Estimating the Income Shares of a Grouped Frequency Distribution of Incomes(*)

Agostino Tarsitano
Università degli studi della Calabria
Dipartimento di Economia e Statistica
87030 Arcavacata di Rende (Cs)
agotar@unical.it

Summary
Several techniques of income distribution analysis assume that the income shares of the classes into which the income range has been partitioned are known. These shares are often omitted from the income data and must be interpolated. The purpose of this article is first, to review three procedures for piecewise interpolation of income shares; and second, to compare their performances with a set of observed distributions. The results suggest that a spline function of loglogistic densities provides plausible estimates of income shares also with a very limited amount of information.
.

keywords: income indequality, class marks, concentration indices

(*) Lavoro apparso in Statistica applicata, Vol. 21,n.3, pp. 307-319

1. Introduction

A common way to describe the size distribution of income is the frequency table

$$\{[a_{i-1},a_i),\ n_i;\ i=1,2,...,h\} \tag{1}$$

where $\{a_i, i=0,1,2,...,h;\ a_{i-1} < a_i\}$ are the boundaries, set in advance, of h exhaustive and mutually exclusive classes. The symbol ni indicates the number of units whose income falls in the i-th class.

Many techniques for analyzing a size distribution of incomes (particularly, the study of its Lorenz curve and the derivation of bounds for the Gini and other measures of inequality) require the availability of the income share attributable to each class. These are often omitted from the income data and must be interpolated. Haytovsky (1983) d i d not include such an important econometric problem in his fine summary on grouped data.

Traditionally, it is assumed that the midpoints are reasonable approximators of the class means (provided that the classes are closed and that the class widths are adequately small). However, because of the skewness of the income density, estimation of the class means (and thus of the income shares) will be biased to some extent. In addition, midpoints underestimate the true inequality level by assuming perfect equality within classes (the effect of this error on some inequality measures has been studied by Seiver (1979)).

To improve the accuracy of calculation for the income shares I consider the method of piecewise interpolation i.e. the approximation of the true model by a sequence of submodels each joined to its neighbors at the boundaries delimiting that region of the income variate corresponding to a single submodel.

The underlying idea is that the distribution of income is a heterogeneous phenomenon behaving differently at different income levels; thus it cannot, in general, be described by a single density function. In this sense Mandelbrot (1960) remarked "... it is unlikely that a single empirical formula could ever represent all the data"; Budd (1 970) wrote "*We know that it is virtually impossible to describe empirical distributions accurately by just one function"*. Yet the research on income distribution deals almost exclusively with the specification of single models to fit the whole range of income (see, for example, Dagum (1984)).

The primary contribution of the present paper is to propose a new method for estimating the income shares of a grouped frequency distribution of incomes. The method, based on a spline of density functions, possesses several advantages over the usual methods:

1) Provides a framework for modeling heterogeneity across the income classes.

2) Does not assume perfect equality within classes.

3) Expresses both the midpoints and the fitting of a single distribution in a unified manner.

The plan of the paper is as follows: in section 2, I present a new formulation of the income model. In section 3, I discuss the selection of submodels and in section 4 three families of submodels (LogLaplace, Loglogistic, and Lognormal) are introduced. Finally, in the fifth section, the results of the various procedures are compared to the income quantiles contained in six published sources

A new formulation of the income model.

The fitting of a particular density function to the frequency table (1) has rarely produced acceptable approximations over the entire income domain. In this section I develop a more flexible approach centered upon the use of possibly different density functions for different income levels.

First, classes $\{[a_{i-1},a_i),\ n_i;\ i=1,2,...,h\}$ are assembled into the intervals $\{[A_{i-1},A_i),i=1,2,...,k\}$ with $A_0=a_0$, $A_k=a_h$, and $k\leq h$. The i-th interval is the union of k_i adjacent classes so that $A_i=a_{r_i}$ with

$$r_i = \sum_{j=1}^{i} k_j$$

Figure 1 and table 1 illustrate the idea.

