SALDRU

Southern Africa Labour and Development Research Unit

Creating household weights for NIDS-CRAM

Martin Wittenberg and Nicola Branson

N.i.D.S.

NATIONAL INCOME DYNAMICS STUDY

CRAM

CORONAVIRUS RAPID MOBILE SURVEY 2020

Working Paper Series

Number 280, Version 1

About the Author(s)

Martin Wittenberg: School of Economics and DataFirst, University of Cape Town
Nicola Branson: SALDRU and School of Economics, University of Cape Town

Acknowledgments

We would like to thank Tim Brophy, Reza Daniels and Kim Ingle from NIDS-CRAM for access to the sampling frame and the sampling files and to Cally Ardington, Rulof Burger and Andrew Kerr for access to the stratification codes.
This paper was funded by the CRAM study and is available on their website.

Recommended citation

Wittenberg, M., Branson, N. (2021). Creating household weights for NIDS-CRAM. Cape Town: Southern Africa Labour and Development Research Unit, University of Cape Town. (SALDRU Working Paper Number 280)

ISBN: 978-1-928516-42-2

© Southern Africa Labour and Development Research Unit, UCT, 2021

Working Papers can be downloaded in Adobe Acrobat format from http://opensaldru.uct.ac.za. A limited amount of printed copies are available from the Office Administrator: SALDRU, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 1808, Fax: (021) 650 5697, Email: amy.jephthah@uct.ac.za

Creating Household Weights for NIDS-CRAM*

Martin Wittenberg and Nicola Branson
School of Economics, SALDRU and DataFirst
University of Cape Town
Cape Town, South Africa

July 2021

Abstract

NIDS-CRAM is widely used to investigate the impact of the COVID pandemic on individuals and households. However, because NIDS-CRAM is a survey of **individuals** it is difficult to make accurate statements about **households**. Nevertheless many issues of interest, such as the hunger questions in NIDS-CRAM, are about the household and not just the respondent.

The problem with using the existing NIDS-CRAM weights for these analyses is that there is double-counting: there are potentially many individuals from the same household in the NIDS-CRAM survey. We show that overlapping membership affects between 40% to 50% of the observations.

In this paper we lay out the theory for dealing with this problem and generate a set of "household weights" to reduce the double-counting. We use these weights to produce some initial estimates of how prevalent hunger might have been during the lockdown.

Paradoxically estimates of the fraction of households affected by hunger are not changed much by using the household weights rather than the person weights released with NIDS-CRAM. The reason for this is that hunger is only very weakly associated with household size, so the double-counting implicit in using the person weights does not skew the estimates much. However if one wants to generate estimates of the **number** of households or people affected by hunger the household weights make a much bigger difference. Indeed, we generate a first set of numbers that quantify the problem. For instance somewhere between 1.5 million and 3.1 million children were

*We would like to thank Tim Brophy, Reza Daniels and Kim Ingle from NIDS-CRAM for access to the sampling frame and the sampling files and to Cally Ardington, Rulof Burger and Andrew Kerr for access to the stratification codes.

affected by hunger at the time of the field work for NIDS-CRAM wave 5. These estimates have to be treated with some caution, because our weights do not properly deal with changes in the distribution of households since 2017, in particular new household formation.

1 Introduction

NIDS-CRAM (2020a, 2020b, 2021a, 2021b, 2021c) has been one of the most widely used sources of information about the impact of the COVID pandemic on individuals and households. Nevertheless as the panel user guide points out (Ingle, Brophy and Daniels 2020, section 7.1, p.5), the household level information has to be treated with caution:

> 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 [state] that Y% of households received a government grant.

The situation is schematically depicted in Figure 1. Information is collected at two levels in a survey: individual (age and gender) as well as household (household size and prevalence of hunger). If we analyse the household level information (e.g. household size) over the distribution of individuals we get misleading information. For instance, adding up household size over the **individual** file we would get a "total population" of 30 (in the example of Figure 1) whereas the true count is 10. The problem is the double-, triple- and quadruple-counting of the information in large households.

The problem, of course, is that in NIDS-CRAM we only have a sample of **individuals**. So, although we have information on the households of those individuals, standard estimators will yield estimates for the frame from which those samples were drawn, i.e. they will provide estimates for the population of individuals[1] not the population of households. This is schematically depicted in Figure 2.

Although this problem is well understood (as indicated in the quote above), the temptation to slide from individual level statistics to household level ones is very strong. For instance van der Berg, Patel and Bridgman (2021, p.1) report:

[1] actually, only adult individuals

Household	Person	Age	Male	Size	Hunger
1	1	55	0	3	1
1	2	33	0	3	1
1	3	5	1	3	1
2	1	60	1	4	1
2	2	55	0	4	1
2	3	25	0	4	1
2	4	3	0	4	1
3	1	30	1	1	0
4	1	24	1	2	0
4	2	22	0	2	0

Household	Size	Hunger
1	3	1
2	4	1
3	1	0
4	2	0

Figure 1: Information in surveys is often collected at different levels – individuals and households. This does not cause problems if the data is analysed at the appropriate level.

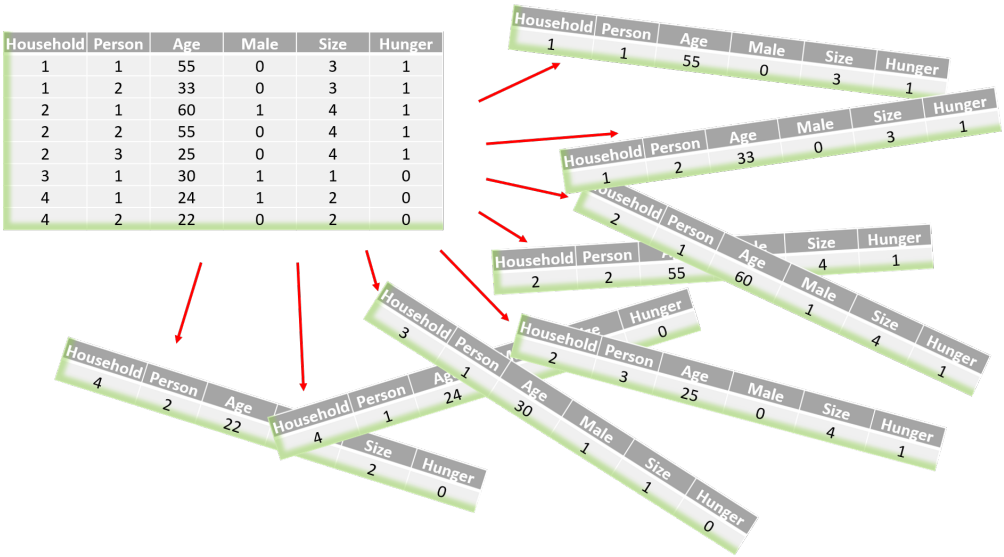

Figure 2: Sampling in NIDS-CRAM is from a frame of individuals, so the household level information cannot be connected directly back to a frame of households.

> Overall, the proportion of households who reported running out of money for food went from 47% in Wave 1 to 38% in Wave 2 – a significant decrease. In Wave 3, however, the proportion rose again to 41%

One could rewrite this in terms of "the number of respondents who reported that their households ran out of money" but policy makers and the general public are unlikely to understand the distinctions. Furthermore, what respondents report on their households is typically not the real object of interest.

In this paper we do several things:

- We document the extent to which the NIDS-CRAM survey is subject to "double-counting" of households.
- We present an approach for how to calculate statistics that are more informative about the household level. This weighting strategy has to be treated with some caution, because it cannot deal with household changes that have occurred since 2017. It is, however, more defensible than using the NIDS-CRAM weights for this purpose.
- We sketch out what an "optimal" weighting strategy might look like and take some steps towards implementing it.
- We show that individuals who appear to be from the same household sometimes report different information. That can be due to changes in living arrangments. But we observe differences even when such changes do not seem to have occurred, suggesting that researchers need to consider the possible impact of measurement error on their findings.
- Despite the potential problems of "double-counting" and mismeasurement it turns out that the statistics reported in the literature, e.g. in the quote from van der Berg et al. (2021) cited earlier, are not far off from the "better" results that we calculate.

We begin our discussion, in the next section, with a quick examination of the extent of double-counting in NIDS-CRAM. We then present the general approach to estimating household level statistics. It is based on the straightforward idea that the household level quantity needs to be "shared out" between its constituent members to eliminate double counting. There are many ways of doing so. We argue that the "optimal" rule for doing so, in an ideal world, is proportional to the odds of an individual being selected.

However, this is not an ideal world. The first major complication is that we do not have proper design weights. Instead, nonresponse is a centrally important

issue. NIDS-CRAM dealt with this problem at two levels: increasing the sampling rate in strata that were hard to reach and a non-response adjustment to the weight. We show in section 4 how to incorporate such adjustments into the calculation of household level statistics.

