SAMPLING BIAS IN *FIRMING UP INEQUALITY*

MARSHALL I. STEINBAUM*

ABSTRACT. *Firming Up Inequality* [Song, Price, Guvenen, and Bloom (2015)] estimates the extent to which increasing inequality in the distribution of earnings from labor is caused by rising within-firm vs. between-firm inequality. But its statistical sampling from the Social Security Master Earnings File (MEF) is biased in a way that reduces inequality in the sample relative to the population, artificially limiting the scale of the phenomenon the paper purports to investigate. This note explains the two biases the authors introduce into the paper: first, they draw one 1/16th random sample of the MEF, which has become increasingly skewed over the period they study. Second, they winsorize individual income at the top 0.001%, under-representing earnings of the highest earners. Both procedures pose a particular danger to studying highly unequal distributions. Simulation results of their sampling technique show that the two biases are substantial and relevant to empirical research on inequality using both administrative and survey data.

June 15, 2015

* Research Economist, Washington Center for Equitable Growth. Many thanks to Enghin Atalay, John Schmitt, and Branko Milanovic for helpful advice.

1. Introduction

Song, Price, Guvenen, and Bloom (2015) studies the rise in labor market earnings inequality in the Social Security Administration's Master Earnings File (MEF). The paper's conclusion, that rising inequality in the individual earnings distribution is driven entirely by rising inequality in the distribution of firm average wages, has been subjected to several noteworthy critiques (see Mishel (2015)). This note raises a different concern: the authors' sampling from the MEF introduces statistical bias against rising inequality into their estimates.

There are two distinct reasons why the paper's sampling procedure is biased: first, they analyze a single 1/16th sample of the MEF. Second, they truncate or "winsorize" earnings at the 99.999th percentile, which means that anyone in the top 0.001% of earners is reported as receiving the earnings at the 0.001% cutoff, rather than what they actually earn in excess of that cutoff. The authors write "it is both challenging to analyze the universe of all workers given the substantial sample size and it is not necessary given that the results are unlikely to change if we were to work with a subsample." On the contrary, the results are very likely to change under their procedure, as the simulations reported in this note show.

In combination, the two sources of bias reduce the estimate of the top 1% share by 1-2 percentage points if the true top 1% share is about 20% (as reported by Alvaredo, Atkinson, Piketty, and Saez (2015)) and the top 0.01% share by 2-3 percentage points if the true top 0.01% share is about 6%. The critical point to make is that it's because of rising inequality that these procedures introduce bias. If inequality remained where it was at the start of this study in 1980, the procedures wouldn't be biased substantially.

The reason why inference on a single 1/16th sample of a fat-tailed distribution is biased is precisely that drawing a sample at random from the population is likely to underweight the tails, and as inequality in the tails increases, the consequences of that small-sample bias worsen. Consider a population of sixteen people, one of whom earns all of the income– the other fifteen earn nothing. Drawing one person at random from that population and estimating the mean wage based on that draw will yield a downward-biased estimate of the true mean. In fifteen out of sixteen samples (probabilistically), the inference on the mean

wage is zero. In one out of sixteen samples, the inference is greater than the truth. The key point is that the small-sample bias gets worse the larger is the income of the one person with positive earnings.

The reason why winsorizing top incomes introduces bias is obvious: it reduces the largest outliers to some limit on income. The higher is tail inequality, the larger the share of total income above any limit. It has been widely noted that tail inequality is fractal in nature, meaning that the pattern stays roughly constant no matter how far up the distribution you look. Within the top 0.001% of earners, earnings are distributed approximately as unequally as they are within the top 0.01, 0.1, or 1%.[1] Truncating the earnings distribution anywhere thus introduces bias.