Table 1. Class and interval boundaries

i	k_i	r_i	Interval Limits	Class Limits	Composition
1	3	3	A_1	a_3	$[A_0,A_1)=[a_0,a_1)\cup[a_1,a_2)\cup[a_2,a_3)$
2	2	5	A_2	a_5	$[A_1,A_2)=[a_3,a_4)\cup[a_4,a_5)$
3	1	6	A_3	a_6	$[A_2,A_3)=[a_5,a_6)$
4	2	8	A_4	a_8	$[A_3,A_4)=[a_6,a_7)\cup[a_7,a_8)$

If $k=1$, i.e. if $k_1=h$ then we deal with the observed income range $[a_0,a_h)$ as a whole and if $k=h$ then each class is separately treated. These are the cases most frequently encountered, but other arrangements could be found. The "optimal" aggregation of classes in interval is an important aspect of piecewise interpolation, but it will not be discussed in the present paper (see Bellman (1969) for a dynamic programming approach to this problem).

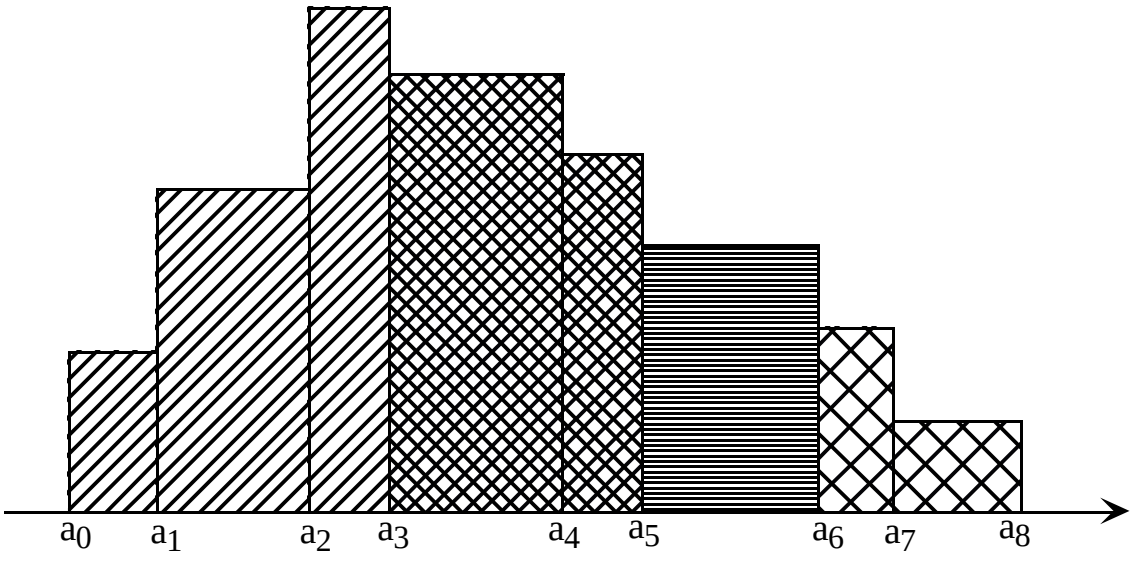

Figure 1- A partition of the income range

Next, I suppose that the behavior of the income variate *y* over the i-th interval, can be described by the density $f_i(y)$ about which I assume

$$f_i(A_{i-1}) = f_{i-1}(A_{i-1}) \tag{2}$$

$$\int_{A_{i-1}}^{A_i} f_i(y)dy = \pi_i \tag{3}$$

where

$$0 \leq \pi_i = \frac{\sum_{j=r_{i-1}+1}^{r_i} n_j}{\sum_{j=1}^{h} n_j} \leq 1, \ with \ r_0 = 0 \tag{4}$$

is the fraction of units whose income falls in $[A_{i-1},A_i)$. It follows that the density of *y* over the entire income range $[A_0,A_k)$ is

$$f(y) = \sum_{i=1}^{k} \frac{\alpha_i(y)}{f_i(y)} \tag{5}$$

where

$$\alpha_i(y) = \begin{cases} 1 & if \ \ A_{i-1} \leq y \leq A_i \\ 0 & Otherwise \end{cases} \tag{6}$$

Conditions (2) and (3) ensure that (5) is a continuous density function (5) is a finite mixture of the f_i's formed by truncating $f_i(y)$ outside $[A_{i-}1,A_i)$ and splicing together the various components. From another point of view $f(y)$ can be considered a spline function (see Wold ,1974) whose components are density functions instead of the usual polynomials.

