Estimation of the distribution of income from survey data, adjusting for compatibility with other sources. [1]

Víctor Alfredo Bustos y de la Tijera[2]

INEGI

Abstract

In this paper we present an approach for the estimation of income distributions, which is aimed at dealing with survey data shortcomings through simultaneous consideration of other statistical sources and through adjustment for compatibility with all of them. Our proposal is based on well-established statistical criteria and methods and thus reduces the need for subjective or arbitrary choices. It has the purpose of selecting the distributional model that best fits the data from the survey, using a Constrained Pseudo Log-likelihood criterion. We show how our proposal deals both with income under-reporting and with truncation which are known to be present in the survey. We then apply our procedure to Mexican data from the National Survey on Household Income and Expenditure for the year 2012, and from Mexico's System of National Accounts, sources that produce widely differing results regarding total household current income for the country. We show that, among all fitted models, a satisfactory explanation is given by a 4-parameter Generalized Beta Type 2 distribution. The chosen distribution has little impact on the official poverty measurement. The Gini coefficient, however, reaches a value as high as 0.803.

Keywords: Income distribution, Log-Normal, Gamma, Generalized Gamma, Generalized Beta, pseudo-likelihood, numerical optimization, nonlinear constraints.

[1] The results shown in this paper do not necessarily reflect those of Mexico's National Institute of Statistics and Geography (INEGI) and are not to be considered official in any way.

[2] Dr. Víctor Alfredo Bustos y de la Tijera, researcher, INEGI, Av. Héroe de Nacozari Sur No. 2301, Aguascalientes, Ags., 20276 México, Tel. (+52) 449 910-5431, e-mail: alfredo.bustos@inegi.org.mx

1. INTRODUCTION

In studies regarding economic inequality based on the distribution of income of Mexican households it is customary to consider several statistical sources which, in view of their diverse nature, lead to different results. In particular, the National Survey on Household Income and Expenditure (ENIGH) collects information about various sources of income which, when added, form the total current income for each household. On the other hand, the Mexican System of National Accounts (SCNM) which, following UN recommendations, produces the institutional sector accounts among which the household sector is included. Among other results, SCNM produces the yearly national gross disposable household income. It has been pointed out by many authors that the total household income estimates produced by each of the two sources show a significant discrepancy and indicate as the two main reasons the under-reporting of income by households in the survey, and the truncation effect resulting from the unintended exclusion from the sample of households with very high incomes. This has led to numerous proposals that try to make both results compatible. However, those who have evaluated these proposals agree that, so far, they exhibit a high degree of arbitrariness since they are based on assumptions that are far from becoming conventions any time soon.

This brief background account is based primarily on Leyva [7] in which the author analyzes various proposals that have appeared in the literature on how to adjust household income survey data to results from the System of National Accounts. From such an analysis he suggests that the adjustment proposals are, in general, based upon the following basic assumptions:

1. Income concepts used by both sources are comparable.

2. Revenue figures produced by the national accounts are at least as plausible as those from household income surveys.
3. Differences between the two sources are mainly due to income under-reporting problems rather than to truncation issues.
4. There exists an optimal allocation rule that allows distributing the differences in household income, at the macro level, to the income (expanded) of each household in the sample of the income survey (micro level).

Of course some of these assumptions are more sensible than the rest while others do not resist the weight of the evidence. For instance, if the small group of people whose income is very important in relation to the rest, is underrepresented or not represented at all in the sample, "the value of total expanded income derived from the survey, even if income was not under-reported, should be lower than that obtained from the National Accounts, whose methodology and coverage includes in principle all income earners, without exception." Absence of such a group of households in the sample, results in truncation censorship which can be substantial. Consequently, the third hypothesis does not correspond with reality. Furthermore, if in adjusting to SCNM this truncation is ignored, the result may be the redistribution of an unknown and perhaps important amount among sampled households. In fact, the observed income will increase in varying proportions which will lead to unpredictable consequences in the implementation of social policies; for example, by artificially reducing the number of households in poverty conditions. Leyva concludes that the distinction between under-reporting and truncation is not trivial and that there is currently no robust procedure for carrying it out. Incidentally, Leyva [7] questions the validity of the second assumption regarding the use of figures from the SCNM as a reference for the correction of survey data on household income, an idea to which we shall briefly return later on.

Although in a different context, the recent controversy over the book by French economist Thomas Piketty [8] illustrates that the problem discussed above for the case of Mexico, has been and is present in other latitudes. One of the most widespread criticisms to the results presented in the book came from Chris Giles [2], economics editor of the Financial Times (FT), who reviews Piketty's data and results, and argues that they contain a series of errors of various kinds that skew the book's findings. In fact, Giles concludes that, after correcting apparent errors, he does not find evidence that the concentration of wealth has increased in the most recent 30 years in the UK. In his response, Piketty [9] argues that some of the corrections made by FT are minor and do not alter his conclusions; others are based on methodological choices that are debatable ("to say the least"). Piketty found that among the methodological choices made by FT, particularly problematic is the initial use of fiscal sources for the earlier decades and the change to sample estimates for more recent periods. He goes on to say that "(this) is problematic because we know that in every country surveys tend to underestimate the wealth share of the richest in contrast to administrative data based on tax estimators. Therefore, these methodological choices may bias the results toward declining inequality".

In turn, Krugman [6] suggests that the alleged errors raised by Giles were essentially "the kinds of data adjustments that are normal in any research that relies on a variety of sources. And the crucial assertion that there is no clear trend toward increased concentration of wealth rested on a known fallacy, an apples-to-oranges comparison that experts have long warned about". He goes on to say: "We have two sources of evidence on both income and wealth: surveys, in which people are asked about their finances, and tax data. Survey data, while useful for tracking the poor and the middle class, notoriously under-report top incomes and wealth — loosely speaking, because it's hard to interview enough billionaires. So studies of the 1 percent, the 0.1 percent, and so on rely mainly on tax data. The Financial Times critique, however, compared older estimates of wealth concentration based on tax data with more recent estimates based on surveys; this produced an automatic bias against finding an upward trend".

Leyva [7] concludes that "… even if figures on household revenue provided by the national accounts could effectively be seen as more reliable than those from ENIGH, the problem persists that no criteria have been developed yet for optimal allocation to bridge the gap and to distribute the macroeconomic differences at the micro-level; in the apples-and-oranges allegory, for turning apples into oranges. This problem is compounded when one considers the fact that National Accounts and ENIGH refer to different household universes, given that there is a fraction, which could be large, of income reported in macroeconomic statistics (assumed to include all household groups), which is not captured in ENIGH-type surveys, which implies the existence of a truncation (at least) in the upper part of the income distribution. Thus, even if an optimal allocation rule to move from macro to micro were available, it would be necessary to know which part of the discrepancy between National Accounts and ENIGH corresponds to under-reporting and which to truncation." He concludes "it is therefore desirable to carry out research to begin to shed some light on this matter."

The purpose of this paper is precisely to cater to this last recommendation. The following section presents our approach, together with some necessary prerequisites. This is followed by a numerical example where, under the understanding that there is a significant discrepancy between the total household income according to each source (ENIGH or National Accounts), we estimate the distribution of household income in Mexico by combining information from both sources (ENIGH 2012 will be used for exemplification); in other words, to produce an apple and orange salad with each ingredient contributing its essence. The results are then briefly discussed, together with some graphical assistance to be presented, and the relevant differences are commented on. A final section which presents some closing remarks is then included.

2. THE CRITERION

2.1 The Constrained Pseudo Log-likelihood criterion (CPLL)[3].

The pseudo log-likelihood maximization takes into account only the available data and the sample design that gave rise to them but is still short of achieving the purpose of reconciling the result with other information sources. To achieve the latter purpose, we consider additional information in the form of constraints on the values that one or more, possibly non-linear, functions of the parameters may take. For example, if among other constraints, we want the average of the fitted distribution to match a specific value from an alternative source, we express the former as a function of the parameters $E(\underline{Y} \mid \underline{\theta})$ and force it to assume a particular value $\underline{c}$, from the alternative source; in other words, we ask that $E(Y \mid \underline{\theta}) \equiv c$. In this fashion, parameter estimates are obtained from the solution of the optimization problem posed in (1).

$$\underset{\underline{\theta},\underline{\lambda}}{Max}\left\{\sum_{i=1}^{n}\frac{1}{\pi_{(i)}}\ell\left(\underline{\theta};\underline{Y}_{(i)}\right)-\underline{\lambda}'\left[\underline{h}\left(\underline{\theta}\right)-\underline{c}\right]\right\} \tag{1}$$

where

$\ell\left(\underline{\theta};\underline{Y}_{(i)}\right)$ represents the natural logarithm of the density function, evaluated at the i-th sample value;

[3] For details about its derivation refer to the Technical Appendix

$\pi_{(i)}$, inclusion probability for that sample unit;

$\underline{h}(\underline{\theta})$, one or more functions of the parameter vector whose values are to coincide with those in vector $\underline{c}$. Note that the dimension of $\underline{h}(\underline{\theta})$ cannot equal or exceed that of $\underline{\theta}$ because the survey data would become irrelevant.

