Eur. Phys. J. B **20**, 585–589 (2001)

THE EUROPEAN PHYSICAL JOURNAL B

EDP Sciences

Evidence for the exponential distribution of income in the USA

A. Drăgulescu and V.M. Yakovenko[a]

Department of Physics, University of Maryland, College Park, MD 20742-4111, USA

Received 21 August 2000

Abstract. Using tax and census data, we demonstrate that the distribution of individual income in the USA is exponential. Our calculated Lorenz curve without fitting parameters and Gini coefficient 1/2 agree well with the data. From the individual income distribution, we derive the distribution function of income for families with two earners and show that it also agrees well with the data. The family data for the period 1947–1994 fit the Lorenz curve and Gini coefficient $3/8 = 0.375$ calculated for two-earners families.

PACS. 87.23.Ge Dynamics of social systems – 89.90.+n Other topics in areas of applied and interdisciplinary physics – 02.50.-r Probability theory, stochastic processes, and statistics

1 Introduction

The study of income distribution has a long history. Pareto [1] proposed in 1897 that income distribution obeys a universal power law valid for all times and countries. Subsequent studies have often disputed this conjecture. In 1935, Shirras [2] concluded: "There is indeed no Pareto Law. It is time it should be entirely discarded in studies on distribution". Mandelbrot [3] proposed a "weak Pareto law" applicable only asymptotically to the high incomes. In such a form, Pareto's proposal is useless for describing the great majority of the population.

Many other distributions of income were proposed: Levy, log-normal, Champernowne, Gamma, and two other forms by Pareto himself (see a systematic survey in the World Bank research publication [4]). Theoretical justifications for these proposals form two schools: socio-economic and statistical. The former appeals to economic, political, and demographic factors to explain the distribution of income (*e.g.* [5]), whereas the latter invokes stochastic processes. Gibrat [6] proposed in 1931 that income is governed by a multiplicative random process, which results in a log-normal distribution (see also [7]). However, Kalecki [8] pointed out that the width of this distribution is not stationary, but increases in time. Levy and Solomon [9] proposed a cut-off at lower incomes, which stabilizes the distribution to a power law.

In this paper, we propose that the distribution of individual income is given by an exponential function. This conjecture is inspired by our previous work [10], where we argued that the probability distribution of money in a closed system of agents is given by the exponential Boltzmann-Gibbs function, in analogy with the distribution of energy in statistical physics. In Section 2, we compare our proposal with the census and tax data for individual income in USA. In Section 3, we derive the distribution function of income for families with two earners and compare it with the census data. The good agreement we found is discussed in Section 4. Speculations on the possible origins of the exponential distribution of income are given in Section 5.

[a] e-mail: yakovenk@physics.umd.edu
http://www2.physics.umd.edu/~yakovenk

2 Distribution of individual income

We denote income by the letter r (for "revenue"). The probability distribution function of income, $P(r)$, (called the probability density in book [4]) is defined so that the fraction of individuals with income between r and $r + \mathrm{d}r$ is $P(r)\,\mathrm{d}r$. This function is normalized to unity (100%): $\int_0^\infty P(r)\,\mathrm{d}r = 1$. We propose that the probability distribution of individual income is exponential:

$$P_1(r) = \exp(-r/R)/R, \tag{1}$$

where the subscript 1 indicates individuals. Function (1) contains one parameter R, equal to the average income: $\int_0^\infty r\,P_1(r)\,\mathrm{d}r = R$, and analogous to temperature in the Boltzmann-Gibbs distribution [10].