When we turn from household level variables measured on the 2017 sampling frame (i.e. where there is no uncertainty) to household variables measured in NIDS-CRAM itself, we also need to deal with the possibility of measurement error. In section 7 we argue that its effect is to shift the "optimal" household sharing rule towards sharing the household quantities proportional to the probability of an individual being selected.

The final, and most difficult, set of complications arise from the fact that that the household distribution has itself changed between 2017 and 2020. In section 8 we sketch out what would be required to deal with this issue. We hope to tackle that topic in future work. In the conclusion we attempt to draw some lessons from our empirical work.

We try to keep the technical details in the main text to a minimum. A more technical discussion of the procedure as well as how it relates to the NIDS-CRAM sampling scheme is given in the appendix (section A).

2 Double-counting in NIDS-CRAM

Table 1 presents the basic information on the level of potential double counting in NIDS-CRAM. Of the 26 889 individuals on the frame (i.e. aged 15 or older in NIDS wave 5) 20 646 were sampled in one form or another in NIDS-CRAM. 81% of these sampled individuals were co-resident in 2017 with another individual who got sampled. However we see (in panel C of Table 1) that this overlap reduces to 53% when we consider only respondents in NIDS-CRAM wave 1. The non-response process could either have aggravated or ameliorated the problem of double-counting. We see that in this case it substantially reduced the problem. Nevertheless the overlap is still substantial. Of course part of this "overlap" is spurious – two respondents from the same 2017 household could now very well be living in different households. Panel D of Table 1 now considers only respondents who also indicated that they were still living in the same place where they were interviewed in Wave 5 of NIDS. Paradoxically the overlap is reduced, but it is still sizable.

Given the fact that double-counting of households affects a minimum of 40% of sampled cases and probably more, it is clear that using NIDS-CRAM respondent information to generate "household" level statistics will be dubious. We turn now to consider how this might be done.

Table 1: Extent of double counting: Percentage and number of NIDS-CRAM individuals in shared NIDS wave 5 households.

NIDS-CRAM		Percent from shared NIDS w5 household	Observations
A. Frame			26889
B. Sampled	All	0.809	20646
	Original	0.756	17568
	Top-up	N.A.[1]	3078
C. Respondents	Wave 1	0.527	7073
	Wave 2	0.477	5676
	Wave 3	0.511	6130
	Wave 4	0.495	5629
	Wave 5	0.495	5862
D. Respondents in "same" household[2]	Wave 1	N.A.[3]	
	Wave 2	0.419	4272
	Wave 3	0.428	4035
	Wave 4	0.444	4215
	Wave 5	0.439	4395

Notes:
1. 27.6% of top-up respondents shared households with other top-up respondents, but many of them also shared households with original sample members so the more relevant statistic is the one given for "All."
2. These are respondents who indicated that they are still living in the same 2017 dwelling unit
3. Information on whether living in same dwelling as when interviewed in NIDS wave 5 not asked in NIDS-CRAM wave 1

3 Estimating household level information

There is, of course, a straightforward fix for estimating household level quantities from an individual level dataset. To get from the individual frame to the population frame (e.g. in Figure 1) we need to keep one individual per household. We could achieve this by marking that individual (e.g. the "Household Head" or the oldest individual) on the frame and then estimating household quantities on the sample only over the marked individuals. We show in section A.1 that this procedure produces an unbiased estimate, provided that every household is represented on the frame.[2]

Observe that

- This is **not** equivalent to keeping one individual per household in the **realised** sample. There is no guarantee that the individuals in the sample include the "Household Head" or any other individual designated *ex ante* as household representative. Picking another individual based on the realised sample would make the variable designating the household representative stochastic. Indeed the estimator would no longer be unbiased. Households with more members would have many more ways of contributing to the sample estimates than households with just one individual.

- It is clear from the previous point that this would be a very noisy estimate, since there will be a lot of households that are represented in the sample that will not contribute to the estimate, because the sampled individual is not the "Head" or *ex ante* designated representative. We are throwing away a lot of information.

One way of getting around that problem is to "share out" the household level variable x_h between all the household members that are represented on the frame. In the case of NIDS-CRAM that would be adult individuals. In fact any set of shares s_{ih} that add up to one for individuals on the frame from the same household will work, as we show in section A.2. Figure 3 shows three possible rules.

The simplest rule would be to allocate the shares equally across members from the same household on the frame. Note again that this rule needs to be applied on the **frame** and not on the **realised sample**.

It turns out that there is an even more efficient sharing rule. We show in section A.3 that the approximately optimal sharing rule is to allocate shares proportional to the odds of selection:

$$s_{ih} \propto \frac{\pi_{ih}}{1 - \pi_{ih}} \tag{1}$$

[2]In the context of NIDS-CRAM this means that we assume that there are no all-children households.

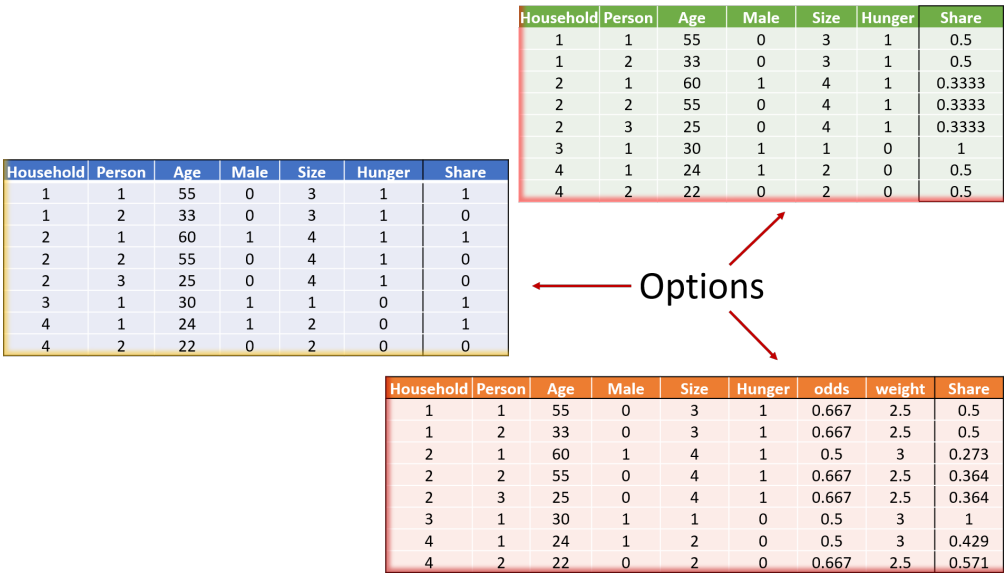

Figure 3: Any set of shares that add up to one within households on the frame will lead to unbiased estimates of household quantities. Picking one representative per household (left-most box) will do so, but will lead to more noisy estimates than creating equal shares (top box). Theoretically the most efficient sharing rule is to do so inversely proportional to the odds of selection (bottom right: assumed sample design – 1 male and 2 females by simple random sampling).

The intuition is that observations that have a **low** probability of selection will have a high weight. That, however, will induce greater variance in the estimates, depending on whether or not that observation is included. By allocating it a smaller share we are making the estimates more stable.

In the empirical results we use four household sharing rules to generate estimates of household level statistics:

- A household "representative" rule, i.e. picking the oldest member on the frame
- Equal shares within households
- Shares proportional to the odds of selection
- Shares proportional to the probability of selection

The motivation for using the last rule is sketched out in section 7 below (and in more detail in section A.5).

4 Dealing with NIDS-CRAM sampling and Non-response

The NIDS-CRAM sampling and weighting strategy is more complex than is the case in standard surveys (Kerr, Ardington and Burger 2020). In particular it is impossible to separate out a fixed sampling probability from the nonresponse correction (see section A.4).

4.1 Assuming uniform non-response within NIDS-CRAM strata

Nevertheless, we show in section A.4.1 that it is possible to define a set of weights on the sampling frame that will produce unbiased results, provided that the probability of response is constant within NIDS-CRAM strata. It leads to a very simple "sharing rule" within households:

$$s^*_{ih(t)} \propto \frac{n_t}{N_t - n_t} \tag{2}$$

Here individual i in household h is of type t (i.e. this is the NIDS-CRAM stratum t); n_t is the number of NIDS-CRAM sample members of type t and N_t is the number of people on the original frame of type t.

4.2 Dealing with non-uniform non-response

The assumption of a constant probability of non-response within NIDS-CRAM strata turns out to be empirically dubious. Consequently the NIDS-CRAM sampling team made an additional non-response adjustment, estimating a probit model for response on the final sample (Kerr et al. 2020). We present a discussion of this procedure in section A.4.2. We use these weights in our empirical work, although we cannot use them to calculate the household shares, since we do not have them for every individual (including non-sampled ones) on the sampling frame.