In public commentary responding to criticism of the paper, Professor Guvenen wrote "my view is that focusing too much on 350 people (CEOs) and brushing aside what is happening to 300 million people is not good economics."[2] But that misses the point of rising inequality: a tiny fraction of earners accounts for a large share of total earnings. That is why Alvaredo, Atkinson, Piketty, and Saez (2015) is a crucial resource for studying inequality, and why properly-conducted survey-based studies of inequality employ some means of over-weighting responses from the wealthy.[3]

2. Simulation

Song, Price, Guvenen, and Bloom (2015) draw 1/16th of the individuals in the MEF based on a cryptographic transformation of individuals' Social Security Numbers. They then create an individual-earnings distribution year-by-year and calculate the percent income growth at each percentile between 1982 and 2012. They link the workers at each percentile to the firms where they work and calculate growth of average wages at those firms over the same period. The conclusion of the paper is that growth in firm average wages follows the same pattern

[1] In fact, Alvaredo, Atkinson, Piketty, and Saez (2015) reports that inequality increases as you go up the distribution: the top 1% earns 20% of the income, but the top 0.01% earns 6%. Whereas if inequality were perfectly replicated at each point in the distribution's tail, the top 0.01% would earn 4% if the top 1% earns 20%.

[2] See Guvenen's tweet on June 2, 2015 timestamped 9:34 AM. https://twitter.com/fatihguvenen/status/605774409639419905.

[3] See Eckerstorfer, Halak, Kapeller, Schutz, Springholz, and Wildauer (2015) for an illuminating discussion.

across the centiles of the firm distribution of widening inequality as exhibited by the centiles of the individual distribution.[4]

In order to replicate their procedure, I created synthetic populations by drawing from a Pareto distribution and varying the shape parameter, thus varying the degree of tail inequality. I calculate the true top 1% and top 0.01% share for the population, then draw a 1/16th sample of that population and estimate the top 1% and top 0.01% share using the sample. I winsorize the top 0.001% of either the population (i.e, before drawing the 1/16th sample) or of the sample, since the text of the article is unclear exactly which order this is done.[5] For each Pareto shape parameter setting, I draw 100 different populations and one 1/16 sample from each population. I calculate the share estimates from the sample and then the bias of those estimates relative to the truth. I then average across the 100 bias measures to estimate the bias as a function of true tail inequality. I repeat the procedure over a fixed grid of 50 shape parameter settings.[6]

The results are shown in figures 2.1-2.4. The bias is larger the further up the income distribution you look (ie, estimates of the top 0.01% share are biased more than estimates of the top 1% share), and they are larger the more unequal is the distribution. Practically speaking, small-sample bias alone probably doesn't matter all that much for the level of inequality present in the MEF, but in combination with winsorizing the top 0.001% of earnings, the bias does matter.

3. Implications for Song, Price, Guvenen, and Bloom (2015)

Section 2 establishes that the sampling procedure in Song, Price, Guvenen, and Bloom (2015) would be a biased estimate of top income shares. But those authors aren't estimating top income shares per se. They are attributing unequal income growth over time to within-firm vs. between-firm components. Nonetheless, the bias is quite relevant: if the procedure

[4]Again, the firms are ranked depending on which firms individuals at each level of the individual earnings distribution work.

[5]The text says the variables are winsorized "immediately before analysis," which suggests that it is the sample that is winsorized. Nonetheless, I report the simulation results for both a winsorized population and a winsorized sample. The results for a winsorized population are reported in Appendix A.

[6]The shape parameter varies from 1 to ∞ in the simulations.

samples the top of the income distribution insufficiently, it's leaving out the individuals with the largest income growth over the study period, and hence, "their" firms as well.

In a presentation slide deck dated May 2015, Professor Guvenen wrote of the results from this project:

> "The pay of workers in the top 0.01% increased by 500% from 1982 to 2012. The pay gap between these top earners and the average employee at the same firm has increased by only 20% during the same time. Alternatively put: the rise in CEO to average employee wage ratio explains a very small part of rising inequality. The bulk of the action comes between firms."[7]

However, the biased sampling procedure under-represents the top 0.01% of the population about which this claim is made. Winsorizing the top 0.001% eliminates distinguishing information about the top 10% of that top 0.01%. If the population in the MEF (under the authors' restrictions) is 100 million, the top 0.01% is 10,000 individuals and the winsorized top 0.001% is 1,000 individuals. It is likely that Song, Price, Guvenen, and Bloom (2015) simply don't have the highest-paid CEOs represented in the sample they analyze.