From the Survey of Income and Program Participation (SIPP) [11], we downloaded the variable TPTOINC (total income of a person for a month) for the first "wave" (a four-month period) in 1996. Then we eliminated the entries with zero income, grouped the remaining entries into bins of the size 10/3 k$, counted the numbers of entries inside each bin, and normalized to the total number of entries. The results are shown as the histogram in Figure 1, where the horizontal scale has been multiplied by 12 to convert monthly income to an annual figure. The solid line represents a fit to the exponential function (1). In the inset, plot A shows the same data with the logarithmic vertical scale. The data fall onto a straight line,

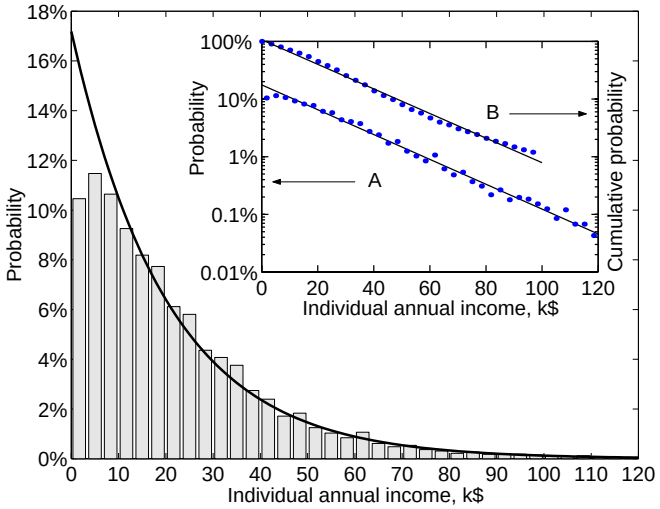

Fig. 1. Histogram: Probability distribution of individual income from the US. Census data for 1996 [11]. Solid line: Fit to the exponential law. Inset plot A: The same with the logarithmic vertical scale. Inset plot B: Cumulative probability distribution of individual income from PSID for 1992 [12].

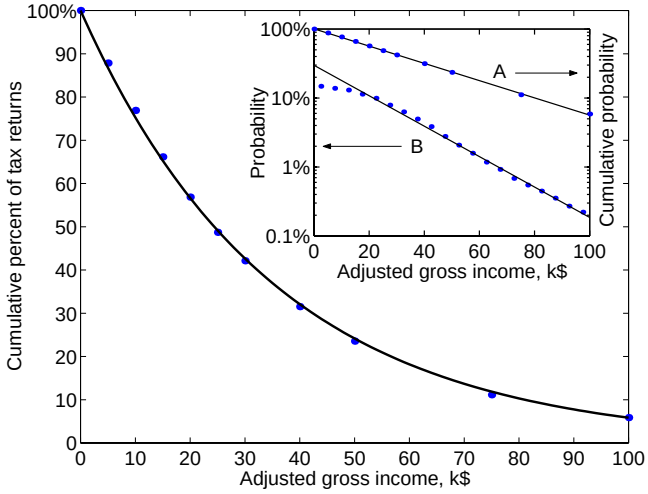

Fig. 2. Points: Cumulative fraction of tax returns *vs.* income from the IRS data for 1997 [13]. Solid line: Fit to the exponential law. Inset plot A: The same with the logarithmic vertical scale. Inset plot B: Probability distribution of individual income from the IRS data for 1993 [14].

whose slope gives the parameter R in equation (1). The exponential law is also often written with the bases 2 and 10: $P_1(r) \propto 2^{-r/R_2} \propto 10^{-r/R_{10}}$. The parameters R, R_2 and R_{10} are given in line (c) of Table 1.

Plot B in the inset of Figure 1 shows the data from the Panel Study of Income Dynamics (PSID) conducted by the Institute for Social Research of the University of Michigan [12]. We downloaded the variable V30821 "Total 1992 labor income" for individuals from the Final Release 1993 and processed the data in a similar manner. Shown is the cumulative probability distribution of income $N(r)$ (called the probability distribution in book [4]). It is defined as $N(r) = \int_r^\infty P(r')\,\mathrm{d}r'$ and gives the fraction of individuals with income greater than r. For the exponential distribution (1), the cumulative distribution is also exponential: $N_1(r) = \int_r^\infty P_1(r')\,\mathrm{d}r' = \exp(-r/R)$. Thus, R_2 is the median income; 10% of population have income greater than R_{10} and only 1% greater than $2R_{10}$. The points in the inset fall onto a straight line in the logarithmic scale. The slope is given in line (a) of Table 1.

