Income Interpolation from Categories Using a Percentile-Constrained Inverse-CDF Approach

G. Lance Couzens[1], Marcus E. Berzofsky[1], Kimberly C. Peterson[1]
[1]RTI International, 3040 E Cornwallis Rd, Durham, NC 27709

Abstract

It is often the case that surveys of persons and households collect income data along with other demographic and socioeconomic questions. When income level is not the primary focus of the survey, it may be used in domain estimation or as a covariate in multivariable analyses. In these instances, it is common practice for income to be collected in a categorical form, with nonstandard category boundaries that vary from one survey to another. Though these categories may be appropriate for their originally-intended purposes, they often are not ideal for analyses not considered when the survey instrument was developed (e.g., for determining a household's percent of the federal poverty level). This paper describes a method for estimating a continuous income measure based on observed categorical responses with arbitrary category boundaries. The authors present this method in general terms and provide validation results both from simulation and comparison with federal benchmark surveys.

Key Words: Income, Interpolation, Distribution Estimation, National Crime Victimization Survey (NCVS), Current Population Survey (CPS)

1. Introduction

Surveys of persons and households are implemented to answer any number of research questions and to track populations over time. Though the primary objectives of any two surveys may differ widely, they will typically feature a battery of questions relating to demographic, geographic, and socioeconomic characteristics. Responses to these questions can be used to define estimation domains, to compare or calibrate samples, and as controls for potentially confounding effects in multivariable analyses. While many of these core variables (e.g., gender, race, state, etc.) may be compared from one survey to another with little or no adjustment – through collapsing of categories, for example – income often presents a more difficult challenge. This is because household or personal income is in many cases asked in the form of categories, often as a tool to mitigate against nonresponse. The categorical nature of these questions can prove problematic, however, because the values used to define category boundaries are nonstandard and can vary from one questionnaire to another. This complicates comparisons between surveys as well as analyses that may require or benefit from a continuous income measure. In this paper the authors describe a method for the estimation of continuous income distributions based on existing categories as well as for interpolation between category boundaries according to estimated underlying continuous distributions.

1.1 Purpose

Ideally, in every circumstance, income responses would be provided in actual dollar amounts. Even if categories are required for a particular research purpose, it is preferable for the researcher to form the categories himself according to his own requirements. This is very often not the case, however, and researchers must deal with income data with non-ideal and predetermined category boundaries (e.g., the National Crime Victimization Survey (NCVS; Truman & Langton, 2014), or, in some cases, even with income data that is mixed type – both continuous and categorical (e.g., the Ohio Medicaid Assessment Survey (OMAS; Ohio Medicaid Assessment Survey, 2015). The latter scenario is commonly encountered when survey instruments are designed to provide the opportunity for respondents to provide income ranges after initially refusing a specific dollar amount question. Regardless of the motivation for initial collection of income in categories, in many cases continuous values are required. This poses a methodological question regarding the manner in which data users should convert categorical responses to actual dollar values.

In practice, data users have some options in how to interpolate categorical income responses, though there is no clear guidance in the survey literature as to how this should be achieved. The obvious choice, and by far the simplest to implement is linear interpolation. Linear interpolation is simply a matter of randomly selecting a dollar value between a respondent's category boundaries. This approach is attractive for ease-of-implementation but requires a very strong and likely false assumption about the underlying continuous distribution. Specifically, the researcher is assuming that every value in a given category is equally probable. For narrow and relatively central categories, an equal-probability assumption may not deviate very far from reality. For non-central categories, or for categories that are especially wide, however, it's much less safe to assume linearity. For example, it may not be unreasonable to assume that a respondent indicating an income value in the range of $30,000 to $35,000 was just as likely to have a true value of $30,001 as she was $34,999. It is much less reasonable to assume that a respondent indicating an income value in the range of $0 to $10,000 is equally likely to have a true value of $1 as he is $9,999. In addition to the potential for erroneously inflating lower income values through the use of linear interpolation, there is also the issue of how to address the highest category. Due to the nature of income, it is inevitable that the highest category will be unbounded to the right. With a linear approach it is impossible to interpolate individuals in the highest category without imposing an artificial upper boundary, and doing so would introduce a similar problem encountered when interpolating the lowest category.

