8-2022

REDI for Binned Data: A Random Empirical Distribution Imputation Method for Estimating Continuous Incomes

Molly M. King
Santa Clara University, mmking@scu.edu

Recommended Citation

King, M. M. (2022). REDI for Binned Data: A Random Empirical Distribution Imputation Method for Estimating Continuous Incomes. Sociological Methodology, 52(2), 220–253. https://doi.org/10.1177/00811750221108086

REDI for Binned Data: A Random Empirical Distribution Imputation Method for Estimating Continuous Incomes

Molly M. King

Department of Sociology, Santa Clara University

Keywords: imputation, distribution free, income brackets, grouped data, inequality, interval censored, top-coded

Corresponding Author: Molly M. King, Department of Sociology, Santa Clara University, 500 El Camino Real, Santa Clara, CA, 95053. Email: mollymkingphd at gmail.com

ABSTRACT

Researchers often need to work with categorical income data. The typical nonparametric (including midpoint) and parametric estimation methods used to estimate summary statistics both have advantages, but they carry assumptions that cause them to deviate in important ways from real-world income distributions. The method introduced here, Random Empirical Distribution Imputation (REDI), imputes discrete observations using binned income data, while also calculating summary statistics. REDI achieves this through random cold-deck imputation from a real-world reference dataset (demonstrated here using the Current Population Survey ASEC). This method can be used to reconcile bins between datasets or across years and handle top incomes. REDI has other advantages for computing values of an income distribution that is nonparametric, bin consistent, area- and variance-preserving, continuous, and computationally fast. I provide proof of concept using two years of the American Community Survey. The method is available as the *redi* command for Stata.

Categorical income data often pose a problem for researchers who seek to include them in quantitative analyses. Income category bins may differ between surveys—or even within the same survey. This results in groupings of income that may not be socially meaningful:

> The truth is that any representation of inequality that relies on a small number of categories is doomed to be crudely schematic, since the underlying social reality is always a continuous distribution. At any given level of wealth or income there is always a certain number of flesh-and-blood individuals, and the number of such individuals varies slowly and gradually in accordance with the shape of the distribution in the society in question. There is never a

discontinuous break between social classes. (Piketty 2014:252)

To better approximate the nature of the social world, researchers often wish to convert the categorical income measures employed in surveys to a continuous measure of income for use as a gradational measure of socioeconomic class (Hout 2004). Some researchers have used categorical questions to impute truly continuous measures of income, but such transformations are uncommon (Bhat 1994; Mellon and Prosser 2018; Schenker et al. 2006). Even within the same dataset, researchers may be presented with overlapping income bins, the result of survey questions probing reluctant respondents with wider categories. When working with longitudinal data, not only do the income categories often change over time, but inflation changes the meaning of income bins (Hout 2004; Ligon 1989). Additionally, the top open-ended bracket poses a challenge for researchers, given the vast inequalities in income and wealth in the United States and around the world (Piketty and Saez 2014). Binned data prevent accurate estimation of such metrics as the median, mean, and inequality statistics. Most researchers seek continuous rather than discrete values for ease of analysis (Hout 2004). Converting categorical to continuous income provides a way to make income comparable across surveys and across years. The few earlier attempts impute incomes from an estimated distribution fit to the bins (Bhat 1994; Mellon and Prosser 2018). Given the vast increase in online trace and simple survey data (Evans and Foster 2019), tools to impute missing or categorical values are increasingly important. Appendix A provides an incomplete list of publicly available, nationally representative datasets that provide binned income data; these are popular datasets for social science researchers to which my method might be applied.

The REDI method is novel in that it imputes income from an independent reference dataset with continuous incomes. At least when it comes to nationally representative surveys, we have no need to approximate continuous distributions of common variables using summary values of binned data. Thanks to comprehensive national surveys, we know the real distribution of incomes in the United States. Similar surveys exist for most countries (e.g., Bailey, Saperstein, and Penner 2014; Donnelly and Pop-Eleches 2018; Wang, Xie, and Hao 2014). Rather than using categorical income bins maintained on smaller national surveys for respondents' confidentiality, the REDI method imputes from the real national distribution of incomes from a real-world reference dataset. This minimizes assumptions about the shape of the distribution and aligns them with empirical knowledge. As I will show, this method also outperforms the other

methods of robust Pareto midpoint estimator, multimodal generalized beta estimator, and cumulative distribution function interpolation in terms of summary statistics that most closely match the underlying data.

I begin by reviewing properties and assumptions of estimators in general. Next, I briefly review the broad categories of approaches for categorical to continuous conversions—nonparametric, including midpoint, and parametric estimation methods—and their respective statistical advantages. For a more detailed discussion, I recommend the reviews by Von Hippel and colleagues (von Hippel, Hunter, and Drown 2017; von Hippel, Scarpino, and Holas 2016). I then describe the four most recent methods introduced for estimating continuous incomes from binned data: the robust Pareto midpoint estimator (von Hippel et al. 2016), cumulative distribution function interpolation (von Hippel et al. 2017), the multimodel generalized beta estimator (von Hippel et al. 2016), and mean-constrained integration over brackets (Jargowsky and Wheeler 2018).

Next, I present my method for empirical estimation of continuous incomes from binned data, Random Empirical Distribution Imputation (REDI), together with details about its advantages and assumptions. I then provide a proof-of-concept demonstration using two nationally representative datasets. I demonstrate the usefulness of the resulting imputed income in some model regressions using state-level data. Finally, I conclude with discussion of directions for future research, advantages, and scope.

PROPERTIES AND ASSUMPTIONS OF ESTIMATORS

In a research dataset with binned data, there are M income brackets,

$$B = \{1, 2, ..., ..., M\},$$

in ascending order of income level. Each bracket has an upper income limit (U_b) and a lower income limit (L_b), except the top and bottom brackets, which are open-ended.

Most methods treat the bottom of the lowest bracket (L_1) as 0. This appears to be a generally justified assumption. In the Integrated Public Use Microdata Series, for instance, only 0.05 percent of all households have negative income (von Hippel et al. 2016). However, aside from in the standard calculation of the Gini index, the REDI method does not require even this assumption.

A popular approach is to extrapolate from the next-to-top bin using the frequencies of the

top two categories in a formula based on a Pareto distribution. This is an arbitrary but convenient assumption (von Hippel et al. 2016). Nonetheless, prior work shows the Pareto distribution is an inexact fit to this upper tail (Blanchet et al. 2018; Hout 2004). This is important because even small errors in defining this distribution will result in large errors in estimating the variance and other parameters (Jargowsky and Wheeler 2018).

Nonparametric Estimators

Nonparametric estimators make no assumptions about the exact form or function underlying the distribution of incomes. First, I present the most fundamental form of nonparametric estimator: the basic midpoint estimator. Next, I review a more recent modification of this midpoint method, the robust Pareto midpoint estimator. Then, I present a nonparametric estimator based on a different premise: CDF Interpolation.

Midpoint estimators

The basic midpoint estimation method assigns the midpoint value of the income bin to all respondents who reported income in that bin (Hout 2004). The midpoint of each bracket (other than the top) is assigned such that

$$\text{midpoint}_b = (L_b + U_b)/2.$$

Midpoint estimators have the advantage of being *bin consistent*: because they are nonparametric (and therefore make no assumptions about the underlying distribution), the estimate of the midpoint approaches the real median income of that bin as the bins get narrower (von Hippel et al. 2016). Midpoint estimation methods are also very fast to compute. Midpoint estimation methods are most accurate when the dataset has many bins (von Hippel et al. 2017).

The midpoint can also be a good approximation of the mean value for most income bins, except for the most extreme categories (Ligon 1989). Even when the lowest and highest income categories are not treated as unbounded on each end, the distributions within these bins often demonstrate unusual properties.

Robust Pareto midpoint estimator

The simple Pareto midpoint estimator assigns the value of the bin midpoint to observations, except for the top bracket, where cases are assigned the arithmetic mean of a Pareto distribution (Henson 1967). The robust Pareto midpoint estimator (RPME) improves on the simple Pareto midpoint estimator by ensuring the mean of the top bin is defined. The RPME is made robust by constraining the Pareto parameter and/or replacing the arithmetic mean with

the median, the geometric mean, or the harmonic mean.

With a limited number of bins, this method is less accurate than the multimodel generalized beta estimator method (MGBE, explained below) at estimating the median and Gini index, but can be equally accurate with many (more than 16) bins. The RPME performs almost as well as the MGBE method when estimating mean values, as long as the data have at least four bins. The RPME is also much faster and particularly robust in estimating small samples (von Hippel et al. 2016).

CDF interpolation

Cumulative distribution function (CDF) interpolation is a nonparametric method that fits a continuous density function to binned income data. This method defines "a nonparametric density that fits the bin counts exactly" (von Hippel et al. 2017:646). The function used for the top can either be Pareto, rectangular, or exponential (von Hippel et al. 2017). A major advantage of the CDF interpolation method is that it *preserves areas*: the fraction of incomes that should be in the bin according to the fitted distribution is equal to the observed fraction of incomes actually in the bin (von Hippel et al. 2017). The package developed to implement this function in R is called *binsmooth* (Hunter and Drown 2020; von Hippel et al. 2017).

However, like the RPME method and the multimodel generalized estimator, CDF interpolation only computes summary distributions and does not assign unique incomes to observations. Recent work incorporates external data within subgroups into this CDF interpolation approach to improve individual-level imputation of missing data developed using the British Election Study. This approach relies on some variables not available in all nationally representative surveys (e.g., newspaper readership), does not provide implementation code, and has not been peer reviewed at the time of this writing (Mellon and Prosser 2018).

Parametric Estimators

Perhaps the primary feature characterizing methods estimating income from binned data is whether they assume that a set of parameters or a certain distribution underlies the income estimations. Parametric estimators follow probability distributions based on a set group of parameters (a function). One strength of parametric functions is that they treat income as continuous (von Hippel et al. 2017), reflecting real-world conditions. The traditional parametric approach method fits a single probability distribution to the entire set of income bins (McDonald 1984; McDonald and Ransom 1979).

Parametric estimates perform well with fewer than eight bins because of their smoothness; nonparametric midpoint estimates perform better with many bins (von Hippel et al. 2016). Parametric and nonparametric methods perform similarly well with 15 to 25 income bins (von Hippel et al. 2016). The parametric distribution may not produce accurate estimates if it does not reflect the underlying population well (von Hippel et al. 2017). Parametric functions are also computationally slow (von Hippel et al. 2016).

Multimodel generalized beta estimator

Generalized beta densities are some of the most popular parametric estimators for modeling income. The multimodel generalized beta estimator (MGBE) fits 10 different income distributions from the generalized beta family. It then uses the AIC and BIC fit statistics to choose among them or to average estimates across models. This is an improvement over the traditional parametric approach because there is no a priori method of determining which distribution will provide the best fit for the data. The MGBE method is slower than nonparametric methods, but with small numbers of bins, it provides more accurate estimates for the median and Gini coefficient (von Hippel et al. 2016).

Mean-constrained integration over brackets

The mean-constrained integration over brackets (MCIB) method estimates parameters of the income distribution by computing integrals over each income bracket. It improves on the traditional midpoint estimator, the robust Pareto midpoint estimator, and the multimodel generalized beta estimator by making more complete use of the available information. Mean-constrained integration over brackets allows densities of incomes to change continuously within and across brackets. MCIB uses the Pareto distribution for the upper-tail income and allows for inclusion of zero and negative incomes (Jargowsky and Wheeler 2018). One disadvantage is that there is currently no implementation of MCIB for any of the popular statistical packages, so it remains theoretical. Therefore, no results of MCIB are included for comparison in the evaluations presented here.

