Bulletin of Applied Economics, 2020, 7(1), 107-122

The Econophysics of Labor Income

Nikolaos Papanikolaou[1]

Abstract

This paper examines the Census Bureau's Current Population Survey (CPS) of full-time wage and salary workers to determine the underlying functional form of the size distribution of income from 1996 to 2008. There has been a lot of attention on income inequality Pre and Post Great Recession of 2008-2009. This paper applies the tools developed in a new field of economics called Econophysics. The analysis uses parametric and nonparametric methods to determine the size distribution of wage and salary income. The findings suggest that the underlying functional form of labor income is approximately distributed as an exponential distribution, while non-labor income is underscored by a Pareto distribution.

JEL classification numbers: C19, D31, D33, D39, J01
Keywords: Size Distribution of Labor Income, Income Inequality, Boltzmann-Gibbs Distribution, Optimal Bandwidth, Kernel Density Estimator, Census Bureau, Exponential Distribution, Pareto Distribution.

1 Introduction

Classical economist such as Adam Smith (1776) incorporated a scientific approach visa-a-vise quantitative method to explain the real world. Leon Walras (1874) introduced rudimentary calculus in economic analysis to explain the real world. Vilfredo Pareto (1897) applied mathematical tools developed in the field of engineering to economic analysis. Contemporary economics has advanced beyond the Classical and Walrasian general equilibrium theory and implemented advanced mathematical and Statistical methods which are applied in physics such as thermal dynamics and nonparametric models. These methods use probability distributions to explain large number of economic phenomena. According to Chatterjee et al. (2008), probability distributions can be thought of as representing some sort of income "inequality" among the large number a agents (objects) of a given system. This is in a sense that different agents have different values of a given variable. Hence, studying probability distributions within the context of economics is also a study of income inequality developing from a system based on statistical reasoning. Therefore, using statistical models developed in Econophysics will provide additional tools for policy makers to address the rise in income inequality.

The remainder of the paper is as follows. Section 2 is the literature review. Section 3 is the methodology which is broken-down into subsections, 3.1 top-coding current population survey (CPS) income data, 3.2 CPS data variables, 3.3 optimal bandwidth, 3.4 kernel density estimator, 3.5 Goodness-

[1] Lehman College, Bronx, New York, USA

Article Info: *Received*: March 25, 2020 *Published online*: April 15, 2020

of-fit, 3.6 Komogorov-Simirnov test, 3.7 Boltzmann-Gibbs distribution. Section 4 the empirics: full-time wage and salary respondents. Section 5 conclusion.

2 Literature Review

The roots of inquiry on the statistical size distribution of income and wealth can be traced as far back as 1897, Vilfredo Pareto's "Cours d'economie politique." (Pareto, 1897). Thus, Pareto's empirical research in the mid-to-late nineteenth Century in England applied mathematical and engineering techniques to derive "laws" which describe the distribution of income and wealth subsequently, coined the "Law of the Vital Few", "Pareto's Principle", the "80-20 Rule" or "Pareto's Power Law." Pareto's index[2] α measured the proportion of the population with very high incomes and Pareto's research based on the parameter α exhibited that twenty percent of the population in England owned eighty percent of the wealth. Hence, Pareto's pioneering work identified the first power-law distribution in any field. Subsequently, it established the groundwork for future research and analysis in the statistical size distributions and functional forms of income from economists to a selected number of physicists. Furthermore, the fact that empirically, the values of parameter α remain stable, specifically, in economies such as semi-feudal Prussia, Victorian England, capitalist but highly diversified Italian cities circa 1887, and communist-like regime of the Jesuits in Peru during Spanish rule (1556-1821) – caused Pareto to conclude that human nature, that is, humankind's varying capabilities, is the main cause of income inequality, rather than the organization of the economy and society Kleiber and Kotz, (2003).

Pareto's Power Law[3] is a powerful tool in the analysis of income for the reason that it recognized different distribution functions along a given range. Therefore, economists are able to approximate with some confidence (i.e., probability) the underlying true distribution and functional form in the whole income range with less ambiguity and greater preciseness.

The revival of interest in the statistical size distribution of income and income inequality has produced a wide range of research papers in the current literature. This paper adds to the current literature and builds on the work of Dragulescu and Yakovenk (2000); Dragulescu and Yakovenko (2001a; 2001b); Silva and Yakovenko (2005a; 2005b); Yakovenko (2009); Yakovenko and Rosser (2009); Shaikh and Papanikolaou and Wiener (2014) that the distribution of labor income follows an exponential (thermal) distribution while that of property (and/or financial assets) income follows a Pareto (super-thermal). Therefore, the bottom 97-99% of the distribution of personal incomes is exponential (and thereby dominated by labor income) while the top 1-3% is Pareto.

---

[2] Pareto's index $(\frac{r_n}{r})^{\alpha}; r > r_n$, where r_n is any positive income. Therefore α measures the proportion of the population with very high incomes. The larger $\alpha > 1$ the smaller the proportion of high incomes relative to the rest of the population. Pareto exhibited that parameter α was stable over time and space and subsequently, Yakovenko and Rosser (2009). expanded Pareto's work and argued that incomes below the Pareto part (there is no one specific income that divides the Pareto and exponential distribution) of the income distribution) are also stable over time and incomes are distributed exponentially.

[3]Mandelbrot's (1960) and (1963), contribution to the statistical size of income distributions led to numerous and more recent empirical studies, for example Levy and Solomon (1997); Dragulescu and Yakovenko (2001a; 2001b); Souma (2001), and Souma (2002) have all shown that the power law tail is ubiquitous feature of income distributions for asset-based incomes.

Income inequality is nothing new in the United States. According to the Congressional Budget Office 2014 report[4], income inequality is on the rise. Furthermore, the income gap between the rich and the poor has drastically widened in the advent of the Great Recession of 2008-2009. Furthermore, an important question that could be proposed is why does income inequality matter in the first place from an economic growth standpoint? In other words, poverty itself has risen as low and middle-class incomes have stagnated, which have contributed to expanding inequality. It should also be emphasized that younger workers, male and female, are doing worse than their elders in terms of rates of improvement. Median wages and salaries for those in their late twenties and early thirties have fallen or hardly risen at all over thirty-six years Madrick and Papanikolaou (2010). Therefore, a growing economy does not necessary translate into a reduction in poverty or income inequality.