The cumulative distribution function and the incomplete first moment distribution function corresponding to (5) are

$$F(y) = \sum_{i=1}^{s-1} \pi_i + \int_{A_{s-1}}^{y} f_s(t)dt \tag{7}$$

$$F_1(y) = \sum_{i=1}^{s-1} \theta_i + \int_{A_{s-1}}^{y} \left(\frac{t}{\mu}\right) f_s(t)dt \tag{8}$$

where *s* is an index such that $A_s=min\{A_i| \ A_i \geq y\}$ and with

$$\theta_i = \int_{A_{i-1}}^{A_i} \left(\frac{t}{\mu}\right) f_i(t)dt = \left(\frac{\mu_i}{\mu}\right)\left[F_{1i}(A_i) - F_{1i}(A_{i-1})\right] \tag{9}$$

In (9) μ is the total mean income whereas μ_i and $F_{1i}(y)$ are, respectively, the complete and incomplete first moment of the i-th submodel. Such a formulation allows us to take heterogeneity into account across the income intervals. In fact, we now have the possibility to select the f_i's either from a family indexed by a vector of parameters or from different families (LogLaplace, Lognormal, etc.).

Furthermore, several techniques of piecewise interpolation can be considered as particular cases of (5). For instance, Leibenberg and Kaitz (1951) applied a parabolic density for the first classes, straight line density for middle classes and the Pareto curve in the last classes; Aigner (1968) assumed $k=h$ and that fi(y) was a parabola through the i-th interval; Budd (1970) used $k=2$, $_1(y)$ polynomial and $f_2(y)$ exponential; Campa and Visco (1973) studied a combination of a Lognormal with a Pareto; Petersen (1979) discussed $k=4$, and $f_i(y)=a_iy + b_i$; Krieger (1983) fitted uniform densities separately in each class. Mehran (1975a) and Spiers (1978) applied a Pareto density in all the classes.
Formula (5) is not a succinct description of a size distribution of income in Bartels and Vries (1977) terms. Potentially, it has a greater number of parameters than any other specification currently used. However, the number of parameters in itself is not an obstacle to interpolation and simulation.

3. Choice of a submodel and estimation of its parameters

The choice of a submodel must take several factors into account. Ideally, it would be desirable to have a definite, hopefully simple expression of the general income model (6). Therefore, the f_i's should be able to indicate whether representing the size distribution of income by a single functional form is appropriate or not. This could be achieved by imposing

$$f_i(y) = f(y,\beta_i) \qquad i=1,2,\ldots,k \tag{10}$$

which implies that all the k submodels belong to the same system of frequency curves. If the parameters turn out to be stable across the intervals, then a single parametric curve $f(y,b)$ should be fitted over the whole range. If this is the case, the associated density should have the usual shape of a size distribution of incomes: unimodal, highly leptokurtic, and with a positive skewness.

In addition, the cumulative distribution function $F(y,\beta_i)$ must have high contact with its upper asymptote. Specifically, it is required that

$$\lim_{y\to\infty} y^m\left[1-F(y,\beta_i)\right] = constant \qquad i=1,2,\ldots,k \tag{11}$$

for some integer $m>0$.

Condition (11) has two important implications (Kendall and Stuart (1977), p. 333). First, the distribution $F(y,\beta_i)$ has no moments of order m and, second, the proportional failure rate

$$PFR(y) = -\frac{dLn\left\{\left[1-F(y,\beta_i)\right]\right\}}{dLn(y)} = \frac{yf(y,\beta_i)}{1-F(y,\beta_i)} \qquad i=1,2,\ldots,k \tag{12}$$

tends to the finite limit *m* as income becomes larger. In order to compute income shares we need mean income; thus we must choose models for which *m > 1*. According to (11) models like the Gamma and the Lognormal should not be applied to the last class. Once the f_i's are specified, the only elements that need to be estimated are the parameters β_i. Given the interval $[A_{i-1},A_i)$, conditions (2) and (3) can be restated in the following form

$$F\left(a_{r_{i-1}+j},\beta_i\right)=p_j \quad j=1,2,\ldots,k; \quad i=,2,\ldots,k \tag{13}$$

where $p_i=\sum_{j=1}^{i}\pi_j$ is the cumulative fraction of units whose income is at most A_i.

