CHAPTER 2

A Function for Size Distribution of Incomes[†]

S. K. Singh and G. S. Maddala [‡]

Abstract

The paper derives a function that describes the size distribution of incomes. The two functions most often used are the Pareto and the lognormal. The Pareto function fits the data fairly well towards the higher levels but the fit is poor towards the low income levels. The lognormal fits the lower income levels better but its fit towards the upper end is far from satisfactory. There have been other distributions suggested by Champernowne, Rutherford, and others, but even these do not result in any considerable improvement. The present paper derives a distribution that is a generalization of the Pareto distribution and the Weibull distribution used in analyses of equipment failures. The distribution fits actual data remarkably well compared with the Pareto and the lognormal.

1 Introduction

The derivation of a function that describes the size distribution of incomes and various other distributions that show similar shapes is the purpose of this paper. The two functions most often used are the Pareto function and the lognormal. The Pareto function fits the data fairly well toward the higher levels but the fit is poor toward the lower income levels. If one considers the entire range of income, perhaps the fit may be better for the lognormal but the fit toward the upper end is far from satisfactory (Cramer, 1971).

Earlier, some efforts have been made by Champernowne (1953), Rutherford (1955), Mandelbrot (1960), Fisk (1961) to derive functional forms to describe the size distribution of incomes, based on reasoning about processes of income generation. The present paper derives a function based on the concept of hazard rate or failure rate which has been widely used for deriving distributions in reliability theory and for the analysis of the distribution of life times (see Barlow and Proschan (1965) and Lotka (1956)). The function derived here was also suggested by Burr (1942) though with a different purpose and reasoning. Also, there is a discussion of hazard rates in Gastwirth (1972), though again with a different purpose.

The plan of the paper is as follows: In Section 2 we present a derivation of the function through a discussion of failure rates. Section 3 presents an alternative derivation of the same function. Section 4 presents an empirical illustration, and the final section gives the conclusions.

2 The Genesis of the Function and Characterization through Failure Rate

If the life time of a person is distributed over the random variable x with probability density function $f(x)$, the probability of surviving at least up to time x is $R(x) = \int_x^\infty f(x)dx = 1 - F(x)$. The probability of death in a small interval of time dx is $f(x)$. After one has survived up to age x, the instantaneous death rate at age x, or the force of mortality, is $r(x) = f(x)/(1 - F(x))$. This ratio is variously known as the failure rate or the hazard rate and considerable work has been done to study the characterization of distribution functions from this point of view. Distributions are characterized as IFR (increasing failure rate) or DFR (decreasing failure rate distribution) depending upon whether $f(x)/(1 - F)$ rises or decreases with x. Generally speaking, one would not expect decreasing failure rate since time is most often the random variable and one does not expect a priori in most of the situations any particular kind of benefit to accrue with time to reduce the failure rate. Most of the distributions used - exponential, gamma function, normal - give IFR. Lognormal gives an increasing section of failure rate, followed by a decreasing section. This property, which appears questionable (see Barlow and Proschan (1965) and Jorgenson *et al.* (1967)) for other situations, is perhaps precisely the reason why it fits, to some extent, the income distribution.

When we change the random variable from time to income, a priori plausibility on theoretical reasoning for DFR after a point is obvious. While aging, as such, may not confer any advantage for living longer or the reduction of the hazard rate, income may help in earning more. The ability to make more money might increase with one's income. The various reasons are just a bit too obvious to be enumerated here. Therefore, it is appealing to consider distributions which are DFR at least after a point for income distribution. While Pareto is a DFR throughout the range, lognormal becomes a DFR only beyond a point.

For certain situations, it is perhaps more instructive to consider the hazard rate in terms of a transform of x rather than x itself. Consider the transform $z = \log x$. We may then try to find out the hazard rate with respect to this transform of $x : r * (z) = (dF/dz)/(1 - F)$.

The Pareto diagram in $(\log(1 - F), \log x)$ plane can be interpreted from this point of view. The first derivative of the Pareto transform is the hazard rate with respect to z.

A probability density function is defined to be IFR (increasing failure rate) if $(dr(x))/dx \geq 0$. It is called DFR (decreasing failure rate) if $(dr(x))/dx \leq 0$.

Similarly, a probability density function is defined to be IPFR (increasing proportionate failure rate) if $(dr * (z))/dz \geq 0$ and DPFR (decreasing proportionate failure rate) if $(dr * (z))/dz \leq 0$.