Note that, as in all cases of inverse probability weighting, the NIDS-CRAM procedure works if, and only if, the distribution of the variable of interest X_h is the same among respondents as among non-respondents, conditional on the explanatory variables $\mathbf{x}'_{ih}$, i.e. we need

$$X_h^0, X_h^1 \perp\!\!\!\perp D_{r\,ih} | \mathbf{x}'_{ih} \tag{3}$$

where $D_{r\,ih}$ is a dummy variable equal to one if individual ih responded and X_h^0 is the outcome variable on X_h if $D_r = 0$ and X_h^1 the case among individuals where $D_r = 1$. In the case of NIDS-CRAM the propensity scores were estimated on a mix of individual and household level variables.

As we do not observe the distribution of X_h^0, we cannot check whether condition 3 holds. It is also likely to depend on the type of variable that is examined.[3]

In summary the weights that we use in the work below are a combination of the within household "sharing rules" s_{ih} and the NIDS-CRAM official weights w_{ih}^{NCp}. Our preferred set of weights (and the ones released publicly) use the "optimal" shares, i.e.

$$w_{ih}^{NC\,hh} = w_{ih}^{NCp} s_{ih}^{*} \tag{4}$$

5 How well do the "Household weights" perform?

The first key issue is the extent to which the household weights manage to reproduce various summary statistics of the NIDS wave 5 household distribution. A first cut is provided by Table 2. It shows that all of the four sharing rules that we consider produce estimates that are close to the ones that we produce with the full NIDS wave 5 sample. By contrast if we had used the NIDS-CRAM person

[3]We could investigate this in the context of the 2017 household variables, since we do have these available on the frame for responders and non-responders. The diagnostics that we show in the next section show that at least for some of these 2017 variables, the combination of household sharing rule and the NIDS-CRAM propensity score adjusted weights do a reasonable job of recovering the original distribution.

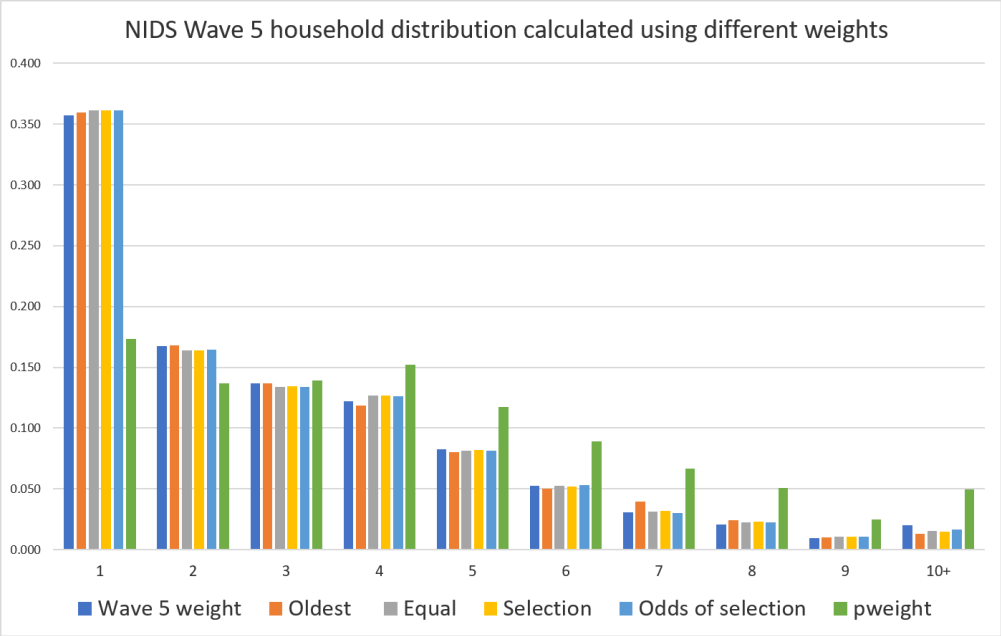

Figure 4: The household size distribution in 2017 using NIDS wave 5 and various NIDS-CRAM wave 1 weights.

weights we would get seriously misleading estimates of average household size and the proportion of single person households. The cost of double-counting is also evident when we look at the estimates of the number of two person households.

In Figure 4 we show the impact of the different weights on estimates of the distribution of household sizes in 2017. We already saw (in Table 2) that using the NIDS-CRAM person weights would lead to a major underestimate of the proportion of one person households. This Figure shows that we would concomitantly markedly over-estimate the proportion of larger households. By contrast, using any of the household sharing rules provides reasonable estimates.

Consulting Table 2 again, we observe that although there are no major differences between the different sharing rules, the estimated standard errors associated with s^*_{ih}, the sharing rule based on the odds of appearing in the sample, seem to be smallest. Since our theory also favours this rule, this is the version of the NIDS-CRAM "household weight" that is distributed.

Table 2: Estimates of the 2017 distribution of households using NIDS wave 5 and various NIDS-CRAM wave 1 weights

		NIDS-CRAM household weights/sharing rule				**Person**
	NIDS Wave 5	Oldest[1]	Equal[2]	Odds[3]	Probability[4]	**Weight**[5]
	Number of households					
b	18 615 108	17 933 652	17 854 518	17 842 200	17 867 488	
se	920 881	814 186	776 435	773 561	780 274	
ll	16 806 146	16 336 249	16 331 182	16 324 501	16 336 619	
ul	20 424 072	19 531 056	19 377 856	19 359 898	19 398 358	
	Number of two person households					
b	3 114 728	3 010 418	2 931 563	2 939 283	2 925 858	5 094 640
se	209 532	227 306	188 023	185 776	191 588	338 846
ll	2 703 126	2 564 453	2 562 669	2 574 798	2 549 969	4 429 838
ul	3 526 330	3 456 384	3 300 457	3 303 768	3 301 747	5 759 443
	Average household size					
b	3.029	2.999	3.000	3.003	2.999	4.306
se	0.048	0.073	0.066	0.066	0.067	0.087
ll	2.935	2.856	2.870	2.873	2.867	4.137
ul	3.123	3.143	3.130	3.132	3.131	4.476
	Proportion single person households					
b	0.357	0.360	0.361	0.362	0.361	0.174
se	0.012	0.017	0.017	0.017	0.017	0.011
ll	0.335	0.326	0.329	0.329	0.328	0.153
ul	0.380	0.394	0.394	0.394	0.394	0.195

Notes:
1. Oldest member "represents" household
2. Equal shares among all members
3. Shares proportional to **odds** of appearing in sample, i.e. "optimal" shares s^*_{ih}
4. Shares proportional to probability of being in sample
5. Using NIDS-CRAM person weights

b: point estimate, **se:** standard error, **ll,ul:** 95% C.I. bounds

6 Applying the household weights to NIDS-CRAM measures

The fact that our NIDS-CRAM household weights manage to reproduce the 2017 household distribution is reassuring. Of course we are interested mainly in applying the weights to NIDS-CRAM variables. Our first cut at this analysis is shown in Table 3. There are several interesting results:

- The average household size according to the NIDS-CRAM results is considerably larger than the average household size in NIDS wave 5. A different look at this problem is provided by Table 4 which suggests that respondents in NIDS-CRAM did, indeed, report more residents in their households than there should have been. There are several potential explanations for this. Firstly, the NIDS-CRAM variable does not impose the stringent conditions that listing in a household roster requires.

  Secondly, it is possible that the NIDS-CRAM sampling process disproportionately picked up individuals whose households **gained** members rather than individuals in households **losing** members.

  Thirdly, it should be noted that if the members gained are "temporary sample members" in NIDS terminology, i.e. individuals who could have been sampled in NIDS but were not, then these additional people should not really be counted. In NIDS this is handled by "sharing out" the original household weight among all the new household members, so that the total population count does not go up. This is obviously not possible to do with the information that we have in NIDS-CRAM.

  Finally, it is possible that the lockdown itself may have persuaded people who were previously living apart to move "back" to their households of origin.

- While all the weights suggest a bigger household size in NIDS-CRAM wave 1, Table 3 shows that the problem is much more pronounced with the double-counting characterising the NIDS-CRAM person weights. Conversely, the household weights also provide more reasonable estimates of the proportion of single person households.

- Remarkably, however, the "hunger" and "no money" means are not really different whether one uses the person weights or the household ones. It suggests that the prevalence of hunger is only weakly related to household size, so that the extra attention paid to large households when the person weights are used does not distort the overall picture.

Table 3: Household variables in NIDS-CRAM wave 1 calculated with different weights.