According to Horowitz and Igielnik and Kochhar of the Pew Research Center of Social and Demographic Trends[5], barely 10 years past the end of the Great Recession in 2009, the U.S. economy is doing well on several fronts. The unemployment rate in November 2019 was 3.5%, a level not seen since the 1960s. The labor market is on a job-creating streak that has rung up more than 110 months straight of employment growth, a record for the post-World War II era. But economic inequality, whether measured through the gaps in income or wealth between richer and poorer households, continues to widen.

Monetary policy did not fare any better in addressing poverty and income inequality. Prior to Great-Recession of 2008-2009, the Federal Reserve had a strong dislike for inflation. Monetary policy was reinforced by a model of long-term positive trade-off between inflation and output growth which ushered in the highest levels of income inequality post-1943 era Papanikolaou (2020).

In the advent of the Great-Recession unconventional monetary policy was used to address the financial crisis and subsequent spike in unemployment. According to Momtaz et al. (2017) states that quantitative easing may have contributed to the increase in inequality over the Great Recession. Furthermore, Guerello (2018) shows that unconventional monetary policy had poor redistributive fiscal policy and highly sensitive households' portfolio might trigger these results. In the case of Japan, Saiki (2014) argues that results that unconventional monetary policy widened income inequality as the Bank of Japan (BoJ) resumed its zero-interest rate policy and reinstated unconventional monetary policy. While Sima et al (2020) apply a vector error correction model and their findings show that an increase in money stock (m1) through Quantitative Easing (QE) and Quantitative and Qualitative Easing (QQE) policies of the Bank of Japan (BOJ) significantly increases the income inequality.

3 Methodology

To better understand income distributions in the whole range, nonparametric and parametric[6] models are often applied to fit economic data. In this paper, I use nonparametric models to analyze the size distribution of income using the United States Census Bureau's March Current Population Survey (CPS) of personal wage and salary income for all full-time respondents. It's important to note that the nonparametric density estimation techniques allow us to provide full information on the entire income distribution. Moreover, the nonparametric approach has the benefit of letting the data speak for itself, eliminating ad hoc and/or arbitrary methods to derive underlying distributional specifications. More

[4] https://www.cbo.gov/system/files/115th-congress-2017-2018/reports/53597-distribution-household-income-2014.pdf

[5] https://www.pewsocialtrends.org/2020/01/09/trends-in-income-and-wealth-inequality/

[6] The only parametric goodness-of-fit test used in the paper is the Kolmogorov-Smirnov.

importantly, assumptions on the data are kept at a minimum and simply assume that the income density exists and satisfies some smoothness properties. Lastly, the nonparametric approach avoids many theoretical difficulties and empirical fragility encountered in tradition measures of income distributions.

The empirical research in this paper and its subsequent results are derived from raw income data obtained from the Current Population Survey (CPS). The survey gathers a wealth of information on income and other aspects of the United States population. The March surveys contain the Annual Demographic File and Income Supplement, which report the income-related aspects of individuals, families and households in detail. In this paper, I specifically analyze the distribution of personal wage and salary income of full-time respondents from 1996 to 2008. The Census Bureau defines personal wage and salary income as workers who receive wages, salary, commission, tips or pay in kind from private employer or from a government unit. Also included are persons who are self-employed in an incorporated business[7].

The Census Bureau conducts several household surveys that measure the economic situation of people (classified as personal or individual), families, and households in the United States. The basic Current Population Survey (CPS) takes place every month. Its primary focus is to collect information on current employment status. In March of every year, a supplementary questionnaire gathers information about income received during the previous calendar year. The March CPS interviewed people in approximately 60,000 households from 1991 until 1996, when the sample size decreased to 50,000 households. Besides the change in sample size, a new sample design was introduced, and the survey was converted from a paper questionnaire to a computerized instrument in March 1994. In addition, weights based on the results of the 1990 Census were introduced in 1993. The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Census Bureau primarily for the Bureau of Labor Statistics.[8]

3.1 Top-Coding CPS Income Data

The March Current Population Survey public use file has been top-coded since 1976 by the Census Bureau in order to protect the identity of individuals with high incomes. The Census Bureau uses top-coding for confidentiality purposes, usual hourly earnings from current job and earnings from longest job are top-coded (i.e., cut off at a particular amount).[9] In the paper, I have accounted for top-coding for 1996 to 2008. According to Burkhauser et al., (2008) top-code income cutoff vary among indicated years and specified by the Census Bureau. The personal full-time wage and salary income data analyzed in this paper are top-coded at $150,000 dollars for 1996 to 2002 and $200,000 dollars for 2003 to 2008.[10]

The data is analyzed and individuals earning zero wage and salary income were removed from the data set. Hence, the data is top-coded according to the Census Bureau defined income level and individuals earning an annual income greater than $150,000 for 1996 to 2002 and an annual income greater than $200,000 dollars for 2003 to 2011 were removed from the data set.

---

[7] Note, even though Census Bureau wage and salary income includes individuals which are self-employed and /or incorporated business in the paper we only include individuals that work primarily for an employer. Therefore self-employed and incorporated business individuals are not included in our CPS March survey data sets.

[8] http://www.census.gov/cps/

[9] http://www.nber.org/papers/w13941

[10] It's important to note that because of top-coding our analysis examines income below the top-code the non-Pareto part of the distribution of income, i.e., the bottom 97% of personal wage and salary income.

3.2 CPS Data Variables

The personal wage and salary incomes are extracted from the raw March Income Supplement of the Current Population Surveys (CPS). The CPS March Survey data income variables used for analysis are A-RACE, A-SEX, WSAL-VAL, WKSWORK, A-USLHRS, OCCUP, LJCW, WAGEOTR, ERN-SRCE, ERN-VAL, and WS-VAL.[11] As noted earlier personal wage and salary income consists of workers which are not self-employed and non-business incorporated. The reasoning is that we wanted to specifically examine by race and gender the distribution of personal wage and salary incomes from workers, individuals who earn a wage or salary from an employer.[12]

3.3 Optimal Bandwidth

The key for kernel density estimates is the choice of bandwidth h which determines the amount of smoothing.[13] One major drawback of kernel density estimation, when applied to long-tailed distributions (i.e., exponential distribution) are that the fixed bandwidth suffers because h is constant across the entire income data. If h is too small, there is a tendency for spurious noise to appear in the tail estimates. In the other hand, if one increase h to eliminate the noise, over-smoothing is often is the case and the essential detail in the main part of the distribution is sacrificed. Irrespective of the drawback, the following steps were taken to derive an objective and optimal bandwidth for CPS personal full-time wage and salary income data. There are numerous methods suggested for analytical purposes to bin data (beyond the scope and purpose of this paper).