$\underline{\lambda}$, vector of so called Lagrange multipliers.

The functional forms of ℓ and $\underline{h}$ vary according to the distribution being considered.

3. NUMERICAL EXAMPLE

3.1 Sources

The National Survey on Household Income and Expenditure (ENIGH) (see INEGI [3, 4]) was preceded by several surveys conducted by different public agencies from 1956. Specifically, in 1977 the Ministry of Planning and Budget (SPP) developed the National Survey on Household Income and Expenditure, a statistical exercise that was the immediate predecessor of the surveys carried out by the National Institute for Statistics and Geography (INEGI) for the periods: 1984, 1989, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2005, 2006, 2008, 2010 and 2012.

It is during this final stage that its objectives are stated as: "*To provide a statistical overview of the behavior of income and expenditure by households in terms of amount, origin and distribution; ENIGH additionally provides information on occupational and socio-demographic characteristics of household members, as well as the characteristics of their dwelling, including its infrastructure, and durable household appliances.*"

The total number of households considered in the ENIGH 2012 sample reached 9,005, which expanded to 31’559,379 for the whole country. Table 1 shows a summary of results from the 2012 survey that is relevant to our discussion. In it, that year's results are compared with those from the three previous surveys for 2006, 2008 and 2010. It would seem that only deciles I, II, III and X have reversed the declining trend due to the economic crisis that began in 2008 but have not reached yet that year estimated levels.

<<Table 1. Quarterly average total current income per household in deciles of households by year of survey and Gini coefficients (2012 Constant Prices)>>

Meanwhile the System of National Accounts of Mexico (SCNM) (INEGI [5]) already incorporates the latest guidelines of the 2008 SNA by the United Nations, by the International Monetary Fund, the World Bank, the Organization for Economic Cooperation and Development and by Eurostat, at the same time shows the results of changing base year to 2008; thus the economic structure of the country and the prices used to measure the macroeconomic variables without the effect of inflation is updated.

In general, National Accounts provide an organizational scheme for developing knowledge about various macroeconomic aspects of the country, such as: production, consumption, savings, investment by sector of economic activity, and primary and secondary distribution of income; as well as foreign financial transactions and economic relations by institutional sectors.

In turn, the institutional sector accounts record transactions on the redistribution of national income, through taxes, contributions and transfers that occur between institutional sectors. Also record all actual transactions concerning the accumulation of financial and non-financial assets, as well as liabilities, allowing us to know the net saving position of each sector and of the economy as a whole. The vast majority of the recommendations contained in the new SNA 2008 manual of methods fall in this subsystem.

<<Table 2. Accounting balances by institutional sector, Year 2012, Base 2008, accounts for institutional sectors, System of National Accounts of Mexico.>>

It is to be stressed that while ENIGH's estimate for Quarterly Household Total Current Income reaches $1‘199,245‘100,000, SCNM figures in Table 2 show a Yearly Household Total Disposable Income of $10,908,255,712,944.10. When referred to the same time period, and consequently when we ignore the possible presence of seasonal influences, the SCNM value reaches to about 2.27[4] times its ENIGH equivalent, in 2012.

4. Results[5]

On the basis of quarterly incomes collected by ENIGH 2012 for households, the following 4 distributional forms were fitted (see Technical Appendix for details):

a) Gamma Distribution (G);

b) Log-normal Distribution (L-N);

c) Generalized Gamma Distribution (GG);

d) Generalized Beta, Type 2, Distribution (GB2).

The problem whose solution we seek in order to obtain parameter estimates is given in (1). In general, therefore, we seek to determine the value of the parameter vector that maximizes the pseudo log-likelihood, subject to suitable constraints[6]. Unconstrained weighted fits were also

[4] Leyva [7] found that for the years 2000 and 2002 this ratio reached 2.55 and 2.88, respectively.

[5] Calculations were performed by Miriam Romo (miriam.romo@inegi.org.mx), whose diligent assistance the author gratefully acknowledges.

[6] In all cases, R libraries were used; Alabama being prominent for numerical optimization, (http://cran.r-project.org/web/packages/alabama/alabama.pdf).

performed in order to help us understand through comparisons the effect of introducing constraints.

Note that, in our numerical example, the only constraint imposed assumes the following form:

$\underline{h}(\underline{\theta})$=E[Y|$\underline{\theta}$]= \$92,733.62[7]=Quarterly Average Household Income according to SNCM 2012.

In other words, all fitted distribution functions are adjusted so that their averages equal this figure. It is to be noted that the average quarterly household income reported by ENIGH for 2012 reaches \$38,125.00. Results are summarized in Table 3.

<< **Table 3. Summary maximum pseudo log-likelihood fits of 4 types of distributions to ENIGH 2012 data, both without and with restrictions on the mean value.>>**

Table 3 shows the impact of introducing constraints. Since these affect only the average value of the distributions, all of them are shown to satisfy the constraint. The values of the log-likelihood show an increasing trend. These values show major differences between them, which could lead us to conclude that it is necessary to consider at least 3 parameters. According to the above, the distribution GB2 is the one that gives the best fit and should be our chosen model. However, it is closely followed by GG and we are unable at this point to make a precise statement about the significance or not of their difference. This is the reason why we try to look at the quality of the fit achieved though alternative criteria. We will maintain the other two distributions since they help to understand these criteria.

<< **Figure 1 Unconstrained and Constrained Pseudo Log-likelihood Optima for 4 distributions.**>>

Columns 1 to 3 in Tables 4.a and 4.b show selected percentiles for the expanded empirical distribution, along with their 95% confidence intervals (CI). The remaining columns show the corresponding values for each of the fitted distributions. Cells shaded in red contain values smaller

[7] All monetary figures given in Mexican Pesos.

than the CI's lower limits; in yellow, values within the corresponding intervals; and in green, values larger than the CI's upper limits[8].

<< **Table 4a. Percentiles of distributions fitted by unconstrained maximum pseudo log-likelihood.>>**

<< **Table 4b. Percentiles of distributions fitted by constrained maximum pseudo log-likelihood.>>**

In the un-weighted case, in agreement with the criterion values shown in fig. 1, three of the four tested distributions show small differences among themselves, with percentile values mostly within the empirical confidence limits. In the weighted case, for GB2 the values of the lower percentiles (under 1%) lie outside and to the left of the empirical confidence intervals which would indicate income over-reporting in the sample. This is reversed and the values of percentiles between the 2nd and the 60th lie within the confidence limits. All greater percentiles have values that lie outside and to the right of the CIs; this would imply that the initial over-reporting is rapidly compensated by under-reporting in the mid-percentiles which remains low until, at the upper end, under-reporting becomes important enough for the percentile values to lie outside and to the right of the CIs.

A graphic summary of results is also possible and is introduced in order allow further insight into the results. Figures 2 and 3, next, are included as an aid in explaining how figures 4 and 5 are to be interpreted. In figure 2, empirical (dotted line) and the fitted Gamma distributions are included. Even though this distribution yields the poorest fit, we consider it a good example of the usefulness our graphic approach.

[8] We also fitted all four distributions (results not shown) using an unweighted and unconstrained version of the criterion. Note that optimal values of the criteria are not comparable. However, since in the unweighted case over 80% of the cells were coloured in red, we concluded that weighting was the right way to proceed.

<< **Figure 2 Interpretation of discrepancies between graphic representations of the empirical and the constrained maximum pseudo log-likelihood fitted distributions.**>>

When there is total agreement between the two curves we will say that the adjusted distribution does not provide evidence of either under or over-reporting in any of the income deciles. When the horizontal distance between the curves is constant, either under- or over-reporting is proportional to income at all percentiles. If this distance increases monotonically, the same phenomenon (under- or over-reporting) is present at all percentiles to a greater or lesser degree. When the curves intersect, there is an initial under- or over-reporting which later on is offset by the opposite effect on subsequent percentiles. If the fitted distribution lies always to the left (right) of the empirical one, we will say that we have evidence of over-reporting (under-reporting) for some of the initial percentiles, or for all of them. Finally, the presence of fitted percentiles with greater values than the largest income observed in the sample will be indication of truncation. Of course, combinations of these behaviors may be found in practice.

From Figure 2 we can conclude that, when the low percentiles for the fitted Gamma are compared to those of ENIGH, income over-reporting seems to be present. However, this is reversed given that it is fully compensated by the 15th percentile. From this percentile on, under-reporting grows, although at a declining pace. Indeed, from the 90th percentile the discrepancy decreases until it almost vanishes. In other words, the 10% highest income households' over-reporting nearly offsets the under-reporting accumulated over all previous deciles.

An alternative way of assessing the presence of reporting errors is exemplified in Figure 3 and is given in terms of the horizontal difference between the curves along the range of values of the (log) income variable. The cases discussed in the previous paragraph are transformed into the following behaviors. A vertical line which coincides with the y-axis shows lack of evidence of statement errors.