	Person	**Household weight/sharing rule**			
	Weight	Oldest	Equal	Probability	Odds
	NIDS CRAM Average household size				
b	4.976	4.364	4.309	4.308	4.308
se	0.078	0.083	0.068	0.070	0.067
ll	4.824	4.201	4.175	4.171	4.177
ul	5.128	4.528	4.443	4.445	4.440
	Proportion single person households				
b	0.083	0.139	0.140	0.139	0.140
se	0.007	0.012	0.011	0.011	0.011
ll	0.070	0.116	0.118	0.118	0.118
ul	0.096	0.162	0.162	0.161	0.162
	Hunger in household[1]				
b	0.223	0.222	0.219	0.217	0.221
se	0.008	0.011	0.010	0.010	0.010
ll	0.206	0.201	0.199	0.197	0.201
ul	0.239	0.244	0.238	0.237	0.240
	No money for food[2]				
b	0.470	0.465	0.459	0.455	0.464
se	0.011	0.016	0.014	0.014	0.014
ll	0.448	0.434	0.432	0.429	0.437
ul	0.491	0.496	0.485	0.482	0.491

Notes:
1. In last 7 days, has anyone in your HH gone hungry due to lack of food? (proportion answering yes)
2. In the month of April, did your household run out of money to buy food? (proportion answering yes)
b: point estimate, **se:** standard error, **ll,ul:** 95% C.I. bounds

Table 4: Estimates of the total population in NIDS-CRAM[1]

Age[2]	MYPE[3]	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
All	58 667 434	76 236 976	74 897 936	74 794 272	72 906 432	71 581 520
0-6	8 029 831	10 119 548	10 255 757	12 064 054	11 223 592	11 069 304
7-17	10 930 956	8 386 181	8 800 817	15 988 256	15 294 999	14 843 844
18-60	34 639 421	38 546 772	30 916 134	40 932 388	39 868 828	39 617 384
> 60	5 067 227	7 977 114	6 383 144	6 129 779	6 510 891	6 050 929

Notes:
The totals in the age categories need not add up to the overall total; they are derived from different questions.
1. Using household weights with shares proportional to the odds of being in the sample.
2. Age categories changed in different waves of NIDS-CRAM:
Wave 1: All, over 60, under 7, under 18. **Wave 2:** All, over 60, under 7, under 18, 18-60. **Waves 3-5:** All, under 7, 7-17, 18-60, over 60.
3. 2020 Mid-Year population estimates from Statistics South Africa

These findings are provocative and require further investigation. Indeed one of the payoffs of having a framework for thinking about households is that new avenues of enquiry become possible.

While the NIDS-CRAM person weights do not distort the proportions of households affected by hunger, they will obviously distort attempts to arrive at sensible counts of people affected by hunger. And policy makers are often interested in getting estimates of these quantities – even if they are rough.

Table 5 presents a first cut at these questions. Panels A and B of that table present estimates of the total number of households affected by hunger in general or child-hunger more specifically. This is likely a **lower bound** on hunger, since it would also be the point estimate if there was only one person (or child) in the household experiencing hunger. Panels C and D present estimates of the **upper bound** of hunger, since these estimates assume that every member of the affected household experiences hunger. Given that the NIDS-CRAM population estimates are too high, as shown in Table 4, and that the likely cause of this is our inability to correct for new "temporary sample members," we rescaled the estimates to align the NIDS-CRAM population totals with those of the Mid-Year Population Estimates.

The numbers suggest that hunger affected between 2.7 million and 10.6 million people at the time of the NIDS-CRAM fieldwork in April 2021. Somewhere between 1.5 million and 3.1 million children were affected. Table 5 also suggests that the overall levels of hunger have come down somewhat since the beginning of

lockdown in March 2020, although perhaps not as much as one might have hoped.[4]

7 Measurement error

Of course our discussion thus far has assumed that all measures of the household variable that we are interested in are equal. That is true by definition when we use any of the 2017 household distribution measures as we did in Figure 4 or Table 2. Once we turn to NIDS-CRAM variables, however, as we did in Tables 3 and 5 this is no longer the case.

Indeed, there are some differences in the answers given by respondents from the same 2017 NIDS household, as is shown in Table 6. Panel A of the table breaks down how the question on hunger in the household was answered. The columns labelled "Overall" show that about a quarter of the 7016 respondents indicated that someone in their household had gone hungry in the last week. The columns labelled "Between" indicate that hunger was recorded in 1 605 of the 4 903 NIDS wave 5 households, and in 3 872 of the households, a respondent denied that there had been hunger. Some households were obviously put into both categories, i.e. different respondents from the "same" household reported divergent results. The last column ("Within") quantifies the mismatch. It shows that there was only 81% agreement among respondents from households that seem to have been affected by hunger. By contrast, there was much more agreement (93%) in "households" where no hunger was reported.

Panel B shows that there was also at least some disagreement among respondents from the same NIDS wave 5 household whether the household had run out of money to buy food.

There are several reasons why such mismatches might occur. Firstly, it is not evident that every household member is equally well informed about the experiences of other household members or, indeed, whether the household ran out of money. Secondly it is possible that at the time of the NIDS-CRAM survey respondents from the same NIDS wave 5 households might have been in different locations and reporting on different households. There is at present no way to separate out measurement error from real changes in the household.

In the presence of measurement error it is no longer self-evident that the optimal estimator defined in equation 1 will still be optimal. In the appendix we suggest that it might be more robust to pick shares within the household proportional to the probability of selection:

$$s_{ih} \propto \pi_{ih} \tag{5}$$

[4]The numbers for wave 2 look uniformly on the low side. This requires further analysis.

Table 5: Estimates of the prevalence of hunger

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
	A. Hunger in household[1]				
b	3 389 325	2 268 539	3 254 408	2 785 905	2 755 242
se	235 671	180 851	246 199	224 860	237 345
ll	2 926 279	1 913 181	2 770 664	2 344 074	2 288 881
ul	3 852 372	2 623 896	3 738 153	3 227 736	3 221 603
	B. Hungry child in household[2]				
b	1 787 811	1 319 222	1 747 312	1 493 791	1 536 921
se	144 075	120 557	158 406	147 452	158 730
ll	1 504 733	1 082 337	1 436 067	1 204 059	1 225 031
ul	2 070 889	1 556 108	2 058 556	1 783 522	1 848 810
	C. People in household affected by hunger[3]				
b	13 453 487	9 564 752	12 202 783	11 180 678	10 570 047
se	963 891	837 901	1 042 865	1 103 910	1 091 721
ll	11 559 637	7 918 344	10 153 713	9 011 584	8 424 916
ul	15 347 337	11 211 159	14 251 854	13 349 772	12 715 179
	D. Children in household affected by hunger[4]				
b	3 421 147	2 697 636	3 499 001	3 148 362	3 137 427
se	304 247	274 671	350 747	378 330	403 564
ll	2 823 364	2 157 930	2 809 836	2 404 973	2 344 461
ul	4 018 930	3 237 342	4 188 166	3 891 750	3 930 393

Notes:
Household weights with shares proportional to the odds of being in the sample were used.
1. Count of households with a "yes" answer to "In the last 7 days has anyone in your household gone hungry because there wasn't enough food?"
2. Count of households with a "yes" answer to "In the past 7 days, has any child in your household gone hungry because there wasn't enough food?"
3. Count of household members in the households affected by hunger. The overall count scaled down to bring the NIDS-CRAM population in line with the MYPE as shown in Table 4.
4. Count of children in households affected by child hunger. The overall count scaled down to bring the NIDS-CRAM under-18 population in line with the MYPE as shown in Table 4.

Table 6: Reporting differences on hunger by respondents from the same NIDS wave 5 household

	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
A. Hungry[1]					
0	5160	73.55	3872	78.97	92.99
1	1856	26.45	1605	32.74	81.14
Total	7016	100	5477[2]	111.71	89.52
B. No money[3]					
0	3392	48.33	2766	56.29	86.3
1	3626	51.67	2897	58.95	87.23
Total	7018	100	5663[4]	115.24	86.77

Notes:
1. In last 7 days, has anyone in your HH gone hungry due to lack of food?
2. Distinct number of NIDS wave 5 households=4903
3. Did your household run out of money to buy food?
4. Distinct number of NIDS wave 5 households=4914

This accounts for the fourth "sharing rule" that we used in Figure 4 and Tables 2 and 3. In practice it seems to perform very similarly to our preferred sharing rule.

8 Changes in households

As suggested above, the biggest problem in the case of NIDS-CRAM, however, is the fact that household composition will have changed between 2017 and the time of the surveys. In order to think through the impact of such changes, the following issues would need to be dealt with:

1. The distribution of individuals across households would have changed, i.e. households could have split (or joined up) since 2017

2. Other individuals ("temporary sample members" in NIDS terminology) could have become co-resident with NIDS sample members. We have already noted that this is a possilbe culprit for the large average household size in NIDS-CRAM.

In the absence of other evidence one would need to model this. We sketch out an approach for how to deal with these issues in section A.6. For the moment these are beyond the scope of this paper.