Rectifying Bias. Song, Price, Guvenen, and Bloom (2015) could improve their sampling fairly straightforwardly. First of all, if they're not required to do so, there's no reason to winsorize any variables since the data is confidential. If confidentiality remains an issue, better than simply truncating top earnings would be to estimate them using an assumption about the shape of the income distribution. Eckerstorfer, Halak, Kapeller, Schutz, Springholz, and Wildauer (2015) describe a procedure for doing that, in the case of a national wealth survey in which the sample is a tiny fraction of the population. In summary, the idea is to fit a Pareto distribution to the top section of the data, estimate the shape parameter, and then fill in the top, truncated or thinly-sampled tail by simulating synthetic observations using the fitted distribution. The problem in this case would then be deciding at which firms the simulated observations work. Another possibility that would be simpler though less accurate is simply to give the winsorized top earners the sample mean of all of their earnings, rather than the minimum. In that case, they would still be linked to the firms where they work.

[7] Guvenen (2015).

Since Song, Price, Guvenen, and Bloom (2015) do in fact have, in effect, the entire population, the more straightforward solution is to either draw a larger sample than 1/16th or to draw multiple smaller samples and conduct the analysis on each of them.

4. Conclusion

The sampling in Song, Price, Guvenen, and Bloom (2015) is biased for at least two reasons, in such a way that it casts doubt on the results in the paper and the authors' public statements about their results. There are several implementable ways to address these issues.

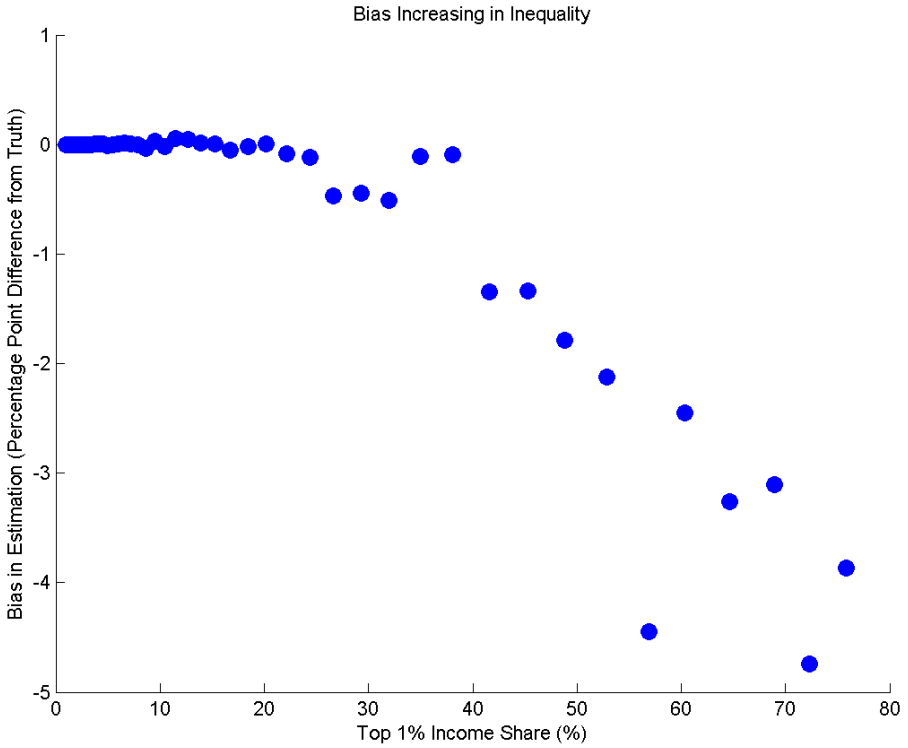

FIGURE 2.1. The x-axis in this chart shows the top 1% share in the population, and the y-axis shows the bias between the sample-derived top 1% share and the truth. For instance, if the true top 1% share were 50%, then a bias of -2 would mean the estimated share is 48%. Without the effect of winsorizing (only the small-sample bias is in effect), the impact isn't really felt until inequality is such that the top 1% share reaches 30%. In other words, if the current level of income inequality is that the top 1% share is 20%, a 1/16 sample of the MEF would likely be sufficient to estimate the top 1% share without introducing bias.