It can be easily checked that the Pareto distribution is monotone DFR for $r(x)$ though $r * (z)$ is constant. Lognormal has an $r(x)$ which has an IFR section followed by a DFR section. However, what is interesting is that $r * (z)$ is monotone increasing. This is the reason why the lognormal does not fit well at the high income level. As an empirical regularity, $r * (z)$ approaching constancy for high incomes appears to be fairly well accepted.

The intuitive economic meaning of $r * (z)$ is clear. At any income, it measures the odds against advancing further to higher incomes in a proportionate sense. It is a variable that should be allowed considerable flexibility because one would be interested in finding out its precise shape at varying levels of income. The restrictions imposed both by the shape of the Pareto function and the lognormal are rather severe. In the interpretation given above, lognormal asserts that it is easiest for one to improve one's relative position at low income groups, and the odds go on increasing monotonically tending to infinity as one's income increases. The Pareto distribution implies a constant value of the odds in the $r * (z)$ sense throughout at all income ranges.

A good starting point for deriving the distribution function is then the following: We accept the behaviour of $r * (z)$ toward the upper end of the income, i.e., asymptotic constancy on the basis of accumulated findings and received opinion. However, one must provide for lower $r * (z)$ at the lower income levels. This would mean allowing $r * (z)$ to rise with z and let it reach an asymptote. This can be done again variously. Should $r * (z)$ rise throughout with decreasing rate? Or should it rise first with z at an increasing rate, then a decreasing rate, and then asymptotically reach constancy? We will make the latter assumption.

For purposes of exposition, it is easier to take the negative of the Pareto transform, which is henceforth called $y : y = -\log(1 - F)$, $z = \log x$, and $y = f(z); y' > 0$, $y'' > 0$.

We advance the following assumption:

$$y'' = a \cdot y'(\alpha - y'), \tag{2.1}$$

a being constant. We solve this differential equation to get the distribution function. The composite assumption consists of three parts: (A-1) $r * (z)$ reaches

asymptotically a constant value a. (A-2) It first increases with an increasing rate, and then with a decreasing rate. (A-3) The rate of increase of $r*(z)$ is zero when the value of $r*(z)$ is zero. Rearranging (2.1) we get

$$\frac{y''}{y'} + \frac{y''}{\alpha - y} = a\alpha. \tag{2.2}$$

Integrating, we get

$$\log y' - \log(\alpha - y') = a\alpha z + c_1 \tag{2.3}$$

where c_1, is a constant of integration. This can be written as

$$\frac{y'}{\alpha - y'} = e^{a\alpha z + c_1}$$

or

$$y' = \frac{\alpha e^{a\alpha z + c_1}}{1 + e^{a\alpha z + c_1}}. \tag{2.4}$$

We note that y', which is the proportional failure rate, is the three-parameter logistic. Integrating (2.4) again we get

$$\log y = \frac{1}{a} \log(1 + e^{a\alpha z + c_1}) + c_2 \tag{2.5}$$

where c_2, is another constant of integration. After we substitute $-\log(1 - F)$ for y and $\log x$ for z in (2.5) we get, with some algebra,

$$\log(1 - F) = c - \frac{1}{a} \log(b + x^{a\alpha}), \tag{2.6}$$

where $c = (-c_2 - c_1)/\alpha$ and $b = 1/e^{c_1}$. Equation (2.6) gives the distribution function

$$F = 1 - \frac{c}{(b + x^{a\alpha})^{1/a}}. \tag{2.7}$$

The function in (2.7) has four constants. But since $F = 0$ for $x = 0$ we get $c = b^{1/a}$. Thus the three-parameter function is

$$F = 1 - \frac{b^{1/a}}{(b + x^{a\alpha})^{1/a}} \tag{2.8}$$

or

$$F = 1 - \frac{1}{(1 + a_1 x^{a_2})^{a_3}}, \tag{2.9}$$

where $a_1 = 1/b$, $a_2 = a\alpha$, and $a_3 = 1/a$. Note that $F = 0$ for $x = 0$ and, as $x \to \infty$, $F \to 1$.

In summary, F as in (2.9) is characterized by a PFR which is a logistic with respect to "income power", or z. Also, given that characterization, F as derived in (2.9) is unique. In upper income tail, the PFR is the same as for Pareto; at lower incomes it differs.