RANDOM EMPIRICAL DISTRIBUTION IMPUTATION

Previous recommendations for improvements on these existing methods include the "idea of fitting a density that is as smooth as the [generalized beta] family densities but as flexible and assumption free as the [robust Pareto midpoint estimator]" (von Hippel et al. 2016:245). The Random Empirical Distribution Imputation (REDI) method, introduced here, achieves this by

being both nonparametric and continuous. Rather than fitting the bins to an artificial distribution or averaging among those that fit from a generalized beta family, the REDI method uses a real-world distribution of income (the reference dataset) to estimate the distribution within each income bin (of a research dataset). Furthermore, it not only supplies a distribution fit to the research dataset, it does so by assigning a discrete value from that distribution to every observation. This direct assignment makes it straightforward to use these data in regression or other analyses. To enable ready use, I made the method available as the *redi* command for Stata, which can be installed using Stata's *ssc* command.

Converting Categorical to Continuous Income

REDI starts by taking each income category of the research dataset, with bounds L_b to U_b, and counting the number of observations in that income category, N_b. REDI then draws at random the same number of observations (N_b) from within those same bounds (L_b to U_b) in the corresponding reference income distribution. Within the boundaries of each bin, REDI draws observations from the reference dataset using simple random sampling with replacement. Sampling with replacement ensures the values for the observations are independent and identically distributed (IID).

The method then applies these exact household income values from the reference dataset to observations from the corresponding set of categorical values within the research dataset. Thus, within a given bin, we get a distribution of numerical values for the research dataset that corresponds to their probability of occurrence in the same income range of the reference dataset. Repeating this for all income categories B within the research dataset provides income values for every observation based on a random draw from within that same income bracket in the reference dataset.

In other words, for every observation between L_b and U_b, REDI draws at random an income value between L_b and U_b from the reference dataset. This value is assigned to the original observation.

Figure 1 illustrates the concepts underlying the method. A given research dataset has a certain proportion of observations within each income bin, as represented by the horizontal line N_b / N. The REDI method starts with the income bin that each individual is in for the research dataset of interest (e.g., the third bin between L_3 of \$100,000 and U_3 of \$150,000). It takes a

sample of N_3 observations from the reference dataset, as represented by the small circles in this income bin. REDI assigns a random income value within that range (represented by one of the circles, say $123,000) to each observation from the research dataset. In the aggregate, these randomly sampled income values approximate a continuous distribution for the research dataset (represented by the gray line). Income values from the reference dataset are therefore randomly assigned to the research observations.

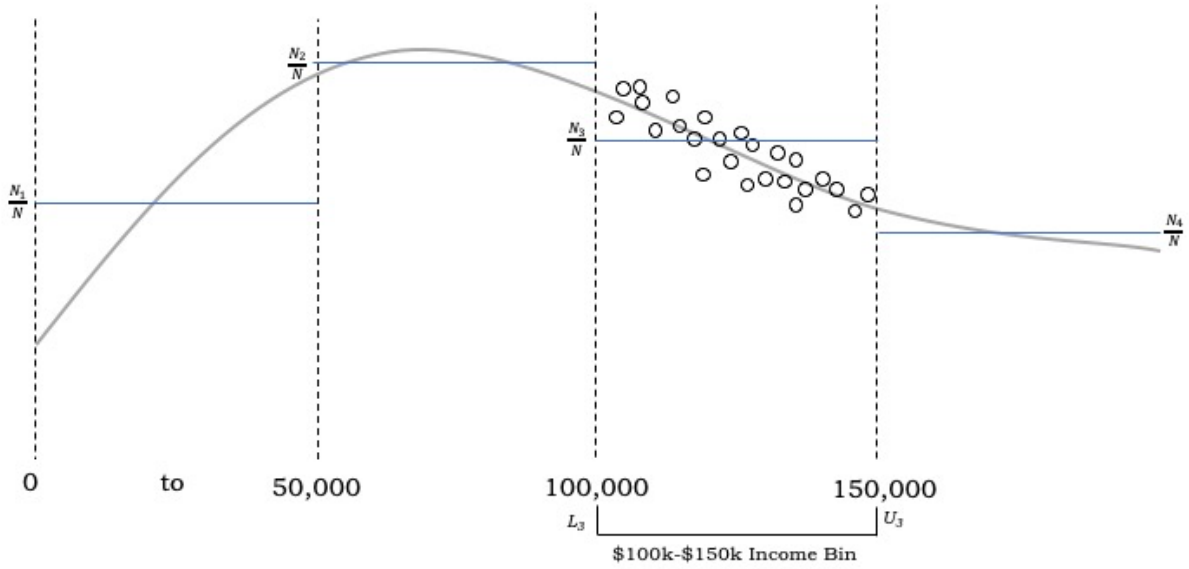

Figure 1. Illustration of the REDI Method
Note: The horizontal lines within each bin marked with N_b / N indicate the proportion of the total number of income observations within each bin. The horizontal axis reflects a continuous distribution of income, with vertical dashed lines indicating upper and lower boundaries for income bins in the research dataset. The open circles represent the likelihood of drawing a given income value from the reference dataset. The curved line (a scaled probability density function) represents the probability of sampling a given specific income, smoothed across the entire research dataset. Thus, the number of income values sampled from all income bins approximates a continuous distribution.

REDI estimates top incomes using the same method as all other values. For this reason, it is important that the reference dataset not be top-coded. Top-coded incomes in the research dataset are not a problem. When the method applies values from the reference dataset to the research dataset, top-coded values are treated as belonging to the range (L_M, $\max_I$), where $\max_I$ is the maximum income in the reference dataset, rather than the range (L_M, ∞), as they would be in a Pareto distribution.

Adjusting for Inflation

The REDI method has the option to adjust all dollar amounts for inflation. The inflation adjustment built into the REDI program uses the Consumer Price Index retroactive series using current methods with all items (R-CPI-U-RS). However, in theory, researchers could choose any inflation index. This yields the equivalent purchasing power of the income for any year in the research dataset in terms of the reference year (Fontenot, Semega, and Kollar 2018; Hout 2004). To do this, I create a conversion factor for each research year relative to the reference year, in this case based on the R-CPI-U-RS for all items (U.S. Bureau of Labor Statistics 2015, 2020). The continuous income obtained from the steps above is then converted to current reference dollars by dividing the income by this conversion factor. This enables comparison of income measures across surveys and across years.

Advantages of the Method

The REDI method is particularly valuable under certain conditions because it is continuous, nonparametric, area- and variance-preserving, bin consistent, and computationally fast. REDI is useful when researchers have binned income data they are interested in converting to continuous income data. The method has the advantage of a parametric approach in that it treats income as *continuous*. Yet it assumes no parameters or particular shape of distribution underlying the dataset of interest; REDI is a *nonparametric* method. Rather, REDI requires a reference dataset that can be used to draw the income distribution. Because it samples the same number of incomes from the reference distribution as the number of observations in the original bin of the research dataset, the REDI method *preserves areas and variances*.

Most notably, REDI allows researchers to reconcile different income brackets across multiple datasets or years using continuous income estimates. Income category bins differ between surveys. For example, one dataset may have income bins ranging from $0 to $29,999 and $30,000 to $49,999; another may have bins that span $0 to $19,999 and $20,000 to $59,999. REDI adapts easily to both datasets, allowing researchers to create comparable point estimates for respondents in each dataset. This is an improvement on midpoint estimation methods, where the bounds of each bin result in midpoint estimates that are *bin consistent* but not necessarily comparable with other midpoint estimates from different datasets. As a result, REDI can be particularly helpful for reconciling bins of different values across years of the same dataset.

Because it is fully nonparametric, REDI is a resource for researchers particularly interested in top incomes. The method estimates top incomes without relying on the Pareto

distribution, an advantage because true income distributions deviate in important ways from this idealized model (Blanchet et al. 2018; Hout 2004). By treating top-coded incomes as belonging to the range of empirical data (L_M , $\max_I$), available from a reference dataset, REDI maintains more realistic assumptions about the distribution of top incomes.

The method also allows for the inclusion of zero and negative incomes without recoding, so the distribution of variance of the final result is the most accurate reflection of the world as possible. REDI preserves variance and retains information from the extreme income categories. Furthermore, it does not need additional information about the bins themselves, such as the mean or median of each bracket (e.g., Jargowsky and Wheeler 2018).

REDI is also a good alternative when hot-deck single or multiple regression imputation are inappropriate or impossible. In single regression imputation, a regression equation is used to predict missing observations from the complete observations, assuming perfect correlation. Hot-deck imputation uses methods to match observations with missing values to otherwise similar cases in the same research dataset (Gelman and Hill 2006; Roth 1994). In some situations, hot-deck single and multiple regression imputation may not be the best methods for research needs: single regression imputation underestimates standard errors, and the validity of multiple imputation relies on a well-specified model that includes covariates (Allison 2000, 2002). In contrast, cold-deck imputation draws data for missing cases from otherwise similar observations in a different dataset (Hu and Salvucci 2001; Yan 2011). The proof of concept example discussed below uses the Current Population Survey March Annual Social and Economic Supplement (CPS ASEC) (U.S. Census Bureau and U.S. Bureau of Labor Statistics 2018) as this reference dataset, which would be appropriate for most researchers dealing with nationally representative research datasets.

Assumptions of the Method

REDI operates under three key assumptions: the sample frame and sampling method are identical in the research and reference datasets; the income question is the same in the datasets; and income is not top-coded in the reference dataset.

First, the sample frames and sampling methods must be the same between the reference and research datasets. Although using REDI to sample from a reference dataset will result in more realistic imputed values, it is only practical if the reference dataset and the binned research dataset represent approximately the same income distribution. Such continuous reference

distributions may be difficult to find for small sampling areas (e.g., counties, school districts). Nonetheless, they are available for individual U.S. states in the CPS ASEC and for many countries.

Second, the income question must be the same, both in phrasing and unit of observation. For example, some surveys ask about before-tax income, and others ask for after-tax income. Furthermore, there must be a match between research and reference datasets in asking about individual, household, or family income. If there are differences in any of these, then the underlying distributions differ and any inference may be biased.

Finally, if income is top-coded in the reference dataset, the REDI method may still be used, but additional assumptions will need to be added (e.g., the application of a Pareto distribution to top incomes). Overall, the assumptions are practical compared to other methods.

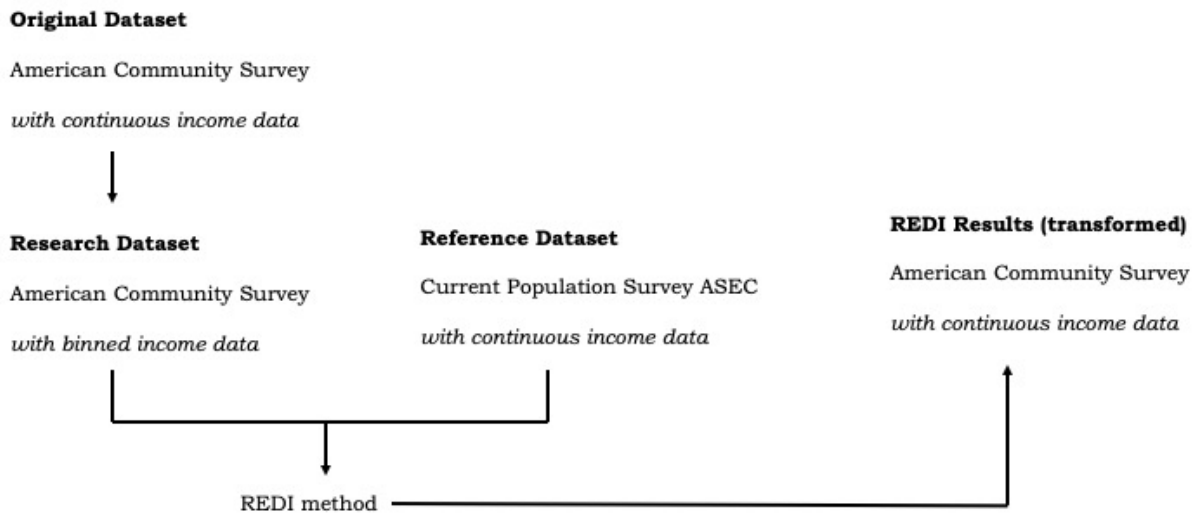

Figure 2. Workflow of REDI Method Using Original, Research, and Reference Datasets
Note: Both the American Community Survey (ACS) and the Current Population Survey March Annual Social and Economic Supplement (CPS ASEC) are publicly available with continuous income data. To create the research dataset, I added artificial bins to the ACS. The CPS ASEC serves as the reference dataset for the REDI method, which is used to impute continuous incomes for the research dataset.