9 Conclusion

The main purpose of this document has been to lay out a framework for thinking about households in the NIDS-CRAM sample. Our results suggest that the household weights that we constructed do permit interesting analyses. Nevertheless they should also be used with caution. While the weights reduce the problem of double-counting, they cannot provide proper estimates for the population of households in 2020 or 2021. To some extent they provide estimates for 2017 households but we know that "normal" social and economic processes will have reshaped the distribution since then. More to the point, the "abnormal" processes of the COVID pandemic and lockdowns will have had massive impacts on co-residency arrangements. The weights that we have discussed here will not help in capturing those except, perhaps, indirectly.

A Technical discussion

We follow the general procedure in Horvitz and Thompson (1952), i.e. we assume that the population information is fixed and only the sampling is stochastic. Let the total on variable X measured on the population of households be T_{pop}, i.e.

$$T_{pop} = \sum_{h=1}^{H} x_h \tag{6}$$

where x_h is the outcome of X on household h. Assume that the household level variable x_h is attached to every individual in that household so that

$$x_{1h} = x_{2h} = \cdots = x_{kh} = x_h$$

where we assume that there are individuals $1, 2, \ldots, k$ in household h.

A.1 The household representative estimator

For our first estimator we will pick one individual per household. Without loss of generality we assume that this is the first listed individual on the frame, i.e. let variable D_1 be defined so that it is equal to one for the first observation in every household on the frame and zero otherwise, i.e.

$$D_{1ih} = \mathbf{1}\left(i = 1\right)$$

where $\mathbf{1}\left(\right)$ is the indicator function.

Now define a_{ih} as the indicator variable indicating whether individual i in household h is sampled. Let π_{ih} be the probability of selection, given the fixed sampling design and $w_{ih} = \frac{1}{\pi_{ih}}$ be the corresponding Horvitz-Thompson weights.

Our sample estimator of the population total is

$$\hat{T}_{pop}^{rep} = \sum_{a_{ih}=1} w_{ih} D_{1ih} x_{ih} \tag{7}$$

We rewrite this as a sum over the entire population

$$\begin{aligned} \hat{T}_{pop}^{rep} &= \sum_{h=1}^{H} \sum_{i=1}^{k} w_{ih} a_{ih} D_{1ih} x_{ih} \\ &= \sum_{h=1}^{H} w_{1h} a_{1h} x_h \end{aligned}$$

In the last line we have used the fact that D_1 is zero except for the first individual per household on the frame and the measure x_{ih} is equal to x_h for all i.

Taking expectations

$$\begin{aligned} \mathbb{E}\left(\hat{T}_{pop}^{rep}\right) &= \sum_{h=1}^{H} w_{1h} \mathbb{E}\left(a_{1h}\right) x_h \\ &= \sum_{h=1}^{H} w_{1h} \pi_{ih} x_h \\ &= \sum_{h=1}^{H} x_h \end{aligned}$$

It is therefore evident that the estimator given in equation 7 is unbiased.

A.2 The household sharing estimator

Define the share s_{ih} for individual i in household h such that

1. $s_{ih} \in [0, 1]$ for all individuals i in household h
2. $\sum_{i=1}^{k} s_{ih} = 1$

The "household sharing" estimator of T_{pop} can be defined as:

$$\hat{T}_{pop}^{s} = \sum_{a_{ih}=1} w_{ih} s_{ih} x_{ih} \tag{8}$$

Again we write this as a sum over the population:

$$\begin{aligned}\hat{T}^s_{pop} &= \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}a_{ih}s_{ih}x_{ih} \\ &= \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}a_{ih}s_{ih}x_h \end{aligned} \tag{9}$$

where we have used the fact that $x_{ih} = x_h$ for all i.

Taking expectations

$$\begin{aligned}\mathbb{E}\left(\hat{T}^s_{pop}\right) &= \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}\mathbb{E}\left(a_{ih}\right)s_{ih}x_h \\ &= \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}\pi_{ih}s_{ih}x_h \\ &= \sum_{h=1}^{H}\sum_{i=1}^{k} s_{ih}x_h \\ &= \sum_{h=1}^{H} x_h \end{aligned}$$

Note that this derivation presumes that s_{ih} is non-stochastic, i.e. fixed before sampling.

Observe also that the "household representative" estimator $\hat{T}^{rep}_{pop}$ is just a special case of the "household sharing" one, with $s_{1h} = 1$ and $s_{jh} = 0$ if $j \neq 1$.

A.3 Variance with different sharing rules

Since **any** sharing rule among individuals on the frame belonging to the same household will yield an unbiased estimator, it makes sense to pick a sharing rule that will reduce the variance.

Using the "population" version of the equation for $\hat{T}^s_{pop}$ as given in equation 8, it is clear that its variance will depend *inter alia* on the variances and covariances of the a_{ih} selection terms. In the case of NIDS-CRAM individuals were sampled from strata defined by individual level characteristics (in particular age, gender and "race") and sampling was done by simple random sampling without replacement. The latter implies that individuals from different strata (even if they lived in the same household) would have been sampled essentially independently of each other. Within the same stratum the selection process is obviously not independent, since sampling happened without replacement. Nevertheless to the extent that the

strata on the frame are big relative to sample, the probability of individuals ih and jm being jointly included, i.e. $\mathbb{E}(a_{ih}a_{jm})$ will be approximately the product of the individual inclusion probabilities, i.e. $cov(a_{ih}, a_{jm}) \approx 0$.[5]

Consequently

$$\begin{aligned}\mathbb{V}\left(\hat{T}^s_{pop}\right) &\approx \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}^2 \mathbb{V}(a_{ih}) s_{ih}^2 x_h^2 \\ &= \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}^2 \pi_{ih}(1-\pi_{ih}) s_{ih}^2 x_h^2 \\ &= \sum_{h=1}^{H}\sum_{i=1}^{k} (w_{ih}-1) s_{ih}^2 x_h^2\end{aligned}$$

Without loss of generality, we will consider how to minimise this by picking a sharing rule within household h, i.e. we want to choose s_{1h}, s_{2h}, s_{kh} so as to minimise $\mathbb{V}\left(\hat{T}^s_{pop}\right)$ subject to the constraint that $\sum s_{ih} = 1$.

The Lagrangian for this problem is

$$\mathfrak{L} = \sum_{i=1}^{k} (w_{ih}-1) s_{ih}^2 x_h^2 - \lambda\left(\sum s_{ih} - 1\right)$$

With first order conditions

$$\frac{\partial \mathfrak{L}}{\partial s_{ih}} = 2 s_{ih}(w_{ih}-1) x_h^2 - \lambda = 0 \tag{10a}$$

$$\sum s_{ih} = 1 \tag{10b}$$

From the first set of conditions (equations 10a) we get

$$\begin{aligned} s_{ih}(w_{ih}-1) &= c \\ s_{ih} &= \frac{c}{w_{ih}-1} \\ &= c\frac{\pi_{ih}}{1-\pi_{ih}}\end{aligned}$$

[5]Unfortunately, the information in Table 1 refutes this claim for most strata. Accounting for correlation within strata really complicates the analysis. A simple across-the-board rule like the one we derive below will not work. Of course to the extent to which households are composed of individuals from different strata, the analysis still works. In section A.4.1 we adapt the procedure to account for nonresponse. The realised sample from each stratum is, indeed, small relative to the frame, so if nonresponse is independent within households, this approximation is much more reasonable when considering the final sample.

where c is a constant. The (approximately) variance minimising rule is therefore to pick the share s_{ih} inversely proportional to $w_{ih} - 1$ or, equivalently, proportional to the odds of selection, which is given as equation 1 in the main text, i.e.

$$s_{ih} \propto \frac{\pi_{ih}}{1 - \pi_{ih}}$$

Intuitively, the closer p_{ih} is to one, the more we gain by putting all our attention on x_{ih} – that observation will appear in almost every sample and it will have a low weight, so the between-sample variability will be reduced.

A.4 NIDS-CRAM sampling

One immediate complication is that NIDS-CRAM sampling did not fix the probability of selection *ex ante*. This means that there are no fixed "design weights" for the procedure outlined in the previous sections.

A.4.1 Sampling with top-up

Instead, the NIDS-CRAM procedure was to divide the individual frame up into separate "strata" (types of individuals) and to make repeated attempts to augment the original sample in order to achieve a targeted sample size within any particular stratum. In order to think about the implication of this for a weighting strategy, it is useful to trace through what happens in a very simple example, outlined in Figure 5. Here we have assumed a population (stratum) of size three (consisting of unit A,B and C) and a targeted sample of size two, subject to supplemental sampling if there is nonresponse at the first stage. We assume that the probability of nonresponse is a constant ϕ in this population (stratum) and that sampling is by simple random sampling. There are three possible samples at the outset, but twelve possible outcomes of realised samples once fieldwork has happened (non-responding units are shown by a blank box).

