ELSEVIER

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*

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.
© 2006 Elsevier B.V. All rights reserved.

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 $x + dx$ is equal to $P(x)\,dx$. 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).

discussed above have the following functional forms:

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

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 $x = me^{-s^2}$ 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 $1.3-1.5 \times 10^7$ 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) $C(x) = \int_x^\infty P(x')\,dx'$. 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),

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

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

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 $1\,k\$ = 10^3\,\$$). T is also equal to the average income $\langle x \rangle$ for the exponential distribution. The parameters m and β for the log-normal