PROOF OF CONCEPT

As discussed earlier, the REDI method requires two datasets: a research dataset with binned income categories and a reference dataset to draw a sample from a similar population with continuous income data. In the demonstration illustrated here, I use an artificially-binned version of the American Community Survey (ACS) as the research dataset (for demonstration) and the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) as the

reference dataset. Figure 2 illustrates how these datasets are used in this article as inputs for the REDI method to produce the transformed research dataset with continuous income data. The ACS and CPS ASEC are both weighted to yearly population estimates.

The following section proceeds by first converting the ACS from the original dataset with continuous income data to the research dataset with binned income data. I then use this binned dataset as my research dataset for the REDI method (allowing the original, continuous version to be used as a baseline for comparison of performance). Results from the method are transformed using inflation measures (here in 2017 dollars) to be comparable across years. The following sections lay out tools for diagnostics, variable transformation, and performance metrics.

The Research Dataset (ACS)

For demonstration purposes, I use the American Community Survey (ACS) as my research dataset (Ruggles et al. 2021). When using the REDI method, researchers would substitute their own datasets of interest for the ACS.

The ACS is currently the largest household survey in the United States, sampling 3.5 million households (approximately 295,000 per month) on a rolling basis. The ACS questions about respondents' income—including wage and salary, self-employment, property income, Social Security, family assistance, retirement, and other—are less detailed than those of the CPS ASEC. The ACS asks for respondents' income during the previous 12 months (rather than the CPS's request for details about the previous calendar year) (Rothbaum 2015).[1]

Again, the ACS stands in as an example dataset on which a researcher could use the REDI method. The only requirement for the research dataset is that it have binned income data. Total income in the ACS varied from 84.2 to 87.5 percent of the administrative benchmark (these totals are not significantly different from the CPS ASEC), but the aggregate wage and salary data in the ACS are less accurate (Rothbaum 2015). The ACS is a useful dataset for proof of concept because it provides continuous income data to check the resulting summary statistics.

The Reference Dataset (CPS ASEC)

This illustration of REDI uses the Current Population Survey Annual Social and Economic Supplement (CPS ASEC) as its reference dataset. The CPS ASEC interviews around 100,000 households of the non-institutionalized U.S. population per year. These interviews of 200,000 individuals asked respondents about their income in the previous calendar year, family and household characteristics, and program participation (U.S. Census Bureau and U.S. Bureau of Labor Statistics 2018). Here, to eliminate issues of collinearity among members of the same

household, I selected only one member from each household and weighted household income for the individual using household survey weights.

The CPS ASEC asks about income from over 50 sources, including wage and salary, self-employment, interest and dividends, Social Security, pensions, family assistance, worker compensation, and unemployment compensation (Rothbaum 2015). This provides detailed coverage of all individual potential income sources. Importantly, both the CPS ASEC and the ACS use the same definition of earned pre-tax income, excluding capital gains and certain lump-sum payments (Rothbaum 2015:12).

The CPS ASEC is a particularly strong choice as a reference dataset for nationally representative studies in the United States. It is publicly available to all researchers, yet matches total income data at 83.0 to 86.3 percent of benchmarks generated using administrative data. Self-reported earnings in the CPS ASEC are particularly close to the administrative benchmark, at 96.2 to 99.0 percent between 2007 and 2012 (Rothbaum 2015).[2] The CPS ASEC also includes income data for individuals, families, and households (U.S. Census Bureau 2018). Perhaps most importantly for REDI, after 2011, incomes in the CPS ASEC are no longer top-coded (Minnesota Population Center 2018). Hence, I chose the CPS ASEC as the reference dataset for my demonstration of the REDI method, even over the larger ACS, and I recommend it to other researchers looking for a U.S. nationally representative source of reference income data for use with the REDI method.

Binning Incomes

The original ACS dataset does not use categorical bins for income data, so I impose categorical breakdowns on the ACS dataset prior to implementing the REDI method. One measure of dispersion is household income in selected categories that the Census Bureau often uses in summarizing data from the CPS ASEC (Fontenot et al. 2018:27). These categories, presented in Appendix Table B1, divide the population into approximately tenths. The bins are useful for demonstration purposes here, in that they are commonly used options on surveys. Appendix B also illustrates the match between the distributions of these income bins when imposed on the actual CPS ASEC public-use dataset and the ACS dataset.

Drawing Observations from the Reference Dataset

REDI draws the distribution of population incomes from between the lower bound, L_b, and upper bound, U_b, of each bracket of the CPS ASEC. In practice, because the research

dataset (the ACS) has a much larger N than the reference dataset (the CPS ASEC), observations from the reference dataset will be sampled multiple times. I implement repeated sampling by generating uniformly distributed random integer values on the interval $(1, N_b)$.[3] Each value from within the income category in the reference dataset has probability of being assigned to each observation in the research dataset drawn from a $(1, N_b)$-uniform variable. Each observation of a continuous income drawn from the CPS ASEC is assigned to each original observation in the ACS research dataset between L_b and U_b.

Handling Top Incomes

For the top-coded income category in the research dataset, I again use the distribution from the CPS ASEC reference. Starting in 2011, the CPS ASEC uses a rank proximity procedure to replace top incomes with near-neighbor approximate values.[4] Individuals at risk of being identified in the data have their incomes rounded to two significant digits and then swapped within certain bounds (Minnesota Population Center 2018). This is not an issue for REDI because it uses the total CPS ASEC dataset as a population from which to draw the random sample. Any incomes that ought to have been included in one category but were swapped with incomes in the second-highest bin by the top-coding rank proximity procedure will still be included with probability equal to their occurrence in the reference dataset. Therefore, I treat top-coded income categories like all others. This approach might be further improved with the enrichment of top incomes in the reference dataset to better reflect the true distribution of high incomes (Fixler, Gindelsky, and Johnson 2019), or adjustments to reflect nonresponse bias in the upper and lower tales of the income distribution (Bollinger et al. 2019).

Performance Metrics

I compare estimates from each method with estimates calculated directly from the original datasets.[5] Although these are, of course, population estimates, for ease of reference I refer to them as the estimand or the true value, because they are as close to true population values as possible, without access to administrative records.

Again following von Hippel and colleagues (2016), I use the values of median and mean income from a single population random sample, as well as Gini index, to measure how close REDI and other methods come to estimating the true population values from the original ACS research dataset. I obtained median variance estimates with survey weights using the user-written Stata package *epctile* (Francisco and Fuller 2008). Note that to obtain only the median value for

the research dataset of interest, only the application of REDI to the income bin containing the median is important; therefore, researchers may avoid any assumptions about the distribution of the top-coded category for this purpose. The Gini index is a summary of income inequality. The index varies from zero to one: zero indicates perfect equality, where there is equal distribution of income, and one indicates complete inequality (Guzman 2018). These measures of REDI results, including standard deviation of a single sample, provide an estimate of the variability of incomes among individuals.

Because the REDI method involves randomness, I also present results of the mean and median of repeated samples from the population. The resulting measures of central tendency and standard errors evaluate how precise the estimates are from REDI (and thus, how the method performs relative to other methods). In this way, the standard error measures the variability of the method across samples. To demonstrate this and better understand the variability of REDI across repeated samples, I repeated the method 500 times for each year. Based on these results, I calculated the mean, standard error, and median across these 500 repetitions.

Finally, I model artificial multivariate regressions using the original continuous values from the research dataset and the REDI-calculated income values. The similarity of these regression coefficients serves as a measure of the usefulness of the method for generating values that can be used not just for aggregate statistical measures but also for direct correlation analysis. In these model regressions, I use nationally representative samples and subnational samples. With the latter, I demonstrate the potential for the method to be used with any research dataset so long as an appropriate reference dataset exists (in this case, state subsets of the CPS ASEC).

RESULTS

Here I present the results of each method for estimating income from binned data to the original continuous values from the ACS research dataset. The measure of accuracy in evaluating each method is how close the generated summary statistics are to the original ACS estimands. First, I compare calculations using the REDI method; for illustrative purposes, REDI uses the Current Population Survey March Annual Social and Economic Supplement (CPS ASEC) as the reference dataset, as described earlier. I demonstrate the method using a single sample, then illustrate precision in repeated samples. Next, I apply the robust Pareto midpoint estimator, the multimodel generalized beta estimator, and CDF interpolation each separately to estimate median, mean, and Gini index values from the artificially-binned ACS research dataset. I

compare the results of these models to each other and to the original ACS estimands.

REDI Results: Single Sample

First, I demonstrate the use of REDI using a single sample from the ACS research dataset. Because there is some stochasticity built into the method, each new instance of results will differ slightly; the results presented here are a single representative set using a reproducible seed. In the next section, I calculate the average of many such repetitions. Table 1 presents the results of a single repetition of the REDI method compared to the original ACS research dataset and the CPS ASEC reference dataset from which the incomes for the bins were drawn.

In this set of results, REDI estimates median household income for 2016 at $68,823 (95 percent confidence interval: $68,503, $69,142). The method estimates median household income for 2017 at $70,186 (95 percent confidence interval: $70,002, $70,369). These are very close to the true median values for the ACS research dataset for both 2016 ($69,687) and 2017 ($71,000). In this instance, the REDI-calculated mean for 2016 is $90,062 (95 percent confidence interval: $89,841, $90,285); for 2017 it is $91,820 (95 percent confidence interval: $91,592, $92,049). These values are significantly lower than the original ACS mean values of $93,825 for 2016 (Wald test, $t = 58.77$, $p < 0.001$) and $95,442 for 2017 (Wald test, $t = 51.16$, $p < 0.001$) (UCLA: Statistical Consulting Group 2020). All estimates are in 2017 dollars. The Gini index estimated by REDI are also close to the true values for the ACS dataset, varying by less than 0.015 index points.

Table 1. Comparisons of Reference and Research Datasets with Single Instance of REDI Results

Data	Median		Mean		Gini	
Method	2016	2017	2016	2017	2016	2017
CPS ASEC	57,230	58,849	80,822	82,957	0.478	0.481
ACS	69,687	71,000	93,825	95,442	0.455	0.454
REDI	68,823	70,186	90,062	91,820	0.442	0.441

Note: All results are for household income values in 2017 dollars. CPS ASEC values are noted for reference from direct calculations on the dataset. The CPS ASEC values also differ from those in the corresponding official report: medians of $60,309 (2016) and $61,372 (2017); means of $84,931 (2016) and $86,220 (2017); Gini indices of 0.481 (2016) and 0.482 (2017) (Fontenot et al. 2018). ACS original median, mean, and Gini index values are calculated directly from the 2016 and 2017 datasets. These values differ substantially from the median values of $58,820 (2016) and $60,336 (2017); the mean values of $85,995 (2016) and $89,294 (2017); and the Gini indices of 0.482 (2016 and 2017) reported in the American Community Survey Brief on Household Income (Guzman 2018). REDI estimates a continuous income distribution based on artificial bins created from the ACS, with reference incomes drawn from the CPS ASEC. Inflation to 2017 dollars was calculated using the R-CPI-U-RS (U.S. Bureau of Labor Statistics 2015). Values are all calculated in Stata/MP version 16.1.