System (13) contains k_i equations. Therefore, β_i must contain no more than ki independent parameters. Note, incidentally, that piecewise interpolation allows a maximum of *2*h* parameters. If the number of equations is greater than the number of parameters then several techniques of estimation could be considered (McDonald,1979). The case *k=h* and known group means has been treated by Cowell and Mehta (1982). The present paper focuses on the same case but assumes that the available information for a class is just the number of observations falling into it. As a result, our choice for the f_i's must be limited to a two-parameter density function.

By solving system (13) we obtain the quantile estimates for the β_i's, and the method of quantiles will be our method of estimation although, for *k=h*, there is no possibility for optimum spacing. Properties of quantile estimators are discussed in Bury (1975).

4. The submodels

In this section I examine three types of submodels and the quantile estimates of their parameters. The submodels are based on the method of Edgeworth-D'Addario (see Chipman, 1985).

Let *u* denote a measure of aptitude with density function *g(u)* and let *y* be the income derived by a continuous monotonic translation of aptitudes *y=T(u)*. The density of the income variate is given by

$$dF(y)=g\left[r(y)\right]r'(y)dy \tag{14}$$

where $r(y)=T^{-1}(y)$. For this article I have chosen the exponential translation

$$T(u)=\beta_1 e^{\frac{u}{\beta_2}} \tag{15}$$

and a standard (i.e. which does not include unknown parameters symmetric density for aptitude. Specifically, I selected the Laplace, the Normal, and the Logistic model. Other alternatives can be derived from the generalized Gaussian model (see Vianelli, 1982).

$$g(u)=\left[2^{\gamma+1}\Gamma(\gamma+1)\right]e^{-\frac{|u|^{\frac{1}{\gamma}}}{2}} \tag{16}$$

The LogLaplace
The i-th cumulative distribution function is

$$F(y,\beta_i)=\begin{cases}\frac{1}{2}\left(\frac{y}{\beta_{1i}}\right)^{\beta_{2i}} & if \quad y<\mu_e\\ 1-\frac{1}{2}\left(\frac{y}{\beta_{1i}}\right)^{\beta_{2i}} & if \quad y\geq\mu_e\end{cases} \tag{17}$$

with $\beta_{1i}>0$ and $\beta_{2i}>1$ and where μ_e is the median income.

The PFR of (17) increases for incomes lower than β_{1i} and becomes constant and equal to β_{2i} for incomes greater than β_{1i}. Mehran (1975a) observed that the first branch of (17) gives good approximations for the poor classes whereas the second branch performs well for the rich classes. However, the LogLaplace model should not be used if the histogram of the observed data has one or more submode(s).The incomplete first moment distribution is

$$F_1(y,\beta_i)=\begin{cases}\left(\frac{y}{\beta_{1i}}\right)^{\beta_{2i}+1} & if \quad y<\mu_e\\ 1-\left(\frac{y}{\beta_{1i}}\right)^{\beta_{2i}+1} & if \quad y\geq\mu_e\end{cases} \tag{18}$$

After inserting (18) in system (13) we obtain the estimates for the parameters of the first branch

$$\begin{cases}\bar{\beta}_{1i}=e^{\left(\frac{t_i}{1+t_i}\right)A_{i-1}^{+}-\left(\frac{1}{1+t_i}\right)A_i^{+}}\\ \bar{\beta}_{2i}=\dfrac{Ln\left(\frac{p_i}{p_{i-1}}\right)}{A_i^{+}-A_{i-1}^{+}}\end{cases} \qquad if \quad y\leq\mu_e \tag{19}$$

where $A_i^{+}=Ln(A_i)$ and $t_i=\text{Log}(2p_i)/\text{Log}(2p_{i-1})$,and for the second branch