When the curve is vertical but located to the left or right of the vertical axis would indicate over or under-reporting proportional to income. If the curve falls to the left (right) of the vertical axis with negative (positive) slope, the same phenomenon has occurred along the income range. When the curve crosses the vertical axis at one or more points, the initial effect is completely reversed by its counterpart. For example, in Figure 3 the adjustment of a Gamma distribution to ENIGH 2012 data provides evidence of over-reporting in the initial deciles; this is gradually compensated and reversed by the opposite effect to accumulate a larger under-reporting between the 15th and 90th percentiles; the comparison again suggests the presence of growing over-reporting in the remaining percentiles, so that it almost reverses the accumulated under-reporting up to that point. This last statement, in particular, suggests that there is hardly any or no evidence of truncation. In my opinion, this is a highly improbable behavior in practice, suggesting a deficient fit.

<< **Figure 3 Horizontal differences between logarithms of empirical and fitted percentiles >>**

Graphically, the four distributions fitted to the data are presented in Figure 4, including the empirical distribution (dashed line). For the sake of completeness, we have included again the fitted Gamma distribution.

<< **Figure 4 Empirical and fitted distributions>>**

When a Log-Normal distribution is fitted to ENIGH 2012 data (Figs. 4.b and 5.b), over-reporting for the lower percentiles is smaller than in the Gamma case; something similar happens with the under-reporting in the immediate upper percentiles since it is until the second decile that both compensate. In contrast, since the fitted distribution must satisfy the restriction, greater under-reporting and greater evidence of truncation are implied. In practical terms, it seems that this second family of distributions improves the achieved fit, as the values of the pseudo log-likelihoods indicated. In fact, in many references on the subject this is a highly favored family of distributions.

<< **Fig. 5 Horizontal differences between empirical and fitted distributions>>**

When GB2, the best fitting distribution, is considered, 0.1 to 1 percentiles show an atypical behavior with respect to their immediate neighbors. Thereafter the graph of horizontal differences (Fig. 5.d) shows a nearly vertical growth along the vertical axis all the way until approximately the sixth decile. From then on under-reporting prevails and grows resulting in a marked truncation effect. This effect may be expressed both as a proportion of total income, as in (A.13), which yields 47.71% or as a proportion of the difference between total current income according to both sources, as in (A.14), which equals 80.83%. Therefore, income under-reporting amounts to 20.79% of such difference.

5. Consequences on the measurement of inequality and poverty in Mexico[9]

The choice of the Type 2 Generalized Beta model to describe the way in which income was distributed among Mexican households in 2012 may have implications on some of the best known official measurements of inequality and poverty in Mexico. For instance, for our fitted distribution the estimated value for the Gini coefficient for household income reaches 0.803. Note that for 2012, ENIGH produced a value of 0.440 and the unconstrained fit of GB2 resulted in a value of 0.449 for the same Gini coefficient. In turn, the aggregated income for the highest income 10% of households is about 90 times that of the lowest income 10% of households; for ENIGH the same ratio equals only 19.00 (see Table 1). Under our fitted GB2 model, about 99.09% of households receive an income smaller than or equal to the largest observed income in the sample. Therefore, the average quarterly income for this group reaches nearly MXN$48,977.16, which results in an average under-

[9] It should be noted that these estimates depend strongly on the value provided by SCNM. Leyva's observation emphasizing the need to review the SCNM results for possible biases becomes relevant, see Leyva (2004).

reporting close to MXN$10,800.00. The remaining 0.91% of households gets slightly over MXN$4'846,790.00 on average over the same period. Therefore, under present conditions measurements of income inequality undergo important changes but the explanation of their causes and consequences lies beyond the scope of this paper.

The above situation changes when the official measurement of poverty is concerned since only lower incomes are considered. The Mexican National Council for the Evaluation of Social Development Policies (CONEVAL), when measuring poverty, decided against adjusting survey income data in view of the arbitrariness already pointed out. Mexico's poverty measure follows a multifactor approach in which income plays a role along with the fulfillment of 6 aspects considered necessary for adequate social development (Household educational backwardness average; access to health services; access to social security; dwelling quality; access to basic dwelling services; and degree of social cohesion). Two welfare lines are defined. The first case is identified with an income considered sufficient to meet nourishment as well as other goods and services needs. The second one, termed "Minimum welfare line (MWL)", corresponds to an income high enough to guarantee adequate nourishment for a person. In this fashion (see fig. 6), when income lies below the first welfare line and one or more social needs are not met, a person is considered poor. People in extreme poverty would be those with incomes below MWL and for whom 3 or more social needs are not met.

<< Figure 6. Poverty multidimensional measurement in México.>>

Table 5 shows as percentages the result of applying the above definitions in 2012 to the Mexican population and compares them with our findings for households.

<< Table 5. Percentages of people and households whose income lie below each welfare line, 2012 >>

Two immediate differences become apparent between CONEVAL's approach and the one we have followed in this paper. First, CONEVAL deals with people and we have been referring to households; second, consideration of social factors other than income may preclude reference to income distribution percentiles. Therefore, direct comparison becomes impossible. However, since in general lower income households in Mexico are also formed by a larger number of people, it is safe to assume that households with incomes below the 50/60 percentile cover over 50/60% of the Mexican population. A welfare line at any of these household income percentiles would encompass at least all individuals with incomes below the corresponding welfare line. Fig. 7 shows, through the comparison of empirical and fitted distributions, how welfare lines up to the 60 percentile have little or no effect on poverty measurement.

<< Figure 7. Empirical and fitted GB2 distributions and poverty line.>>

6. Summary and conclusions

We have presented an alternative methodological approach that seeks to address the deficiencies reported in the study of income through household surveys; namely, those identified with both the under-reporting in household incomes, and the truncation in observed earnings due to households with very high income that rarely, if ever, are included in the sample. There is evidence that both problems exist when sample results are compared with estimates from other sources, such as the SNA or the tax system. The proposal consists of choosing the distribution that significantly best fits sample data, using numerical optimization techniques, from among various distribution functions that have been proposed in the literature for the study of income, taking into account the sample design that gave rise to the data and giving consideration to as many restrictions as necessary to achieve compatibility between the fitted results and those provided by alternative sources.

The implementation of the proposal has been exemplified by estimating 4 distributional forms using data from ENIGH 2012. For exemplification purposes, only one restriction on the average value of the fitted distribution was imposed. From the values of the criterion function, tables and graphs, as well as other statistics the choice is made. For the example, it was concluded that the most appropriate distribution was a Generalized Gamma with 3 parameters. In the absence of accurate statistical statements, in the form of hypothesis tests that take into account the conditions under which the fit is carried out, it becomes necessary to incorporate the opinion of the investigator, in particular to determine whether inclusion of additional parameters is justified or not.

While the proposal seeks to reduce dependence on arbitrary assumptions that are not supported by the available information, as mentioned at the end of the previous paragraph, the methodologies to achieve our purposes may not be available and the discretion of the investigator is still important. Something similar happens with respect to the quality and quantity of the constraints that must be considered. For example, given the availability of tax information, and always assuming that it is reliable, you may consider imposing restrictions related to the values of some percentiles. However, it is not possible to give a general rule for these or other conditions.

The numerical results exhibit behaviors that may seem extreme but are consequence of accepting without change, information from other sources. Clearly, if revision of the system of national accounts figures were to lead to a lower average quarterly income for households, the Gini coefficients could show versions closer to the sample estimates but the proposed methodology would remain unchanged. The review of methods to generate the results for institutional sector accounts is, however, beyond the scope of this paper.

REFERENCES

[1] Bandourian, R., McDonald, J., Turley, R., “A *comparison of parametric models of income distribution across countries and over time*”, Dept. of Economics, Brigham Young University, 2002.

[2] Giles, Ch., “Data problems with Capital in the 21st Century”, Financial Times, May 23, 2014. Retrieved June 30, 2014 from http://blogs.ft.com/money-supply/2014/05/23/data-problems-with-capital-in-the-21st-century/

[3] Instituto Nacional de Estadística y Geografía (INEGI), “Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH)”, retrieved on June 15th, 2014 from http://www.inegi.org.mx/est/contenidos/Proyectos/encuestas/hogares/regulares/enigh/presentacion.aspx

[4] Instituto Nacional de Estadística y Geografía (INEGI), “Nueva Construcción de la ENIGH 2012: Tabulations”, Downloaded on June 15th, 2014, from http://www3.inegi.org.mx/Sistemas/TabuladosBasicos/tabdirecto.aspx?s=est&c=33501

[5] Instituto Nacional de Estadística y Geografía (INEGI), “PIB y Cuentas Nacionales”, downloaded on June 15th, 2014, from http://www.inegi.org.mx/est/contenidos/proyectos/cn/

[6] Krugman, P., “On inequality denial”, The New York Times, June 2, 2014, page A21 of the New York edition. Retrieved June 30, 2014 from http://www.nytimes.com/2014/06/02/opinion/krugman-on-inequality-denial.html?_r=0

[7] Leyva-Parra, G., “El ajuste del ingreso de la ENIGH con la Contabilidad Nacional y la medición de la pobreza en México”, Serie: Documentos de Investigación, No. 19, SEDESOL, México, 2004.

[8] Piketty, T., “Capital in the Twenty-First Century”, translated by Arthur Goldhammer, *Harvard University Press,* 2014.