Figure 3 presents a comparison of the 2017 income distributions from the original ACS

dataset and results of the REDI method. This depiction of the cumulative distribution function shows that the income distribution calculated by REDI is very similar to that of the original ACS distribution throughout the entire range. If I were to plot histograms of the two distributions on top of each other, we would see identical proportions (and numbers of observations) of the distributions in each income bracket in the original ACS research and REDI-calculated datasets, since by design REDI assigns the same number of incomes as original observations in each income bracket.

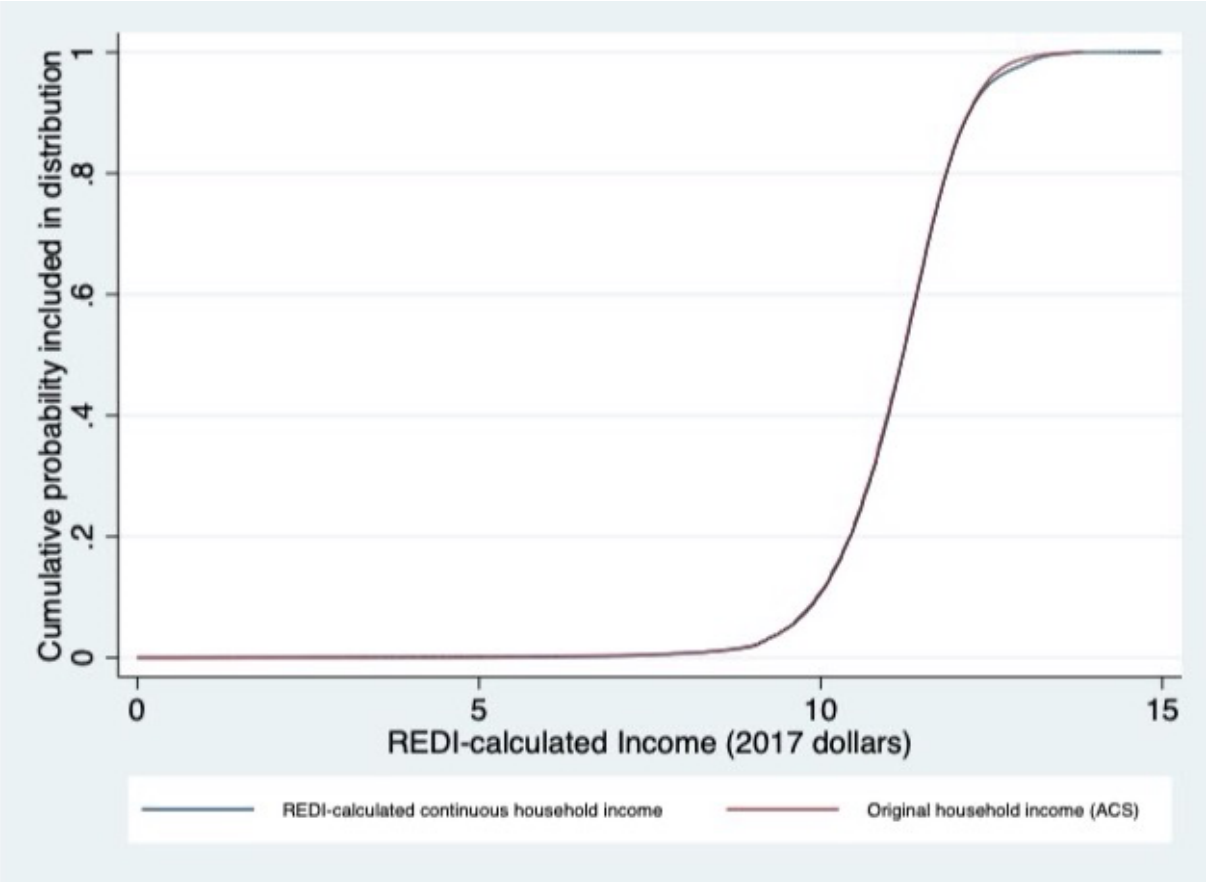

Figure 3. Cumulative Distribution Functions Comparing Income Distributions from Original ACS Data and REDI-Computed Results
Note: Household income from 2017 is natural log-transformed. Plot produced using Stata package *distplot* (Cox 1998).

Still, a two-sample Kolmogorov-Smirnov nonparametric test for equality of distribution functions finds that the two distributions are significantly different. With a *p*-value of less than 0.001, we can reject the null hypothesis that both samples were drawn from the same population. This also holds true with an Epps–Singleton two-sample nonparametric test for equality of distribution functions, despite the slightly relaxed assumptions of this test (Goerg and Kaiser

2009).

REDI Results: Repeated Samples

Because the REDI method draws randomly from the reference distribution, the resulting dataset (and summary measures) will always involve some amount of randomness. The degree of variation in the resulting estimates, of course, is a function of the initial variation in the reference dataset. To get a sense of reliability across samples, I repeated the method 500 times, each time resampling from the CPS ASEC reference dataset but using the same income categories from the ACS research dataset.

The grand mean (the mean of the means) income of the 500 repeated samples calculated using REDI for 2016 was $92,451, with a standard error of less than $2. The range of means calculated during repeated applications of REDI on the reference dataset ranged from $92,382 to $92,548. The grand median calculated for 2016 is $70,360. The grand mean income for repeated samples in 2017 is $94,104, ranging from a minimum of $94,006 to $94,214. The standard error is also under $2. The grand median income for 2017 is $71,565.

These grand means are closer to the mean values of the true ACS mean household incomes then the means calculated by a single deployment of the REDI method. Repeating the method averaging 500 different samples per year from the CPS ASEC results in a difference between the REDI grand means and the ACS estimand of an average of $839 (1.2 percent); the single REDI repetition results in an average difference of $3,693 (3.9 percent) between the means. However, the median values using only a single draw from the CPS ASEC are only about $200 (0.03 percent) closer to the ACS median estimands than the grand median values using repeated sampling. Repeating the REDI method and averaging the results does produce more accurate mean estimates, likely a result of better representation of the more rare very high income values in the reference dataset. However, for researchers interested primarily in median summary statistics, a single sample from the reference dataset is quite adequate.

Robust Pareto Midpoint Estimator Results

I used the Stata package *rpme* (von Hippel and Powers 2015; von Hippel et al. 2016) to calculate the robust Pareto midpoint estimates for the ACS research dataset. First, I performed the calculation including the grand mean (von Hippel et al. 2017). When including a grand mean in RPME estimates, I calculated the mean of the top bin such that, when it is included in calculations with all bin midpoints, it will equal the grand mean. Using the grand mean, the robust Pareto midpoint estimation method calculated the 2016 median income to be $51,287 and

the 2017 median income to be $49,885 (both in 2017 dollars). This specification is substantially different from the ACS median income estimand, especially in 2017 (see Table 2).

Although the mean-constrained specification is recommended, I also performed the variation of the robust Pareto midpoint estimator without specifying the grand mean. Under this specification, the mean of the top bin is calculated by fitting a Pareto curve (using a harmonic mean) to the top two income bins. Under this set of assumptions, the 2016 median income is $63,844 and the 2017 median income is $62,500. Gini index estimations under both specifications for both years are close to or slightly above 0.5.

Multimodel Generalized Beta Estimator Results

The multimodel generalized beta estimator can be calculated using the Stata package *mgbe*. The *mgbe* package fits up to 10 generalized beta family distributions to the research dataset, either averaging estimates across distributions or choosing the best-fitting models using multimodel inference (von Hippel et al. 2016).

The *mgbe* package does not allow negative incomes as lower bounds of categories. Therefore, unlike calculations for the other procedures, the MGBE values are calculated with a lower limit of $0 for the lowest income bracket. If anything, this should bias the mean upward toward the original ACS values, making the substitution of $0 for negative values conservative, given the results.

The MGBE method has the advantage of allowing the researcher to average the estimates across beta family distributions. I used the *mgbe* package for Stata to produce average parameters from all 10 generalized beta family distributions fit to the two years of ACS data. Besides REDI, the MGBE was the method that came closest to the original ACS values, with medians of $71,196 in 2016 and $72,403 in 2017 and mean values of $92,519 in 2016 and $93,532 in 2017. The Gini index estimates were a bit further from the ACS values than the REDI and CDF interpolation calculations, with MGBE approximating the Gini index at 0.433 in 2016 and 0.430 in 2017.

CDF Interpolation Results

Cumulative distribution function (CDF) interpolation is performed using the *binsmooth* package in R (Hunter and Drown 2020). For CDF interpolation by line segments, the CDF is polygonal, and the method is carried out using the *stepbins* function from the *binsmooth* R package (Hunter and Drown 2020; von Hippel et al. 2017). The *binsmooth* package allows the top bin of the step-function CDF to follow the shape of either a rectangular, exponential, or

Pareto distribution; I chose the Pareto distribution to best compare with the RPME estimates.

For CDF interpolation using a continuous, monotonic cubic spline, I used the *splinebins* function from the *binsmooth* R package (Hunter and Drown 2020; von Hippel et al. 2017). I used the "hyman" method for constructing the monotonic spline, as the authors suggest this allows the function to integrate faster and produce smoother density functions (Hunter and Drown 2020).

This mean-constrained CDF interpolation using the cubic spline produced the summary values second-closest (after REDI) to those from the original research dataset. I calculated Gini coefficients and means using numerical integration after calculation of the CDFs.[6] The Gini index for the CDF interpolation using the cubic spline was 0.44 in both 2016 and 2017. The median is also calculated numerically.[7] The medians calculated using the mean-constrained CDF interpolation method using the cubic spline were $71,503 for 2016 and $72,700 for 2017. These values are further from the ACS median estimands than those of REDI, on average, although not by much.

The CDF interpolation method produces an overall distribution. The summary statistics produced by *binsmooth* were very similar to the original ACS estimands. The empirically tested version of *binsmooth* does not impute individual observations, instead fitting a distribution to income bins. The *binsmooth* package used to calculate CDF interpolation does have a function for drawing a random sample from this distribution, but the authors warn it should be regarded as experimental (Hunter and Drown 2020). Researchers would also have to assign these income points manually to observations within their income bins, making it impractical for the repeated application necessary for regression analysis. REDI has the advantage of being able to quickly produce discrete values tied to each observation for use in later analyses.

Performance of REDI in Summary Statistics

Table 2. Comparisons of REDI Grand Means with Five Alternative Estimators for a Continuous Income Distribution Based on Artificial Bins Created from the ACS

Method	Parameter	Median		Mean		Gini	
		2016	2017	2016	2017	2016	2017
ACS data	(original estimand)	69,687	71,000	93,825	95,442	0.455	0.454
REDI	(grand mean of repetitions)	70,360	71,565	92,451	94,104	0.442	0.441
RPME	(mean-constrained)	51,287	49,885	93,825[a]	95,442[a]	0.505	0.506
RPME	Pareto curve (harmonic mean)	63,844	62,500	116,798	119,577	0.500	0.510

MGBE	average of 10	71,196	72,403	92,519	93,532	0.433	0.430
CDF Interpolation	Polygonal (mean-constrained)	71,917	72,890	93,825 [a]	95,442 [a]	0.438	0.438
CDF Interpolation	smooth spline (mean-constrained)	71,503	72,700	93,825 [a]	95,442 [a]	0.440	0.440

Note: All results are for household income values in 2017 dollars. REDI, RPME, and MGBE summary statistics are all calculated from performing the procedures on the ACS 2016 and 2017 datasets in Stata/MP version 16.1. REDI values are grand means from 500 repetitions of the method per year (results from a single repetition are in Table 1). CDF interpolation values were calculated using RStudio version 1.2.5019 and R version 3.6.1. Inflation to 2017 dollars was calculated using the R-CPI-U-RS (U.S. Bureau of Labor Statistics 2015, 2020). ACS original estimands of median, mean, and Gini values are calculated directly from the 2016 and 2017 datasets. These values differ substantially from the median values of $58,820 (2016) and $60,336 (2017); the mean values of $85,995 (2016) and $89,294 (2017); and the Gini indices of 0.482 (2016 and 2017) reported in the American Community Survey Brief on Household Income (Guzman 2018).