$$\begin{cases}\bar{\beta}_{1i}=e^{\left(\frac{1}{1-t_i^{*}}\right)A_i^{+}-\left(\frac{t_i^{*}}{1-t_i^{*}}\right)A_{i-1}^{+}}\\ \bar{\beta}_{2i}=\dfrac{Ln\left(\frac{1-p_{i-1}}{1-p_i}\right)}{A_i^{+}-A_{i-1}^{+}}\end{cases} \qquad if \quad y>\mu_e \tag{20}$$

where $t_i^* = Ln[2(1-p_{i-1})]/Ln[2(1-p_i)]$. Usually, the median class $[A_{m-1},A_m)$, has to be split into two subclasses: $[A_{m-1},\mu_e)$ and $[\mu_e,A_m)$. I applied (19) to the former and (20) to the latter. In both cases $ß_{1m}=\mu_e$.

Since $\beta_{2i}>1$, model (18) cannot be used unless the following conditions

$$\begin{cases} A_{i-1}p_i > A_i p_{i-1} & if \quad y \le \mu_e \\ A_{i-1}(1-p_{i-1}) > A_i(1-p_i) & if \quad y > \mu_e \end{cases} \tag{21}$$

are satisfied. The selection of the appropriate branch for a given class and the estimation of the income shares requires the knowledge of μe and μ. Usually their values are given as auxiliary information of table (1).

The Loglogistic

The i-th cumulative distribution function is

$$F(y,\beta_i) = 1 - \left(1 + \frac{y^{\beta_{2i}}}{\beta_{1i}}\right)^{-1} \tag{22}$$

with $\beta_{1i}>0$ and $\beta_{2i}>1$. It is universally claimed that the Loglogistic model provides a close approximation of the income distribution for all the intervals. The PFR of (22) is an increasing function of income and tends to the limit β_{2i} as income goes to infinity. The incomplete first moment distribution is

$$F_i(y,\beta_i) = \frac{e^{\frac{Ln(\beta_{1i})}{\beta_{2i}}} B\left(1+\frac{1}{\beta_{2i}},1-\frac{1}{\beta_{2i}}\right)}{\mu}\left[IB\left(F_i;1+\frac{1}{\beta_{2i}},1-\frac{1}{\beta_{2i}}\right) - IB\left(F_{i-1};1+\frac{1}{\beta_{2i}},1-\frac{1}{\beta_{2i}}\right)\right] \tag{23}$$

with $F_i=F(y;\beta_i)$. In (23) $B(a,b)$ and $IB(y;a,b)$ are, respectively, the Beta function and the Beta function ratio (which can be computed by using the algorithm designed by Majumbder and Bhattacharjee,1973).

The quantile estimates of the parameters are

$$\begin{cases} \bar{\beta}_{1i} = e^{\left(\frac{1}{d_i-1}\right)Ln(S_i) - \left(\frac{d_i}{d_i-1}\right)Ln(S_{i-1})} \\ \bar{\beta}_{2i} = \dfrac{Ln\left(\frac{S_i}{S_{i-1}}\right)}{A_i^+ - A_{i-1}^+} \end{cases} \tag{24}$$

with $S_i=p_i/(1-p_i)$ and $d_i = A_i^+ / A_{i-1}^+$. An immediate check for the applicability of the Loglogistic is the relation: $S_iA_{i-1} > S_{i-1}A_i$. If this is not true, model (22) cannot be used.

It is worth noting that to compute (24) the knowledge of median income is superfluous; also, if an estimate of $F_1(y,\beta_1)$ is required, the additional problem of A_0 unknown can be skipped.

The Lognormal

The cumulative distribution function over the i-th class is

$$F(y,\beta_i)=\Phi\left(\frac{Ln(y)-\beta_{1i}}{\beta_{2i}}\right) \tag{25}$$

where $\Phi(.)$ is the standard normal integral and $\beta_{1i},\beta_{2i}>0$.

The Lognormal model has a *PFR* that does not tend to a finite limit as income increases; thus the Lognormal violates condition (11). However, according to well established empirical findings, model (25) should provide a good fit in central intervals.