Table 1. Parameters R, R_2, and R_{10} obtained by fitting data from different sources to the exponential law (1) with the bases e, 2, and 10, and the sizes of the statistical data sets.

	Source	Year	R (\$)	R_2 (\$)	R_{10} (\$)	Set size
a	PSID [12]	1992	18,844	13,062	43,390	1.39×10^3
b	IRS [14]	1993	19,686	13,645	45,329	1.15×10^8
c	$\mathrm{SIPP_p}$ [11]	1996	20,286	14,061	46,710	2.57×10^5
d	$\mathrm{SIPP_f}$ [11]	1996	23,242	16,110	53,517	1.64×10^5
e	IRS [13]	1997	35,200	24,399	81,051	1.22×10^8

The points in Figure 2 show the cumulative distribution of tax returns *vs.* income in 1997 from column 1 of Table 1.1 of reference [13]. (We merged 1 k\$ bins into 5 k\$ bins in the interval 1–20 k\$.) The solid line is a fit to the exponential law. Plot A in the inset of Figure 2 shows the same data with the logarithmic vertical scale. The slope is given in line (e) of Table 1. Plot B in the inset of Figure 2 shows the distribution of individual income from tax returns in 1993 [14]. The logarithmic slope is given in line (b) of Table 1.

While Figures 1 and 2 clearly demonstrate the fit of income distribution to the exponential form, they have the following drawback. Their horizontal axes extend to $+\infty$, so the high-income data are left outside of the plots. The standard way to represent the full range of data is the so-called Lorenz curve (for an introduction to the Lorenz curve and Gini coefficient, see book [4]). The horizontal axis of the Lorenz curve, $x(r)$, represents the cumulative fraction of population with income below r, and the vertical axis $y(r)$ represents the fraction of income this population accounts for:

$$x(r) = \int_0^r P(r')\,\mathrm{d}r', \quad y(r) = \frac{\int_0^r r'P(r')\,\mathrm{d}r'}{\int_0^\infty r'P(r')\,\mathrm{d}r'}. \qquad (2)$$

As r changes from 0 to ∞, x and y change from 0 to 1, and equation (2) parametrically defines a curve in the (x, y)-space.

Substituting equation (1) into equation (2), we find

$$x(\tilde{r}) = 1 - \exp(-\tilde{r}), \quad y(\tilde{r}) = x(\tilde{r}) - \tilde{r}\exp(-\tilde{r}), \qquad (3)$$

where $\tilde{r} = r/R$. Excluding $\tilde{r}$, we find the explicit form of the Lorenz curve for the exponential distribution:

$$y = x + (1-x)\ln(1-x). \qquad (4)$$

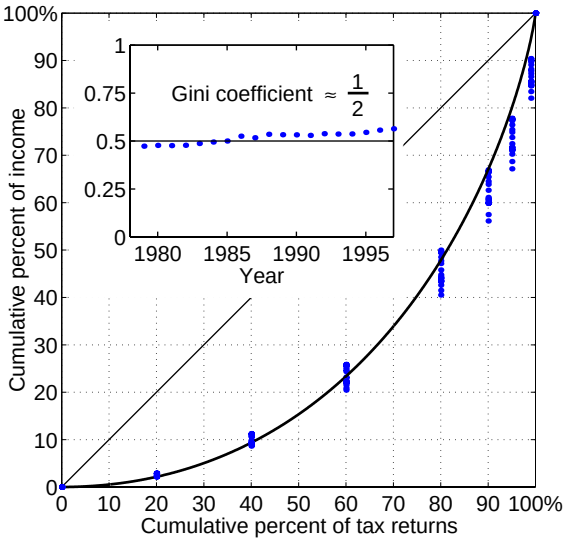

Fig. 3. Solid curve: Lorenz plot for the exponential distribution. Points: IRS data for 1979–1997 [15]. Inset points: Gini coefficient data from IRS [15]. Inset line: The calculated value 1/2 of the Gini coefficient for the exponential distribution.

