Physica A 415 (2014) 54–60

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Race, gender and the econophysics of income distribution in the USA☆

Anwar Shaikh [a,*], Nikolaos Papanikolaou [b], Noe Wiener [a]

[a] Department of Economics, New School for Social Research, New York, NY 10003, USA

[b] Department of Economics, Lehman College of The City University of New York, New York, NY 10468, USA

HIGHLIGHTS

- Shows that labor income distributions by race and gender are roughly exponential.
- Results support Yakovenko econophysics "two-class" income distribution hypothesis.
- Regressions yield $R^2 \approx 0.99$ and actual/predicted mean incomes from 1.03 to 1.10.
- Weibull probability distribution provides only marginally better results.
- Social policy does not seem to affect intra-group race and gender distributions.

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form 10 July 2014
Available online 1 August 2014

Keywords:
Economics
Econophysics
Income distribution
Classical statistical mechanics
Gender and race

ABSTRACT

The econophysics "two-class" theory of Yakovenko and his co-authors shows that the distribution of labor incomes is roughly exponential. This paper extends this result to US subgroups categorized by gender and race. It is well known that Males have higher average incomes than Females, and Whites have higher average incomes than African-Americans. It is also evident that social policies can affect these income gaps. Our surprising finding is that nonetheless intra-group distributions of pre-tax labor incomes are remarkably similar and remain close to exponential. This suggests that income inequality can be usefully addressed by taxation policies, and overall income inequality can be modified by also shifting the balance between labor and property incomes.

1. Introduction

The study of income distribution in capitalist societies can be traced back to Pareto's (1897) finding that they seem to follow a particular (Pareto) power law. Modern evidence confirms that this law only applies to the upper tail of the distribution [1, p. 1], [2]. At the same time, it has been repeatedly shown that the lower bulk of the income distribution does not follow a Pareto law [3, p. 585].

One way to address these results is to try to encompass the whole range of incomes by means of a single distribution such as the log-normal, Levy, Gamma, Champernowne and others [3]. An alternate path has been proposed in the econophysics "two-class" theory of income distribution (EPTC) of Yakovenko and his co-authors. They argue on theoretical grounds that labor income approximately follows an exponential (thermal) distribution while property income follows a Pareto (superthermal) distribution. Personal incomes may of course encompass both labor and property incomes, but it is plausible

☆ We wish to thank to two anonymous referees for their insightful and helpful comments on earlier drafts.

* Corresponding author. Tel.: +1 7187683935.
E-mail addresses: shaikh@newschool.edu, ashaikh606@gmail.com (A. Shaikh).

http://dx.doi.org/10.1016/j.physa.2014.07.043
0378-4371/

that the former type dominates at lower income levels and the latter at the highest ones. Yakovenko et al. provide substantial empirical evidence for their claim that the bottom 97%–99% of the distribution of personal incomes is exponential while the top 1%–3% is Pareto. They also provide an ingenious method of marrying the two types of distributions, thereby providing a powerful yet parsimonious approximation to overall income distribution data [1].

For individual incomes (r) an exponential distribution function has a probability distribution $P(r) = (1/R)\,e^{-(\frac{r}{R})}$. The parameter R represents the mean income of the theoretical exponential distribution, to which we compare the actual mean incomes of the observed distributions. The cumulative probability distribution for incomes *above* r is $C(r) = e^{-(\frac{r}{R})}$, which is parameter-free in normalized income $\rho \equiv r/R$. Since $\ln C(r) = -(1/R)\,r$, we can estimate R by means of a regression of $\ln C(r)$ on r.

The EPTC hypothesis predicts that labor incomes will cluster around the straight line $\ln C(r) = -\rho$. This paper tests the hypothesis against the gender and race distributions of wage and salary incomes in the United States. The test is of interest because even though Males are known to have higher average incomes than Females, Whites higher incomes than African-Americans, and all gaps are known to change over time in response to social forces,[1] there seems to have been little attention paid to the income distribution within each group. We find that despite all the differences between groups, intra-group distributions continue to be *approximately exponential*.