In most cases, except those where both sampled units respond, the sample is topped up. Since there is only one way of topping up this doesn't change the probabilities of the sample at that stage. The final realised sample, of course also depends on the nature of non-response on the top-up. Once fieldwork has been completed, there are altogether twenty-one possible outcomes, although there are only seven possible samples: $\{A, B\}, \{A, C\}, \{B, C\}, \{A\}, \{B\}, \{C\}, \varnothing$

Looking across all possible samples, we see that each of them (e.g. the sample $\{A, B\}$) can be reached in three ways. The inclusion possibility for any particular element will be the sum of the probabilities of all the samples in which it can appear. Calculating conventional design weights is therefore possible (conditioning

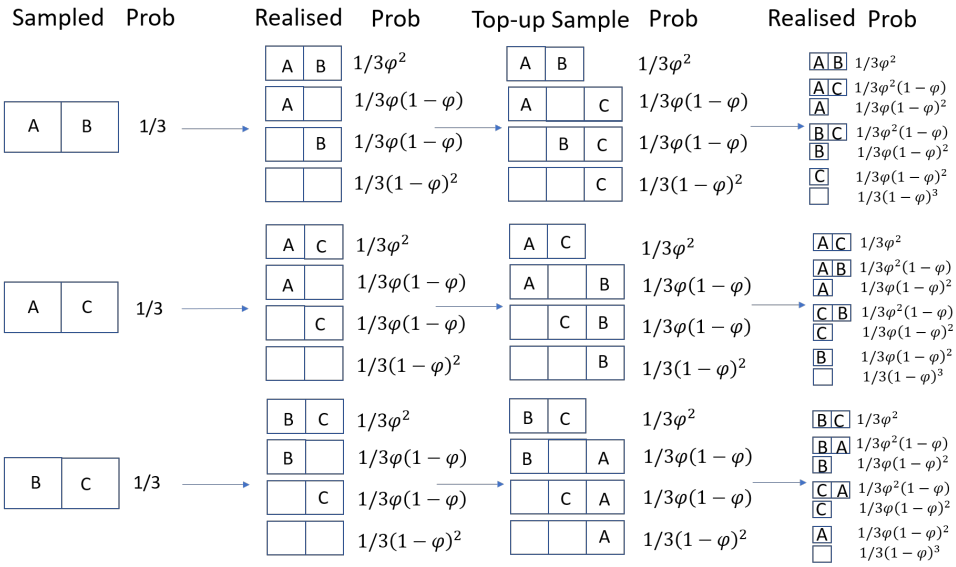

Figure 5: A schema for sampling with refreshment and nonresponse

on getting a sample other than the empty set), but this is not actually the most sensible way to progress.

The most sensible way of analysing the data is to condition on the sample size that was actually achieved. It then remains to observe that the situation is perfectly symmetrical, so there are exactly as many ways of getting the sample $\{A, B\}$ as there are of getting $\{A, C\}$ or $\{B, C\}$. This means that $\Pr(\{A, B\} | n = 2) = \Pr(\{A, C\} | n = 2) = \Pr(\{B, C\} | n = 2) = \frac{1}{3}$. This means that the inclusion probabilities of every unit, conditioning on the realised sample size, will all be equal, i.e. $\Pr(A|n = 2) = \frac{2}{3}$, while $\Pr(A|n = 1) = \frac{1}{3}$.

The weights correcting for design **and** non-response and conditioning on realised sample size, are therefore just $\frac{N}{n}$. It is easy to verify that these weights will give unbiased estimates of totals estimated from the population (using the usual Horvitz-Thompson arguments).

Note that because we know which stratum each of the individuals on the sampling frame belong to, we can calculate *ex post* what these weights should be for every member of the frame, i.e. for individual i in household h of type t

$$w^{nr}_{ih(t)} = \frac{N_t}{n_t} \tag{11}$$

where N_t is the size of the stratum in the population and n_t the size of the realised sample.

Instead of the design weights we can use these design-and-non-response ad-

justed weights $w^{nr}_{ih(t)}$ to calculate the household sharing rules (e.g. equation 1).[6] Denote these as s^*_{ih}, we get equation 11 in the main text, i.e.

$$s^*_{ih(t)} \propto \frac{n_t}{N_t - n_t}$$

To get an optimal unbiased estimator of any population statistic, we however require additional conditions. Firstly, we need to assume that the probability of responding is independent of the household outcome measure x_h. This will be the case if the response probability is constant within individual strata (as was assumed in the argument above). Secondly, we need to assume that the probability of response is independent within households. If it is not, then we cannot derive an optimal sharing rule without knowing more about the response mechanism. Both of these assumptions are questionable given the empirical evidence that we produce.

A.4.2 Non-constant non-response within strata

In practice it transpires that assuming a constant non-response probability within strata does not provide plausible estimates of population statistics from the realised sample. Consequently, the NIDS-CRAM team made an additional adjustment for individually differing probabilities of response (Kerr et al. 2020).

To understand the logic, we can rewrite equation 11 as

$$w^{nr}_{ih(t)} = \frac{N_t}{n^s_t} \frac{n^s_t}{n_t}$$

where n^s_t is the number of individuals from stratum t that were sampled, and N_t and n_t are the stratum size and size of realised sample, as before. The quantity $\frac{n^s_t}{N_t}$ is the sampling fraction of stratum t.

Using the full sample (i.e. including the non-responders) the NIDS-CRAM sampling team estimated a probit model for response using "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" (Kerr et al. 2020, p.2). We can write the estimated probability as

$$\hat{p}_{ih(t)} = F\left(\mathbf{z}'_{ih}\hat{\beta} + \mathbf{z}'_{h}\hat{\gamma} + \sum_{stratum=j}^{T} \alpha_j \, D_{ih(j)}\right)$$

[6] Implicit in this procedure is the assumption that the nonresponse process is random within households, conditioning on the strata of its members. Given that sampled individuals were telephoned directly this is not a wild assumption.

where F is the cumulative normal distribution, $\mathbf{z}'_{ih}$ is a vector of individual characteristics, $\mathbf{z}'_h$ a vector of household characteristics and $D_{ih(j)}$ the set of stratum dummy variables.

The NIDS-CRAM non-response adjusted weights then become

$$w^{NCs}_{ih(t)} = \frac{N_t}{n^s_t} \frac{1}{\hat{p}_{ih(t)}} \tag{12}$$

Note that this reduces to the case in equation 11 if the vectors $\hat{\beta}$ and $\hat{\gamma}$ are both zero, since then $\hat{p}_{ih(t)}$ is just $\frac{n_t}{n^s_t}$. Empirically this is decisively not the case (personal communication, Cally Ardington).

One could interpret $n^s_t \, \hat{p}_{ih(t)}$ as the expected number of individuals of type t in the realised sample, assuming that the response mechanism is described by the probit model.[7]

We cannot implement the optimal sharing rule (equation 1) using the NIDS-CRAM released weights, since we do not have these for individuals on the frame other than those that ended up in the realised sample. Furthermore the weights went through an additional process of trimming.

But, as noted in section A.2, any sharing rule s_{ih} produces a consistent population estimate, provided that the weight w_{ih} corrects properly for the sampling design and (in this case) non-response.

In our empirical work we therefore combine the NIDS-CRAM weights w^{NCs}_{ih} with shares given by equation 2. The estimator for a population household quantity can therefore be written as

$$\hat{T}^s_{pop} = \sum_{a_{ih}=1} w^{NCs}_{ih} s^*_{ih} x_{ih}$$

Using the sharing rule s^*_{ih} rather than some other plausible ones (e.g. equal shares) will be better to the extent to which the "stratum response probability" $\frac{n_t}{n^s_t}$ is at least a partial indicator of the individual response probability $\hat{p}_{ih}$. Let the ratio of the NIDS-CRAM weights and the weights defined in equation 11 be $r^{adj}_{ih(t)}$, i.e.

$$\begin{aligned} r^{adj}_{ih(t)} &= \frac{w^{NCs}_{ih}}{w^{nr}_{ih(t)}} \\ &= \frac{n_t}{\hat{p}_{ih(t)} \, n^s_t} \end{aligned}$$

[7] In the example of Figure 5 this procedure gives the right estimates – although the probit does not actually estimate the response propensity consistently. The underlying problem is that the size of the sample within a stratum, i.e. n^s_t, is itself a function of nonresponse and so depends on the response probability. Estimating the probability on an endogenous sample creates problems. A preferable procedure would be to estimate the probability of being sampled and responding over the entire stratum. In practice this is unlikely to make much difference.

