Detecting Politically Motivated Tampering With Workers' Labor Income In Survey Data[1]

John Angle
The Inequality Process Institute, Post Office Box 215, Lafayette Hill, PA 19444

Abstract

There is no reason to think that labor income data in public use samples of the March Current Population Survey (CPS), a series dating back to March 1962, have been tampered with. So that series presents an opportunity to test a method of detecting tampering with labor income data. A tampering scenario: a recently elected national chief executive, aspiring to autocracy, must win the next election before dispensing with fair elections. Good news about the labor incomes of less skilled workers, the aspiring autocrat's electoral base, is a prerequisite. The government's statistics agency, controlled by the aspiring autocrat, continues to issue public use samples while it tampers with labor income data. Economists and survey statisticians may discern tell-tale evidence of tampering. They and it are politically dismissible. The Inequality Process (IP) is a model of labor income statistics. It implies there are only two drivers of change in the labor income distribution of a national population and how the distribution changes. Ad hoc tampering with labor income data is likely to be clearly highlighted by violation of IP invariances.

Key Words: data tampering, forensic statistics, Inequality Process, labor income, March Current Population Survey, particle system

1.0 Introduction

This paper shows that invariances of the Inequality Process, a mathematical model of income and wealth statistics (Angle, 1983-2013), can be used to detect an example of politically motivated tampering with labor income data in public use samples of a national survey. The requirements for detection are i) a periodic national household surveys with a large N, ii) no published longitudinal information, iii) respondent education and labor income, and iv) a public use sample, *before* an unscrupulous chief executive of the national government orders tampering with labor income in that survey.

2.0 A Scenario of Data Tampering[2]

A recently elected leader of a national government aspires to be an autocrat. He[3] has to win another election or two before he can re-write the constitution, start a war, stage a coup, silence journalists, opponents, and critics, or otherwise instigate autocracy. His core electoral support is the tory working class. The tory working class electorate is largely identifiable in a national survey as adults without post-secondary education. Labor income is their lifeline. The aspiring autocrat, "***The Boss***", has appointed a personally loyal crew to manage the government statistics agency. Their mission: "***Great Stats!***" to get "***The Boss***" re-elected. The Agency Director assembles a small team of statisticians to tamper with labor income data in the annual national survey, released as a public use sample.

---

[1] This article is dedicated to Dr. Andreas Georgiou, former president, Helenic Statistical Authority (ELSTAT) [chief statistician], Government of Greece. See Langkjaer-Bain (2017).
[2] Alan Alda (2017), a science communicator, suggests story telling as an expository device. The "***Great Stats! Team***" scenario is fiction with no particular geo-political reference.
[3] All examples of such national chief executives known to the author are male.

Excited to work with the Agency Director, chairing the "***Great Stats! Team***", they hope to be "leapfrog" promoted.

3.0 Data For Simulation Of Tampering

The data for this simulation are the March Current Population Survey conducted by the U.S. Bureau of the Census with a supplement, the Annual Social and Economic Supplement, funded by the U.S. Bureau of Labor Statistics. The dataset was purchased from the Unicon Corporation (Current Population Surveys, 2011). The years of labor income to be tampered with are 2006 through 2010, worker annual wage and salary incomes in the 2007 through 2011 March Current Population Surveys (CPS'). These are the years of the "Great Recession" in the U.S., the sort of labor income data that "***The Boss***" would want "touched up". Dollar values have been adjusted to constant 2010 dollars using the Council of Economic Advisers' personal consumption expenditure (PCE) price index. (Table B-7 Chain type price indexes for gross domestic product 1959–2010, Economic Report of the President, February 2012 (Council of Economic Advisers, 2012). The target group for tampering is people with at least $1 in annual wage and salary income, 25+ years old, with no post-secondary education. Data analysis is done with GAUSS (Aptech Systems, 2012).