[9] Piketty, T., “Response to FT”, retrieved on June 15, 2014 from http://www.voxeu.org/article/factual-response-ft-s-fact-checking.

[10] Stasinopoulos D. M., Rigby R.A. and Akantziliotou C. (2008) Instructions on how to use the GAMLSS package in R. Second Edition. http://www.gamlss.org/wp-content/uploads/2013/01/gamlss-manual.pdf

Table 1. Quarterly average total current income per household in deciles of households by year of survey and Gini coefficients (2012 Constant Prices)

HOUSEHOLD DECILES	COLLECTION YEAR			
	2006	2008	2010	2012
TOTAL CURRENT INCOME	**43 698**	**42 865**	**37 574**	**38 125**
I	7 796	7 136	6 633	6 997
II	13 506	12 460	11 673	11 794
III	17 780	16 792	15 611	15 734
IV	22 161	20 986	19 650	19 513
V	27 072	25 628	23 973	23 914
VI	32 611	31 501	29 059	28 862
VII	40 357	39 381	35 605	35 570
VIII	50 788	50 084	45 089	44 849
IX	69 194	69 159	61 133	61 014
X	155 715	155 525	127 313	133 003
GINI COEFFICIENT	**0.445**	**0.457**	**0.435**	**0.440**

SOURCE: INEGI [4]. National Survey on Household Income and Expenditure, public file (2012).

Table 2. Accounting balances by institutional sector, Year 2012, Base 2008, accounts for institutional sectors, System of National Accounts of Mexico.

Millions of pesos

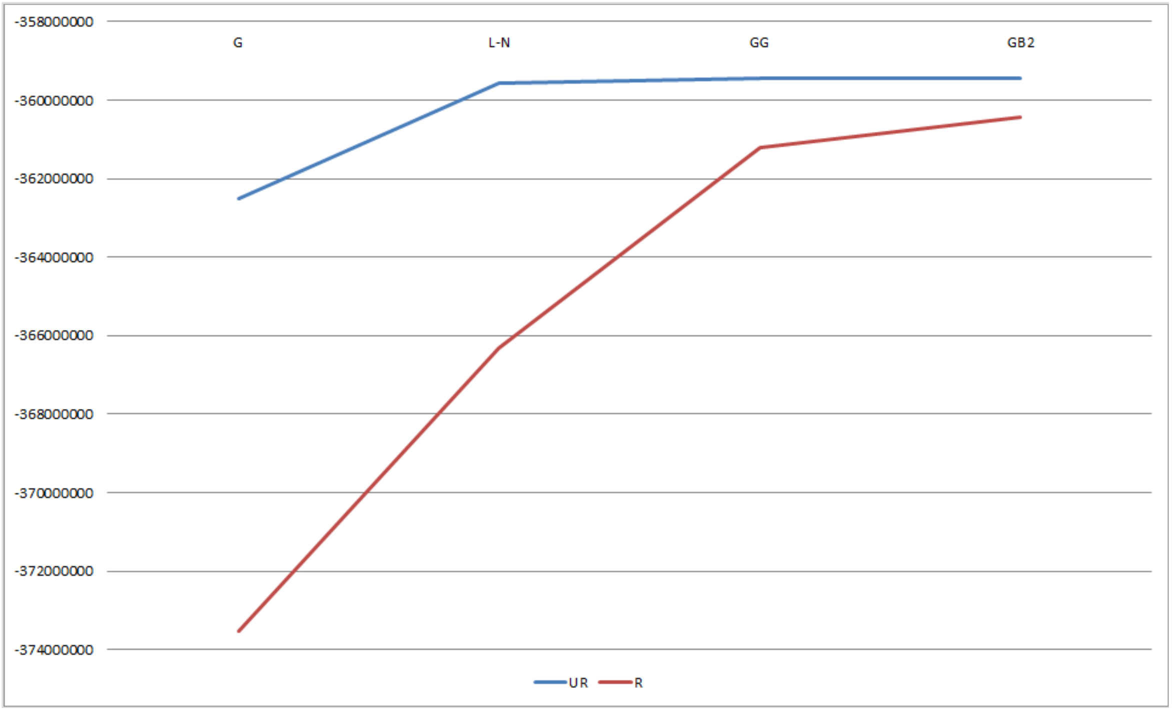
Code	Concept	S.11 Non-financial corporations	S.12 Financial corporations	S.13 General government	**S.14** **Households**	S.15 Non-profit institutions serving households (ISFLSH)	S.1 Total	S.2 Rest of the World
B.1b	Gross Value Added	9,792,275	517,267	1,378,228	**3,231,073**	159,432	15,561,472	
B.2b	Gross operating surplus	7,822,128	360,718	7,119	**1,278,929**	58,116	9,527,010	0
B.3b	Gross mixed income				**1,268,856**		1,268,856	
B.6b	Gross disposable income	561,397	873,788	2,175,939	**11,696,850**	273,487	15,581,461	0
B.8b	Gross savings	561,397	463,440	359,335	**1,796,991**	65,133	3,246,296	152,153
B.9	Net lending (+) / net indebtedness (-)	-1,016,405	442,754	-91,466	**498,721**	14,243	-152,153	152,153

Table 3. Summary maximum pseudo log-likelihood fits of 4 types of distributions to ENIGH 2012 data.

		Fitted Distributions			
		Gamma (G)	**Log-normal (L-N)**	**Generalized Gama (GG)**	**Generalized Beta, Type II (GB2)**
		2 Pars.	**2 Pars.**	**3 Pars.**	**4 Pars.**
OPTIMAL VALUE OF UN-RESTRICTED PSEUDO LOG-LIKELIHOOD		-362,503,029	-359,553,923	-359,439,551	-359,431,519
GINI COEFFICIENT		0.41665	0.43182	0.447752	0.449056
		FITTED PARAMETERS			
UNRESTRICTED FIT.	E(X|θ)	38155.08	37172.35	37876.76	38073.03
	μ	38155.08	10.197522	24789.4844	20394.4242
	σ	0.798911	0.8072155	0.79909	1.198802
	ν			-0.2469259	3.008914
	τ				2.2948
OPTIMAL VALUE OF RESTRICTED PSEUDO LOG-LIKELIHOOD		-373513661	-366,300,000	-361,204,049	-360,433,838
GINI COEFFICIENT		0.520955	0.624607	0.781296	0.802759
		FITTED PARAMETERS			
RESTRICTED FIT. CONSTRAINT: E[X|**θ**]=86,410.57	E(X|θ)	92733.62	92733.62	92733.62	92733.62
	μ	92733.62	10.6516	19720.8393	17175.97
	σ	1.0554	1.2563	0.85696	3.25349
	ν			-1.08153	0.7905
	τ				0.36741

Source: Own calculations from ENIGH, 2012.

Figure 1. Unrestricted (UR) and Restricted (R) Pseudo Log-likelihood Optima for 4 distributions.



Source: Own calculations from ENIGH 2012 and SCNM

Table 4a. Percentiles of distributions fitted by unconstrained maximum pseudo log-likelihood.

				Fitted Distributions			
				Gamma (G)	Log-normal (L-N)	Generalized Gamma (GG)	Generalized Beta Type 2 (GB2)
				2 Pars.	2 Pars.	3 Pars.	4 Pars.
%	Sample (Expanded Data)			Percentiles from fit			
	Percentile	Lower CL	Upper CL				
0.10	2,639.99	2,062.04	2,896.80	1.25	2,215.07	2,744.98	2,009.12
0.50	3,674.45	3,220.91	4,085.41	12.59	3,355.22	3,879.15	3,303.63
1.00	4,473.03	4,128.00	4,735.81	34.10	4,103.76	4,605.47	4,142.13
2.00	5,416.27	5,091.37	5,732.98	92.44	5,113.83	5,572.17	5,252.88
5.00	7,286.62	6,816.13	7,725.45	346.70	7,113.65	7,460.79	7,384.25
10.00	9,798.45	9,201.94	10,333.49	950.21	9,537.95	9,730.96	9,870.27
20.00	13,795.27	13,133.99	14,531.68	2,661.90	13,604.45	13,534.97	13,879.28
30.00	17,704.09	16,867.73	18,344.26	4,989.01	17,574.74	17,271.22	17,677.82
40.00	21,571.42	20,630.26	22,660.21	7,989.71	21,873.18	21,356.09	21,724.91
50.00	26,117.90	25,299.14	27,206.53	11,827.83	26,836.60	26,132.95	26,366.04
60.00	32,002.08	30,669.51	33,442.48	16,817.78	32,926.32	32,087.26	32,068.41
70.00	39,441.09	37,971.37	41,293.72	23,578.51	40,979.46	40,122.30	39,694.16
80.00	50,957.22	48,686.52	53,267.62	33,530.32	52,938.79	52,388.41	51,317.73
90.00	75,092.46	69,924.69	81,411.27	51,265.57	75,509.24	76,574.14	74,569.96
95.00	105,637.41	99,742.69	114,505.58	69,567.03	101,242.44	105,686.35	103,580.19
98.00	153,168.07	139,457.98	177,640.20	94,295.86	140,834.39	153,391.08	154,064.70
99.00	207,648.04	183,716.35	248,933.74	113,276.26	175,498.55	197,816.72	204,556.40
99.50	267,552.62	233,894.88	402,277.30	132,426.13	214,651.56	250,751.32	269,114.35
99.90	466,728.31	415,389.61	792,894.38	177,353.73	325,138.14	414,438.70	498,098.88