[a] The mean values in the grand mean specification of RPME and the CDF interpolation methods are definitionally equal to the true mean values of the ACS, because they take the mean of the research dataset as a parameter in their calculations.

Here I present evidence of an application of REDI and its results on the estimation of continuous incomes compared to several other methods for estimating income summary statistics. REDI estimates the median and Gini indices closer (on average) to the real ACS estimands than does any other estimation method. This is true for a single instance of the REDI method (Table 1) and for the grand mean of 500 repetitions per year (Table 2).

Without being constrained by any prior information about the overall mean of the research distribution, the REDI method using 500 repetitions came closer than either of the other two methods (RPME with the Pareto estimator for upper incomes and MGBE average across 10 methods) that make no assumptions about the mean. Further improvements on the method could be made by adding a mean constraint, which might enable even the REDI method using a single draw from the reference dataset to be equally or more accurate.

Overall, the quality of the summary estimates produced by REDI are on par or exceed those of existing methods. REDI has the added advantages of needing no additional assumptions and producing estimates for individual observations. Also, without provision of a known mean for the research dataset, several existing methods would produce far less accurate summary statistics.

Performance of REDI in Multivariate Analyses

Another advantage of the REDI method is that it provides individual values for each observation that can be used in bivariate and multivariate analyses. This is a distinct feature of REDI compared to RPME, MGBE, and CDF interpolation, which can only provide aggregate statistical measures of distributions. Here I provide examples of the use of REDI-calculated

values as a dependent variable and as an independent variable in multivariate regression, followed by a concrete example of workflow from my own research.

REDI-Calculated Values as a Dependent Variable

To test the performance of REDI-estimated values in regression analyses, I performed two simple multivariate regressions, both predicting household income as a function of the respondent's education level, gender, race/ethnicity, disability, and marital status. The regressions predict the REDI-estimated values (Model 1), the original continuous ACS research dataset values (Model 2), and the CPS ASEC reference income values (Model 3).

Table 3 shows that when used as a dependent variable in multivariate regression calculations, REDI-generated income values are comparable to the complete CPS ASEC reference dataset from which they were derived. Most differences in coefficients between the two models are relatively small, especially considering the only factor used to estimate the household income values in the REDI dataset were the artificially-derived income bins from the CPS ASEC reference dataset. One strength of the REDI method here is that no additional information (e.g., about race or gender) from the CPS ASEC data is needed to impute REDI household income values. Still, REDI-generated income values produce regression results of the same direction and significance as the original ACS research dataset and the original CPS ASEC reference dataset.[8] These comparisons show the REDI-generated income variable performs reliably as a dependent variable.

Table 3. Multivariate Regressions Comparing (1) REDI-Computed Income with (2) ACS Continuous Incomes and (3) CPS ASEC Reference as Dependent Variables

Predictor of Household Income	**(1) REDI Estimate** Coefficient (S.E.)	**(2) ACS Continuous** Coefficient (S.E.)	**(3) CPS ASEC Reference** Coefficient (S.E.)
Woman	-6338.28***	-6752.02***	-7950.76***
	(74.90)	(77.74)	(557.29)
Race [Reference: White]			
Black	-18828.58***	-21516.79***	-12902.57***
	(175.83)	(184.32)	(645.58)
Asian	10417.75***	9947.76***	1994.29
	(371.67)	(406.31)	(1454.16)
Hispanic	-11140.51***	-13610.93***	-6655.50***
	(189.53)	(195.43)	(714.41)
Other Race	-3424.50***	-4407.36***	-6547.78***
	(388.83)	(428.55)	(1529.82)
Education [Reference: less than High School]			
High School	-2580.18***	-4559.46***	12788.09***
	(184.38)	(192.74)	(668.23)
Some College	8115.26***	6406.21***	26424.27***
	(179.68)	(195.38)	(689.91)
College	38762.44***	40946.89***	58257.43***

	(225.78)	(246.63)	(883.67)
Graduate / Prof.	60547.50***	70457.83***	86862.51***
	(240.57)	(308.65)	(1271.47)
Married	20959.59***	22960.94***	43890.41***
	(102.58)	(109.54)	(553.84)
Disability	-21127.78***	-22399.6***	-23014.23***
	(143.92)	(181.46)	(627.73)
N	4,980,959	4,980,959	138,757
R-squared	0.1181	0.1214	0.1773

Note: Multivariate regressions demonstrating comparison of (1) covariates predicting REDI-estimated values of household income; (2) covariates predicting ACS original continuous research values of household income; and (3) covariates predicting CPS ASEC reference values of household income. Multivariate linear regressions use survey-weighted values, predicting household income in 2017 standardized dollars from pooled years 2016 and 2017.

REDI-Calculated Income Values as an Independent Variable

Table 4 demonstrates the use of REDI-generated income as an independent variable in a logistic model predicting the odds of changing residence (Model 1). The REDI-generated income variable is compared to the original ACS household income variable before it was artificially binned (Model 2) and to the artificial categories generated (Model 3). The 95 percent confidence interval for the (natural log of the) REDI-generated income (0.772 to 0.894) as a predictor variable overlaps almost with the 95 percent confidence interval for the (natural log of the) original continuous ACS research income predictor (0.773 to 0.930). This indicates the REDI-generated continuous income—derived from the CPS ASEC and the artificially-binned ACS dataset—is performing similarly as original continuous income values as an independent variable in this logistic regression.

Table 4. Multivariate Logistic Regressions Comparing (1) REDI-Computed Income with (2) ACS Continuous Incomes and (3) ACS Categorical as Predictors of Changing Residence

Predictor of Changing Residence	**(1) REDI Estimate** Odds Ratio (S.E.)	**(2) ACS Continuous** Odds Ratio (S.E.)	**(3) ACS Categorical** Odds Ratio (S.E.)
ln(HH income)	.845***	.864**	
	(.035)	(.043)	
HH income category			
$15000 to $24999			.816
			(.271)
$25000 to $34999			1.30
			(.384)
$35000 to $49999			.834
			(.211)
$50000 to $74999			.511*
			(.134)
$75000 to $99999			.365*
			(.118)
$100000 to $149,999			.420*
			(.119)
$150000 to $199,999			.313**

			(.104)
$200000 and above			.475*
			(.167)
Woman	1.07	1.07	1.02
	(.094)	(.092)	(.088)
Race [Reference: White]			
Black	4.36**	4.27**	3.84*
	(2.24)	(2.19)	(2.15)
Asian	3.25	3.24	2.96
	(1.96)	(1.94)	(1.91)
Hispanic	2.07***	1.98**	2.06**
	(.435)	(.418)	(.454)
Other Race	1.19	1.18	1.23
	(.298)	(.297)	(.288)
Education [Reference: less than high school]			
High School	2.27**	2.26**	2.13**
	(.577)	(.571)	(.543)
Some College	2.44**	2.43**	2.26**
	(.702)	(.695)	(.661)
College	1.92*	1.96*	1.95*
	(.560)	(.562)	(.576)
Graduate / Prof.	2.34**	2.25*	2.30*
	(.740)	(.713)	(.751)
Married	.455***	.449***	.508***
	(.072)	(.068)	(.080)
Disability	.762	.776	.712
	(.231)	(.232)	(.225)
N		4,694	

Note: Multivariate logistic regressions use survey-weighted values, predicting changing residence in 2019 among Wyoming respondents. Models compare the covariates of (1) REDI-estimated values (using CPS ASEC reference) of household income; (2) ACS original continuous research values of household income; and (3) ACS categorical values of household income. ACS top-coded incomes are omitted in all models, since they cannot be predicted in model (2).

In other models (see Appendix Table C1), the significance of income as a predictor disappears when homeownership is added as an additional predictor. This is consistent across the three models, for the REDI-generated income and the ACS continuous and categorical values of income. In another test using California data and continuous REDI and ACS incomes (see Appendix Table C2), the pattern holds: income is a significant predictor in both models without homeownership and loses its significance in the model where homeownership is added. I chose Wyoming and California as the least and most populous states, respectively, to demonstrate applicability across a range of dataset sizes. These models show that the REDI-generated income performs very similarly to the original ACS continuous research income, even when a control variable is added that draws almost all variance from the model.

These tests are notable because they demonstrate the method across multiple independent samples, as well as the ability to use the method on subpopulations of the U.S. nationally

representative data for which it was designed. The REDI method can be deployed on smaller statistical areas, although researchers will need to determine the most appropriate reference data for non-national samples. That the REDI-generated income produces such consistent performance as an independent variable, even in smaller samples, provides strong evidence for the reliability of the method across contexts.

Example: Using REDI in Research

The particular value in the REDI method is demonstrated most clearly when a researcher has one or several datasets with a unique variable that cannot be found elsewhere that also has a categorical income variable the researcher wishes to use as a continuous variable in analysis. One might want to convert a categorical to a continuous income variable for one or several reasons: it will simplify multivariate regression; it will enable intersectional analyses; and it will enable longitudinal or cross-dataset comparisons.

As a final example, I describe a situation from my own research. I was interested in comparing performance on a number of health factual knowledge questions in the Health Information National Trends Survey (HINTS). HINTS always asks about household income, but the income categories available to respondents vary over time: in the 2003 survey, the topcode for income was $75,000, whereas in 2014, the topcode was $200,000.[9]

To compare answers to the same questions between 2003 and 2014, I wanted to perform logistic regressions predicting the probability of a correct answer based on demographic variables, including income, gender, race and ethnicity, and education. However, with the income categories changing over time, the regressions would not be truly comparable without converting the binned incomes into continuous incomes. This also allowed for the important step of adjusting for inflation from 2003 to 2014 dollars.

Figure 4 illustrates the steps using the REDI method to convert from categorical to continuous income, before completing further analyses. As depicted in this process, both research datasets (with categorical income data) are transformed using the REDI method and the reference dataset containing continuous income data. This transformation can also include adjustments for inflation to account for different years of data. After these transformations, the two (or more) research datasets can be used in further analyses, including regression analyses with continuous income data.

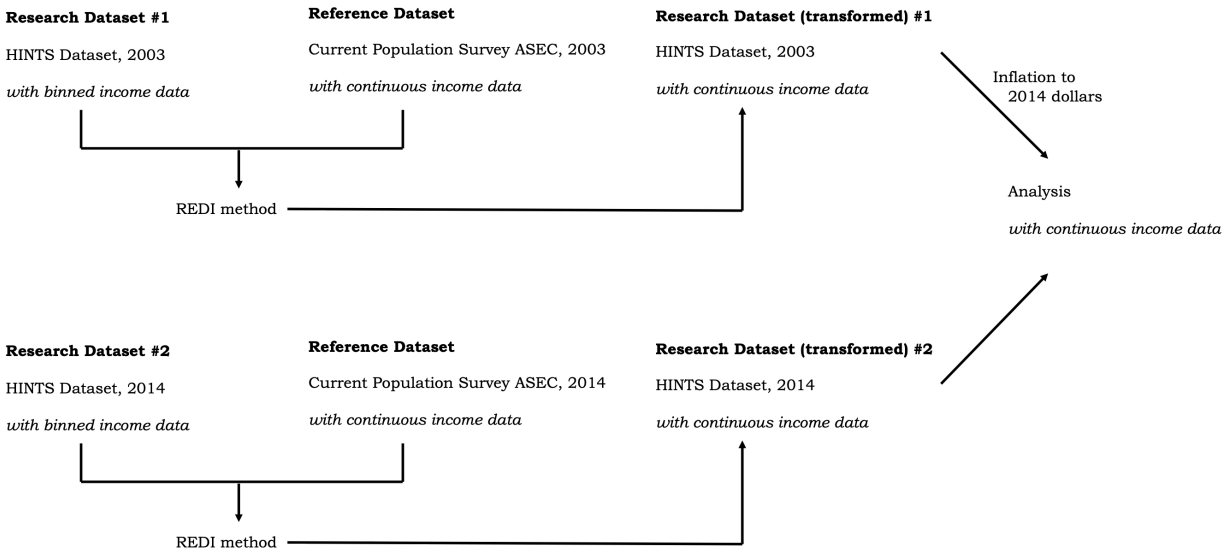

Figure 4. Workflow of Research Using the REDI Method with Two Research Datasets
Note: The CPS ASEC serves as the reference dataset for the REDI method, which is used to impute continuous incomes back to the HINTS research datasets. After continuous incomes have been imputed for the HINTS data, researchers may perform any analyses they like.