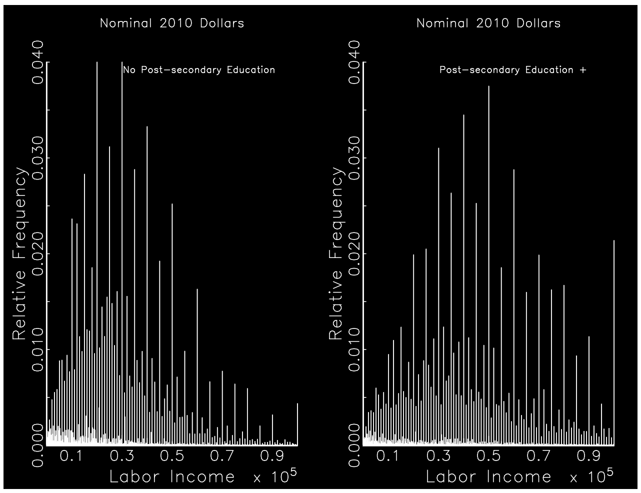

Figure 1

4.0 The "*Great Stats!* Team" Gets Down To Business

The Agency Director announces that a random percentage will be added to the reported labor incomes of workers without post-secondary education in the public use sample of the next national survey. Team Member A points out that doing so eliminates the characteristic frequency spiking structure of labor income. Most respondents "ballpark" their labor income when asked. In the U.S. that means a multiple, particularly an even multiple, of $5,000. "It'll be easy to spot!", says Member A. See Figure 1 for the frequency spikes in March 2011 CPS in 2010 annual wage and salary income (in 2010 dollars).

4.1 Team Member B's Tampering Plan

Team Member B proposes a data tampering plan. Like the Agency Director, Member B proposes to tamper with labor income data after the completion of the survey. B explains: "We need "unenhanced" data to do the "enhancing" [the language of the "***The Boss'*** " secret directive, "data enhanced for national needs"]. Military and economic planners need "unenhanced" data. Market researchers too. Access to the "unenhanced" data is something "***The Boss***" can offer loyal businesses. There'll be two versions of the data. We'll do routine post-stratification and respondent confidentiality work as well as "enhancement". Other agency employees will work in data collection and routine processing as before.".

B's plan randomly separates records of the target group into two groups, 75% to "enhancement" (tampering), 25% "unenhanced". B says "The "enhanced" group has its survey weights multiplied by 1.1 if their labor income exceeds the "unenhanced" median labor income, or by the reciprocal of 1.1 otherwise. In next year's survey, that year's "enhanced" target group who exceed next year's "unenhanced" median will have their survey weights multiplied by $(1.1)^2$, otherwise by its reciprocal. And so, until the next election.". B's plan leaves 25% of the records in the target group "unenhanced" for more realistic post-tampering "raking" (iterative adjustment of record weights to pre-tampering weight totals for target cases in a few high visibility categories).

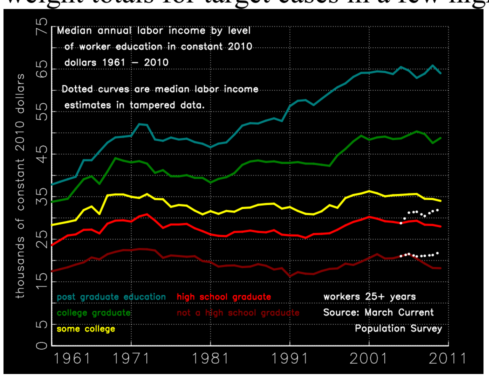

Figure 2

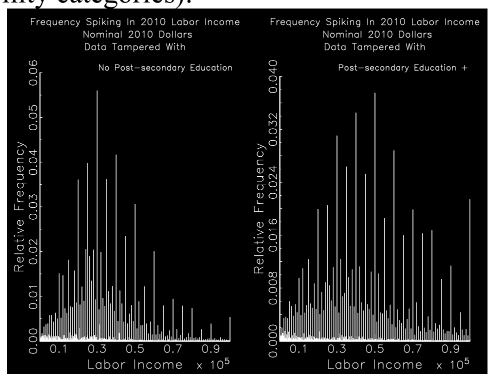

Figure 3

4.2 Simulating B's Tampering Plan

Figures 2 and 3 show how Member B's tampering plan works on labor incomes in the 2007 through 2011 March CPS records of people in B's target group. The white dotted curves of Figure 2 show the medians of the tampered data. The tampered medians show healthy but not utterly implausible gains for workers in the target group, while their actual medians decreased and those of other workers stayed essentially flat. Comparing Figures 1 and 3 shows that B's plan preserves income frequency spikes in a way that appears consistent with a general rise in the labor incomes of the target group.