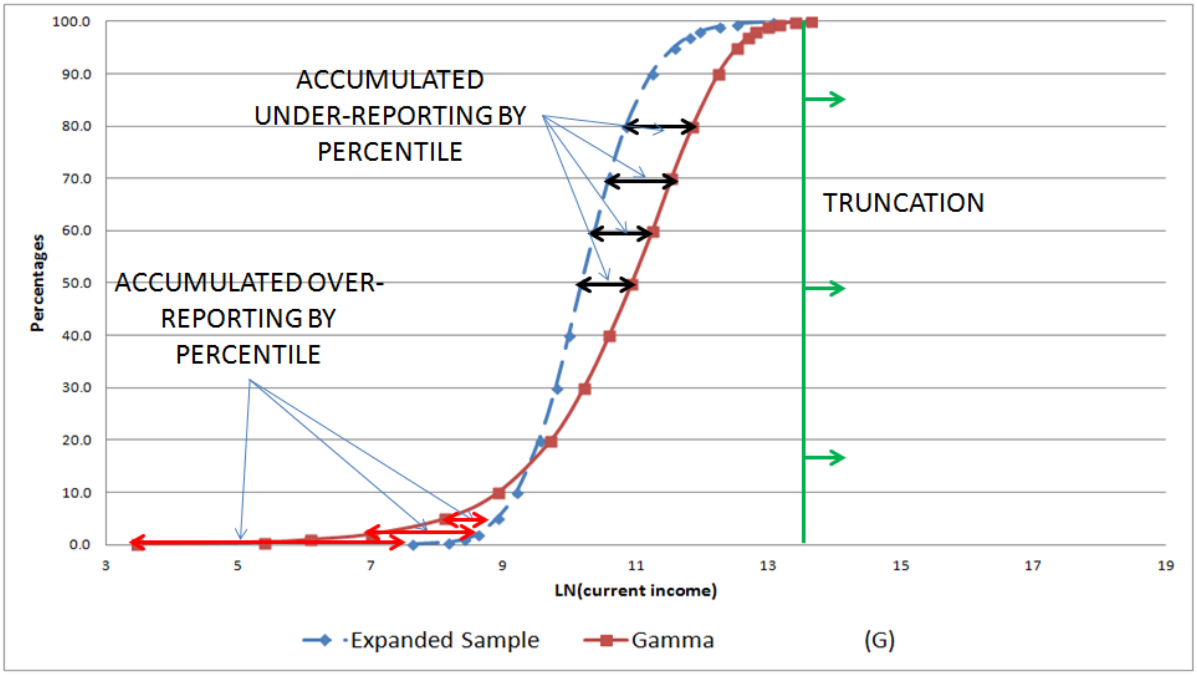
Source: Own calculations from ENIGH, 2012.

Table 4b. Percentiles of distributions fitted by constrained maximum pseudo log-likelihood.

				Fitted Distributions			
				Gamma (G)	Log-normal (L-N)	Generalized Gamma (GG)	Generalized Beta Type 2 (GB2)
				2 Pars.	2 Pars.	3 Pars.	4 Pars.
%	Sample (Expanded Data)			Percentiles from fit			
	Percentile	Lower CL	Upper CL				
0.10	2,639.99	2,062.04	2,896.80	45.01	878.00	3,598.49	1,650.32
0.50	3,674.45	3,220.91	4,085.41	270.62	1,673.25	4,552.64	3,087.55
1.00	4,473.03	4,128.00	4,735.81	586.64	2,287.65	5,152.43	4,046.79
2.00	5,416.27	5,091.37	5,732.98	1,274.10	3,219.63	5,947.37	5,311.92
5.00	7,286.62	6,816.13	7,725.45	3,577.19	5,375.62	7,511.73	7,658.46
10.00	9,798.45	9,201.94	10,333.49	7,913.40	8,476.72	9,442.45	10,232.39
20.00	13,795.27	13,133.99	14,531.68	18,010.90	14,714.56	12,859.19	14,118.25
30.00	17,704.09	16,867.73	18,344.26	30,000.18	21,900.67	16,481.87	17,688.45
40.00	21,571.42	20,630.26	22,660.21	44,237.03	30,763.24	20,781.50	21,556.13
50.00	26,117.90	25,299.14	27,206.53	61,418.35	42,263.60	26,294.66	26,236.03
60.00	32,002.08	30,669.51	33,442.48	82,779.70	58,063.19	33,950.42	32,520.72
70.00	39,441.09	37,971.37	41,293.72	110,682.19	81,559.67	45,760.10	42,082.28
80.00	50,957.22	48,686.52	53,267.62	150,468.41	121,390.80	67,352.35	59,624.73
90.00	75,092.46	69,924.69	81,411.27	219,258.84	210,719.68	124,224.35	106,883.80
95.00	105,637.41	99,742.69	114,505.58	288,653.63	332,280.07	222,248.37	190,979.66
98.00	153,168.07	139,457.98	177,640.20	380,960.40	554,787.48	469,410.97	411,076.61
99.00	207,648.04	183,716.35	248,933.74	451,080.97	780,805.21	819,994.93	734,092.90
99.50	267,552.62	233,894.88	402,277.30	521,383.81	1,067,511.23	1,427,687.13	1,310,921.30
99.90	466,728.31	415,389.61	792,894.38	685,122.45	2,034,409.21	5,144,885.33	5,038,445.08

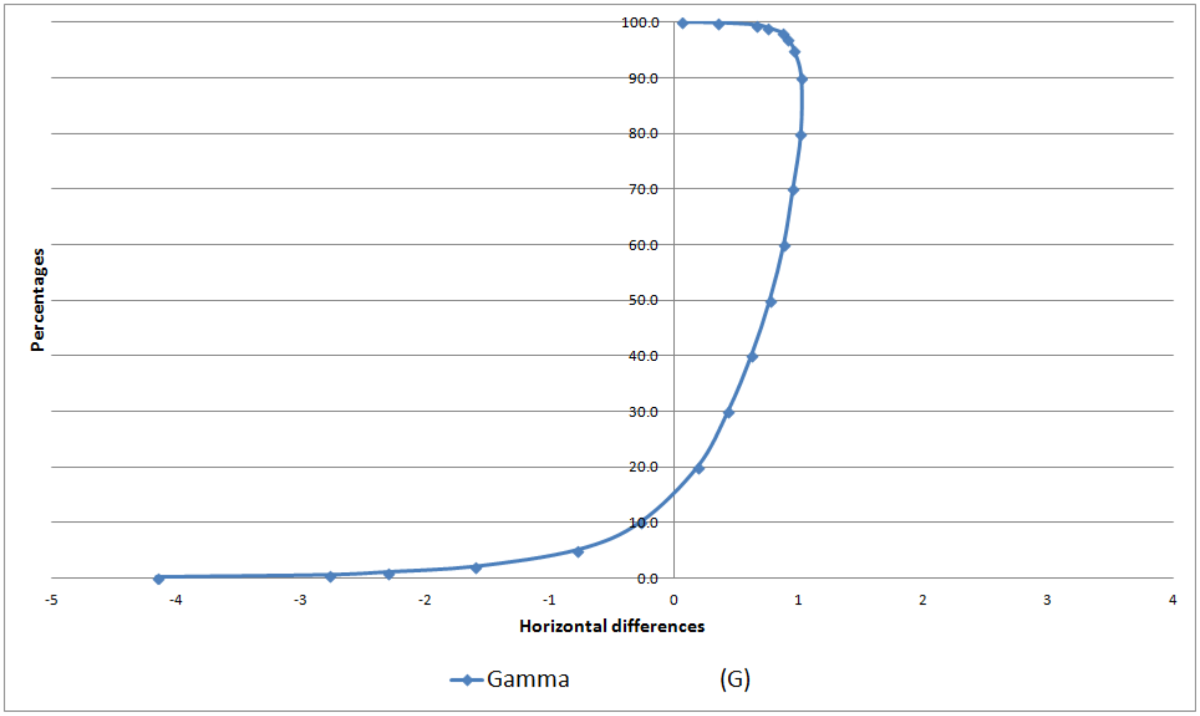
Source: Own calculations from ENIGH, 2012.

Figure 2. Interpretation of discrepancies between graphic representations of the empirical and the constrained maximum-likelihood fitted distributions.