In this case, REDI is valuable because the health variables of interest only exist in the HINTS data, but the HINTS data do not include a continuous income variable. Furthermore, the income categories in the two years of HINTS data were incommensurate, making comparisons across years using regressions challenging at best. Therefore, REDI provides a valuable intervention here, allowing for translation of individual income observations into continuous income values while preserving the relationship between health knowledge and income.

DISCUSSION

The method introduced here, Random Empirical Distribution Imputation (REDI), converts categorical numerical data into a continuous distribution. In this way, REDI is valuable for the ease with which it allows researchers to build a distribution of discrete values, not just to compute summary statistics. This is critical when a continuous variable is needed as an input for analysis. To do this, the REDI method needs only one resource: a separate dataset, drawn from a similar population, to use as the reference distribution. REDI performs well with very few assumptions about the research dataset.

Extensions

One possibility for improving the performance of the method in some circumstances

would be to include auxiliary variables in a multiple imputation implementation of REDI. In extending the method, researchers could include auxiliary variables from the research dataset when deciding which income observations to draw from the reference dataset (CPS ASEC). Such auxiliary variables are "not part of the model of substantive interest, but [are] highly correlated with the variables in the substantive model" (Graham 2009:560). For example, researchers who are not including education in their final model could include this demographic variable as a predictor of income when drawing observations from the reference dataset. Such an approach would allow the researcher to predict with greater accuracy which observations in the research dataset should be assigned particular income values from the reference distribution of incomes. Multiple imputation with auxiliary variables also allows the researcher to reduce estimation bias when data are missing not at random (MNAR) (Collins, Schafer, and Kam 2001).[11] Additionally, multiple imputation would allow the researcher to average the estimate for each value chosen for each observation, thereby providing more accurate estimates for the research dataset (with respect to the reference data). Schenker and colleagues (2006) carried out multiple imputation on categorical income data in the National Health Interview Study, which corrected biases and increased efficiency. Such efforts at multiple imputation require auxiliary variables and well-developed models (Mellon and Prosser 2018; Schenker et al. 2006). A core strength of the REDI method is that it can be used for a research dataset that has no overlapping variables with the reference dataset other than the variable being imputed. In our example, we could use REDI to calculate continuous income values for the ACS research dataset even if the only variable in the CPS ASEC were income.

Performance of the REDI method might also be improved by calculating the correlation between the REDI-generated values and the actual incomes from the reference data. This correlation would provide an estimate of the amount of the random measurement error that REDI introduces to the research dataset. Such an estimate could be used as a correction in improving bivariate and multivariate estimates involving a REDI-generated income and other variables.

The REDI method may also be used to calculate binned values other than income. For example, test scores are sometimes reported in ranges or percentiles, and educational researchers may wish to use a more continuous measure. Respondent ages may also be presented as categorical data. Age categories could be converted to continuous age data using a reference dataset.

Many sociological researchers use nationally representative data and are limited by the categorical feature of these data. However, provided appropriate reference data, there is no reason REDI could not be applied to different sampling frames. For example, state- or county-level CPS data could be used as a reference for smaller areas. Indeed, I have provided one such example. The CPS ASEC is a robust reference dataset for the United States for all the reasons outlined here, but REDI may be used with another reference dataset if it is more useful for the research question at hand. Again, the REDI method assumes that both the unit of observation (e.g., family, household, or individual income) and the sample frame and method are the same in both the research and reference datasets. For instance, a researcher may be interested only in very local-level effects or in international data, for which reference distributions in the CPS ASEC are not available. In such cases, REDI is flexible enough that it can be used with more relevant reference data that fit the above criteria, if they are available. The large sample size of the ACS may make it a better choice as a reference for state-level analyses. Additionally, the Small Area Income and Poverty Estimates Program (SAIPE) combines ACS with other data to provide highly accurate subnational estimates at the county- and school-district levels. Because the CPS ASEC also provides state and county FIPS codes, it may be an appropriate reference for smaller scale studies. Future elaborations might consider how researchers working with binned data where no continuous reference dataset exists might further adapt the method.

Another extension is the potential of REDI for de-identifying data. A researcher interested in de-identifying continuous data for public release could use the REDI method, using the dataset of interest as both the research and reference datasets. The continuous measure to be masked (e.g., income) would first be divided into categorical bins. Then, the researcher would draw individual observations from the original continuous set of values to replace these categorical values. Advantages for this method over categorical data are much the same as using REDI in any context: in subsequent analyses, the continuous measure preserves variance. Additionally, if the dataset were large enough, correlated variables (e.g., race and ethnicity or gender in the case of income) could be used in the imputation to preserve more information about the social world without revealing participants' identity. This proposed extension is similar in many ways to the rank-based proximity swap algorithm used by the Census to preserve multivariate correlations and means of subsets (Moore 1996:7).

Advantages and Scope

Overall, REDI has several advantages over alternative methods of converting categorical values to continuous values. Several advantages concern assumptions about data distributions: REDI is nonparametric; it can adjust for inflation; it does not impose artificial lower or upper bounds, making it variance-preserving; and it produces a continuous distribution while still providing individual estimates for observations. These features make REDI particularly valuable when a researcher must reconcile data across time or competing categories. Finally, the resulting values from this form of cold-deck imputation are independent and identically distributed. I discuss each of these advantages in turn.

First, the method is nonparametric, meaning researchers need make no assumptions about the underlying distribution of continuous values. This is important because we know the distributions of values of sociological interest in the United States do not always follow a normal distribution—or even another simply parameterized distribution, in the case of income (Morris and Western 1999; Piketty and Saez 2006). Furthermore, these distributions change over time (Chetty et al. 2014). The REDI method also allows distributions to change over time without issue. As a result, it can easily be adjusted for inflation.

This nonparametric feature also allows REDI to overcome the challenge of top codes without relying on a Pareto distribution (assuming the reference dataset has no artificial top code). There are no artificial lower or upper bounds, except those imposed by the reference dataset. Maintaining a spread of values that reflects real-world conditions enables the method to preserve variance. REDI does not require the researcher to provide the mean or median of categorical brackets to produce accurate results.

Finally, REDI produces an individual estimate for each observation, which is important for further multivariate analyses, including structural equation modeling (Bauldry 2015). While performing this conversion, REDI is area preserving, ensuring the number of observations within each income category remains the same.

One can think about REDI as a within-class random cold-deck imputation method. Cold-deck imputation is a deductive method that selects observations from an external, similar dataset to replace missing data (Hu and Salvucci 2001; Yan 2011). A study by the National Center for Education Statistics recommends deductive imputation methods whenever possible because of their accuracy (Hu and Salvucci 2001:4). REDI is valuable in cases where cold-deck multiple imputation is not appropriate because all variables of interest are not available in the reference

dataset. Unlike classic cold-deck imputation, which uses data collected on the same individuals at an earlier period in time, REDI is random rather than deterministic; within each income bin, there is more than one possible value for imputing each missing case. REDI is a form of random cold-deck imputation, so the same limitations also apply: the performance of the method depends on the quality of the reference data (Hu and Salvucci 2001). However, because REDI draws randomly from the reference dataset, the resulting values escape a common failure of imputation; the researcher can assume that the observations resulting from the REDI method are independent and identically distributed (IID), as long as values are drawn with replacement from the reference dataset.[12]

This makes REDI valuable in specific data contexts. Importantly, REDI is only possible when researchers have an available dataset for drawing the continuous reference distribution.[13] Second, researchers need to have determined that hot-deck multiple imputation is not appropriate or possible because a well-specified model is not yet available or auxiliary variables have not been identified. REDI does not account for correlated predictors of income, but this also means it does not demand them. Finally, the REDI method is appropriate and useful when researchers have categorical data they are interested in converting into continuous data. Alternatively, the method can reconcile different ranges of categorical brackets across different years and even across different datasets and convert them to comparable continuous data.

DATA AND CODE AVAILABILITY

In keeping with recommendations on transparent and open social science (Freese and King 2018), code for replication of the REDI demonstration can be found at the Open Science Framework repository at https://osf.io/cdysr/.

The research dataset (American Community Survey) is available for download from the IPUMS USA website (http://usa.ipums.org). Sample years selected were 2016 and 2017. Variables selected for the basic REDI conversions were YEAR, REPWTP, PERWT, and HHINCOME. The additional variable used in the regression analyses was STATEFIP.

The reference dataset (Current Population Survey ASEC) is available for download from the IPUMS CPS website (http://cps.ipums.org). Sample years selected were 2016 and 2017. Variables selected for the basic REDI conversions were YEAR, ASECWTH, HHINCOME, and PERNUM. Additional variables used in the regression analyses were SEX, RACE, HISPAN, EDUC, MARST, DIFFMOB, OWNERSHIP, and STATEFIP.

The Consumer Price Index retroactive series using current methods with all items (R-CPI-U-RS) can be downloaded as a spreadsheet from the U.S. Bureau of Labor Statistics website (https://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm). Details on how to import and use these data can be found along with the rest of the code at the Open Science Framework repository.

ACKNOWLEDGEMENTS

The author gratefully acknowledges Christof Brandtner, Jeremy Freese, David Grusky, Michelle Jackson, Sarah Burgard, Shelley Correll, Rita King, Armand Rundquist, and anonymous reviewers for helpful comments on earlier drafts of this paper. Minha Khan and Anshu Kripalani provided valuable research assistance. The Santa Clara University Wiegand Advanced Visualization Environment (WAVE) HPC Cluster provided computing resources. This research was supported by funding from a National Science Foundation Graduate Research Fellowship (grant DGE-1147470), the Institute for Research in the Social Sciences and the Clayman Institute for Gender Research at Stanford University, and Santa Clara University.

AUTHOR BIOGRAPHY

Molly M. King is an Assistant Professor of Sociology at Santa Clara University. Her research focuses on the sociology of knowledge and information, inequality, gender, and science. She uses mixed methods, paired with a methodological commitment to open science. Her full

CV is at https://www.mollymking.com/cv.

NOTES

1. For simplicity in comparing the ACS with the CPS ASEC, I treat all responses about a given year in the ACS as income for that calendar year; however, this assumption may affect comparisons of the aggregates (Rothbaum 2015). The fact that REDI-generated income values perform well in multivariate analyses (see Table 3) despite this relaxed assumption demonstrates the robustness of the method.

2. Self-reported income is generally unreliable and underreported. The reliability of self-reported income typically ranges from 71 to 98 percent, with a mean of 86 percent (Marquis, Marquis, and Polich 1986). In other words, if the same individuals are asked about their income at two different points in the survey using two slightly different measures to get at the same question, the results will agree, on average, within about 0.86. Compared to tax or Social Security Administration records, self-reported measures of income tend to be about 70 to 87 percent of official amounts (Coder and Scoon-Rogers 1996; Reardon 2011; Rothbaum 2015). This makes the fairly valid and reliable self-reported earnings measures in the CPS ASEC particularly valuable. The CPS ASEC dataset is also the official source of data for national poverty estimates in the United States (U.S. Census Bureau 2018).

3. These random integer values are generated using the Stata command *runiformint.* This allows me to sample, with replacement, income values from the CPS ASEC dataset according to a uniform distribution where the probability of being sampled is proportional to the representation of that income value in the reference dataset.