B's plan conceals tampering via post-tampering raking. Technical Bulletin 66 (U.S. Bureau of the Census, 2006) explains raking in the CPS, as the sequential, iterative adjustment of the weights of certain categories of respondents in the March CPS so they have the same proportions in both survey sample and national population (as projected from the last national census). The ***"Great Stats!*** Team" keeps those projections secret so potential critics have to rely on comparisons to previous surveys to detect implausible changes in the survey weights. Leaving some survey weights of people in the target group untampered makes for more plausible post-tampering raking results.

In simulating B's post-tampering raking with March CPS data, the ratio of the target group's pre-tampered weight totals in certain categories to the post-tampered totals (which include the untampered target group weights) is taken. These categories are 1) the whole target group (respondents with positive labor income, no post-secondary education, and 25+ in age), 2) females in the target group, 3) males in the target group, 4) African-Americans in the target group, 5) non-African-Americans in the target group, 6) 25 to 34 year olds in the target group, 7) 35 to 44 year olds in the target group, 8) 45 to 54 year olds in the target group, 9) 55 to 64 year olds in the target group, and 10) 65+ year olds in the target group. All target group survey weights in each category are multiplied by the category's ratio of sum of weights, one category at a time. Then the sums of weights and their ratios are recalculated and the procedure repeated. Convergence of the ratios to 1.0 in all categories occurs after a few iterations. For example, the initial ratio of men's pre-tampering sum of weights to the sum of their post-tampering sum of weights is .9763. So in the first round of adjustments the survey weights of all men in the target group (both tampered and untampered) are multiplied by .9763.

Table 1 Employee B's Tamper Factors And Start Values For Post-Tampering Raking

Tamper factors (applied to survey weights of 75% of target group, those with labor incomes greater than true median (1.1) and those below it (1.1^{-1}) in year 1)	Year 1 (1.1), $(1.1)^{-1}$	Year 2 $(1.1)^2$, $(1.1)^{-2}$	Year 3 $(1.1)^3$, $(1.1)^{-3}$	Year 4 $(1.1)^4$, $(1.1)^{-4}$	Year 5 $(1.1)^5$, $(1.1)^{-5}$
Ratios of Sum of Weights Of Target Group Pre-Tampering To That Sum Post-Tampering					
Whole target group (tampered and untampered)	0.9866	0.9701	0.9561	0.9333	0.9107
Women in target group	1.0001	0.9933	0.9935	0.9769	0.9588
Men in target group	0.9763	0.9529	0.9291	0.9028	0.8776
African-Americans in target group	0.9922	0.9832	0.97801	0.9607	0.9382
Non African-Americans in Target group	0.9857	0.9682	0.9529	0.9294	0.9067
25 to 34 year olds in target group	0.9920	0.9836	0.9779	0.9649	0.9453
35 to 44 year old in target group	0.9835	0.9624	0.9483	0.9199	0.8970
45 to 54 year olds in target group	0.9812	0.9576	0.9378	0.9141	0.8889
55 to 64 year olds in target group	0.9843	0.9698	0.9466	0.9214	0.8931
65+ year olds in target group	1.0086	1.0093	1.0171	0.9940	0.9877