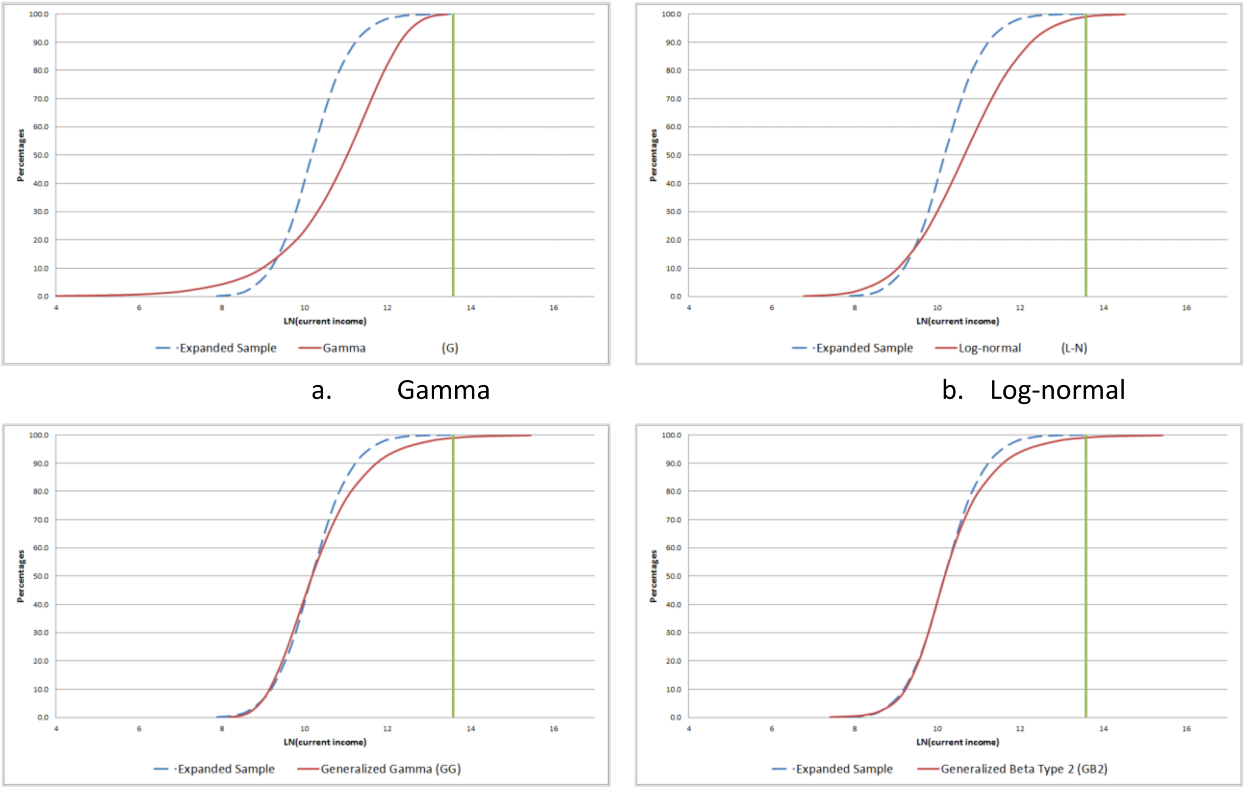
Note: horizontal axis in logarithmic scale.

Figure 3. Horizontal differences between logarithms of empirical and fitted percentiles.



Source: Own calculations from ENIGH, 2012.
Note: Horizontal axis in logarithmic scale.

Figure 4. Empirical and fitted distributions



a. Gamma

b. Log-normal

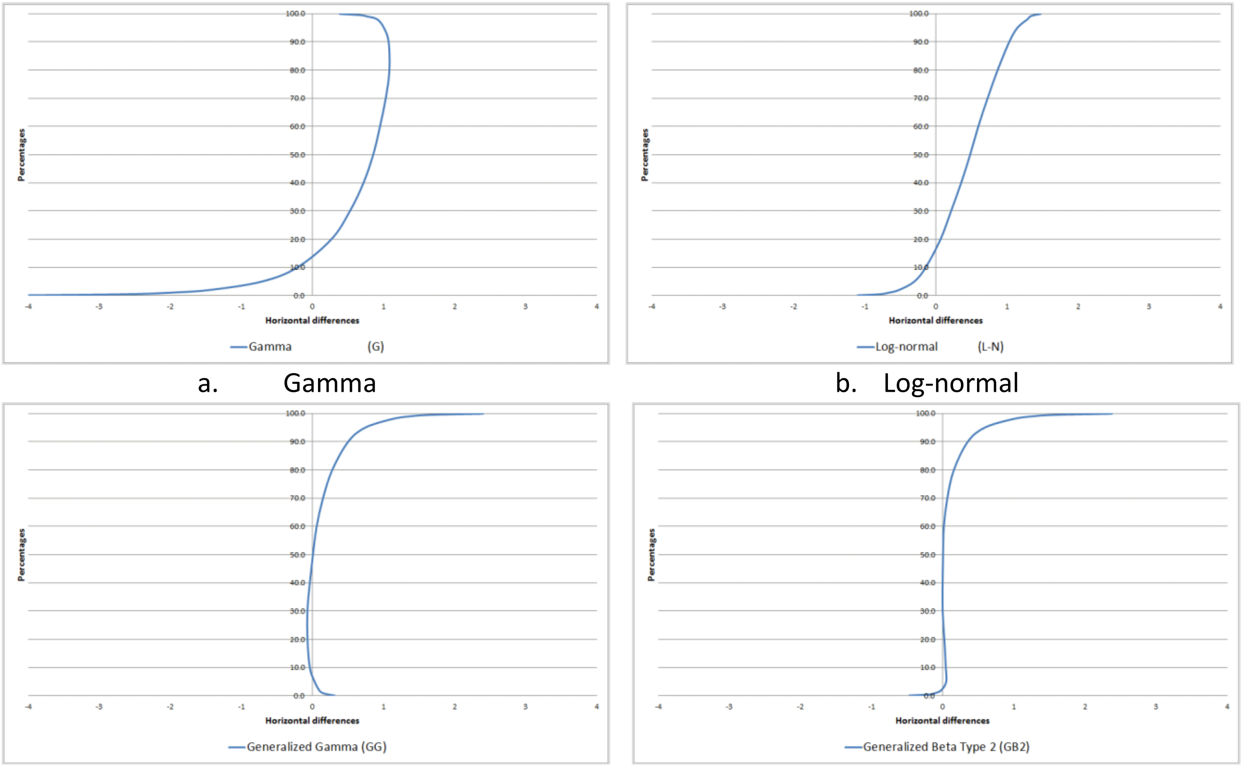
c. Generalized gamma

d. Generalized Beta, type 2

Source: Own calculations from ENIGH, 2012.

Note: Horizontal axis in logarithmic scale.

Figure 5. Horizontal differences between empirical and fitted distributions



a. Gamma

b. Log-normal

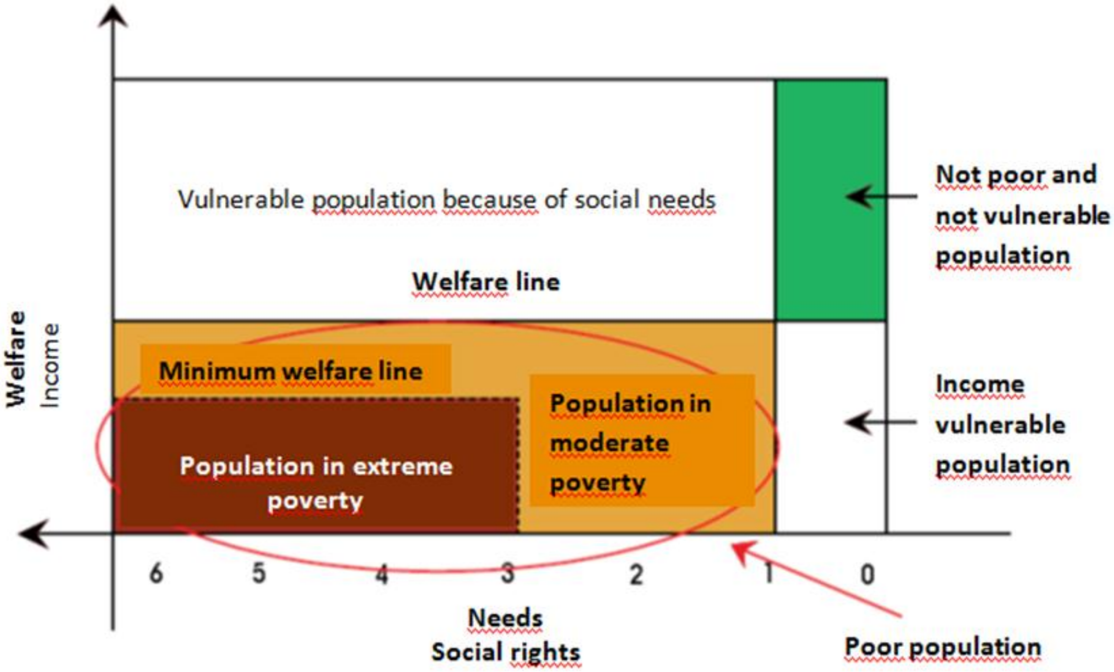
c. Generalized gamma

d. Generalized Beta, type 2

Source: Own calculations from ENIGH, 2012.

Note: Horizontal axis in logarithmic scale.

Figure 6. Poverty multidimensional measurement in México.


