

# HelpAge International: 2011 India NSS Analysis

Stewart Kerr & Rick Griffin

2021-03-01

## 1. 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 weekly earnings in rupees. 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, our analysis focuses on sample size differences across varied groupings of age, sex, employment industry, and region (urban/rural) stratifiers. Additionally, we also provide preliminary findings on how average weekly earnings vary across these groupings. 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 weekly 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 does sample size 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 disaggre-

gation for similar surveys?

## 2. 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 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 in rupees 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 final 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.** This was done because we have no way of knowing whether those employed individuals reporting 0 income in the previous week actually made no earnings or if there was a data collection error. Therefore, the variable of interest should be regarded as the average earnings reported in the week prior to being interviewed. 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		Overall
	Male	Female	Male	Female	Male	Female	
Under 60	113848	37748	4828	2452	96191	163904	<b>418971</b>
60-64	4898	1468	46	21	2368	5782	<b>14583</b>
65-69	2825	715	19	11	2291	4433	<b>10294</b>
70-74	1298	258	10	0	2050	2905	<b>6521</b>
75-79	448	65	2	0	1146	1614	<b>3275</b>
80+	248	33	4	0	1284	1786	<b>3355</b>
Farming, forestry, or fishing	36486	20314	0	0	0	0	<b>56800</b>
Manufacturing	14880	5562	0	0	0	0	<b>20442</b>
Construction	16081	2301	0	0	0	0	<b>18382</b>
Other	56118	12110	0	0	0	0	<b>68228</b>
Urban	47723	11969	2079	1146	40926	72393	<b>176236</b>
Rural	75842	28318	2830	1338	64404	108031	<b>280763</b>
<b>Total</b>	<b>123565</b>	<b>40287</b>	<b>4909</b>	<b>2484</b>	<b>105330</b>	<b>180424</b>	<b>456999</b>

## Statistical Analysis and Results

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 (SE) of these weighted estimates was calculated using a simple bootstrap method<sup>1</sup>. 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 is that 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, both 90% and 95% confidence intervals for the estimate of weekly earnings were calculated using this formula:

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

where  $\bar{Y}$  is the weighted estimate from the original sample,  $z$  is the z-score reflecting the size of the confidence interval we want ( $z = 1.645$  for 90% or  $z = 1.96$  for 95%), 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. The results of this analysis and all subsequent analysis are shown in table 2 in the appendix.

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. 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.

Now, we provide instructions and an example on the comparison of any two particular groups. For any two groups, if the 90%/95% confidence intervals do not overlap then it is clear that the average earnings in the two groups are significantly different at the 90%/95% confidence level. 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 1.645/1.96 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  $\sim 567$  and the standard error is  $\sim 92$ . The estimated wage for group 2 is  $\sim 493$  and the standard error is  $\sim 65$ . 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.

Note that if one group is a subset of the other group, (e.g. comparing men aged 60-64 to men aged 60-69), then it is equivalent to compare the smaller group with the part of the larger group that is not contained in the smaller group. In the example above, comparing men aged 60-64 to men aged 65-69 is the same as comparing men aged 60-64 to men aged 60-69; if a statistically significant difference is observed when comparing men aged 60-64 and men aged 65-69 then men aged 60-64 are also significantly different from men aged 60-69. These comparisons of a larger group to its subset get more difficult as you have more levels of disaggregation and the average weekly earnings for all subsets of a larger group may not be reported in the tables below. Therefore, if you desire to compare the average weekly earnings of a group to a subset of that group, proceed with caution.

### 3. Discussion

#### Most Granular Disaggregation and Age

The most granular level of disaggregation that we can recommend as being statistically robust is the following:

**For men:**

- Under 60: All 4 industries and region
- 60-64: All 4 industries and region
- 65-69: All 4 industries, no region
- 70-74: “Farming, forestry, and fishing” and other industries, no region
- 75-79: “Farming, forestry, and fishing” and other industries, no region
- 80+: No industry, no region

**For women:**

- Under 60: All 4 industries and region
- 60-64: “Farming, forestry, and fishing”, “Manufacturing or construction”, and other industries, no region
- 65-69: “Farming, forestry, and fishing” and other industries, no region
- 70-74: No industry, no region
- 75+: No industry, no region