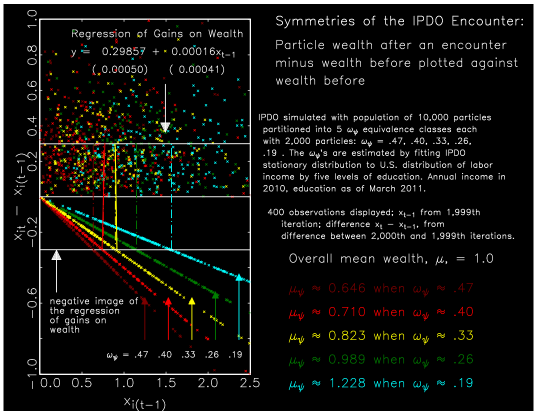

Figure 4

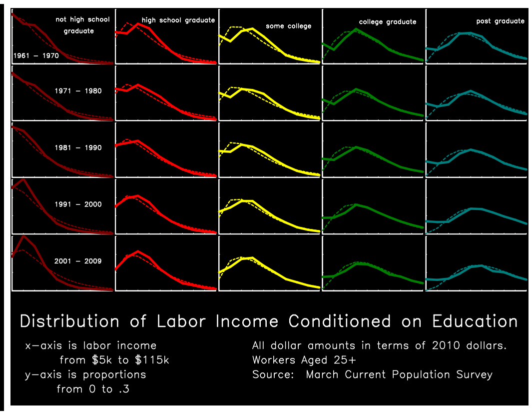

Figure 5

5.0 B's Tampering Violates Inequality Process (IP) Invariances

The Inequality Process (IP) (Angle, 1983-2013) is a stochastic interacting particle system model in which particles exchange a positive quantity (wealth). Each particle has one parameter, ω, the proportion of wealth it loses in an encounter with another particle in which it loses. See (1a,b) and Figure 4.

$$x_{it} = x_{i(t-1)} + d_t \omega_\theta x_{j(t-1)} - (1-d_t)\omega_\psi x_{i(t-1)}$$
$$x_{jt} = x_{j(t-1)} - d_t \omega_\theta x_{j(t-1)} + (1-d_t)\omega_\psi x_{i(t-1)} \quad (1a,b)$$

Where x_{it} is particle i's wealth at time t. Particle i has ω parameter, ω_ψ, $0 < \omega_\psi < 1$. Other particles may have a different value of ω. $d_t = 0$ with probability .5, 1 with probability .5. The verbal theory from which the Inequality Process (IP) was derived implies that more productive workers are more sheltered from competition. In the IP "more sheltered from competition" means smaller ω, and consequently in testing the IP's relevance to labor income data, smaller ω indexes greater worker productivity. As is common in economics, worker productivity is operationalized by worker education. So the stationary distribution of the IP conditioned on particular values of ω, (2), is hypothesized to model the distribution of labor income conditioned on level of worker education. Part of this hypothesis is that the ω's estimated in the fit scale inversely with worker education, treating workers with the same level of education as IP particle equivalence class of ω's. Figure 5 shows (2) approximately fits the distribution of labor income conditioned on five levels of worker education in March CPS data (1962-2011). The ω's estimated in these fits, incomes from 1961-2010, are given in Figure 6. 1962's values are interpolated. The white dotted curves of Figure 6 are from fits of (2) to tampered income data (2006-2010). Note that the ω's (estimated from untampered data) remain ordered inversely with education, as implied by the theory from which the IP was abstracted, and close to their 50 year averages, in every year in the untampered data. Figure 6 shows that when (2) is fitted to the tampered data, estimated ω's (white dotted curves) in the target group deviate from their long term averages and, in the target group, either step out of the IP hypothesized order or are headed that way. Figure 7 shows (2)'s fits to the tampered data becoming quickly poorer with more aggressive tampering.