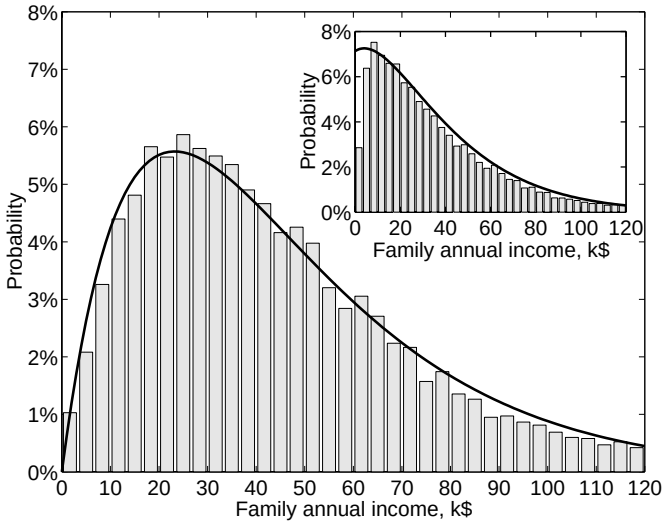

Fig. 4. Histogram: Probability distribution of income for families with two adults in 1996 [11]. Solid line: Fit to equation (5). Inset histogram: Probability distribution of income for all families in 1996 [11]. Inset solid line: $0.45P_1(r) + 0.55P_2(r)$.

R drops out, so equation (4) has no fitting parameters.

The function (4) is shown as the solid curve in Figure 3. The straight diagonal line represents the Lorenz curve in the case where all population has equal income. Inequality of income distribution is measured by the Gini coefficient G, the ratio of the area between the diagonal and the Lorenz curve to the area of the triangle beneath the diagonal: $G = 2\int_0^1 (x - y)\,\mathrm{d}x$. The Gini coefficient is confined between 0 (no inequality) and 1 (extreme inequality). By substituting equation (4) into the integral, we find the Gini coefficient for the exponential distribution: $G_1 = 1/2$.

The points in Figure 3 represent the tax data during 1979–1997 from reference [15]. With the progress of time, the Lorenz points shifted downward and the Gini coefficient increased from 0.47 to 0.56, which indicates increasing inequality during this period. However, overall the Gini coefficient is close to the value 0.5 calculated for the exponential distribution, as shown in the inset of Figure 3.

3 Income distribution for two-earners families

Now let us discuss the distribution of income for families with two earners. The family income r is the sum of two individual incomes: $r = r_1 + r_2$. Thus, the probability distribution of the family income is given by the convolution of the individual probability distributions [16]. If the latter are given by the exponential function (1), the two-earners probability distribution function $P_2(r)$ is

$$P_2(r) = \int_0^r P_1(r')P_1(r - r')\,\mathrm{d}r' = \frac{r}{R^2}\mathrm{e}^{-r/R}. \qquad (5)$$

The function $P_2(r)$ (5) differs from the function $P_1(r)$ (1) by the prefactor r/R, which reflects the phase space available to compose a given total income out of two individual ones. It is shown as the solid curve in Figure 4. Unlike $P_1(r)$, which has a maximum at zero income, $P_2(r)$ has a maximum at $r = R$ and looks qualitatively similar to the family income distribution curves in literature [5].

From the same 1996 SIPP that we used in Section 2 [11], we downloaded the variable TFTOTINC (the total family income for a month), which we then multiplied by 12 to get annual income. Using the number of family members (the variable EFNP) and the number of children under 18 (the variable RFNKIDS), we selected the families with two adults. Their distribution of family income is shown by the histogram in Figure 4. The fit to the function (5), shown by the solid line, gives the parameter R listed in line (d) of Table 1. The families with two adults and more than two adults constitute 44% and 11% of all families in the studied set of data. The remaining 45% are the families with one adult. Assuming that these two classes of families have two and one earners, we expect the income distribution for all families to be given by the superposition of equations (1) and (5): $0.45P_1(r)+0.55P_2(r)$. It is shown by the solid line in the inset of Figure 4 (with R from line (d) of Tab. 1) with the all families data histogram.

