Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Physica A 370 (2006) 54–59
PHYSICA A
www.elsevier.com/locate/physa

A study of the personal income distribution in Australia

Anand Banerjee[a], Victor M. Yakovenko[a,*], T. Di Matteo[b]

[a]Department of Physics, University of Maryland, College Park, MD 20742-4111, USA
[b]Department of Applied Mathematics, The Australian National University, Canberra, ACT 0200, Australia

Available online 12 May 2006

Abstract

We analyze the data on personal income distribution from the Australian Bureau of Statistics. We compare fits of the data to the exponential, log-normal, and gamma distributions. The exponential function gives a good (albeit not perfect) description of 98% of the population in the lower part of the distribution. The log-normal and gamma functions do not improve the fit significantly, despite having more parameters, and mimic the exponential function. We find that the probability density at zero income is not zero, which contradicts the log-normal and gamma distributions, but is consistent with the exponential one. The high-resolution histogram of the probability density shows a very sharp and narrow peak at low incomes, which we interpret as the result of a government policy on income redistribution.

Keywords: Econophysics; Income distribution; Australia

1. Introduction

The study of income distribution has a long history. More than a century ago, Pareto [1] proposed that income distribution obeys a universal power law, valid for all time and countries. Subsequent studies found that this conjecture applies only to the top 1–3% of the population. The question of what is the distribution for the majority (97–99%) of population with lower incomes remains open. Gibrat [2] proposed that income distribution is governed by a multiplicative random process resulting in the log-normal distribution. However, Kalecki [3] pointed out that such a log-normal distribution is not stationary, because its width keeps increasing with time. Nevertheless, the log-normal function is widely used in literature to fit the lower part of income distribution [4–6]. Yakovenko and Drăgulescu [7] proposed that the distribution of individual income should follow the exponential law analogous to the Boltzmann–Gibbs distribution of energy in statistical physics. They found substantial evidence for this in the statistical data for USA [8–11]. Also widely used is the gamma distribution, which differs from the exponential one by a power-law prefactor [12–14]. For a recent collection of papers discussing these distributions, see the book [15].

Distribution of income x is characterized by the probability density function (PDF) $P(x)$, defined so that the probability to find income in the interval from x to $x + \mathrm{d}x$ is equal to $P(x)\,\mathrm{d}x$. The PDFs for the distributions

*Corresponding author. Tel.: +1 301 4056151; fax: +1 301 3149465.
E-mail address: yakovenk@physics.umd.edu (V.M. Yakovenko).

0378-4371/$ - see front matter
doi:10.1016/j.physa.2006.04.023

discussed above have the following functional forms:

$$P(x) = \begin{cases} \frac{1}{T}\exp(-x/T) & \text{exponential,} \\ \frac{1}{xs\sqrt{2\pi}}\exp\left[\frac{-\log^2(x/m)}{2s^2}\right] & \text{log-normal,} \\ \frac{(\beta)^{-(1+\alpha)}}{\Gamma(1+\alpha,0)}x^{\alpha}\exp(-x/\beta) & \text{gamma.} \end{cases} \tag{1}$$

The exponential distribution has one parameter T, and its $P(x)$ is maximal at $x = 0$. The log-normal and gamma distributions have two parameters each: (m, s) and (β, α). They have maxima (called modes in mathematical statistics) at $x = m\mathrm{e}^{-s^2}$ and $x = \alpha\beta$, and their $P(x)$ vanish at $x = 0$. Many researchers impose the condition $P(0) = 0$ a priori, "because people cannot live on zero income". However, this assumption must be checked against the real data.

In this paper, we analyze statistical data on personal income distribution in Australia for 1989–2000 and compare them with the three functions in Eq. (1). The data were collected by the Australian Bureau of Statistics (ABS) using surveys of population. The anonymous data sets give annual incomes of about 14,000 representative individuals, and each individual is assigned a weight. The weights add up to $1.3-1.5\times10^7$ in the considered period, which is comparable to the current population of Australia of about 20 million people. In the data analysis, we exclude individuals with negative and zero income, whose total weight is about 7%. These ABS data were studied in the previous paper [4], but without weights and with the emphasis on the Pareto tail at high income. Here we reanalyze the data in the middle and low income range covering about 99% of the population, but excluding the Pareto tail. The number of data points in the Pareto tail is relatively small in surveys of population, which complicates accurate analysis of the tail.