These groupings reflect a guiding principle of having a sample size of at least 150 in order to accurately estimate average earnings. We also examined using 10 year age bands in our analysis. Ultimately, for this survey, while 10 year age bands generally provide adequate sample size for females in the 60-69 year age group to add additional disaggregation by industry, it is insufficient in the 70-79 year age group or older. Thus, for women, 10 year age bands may be useful if disaggregation by industry is an area of primary interest. For men, there are enough employed individuals that we can fully disaggregate by industry using 5 year age bands in the 60-69 age range. At older ages, if disaggregation by industry is of primary interest, 10 year age bands might allow for additional disaggregation by industry. In general, at older ages, there are few employed people in this survey. There are only 696 employed men aged 75 or older and 98 employed women aged 75 or older. With this sample size, it is difficult to recommend any further disaggregation beyond one or two industry categories, age, and sex.

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 each age group/sex combination, we have included disaggregation by various industries as well as no disaggregation by industry (rows in the “all” category in table 2). Increased disaggregation by industry appears to be useful primarily at younger ages where sample size is sufficient to detect differences in average weekly earnings. For example, “farming, forestry, or fishing” has consistently (and statistically significant) lower earnings than other industry categories across many age group/sex combinations. At older age groups, for example men aged 75-79, there is inadequate sample size to make this same determination.

## **Including Region as a Secondary Level of Disaggregation**

We find that region (urban/rural) can only be used as a stratification for males and females under age 60 and males aged 60-64. Other age and sex categories simply do not have enough employed persons to accurately estimate average earnings when we use region as a secondary level of disaggregation.

Nevertheless, for the groups we were able to disaggregate by region, we find a statistically significant difference in average weekly earnings between urban and rural dwellers with urban individuals earning more than their rural counterparts. The differences appear to be smaller or nonexistent for individuals working in the “farming, forestry, or fishing” or construction industries.

## **Generalization of Results to Similar Surveys**

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. This is a consideration that the survey designers should keep in mind from the beginning of the survey process. In estimating income from working, it becomes difficult to perform much disaggregation at older ages (75/80+) where there are fewer working individuals. Out of approximately 457,000 individuals surveyed, ultimately only around 12,000 employed persons older than 60 were surveyed. This was sufficient for disaggregation at younger ages but we were unable to disaggregate much at older ages. In a study of income, if highly granular disaggregation at these older ages is desired then it may be beneficial to explore methods for increasing sample size in those older age groups.

## **4. Conclusion**

In this report, we examined data from the employment and unemployment surveys included in the 2011 India National Sample Survey (NSS) to examine disaggregation in the context of computing statistically robust estimates of average earnings. For various combinations of age/sex/industry/region groupings, we

calculated weighted estimates of average weekly earnings. In instances when sample size is deemed to small for an accurate estimate, we consolidated groups until we had enough persons in the group. Due to a complex sample design, we used a bootstrap method to estimate standard errors and 95% confidence intervals for these weighted estimates. Our results are included in the appendix and we provide analysis with particular consideration for generalization to similar surveys.



## References

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

## Appendix

Table 2. Average reported weekly earnings (rupees) for various groups surveyed in the 2011 India NSS survey.