By substituting equation (5) into equation (2), we calculate the Lorenz curve for two-earners families:

$$x(\tilde{r}) = 1 - (1 + \tilde{r})\mathrm{e}^{-\tilde{r}}, \quad y(\tilde{r}) = x(\tilde{r}) - \tilde{r}^2\mathrm{e}^{-\tilde{r}}/2. \qquad (6)$$

It is shown by the solid curve in Figure 5. Given that $x - y = \tilde{r}^2\exp(-\tilde{r})/2$ and $\mathrm{d}x = \tilde{r}\exp(-\tilde{r})\,\mathrm{d}\tilde{r}$, the Gini

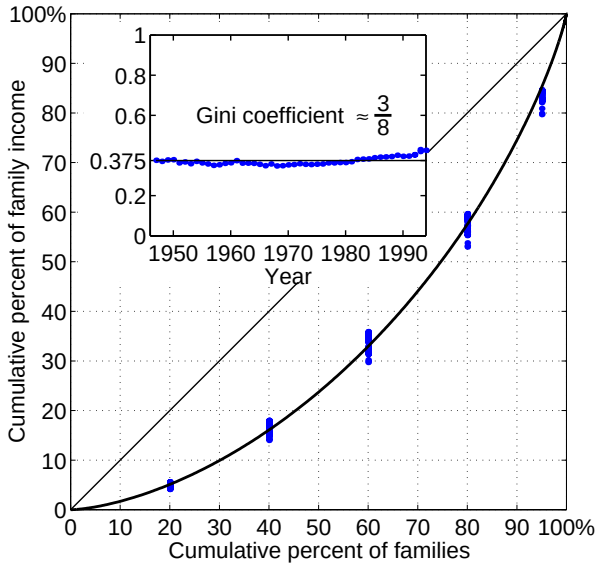

Fig. 5. Solid curve: Lorenz plot (6) for distribution (5). Points: Census data for families, 1947–1994 [17]. Inset points: Gini coefficient data for families from Census [17]. Inset line: The calculated value 3/8 of the Gini coefficient for distribution (5).

coefficient for two-earners families is: $G_2 = 2\int_0^1 (x - y)\,\mathrm{d}x = \int_0^\infty \tilde{r}^3 \exp(-2\tilde{r})\,\mathrm{d}\tilde{r} = 3/8 = 0.375$. The points in Figure 5 show the Lorenz data and Gini coefficient for family income during 1947–1994 from Table 1 of reference [17]. The Gini coefficient is very close to the calculated value 0.375.

4 Discussion

Figures 1 and 2 demonstrate that the exponential law (1) fits the individual income distribution very well. The Lorenz data for the individual income follow equation (4) without fitting parameters, and the Gini coefficient is close to the calculated value 0.5 (Fig. 3). The distributions of the individual and family income differ qualitatively. The former monotonically increases toward the low end and has a maximum at zero income (Fig. 1). The latter, typically being a sum of two individual incomes, has a maximum at a finite income and vanishes at zero (Fig. 4). Thus, the inequality of the family income distribution is smaller. The Lorenz data for families follow the different equation (6), again without fitting parameters, and the Gini coefficient is close to the smaller calculated value 0.375 (Fig. 5). Despite different definitions of income by different agencies, the parameters extracted from the fits (Tab. 1) are consistent, except for line (e).

The qualitative difference between the individual and family income distributions was emphasized in reference [14], which split up joint tax returns of families into individual incomes and combined separately filed tax returns of married couples into family incomes. However, references [13] and [15] counted only "individual tax returns", which also include joint tax returns. Since only a fraction of families file jointly, we assume that the latter contribution is small enough not to distort the tax returns distribution from the individual income distribution significantly. Similarly, the definition of a family for the data shown in the inset of Figure 4 includes single adults and one-adult families with children, which constitute 35% and 10% of all families. The former category is excluded from the definition of a family for the data [17] shown in Figure 5, but the latter is included. Because the latter contribution is relatively small, we expect the family data in Figure 5 to approximately represent the two-earners distribution (5). Technically, even for the families with two (or more) adults shown in Figure 4, we do not know the exact number of earners.