The most frequently applied binning method is the histogram. Although using a histogram and/or eyeballing the data to derive the true underlying distribution function of personal full-time wage and salary income can be useful in a pinch (i.e., simplicity) but for practical purposes this approach can be problematic because the probability density function (PDF) can differ dramatically depending on the number of bins used. Furthermore, the histogram has two distinct disadvantages. According to Tarozzi (2009),

> *"First, the number of bins is somehow arbitrary; second, the estimated density is a step function, therefore not differentiable. The slope of the density can be of interest and it is generally more useful to deal with differentiable functions. A more practical approach to the binning problem can be overcome to a certain extent by using nonparametric density estimation."*

Density estimation is an attempt to estimate the PDF based on a given sample. In other words, it can be thought of as a way of averaging and smoothing the histogram. This method is applicable because the total probability encloses an area of one. Subsequently, the process rescales the area within the bins of the histogram so that the total area under the smoothed line equals one. This results in the proportion of incomes at specific points in the histogram rather than the frequency counts.

---

[11] The income variables analyzed in this paper as defined by the Census Bureau CPS: race of income earner (A-RACE), gender (A-SEX), total wage and salary earnings value (WSAL-VAL), weeks worked (WKSWORK), usual hours per week (A-USLHRS), occupation (OCCUP), class of worker (LJCW), other wage and salary earnings (WAGEOTR), source of earnings from longest job (ERN-SRCE), earnings before deduction value (ERN-VAL), and wage and salary earnings, other, amount (WS-VAL).

[12] The data in this paper are analyzed using mathematical and statistical programs, Mathematica, MathStatica and STATA.

[13] For a comprehensive introduction to kernel density estimation see Silverman (1986) and Jones et al. (1996).

The method of kernel density estimation applies weighted averaging of the distribution by applying a weight function or kernel that ensures the enclosed area of the curve equals one. Specifically, the kernel density estimator creates an estimate of the density by placing a "marker" at each data point and then it sums the "markers." Therefore, *K* is the kernel density function (the "marker" function); *x* is the point where the density is estimated; X_i is the center of the interval; and *h* is the bandwidth or window half-width.

In the applied statistics literature, the kernel density function that is recommended in Silverman (1986) is the Epanechnikov kernel. The Epanechnikov kernel (EPk) is the most efficient bandwidth estimator because it minimizes the mean square error (MSE) and subsequently, the mean integrated square error (MISE) more efficiently than most if not all of the kernel density estimators, irrespective of the distribution. The EPk approximates the true underlying probability density function and calculates the optimal bandwidth by employing MSE and MISE.

In most cases researchers derive bandwidth by arbitrary tactics and select the optimal bandwidth by trial and error. In other words, the smaller bandwidth is, the more details (information) are shown on the graph of the kernel density (over-smoothing). As the bandwidth increases, the density curve becomes smoother and less detail is shown on the graph (under-smoothing). The difficulty is selecting the appropriate size (i.e., small bandwidth or large bandwidth) of the bandwidth. As noted earlier, for the majority of the time deriving a given bandwidth for a distribution is done arbitrarily, either by a predetermined number of bins and/or eye-balling the histogram. Moreover, once the bandwidth is derived it specifies a unique number of bins for the density function. Consequently, the choice of the optimal bandwidth becomes ambiguous and impractical for research purposes because the optimal bandwidth essentially combines both-that is small enough to reveal detail in the graph, but large enough to produce random noise. Therefore, without an efficient and objective method to derive the optimal bandwidth, researchers will derive subjective and incorrect conclusions from the data. The EPk provides an objective, robust and efficient method based on the MSE and MISE that is simple and easy to implement (due to sophisticated statistical programs and computers) to derive the optimal bandwidth for a distribution function.

3.4 Kernel Density Estimator

We can estimate the density of ***X*** evaluated at ***x*** by using the standard kernel density estimator:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right) \quad (1)$$

The estimator, then, instead of putting a weight equal to one on all the observations in an interval width ***h*** centered around ***x***, assigns higher weights to observations that are closer to ***x***. When the bandwidth ***h*** increases, the argument of the kernel decreases, and then the assigned weights increase. Then we will get smoother estimates. Vice-versa, if h is very small, only observations very close to each point will receive high weight and the density can become very jagged. It's important to note that kernel estimators are typically biased because we are estimating the density using observations that can be from ***x.*** The objective is to derive a bandwidth that becomes smaller when $n \to \infty$. As a result, only observations close to ***x*** will be used, and the bias is eliminated. Smaller ***h***, then, means smaller bias, but it means larger variance, as curves become more jagged, and different sample can produce very different densities. Therefore, there is a trade-of between bias and variance. This is the motivations to derive the "optimal" bandwidth by minimizing the approximate mean integrated squared error (AMISE), which is an approximate estimate of the mean square error (MSE) of the estimate over the whole range of ***x***.

$$MISE = E\{\int[\hat{f}(x) - f(x)]^2 dx \quad (2)$$

Thus, approximate mean integrated squared error (AMISE) of equation (1) will be

$$\frac{1}{4}h^4\mu^2 \int_x f''(x)]^2\, dx + \frac{1}{nh}[\int k^2(u)du] \quad (3)$$

The optimal bandwidth can be chosen to minimize equation (3). The first order condition (FOC) is $h_{optimal} = cn^{-1/5}$where, $c = \left(\frac{\int k^2(u)du}{\mu^2 \int_x [f''(x)]^2 dx}\right)^{1/5}$, therefore, the optimal bandwidth becomes smaller when the sample size increases. The rule of thumb to derive optimal bandwidth [19] suggests the Epanechnikov density kernel,

$h = b\, min\{\sigma of x, \frac{InterquartileRange}{1.34}\}n^{-1/5}$, where ***b*** depends upon the kernel which is the default kernel in statistical program STATA.