An alternative to an individual respondent-based linear approach is to fit a function to the cumulative densities observed at the category boundaries, and use that function to interpolate individual respondents. Using a purely empirical approach that makes no assumptions regarding the nature of the underlying continuous income distribution, one could attempt to employ polynomial or spline interpolation based on the observed densities. This approach is unappealing in its basic form, however, in that it either precludes implementation in the highest category or makes potentially naïve assumptions about the behavior of the population in that category based solely on observed densities in lower categories. Dikhanov and Ward (2001) overcame this limitation by using a so-called quasi-exact rendering technique based on the use of fourth-order polynomials to interpolate categorical income data with the lowest and highest groups being forced to be lognormal.

The mixed polynomial and lognormal method used by Dikhanov and Ward is appealing in certain contexts in that the fitted functional form of the distribution is exact at category

boundaries. In the context of a survey sample, however, this implies that sample-based quantiles are accepted without regard for potential sample variation. The authors instead seek to determine which single lognormal distribution is implied by the sample without requiring exact equality at observed boundaries. Doing so does not preclude consistency between interpolated values and observed category boundaries, though it implies that boundary percentile values must be allowed to deviate between the sample and the population (the level of deviation is minimized by the algorithm used to estimate the distribution). To this end, this paper describes a method for using observed boundary densities to estimate a lognormal distribution which may then be used to interpolate income categories to continuous values that are consistent with the observed category definitions. This method is not computationally intense1 and can be easily implemented with basic software.

2. Methods

The following sections detail a process for estimating a lognormal income distribution based on empirical cumulative mass at category boundaries and for drawing random variates for individual respondents from that distribution that are consistent with reported income categories. The validity of this method as presented here depends on the assumption that income is lognormally-distributed. The literature shows that this assumption is reasonable – Pinkovskiy and Sala-i-Martin (2009) in particular provides a good overview of previous research efforts to validate lognormality of income, and the authors themselves show that the lognormal distribution provides superior fit to other common parametric alternatives.

2.1 Empirical Cumulative Mass at Boundary Points as Proxy for Lognormal Percentiles

In order to estimate lognormal parameters based on a sample of categorical responses, it is first necessary to make an assumption about the nature of the categorical responses and how they relate to the underlying continuous distribution. Specifically, we assume that we are observing individuals within an ordinal classification of income and that the cumulative mass of observations at a category boundary (dollar value) is equivalent to the boundary point's percentile value that would have been observed had the data been collected on a continuous scale. For example, if we have a five-level income variable and 63% of individuals indicated income values less than or in the third category ($35,000-$50,000), we are assuming that $50,000 is the 63rd percentile of the true lognormal distribution for our sample.

2.2 Minimization of Percentile Vector Distance for Estimation of Lognormal Parameters

To estimate the true underlying lognormal distribution, we choose a simple and computationally efficient algorithm based on grid-searching over a reasonable parameter space. The search grid is defined in one dimension by potential log-mean values at a specified granularity, while the other dimension is similarly defined by potential log-standard deviation values. In practice, it is important to acknowledge that a single distribution may not best represent all individuals in a given survey's sample. The

[1] The basic method as presented is based on a lognormal assumption – deviation from this (e.g., use of a mixture distribution with unknown quantile function) is possible through extension based on simulation at the loss of simplicity and computational efficiency.

algorithm is therefore implemented across strata defined by one or more characteristics associated with income (e.g., age group, educational attainment, etc.). Our notation is defined as follows:

I = The number of income categories

u_i = Upper bound (in dollars) for the i^{th} income group, with $i < I$

d_{hi} = Observed proportion of stratum h households in income group i

c_{hi} = Cumulative density at boundary point u_i for stratum h

$= \sum_{j=1}^{i} d_{hj}$

$\vec{c}_h$ = The vector of values c_{hi}

m_{min} = Minimum potential log-mean value for candidate lognormal distributions

m_{max} = Maximum potential log-mean value for candidate lognormal distributions

s_{min} = Minimum potential log-standard deviation value for candidate lognormal distributions

s_{max} = Maximum potential log-standard deviation value for candidate lognormal distributions

δ = The absolute difference between log-mean values in the set $[m_{min}, \ldots, m_{max}]$

φ = The absolute difference between log-standard deviation values in the set $[s_{min}, \ldots, s_{max}]$

K = The number of candidate log-normal distributions (parameter pairs) in the grid-space:

$= \left(\frac{m_{max} - m_{min}}{\delta}\right) * \left(\frac{s_{max} - s_{min}}{\varphi}\right)$

m_{kh} = The log-mean value for candidate lognormal distribution k in stratum h, with k=1, 2,… K

s_{kh} = The log-standard deviation value for candidate lognormal distribution k in stratum h, with k=1, 2,… K

p_{khi} = The percentile corresponding to u_i – expressed as a proportion – for candidate lognormal distribution k in stratum h

$\vec{p}_{kh}$ = The vector of values p_{khi}

F_k^{-1} = Inverse cumulative density function (CDF) for normal distribution corresponding to candidate parameter pair k

For each candidate distribution k in a given stratum h, calculate the Euclidean distance between the vectors $\vec{c}_h$ and $\vec{p}_{kh}$ as:

$$E_{kh} = \sqrt{\sum_{i=2}^{I} (c_{hi} - p_{khi})^2}$$

The final distribution *k* for stratum *h* is chosen such that the corresponding distance E_{kh} is minimum in the set $[E_{1h}, \ldots, E_{Kh}]$.

By estimating a lognormal distribution for income according to the method described above, we ensure that the selected distribution reflects what we know about the way income is distributed in general (a lognormal distribution), while allowing the distribution's location and scale to be determined by the sample. Additionally, forgoing the requirement that a given boundary point have the same percentile value in the population distribution as observed in the sample allows for sample variation that could lead to no single lognormal distribution achieving equality at every boundary point.

Clearly, the minimum achievable distance E_h is directly related to the choice of granularity parameters δ and φ and the possibility that the true log-mean and log-standard deviation values are contained in the sets used to define the grid-space. For these reasons, it's important to leverage prior knowledge of the target population by using a reasonable range of values m_{kh} and s_{kh} that are sure to contain the true parameters – use of auxiliary data sources can be informative here. In the absence of good starting parameters for construction of the grid-space, a two-step approach can be used. In this two-step approach, one first applies the algorithm to a wide range of parameters with low granularity. The best-fitting distributions parameters may then be used as central points to define narrower but more granular grid axes.

2.3 Percentile-Constrained Interpolation

Once a best-fitting distribution has been identified, random variates may be drawn from it in such a way that the resulting values lie between the boundary points bordering each respondent's categorical response. For each respondent *j* in stratum *h*, interpolate categorical income response to a continuous value according to the distance-minimizing parameters identified in §2.2 above as:

$$y_{hj} = e^{F^{-1}(x)}$$

$$where$$

$$x \sim Uniform(g(i), h(i));$$

$$g(z) = \begin{cases} 0, & z = 1 \\ p_{hi-1}, & z > 1 \end{cases};$$

$$h(z) = \begin{cases} p_{hi}, & z < I \\ 1, & z = I \end{cases}$$

3. Results

In order to validate the proposed approach, two analyses were completed. The first was a simulation study that sought to determine how well the method performed at identifying lognormal parameters from categories derived from random lognormal variates. The second was a case-study analysis that compared percent of the federal poverty limit (%FPL) distributions between two nationally-representative surveys where one survey reports continuous income values and the other categorical.

3.1 Simulation Study

The goal of the simulation study was to establish how well the algorithm performs at identifying the correct lognormal parameters from categories when a known distribution is used to generate the categorical responses. To assess the method's performance, two simulation parameters were introduced: (1) the number of income categories (ranging from 4 to 15 with equidistant boundaries between $0 and $100,000), and (2) the range of data-generating lognormal parameters ($[m, s]$ pairs centered at $[10.5,1,2]$ and ranging ±40% in a common direction). ***Figure 1*** shows the results of the simulation and is based on one million simulated data points per simulation parameter combination.