2. Cumulative distribution function

In this section, we study the cumulative distribution function (CDF) $C(x) = \int_x^\infty P(x')\,\mathrm{d}x'$. The advantage of CDF is that it can be directly constructed from a data set without making subjective choices. We sort incomes x_n of N individuals in decreasing order, so that $n = 1$ corresponds to the highest income, $n = 2$ to the second highest, etc. When the individuals are assigned the weights w_n, the cumulative probability for a given x_n is $C = \sum_{k=1}^{n} w_k / \sum_{k=1}^{N} w_k$, i.e., $C(x)$ is equal to the normalized sum of the weights of the individuals with incomes above x. We fit the empirically constructed $C(x)$ to the theoretical CDFs corresponding to Eq. (1),

$$C(x) = \begin{cases} \exp(-x/T) & \text{exponential,} \\ \frac{1}{2}\left[1 - \mathrm{Erf}\left(\frac{\log(x/m)}{s\sqrt{2}}\right)\right] & \text{log-normal,} \\ \Gamma(1+\alpha, x/\beta)/\Gamma(1+\alpha,0) & \text{gamma,} \end{cases} \tag{2}$$

where $\mathrm{Erf}(x) = (2/\sqrt{\pi})\int_0^x \mathrm{e}^{-z^2}\,\mathrm{d}z$ is the error function, and $\Gamma(\alpha, x) = \int_x^\infty z^{\alpha-1}\mathrm{e}^{-z}\,\mathrm{d}z$.

To visualize $C(x)$, different scales can be used. Fig. 1(a) uses the log–linear scale, i.e., shows the plot of $\ln C$ vs. x. The main panel in Fig. 1(b) uses the linear–linear scale, and the inset the log–log scale, i.e., $\ln C$ vs. $\ln x$. We observe that the log–linear scale is the most informative, because the data points approximately fall on a straight line for two orders of magnitudes, which suggests the exponential distribution. To obtain the best fit in the log–linear scale, we minimize the relative mean square deviation $\sigma^2 = (1/M)\sum_{i=1}^{M}((C_e(x_i) - C_t(x_i))/C_e(x_i))^2 \approx (1/M)\sum_{i=1}^{M}\{\ln[C_e(x_i)] - \ln[C_t(x_i)]\}^2$ between the empirical $C_e(x)$ and theoretical $C_t(x)$ CDFs. For this sum, we select $M = 200$ income values x_i uniformly spaced between $x = 0$ and the income at which CDF is equal to 1%, i.e., we fit the distribution for 99% of the population. The minimization procedure was implemented numerically in Matlab using the standard routines.

For the exponential distribution, the fitting parameter T determines the slope of $\ln C$ vs. x and has the dimensionality of Australian dollars per year, denoted as AUD or simply \$ (notice that $1\,\mathrm{k\$} = 10^3\,\$$). T is also equal to the average income $\langle x\rangle$ for the exponential distribution. The parameters m and β for the log-normal