Smoothing techniques are powerful and widely used methods in nonparametric estimation. The CPS wage and salary income data in this paper for all individuals as well as race and gender are binned using the Epanechnikov kernel density function.[14]

3.5 Goodness-of-Fit Test

There are two approaches to test the statistical size distribution of income. The first is the parametric approach which uses goodness-of-fit (GoF) methods (i.e., theoretical constructs) such as the Chi-Square, Anderson-Darling and Kolmogorov-Smirnov (K-S) test and compares an empirical distribution to a theoretical or reference distribution. The tests are based on complex statistical theory and require sophisticated statistical computer programs. The second is the nonparametric approach which is more intuitive and utilizes graphical properties of a given distribution to test the true underlining nature of a distribution such as probability-probability plots (P-P plots) and quantile-quantile plots (Q-Q plots). Q-Q plots are frequently applied to empirical income data for evaluating the fit of an assumed distribution visa-a-via visual assessment of the linearity of the pattern of points on the plot. I apply both approaches parametric (Kolmogorov-Smirnov test and non-parametric (Q-Q plot) procedures to test whether CPS personal full-time wage and salary income is exponentially distributed.

Parametric goodness-of-fit statistical methods are grounded on a number of assumptions that must be held in order to get valid results. In the case of the exponential distribution (equation 4) the assumptions that form the foundation are the theoretical (scale parameter) the mean, λ; standard deviation λ (equal to the mean); coefficient of skewness (equal to 2); coefficient of kurtosis (equal to 9); and coefficient of variation (equal to 1), respectively.

$$f = \frac{1}{\lambda}e^{-x/\lambda}; domain[f] = \{x, 0, \infty\}\&\&\{\lambda > 0\}; \quad (4)$$

[14] The procedure of binning and randomly choosing bin width is argued by some researchers as misleading and an inappropriate analytical tool for adequately determining statistical distributions. In order to eliminate the subjective process of choosing a bin width, we applied the Epanechnikov kernel density function. Thus the authors' remove any subjective input which some researchers refer to as "eye balling" the data to determine bin width.

3.6 Kolmogorov-Simirnov Goodness-of-FitTest

I apply the Kolmogorov (K-S) goodness-of-fit test to determine whether or not the size distribution of personal full-time wage and salary income (FTWSI) respondents is exponential. Therefore, establishing the true underlying distribution of FTWSI depends on the correct implementation of statistical procedures. The GoF tests are for the most part grounded on two types of distributional elements: the density function (PDF) and the cumulative function (CDF). The K-S test uses the cumulative distribution function method and is classified as a distance test.

Thus the K-S test uses both a theoretical (assumed, expected or reference) CDF distribution F_0 and actual (empirical) CDF distribution F_n at each income data point. The K-S test defines the maximum distance $|F_0 - F_n|$ between the theoretical and empirical distribution. The approach is twofold:

$$F_0(r_i) = P_0(r \leq -r_i) = CDF(r_i) \tag{5}$$

Thus $F_0(r_i)$ is the assumed cumulative distribution function evaluated at r_i and $F_n(r_i)$ is the empirical distribution function obtained by the proportion of the data smaller than r_i in the individual wage and salary income data of size n.

$$F_n(r_i) = \frac{\#ofr's \leq r_i}{n} = \frac{i}{n}; i = 1, \ldots, n \tag{6}$$

Hence, we can define: $D+= F_n - F_0$ and $D-= F_0 - F_{n-1}$ for each individual wage and salary income data point r_i.

The K-S test statistics is defined as:

D = Maximum of all D+ and D- (≤ 0); for I = 1,…,n (7)

The test statistic D defines the maximum distance between the theoretical distribution function (CDF) which is assumed and the empirical distribution function (CDF) which is based on the CPS personal full-time wage and salary income data. If D is small the two CDFs are said to derive from the similar populations (i.e., exponentially distributed population).

3.7 Boltzmann-Gibbs Distribution

The Boltzmann-Gibbs probability density function (equation 8) is used to demonstrate that wage and salary income is exponentially distributed and has the property that particle velocities are distributed in an average sample of particles; domain$[f] = \{r, 0, \infty\} and \{\sigma > 0\}$, where $\sigma = \sqrt{Tk_B/r}$ and the average temperature is $2\sqrt{\frac{2}{\pi}}\sigma$. Where, (Equation 10) is the cumulative distribution function.

$$f = \frac{\sqrt{2/\pi}}{\sigma^3} r^2 e^{-\frac{x^2}{2\sigma^2}} \tag{8}$$

$$CDF = \int_0 + [-\frac{e^{-\frac{x^2}{2\sigma^2}}\sqrt{\frac{2}{\pi}}r}{\sigma} + Erf[\frac{r}{\sqrt{2}\sigma}] \tag{9}$$

The Boltzmann-Gibbs distribution for full-time wage and salary workers in the United States defined as N >> 1 workers and each full-time wage and salary worker earns an income r_i and the probability density function is P(r) and each full-time wage and salary worker 's income is between r and d(r) equal to NP(r)d(r). Thus, the stationary distribution P(r) corresponds to the state of thermodynamic equilibrium. In this state full-time wage and salary workers income strongly fluctuates but the overall probability distribution P(r) does *not* change. Therefore, the equilibrium function P(r) can be obtained in the same manner as the equilibrium distribution function of energy P(ε) in physics. Moreover, if we divide the system into two sub-systems one and two and taking into account the full-time wage and salary income is conserved as well as additive; $r = r_1 + r_2$, whereas the probability is multiplicative; $P = P_1 P_2$, then $P(r_1 + r_2) = P(r_1)P(r_2)$ and its solution is $P(r) = Ce^{-r/T_r}$ and is analogous to the exponential function, $P(r) \propto exp(-r/T_r)$, where T is the average temperature and defined as $T_r = \langle r \rangle = \int_0^\infty dr' r' p(r')$. The paper analyzes full-time wage and salary income is defined by ***r***, so as a fraction of individuals with wage and salary income between ***r*** and ***r*** + dr is P(r) dr. The Boltzmann-Gibbs temperature (average personal full-time wage and salary income fitted by MLE regression) is defined by T_r and ***c*** is a normalizing constant, $\int_0^\infty P(r)\, dr = 1$, the probability P(r) of a physical system or sub-system in a state with the income ***r*** is given by equation (11). The expected value of any given wage and salary income is derived by equation (11). Moreover, the average T_r is equal average personal wage and salary income. The sum of the probabilities of T_r is equal to 1.

$$P(r_1) = exp(\tfrac{-r}{T_r})/T_r \tag{10}$$

$$EXP[r] = \frac{\sum_i r_i e^{-r_i/T_r}}{\sum_i e^{-\varepsilon_i/T_r}} \tag{11}$$

Therefore, $P(r_1)$ represents each individual's wage and salary income and the average wage and salary income are defined as $\int_0^\infty r P_1(r)\, dr = T_r$ which is analogous to "temperature T " in the Boltzmann-Gibbs distribution. The cumulative probability $c(r) = \int_r^\infty dr' P(r')\, dr$ is the probability to have income above ***r,*** C(0)=1. As noted previously, the Boltzmann-Gibbs distribution is analogous to the exponential function $P(r) \propto exp(-r/T_r)$, where T_r is the average income temperature and defined as $T_r = \langle r \rangle = \int_0^\infty dr' r' p(r')$, and defined [3] as the "income temperature," when P(r) is exponential, $C(r) \propto exp(-r/T_r)$ is also exponential.