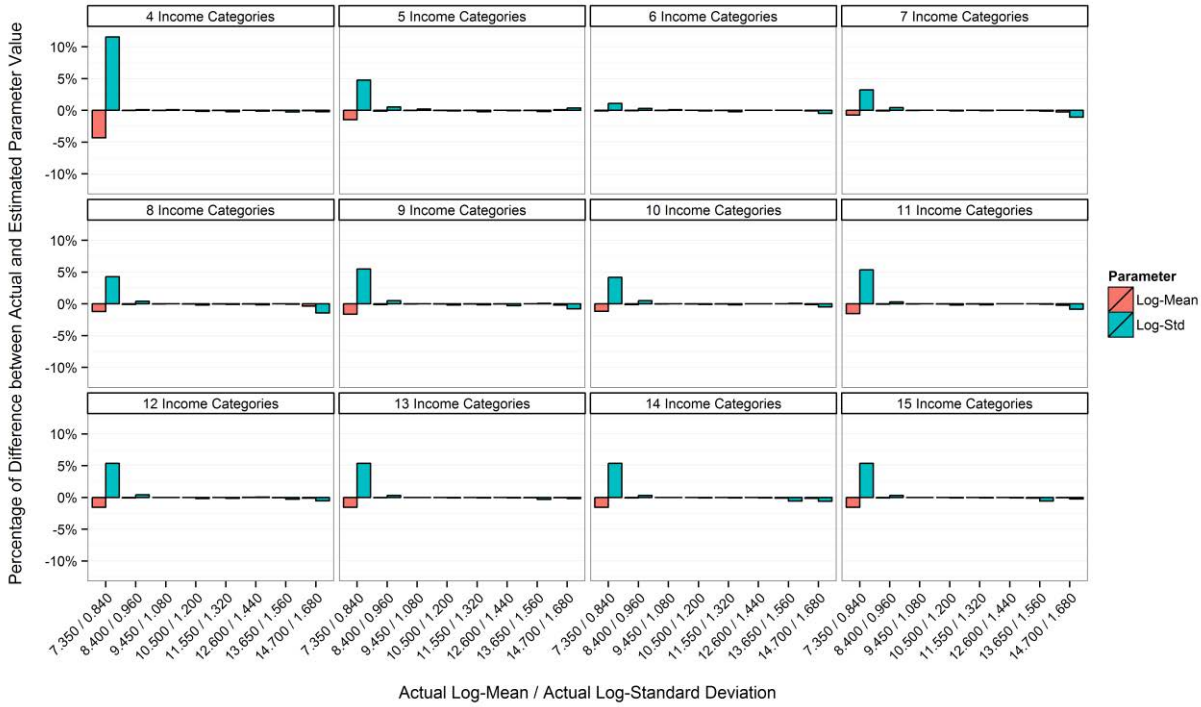

Figure 1: Performance over a Range of Lognormal Parameters and Category Numbers

As shown in ***Figure 1***, the algorithm estimated log-mean and log-standard deviation parameters very close to the true values. In most cases estimated values were within 1% of the true values, and all others were well-controlled. Interestingly, the number of categories had very little impact on accuracy, and no impact for five or more categories.

3.2 Case Study

In application, the proposed method yields interpolated income values that may be used for many statistical purposes. A specific instance of application is an ongoing analysis of the NCVS conducted by the Bureau of Justice Statistics (BJS) that focuses on criminal victimization among individuals across a range of %FPL categories. Since the NCVS collects income data as categories[2], and since thresholds for the federal poverty limit change annually and do not conform to the fixed category boundaries used in the NCVS, a continuous measure of income is required. For this analysis, the method presented above was applied with strata defined as the cross-classification of householder age and race categories[3]. Respondents were then classified according to calculated %FPL and their victimization rates compared. To validate the interpolation technique, NCVS %FPL categories were compared to equivalently-defined categories using data from the Current Population Survey (CPS), which collects income as a continuous measure. ***Figure 2*** shows the results of this comparison.