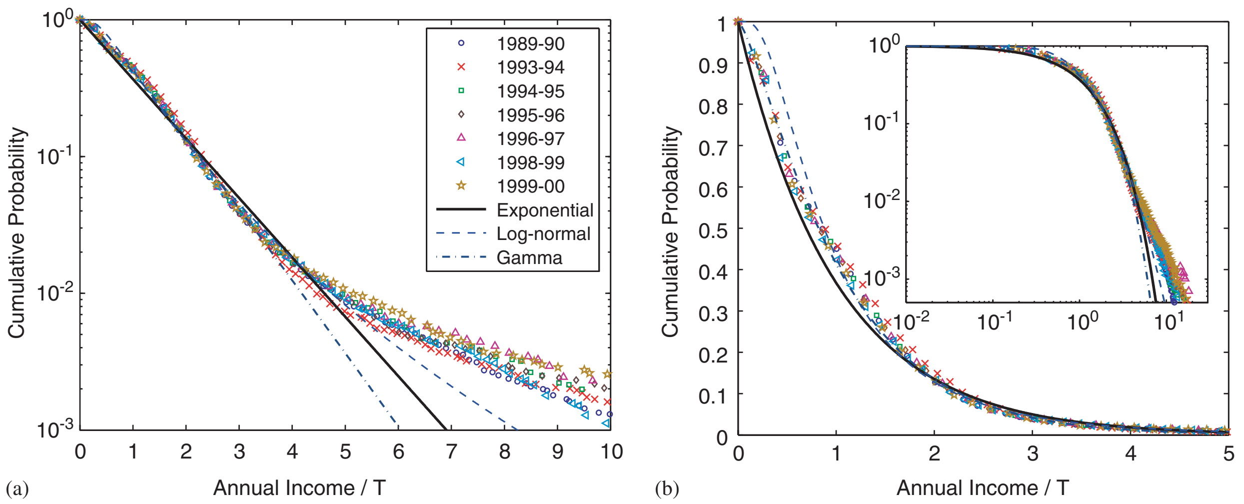

Fig. 1. The CDF of income, shown in the log–linear (a), linear–linear (b), and log–log (inset) scales. The income values for different years are normalized to the parameter T of the exponential distribution, given in Table 1. The lines show fits to different theoretical distributions in Eq. (2).

Table 1
Parameters of the distributions (1) and (2) obtained by minimization of the relative mean square deviation σ^2 between the empirical and theoretical CDFs

Year	T (k\$)	m (k\$)	s	β (k\$)	α	σ			Peak
						Exp (%)	L–N (%)	Gamma (%)	\$
1989–1990	17.8	15.1	0.74	13.4	0.39	13	11	6.8	6196
1993–1994	18.5	18.8	0.63	13.1	0.59	18	9.6	5.7	7020
1994–1995	19.6	17.7	0.71	14.9	0.40	15	9.4	5.5	7280
1995–1996	20.5	18.2	0.72	15.7	0.39	14	8.6	6.5	7280
1996–1997	21.2	18.9	0.72	16.5	0.37	14	8.4	7.7	7540
1998–1999	23.7	19.0	0.79	19.6	0.25	10	11	7.1	7800
1999–2000	24.2	19.6	0.78	19.3	0.30	11	11	7.2	7800

The last column gives position of the sharp peak in Fig. 2(b).

and gamma distributions also have the dimensionality of AUD, and the average incomes $\langle x \rangle$ for these two distributions are $me^{s^2/2}$ and $\beta\Gamma(\alpha+2,0)/\Gamma(\alpha+1,0)$. The parameters s and α are dimensionless and characterize the shape of the distributions. The values of these parameters, obtained by fits for each year, are given in Table 1. Using the values of T, we plot C vs. x/T in Fig. 1. In these coordinates, the CDFs for different years (shown by different symbols) collapse on a single curve for the lower 98% of the population. The collapse implies that the shape of income distribution is very stable in time, and only the scale parameter T changes in nominal dollars. The three lines in Fig. 1 show the plots of the theoretical CDFs given by Eq. (2). In these coordinates, the exponential CDF is simply a straight line with the slope -1. For the plots of the log-normal and gamma CDFs, we used the parameters $\overline{s} = 0.72$, $\overline{m/T} = 0.88$, $\overline{\alpha} = 0.38$, and $\overline{\beta/T} = 0.77$ obtained by averaging of the parameters in Table 1 over the years. We observe that all three theoretical functions give reasonably good, albeit not perfect, fits of the data with about the same quality, as confirmed by the values of σ in Table 1. Although the log-normal and gamma distributions have the extra parameters s and α, the fitting procedure selects their values in such a way that these distributions mimic the exponential shape. Actually, we constructed the gamma fit only for 98% of the population, because the fit for 99% gives $\alpha = 0$, i.e., the exponential. We conclude that the exponential distribution gives a reasonable fit of the empirical CDFs with only one fitting parameter, whereas the log-normal and gamma distributions with more fitting parameters do not improve the fit significantly and simply mimic the exponential shape.

However, by construction, $C(x)$ is always a monotonous function, so one may argue that different CDFs look visually similar and hard to distinguish. Thus, it is instructive to consider PDF as well, which we do in the next section.

3. Probability density function

In order to construct $P(x)$, we divide the income axis into bins of width Δx, calculate the sum of the weights w_n of the individuals with incomes from x to $x + \Delta x$, and plot the obtained histogram. However, there is subjectiveness in the choice of the width Δx of the bins. If the bins are too wide, the number of individuals in each bin is big, so the statistics is good, but fine details of the PDF are lost. If the bins are too narrow, the number of individuals in each bin is small, thus relative fluctuations are big, so the histogram of PDF becomes noisy. Effectively, $P(x)$ is a derivative of the empirical $C(x)$. However, numerical differentiation increases noise and magnifies minor irregularities of $C(x)$, which are not necessarily important when we are interested in the universal features of income distribution. To illustrate these problems, we show PDFs obtained with two different bin widths in Fig. 2.