4 The Empirics: Full-Time Wage and Salary Respondents

This paper tests the hypothesis that size distribution of labor incomes in the United States from 1996 to 2008 is approximately distributed as an exponential. For individual wage and salary incomes (r), an exponential distribution function has a probability distribution $(r) = \left(\frac{1}{T_r}\right) e^{-\frac{r}{T_r}}$. The cumulative probability distribution for incomes above r is $C(r) = e^{-\frac{r}{T_r}}$, which is parameter free in relative income. Since $\lambda \equiv \mathrm{r}/\mathrm{T_r}$, $log\ C\ (r) = -\left(\frac{1}{T_r}\right) r$, we can estimate, "$T_r$" by means of a maximum likelihood regression of $\log \mathrm{C}(\mathrm{r})$ on r, and use this to construct "relative" income.

I argue that full-time wage and salary income of all individuals will "cluster*"* around the line $log\ C\ (r) = -\lambda$. Thus $\lambda = \frac{r}{T_r}$ is the relative personal full-time wage and salary income of individual

i with respect to the income temperature T_r in the exponential Boltzmann-Gibbs Law. In addition, λ being a purely scale transformation, T_r transformation has the advantage of making personal full-time wage and salary income data unit-free while not affecting the shape of the original empirical distribution. Therefore, CPS respondents with a full-time wage and salary income are exponentially distributed and follow the exponential Boltzmann-Gibbs Law[15].

In Figure 1, I present all CPS personal full-time wage and salary income in the United States for years 1996, 2000, 2004 and 2008 and rescaled full-time wage and salary income r is normalized by the average full-time wage and salary income (temperature) T_r in the exponential part of the distribution. Figure 1is in log-linear scale. The full-time wage and salary income is top-coded at $150,000 dollars for years 1996 and 2000 and top-coded at $200,000 dollars for years 2004 and 2008. For years 1996, 2000, 2004 and 2008, we observe for all CPS respondents, in the original full-time wage and salary income data, follow the exponential Boltzmann-Gibbs law. This is evident by the linearity of the binned full-time wage and salary income data (in log-linear scale) fitted to the exponential function $P(r) = c\,exp(\frac{-r}{T_r})$. The deviation from linearity occurs at "extremely" low probabilities and may be due to artifacts or sampling error in the data and/or high variances among high full-time wage and salary income earners Cockshott et al., (2009). A pattern which emerges is that deviation from linearity occurs at probabilities that range from 0.02 to 0.05. In addition, deviation from linearity occurs within the specified range 0.02 to 0.05, irrespective of top coding. Therefore, Figure 1 demonstrates that full-time wage and salary income for all CPS respondents follow the exponential Boltzmann-Gibbs law for years 1996, 2000, 2004 and 2008.

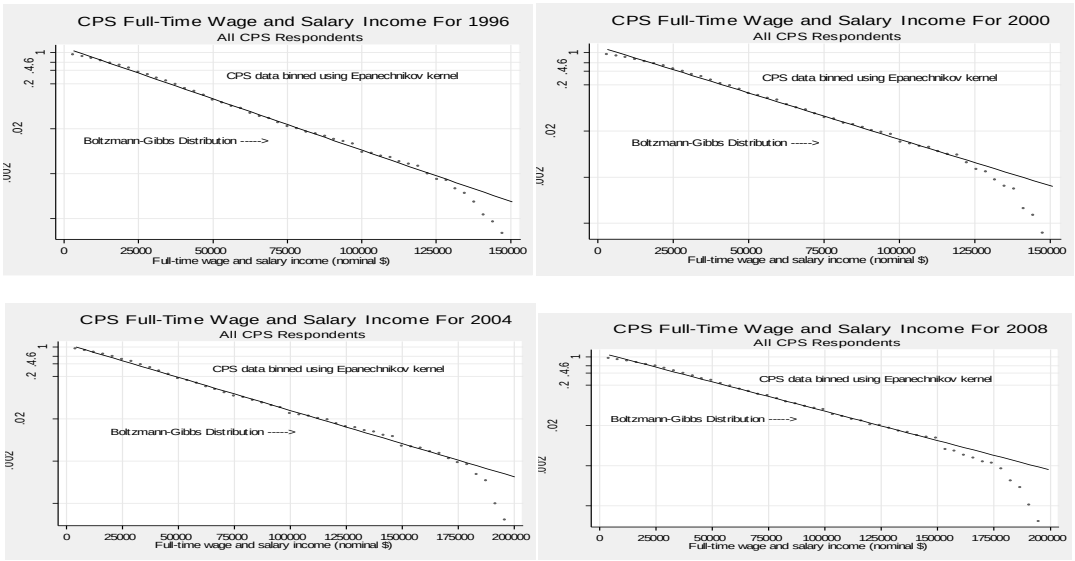

Figure 1: The cumulative probability *C(r)* and probability density *P(r)* plotted in the log-linear scale vs. r/T_r, for all personal full-time wage and salary income *r* normalized by the average income T_r (temperature) in the exponential part of the distribution for years 1996, 2000, 2004 and 2008. Overall, the fit is good for years 1996 and 2000 for incomes below $125,000 and for years 2004 and 2008 for incomes below $150,000. Deviation from exponential is evident for probabilities less than two percent of the population.

[15] I apply the Epanechnikov density kernel to estimate the optimal bandwidth for CPS personal wage and salary income data in the United States for years 1996 to 2008.

