EASYPol

On-line resource materials for policy making

ANALYTICAL TOOLS

Module 081

Policy Impacts on Inequality

Weighting Samples: Introductory Issues

Policy Impacts on Inequality

Weighting Samples: Introductory Issues

by

Lorenzo Giovanni Bellù, Agricultural Policy Support Service, Policy Assistance Division, FAO, Rome, Italy

Paolo Liberati, University of Urbino, "Carlo Bo", Institute of Economics, Urbino, Italy

for the

Food and Agriculture Organization of the United Nations, FAO

About EASYPol
EASYPol is a an on-line, interactive multilingual repository of downloadable resource materials for capacity development in policy making for food, agriculture and rural development. The EASYPol home page is available at: ***WWW.FAO.ORG/TC/EASYPOL.***
EASYPol has been developed and is maintained by the Agricultural Policy Support Service, Policy Assistance Division, FAO.

Table of contents

1. SUMMARY

This module illustrates the concept of weighting samples in survey analysis. In particular, it deals with: a) the concept of weights and inflation factors; b) why weights are important in calculating statistical measures; c) how disregarding weights may lead to misleading results in empirical analysis.

2. INTRODUCTION

Objectives

The objective of the tool is to explain the concept of weights, why they are important and what are the consequences of not using them. This gives analysts the tools required to infer population measures from surveys.

Weighting is a procedure aimed at assigning to each *observed* household (or individual) in a sample a number representing how many households (or individuals) in the *actual* population it represents.

The weighting procedure is important in all cases where the probability of households (individuals) of being sampled is different. For policy analysis, i.e. analysis of impact of policies, microsimulation analysis, etc., weighting is essential.

Target audience

This module targets current or future policy analysts who want to increase their capacities in analysing impacts of development policies on inequality by means of income distribution analysis. On these grounds, economists and practitioners working in public administrations, in NGOs, professional organisations or consulting firms will find this helpful reference material.

Required background

Users should be familiar with basic notions of mathematics and statistics.

Links to relevant EASYPol modules, further readings and references are included both in the footnotes and in section xxxxx of this module[1].

[1] EASYPol hyperlinks are shown in blue, as follows:
a) training paths are shown in **underlined bold font;**
b) other EASYPol modules or complementary EASYPol materials are in ***bold underlined italics***;
c) links to the glossary are in **bold**; and
d) external links are in *italics*.

3. CONCEPTUAL BACKGROUND.

The most common situation when using either household or individual surveys is to have different households or individuals in the survey representing a different number of households or individuals in the actual population.

This happens for two main reasons: the first is that the survey is usually a sample of the actual population, i.e. the number of *observed* households is less than the number of *actual* households. The second is that each household has a different *probability* of being selected, i.e. to be in the sample.

The first reason is obvious. As populations may be extremely large, surveys can only try to represent the population by drawing from a sample.

The second reason is less obvious. Different households may have different probabilities of being selected. This may happen, for example, because some households are difficult to survey, or because a given group of households gives a very low rate of response to the survey. For example, the very poor and the very rich are typically underrepresented in household surveys, as both are difficult to survey. Some specific groups of population can also be underrepresented not because they cannot be potentially surveyed but just because they decide not to answer to the survey.

As a result, we typically end up with household surveys in which each household represents a different number of households in the actual population. In this case, some typical measures may be biased and this issue must be dealt with in policy analysis.

Saying that each *observed* household represents a different number of *actual* households means that each *observed* household is attached a different **weight**. If we observed household represents, for instance, 100 actual households means to say that it has more weight than an observed household representing maybe 20 actual households. In other words, the characteristics of the first household weight more in defining statistical measures.

Where is this weight from? To deal with this issue, we can define an actual population of N households and a probability p_i of being sampled assigned to each household i in the population. Note that the probability p_i is the ex-ante probability that a given household may be selected in a draw. If actual households are in different conditions or the sample is designed to select specific groups of households, the ex-ante probabilities will be different. If the survey is designed to have n sampled (observed) households, the total probability of having a given household in the sample is the product np_i, i.e. the product of the total number of draws n times the probability of being selected in each draw p_i.[2]

Given these parameters, the weight v_i attached to each household can be calculated as the reciprocal of the total probability:

[2] Note that this conclusion holds when the sample is small relative to the population, so that the probability of a household of appearing more than once is small. See Deaton (1997).