2. Empirical evidence

Our data is from the US Census Bureau Current Population Survey (CPS) in the March supplement of personal wage and salary income (r) by race and gender from 1996 to 2008, adjusted to remove persons with zero incomes and those listed as self-employed (including those listed as self-employed incorporated businesses).[2] On the latter issue, it is useful to note that national income accounts classify the income of proprietors and partners entirely as personal income even though it is acknowledged that some part of that income is really the *profit* of unincorporated enterprises [5, p. 15]. Correcting for this at the sectoral level requires splitting sectoral proprietors' income into two parts: a wage-equivalent and a corresponding profit-equivalent [6, pp. 4–5]. Since such a correction is not possible for individual incomes, we excluded income from self-employment.

Probabilities were calculated from data binned over the whole sample, with the bin width being calculated using the Epanechnikov kernel function for density estimation in order to avoid subjective bias. The data is only available in censored form, supposedly to preserve confidentiality. Hence in each year there is a particular topcode income level (e.g. $200,000 in 2007) above which all individual data points are set equal to the average income of a similar cohort.[3] Topcoding does not affect the sample mean but it does change the shape of the distribution in the top-coded region. Still, on the assumption that the true distribution also obtains in the top-coded region, OLS estimates of the distribution parameters will be consistent and unbiased. Following Yakovenko et al., the "income-temperature" R in each year is estimated by regressing $\log C(r)$ on r (with no intercept), and using each annual estimate of R to construct normalized incomes ($\rho \equiv r/R$) in that year.

Fig. 1 compares the actual distributions (depicted by symbols) of *All Incomes* in US CPS Wage and Salary Income data, with the theoretical distribution (depicted by a solid line), on both log-linear and log–log scales over two sets of years: 1996–2002 and 2003–2008, respectively.[4] The log–log curves were added because then observed deviations from the theoretical line represent percentage differences, these being noticeably smaller at upper income levels than the corresponding level differences in the log-linear curves. Inset in each chart is the estimated income-temperature R for each year. Our CPS data displays the same near-exponential distributions found by the EPTC group using IRS data. The correspondence between the data and the theory appears to be quite good, except for a smattering of points at the upper end of the spectrum. This latter effect could be due to the low probabilities associated with the upper points which give rise to random sample fluctuations [8, p. 207], but is more likely due to the fact that salary incomes near the topcode limit include a certain amount of property income in the form of performance bonuses and exercised stock options [9, p. M-8]. These would raise the average incomes in the top bins and raise the observed points above the theoretically expected ones at the tail end of the distribution—just as we find in our charts (see Figs. 4–6).

According the EPTC hypothesis, a parameter-free straight line $\log C(\rho) = -\rho$ should provide a good approximation to the actual data, and actual Gini coefficients should be close to 0.5 which is the Gini of a pure exponential distribution [1,3]. Shown in parentheses in Table 1 are the regression coefficients of determination (R^2 which here is *not* the square of the

[1] From 1979 to 2009, relative to that of White men the median weekly earnings of White women has gone from 61.7% to 79.2%, that of African-American men from 76.2% to 73.5%, and of African-American women from 56.7% to 68.9% (for full time wage and salary workers, 16 years and older) [4].

[2] Our CPS sample was created via the code WSAL_VAL $>$ 0 & ERN_SRCE $=$ 1, where the first condition selects respondents with positive wage and salary earnings, and the second excludes the self-employed. To select Whites, Blacks, Males and Females we use A_RACE(PRDTRACE) $=$ 1 or 2, A_SEX $=$ 1 or 2, respectively.

[3] From 1996 to 2002 topcoding for wages and salaries begins at $150,000 and from 2003 to 2008 at $200,000 [7]. "Records that were topcoded will have a value greater than the topcode value. The topcode was replaced with the mean earnings for topcoded individuals with similar characteristics" (https://cps.ipums.org/cps/topcodes_tables.shtml#1996top). This obviously distorts the distribution in the topcoded regions and reduces inequality by assigning the same income to several people in a given range.

[4] The data in a given CPS publication refers to the incomes in the prior year. We retain the publication year as the reference.