Then we can rewrite the estimator of the population total as

$$\hat{T}^{s}_{pop} = \sum_{a_{ih}=1} r^{adj}_{ih(t)} w^{nr}_{ih(t)} s^{*}_{ih} x_{ih}$$

If r^{adj}_{ih} is equal to one, then this is equivalent to doing the calculation only with the stratum weights, and the sharing rule will then be variance minimising. The further r^{adj}_{ih} is from one, the further the shares s^{*}_{ih} are likely to be from optimality. In our empirical work we provide estimates with different sharing rules. The estimated standard errors are an indicator of the the relative efficiencies of differing sharing rules. It turns out that the "optimal shares" s^{*}_{ih} seem to perform best.

A.4.3 Adjusting to the 2017 population

In practice there is an additional adjustment. The NIDS-CRAM weights are not designed to provide estimates for the NIDS wave 5 sample, but for the population that is covered by that sample (roughly the adult South African population of 2017). The process of generating estimates for the 2017 population is straightforward.

By assumption an unbiased estimator of a population total $\mathbf{X}^{2017}_{h}$ using the NIDS household measures x_h is given by

$$\begin{aligned}\hat{\mathbf{X}}^{2017}_{h} &= \sum_{h=1}^{n_{nids}} w5_wgt_h x_h \\ &= \sum_{h=1}^{n_{nids}} z_h\end{aligned}$$

where $w5_wgt_h$ is the NIDS wave 5 household weight for household h – which also happens to be the person weight – and we let $z_h = w5_wgt_h x_h$.

By the argument of the previous section

$$\begin{aligned}\hat{X}^{2017}_{pop} &= \sum_{a_{ih}=1} w^{NCs}_{ih} s^{*}_{ih} z_{ih} \\ &= \sum_{a_{ih}=1} w^{NCs}_{ih} s^{*}_{ih} w5_wgt_{ih} x_{ih}\end{aligned}$$

will give an unbiased estimate of $\sum_{h=1}^{n_{nids}} z_h$ which, of course, is our unbiased estimate of $\hat{\mathbf{X}}^{2017}_{h}$. It is evident that the weights required to provide estimates for the 2017 population of **households** are

$$w^{NC\,hh}_{ih} = w^{NCs}_{ih} s^{*}_{ih}\, w5_wgt_{ih} \tag{13}$$

Note that the **person weights** released with NIDS-CRAM already incorporate the weighting adjustment to produce estimates for the 2017 population, i.e. the NIDS-CRAM weights are

$$w_{ih}^{NCp} = w_{ih}^{NCs}\, w5_wgt_{ih}$$

So the **household weights** given by equation 13 can also be written in the form given in equation 4 in the main text.

A.5 Measurement error

Up to this point we have assumed that every individual within household h provides the same information, viz. x_h. Let us now consider the case where individuals report the household measure with error, i.e.

$$x_{ih} = x_h + \eta_{ih} \tag{14}$$

The "household share" estimator defined in equation 8 can now be written as

$$\hat{T}_{pop}^{s} = \sum_{a_{ih}=1} w_{ih} s_{ih} \left(x_h + \eta_{ih}\right)$$

It is evident that even in expectation this is not guaranteed to give us the true population total. Rewriting the estimator again as a sum over the population (as in equation 9) but focusing only on the contribution of individuals in household h we see that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{k} w_{ih} a_{ih} s_{ih} \left(x_h + \eta_{ih}\right)\right] &= \sum_{i=1}^{k} w_{ih} \mathbb{E}\left(a_{ih}\right) s_{ih} \left(x_h + \eta_{ih}\right) \\
&= \sum_{i=1}^{k} w_{ih} \pi_{ih} s_{ih} \left(x_h + \eta_{ih}\right) \\
&= x_h + \sum_{i=1}^{k} s_{ih} \eta_{ih}
\end{aligned} \tag{15}$$

This says that, in expectation, the contribution from individuals in household h would give us the true value plus a weighted average of the measurement errors, with weights s_{ih}. Note that in the Horvitz-Thompson framework we treat x_{ih} as fixed, so that we have treated the measurement error η_{ih} likewise as fixed when we took expectations.

It is worth noting that equation 15 implies that different choices of the "sharing rule" change the **population parameter** that is estimated. For household h, the

outcomes will range between $\min\{x_{1h}, x_{2h}, \ldots, x_{kh}\}$ and $\max\{x_{1h}, x_{2h}, \ldots, x_{kh}\}$. Likewise, the estimate $\hat{T}^s_{pop}$ can range between the case where the sum is constructed from the set of minima in each household and the sum of the maxima. Note that this range is not due to sampling noise, but due to measurement error.

Without further knowledge of the error process it is hard to say more about how the "sharing rule" s_{ih} will impact. It seems clear, however, that it is unlikely that picking only one representative per household will deal well with the errors.

To get a little traction, let us assume that the errors η_{ih} are drawn independently of each other from the same distribution with zero mean and variance σ^2_η; and that they are also independent of the sampling process a_{ih}. It then follows that $cov\left(x_{ih}, x_{jm}\right) = 0$ if $i \neq j$ and that $cov\left(x_{ih}, a_{jm}\right) = 0$.

Furthermore we can show that if the random variables X and Y are independent, then

$$\mathbb{V}\left(XY\right) = \mathbb{V}\left(X\right)\mathbb{V}\left(Y\right) + \mathbb{V}\left(X\right)\mu_Y^2 + \mathbb{V}\left(Y\right)\mu_X^2$$

where $\mu_X = \mathbb{E}\left(X\right)$ and $\mu_Y = \mathbb{E}\left(Y\right)$.

Applying this result to the formula for $\hat{T}^s_{pop}$ as given in equation 9, and assuming (as before) that the sampling design for NIDS-CRAM is such that a_{ih} and a_{jm} are approximately independent, then

$$\begin{aligned}
\mathbb{V}\left(\hat{T}^s_{pop}\right) &\approx \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}^2 \mathbb{V}\left(a_{ih}x_{ih}\right) s_{ih}^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}^2 \{\mathbb{V}\left(a_{ih}\right)\mathbb{V}\left(x_{ih}\right) + \mathbb{V}\left(a_{ih}\right)x_h^2 + \mathbb{E}\left(a_{ih}\right)^2\mathbb{V}\left(x_{ih}\right)\} s_{ih}^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{k} w_{ih}^2 \{\pi_{ih}\left(1-\pi_{ih}\right)\sigma_\eta^2 + \pi_{ih}\left(1-\pi_{ih}\right)x_h^2 + \pi_{ih}^2\sigma_\eta^2\} s_{ih}^2 \\
&= \sum_{h=1}^{H}\sum_{i=1}^{k} \{\left(w_{ih}-1\right)\left(\sigma_\eta^2 + x_h^2\right) + \sigma_\eta^2\} s_{ih}^2
\end{aligned}$$

If we again focus on choosing s_{ih} to minimise the variance we get the Lagrangian

$$\mathfrak{L} = \sum_{i=1}^{k} \{\left(w_{ih}-1\right)\left(\sigma_\eta^2 + x_h^2\right) + \sigma_\eta^2\} s_{ih}^2 - \lambda\left(\sum s_{ih} - 1\right)$$

With first order conditions

$$\frac{\partial \mathfrak{L}}{\partial s_{ih}} = 2s_{ih}\left(w_{ih}\left[x_h^2 + \sigma_\eta^2\right] - x_h^2\right) - \lambda = 0 \tag{16a}$$

$$\sum s_{ih} = 1 \tag{16b}$$

Substituting in $w_{ih} = \frac{1}{\pi_{ih}}$ into equation 16a we get

$$s_{ih} = \frac{\pi_{ih}}{1 - \pi_{ih} + \pi_{ih} R_h} \frac{\lambda}{x_h^2 + \sigma_\eta^2}$$

where $R_h = \frac{\sigma_\eta^2}{x_h^2 + \sigma_\eta^2}$. Note that R_h is common within household h although different across different households. Within household h we would therefore like to allocate shares according to the rule

$$s_{ih} \propto \frac{\pi_{ih}}{1 - \pi_{ih} + \pi_{ih} R_h}$$

This is however not a formula that can be implemented since it depends on the unknowns σ_η^2 and x_h.[8] But observe that if $\pi_{jh} \leq \pi_{ih}$ then

$$\frac{\pi_{ih}}{\pi_{jh}} \leq \frac{\pi_{ih} / (1 - \pi_{ih} + \pi_{ih} R_h)}{\pi_{jh} / (1 - \pi_{jh} + \pi_{jh} R_h)} = \frac{s_{ih}}{s_{jh}} \leq \frac{\pi_{ih} / (1 - \pi_{ih})}{\pi_{jh} / (1 - \pi_{jh})}$$

So in this case the optimal sharing rule is bounded between the rule expressed by equation 1 and a rule which allocates shares proportionally to the probability of selection, i.e.

$$s_{ih} \propto \pi_{ih}$$

which is equation 5 in the main text. Intuitively, the formula given by equation 1 pays far too much attention to observations with π_{ih} close to one, and this "blows up" any measurement error associated with those cases. The rule given in equation 5 is not so sensitive.