Fig. 2(a) shows the coarse-grained histogram of $P(x)$ for all years with a wide bin width $\Delta x/T \approx 0.43$. The horizontal axis represents income x rescaled with the values of T from Table 1. The lines show the exponential, log-normal, and gamma fits with the same parameters as in Fig. 1. With this choice of the bin width, the empirical $P(x)$ is a monotonous function of x with the maximum at $x = 0$, and the exponential function gives a reasonable overall fit. The log-normal and gamma fits have maxima at $x/T \approx 0.56$ and $x/T \approx 0.29$. These values are close to the bin width, so we cannot resolve whether $P(x)$ has a maximum at $x = 0$ or at a non-zero x within the first bin.

Fig. 2(b) shows the PDF for the year 1994–1995 with a narrow bin width $\Delta x = 1$ k\$, which corresponds to $\Delta x/T \approx 0.05$. This PDF cannot be fitted by any of the three distributions, because it has a very sharp and narrow peak at the low income 7.3 k\$, which is way below the average income of 19.6 k\$ for this year. This peak is present for all years, and its position is reported in the last column of Table 1. The peak is so sharp and narrow that it cannot be attributed to the broad maxima of the log-normal or gamma PDFs. We speculate that this peak occurs at the threshold income of some sort of government policy, such as social welfare, minimal wage, or tax exemption. Comparing the empirical PDF with the exponential curve, shown by the

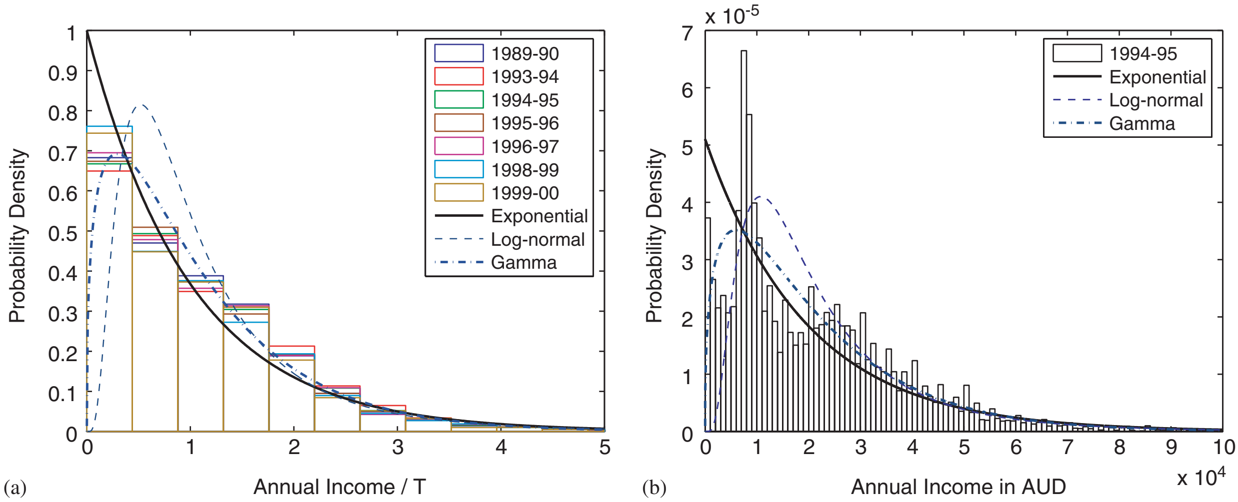

Fig. 2. The PDF of income distribution shown with coarse-grained (a) and high (b) resolutions. The lines show fits to different theoretical functions in Eq. (1).

solid line, we infer that the probability density above and below the peak is transferred to the peak, thus creating anomalously high population at the special income.

We also studied how often different income values occur in the data sets. The most frequently reported incomes for different years are always round numbers, such as 15 k$, 20 k$, 25 k$, etc. This effect can be seen in the periodically spaced spikes in Fig. 2(b). It reflects either the design of the survey questionnaires, or the habit of people for rounding their incomes in reporting. In addition to the round numbers, we also find the income corresponding to the peak position among the five most frequently reported incomes. This income, shown in the last column in Table 1, is not round and changes from year to year, but sometimes stays the same. This again suggests that the sharp peak in Fig. 2(b) is the result of a government-imposed policy and cannot be explained by statistical physics arguments.

