HelpAge International: 2011 India NSS Analysis

Stewart Kerr & Rick Griffin

2021-02-18

1. Summary

2. Introduction

HelpAge International wants to challenge established norms for statistical reporting on older persons by proving that data disaggregation to a lower, more granular level is possible and statistically robust. Nongranular statistics reinforces an oversimplified picture of inequalities and the inadequate data itself becomes a barrier to the inclusion of at-risk and marginalized groups in policy and program responses.

To serve this goal, we analyzed data from the employment and unemployment surveys included in the 2011 India National Sample Survey (NSS) to determine the lowest level of disaggregation that was possible while maintaining statistical robust estimates of average wage. The employment and unemployment surveys of the NSS aim to get estimates of various employment characteristics at the national and state level. In addition to employment related variables, individual characteristics such as region, age, sex, industry, education, and others are collected by the survey. In accordance with HelpAge's statement of work, we focused how earnings of employees varied by an individual's age, sex, employment industry, and region (urban/rural). Disability status was only collected in relationship to employment (i.e. unable to work due to disability) and was not available to analyze in relation to average earnings.

In our analysis, we sought to answer three specific research questions related to data disaggregation:

1. What is the most granular level of disaggregation of age, sex, and employment industry? What are the most appropriate age bands (i.e. 5 year groupings or 10 year groupings) and upper age cohort (i.e. 80+, 85+, etc)? How do results differ going from broader to more granular disaggregation?
2. What is the most granular level of disaggregation of age, sex, and employment industry when we also include geographic location (urban/rural)?

3. Based on these results, what general recommendations or considerations can be made on data disaggregation for similar surveys?

3. Materials and Methods

Data Collection

The 2011 India NSS used a stratified multi-stage design. First villages in rural communities were selected by probability proportional to size with replacement while blocks in urban areas were selected by simple random sampling without replacement. An equal number of villages and blocks were selected. Next, if the village or block contained more than 1200 people, it was divided into subgroups containing roughly the same amount of people. Then, households within each subgroup were stratified into 3 groups according to to measures of wealth. Lastly, households from each strata were selected by simple random sampling without replacement and all individuals within the household were surveyed. Samplings weights were calculated and provided for each individual by the India Ministry of Statistics & Programme Implementation.

We extracted the raw survey data in .sav (SPSS) format from the file provided by HelpAge using the required Nesstar Explorer software. For our analysis, we needed to extract the data files `Block_4_Demographic particulars of household members` and `Block 5_3_Time disposition during the week ended on`. This data was then loaded into R using the `haven` package and processed using the `tidyverse` set of R packages.

Data Processing

First, in order to join demographic data in the "Block 4" (B4) dataset to employment data for the past week in the "Block 5_3" (B53) dataset, we had to create a unique ID. This was accomplished by concatenating the following variables for each dataset: * `FSU_Serial_No`, `Stratum`, `Sub_Stratum_No`, `Hamlet_Group_Sub_Block_No`, `Second_Stage_Stratum_No`, `Sample_Hhld_No`, and `Person_Serial_No`

Then, before joining B4 and B53, the following variables were processed or created from the B53 dataset: * `employment_status` - Takes either "employed", "unemployed" or "not in labor force" depending on the value of `current_weekly_activity_status` * `weekly_earnings` - An individual can have multiple entries in the B53 dataset if they performed multiple jobs during the week. Thus, for each person, we get their total weekly earnings by summing their earnings across the last 7 days. * `industry` - There are many industries reported in the `current_weekly_activity_NIC_2008` variable. We collapsed the industries into 4 groups based on sample size considerations: "farming, forestry, or fishing", "manufacturing", "construction", or "other".

After creating these variables, we joined the B4 and B53 datasets as our "analysis dataset." As we are primarily interested in the average earnings of different groups, we focused only on employed individuals. However, there are many employed people in the dataset that did not report any earnings in the previous 7 days. Nevertheless, **we chose to keep those individuals in our analysis dataset.** Table 1 presents the counts of the people included in our analysis dataset.

Table 1. Counts of people in various groupings of the 2011 India NSS survey.

	Employed		Unemployed		Not in labor force		
	Male	Female	Male	Female	Male	Female	Overall
Under 60	113848	37748	4828	2452	96191	163904	**418971**
60-64	4898	1468	46	21	2368	5782	**14583**
65-69	2825	715	19	11	2291	4433	**10294**
70-74	1298	258	10	0	2050	2905	**6521**
75-79	448	65	2	0	1146	1614	**3275**
80+	248	33	4	0	1284	1786	**3355**
Farming, forestry, or fishing	36486	20314	0	0	0	0	**56800**
Manufacturing	14880	5562	0	0	0	0	**20442**
Construction	16081	2301	0	0	0	0	**18382**
Other	56118	12110	0	0	0	0	**68228**
Urban	47723	11969	2079	1146	40926	72393	**176236**
Rural	75842	28318	2830	1338	64404	108031	**280763**
Total	123565	40287	4909	2484	105330	180424	**456999**