Figures 2 presents the rescaled all *full-time* wage and salary income $\frac{r}{T_r}$. As noted earlier T_r represents wage and salary "income temperature" by fitting the data using MLE for the years 1996 to 2008 as depicted in (Figure 1) in log-linear scale, years 1996, 2000, 2004 and 2008 except for extreme low probability levels (i.e., high incomes), the majority of full-time wage and salary CPS respondents' cluster on the exponential curve when plotted vs $\frac{r}{T_r}$. While the log-log scale shows the rescaled wage and salary incomes fitted to the exponential distribution cluster on a straight-line over a period of twelve years and demonstrate that the shape is "extremely" stable and does not change in time, despite small change in nominal income Yakovenko (2009). Subsequently, CPS income r that deviates away from the exponential distribution16 is threefold. First, it may be due to random sampling fluctuations Silva and Yakovenko (2005a; 2005b). Second, a reasonable explanation of deviation away from the exponential distribution (i.e., whole range of the distribution) might indicate that the CPS income distribution is bimodal and consequently, result in a major mode at the high end (low incomes-high probabilities) and a minor mode at the low end (high incomes-low probabilities). Third, wage and salary income may include high level management or executives that earn very incomes relative to the majority of the population.

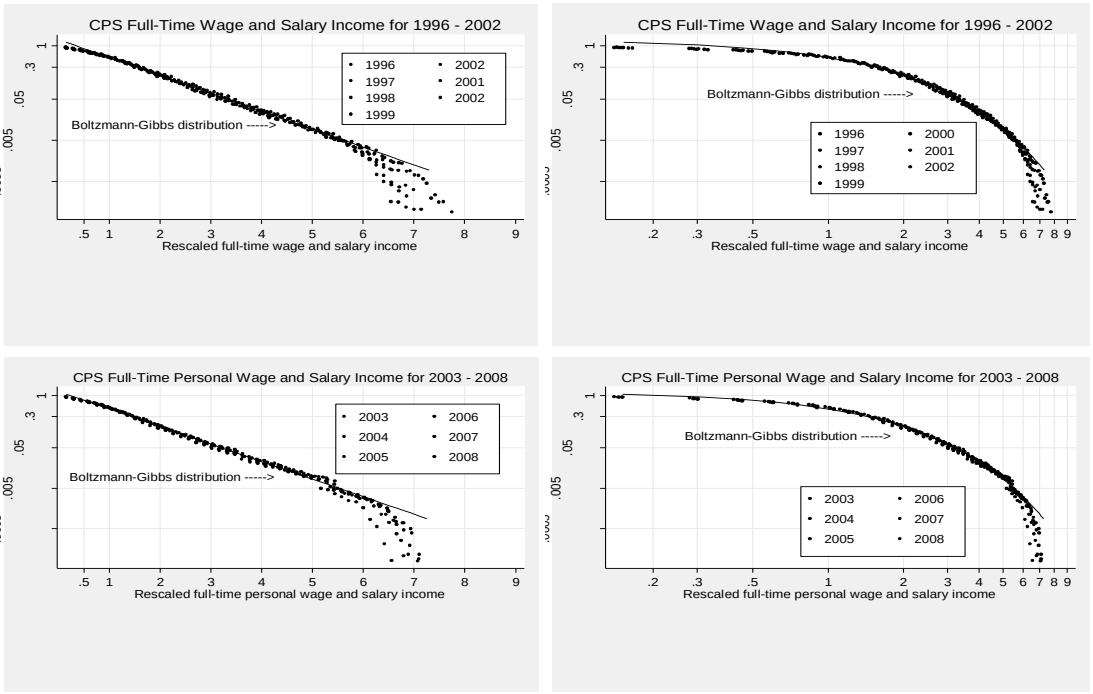

Figure 2: The cumulative probability C(r) and probability density P(r) vs. r/T_r, for all personal full-time (only) wage and salary income r normalized by the average income T_r (temperature) in the exponential part of the distribution. The left-side (log-linear) plots for years 1996 to 2002 and the right-side (log-log) plots for years 2003 to 2008. Rescaled wage and salary incomes cluster on Boltzmann-Gibbs distribution in log-linear and log-log scale for years 1996 to 2008. Deviation occurs for probabilities below two percent of the population.

[16] According to Balakrishman and Basu (1996) there will be deviations from linearity due to random sampling fluctuations, but the larger the sample size the greater the tendency toward a straight-line. The variance of each point in the upper tail of the exponential distribution has a higher variance than those in the restricted lower tail. This fact should be remembered when assessing the deviation from a straight-line and a greater leeway must be given for point in the upper tail.

Figure 3 shows all CPS full-time wage and salary respondents with income $1 to $150,000 dollars for year 1996 and $1 to $200,000 dollars for year 2008, respectively. The points in a QQ-plot lie near the diagonal y = x or 45-degree line. Location differences will lead to Q-Q plots where the points lie above or below this line. Scale differences will lead to Q-Q plots where the points lie first on the one side of the line y = x, next intersect this line, and then lie on the other side of the line, points collapsing on the 45-degree line or with a slope approximately equal to one illustrates that full-time wage and salary income is distributed exponentially. The Q-Q plots display All CPS full-time wage and salary income earners versus the theoretical exponential function. It is evident that full-time wage and salary income for years 1996 and 2007, respectively, diverges somewhat from the theoretical exponential function at the lower tail but demonstrates a reasonable fit for the middle to upper part of the distribution of full-time wage and salary income levels. Therefore, one can make a judicious argument based on Figures 1, 2 and 3 that labor income (CPS Wage and Salary Respondents) are distributed approximately as an exponential for years 1996 to 2008.

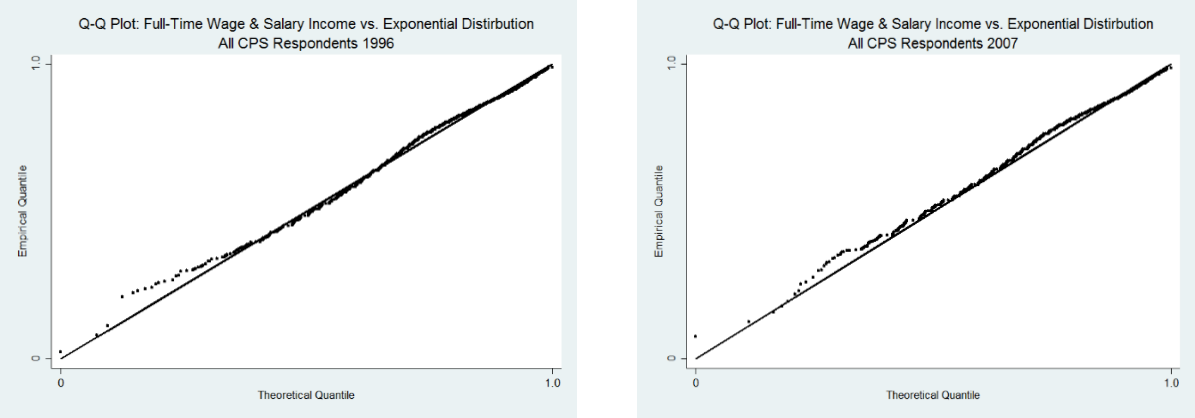

Figure 3: Q-Q plot for All full-time wage and salary income vs. theoretical exponential distribution in the log-log scale for years 1996 and 2007, respectively. The Q-Q plots show a good fit for the majority of the wage and salary income but deviated at low probabilities (i.e., high incomes).