3 An Alternative Approach

An alternative derivation of the function derived in the previous section can be given in terms of models of decay. Let $F(x)$ be a certain mass at point x $(0 \leq x \leq \infty)$ which decays to zero as $x \to \infty$. dF/dx is the rate of decay. We standardize the initial mass to be one. If dF/dx depends only on the left-out mass $(1-F)$ then the process is said to be "memoryless". For the Poisson process, $dF/dx = a(1-F)$. The Pareto process can also be interpreted as memoryless since it implies

$$dF/dx = a(1-F)^{(1+1/a)}. \tag{2.10}$$

A process that introduces memory would be the so-called Weibull process which leads to the Weibull distribution. This implies

$$dF/dx = ax^b(1-F). \tag{2.11}$$

A generalization that combines elements of both (2.10) and (2.11) would be to start with the equation

$$dF/dx = ax^b(1-F)^c. \tag{2.12}$$

It can be readily verified that the solution to (2.12) gives equation (2.9) where (now, in terms of the parameters in (2.12))

$$a_1 = (c-1)(a/(b+1)), \quad a_2 = b+1, \quad \text{and} \quad a_3 = 1/(c-1).$$

The above derivation suggests the relationship between the Pareto, Weibull, and the distribution suggested here. One might wonder what the relationship is between this distribution and that suggested by Champernowne and Fisk. The distribution considered by Fisk (1961) is given by

$$\frac{dF}{d\phi} = \frac{e^\phi}{(1+e^\phi)^2} \quad \text{where } e^\phi = \left(\frac{x}{x_0}\right)^\alpha.$$

It can be easily verified that

$$\frac{dF}{dx} = \frac{a_1 a_2 x^{a_2-1}}{(1+a_1 x^{a_2})^2} \quad \text{where } a_1 = \left(\frac{1}{x_0}\right)^\alpha \quad \text{and } a_2 = \alpha.$$

Thus, putting $a_3 = 1$ we get the function suggested by Fisk.

4 Empirical Results

Salem and Mount (1974) used the method of maximum likelihood because, for the gamma density they considered, the estimating equations involve only the arithmetic and geometric means. For the present distribution, it is not possible to get any such simple expressions. The estimation of the Pareto distribution is customarily done by regressing $\log(1-F)$ on $\log x$. Fisk (1961) estimates the sech^2 distribution by regressing $\log F/(1-F)$ on $\log x$. For the distribution suggested here we have $\log(1-F) = -a_3 \log(1+a_1 x^{a_2})$.

Hence, following the customary procedures we estimated the parameters by using a nonlinear least squares method and minimizing

$$\sum [\log(1-F) + a_3 \log(1+a_1 x^{a_2})]^2.$$

The data used were from US Bureau of the Census (1960-1972) and the program was the nonlinear regression program from the Harvard computing center that uses the Davidon-Fletcher-Powell algorithm. The estimated parameters are shown in Table 2.1. The fit, as judged by the R^2's, was very good (they were all uniformly high around .99). But since this may not be an adequate evidence, we used some other checks with the results.

Salem and Mount (1974) have given the details of the observed and predicted probabilities for two years, 1960 and 1968, for the lognormal and the gamma. For comparison we plot the predicted probabilities from the present function in the same diagram that Salem and Mount (1974, Fig. 3) used. This is shown in Fig. 2.1. Also, the sum of squared deviations between the predicted and observed probabilities were as follows:

Year	Lognormal	Gamma	Present Function
1960	.01187	.00391	.00261
1969	.00752	.00238	.00156

Another check on the fit is to use the procedure suggested by Gastwirth and Smith (1972) which consists of computing bounds on the Lorenz concentration ratio and computing the implied value of this ratio from the estimated values of the parameters. For the years 1967 through 1970, we estimated these values by numerical integration using the estimated values of the parameters. The results are reported in Table 2.2. As can be easily seen, the estimates of the Lorenz ratio fall within the bounds.