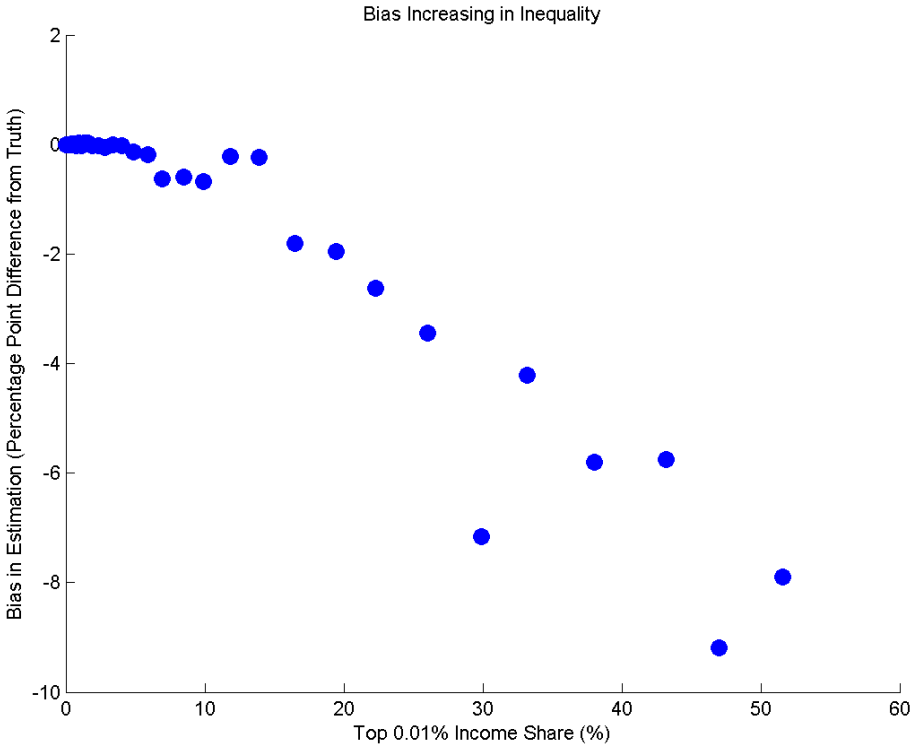

FIGURE 2.2. Since the best estimate of the top 0.01% share is currently 6%, small-sample bias alone is similarly unlikely to afflict the Song, Price, Guvenen, and Bloom (2015) procedure.