	Male			Female		
	Count	Estimate (90% CI)	SE	Count	Estimate (90% CI)	SE
<b>Under 60, Urban</b>						
Farming, forestry, or fishing	2995	412.9 (334.2, 491.6)	47.8	1452	242.9 (216.5, 269.2)	16
Manufacturing	7637	1546.2 (1453.6, 1638.9)	56.3	2666	332 (285.6, 378.4)	28.2
Construction	5133	1189.8 (1137, 1242.6)	32.1	457	976.7 (868, 1085.3)	66.1
Other	29284	1832.8 (1781, 1884.5)	31.4	6779	1870.3 (1767, 1973.7)	62.8
All	45049	1632.4 (1594.8, 1670)	22.8	11354	1250.1 (1186.5, 1313.7)	38.7
<b>Under 60, Rural</b>						
Farming, forestry, or fishing	28072	282 (274.1, 289.9)	4.8	17272	194.9 (186.9, 202.8)	4.8
Manufacturing	6396	781.4 (727, 835.8)	33	2658	171.4 (152.8, 190)	11.3
Construction	10290	957.9 (941.5, 974.2)	10	1725	639.8 (618.3, 661.3)	13.1
Other	24041	1018.1 (986.3, 1049.9)	19.3	4739	840.9 (773, 908.8)	41.3
All	68799	573.6 (563.1, 584.2)	6.4	26394	288.2 (278.5, 297.8)	5.9
<b>Under 60, Urban and Rural</b>						
Farming, forestry, or fishing	31067	286.8 (278.6, 295.1)	5	18724	196.5 (188.9, 204)	4.6
Manufacturing	14033	1197.1 (1143.9, 1250.3)	32.4	5324	237.5 (215.3, 259.7)	13.5
Construction	15423	1016.2 (998, 1034.4)	11.1	2182	688.2 (663.1, 713.3)	15.3
Other	53325	1477.3 (1445, 1509.5)	19.6	11518	1460.9 (1394.5, 1527.3)	40.4
All	113848	898.4 (884, 912.9)	8.8	37748	494.2 (477.8, 510.7)	10
<b>60-64, Urban</b>						
Farming, forestry, or fishing	284	185.5 (132.3, 238.7)	32.3	91	-	-
Manufacturing	223	853.2 (534.8, 1171.7)	193.6	73	-	-
Construction	124	862.5 (705, 1020.1)	95.8	8	-	-
Manufacturing or construction	347	856 (628.7, 1083.2)	138.1	81	-	-

Table 2. Average reported weekly earnings (rupees) for various groups surveyed in the 2011 India NSS survey.  
(continued)

	Count	Estimate (90% CI)	SE	Count	Estimate (90% CI)	SE
Other	801	628.6 (456.4, 800.8)	104.7	183	955.3 (280.9, 1629.6)	410
All	1432	618.5 (497.6, 739.3)	73.5	355	534.5 (188.4, 880.6)	210.4
<b>60-64, Rural</b>						
Farming, forestry, or fishing	2291	196.8 (168.3, 225.3)	17.4	824	169.7 (141.6, 197.9)	17.1
Manufacturing	234	372.7 (206.4, 539.1)	101.1	60	-	-
Construction	275	839.1 (758.2, 920.1)	49.2	62	-	-
Manufacturing or construction	509	620.4 (534.7, 706.2)	52.1	122	-	-
Other	666	303.7 (216.5, 391)	53	167	220.6 (96.6, 344.5)	75.3
All	3466	265.9 (239.5, 292.4)	16.1	1113	190.4 (162.4, 218.4)	17
<b>60-64, Urban and Rural</b>						
Farming, forestry, or fishing	2575	196.3 (170.4, 222.1)	15.7	915	169.8 (142, 197.6)	16.9
Manufacturing	457	566.9 (414.6, 719.2)	92.6	133	-	-
Construction	399	843.8 (773.7, 914)	42.6	70	-	-
Manufacturing or construction	856	693.6 (602.7, 784.5)	55.2	203	244.1 (178.8, 309.4)	39.7
Other	1467	473 (369.2, 576.8)	63.1	350	566.9 (228.3, 905.6)	205.8
All	4898	330.4 (298.8, 362.1)	19.3	1468	238.9 (182.2, 295.6)	34.5
<b>65-69, Urban and Rural</b>						
Farming, forestry, or fishing	1638	178.3 (141.9, 214.8)	22.1	460	186.8 (135.8, 237.7)	31
Manufacturing	239	316.8 (219.3, 414.4)	59.3	61	-	-
Construction	180	940.6 (847.1, 1034)	56.8	33	-	-
Manufacturing or construction	419	590.7 (503.2, 678.1)	53.2	94	-	-
Other	768	297.8 (182.9, 412.8)	69.9	161	256.4 (119.9, 392.8)	82.9
All	2825	259 (223.4, 294.5)	21.6	715	218.9 (172, 265.7)	28.5
<b>60-69, Urban</b>						
Farming, forestry, or fishing	442	168.7 (126.9, 210.6)	25.5	132	-	-
Manufacturing	333	696.9 (481, 912.7)	131.2	103	-	-
Construction	175	913.2 (786.8, 1039.5)	76.8	10	-	-
Manufacturing or construction	508	761.3 (603.2, 919.3)	96.1	113	-	-