The distribution of particle wealth in the IP's ω_ψ equivalence class is closely approximated by:

$$f(x_\psi) \equiv \frac{\lambda_{\psi t}^{\alpha_{\psi t}}}{\Gamma(\alpha_{\psi t})} x_\psi^{\alpha_{\psi t}-1} e^{-\lambda_{\psi t} x_\psi} \qquad (2)$$

where:

$$\tilde{\omega}_t = \left(\sum_{\psi=1}^{\Psi} \frac{w_{\psi t}}{\omega_\psi} \right)^{-1}$$

$\Psi \equiv$ the number of distinct ω_ψ equivalence classes

$x_\psi \equiv$ wealth in the ω_ψ equivalence class

$x_\psi > 0$

$\alpha_{\psi t} \equiv$ shape parameter $\approx \frac{1-\tilde{\omega}_t}{\omega_\psi}$

$\lambda_{\psi t} \equiv$ scale parameter $\approx \frac{1-\tilde{\omega}_t}{\tilde{\omega}_t \mu_t}$

$\tilde{\omega}_t \equiv$ harmonic mean of ω_ψ's at time-step t

$\tilde{\omega}_t \leq 0.5$

μ_t unconditional mean of wealth at time-step t

$$\mu_{\varphi t} \approx \frac{\widetilde{\omega_t}}{\omega_\psi} \mu_t$$

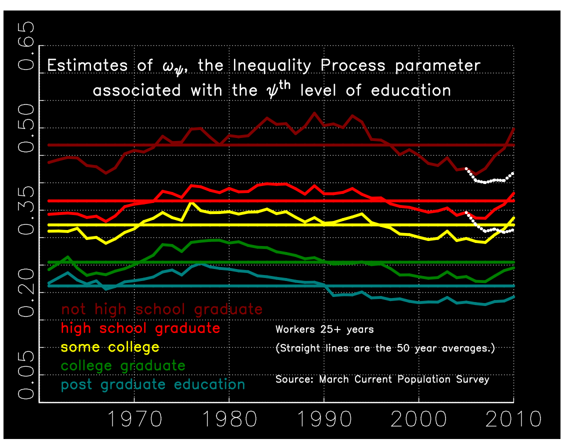

Figure 6

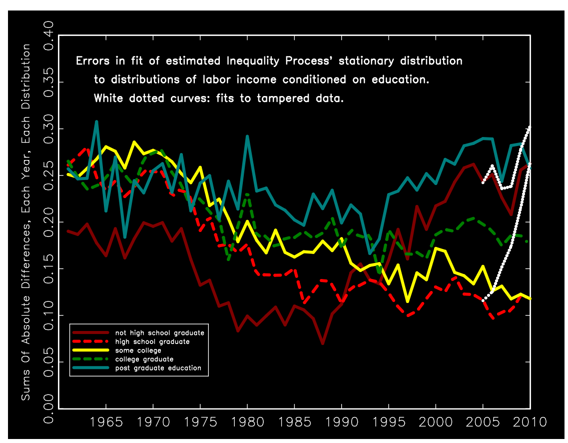

Figure 7

Change in (2) is driven by change in the unconditional mean of particle wealth, μ_t, the operationalization of mean labor income, and change in the harmonic mean of particle ω's, $\tilde{\omega}_t$, whose referent is labor force productivity. The partial derivative of (2) with respect to μ_t is readily interpretable: $\frac{\partial f_{\psi t}(x_0)}{\partial \mu_t} = f_{\psi t}(x_0)\, \lambda_t \left(\frac{x_0 - \mu_{\psi t}}{\mu_t} \right)$ (3). The distribution of labor income in the ψ^{th}

ω equivalence class pin-wheels around its mean, the conditional mean $\mu_{\psi t}$, when μ_t changes. The tails of the distribution change the most proportionally, the right tail in particular since the distribution is right skewed. Angle (2007a) shows that (2)'s two drivers explain changes in the U.S. distribution of labor income and consequently its scalar statistics.