Table 2.1:

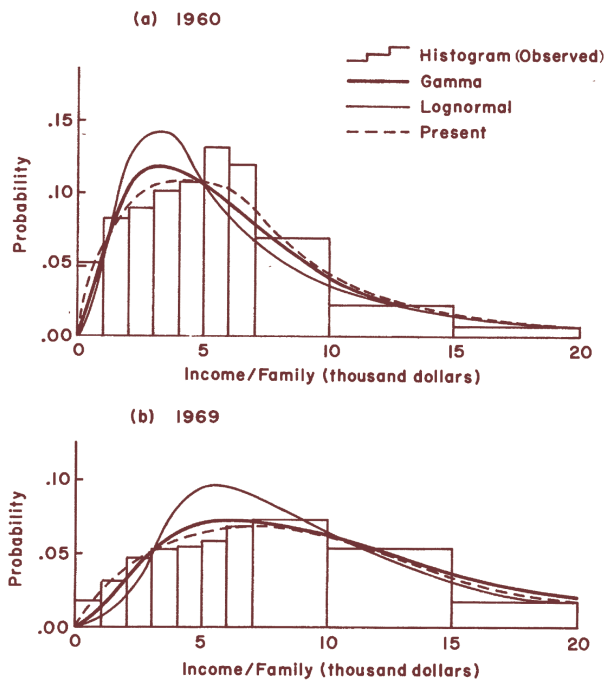| Year	a_1	a_2	a_3
1972	.3070	2.064	2.538
1971	.3125	2.139	2.544
1970	.3102	2.121	2.546
1969	.3101	2.131	2.611
1968	.3071	2.111	2.712
1967	.3120	2.012	2.552
1966	.3109	2.197	2.558
1965	.3082	2.127	2.624
1964	.3184	2.080	2.550
1963	.3084	2.051	2.597
1962	.3079	2.063	5.609
1961	.2735	1.972	3.009
1960	.2931	1.992	2.803

Fig. 2.1: Observed and predicted probabilities of United States families in ten income classes: 1960 and 1969.

Table 2.2:

Year	Bounds Reported by Salem and Mount		Estimates Obtained by the Fitted Function
	Lower	Higher	
1967	3,504	3,556	3,517
1968	3,391	3,457	3,402
1969	3,421	3,506	3,429
1970	3,466	3,565	3,484

5 Conclusions

The paper derives a function to describe the size distribution of incomes based on an analysis of hazard rates or failure rates. The distribution is a generalization of the Pareto and the Weibull distribution studied extensively in the analysis of equipment failures. The $sech^2$ distribution suggested by Fisk can also be considered as a special case of the distribution suggested here. The distribution has been fitted to United States income data and has been found to fit remarkably well. Earlier, Salem and Mount found that the gamma distribution gives a better fit than the lognormal. We find that the function suggested in the paper gives a better fit than the gamma.

References

Aitchison, J. and J. A. C. Brown (1957) *The Lognormal Distribution with Special Reference to its Uses in Enonomics*, Cambridge University Press, Cambridge.

Atkinson, A. B. (1970) On the Measurement of Inequality, *Journal of Economic Theory*, **2**, 244–263.

Barlow, R. E. and F. Proschan (1965) *Mathematical Theory of Reliability*, Wiley, New York.

Burr, I. W. (1942) Cumulative Frequency Functions, *Annals of Mathematical Statistics*, **13**, 215–235.

Champernowne, D. G. (1953) A Model of Income Distribution, *Economic Journal*, **63**(2), 318–351.

Cramer, J. S. (1971) *Empirical Econometrics*, North Holland, Amsterdam.

Fisk, P. R. (1961) The Graduation of Income Distributions, *Econometrica*, **29**(2), 171–185.

Gastwirth, J. L. (1972) The Estimation of the Lorenz Curve and Gini Index, *Review of Economics and Statistics*, **54**, 306–316.

Gastwirth, J. L. and J. T. Smith (1972) A New Goodness-of-Fit-Test, *ASA Proceedings of the Business and Economic Statistics Section*, pp. 320–322.

Jorgenson, D. J., J. J. McCall and R. Radner (1967) *Optimal Replacement Policy*, North Holland, Amsterdam.

Lotka, A. J. (1956) *Elements of Mathematical Biology*, Dover, New York.

Mandelbrot, B. (1960) The Pareto-Lévy Law and the Distribution of Income, *International Economic Review*, **1**, 79–106.

Rutherford, R. S. G. (1955) Income Distributions: A New Model, *Econometrica*, **23**(3), 277–294.

Salem, A. B. Z. and T. D. Mount (1974) A Convenient Descriptive Model of Income Distribution: the Gamma Density, *Econometrica*, **42**(6), 1115–1127.

US Bureau of the Census (1960-1972) *Current Population Reports*, Series P-60 and P-20, US Government Printing Office, Washington.