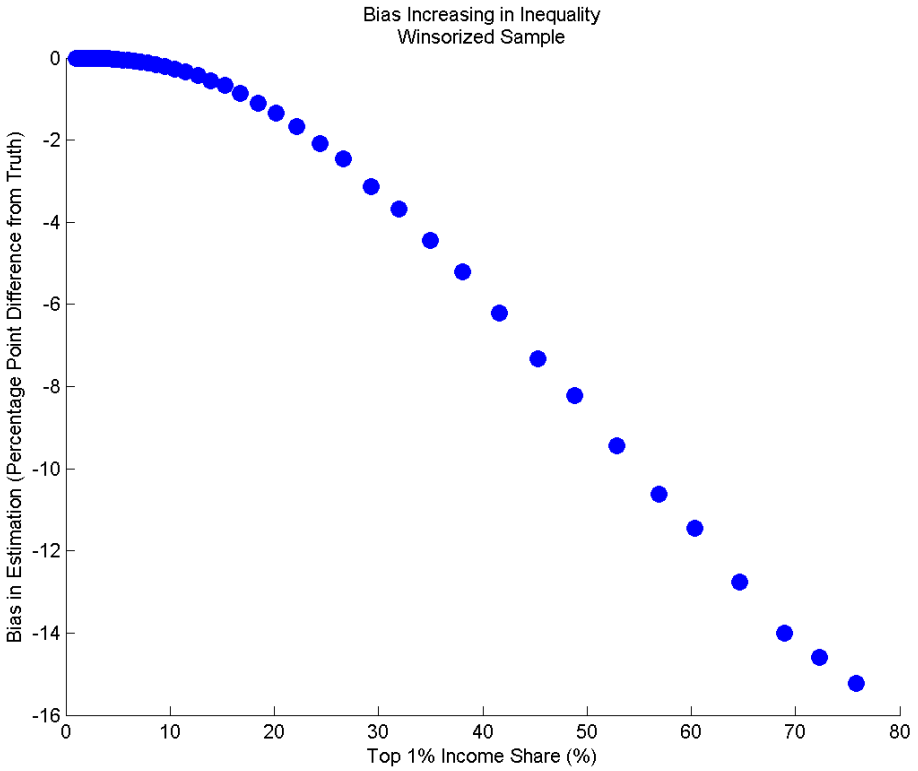

FIGURE 2.3. The story changes when small-sample bias is combined with winsorizing at the top 0.001%. In that case, bias starts to matter when the top 1% income share exceeds 10%, and under current conditions is somewhere between 1-2 percentage points. Note that since the top 1% share was 8% in 1980, the start of the period *Firming Up Inequality* covers, the bias wouldn't have mattered if inequality hadn't increased.

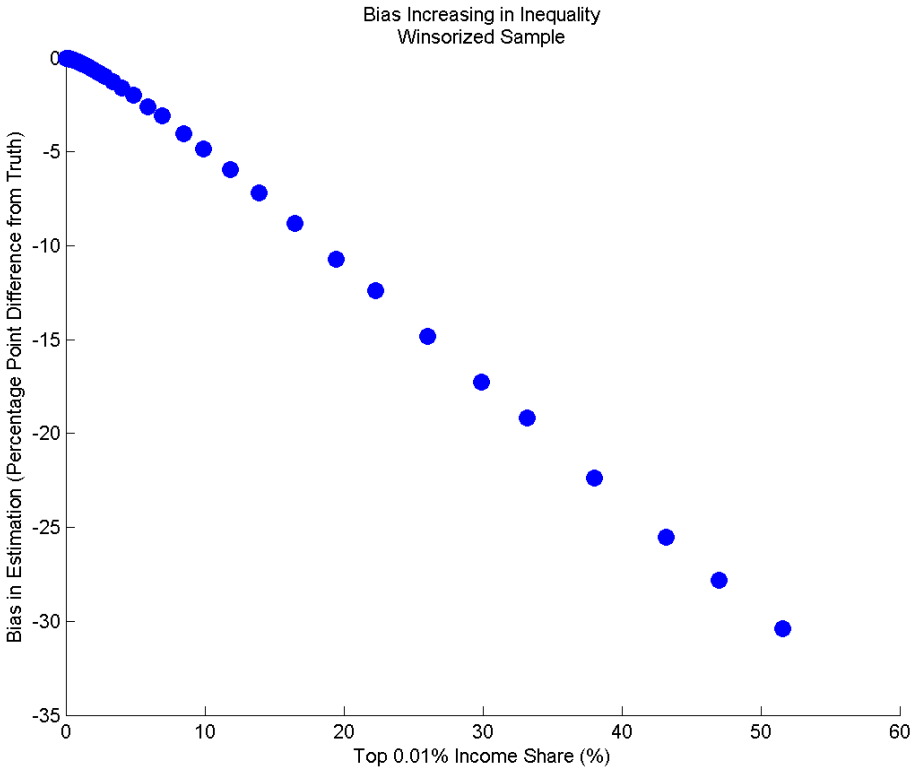

FIGURE 2.4. The bias is more severe looking at the top 0.01% share. Combining small sample and winsorizing, if the top 0.01% share is 6%, the Song, Price, Guvenen, and Bloom (2015) procedure biases it by about 3 percentage points.