Statistical Analysis and Methodology

For each grouping of employed people, we are interested in the average weekly earnings. The average earnings were calculated using the survey weights, which indicate how many individuals in India's population that a sampled person represents. That is, for each group we performed the calculation:

$$\frac{\sum_i w_i Y_i}{\sum_i w_i}$$

where w_i is the weight and Y_i is the weekly earnings for individual i within the grouping. Then, the standard error of these weighted estimates was calculated using a simple bootstrap method[1]. A bootstrap method was chosen because the survey design is highly complex and calculating standard errors while accounting for this design would be difficult and complicated. The cost of using bootstrap to estimate standard errors instead of incorporating the survey design causes our standard errors to be slightly larger, however, we compared our bootstrap standard errors with a simple random sampling variance multiplied by a cluster sampling design effect to verify that the bootstrap variances magnitudes were reasonable.

For each group identified below, 1000 independent bootstrap replicates each of the size of the group were

selected by simple random sampling with replacement. Thus, each bootstrap replicate has the same sample size as the group but some sample persons might be selected more than once and others not at all in each replicate. For each replicate, the weighted average earnings were calculated as described above. Then, the sample variance over these 1000 weighted estimates was calculated and the square root of the variance is the standard error. For each grouping, 95% confidence intervals for the estimate of weekly earnings were calculated using this formula:

$$\bar{Y} \pm 1.96 \times SE(\bar{Y})$$

where $\bar{Y}$ is the weighted estimate from the original sample and $SE(\bar{Y})$ is the standard error of the estimate calculated using the bootstrap procedure.

To produce accurate estimates that are statistically robust, we suggest including at least 150 sampled persons in the group. With this consideration, we first considered the most granular level of disaggregation: stratification by 5 year age groups sex, industry, and geographic location (urban/rural) across the board. In most cases, our sample size was not large enough to include the secondary level of disaggregation of urban/rural. Specifically, we were only able to include urban/rural for employed person's under the age of 60 and men aged 60-64. Additionally, as the count of employed people decreased at older ages, we had to collapse industry categories in order to maintain a sample size that was close to or greater than 150. At our oldest age groups, we had to drop stratification by industry entirely. After collapsing these levels of disaggregation, we arrived at the most granular level of disaggregation that we can recommend as being statistically robust for this particular survey. These results are shown in table 2.

Next, we used the same stratification categories with 10 year age groupings instead of 5 year age groupings to assess the difference of results using different age groupings. The results of this comparison are shown in table 3.

Lastly, we dropped the secondary level of disaggregation, geographic location, entirely to examine how different the results are whenever less disaggregation is performed. As we included urban/rural only for people younger than 60 and men aged 60-64 in our primary analysis, we can only compare results for a small subset of groupings. The results of this comparison are shown in table 4.

Now, we include instructions and an example on the comparison of any two particular groups. For any two groups, if the 95% confidence intervals do not overlap then it is clear that the average earnings in the two groups are significantly different. If the confidence intervals do overlap, then a simple 5% error hypothesis

test can be used to test the null hypothesis that the two groups have the same average wage. The detectable difference between two estimates is 1.96 times the standard error of the difference between the two estimates. The standard error of the difference is calculated as:

$$\sqrt{SE_1^2 + SE_2^2}$$

where SE_1 is the standard error of the estimate for group 1's earnings and SE_2 is the standard error of the estimate for group 2's earnings.

We illustrate this procedure for the comparison of the estimated average earnings of males age 60-64 in the manufacturing industry (group 1) and males age 60-64 in the "other" industry group (group 2). The estimated average wage for group 1 is ~ 567 and the standard error is ~ 92. The estimated wage for group 2 is ~ 493 and the standard error is ~ 65. The square root of 92^2 plus 65^2 is $\sqrt{92^2 + 65^2} = 112.65$, so the detectable difference is $1.96 \times 112.65 = 220.79$. Since the difference is $567 - 493 = 74$, we cannot claim the true values are statistically different.

table1

table2

table3

4. Discussion

Research question 1

For males and females under age 60 as well as males age 60-64, all 4 industry as well as urban/rural categories have enough sample to accurately estimate average wage of employed persons. For males 65-69 it is necessary to eliminate urban/rural but all 4 industry categories can be kept. For males age 70-74 and age 74-79 it is necessary to collapse to two industry categories while for males 80+ there are not enough employed persons to support any industry categories.

For Females age 60-64 and age 65-69 some collapsing of industry categories is necessary while females 70+ and 75+ do not have enough employed for any industry categories,

Research question 2

Urban or rural can only be used for males and females under age 60 and males age 60-64. Other age and sex categories do not have enough employed persons for accurate estimation of average wage.

Research question 3

Data disaggregation for other surveys depend heavily on the sample design and sample allocation. It is likely necessary to have at least 150 sample cases in any age/sex/industry or occupation/ geographic categories to define estimation groups. The survey designers need to keep this in mind. If estimates are desired for 5 year age/sex groups sufficient sample is necessary in those groups.

5. Conclusion

References

1. *Introduction to Bootstrapping in Statistics with an Example - Jim Frost*