Vulnerable population because of social needs
Welfare line
Not poor and not vulnerable population
Welfare
Income
Minimum welfare line
Population in extreme poverty
Population in moderate poverty
Income vulnerable population
6
5
4
3
2
1
0
Needs
Social rights
Poor population

I

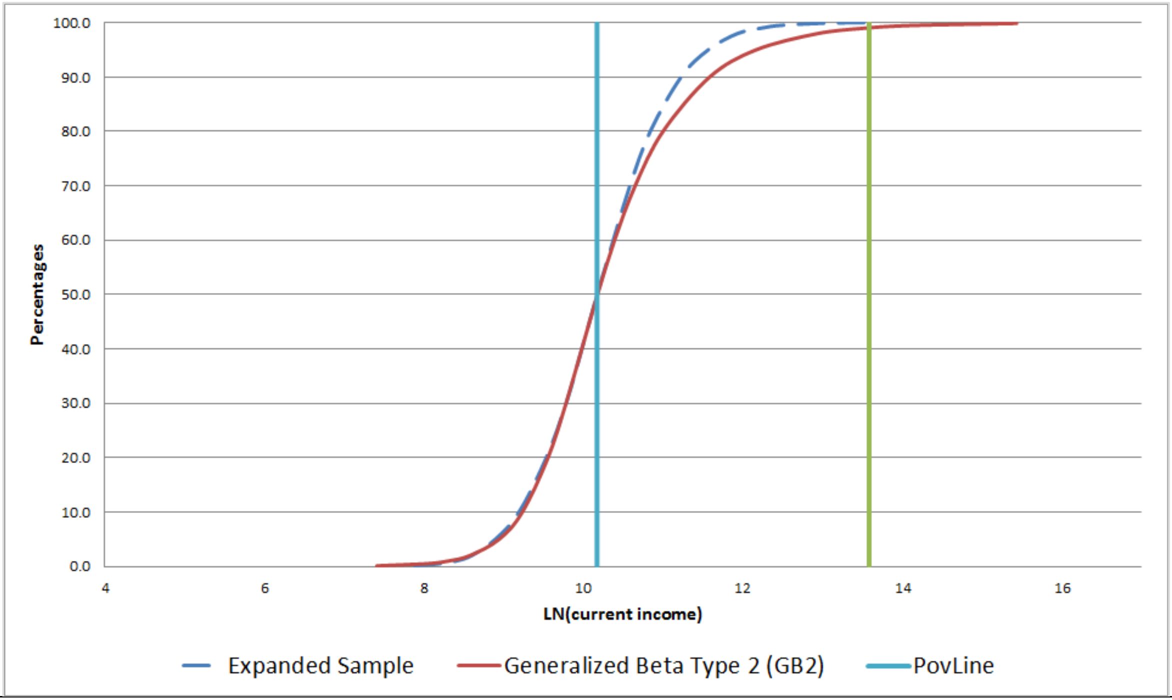
Source: CONEVAL, Construcción de las Líneas de Bienestar, Documento Metodológico, Metodología para la medición multidimensional de la pobreza, México, 2012.

Table 5. Percentages of people and households whose income lie below each welfare line, 2012.

People(1)			**Households(2)**	
Welfare line	Minimum welfare line		Welfare line	Minimum welfare line
51.6%	20.0%		40-50%	10-20%

Source: (1) CONEVAL and (2) own estimations, both based on ENIGH, 2012.

Figure 7. Empirical and fitted GB2 distributions and poverty line.



Source: Own calculations from ENIGH, 2012.
Note: Horizontal axis in logarithmic scale.

TECHNICAL APPENDIX

Context of finite population sampling

In finite population sampling it is common to begin from the concept of a superpopulation Ω of infinite size. The finite population, a collection of units $U = \{u_1, u_2, ..., u_N\}$, is conceptualized as a random sample of size N obtained from Ω. Thus, if the random vector $\underline{Y}$ is distributed according to a superpopulation distribution with density function $f(\underline{Y} \mid \underline{\theta})$, the total log-likelihood for population U may be written as in (A.1).

$$L(\underline{\theta}; \underline{Y}_1, \ldots, \underline{Y}_N) = Ln\left(\prod_{i=1}^{N} f(\underline{Y}_i; \underline{\theta})\right) = \sum_{i=1}^{N} Ln(f(\underline{Y}_i; \underline{\theta})) = \sum_{i=1}^{N} \ell(\underline{\theta}; \underline{Y}_i) \cdot \tag{A.1}$$

In (A.1) it is assumed that the value of vector $\underline{Y}$ is known for each and every unit in U as would be the case in a census.

Sampling:

For those cases in which total population enumeration is not an option, the random selection of a subset of the population units may become the only viable alternative to study some characteristics of vector $\underline{Y}$. The selection mechanism is partially[10] characterized by a N-dimensional vector of binary random variables $\underline{I} = (I_1, I_2, ..., I_N)$ and by the vector $\underline{\pi} = (\pi_1, ..., \pi_N)$, whose entries are known as inclusion probabilities. Letting the random event $\{I_j = 1\}$ indicate inclusion of the j-th population unit in the sample, both vectors are related by $\pi_j = P\{I_j = 1\} = E\lfloor I_j \rfloor,\ j = 1, \ldots, N$.

[10] Since second and higher order moments of vector $\underline{I}$ are not given.

When the sample design provides for more than one selection stage, vector $\underline{I}$ is obtained as the composition of similar vectors each associated with, respectively, primary sampling units (PSU), second sampling units (SSU), tertiary sampling units (TSU), and so on, as shown in (A.2).

$$\underline{I} = \underline{I}_P \otimes \underline{I}_{S|P} \otimes \cdots \otimes \underline{I}_{K|P,S,\ldots}. \tag{A.2}$$

We introduce a slight abuse of notation for the Kroenecker product since units at any stage do not always contain the same number of subunits; however, given that we are dealing with vectors this should not limit its application to the “telescopic” view of the population when multistage sampling is envisioned. When the sample design includes selection of stratified units at some stage, an additional non-random vector, whose entries are all equal to one since each stratum is visited with certainty, is included where appropriate. So to speak, a "census" of strata is performed since units are selected from each and every stratum. Something similar occurs when a census of sub-units takes place within units selected at the previous stage; in this case, the vectors associated with the (censed) subunits have all their entries equal to one with probability one. We shall refer to them as “certainty vectors”.

It is not difficult to see that, under the above conditions, vector $\underline{\pi}$ can also be iteratively calculated as indicated in (A.3.1).

$$\underline{\pi} = \underline{\pi}_P \otimes \underline{\pi}_{S|P} \otimes \cdots \otimes \underline{\pi}_{K|P,S,\ldots}. \tag{A.3.1}$$

Covariance matrices may be similarly obtained as shown in (A.3.2).

$$Cov(\underline{I}) = Cov(\underline{I}_P) \otimes \cdots \otimes Cov(\underline{I}_{K|P,S,\ldots}). \tag{A.3.2}$$

Whenever a certainty vector appears, the corresponding matrix is replaced by an identity of appropriate dimensions.

Some relevant properties

In those cases where the sample size is fixed beforehand, the set of binary random variables satisfies non-stochastic condition (A.4).

$$n = \sum_{j=1}^{N} I_j \ ; \qquad \text{(A.4)}$$

Consequently, the inclusion probabilities satisfy expression (A.5). In other words, inclusion probabilities are not fixed but vary from one survey to the next when sample sizes do not match.

$$n = E\left[\sum_{j=1}^{N} I_j\right] = \sum_{j=1}^{N} E\left[I_j\right] = \sum_{j=1}^{N} \pi_j \ ; \qquad \text{(A.5)}$$

Multiplicative inverses of inclusion probabilities, usually known as expansion factors, are essential to Horvitz-Thompson (HT) estimation of totals, as pointed out below. Note that the expansion factors satisfy another important relationship according to (A.6).

$$E\left[\sum_{j=1}^{N} \frac{I_j}{\pi_j}\right] = \sum_{j=1}^{N} \frac{E\left[I_j\right]}{\pi_j} = \sum_{j=1}^{N} \frac{\pi_j}{\pi_j} = N \ ; \qquad \text{(A.6)}$$

In other words, on average, their sum over the sample equals the size of population U. Therefore it is common to adjust the expansion factors to satisfy this condition; whenever necessary, N is obtained from projections of population size.

With respect to the second moments of the set of indicator variables, it is easy to show that their values may be obtained from (A.7).

$$Cov(I_i,I_j)=\begin{cases}\pi_i\left(1-\pi_i\right), i=j\\ \pi_{ij}-\pi_i\pi_j, i\neq j\end{cases}. \tag{A.7}$$

Coefficients π_{ij} in (A.7) are known as second order inclusion probabilities and refer to the simultaneous occurrence of two specific units in the sample.

Horvitz and Thompson (1952) proposed an estimator for the sum of the values of a variable along the entire population U. Their proposal is given in (A.8).

$$\hat{Y}=\sum_{i=1}^{n}\frac{Y_{(i)}}{\pi_{(i)}}=\sum_{i=1}^{N}\frac{I_iY_i}{\pi_i}. \tag{A.8}$$

In (A.8), we have written the subscripts of sample units within parentheses to distinguish them from their population counterparts. The last equality helps prove unbiasedness of this estimator when $\left\{I_j, j=1,\ldots,N\right\}$ is independently distributed from $\left\{Y_j, j=1,\ldots,N\right\}$.