[1] $$v_i = \frac{1}{np_i}$$

For example, if we want a sample of $\boldsymbol{n}$=100 households out of a population of $\boldsymbol{N}$=1000 and the probability of being selected of a given household is $\boldsymbol{p_i}$=0.05, the corresponding weight will be $v_i = \frac{1}{100 \cdot 0.05} = \frac{1}{5} = 0.2$.

Now, in order to go from the weight to the actual number of households represented by that observed households it is worth having a structure of weights normalised to one. This is done by multiplying each weight $\boldsymbol{v_i}$ by the sum of weights:

[2] $$w_i = \frac{v_i}{\sum_{i=1}^{n} v_i}$$

By multiplying each normalised weight $\boldsymbol{w_i}$ by the actual number of households we can get the number of households represented by each household. Let's call this new number the inflation factor specific to each household ($\boldsymbol{IF_i}$):

[3] $$IF_i = w_i \cdot N$$

Note that if the probability of each household to be selected were equal, and therefore $p_i = \frac{1}{N}$ for all $\boldsymbol{i}$, the total probability for any household to be selected in $\boldsymbol{n}$ trials would be $\frac{n}{N}$. The corresponding weight would be $v_i = \frac{N}{n}$ and the corresponding normalised weight would be $w_i = \frac{1}{n}$, as $\sum_{i=1}^{n} v_i = \sum_{i=1}^{n} \frac{N}{n} = n\frac{N}{n} = N$.

From [3] it is easy to note that:

[4] $$\sum_{i=1}^{n} IF_i = \sum_{i=1}^{n} w_i N = N \sum_{i=1}^{n} w_i = N$$

as, by definition, $\sum_{i=1}^{n} w_i = 1$. This means that, in household surveys, if we want an estimate of the total population, it is enough to sum the inflation factors specific to each household.

4. A NUMERICAL EXAMPLE

Table 1 reports an example of how weights and inflation factors work. Let's assume a total population of N=100 households of the same size and composition.[3] Let's also assume that we want to draw a sample of n=5 individuals and that the probabilities p_i of each household to participate in the sample differ. In the example, household 1 has a 1 per cent probability to participate; household 2 has 40 per cent; household 3 has 32 per cent; household 4 has 23 per cent and household 5 has 4 per cent.

Table 1: Weights and inflation factors

Total population **100**

	Household	p_i	v_i	w_i	IF_i
	1	0.01	20.0	0.74	74.09
	2	0.40	0.5	0.02	1.85
	3	0.32	0.6	0.02	2.32
	4	0.23	0.9	0.03	3.22
	5	0.04	5.0	0.19	18.52
Totals	5	1.00	27.0	1.00	**100.00**

The sum of inflation factors gives total population

According to formula [1], the corresponding weights v_i are calculated. Also, the normalised weights w_i can be calculated by dividing each v_i by the sum of them (27.0). This gives the numbers in the fourth column of table 1, whose sum is 1. Finally, by multiplying each w_i by total population N gives how many households each household must represent. As can be easily seen, and also highlighted in Table 1, the sum of inflation factors gives N, as in formula [4]. The column of inflation factors tells us that household 1 must represent about 74 households in total population; households 2 and 3 must represent about two households each; household 4 must represent about 3 households; and household 5 must represent about 19 households in total population.

The interesting thing to note is that the number of households each household must represent to have a statistically significant sample, is in inverse relation with the probability to participate in the sample. If the probability is low, each participating

[3] This allows us to disregard the issue of equivalence scale and to deal with total household income as a proper measure of household welfare.

household must represent a higher number of households in order to avoid *under-representation*; while if the probability is high, each participating household must represent a lower number of households in order to avoid *over-representation*.

5. AN APPLICATION TO MEAN AND TOTAL INCOMES

An obvious question to ask is why weighting is important. The answer is that only by weighting can we obtain the correct statistical measures of the variables of interest. An example is reported in Table 2 for the same set of households as in Table 1.