References

Alvaredo, F., A. B. Atkinson, T. Piketty, and E. Saez (2015): "The World Top Incomes Database," .

Eckerstorfer, P., J. Halak, J. Kapeller, B. Schutz, F. Springholz, and R. Wildauer (2015): "Correcting for the Missing Rich: An Application to Wealth Survey Data," *Review of Income and Wealth*.

Guvenen, F. (2015): "Income Inequality and Income Risk: Old Myths vs. New Facts," JDP Lecture Series on "Dilemmas in Inequality".

Mishel, L. (2015): "New Research Does Not Provide Any Reason to Doubt that CEO Pay Fueled Top 1

Song, J., D. J. Price, F. Guvenen, and N. Bloom (2015): "Firming Up Inequality," *NBER Working Paper 21199*.

Appendix A. Results from Winsorizing the Population

It matters to some extent which order the sampling and winsorizing is done. Here I report the results from winsorizing the population at the 0.001%, then drawing a 1/16th sample.

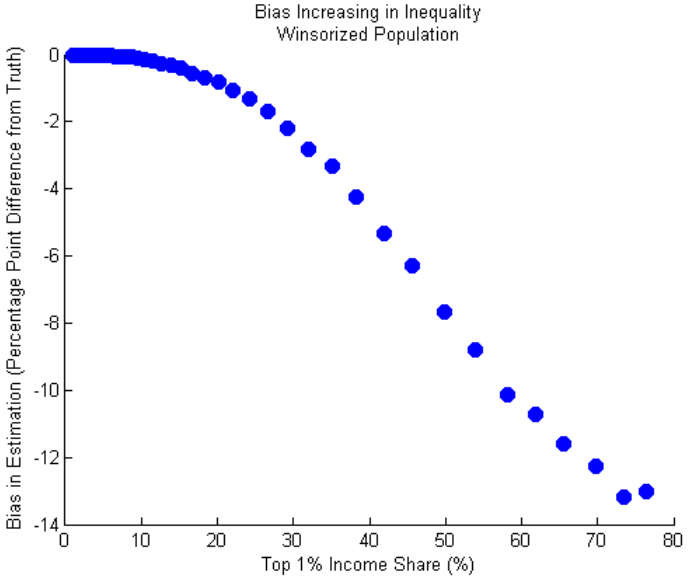

Figure A.1. This plot depicts the result for the top 1% share from winsorizing the population rather than the sample.

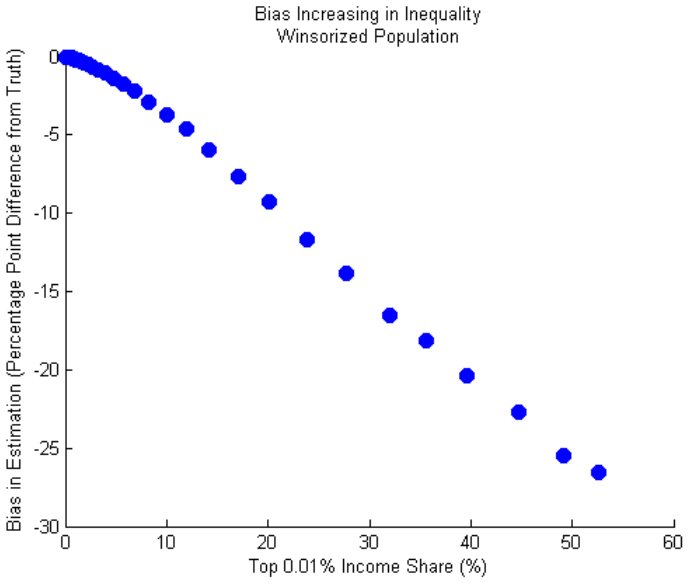

Figure A.2. This figure shows the simulation result for the top 0.01% share from winsorizing the population, then drawing the 1/16th sample.