By definition, $P(x)$ is the slope of $C(x)$ with the opposite sign. Fig. 1 clearly shows that the slope of $C(x)$ at $x = 0$ is not zero, so $P(x = 0) \neq 0$. Fig. 2(b) also shows that the probability density at zero income is not zero. In fact, $P(x = 0)$ is higher than $P(x)$ for all other x, except in the narrow peak. The non-vanishing $P(x = 0)$ is a strong evidence against the log-normal, gamma, and similar distributions, but is qualitatively consistent with the exponential function. However, there is also substantial population with zero and negative income, which is not described by any of these theories.

4. Discussion and conclusions

All three functions in Eq. (1) are the limiting cases of the generalized beta distribution of the second kind (GB2), which is also discussed in econometric literature on income distribution [16]. GB2 has four fitting parameters, and distributions with even more fitting parameters are considered in literature [16]. Generally, functions with more parameters are expected to fit the data better. However, we do not think that increasing the number of free parameters gives a better insight into the problem. We think that a useful description of the data is the one that has the minimal number of parameters, yet reasonably (but not necessarily perfectly) agrees with the data. From this point of view, the exponential function has the advantage of having only one parameter T over the log-normal, gamma, and other distributions with more parameters. Fig. 1(a) shows that $\log C$ vs. x is approximately a straight line for about 98% of population, although small systematic deviations do exist. The log-normal and gamma distributions do not improve the fit significantly, despite having more parameters, and actually mimic the exponential function. Thus we conclude that the exponential function is the best choice.

The analysis of PDF shows that the probability density at zero income is clearly not zero, which contradicts the log-normal and gamma distributions, but is consistent with the exponential one, although the value of $P(x = 0)$ is somewhat lower than expected. The coarse-grained $P(x)$ is monotonous and consistent with the exponential distribution. The high resolution PDF shows a very sharp and narrow peak at low incomes, which, we believe, results from redistribution of probability density near the income threshold of a government policy. Technically, none of the three functions in Eq. (1) can fit the complicated, three-peak PDF shown in Fig. 2. However, statistical physics approaches are intended to capture only the baseline of the distribution, not its fine features. Moreover, the deviation of the actual PDF from the theoretical exponential curve can be taken as a measure of the impact of government policies on income redistribution.

Acknowledgments

T. Di Matteo wishes to thank the Australian Social Science Data Archive, ANU, for providing the ABS data and the partial support by ARC Discovery Projects: DP03440044 (2003) and DP0558183 (2005), COST P10 “Physics of Risk” project and M.I.U.R.-F.I.S.R. Project “Ultra-high frequency dynamics of financial markets”.

References

[1] V. Pareto, Cours d’Économie Politique, Lausanne, 1897.
[2] R. Gibrat, Les Inégalités Économique, Sirely, Paris, 1931.
[3] M. Kalecki, Econometrica 13 (1945) 161.

[4] T. Di Matteo, T. Aste, S.T. Hyde, in: F. Mallamace, H.E. Stanley (Eds.), The Physics of Complex Systems, IOS Press, Amsterdam, 2004, p. 435.
[5] W. Souma, Fractals 9 (2001) 293.
[6] F. Clementi, M. Gallegati, Physica A 350 (2005) 427.
[7] A.A. Drăgulescu, V.M. Yakovenko, Eur. Phys. J. B 17 (2000) 723.
[8] A.A. Drăgulescu, V.M. Yakovenko, Eur. Phys. J. B 20 (2001) 585.
[9] A.A. Drăgulescu, V.M. Yakovenko, Physica A 299 (2001) 213.
[10] A.A. Drăgulescu, V.M. Yakovenko, in: P.L. Garrido, J. Marro, (Eds.), Modeling of Complex Systems: Seventh Granada Lectures, AIP Conference Proceedings, vol. 661, New York, 2003, p. 180.
[11] A.C. Silva, V.M. Yakovenko, Europhys. Lett. 69 (2005) 304.
[12] J. Mimkes, Th. Fruend, G. Willis, cond-mat/0204234, 2002;
G. Willis, J. Mimkes, cond-mat/0406694, 2004.
[13] N. Scafetta, S. Picozzi, B.J. West, Quant. Finance 4 (2004) 353.
[14] J.C. Ferrero, Physica A 341 (2004) 575.
[15] A. Chaterjee, S. Yarlagadda, B.K. Chakrabarti (Eds.), Econophysics of Wealth Distributions, Springer, Berlin, 2005.
[16] J.B. McDonald, A. Mantrala, J. Appl. Econ. 10 (1995) 201.