According to Dragulescu and Yakovenko (2001b) a theoretical exponential distribution has a Gini coefficient equal to 0.50. In Table 1, I show the Gini coefficient of all CPS full-time wage and salary income respondents for years 1996 to 2008. For the majority of the years for All CPS income respondents the Gini coefficient is 0.48, except for years 2000 and 2008 which is 0.47. The average Gini coefficient for the combined years is 0.48. Even though the average Gini coefficient Gini is not equal to the theoretical exponential distribution Gini coefficient of 0.50, it's still within the acceptable bounds with a 0.02 margin of error.

Table 1: Gini coefficient for personal full-time wage and salary income for years 1996 to 2008.

Year	**Gini Coefficient**
1996	0.48
1997	0.48
1998	0.48
1999	0.48
2000	0.47
2001	0.48
2002	0.49
2003	0.48
2004	0.48
2005	0.48
2006	0.48
2007	0.48
2008	0.47
Average	**0.48**

Note: The number in each cell is the Gini coefficient for all CPS respondents.

In addition, I apply the Kolmogorov-Smirnov (K-S) test to CPS full-time wage and salary income to determine if the underlying distribution is exponential. The K-S tests the theoretical exponential cumulative distribution with the empirical cumulative distribution of full-time CPS wage and salary respondents to determine statistical equivalence. The KS-test has the advantage of making no assumption about the distribution of data. In addition, it is distribution free but with a caveat, it's sensitive to the middle of the distribution. The aim of the K-S test is to determine if the cumulative distribution of CPS full-time wage and salary income data produces different results from the theoretical cumulative exponential distribution. If the CPS cumulative distribution and the theoretical cumulative exponential distribution outcomes are statistically "the same", and can reasonably assume that the CPS income data is exponentially distributed. The K-S test assigns the D-statistic and P-value to the test results; P-*values* report if the cumulative distributions differ significantly. If the P-value is "smaller" then the determined level of significance, we reject the null hypothesis that the two cumulative distributions are not distinguishable from one another. The null hypothesis in the K-S test states that the empirical cumulative distribution cannot be distinguished from the theoretical cumulative distribution. The alternative hypothesis is that the two cumulative distributions differ and do not come from the same population.

The Kolmogorov-Smirnov test assumes that the parameter(s) of the test distribution are specified in advance. The duplicate CPS full-time wage and salary income observations were deleted[17] from the

[17] It's important to note that prior to conducting the K-S test (only) duplicate or reoccurring wage and salary income observations were deleted from the CPS income data. For example if income of $100 $100 $100 was in the CPS income data, the last two observations were dropped and only the first income of $100 was kept. CPS income data is rounded-off for a considerable number of observations but not for the total CPS raw income data.

data and the sample mean was used to conduct the K-S test on the CPS income data. In Table 2 we illustrate the K-S test for the cumulative distribution for full-time wage and salary income respondents versus the theoretical cumulative exponential distribution for years 1996, 2000, 2004 and 2008. The results of the K-S test and the corresponding p-values show if a difference exists between the two cumulative distributions at a five percent significance level for all CPS respondents. The K-S test shows that full-time CPS wage and salary respondents are exponentially distributed.

Table 2: Kolmogorov-Smirnov test of empirical cumulative distribution of all, male, female, white and African American full-time wage and salary income to theoretical cumulative exponential distribution for years 1996, 2000, 2004 and 2008.

Full-time CPS Wage and Salary Respondents	**Kolmogorov-Smirnov**	**P-Value**
1996	-0.009	0.366*
2000	-0.007	0.546*
2004	-0.012	0.147*
2008	-0.014	0.119*

Note: duplicate CPS wage and salary income observations were deleted from the data set. (*) defines 5% significance level.

5 Conclusion

In this paper I analyzed the size distribution using parametric and non-parametric methods to test whether full-time CPS wage and salary respondents are exponentially distributed for the general population. I used Q-Q plots, cumulative distribution plots in both log-linear and log-log scale and the Kolmogorov-Smirnov test to examine the true underlying distribution of full-time CPS wage and salary income data.

The weakened labor situation has coincided with a number of significant historical changes in America. Politically, since the inflationary seventies, there has been a shift in attitudes towards government, business and labor unions. Social norms have changed: lay-off s are now more acceptable than they once were as are enormous salaries for CEOs and compensation for Wall Street professionals. Federal oversight has been reduced, business widely deregulated, and labor unions have been weakened. Minimum wages have been raised only modestly. At the same time, tight monetary policies have prevailed to keep inflation low, resulting in high average rates of unemployment, which may well have had a persistent depressing effect on wages and salaries. The most widely accepted explanation is that technology requires more increasingly educated workers, thus pushing wage increases towards better educated workers. Finally, increasing trade-imports are now fifteen percent of GDP in America-and-off

---

Furthermore, in my view deleting duplicate observation or “ties” does not alter the distribution function but does affect the outcome of the K-S test. The area were the K-S test is affected from removing duplicate observation is the middle of the distribution-where the K-S- test is most sensitive. Moreover, the affect is limited because we are comparing and testing the cumulative distributions of CPS income data and theoretical exponential distribution and not the individual income data points.

shoring of jobs have had an impact on wage growth. Stagnating wages may well be a consequence of all and certainly several of these factors Madrick and Papanikolaou (2010).

Understanding changes in the distribution of CPS income data over the entire range is beneficial and provides fruitful information and sheds light on income inequality in the United States. The results indicate that full-time CPS wage and salary respondents are exponentially distributed. The shape of the probability density function reveals important stylized facts: Figures 1 and 2 illustrates that the data sets for different years (1996 to 2008) cluster onto a straight line. It demonstrates that the underlying shape of the distribution of income for all wage and salary income respondents and moreover, is very stable and does not change over time in spite of the small increases in nominal income Dragulescu and Yakovenko (2001a; 2001b). From Figure 3, the Q-Q plots demonstrate for years 1996 and 2007 that wage and salary income cluster onto a theoretical exponential distribution. It demonstrates that the underlying shape of the distribution of income for all wage and salary income respondents follows an exponential distribution except for very low probabilities or very high incomes. Therefore, illustrating that full-time CPS wage and salary income is exponentially distributed.