With all these complications, one should not expect perfect accuracy for our fits. There are deviations around zero income in Figures 1, 2, and 4. The fits could be improved there by multiplying the exponential function by a polynomial. However, the data may not be accurate at the low end because of underreporting. For example, filing a tax return is not required for incomes below a certain threshold, which ranged in 1999 from $2,750 to $14,400 [18]. As the Lorenz curves in Figures 3 and 5 show, there are also deviations at the high end, possibly where Pareto's power law is supposed to work. Nevertheless, the exponential law gives an overall good description of income distribution for the great majority of the population.

5 Possible origins of exponential distribution

The exponential Boltzmann-Gibbs distribution naturally applies to the quantities that obey a conservation law, such as energy or money [10]. However, there is no fundamental reason why the sum of incomes (unlike the sum of money) must be conserved. Indeed, income is a term in the time derivative of one's money balance (the other term is spending). Maybe incomes obey an approximate conservation law, or somehow the distribution of income is simply proportional to the distribution of money, which is exponential [10].

Another explanation involves hierarchy. Groups of people have leaders, which have leaders of a higher order, and so on. The number of people decreases geometrically (exponentially) with the hierarchical level. If individual income increases linearly with the hierarchical level, then the income distribution is exponential. However, if income increases multiplicatively, then the distribution follows a power law [19]. For moderate incomes below $100,000, the linear increase may be more realistic. A similar scenario is the Bernoulli trials [16], where individuals have a constant probability of increasing their income by a fixed amount.

We are grateful to D. Jordan, M. Weber, and T. Petska for sending us the data from references [13,14], and [15], to T. Cranshaw for discussion of income distribution in Britain, and to M. Gubrud for proofreading of the manuscript.

References

1. V. Pareto, *Cours d'Économie Politique* (Lausanne, 1897).
2. G.F. Shirras, Economic Journal **45**, 663 (1935).
3. B. Mandelbrot, Int. Economic Rev. **1**, 79 (1960).
4. N. Kakwani, *Income Inequality and Poverty* (Oxford University Press, Oxford, 1980).
5. F. Levy, Science **236**, 923 (1987).
6. R. Gibrat, *Les Inégalités Économique* (Sirely, Paris, 1931).
7. E.W. Montroll, M.F. Shlesinger, J. Stat. Phys. **32**, 209 (1983).
8. M. Kalecki, Econometrica **13**, 161 (1945).
9. M. Levy, S. Solomon, Int. J. Mod. Phys. C **7**, 595 (1996); D. Sornette, R. Cont, J. Phys. I France **7**, 431 (1997).
10. A. Drăgulescu, V.M. Yakovenko, `cond-mat/0001432`, Eur. Phys. J. B **17**, 723 (2000).
11. The U.S. Census data, `http://ferret.bls.census.gov/`.
12. The PSID Web site, `http://www.isr.umich.edu/src/psid`.
13. *Statistics of Income–1997, Individual Income Tax Returns*, Pub. 1304, Rev. 12-99 (IRS, Washington DC, 1999). See `http://www.irs.ustreas.gov/prod/tax_stats/soi/`.
14. P. Sailer, M. Weber, *Household and Individual Income Data from Tax Returns* (IRS, Washington DC, 1998), `http://ftp.fedworld.gov/pub/irs-soi/petasa98.pdf`.
15. T. Petska, M. Strudler, R. Petska, *Further Examination of the Distribution of Individual Income and Taxes Using a Consistent and Comprehensive Measure of Income* (IRS, 2000), `http://ftp.fedworld.gov/pub/irs-soi/disindit.exe`.
16. W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2 (John Willey, New York, 1966) p. 10.
17. D.H. Weinberg, *A Brief Look at Postwar U.S. Income Inequality*, P60-191 (Census Bureau, Washington, 1996), `http://www.census.gov/hhes/www/p60191.html`.
18. *1040: Forms and Instructions* (IRS, Washington, 1999).
19. H.F. Lydall, Econometrica **27**, 110 (1959).