Given the presence of non-response we will, in our applied work consider the rule which fixes shares in proportion to the probability of being sampled *and* responding, estimated, for instance (see equation 11) as

$$s_{ih}^{**} \propto \frac{n_t}{N_t} \tag{17}$$

A.6 Changes in households

In this section we analyse how changes in the household distribution will affect our estimation strategy. To crystallise the issues, let us adapt the notation that we have used thus far. We assume that the individual identifier i is unique to individuals across waves (akin to the NIDS pid identifier) and that individual i is

[8] One could estimate σ_η^2 as the "within" household variance from a fixed effects estimator and, in big households, estimate x_h as the mean of x_{ih}. This is unlikely to be accurate or worth the effort.

in household h at the time of NIDS wave 5, but in household h' at the time of NIDS-CRAM. The generic estimator of a household total is

$$\begin{aligned}\hat{T}_{pop}^{s\,2020} &= \sum_{a_i=1} w_{ih'}^{NC2020} s_{ih'} x_{ih'} \\ &= \sum_{a_i=1} w_{ih'}^{NCs}\, pw_{ih'}\, s_{ih'}\, x_{ih'} \end{aligned} \tag{18}$$

In the last line we have split the household weight $w_{ih'}^{NC2020}$ into a weight designed to get the NIDS-CRAM sample back to the NIDS frame from which it was drawn ($w_{ih'}^{NCs}$) and the correction to get from that frame to the 2020 population ($pw_{ih'}$), as in equation 4.

To implement this, we need to know:

- which household h' the individual resides in, and
- how to raise the NIDS wave 5 sample to the 2020 population, i.e. which household weights $pw_{ih'}$ to apply.

We discuss these issues in turn.

A.6.1 Households breaking apart

The break-up of existing households will potentially reallocate individuals to households. The first implication of this is that the shares $s_{ih'}$ now need to be calculated over individuals on the frame that are in the same 2020 household h'. If, for example, there were five individuals on the frame from household h in NIDS wave 5 with measures $x_{1h}, x_{2h}, x_{3h}, x_{4h}$ and x_{5h} and this household splits into two households $h' = \{1, 2\}$ and $h'' = \{3, 4, 5\}$, then an equal sharing rule would produce shares $s_{1h'} = s_{2h'} = \frac{1}{2}$ and $s_{3h''} = s_{4h''} = s_{5h''} = \frac{1}{3}$ in formula 18, i.e. $\frac{1}{2}x_{1h'} + \frac{1}{2}x_{2h'}$ would be the expected contribution of new NIDS-CRAM household h' to the 2021 population total and $\frac{1}{3}x_{3h''} + \frac{1}{3}x_{4h''} + \frac{1}{3}x_{5h''}$ would be that of new household h''.

There are two key questions:

- How do wave 5 households break apart?
- Do any wave 5 households "join up"?

The probability of the latter event is so rare, that we exclude it from consideration. To answer the first question we need to work out whether any of the individuals seem to have left their original household. Some of the variables in NIDS-CRAM (about whether they are still in their original dwelling) could be used here.

A.6.2 Are there any new co-residents?

Besides working out how NIDS wave 5 individuals may have changed their residential arrangements, we also need to consider whether any new individuals have joined. New births to CSM mothers will not affect our calculations in any material way, so we consider only the impact of "temporary sample members" (TSMs).

The NIDS rules for updating weights between waves, is that the weights of "continuing sample members" (CSMs) within a household are added up and then "shared out" between all members of that new household. If, for instance, there are k_c CSMs in the NIDS-CRAM household (all from household h in wave 5) and now co-resident with k_t TSMs, then the weight for the 2020 household h' (and all the individuals within them) will be given by:

$$pw_{ih'} = \frac{k_c}{k_c + k_t} w5_wgt_{ih} \tag{19}$$

The term $w5_wgt_{ih}$ is the original, unadjusted wave 5 weight for individual i in household h, i.e. the weight designed to give estimates for the 2017 population of individuals or households. Note that the counts k_c and k_t include children CSMs and TSMs.[9]

Because of the multiple ways in which additional household members could arrive, the most straigthtforward way to get a point estimate would be to substitute the **expected** contribution to the total from individual i, i.e.

$$\mathbb{E}\left(pw_{ih'} s_{ih'} x_{ih'}\right) = \mathbb{E}\left(\frac{k_c}{k_c + k_t}\right) w5_wgt_{ih}\, s_{ih'}\, x_{ih'} \tag{20}$$

We can get estimates about the transition probabilities facing individual i in household type T in terms of the new co-residency relationships. Using the history of previous transitions of course assumes that the changes between 2017 and 2021 have been "normal." It is probable that they were not. In principle one could simulate alternative outcomes and check their impacts on the types of household statistics that are calculated.

References

Horvitz, D. and Thompson, D.: 1952, A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association* **47**(260), 663–685.

[9] Otherwise it would be tempting to use an "equal allocation" share rule for $s_{ih'} = \frac{1}{k_c}$ so that the product term $pw_{ih'} s_{ih'}$ simplifies to $\frac{1}{size} w5_wgt_{ih}$.

Ingle, K., Brophy, T. and Daniels, R. C.: 2020, National Income Dynamics Study – Coronavirus Rapid Mobile Survey (NIDS-CRAM) Panel User Manual, *Release July 2020. Version 2*, Southern Africa Labour and Development Research Unit, Cape Town.

Kerr, A., Ardington, C. and Burger, R.: 2020, Sample Design and Weighting in the NIDS-CRAM Survey, *Wave 1 Technical Paper B*, NIDS-CRAM.

NIDS-CRAM: 2020a, National Income Dynamics Study - Coronavirus Rapid Mobile Survey (NIDS-CRAM), Wave 1, *[dataset]. version 3.0.0*, Allan Gray Orbis Foundation [funding agency], Southern Africa Labour and Development Research Unit [implementer], DataFirst [distributor], Cape Town. DOI: https://doi.org/10.25828/7tn9-1998.

NIDS-CRAM: 2020b, National Income Dynamics Study - Coronavirus Rapid Mobile Survey (NIDS-CRAM), Wave 2, *[dataset]. version 3.0.0*, Allan Gray Orbis Foundation [funding agency], Southern Africa Labour and Development Research Unit [implementer], DataFirst [distributor], Cape Town. DOI: https://doi.org/10.25828/5z2w-7678.

NIDS-CRAM: 2021a, National Income Dynamics Study - Coronavirus Rapid Mobile Survey (NIDS-CRAM), Wave 3, 2020, *[dataset]. version 3.0.0*, Allan Gray Orbis Foundation [funding agency], Southern Africa Labour and Development Research Unit [implementer], DataFirst [distributor], Cape Town. DOI: https://doi.org/10.25828/s82x-nx07.

NIDS-CRAM: 2021b, National Income Dynamics Study - Coronavirus Rapid Mobile Survey (NIDS-CRAM), Wave 4, 2021, *[dataset]. version 2.0.0*, Allan Gray Orbis Foundation [funding agency], Southern Africa Labour and Development Research Unit [implementer], DataFirst [distributor], Cape Town. DOI: https://doi.org/10.25828/y5qj-x095.

NIDS-CRAM: 2021c, National Income Dynamics Study - Coronavirus Rapid Mobile Survey (NIDS-CRAM), Wave 5, 2021, *[dataset]. version beta1*, Allan Gray Orbis Foundation [funding agency], Southern Africa Labour and Development Research Unit [implementer], DataFirst [distributor], Cape Town.

van der Berg, S., Patel, L. and Bridgman, G.: 2021, Hunger in South Africa during 2020: Results from Wave 3 of NIDS-CRAM, *wave 3 working paper 10*, NIDS-CRAM.

Founded in 1975, the Southern Africa Labour and Development Research Unit (SALDRU) is a research-based social responsiveness initiative housed in the School of Economics at the University of Cape Town.

The unit carries out research and capacity building in applied empirical microeconomics with an emphasis on poverty and inequality, labour markets, human capital and social policy. We strive for academic excellence and policy relevance.

SALDRU has implemented a range of innovative surveys in South Africa including the Project for Statistics on Living Standards and Development (PSLSD), Cape Area Panel Study (CAPS) and the National Income Dynamics Study (NIDS), among others. Building on these large data gathering projects, we conduct a range of training and capacity building activities in the use of survey data to analyse social well-being.

Our mission is to challenge inequalities through policy relevant academic research.

Level 3, School of Economics Building,
Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701,
Cape Town, South Africa
Tel: +27 (0)21 650 5696
Fax: +27 (0) 21 650 5797
Web: https://www.saldru.uct.ac.za
Repository: http://opensaldru.uct.ac.za