4. The Census Bureau used a similar procedure between 1996 and 2010. However, researchers interested in using older data should first investigate the details of the different top-coding procedures across years.

5. Note that these differ notably from the true value of the estimand in the population, as reported by official Census statistics. The Census reports ACS median household income (in 2017 dollars) at $58,820 ($102 margin of error) in 2016 and $60,336 ($86 margin of error) in 2017 (Guzman 2018). Estimates for the CPS ASEC are similar: the median income for 2017 was $58,849, and in 2016 it was $57,230 (Fontenot et al. 2018). In both years, the Census measure of the Gini index was 0.482 (with a margin of error of 0.001) (Guzman 2018).

6. Paul von Hippel and David J. Hunter, email exchange with authors, May 15, 2019. Exact code used is available on the Open Science Framework project repository at https://osf.io/cdysr/. The Gini index was calculated as a single integral using the CDF calculated from the *binsmooth* package:

$$G = 1 - \frac{1}{\mu}\int_0^{\infty}(1 - F(x))^2\,dx,$$

where μ is the mean of the CDF and $F(x)$ is the function describing the CDF. The mean of the CDF is calculated by integrating a function of the CDF:

$$\mu = \int_0^{\infty} 1 - F(x)dx.$$

The calculation of the Gini index is now available directly in the R package *binsmooth*.

7. The median is equal to the point at which the cumulative distribution function equals 0.5. To find this point, I took the inverse of either the monotonic cubic spline or the polygonal CDF function $F(x)$ from the lower limit of 0 to the upper limit of ∞ using the R package *GoFKernel* (Pavia 2018). I then evaluated this inverse to find the value of x at which the CDF equals 0.5.

8. The REDI-generated incomes produce regression coefficients that are larger than those in the multivariate model predicting the CPS ASEC reference incomes because the ACS household incomes are, on average, higher than those in the CPS ASEC dataset. As a result, REDI draws more income observations from higher income brackets in the reference dataset, raising the average income of the overall REDI-generated dataset. For example, the mean income among women across both 2016 and 2017 was $72,913 in the CPS ASEC reference dataset and $94,849 in the research ACS dataset. The resulting mean income for women generated by REDI was $90,769. This characteristic could be especially valuable in cases where researchers believe a reference dataset may be more reliable than their research dataset for a continuous variable.

9. Datasets from the Pew Research Center provide another good example of a situation where REDI may be useful. Pew always asks about the respondent's family income, but the categories provided to the respondent change over time.

10. Data are considered to be *missing not at random* (MNAR) when missingness on the dependent variable depends on that characteristic itself. The primary reason for using auxiliary variables in multiple (rather than single regression) imputation is to restore the variance in errors

that is lost from basing the imputed income value on a single regression equation (Graham 2009:556–57). To some degree, the lost error is adjusted by the multiple random draws that REDI executes from the reference dataset; however, random error will be reduced further if one adds auxiliary variables. Multiple imputation counteracts this reduction by restoring this lost variance.

11. I thank Kazuo Yamaguchi for drawing my attention to this.

12. As in the example described here, researchers could use the CPS ASEC for most applications requiring U.S. nationally representative income data.

REFERENCES

Allison, Paul D. 2000. “Multiple Imputation for Missing Data.” *Sociological Methods & Research* 28:301–09.

Allison, Paul D. 2002. “Multiple Imputation: Basics.” Pp. 27–49 in *Missing Data*. Thousand Oaks, CA: Sage Publications.

Bailey, Stanley R., Aliya Saperstein, and Andrew M. Penner. 2014. “Race, Color, and Income Inequality across the Americas.” *Demographic Research* 31:735–56.

Bauldry, Shawn. 2015. “Structural Equation Modeling.” Pp. 615-620 in *International Encyclopedia of the Social and Behavioral Sciences*, 2nd ed., vol. 23, edited by James D. Wright. Oxford: Elsevier.

Bhat, Chandra R. 1994. “Imputing a Continuous Income Variable from Grouped and Missing Income Observations.” *Economics Letters* 46:311–19.

Blanchet, Thomas, Bertrand Garbinti, Jonathan Goupille-Lebret, and Clara Martínez-Toledano. 2018. “Applying Generalized Pareto Curves to Inequality Analysis.” *AEA Papers and Proceedings* 108:114–18.

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2019. “Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch.” *Journal of Political Economy* 127:2143–85.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014. “Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility.” *American Economic Review* 104:141–47.

Coder, John, and Lydia Scoon-Rogers. 1996. “Evaluating the Quality of Income Data Collected in the Annual Supplement to the March Current Population Survey and the Survey of Income and Program Participation.” Working paper, Housing and Household Economic Statistics Division, Bureau of the Census, Department of Commerce (https://www.census.gov/sipp/workpapr/wp215.pdf).

Collins, Linda M., Joseph L. Schafer, and Chi-Ming Kam. 2001. “A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures.” *Psychological Methods* 6:330–51.

Cox, Nicholas J. 1998. “DISTPLOT: Stata Module to Generate Distribution Function Plot.” Boston College Department of Economics (https://econpapers.repec.org/RePEc:boc:bocode:s337502).

Donnelly, Michael J., and Grigore Pop-Eleches. 2018. “Income Measures in Cross-National Surveys: Problems and Solutions.” *Political Science Research and Methods* 6:355–63.

Evans, James A., and Jacob G. Foster. 2019. “Computation and the Sociological Imagination.” *Contexts* 18:10–15.

Fixler, Dennis, Marina Gindelsky, and David Johnson. 2019. “Improving the Measure of the Distribution of Personal Income.” Bureau of Economic Analysis (https://www.bea.gov/research/papers/2019/improving-measure-distribution-personal-income).

Fontenot, Kayla, Jessica Semega, and Melissa Kollar. 2018. “Income and Poverty in the United States: 2017.” *Current Population Reports*. Washington, DC: U.S. Government Printing Office.

Francisco, Carol A., and Wayne A. Fuller. 2008. “Quantile Estimation with a Complex Survey Design.” *The Annals of Statistics* 19:454–69.

Freese, Jeremy, and Molly M. King. 2018. “Institutionalizing Transparency.” *Socius: Sociological Research for a Dynamic World* 4:1–7.

Gelman, Andrew, and Jennifer Hill. 2006. “Missing-Data Imputation.” Pp. 529–44 in *Data Analysis Using Regression and Multilevel/Hierarchical Models*, edited by A. Gelman and J. Hill. New York: Cambridge University Press.

Goerg, Sebastian J., and Johannes Kaiser. 2009. “Nonparametric Testing of Distributions: The Epps–Singleton Two-Sample Test Using the Empirical Characteristic Function.” *The Stata Journal* 9:454–65.

Graham, John W. 2009. “Missing Data Analysis: Making It Work in the Real World.” *Annual Review of Psychology* 60:549–76.

Guzman, Gloria G. 2018. “Household Income: 2017.” American Community Survey Briefs, ACSBR/17-0: U.S. Census Bureau (https://www.census.gov/content/dam/Census/library/publications/2018/acs/acsbr17-01.pdf).

Henson, Mary F. 1967. *Trends in the Income of Families and Persons in the United States, 1947–1964*. Washingtion, DC: U.S. Department of Commerce, Bureau of the Census (https://books.google.com/books?id=tuW2AAAAIAAJ).

Hout, Michael. 2004. “Getting the Most Out of the GSS Income Measures.” GSS Methodological Report #101. Berkeley, CA, University of California, Berkeley Survey Research Center.

Hu, Ming-xiu, and Sameena Salvucci. 2001. “A Study of Imputation Algorithms.” Working paper 2001-17. Washington, DC: U.S. Department of Education, National Center for Education Statistics,

Hunter, David J., and McKalie Drown. 2020. “binsmooth: Generate PDFs and CDFs from Binned Data. R Package Version 0.2.2.” *Comprehensive R Archive Network* (https://cran.r-project.org/web/packages/binsmooth/binsmooth.pdf).

Jargowsky, Paul A., and Christopher A. Wheeler. 2018. “Estimating Income Statistics from Grouped Data: Mean-Constrained Integration over Brackets.” *Sociological Methodology* 48:337–74.

Ligon, Ethan. 1989. “The Development and Use of a Consistent Income Measure for the General Social Survey.” *GSS Methodological Report #64*.

Marquis, Kent H., M. Susan Marquis, and J. Michael Polich. 1986. “Response Bias and Reliability in Sensitive Topic Surveys.” *Journal of the American Statistical Association* 81:381–89.

McDonald, James B. 1984. “Some Generalized Functions for the Size Distribution of Income.” *Econometrica* 52:647–63.

McDonald, James B., and Michael R. Ransom. 1979. “Alternative Parameter Estimators Based Upon Grouped Data.” *Communications in Statistics – Theory and Methods* 8:899–917.

Mellon, Jonathan, and Christopher Prosser. 2018. “Constructing Continuous Household Income Measurement on the British Election Study Internet Panel.” *SSRN Electronic Journal*.

Minnesota Population Center. 2018. “CPS Income and Tax Variables User’s Note: Missing Cases, N.I.U. Cases, Top Codes And Bottom Codes.” *Current Population Survey Documentation*. Minneapolis: University of Minnesota (https://cps.ipums.org/cps/inctaxcodes.shtml#topcodes).

Moore, Richard A. 1996. “Controlled Data Swapping Techniques for Masking Public Use Microdata Sets.” Center for Disclosure Avoidance Research Working Papers, U.S. Census Bureau.

Morris, Martina, and Bruce Western. 1999. “Inequality in Earnings at the Close of the Twentieth Century.” *Annual Review of Sociology* 25:623–57.

Pavia, Jose. 2018. “GoFKernel: Testing Goodness-of-Fit with the Kernel Density Estimator.” (https://www.rdocumentation.org/packages/GoFKernel/versions/2.1-1).

Piketty, Thomas. 2014. *Capital in the Twenty-First Century*. Cambridge, MA: The Belknap Press of Harvard University Press.

Piketty, Thomas, and Emmanuel Saez. 2006. “The Evolution of Top Incomes: A Historical and International Perspective.” *American Economic Review* 96:200–205.

Piketty, Thomas, and Emmanuel Saez. 2014. “Inequality in the Long Run.” *Science* 344:838.

Reardon, Sean F. 2011. “Online Appendix 5: The Widening Academic-Achievement Gap between the Rich and the Poor: New Evidence and Possible Explanations.” In *Whither Opportunity?* edited by G. J. Duncan and R. J. Murnane. New York: Russell Sage Foundation.

Roth, Philip L. 1994. “Missing Data: A Conceptual Review for Applied Psychologists.” *Personnel Psychology* 47:537–60.

Rothbaum, Jonathan L. 2015. “Comparing Income Aggregates: How do the CPS and ACS Match the National Income and Product Accounts, 2007–2012.” Social, Economic, and Housing Statistics Division Working Papers Working Paper 2015-01. Washington, DC: U.S. Census Bureau.

Schenker, Nathaniel, Trivellore E. Raghunathan, Pei Lu Chiu, Diane M. Makuc, Guangyu Zhang, and Alan J. Cohen. 2006. “Multiple Imputation of Missing Income Data in the National Health Interview Survey.” *Journal of the American Statistical Association* 101:924–33.

Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler, and Matthew Sobek. 2021. “IPUMS USA: Version 11.0 [dataset].” Minneapolis, MN: IPUMS (https://doi.org/10.18128/D010.V11.0).

U.S. Bureau of Labor Statistics. 2015. “Consumer Expenditure Surveys.” Washington, DC: United States Department of Labor (https://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm).

U.S. Bureau of Labor Statistics. 2020. “R-CPI-U-RS Homepage.” *Consumer Price Index* (https://www.bls.gov/cpi/research-series/r-cpi-u-rs-home.htm).

U.S. Census Bureau. 2018. “How the Census Bureau Measures Poverty.” Washington, DC (www.census.gov/hhes/www/poverty/about/overview/measure.html).