Table 2: Unweighted and weighted total income

Total population	**100**			
	Household	*IFi*	Observed income	Total income represented by each household
	1	74.09	1,000	74,089
	2	1.85	3,000	5,557
	3	2.32	4,500	10,419
	4	3.22	5,000	16,106
	5	18.52	9,800	181,518
Totals	5	100.00	23,300	287,689

Table 2 shows the observed income attached to each sampled household and the sample total (23,300 income units). As each household represents a given number of actual households, however, an estimate of the *population* total income is reported in the far right column of Table 2, where each income is multiplied by the inflation factor. This gives the population total income equal to 287,689 income units.

Table 3 shows how the use of weights may affect the estimate of average income.

Table 3: Unweighted and weighted means

Total population	100			
	Household	*IFi*	Observed income	Total income represented by each household
	1	74.09	1,000	74,089
	2	1.85	3,000	5,557
	3	2.32	4,500	10,419
	4	3.22	5,000	16,106
	5	18.52	9,800	181,518
Totals	5	100.00	23,300	287,689
Means			4,660	2,877

In Table 3, average income calculated on the sample basis would be equal to 4,660 income units, obtained as 23,300/5. However, given the very different weight of household surveyed, the *population* average income is equal to 2,877 income units, obtained as 287,689/100. Without weighting, we can make serious errors in estimating statistical measure of the variables of interest.

This means that all inequality and poverty measurements should be done by weighting samples, in order to have reliable estimates of *population* variables.

6. AN APPLICATION TO THE HEADCOUNT RATIO

Table 4 reports a simple example where calculating the headcount ratio without weighting may lead to very misleading results.

Table 4: Unweighted and weighted headcount ratio

Total population	100						
	Household	*IFi*	Observed income	Total income represented by each household	Poverty line	Poor in the sample	Poor in population
	1	74.09	1,000	74,089	3,500	1	74.09
	2	1.85	3,000	5,557	3,500	1	1.85
	3	2.32	4,500	10,419	3,500	0	0.00
	4	3.22	5,000	16,106	3,500	0	0.00
	5	18.52	9,800	181,518	3,500	0	0.00
Totals	5	100.00	23,300	287,689		2	75.94

Headcount ratio in the sample		0.4
Headcount ratio in population		0.76

Given a poverty line of 3,500 income units, the sample records two households below the poverty line (the first and the second). If calculated on a sample basis, the headcount ratio would be 2/5=0.4, i.e. 40 per cent of households below the poverty line. However, as the first household represents about 74 households in actual population, weighting the sample gives that the total number of households below the poverty line is about 76, as reported in the far right column of Table 4. This column is calculated by multiplying the indicator of poverty (1 if poor, zero otherwise) by the inflation factor. The fact that the first household observed is poor means that about 74 households in total population are poor. This means that, when measured on total population, the headcount ratio would be about 76/100=0.76, i.e. about 76 per cent of population below the poverty line, a quite different figure compared with the unweighted one!

7. AN APPLICATION TO THE LORENZ CURVE

In order to build a **Lorenz Curve** which represents the entire population from which the sample is drawn, we must again weight each observation with the expansion factor to generate both the correct cumulative income distribution and the cumulative proportion of population.

Recalling the definition of the Lorenz Curve and applying it to a weigthed sample, the *x-axis* records now the weighted cumulative proportion of population while the *y-axis* records the weighted cumulated proportion of income of a given share of the population.

For a household-based analysis, the Lorenz Curve is built as follows:

$$L\left(\frac{k}{P}\right)=\frac{\sum_{i=1}^{k} y_i \cdot IF_i}{Y}$$

where :

$k=\sum_{i=1}^{k} IF_i$ is the position of each weighted household in the income distribution and $i=1\ldots k$ is the position of each household in the income distribution;

$P=\sum_{i=1}^{n} IF_i$ is the total number of weighted households in the distribution;

$\sum_{i=1}^{k} y_i \cdot IF_i$ is the cumulated income up to the k^{th} weighted household (y_i is the income of the i^{th} weighted household in the distribution)

$Y = \sum_{i=1}^{n} y_i \cdot IF_i$ is the total weighted income of the distribution.

Example: There are five households whose Per Capita Equivalent Income are indicated in Table 5 with PCEI and expressed in monetary units[4]. Let's assume variables as households size and expansion factors.

Table 5: Example

Household	household size (hs)	expansion factor (IF)	PCEI
1	3	25	1,000
2	4	20	750
3	2	30	1,200
4	5	15	850
5	5	10	1,200

The first step in building the Lorenz Curve is to calculate the weighted proportion of population. In this case, the sample is drawn on households which have a different size and represent different shares of population.

