata the ***svyset*** command should be used, specifying the original NIDS cluster and stratum, as well as the NIDS-CRAM weight adjusted for non-response.

Stata code to estimate the variance correctly is:

```
svyset cluster [weight = weight], strata(stratum)
```

If there is only 1 observation in the original stratum Stata will report missing variances. The *singleunit(scaled)* option can then be used in the *svyset* command above.

Household Level Analysis

NIDS-CRAM sampled individuals from NIDS wave 5. Unlike previous waves of NIDS, NIDS-CRAM did not attempt to interview or collect information on everyone currently living with the sampled individual. This change in sampling protocol was carefully considered taking into account the main goals and constraints of the NIDS-CRAM. The limits of telephonic surveys with respect to questionnaire length and complexity was a key factor. No attempt was made to check whether successfully re-interviewed individuals resided in the same households as they did in wave 5. Also, individuals from larger households are more likely to be sampled than individuals from smaller households. Researchers can therefore not use NIDS-CRAM to conduct household-level analysis. However, it is possible to estimate statistics at an individual-level about household living conditions. It would be legitimate to state “For a broadly representative sample of adults from 2017, who were re-interviewed in 2020, we estimate that X% of adults live in households receiving a government grant”. However, it is not legitimate to estimate that Y% of households received a government grant. In the same way one could say that Z% of adults live in households where children went hungry, but not that A% of households had children going hungry or that B% of children went hungry.

Further Stratification for NIDS-CRAM

NIDS waves 1-5 was a stratified sample. District councils were the strata. For the selection of individuals for NIDS-CRAM the sample from NIDS wave 5 was further stratified. These new strata should NOT be used to set the complex survey design. Please read the section “Obtaining correct variance estimates” above for the correct way to estimate the variance in Stata, using the original stratum and cluster variables from NIDS.

Further stratification was used in NIDS-CRAM for two reasons. The first is so that the NIDS-CRAM sample would be representative of the NIDS wave 5 sample, given that the final sample size was uncertain at the start of the fieldwork. The second was that the strata would allow for non-response adjustments. The aim was to have strata that were reasonably likely to have responses that were missing completely at random within stratum. In the end 99 strata were created from the NIDS wave 5 respondents eligible to be surveyed for NIDS-CRAM. The strata are formed from five NIDS wave 5 variables, but not on the complete combination of all of them. The sampling rate varied by age category. Individuals between 30 and 50 were sampled at twice the rate of individuals in other age groups.

The stratification variables are age group categories (<30, 30-50, 50-59, 60-69 and 70+), income decile, urban/rural, race and gender. None of the 5 age groups nor the 10 income groups are combined together in a stratum but some of the other groups are combined together.

We first stratify on age and income decile. There is then no further stratification for 70+ age group. There are 10 income deciles for 70+, meaning there are 10 strata.For income deciles 1-6 we further split the age group-income decile strata by urban and rural. There are 4 age groups x 6 income deciles x 2 urban/rural, meaning there are 48 strata.

For income decile 7-9 we further split these age group- income decile strata by Black or other. There are 4 age groups x 3 income deciles x 2 race, meaning there are 24 strata.For income decile 10 we split these age group- income decile strata by White or other. There are 4 age groups x 1 income decile x 2 race, meaning there are 8 strata.

This means we have a total of 90 strata. We then split any strata larger than 600 by further stratifying by gender. There are 9 of these strata, so they become 18. That means we finally have 99 strata in total. The stratum size range is 46- 660. The mean is 261, the median is 205.

REFERENCES

Branson, N. and Wittenberg, M. (2019). Longitudinal and Cross-Sectional Weights in the NIDS Data 1-5. NIDS Technical Paper 9.

For further information please see **cramsurvey.org** and **nids.uct.ac.za**