U.S. Census Bureau and U.S. Bureau of Labor Statistics. 2018. “Current Population Survey March Annual Social and Economic Supplement (CPS ASEC).” Minneapolis, MN (https://cps.ipums.org/cps/asec_sample_notes.shtml).

UCLA: Statistical Consulting Group. 2020. “How Can I Do a *t*-Test with Survey Data?” *Stata FAQ* (https://stats.idre.ucla.edu/stata/faq/how-can-i-do-a-t-test-with-survey-data/).

von Hippel, Paul T., David J. Hunter, and McKalie Drown. 2017. “Better Estimates from Binned Income Data: Interpolated CDFs and Mean-Matching.” *Sociological Science* 4:641–55.

von Hippel, Paul T., and Daniel A. Powers. 2015. “RPME: Stata Module to Compute Robust Pareto Midpoint Estimator.” S457962, Boston College Department of Economics, revised

March 25, 2018, Statistical Software Components (https://ideas.repec.org/c/boc/bocode/s457962.html).

von Hippel, Paul T., Samuel V. Scarpino, and Igor Holas. 2016. "Robust Estimation of Inequality from Binned Incomes." *Sociological Methodology* 46:212–52.

Wang, Weidong, Guihua Xie, and Lingxin Hao. 2014. "Rural Panel Surveys in Developing Countries: A Selective Review." *Economic and Political Studies* 2:151–77.

Yan, Ting. 2011. "Hot-Deck Imputation." Pp. 316–17 in *Encyclopedia of Survey Research Methods*, edited by P. J. Lavrakas. Thousand Oaks, CA: Sage Publications.

APPENDIX A: SELECTED DATASETS USING BINNED INCOME

Here is an incomplete list of popular publicly available datasets that provide binned income (either personal, family, or household):

- Annual American Time Use Survey (ATUS)
- National Health Interview Series (NHIS) prior to 2019
- General Social Survey (GSS)
- Pew Research Center publicly available datasets
- RAND American Life Panel
- Health Information National Trends Survey (HINTS)
- USC Understanding America Study (UAS)
- FINRA National Financial Capability Study (NFCS)
- Annenberg National Health Communication Survey
- American National Election Studies (ANES)
- National Health and Nutrition Examination Survey (NHANES)

APPENDIX B: DIAGNOSTIC: COMPARING BINNED DISTRIBUTIONS

Table B1. Household Total Money Income Categories Used for Artificial Categorization of ACS Dataset in Proof-of-Concept Demonstration

Category	2016 % Population	2017 % Population
Under $15,000	11.0	10.7
$15,000 to $24,999	9.5	9.6
$25,000 to $34,999	9.3	9.2
$35,000 to $49,999	12.6	12.3
$50,000 to $74,999	16.9	16.5
$75,000 to $99,999	12.2	12.5
$100,000 to $149,999	14.4	14.5
$150,000 to $199,999	6.8	7.0
$200,000 and over	7.2	7.7

Note: Income categories come from Census Bureau summary tables of CPS ASEC (Fontenot, Semega, and Kollar 2018:27). Bounds represent the total money income of the noted range in 2017 dollars, earned by households during a given year. Percentages indicate the proportion of the population (as estimated by the March CPS ASEC of the following year) with household incomes in that range during either 2016 or 2017.

As a diagnostic measure prior to any analyses, researchers should compare the distribution of the binned data in the research dataset to the reference dataset. In the case of the ACS and the CPS ASEC, we would expect the distributions to be nearly identical because they are both very large, nationally representative datasets with very similar median incomes. Figure B1 illustrates the overlap between the two datasets of the percentage of the population in each income bin. This figure illustrates the reference distribution (with continuous values) compared to the research distribution, both using the same binned categories of household incomes (transformed into bins for illustration).

Figure B1 shows there is notable variation in the distributions of the two datasets in 2016 and 2017, particularly in the middle of the distribution. However, rather than being cause for concern, this dissimilarity of overlap strengthens my proof of concept: because REDI relies on repeated samples with replacement from the reference dataset, a mismatch in the exact shape of the distributions need not be a cause for concern. In fact, as seen in Figure 3, the REDI-computed cumulative distribution function is a near-identical match to that of the original ACS data, despite being drawn from the CPS ASEC reference dataset. It is still a valuable descriptive exercise to compare the distributions, but the precise degree of overlap is not important for REDI to perform well. In sampling from the reference dataset based on the size of the bins in the

research dataset, the method preserves the relative composition of the research dataset.

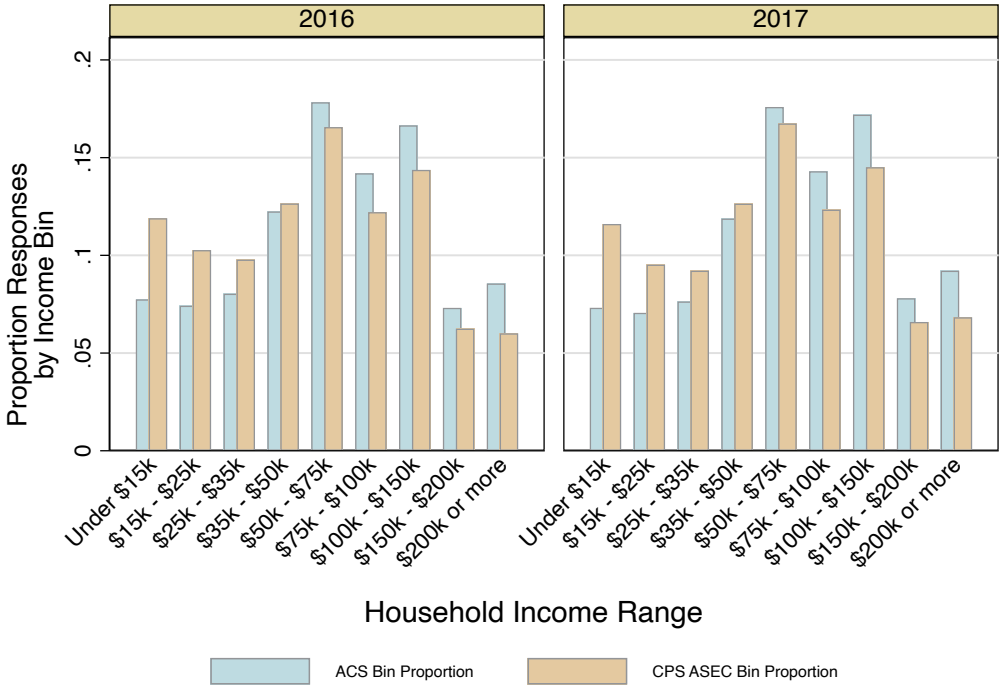

Figure B1. Overlap of Research and Reference Distributions by Year, Segmented into Bins
Note: Bins correspond to categories used by the Census Bureau (see Appendix Table B1).

APPENDIX C: MULTIVARIATE REGRESSIONS DEMONSTRATING REDI-GENERATED INCOMES AS THE INDEPENDENT VARIABLE

Appendix C presents extended multivariate logistic regressions for state subsets of the data, using income to predict change in residence. All models contain gender, race/ethnicity, educational attainment, disability, and marital status. Compared to Table 4 for Wyoming, Table C1 adds homeownership as a control. Table C2 presents similar regressions for data from California, comparing REDI-generated incomes to the original ACS continuous research income values for models with and without homeownership controls. This appendix demonstrates that significance disappears in the same fashion for the REDI-generated income variable as for the original income variables.

Table C1. Extended Multivariate Logistic Regressions Comparing (1) REDI-Computed Income with (2) ACS Continuous Incomes and (3) ACS Categorical as Predictors of Changing Residence among 2019 Wyoming Respondents, Controlling for Homeownership

Predictor of Changing Residence	**REDI Estimate** Odds Ratio (S.E.)	**ACS Continuous** Odds Ratio (S.E.)	**ACS Categorical** Odds Ratio (S.E.)
ln(HH income)	.917	.942	
	(.044)	(.039)	
HH income category			
$15,000 to $24,999			.864
			(.297)
$25,000 to $34,999			1.73
			(.548)
$35,000 to $49,999			1.19
			(.309)
$50,000 to $74,999			.779
			(.217)
$75,000 to $99,999			.623
			(.209)
$100,000 to $149,999			.761
			(.237)
$150,000 to $199,999			.716
			(.271)
$200,000 and above			1.02
			(.339)
Woman	1.09	1.10	1.07
	(.098)	(.093)	(.098)
Race [Reference: White]			
Black	2.87	2.86	2.75
	(1.64)	(2.16)	(1.75)
Asian	2.75	2.74	2.59
	(1.88)	(1.92)	(1.84)
Hispanic	1.71*	1.69*	1.73*
	(.366)	(.416)	(.379)
Other Race	1.16	1.12	1.16
	(.287)	(.295)	(.291)
Education [Reference: less than high school]			

High School	1.95*	1.95*	1.94*
	(.562)	(.566)	(.562)
Some College	2.15*	2.16*	2.13*
	(.694)	(.691)	(.688)
College	1.84	1.86	1.87
	(.607)	(.572)	(.622)
Graduate / Prof.	2.11*	2.08*	2.10*
	(.756)	(.711)	(.765)
Married	.617**	.608**	.645*
	(.107)	(.068)	(.112)
Disability	.729	.725	.724
	(.201)	(.231)	(.207)
Homeowner	.230***	.226***	.233***
	(.036)	(.231)	(.038)
N		4,694	

Note: Multivariate regression models include (1) the covariate of REDI-estimated continuous values of household income; (2) the covariate of original ACS continuous research values of household income; or (3) the artificially-generated ACS household income categories. Multivariate logistic regressions use survey-weighted values. Multivariate logistic regressions use survey-weighted values, predicting changing residence in 2019 among Wyoming respondents.

Table C2. Extended Multivariate Logistic Regressions Demonstrating Comparison of Predictors of Changing Residence among 2019 California Respondents, with or without Controls for Homeownership

	(1) REDI Estimate		**(2) ACS Continuous**	
Predictor of	Odds Ratio	Odds Ratio	Odds Ratio	Odds Ratio
Changing Residence	(S.E.)	(S.E.)	(S.E.)	(S.E.)
ln(HH income)	.946***	1.01	.937***	1.00
	(.005)	(.006)	(.005)	(.007)
Woman	.937***	.949***	.937***	.948***
	(.013)	(.013)	(.013)	(.013)
Race [Reference: White]				
Black	1.05	.878**	1.04	.875**
	(.049)	(.042)	(.049)	(.042)
Asian	1.06*	1.023	1.05*	1.02
	(.023)	(.022)	(.023)	(.023)
Hispanic	.850***	.734***	.851***	.734***
	(.020)	(.017)	(.019)	(.018)
Other Race	1.28***	1.16**	1.28***	1.16**
	(.057)	(.053)	(.057)	(.053)
Education [Reference: less than high school]				
High School	1.38***	1.42***	1.38***	1.43***
	(.046)	(.047)	(.046)	(.048)
Some College	1.48***	1.56***	1.48***	1.56***
	(.043)	(.044)	(.043)	(.045)
College	2.03***	2.10***	2.03***	2.12***
	(.068)	(.070)	(.069)	(.072)
Graduate / Prof.	2.07***	2.23***	2.08***	2.24***
	(.079)	(.085)	(.080)	(.087)
Married	.604***	.710***	.607***	.712***
	(.012)	(.013)	(.012)	(.014)
Disability	.679***	.714***	.680***	.708***
	(.025)	(.026)	(.025)	(.026)
Homeowner		.321***		.322***
		(.007)		(.008)
N		302,150		

Note: Model includes (1) the covariate of REDI-estimated values (using CPS ASEC reference) of household income

and (2) the covariate of original ACS continuous research dataset values of household income in multivariate regression analyses. Multivariate logistic regressions use survey-weighted values, predicting changing residence in 2019 among California respondents.