The cumulative population is therefore the cumulated sum of the weighted households for each observation. To calculate the cumulative proportion of population, for each observation (which ranges between 0 and 1), we have to divide the cumulative population by the total number of households (sse Table 6).

Table 6: Cumulative proportion of population

Household	Cumulative weighted household (Σ *IF*)	Cumulative proportion of weighted household
1	25	0.25
2	45	0.45
3	75	0.75
4	90	0.9
5	100	1

After calculating the weighted proportion of population, let's point out the cumulative proportion of income. Assuming PCIE as the proper welfare measure for each household, the procedure is the same as used for cumulative proportion of population.

The cumulative proportion of income is the cumulated sum of the weighted income divided for the total income in order to have the range (0,1), Table 7.

[4] PCIE is the ratio of total household income and a given equivalence scale. In this case, we are also introducing the complication of having households of different size and composition, which requires to calculate a proper measure of household welfare given by the per capita income equivalent.

Table 7: Cumulative proportion of income

Household	Weighted income	Cumulative weighted income	Cumulative proportion of weighted income
1	25,000	25,000	0.07
2	33,750	58,750	0.17
3	90,000	148,750	0.43
4	76,500	225,250	0.65
5	120,000	345,250	1.00

Summarising results, Table 8 reports the cumulative proportion of households and the corresponding cumulative proportion of income defining the coordinates of the Lorenz Curve.

Table 8: A weighted Lorenz Curve

Household	Cumulative proportion of weighted households	Cumulative proportion of weighted income
1	0.25	0.07
2	0.45	0.17
3	0.75	0.43
4	0.90	0.65
5	1.00	1.00

8. READERS' NOTES

8.1. Time requirements

Time required to deliver this module is estimated in about two hours.

8.2. EASYPol links

Selected EASYPol modules may be used to apply weighting schemes to inequality indexes.

This module belongs to a set of modules that discuss how to compare, on inequality grounds, alternative income distributions generated by different policy options. It is

part of the series of modules composing a training path addressing **Analysis and monitoring of socio-economic impacts of policies.**

8.3. Frequently asked questions

- What is the meaning of weighting observed households or individuals in a survey?
- Why is weighting important?
- What are the consequences of disregarding weighting procedures?

9. REFERENCES

A good discussion on weighting can be found in:

Deaton A., 1997. *The Analysis of Household Surveys*, Baltimora, Johns Hopkins University Press, USA.

Module metadata

1. EASYPol module	081

2. Title in original language	
English	Policy Impacts on Inequality
French	
Spanish	
Other language	

3. Subtitle in original language	
English	Weighting Samples: Introductory Issues
French	
Spanish	
Other language	

4. Summary

This module illustrates the concept of weighting samples in survey analysis. In particular, it deals with: a) the concept of weights and inflation factors; b) why weights are important in calculating statistical measures; c) how disregarding weights may lead to misleading results in empirical analysis.

5. Date

December 2006

6. Author(s)

Lorenzo Giovanni Bellù, Agricultural Policy Support Service, Policy Assistance Division, FAO, Rome, Italy
Paolo Liberati, University of Urbino, "Carlo Bo", Institute of Economics, Urbino, Italy

7. Module type

- ☐ Thematic overview
- ☐ Conceptual and technical materials
- ☒ Analytical tools
- ☐ Applied materials
- ☐ Complementary resources

8. Topic covered by the module

- ☒ Agriculture in the macroeconomic context
- ☐ Agricultural and sub-sectoral policies
- ☐ Agro-industry and food chain policies
- ☐ Environment and sustainability
- ☐ Institutional and organizational development
- ☐ Investment planning and policies
- ☒ Poverty and food security
- ☐ Regional integration and international trade
- ☐ Rural Development

9.Subtopics covered by the module

10. Training path

Analysis and monitoring of socio-economic impacts of policies

11. Keywords

capacity building, agriculture, agricultural policies, agricultural development, development policies, policy analysis, policy impact anlaysis, poverty, poor, food security, analytical tool, inequality, inequality axioms, inequality indexes, inequality measures, income inequality, income distribution, income ranking, welfare measures, social welfare, lorenz curve, lorenz dominance, theil index, gini index, concept of weights, inflation factors