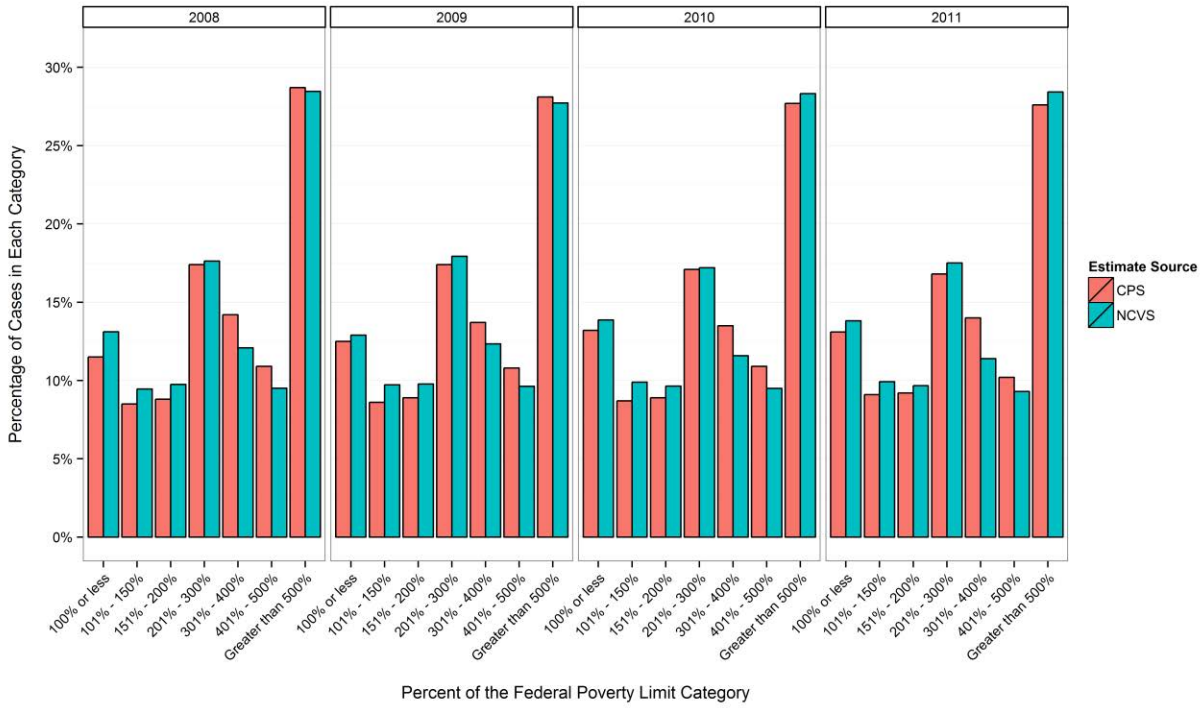

Figure 2: Comparison of NCVS and CPS %FPL Category Distributions

As shown in ***Figure 2***, distributional agreement between the NCVS and CPS was very high, with the vast majority of estimates differing by less than 2 percentage points in any given year. Of particular note is the strong agreement of the 500% or greater category, as the income values used to make this classification often fall well into the uppermost income category of the NCVS which has a lower bound of $75,000. This suggests right tail

[2] The NCVS has 14 income categories that have fine gradations for income levels under $40,000 with ranges of $2,500 or $5,000. Income above $40,000 is split into three categories: $40,000 - $49,999; $50,000 – 74,999; and $75,000 or more

[3] Four categories each for race (non-Hispanic White, non-Hispanic Black, Hispanic, Other) and age (12-29, 30-49, 50-64, 65+) were used, resulting in 16 strata.

estimation is performing well – an area of particular concern for income interpolation methods.

4. Discussion

Collection of income for persons or households in the form of categories is common in the design of survey instruments. Regardless of how well-founded the reasoning for such a choice, it often presents challenges for data users for whom the category boundaries are not ideal. In this paper, the authors have presented a very simple and efficient algorithm for estimating a population lognormal distribution from which the sample of categorical responses is obtained. The method has been shown through simulation to be quite accurate when the lognormal assumption holds, and case study analysis comparing nationally-representative federal surveys demonstrates that interpolated income-based estimates track well with income collected on the continuous scale.

Notably, this method is extensible to address any scenario in which ordinal categories are collected when a continuous measure is required, though the distributional assumptions must be revisited as necessary. Furthermore, the algorithm could be modified to address more complex scenarios when a mixture of distributions would be more appropriate. In such cases, rather than evaluating a vector of percentiles from a known function, one would obtain them empirically through simulation. This flexibility, however, would come at the expense of efficiency. Future research will address these issues as well as potential methods for measuring error in the estimation of distribution parameters.

References

Dikhanov, Yuri, and Michael Ward. "Evolution of the global distribution of income 1970–99." 53rd session of the International Statistical Institute, Seoul, Republic of Korea (2001): 22-29.

Ohio Medicaid Assessment Survey (2015). *2015 Ohio Medicaid Assessment Survey: Methodology Report*.

Pinkovskiy, Maxim, and Xavier Sala-i-Martin. Parametric estimations of the world distribution of income. No. w15433. National Bureau of Economic Research, 2009.

Truman, J. L., and Langton, L. (2014). *Criminal Victimization, 2013*, Government Printing Office, U.S. Bureau of Justice Statistics, Washington, DC.