To properly address the issue of income inequality from a policy perspective, one must first make the distinction between labor income (exponential) and asset income (Pareto). Thus, Fiscal policy which fosters economic growth has not adequately addressed income inequality and the widening gap between rich and poor because no distinction was made to differentiate the two types of income. While monetary policy, especially, unconventional monetary policy which focuses on quantitative easing and asset buybacks has only exacerbated income inequality. Because unconventional monetary policy focused primarily on the asset side while disregarding the labor side of income.

The primary focus of fiscal and monetary policy must be on labor income to reduce poverty and income inequality. Policies that adequately reduce income inequality are increasing the minimum wage to a living wage, an increase the tax rate on higher non-labor income (asset income), stronger unions, affordable housing, student loan forgiveness and strengthen safety net programs.

References

Balakrishnan N. and A.P. Basu. (1996). Exponential Distribution: Theory, Methods and Applications. CRC Press, First (Ed.).

Burkhauser, R.C., SFeng, S.P. Jenkins and J. Larrimore. (2008). Estimating Trends in US Income Inequality using Current Population Survey: The Importance of Controlling for Censoring. Center for Economics Studies (CES) Research Paper.

Chatterjee A. and S. Yarlagadda, & B. K. Chakrabarti. (2008). Econophysics of Wealth Distributions, Springer.

Cockshott W.P. and A.F. Cottrell and G.J. Michaelson and I.P Wright and V.M. Yakovenko. (2009). Classical Econophysics. Routledge advances in experimental and computable economics, No. 12.

Dragulescu, A.A. and V.M. Yakovenko. (2000). Statistical Mechanics of Money. The European Physical Journal B 17, 723-729.

Dragulescu A.A and V.M. Yakovenko. (2001a). Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. Physica A, vol. 299, pp. 213–221.

Dragulescu A.A. and V.M. Yakovenko. (2001b). Evidence for the exponential distribution of income in the USA," The European Physical Journal B, vol. 20, pp.585-589.

Guerello, Chiara, (2018). Conventional and unconventional monetary policy vs. households income distribution: An empirical analysis for the Euro Area, Journal of International Money and Finance, Elsevier, vol. 85(C), pages 187-214.

Jones, M.C. and Marron, J. S. Marron and Sheather, S.J. (1996). A Brief Survey of Bandwidth Selection for Density Estimation (1996). Journal of the American Statistical Association, Vol. 91, No. 433 (Mar., 1996), pp.401-407.

Kleiber C. and S.K. (2003). Statistical Size Distribution in Economics and Actuarial Sciences. WILEY-INERSCIENCE, Chapter 1, pp. 1-19.

Levy, M. and Solomon, S. (1997). Physica A: Statistical Mechanics and its Applications, vol. 242, issue 1, 90-94.

Madrick, J. & Papanikolaou, N. (2010). The stagnation of male wages in the US, International Review of Applied Economics, 24:3, 309-318.

Mandelbrot B.B. (1960). The Pareto-Levy Law and the Distribution of Income. International Economic Review, vol. 1, pp.79-106.

Mandelbrot B.B. (1963). New Methods in Statistical Economics. Journal of Political Economy, vol. 71, pp. 421-440, 1963.

Mumtaz, H., & Theophilopoulou, A. (2017). The impact of monetary policy on inequality in the UK. An empirical analysis. European Economic Review, 98, 410-423.

Papanikolaou, N. (2020). Markov-Switching Model of Family Income Quintile Shares, Atlantic Economic Journal. (pending publication vol. 48, issue 1).

Pareto V. (1897). Cours d'Economie Politique. Lausanne.

Pfeffer, F. T., Danziger, S., & Schoeni, R. F. (2013). Wealth Disparities before and after the Great Recession. The Annals of the American Academy of Political and Social Science, 650(1), 98–123. https://doi.org/10.1177/0002716213497452

Saiki, A., & Frost, J. (2014). Does unconventional monetary policy affect inequality? Evidence from Japan. Applied Economics, 46(36), 4445-4454

Shaikh, A., Papanikolaou, N., and Wiener, N. (2014), Race, gender and the econophysics of income distribution in the USA, Physica A: Statistical Mechanics and its Applications, 415, issue C, p. 54-60.

Silva, A.C. and V.M. Yakovenko. (2005a). Temporal Evolution of the "Thermal" and "Superthermal" Income Classes in the USA during 1983-2001. Europhysics Letter 69, 304-310.

Silva A.C. and V.M. Yakovenko. (2005b). Two-Class Structure of Income Distribution in the USA: Exponential Bulk and Power-Law Tail. In Chatterjee, Yarlagadda, and Chakrabarti, Econophysics of Wealth Distributions, pp. 15-23.

Sima Siami-Namini, Conrad Lyford and A. Alexandre Trindade. (2020). The Effects of Monetary Policy Shocks on Income Inequality Across U.S. States, Economic Papers: A journal of applied economics and policy, 10.1111/1759-3441.12279.

Silverman B.W. (1986). Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability Series.

Smith, Adam. (1776). The Wealth of Nations. (Ed.) Edwin Cannan (1937), Modern Library, New York.

Souma W. (2001). Universal Structure of the Personal Income Distribution, Fractals. World Scientific Publishing Co., Vol. 9, No. 3.

Souma W. (2002). Physics of Personal Income. ATR Human Information Science Laboratories, Kyoto, Japan.

Tarozzi A. (2009). A primer in density estimation. ECON 214 Lecture Notes, http://econ.duke.edu/~taroz/LectureNotes214NP.pdf

Walras, L. (1899). Éléments d'économie politique pure (1899), 4th ed.; 1926, éd. définitive), in English, Elements of Pure Economics (1954), trans. William Jaffé.

Yakovenko, V.M. (2009). Econophysics, Statistical Mechanics Approach to. In R.A. Meyers (Ed.), Encyclopedia of Complexity and System Science. Springer.

Yakovenko V.M. and J.B. Rosser (2009). Colloquium: Statistical Mechanics of Money, Wealth, and Income, Econophysics. Review of Modern Physics, vol. 81, pp. 1703-1725.