Table 2. Average reported weekly earnings (rupees) for various groups surveyed in the 2011 India NSS survey.  
(continued)

	Count	Estimate (90% CI)	SE	Count	Estimate (90% CI)	SE
Other	1212	567.3 (429.8, 704.8)	83.6	274	796.2 (296.6, 1295.7)	303.7
All	2162	555.7 (469.8, 641.5)	52.2	519	472.4 (201.1, 743.7)	164.9
<b>60-69, Rural</b>						
Farming, forestry, or fishing	3771	190.8 (167.7, 213.9)	14	1243	175.1 (149.6, 200.5)	15.5
Manufacturing	363	328 (198.8, 457.3)	78.6	91	-	-
Construction	404	863.1 (798.8, 927.5)	39.1	93	-	-
Manufacturing or construction	767	610.5 (544.4, 676.7)	40.2	184	339.6 (257.5, 421.7)	49.9
Other	1023	243.1 (181.6, 304.6)	37.4	237	185.2 (101.1, 269.3)	51.1
All	5561	248.6 (227.9, 269.3)	12.6	1664	195.5 (171.3, 219.7)	14.7
<b>60-69, Urban and Rural</b>						
Farming, forestry, or fishing	4213	189.8 (168.8, 210.7)	12.7	1375	175.1 (149.1, 201)	15.8
Manufacturing	696	484.7 (371.4, 597.9)	68.8	194	118.6 (53.3, 184)	39.7
Construction	579	874.1 (816.2, 931.9)	35.2	103	-	-
Manufacturing or construction	1275	660.5 (594.1, 726.9)	40.4	297	271.8 (208.1, 335.6)	38.7
Other	2235	408.1 (329, 487.3)	48.1	511	460.3 (233.3, 687.4)	138
All	7723	304.8 (281.1, 328.6)	14.5	2183	232.7 (191.8, 273.5)	24.8
<b>70-74, Urban and Rural</b>						
Farming	768	79.5 (49.7, 109.2)	18.1	157	168.9 (97.2, 240.6)	43.6
Other	530	433.5 (213.1, 654)	134	101	-	-
All	1298	188.2 (111.9, 264.5)	46.4	258	228.1 (154, 302.2)	45.1
<b>75-79, Urban and Rural</b>						
Farming	269	83.4 (25.7, 141.1)	35.1	38	-	-
Other	179	102 (53.7, 150.4)	29.4	27	-	-
All	448	89.7 (48.2, 131.1)	25.2	65	-	-
<b>70-79, Urban and Rural</b>						
Farming	1037	80.3 (53.7, 107)	16.2	195	168.3 (107.2, 229.4)	37.2
Other	709	351.6 (174.6, 528.5)	107.6	128	-	-
All	1746	165.5 (102.1, 228.9)	38.6	323	214.5 (149.9, 279.1)	39.3

Table 2. Average reported weekly earnings (rupees) for various groups surveyed in the 2011 India NSS survey.  
(continued)

	Count	Estimate (90% CI)	SE	Count	Estimate (90% CI)	SE
<b>70+, Urban and Rural</b>						
Farming	1206	82 (56.7, 107.2)	15.3	215	152.8 (92, 213.7)	37
Other	788	343 (185.1, 500.9)	96	141	-	-
All	1994	162.1 (107.9, 216.3)	32.9	356	193.2 (137.5, 248.9)	33.9
<b>75+, Urban and Rural</b>						
All	696	106.1 (59.2, 153)	28.5	98	113.5 (43.6, 183.3)	42.4
<b>80+, Urban and Rural</b>						
All	248	135.6 (33.3, 238)	62.2	33	-	-

<sup>1</sup> Average reported weekly earnings include employed people reporting 0 earned income.

<sup>2</sup> If the sample size of a group is less than 150, we do not provide an estimate of weekly earnings because we are not confident that the estimate is statistically robust. The exception is for women aged 75+ and men aged 60-64 working in construction in urban environments where estimates are included for the sake of completeness.