Consequently, since the H-T estimator represents a weighted sum of all $I_j, j=1,\ldots,N$ entries, the expression for its variance is presented in (A.9).

$$Var\left(\hat{Y}\right)=\sum_{i=1}^{N}\left\{\left(\frac{Y_i}{\pi_i}\right)^2\pi_i\left(1-\pi_i\right)+\sum_{j\neq i}\frac{Y_iY_j}{\pi_i\pi_j}\left(\pi_{ij}-\pi_i\pi_j\right)\right\}. \tag{A.9}$$

When second order inclusion probabilities are known, an unbiased estimator of this variance can be directly obtained, as in (A.10).

$$Var\left(\hat{Y}\right)\mathrel{\hat{=}}\sum_{i=1}^{N}\left\{\frac{I_i}{\pi_i}\left(\frac{Y_i}{\pi_i}\right)^2\pi_i\left(1-\pi_i\right)+\sum_{j\neq i}\frac{I_iI_j}{\pi_{ij}}\frac{Y_iY_j}{\pi_i\pi_j}\left(\pi_{ij}-\pi_i\pi_j\right)\right\}. \tag{A.10}$$

Pseudo log-likelihood:

In order to carry out the fitting of alternative functional forms for the income distribution to data from the sample, we propose to use the approximation to the log-likelihood function of the sample presented in (A.11).

$$
\begin{aligned}
&L\left(\underline{\theta},\underline{\nu};\underline{Y}_{(1)},\underline{Y}_{(2)},\ldots,\underline{Y}_{(n)};I_1,I_2,\ldots,I_N\right)\approx Ln\left(\prod_{i=1}^{N} f^{\frac{I_i}{\pi_i}}\left(\underline{Y}_i;\underline{\theta}\right)\right)\\
&=\sum_{i=1}^{N}\frac{I_i}{\pi_i}Ln\left(f\left(\underline{Y}_i;\underline{\theta}\right)\right)=\sum_{i=1}^{N}\frac{I_i}{\pi_i}\ell\left(\underline{\theta};\underline{Y}_i\right)=\sum_{i=1}^{n}\frac{1}{\pi_{(i)}}\ell\left(\underline{\theta};\underline{Y}_{(i)}\right)
\end{aligned}
\tag{A.11}
$$

Expressions similar to that found immediately to the right of the approximation sign can be found in other contexts. For instance, in the maximum likelihood classification of units; in such case the exponent of each factor does not appear expanded and the binary variable indicates whether or not the unit belongs to a particular class. On the other hand the last two terms in the above expression make explicit that, given a value for $\underline{\theta}$, the pseudo log-likelihood (A.11) becomes an H-T estimator for the population log-likelihood (A.1), for the same value of the parameter vector. In other words, we have an H-T estimator of a function of the parameter vector $\underline{\theta}$. As pointed out above, once the value of the parameter vector that yields the instance of each model that best fits the data is found, it also would be possible to estimate the variance of the optimal value of the pseudo log-likelihood.

The Constrained Pseudo Log-likelihood criterion.

The pseudo log-likelihood maximization takes into account only the available data and the sample design that gave rise to them but is still short of achieving the purpose of reconciling the result with other information sources. To achieve the latter purpose, we consider additional information in the form of constraints on the values that one or more, possibly non-linear, functions of the parameters may take. For example, if among other constraints, we want the average of the fitted distribution to match a specific value from an alternative source, we express the former as a function of the parameters $E(\underline{Y} \mid \underline{\theta})$ and force it to assume a particular value $\underline{c}$, from the alternative source; in other words, we ask that $E(Y \mid \underline{\theta}) \equiv c$. In this fashion, parameter estimates are obtained from the solution of the optimization problem posed in (A.12).

$$\underset{\underline{\theta},\underline{\lambda}}{Max}\left\{\sum_{i=1}^{n} \frac{1}{\pi_{(i)}} \ell\left(\underline{\theta}; \underline{Y}_{(i)}\right) - \underline{\lambda}'\left[\underline{h}(\underline{\theta}) - \underline{c}\right]\right\} \tag{A.12}$$

where

$\ell\left(\underline{\theta}; \underline{Y}_{(i)}\right)$ represents the natural logarithm of the density function, evaluated at the i-th sample value;

$\pi_{(i)}$, inclusion probability for that sample unit;

$\underline{h}(\underline{\theta})$, one or more functions of the parameter vector whose values are to coincide with those in vector $\underline{c}$. Note that the dimension of $\underline{h}(\underline{\theta})$ cannot equal or exceed that of $\underline{\theta}$ because the survey data would become irrelevant.

$\underline{\boldsymbol{\lambda}}$, vector of so called Lagrange multipliers.

Distributions used in the numerical example

Instrumentations and settings contained in Stasinopoulos, et al. [10] were used for the following distributions in the numerical examples.

a) Gamma Distribution (G):

$$f_Y(y \mid \mu,\sigma) = \frac{1}{\left(\sigma^2\mu\right)^{1/\sigma^2}} \frac{y^{\sigma^{-2}-1} e^{-y/(\sigma^2\mu)}}{\Gamma\left(1/\sigma^2\right)}, y > 0;$$

where $\mu > 0$ and $\sigma > 0$.

In this case,

$$E[Y] = \mu$$

and $$Var[Y] = \sigma^2\mu^2.$$

b) Log-normal Distribution (L-N):

$$f_Y(y \mid \mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \frac{1}{y} \exp - \left\{ \frac{(\log(y) - \mu)^2}{2\sigma^2} \right\}, y > 0;$$

Where $\mu > 0$ and $\sigma > 0$.

$$E[Y] = \omega^{1/2} e^{\mu}$$

and $$Var[Y] = \omega(1-\omega)e^{2\mu}$$ with $\omega = \exp(\sigma^2)$.

c) Generalized Gamma Distribution (GG):

$$f_Y(y \mid \mu,\sigma,\nu) = \frac{|\nu|\theta^{\theta} z^{\theta} \exp\{-\theta z\}}{\Gamma(\theta) y}, y > 0;$$

Where $\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$ and where also $z = \left(y/\mu\right)^{\nu}$ and $\theta = (\sigma\nu)^{-2}$.

$$E\left[Y\right]=\frac{\mu\Gamma\left(\theta+{}^{1}\!/_{\nu}\right)}{\left[\theta^{1/\nu}\Gamma\left(\theta\right)\right]},$$

and

$$Var\left[Y\right]=\mu^{2}\ \frac{\Gamma\left(\theta\right)\Gamma\left(\theta+{}^{2}\!/_{\nu}\right)-\Gamma^{2}\left(\theta+{}^{1}\!/_{\nu}\right)}{\theta^{2/\nu}\Gamma^{2}\left(\theta\right)}.$$

d) Generalized Beta, Type 2, Distribution (GB2):

$$f_{Y}\left(y\mid\mu,\sigma,\nu,\tau\right)=\left|\sigma\right|\left(\frac{y}{\mu}\right)^{\sigma\nu-1}\left\{\mu B\left(\nu,\tau\right)\left[1+\left(\frac{y}{\mu}\right)^{\sigma}\right]^{\nu+\tau}\right\}^{-1},y>0;$$

where $\mu>0,-\infty<\sigma<\infty,\ \nu>0$ and $\tau>0$.

$$E\left[Y\right]=\frac{\mu^{2}B\left(\nu+\frac{1}{\sigma},\tau-\frac{1}{\sigma}\right)}{B\left(\nu,\tau\right)},-\nu<\frac{1}{\sigma}<\tau\ ,$$

and

$$Var\left[Y^{2}\right]=\frac{\mu^{2}B\left(\nu+\frac{2}{\sigma},\tau-\frac{2}{\sigma}\right)}{B\left(\nu,\tau\right)},-\nu<\frac{2}{\sigma}<\tau.$$

Some relationships between these models have been studied in, for example, Bandourian, et al. [1]. According to these authors' parameterization, the 2-parameter models considered are special cases of the 3-parameter one and this, in turn, a special case of the 4-parameter distribution.

Truncation effect

The contribution of the truncation effect to total quarterly current household income is obtained as in (A.13).

$$\int_{Max_{i=\overline{1,n}}\left\{Y_{(n)}\right\}}^{\infty} xf\,(x;\underline{\hat{\theta}})dx \Bigg/ \int_{0}^{\infty} xf\,(x;\underline{\hat{\theta}})dx \qquad \text{(A.13)}$$

In turn, this effect may also be presented as a proportion of the difference between sources, as in (A.14).

$$\int_{Max_{i=\overline{1,n}}\left\{Y_{(n)}\right\}}^{\infty} xf\,(x;\underline{\hat{\theta}})dx \Bigg/ \left(\int_{0}^{\infty} xf\,(x;\underline{\hat{\theta}})dx - \frac{\hat{Y}}{N}\right) \qquad \text{(A.14)}$$