References

Alda, Alan. 2017. ***If I Understood You, Would I Have This Look On My Face?*** New York: Random House.
Angle, John. 1986. "The surplus theory of social stratification and the size distribution of Personal Wealth." ***Social Forces*** 65:293-326.
_____. 1992. "The Inequality Process and the distribution of income to blacks and whites". ***Journal of Mathematical Sociology*** 17:77-98.
_____. 1993. "Deriving the size distribution of personal wealth from 'the rich get richer, the poor get poorer'". ***Journal of Mathematical Sociology*** 18:27-46.
_____. 1996. "How the gamma law of income distribution appears invariant under aggregation". ***Journal of Mathematical Sociology***. 21:325-358.
____. 2002a. "The statistical signature of pervasive competition on wages and salaries". ***Journal of Mathematical Sociology***. 26:217-270.
____. 2003a. "The dynamics of the distribution of wage and salary income in the nonmetropolitan U.S.". ***Estadistica***.55: 59-93.
____. 2003b. "Inequality Process, The". An entry in T. Liao, et al., (eds.), ***The Encyclopedia of Social Science Research Methods.*** Volume 2: 488-490. Thousand Oaks, CA: Sage.
____. 2006a. (received 8/05; electronic publication: 12/05; hardcopy publication 7/06). "The Inequality Process as a wealth maximizing algorithm"'. ***Physica A: Statistical Mechanics and Its Applications*** 367:388-414 (DOI information: 10.1016/j.physa.2005.11.017). A draft version of the paper can be downloaded from the Luxembourg Income Study website: [http://www.lisproject.org/publications].
____. 2006b. "A comment on Gallegati et al.'s "Worrying Trends in Econophysics" ". Pp. 247-253 in A. Chatterjee and B.K. Chakrabarti, (eds.), ***The Econophysics of Stocks and Other Markets*** (Proceedings of the Econophys Kolkata II Conference, February, 2006 [http://www.saha.ac.in/cmp/econophys2.cmp/]). Milan: Springer. (ISBN: 978-8847005013).
____. 2007a. "The Macro Model of the Inequality Process and The Surging Relative Frequency of Large Wage Incomes". Pp. 171-196 in A. Chatterjee and B.K. Chakrabarti, (eds.), ***The Econophysics of Markets and Networks*** (Proceedings of the Econophys-Kolkata III Conference, March 2007 [http://www.saha.ac.in/cmp/econophys3.cmp/]). Milan: Springer [ISBN: 978-8847006645]. [on-line at: http://arxiv.org/abs/0705.3430].
____, François Nielsen, and Enrico Scalas. 2009. "The Kuznets Curve and the Inequality Process". Pp. 125-138. In Banasri Basu, Bikas K. Chakrabarti, Satya R. Chakravarty, Kausik Gangopadhyay, editors, ***Econophysics and Economics of Games, Social Choices and Quantitative Techniques***. (Proceedings of the Econophys-Kolkata IV Conference, March 2009, Kolkata, India, jointly sponsored by the Indian Statistical Institute and the Saha Institute of Nuclear Physics.) [http://www.isical.ac.in/cmp/econophys4.com/]. Milan: Springer.
____. 2012. "The Inequality Process v. The Saved Wealth Model: Which is the More Likely to Imply an Analogue of Thermodynamics in Social Science?". ***Journal of Mathematical Sociology***. 36: 156-182.
____. 2013. "How To Win Acceptance Of The Inequality Process As Economics". ***Society And Management Review*** vol. 2 (#2): 117-134. Kozhikode, India: Indian Institute of Management.
Aptech Systems, Inc. (2012). GAUSS (Version 12.0) user guide. Black Diamond, WA: Aptech Systems.
Current Population Surveys, March 1962-2011. [machine readable data files]/ conducted by the Bureau of the Census for the Bureau of Labor Statistics. Washington, DC: U.S. Bureau of the Census [producer and distributor], 1962-2004. Santa Monica, CA: Unicon Research Corporation [producer and distributor of CPS Utilities], 2012.
Council of Advisers. (2005). ***Economic report of the President***. Washington, DC: U.S. Government Printing Office.
Langkjaer-Bain, Robert. 2017. "Trials of a Statistician". ***Significance***. (vol 14, #4, pp. 14-19). London, U.K.: Royal Statistical Society.
U.S. Bureau of the Census and U.S. Department of Labor. 2006 ***Technical Bulletin #66.*** Washington, D.C.: Government Printing Office.