The incomplete first moment distribution is

$$F_i(y,\beta_i)=e^{\beta_{1i}+\frac{\beta_{2i}^2}{2}-Ln(\mu)}\left\{F\left[\frac{\left(Ln(A_i)-\left(\beta_{1i}+\beta_{2i}^2\right)\right)}{\beta_{2i}}\right]-F\left[\frac{\left(Ln(A_{i-1})-\left(\beta_{1i}+\beta_{2i}^2\right)\right)}{\beta_{2i}}\right]\right\} \tag{26}$$

The quantile estimates for the β_i's can be obtained first by computing Z_i and Z_{i-1} where $Z_i=\Phi^{-1}(p_i)$ (see Beasley and Springer,1977) and then by solving the linear system

$$\begin{cases} A_i^+ = \bar{\beta}_{i1}+\bar{\beta}_{i2}Z_i \\ A_{i-1}^+ = \bar{\beta}_{i1}+\bar{\beta}_{i2}Z_{i-1} \end{cases} \tag{27}$$

Since $\bar{\beta}_{i1}$ and $\bar{\beta}_{i2}$ must be positive, a Lognormal submodel cannot be applied to the i-th interval if $A_{i-1}^+Z_i \le A_i^+Z_{i-1}$.

5. Testing the accuracy of the procedures.

To evaluate the accuracy of the procedures of piecewise interpolation discussed in the previous section I compared their results to the quantiles of income from six published distributions: Italy 82, 83, 84 (ISTAT, 1985) and USA 82, 83, 84 (US Bureau of the Census). Data refers to total money income of families and include 14 classes for t he former and 21 classes for the other. For all the distributions, both the first and the last classes, were open-ended (i.e. classes for which no information other than the number of incomes was available). Three major aspects have been considered: the behavior of parameter estimates, the goodness-of-fit and the estimation of the Gini measure.

Table II lists the mean, the coefficient of variation and the range of parameter estimates. Since system (13) is underidentified for the first interval, I avoided interpolating $F_1(A_1)$ and used its true value to compute the other quantiles. This is clearly not a practi-

cable method and ad hoc devices must be arranged to face real situations (e.g. those proposed by Merhan, 1975a and Needleman, 1978). Of course it is not necessary to estimate $F_1(A_k)$ since $F_1(A_k)=1-F_1(A_{k-1})$.

The relative dispersion of the parameter estimates is, broadly speaking, high for β_1 in the Loglogistic (this is probably due to the parametrization adopted in (22)); and low for the Lognormal (this is not surprising if one considers that in (25) there are two shape parameters).

As a rule, the scale parameters have shown higher variability than shape parameters. From the third and sixth column, it is possible to note that the range of estimates is always too high to justify the use of a single functional form. The only exception may perhaps be Italy 82 where a single Lognormal appears recommendable over the entire domain.

Table II: mean, standard deviation, and range of parameter estimates

			$\bar{\beta}$			$\bar{\beta}_2$		
			Mean	C. of V.	Range	Mean	C. of V.	Range
Italy	84	LogLaplace	13.5581	0.1675	8.3795	2.3432	0.3329	2.3448
		Loglogistic	29659.4000	2.3954	271161.0000	3.2346	0.1863	2.1223
		Lognormal	2.6608	0.0316	0.3256	0.5662	0.1648	0.3963
	83	LogLaplace	12.9621	0.1420	7.0682	2.5023	0.3375	2.6808
		LogLogistic	18088.6000	2.2772	158393.0000	3.2559	0.1876	2.1858
		Lognormal	2.4944	0.0316	0.3256	0.5724	0.1544	0.3936
	82	LogLaplace	12.0399	0.1469	6.0882	2.5899	0.3765	3.0180
		Loglogistic	34381.7000	1.7544	178152.0000	3.2643	0.2473	2.4441
		Lognormal	2.3845	0.0699	0.6603	0.5989	0.2295	0.5054
USA	84	LogLaplace	27.3548	0.1161	13.6596	1.6451	0.3477	2.0372
		Loglogistic	12088.1000	2.0764	101179.0000	2.2583	0.2551	1.9836
		Lognormal	3.3930	0.0721	1.0919	0.8396	0.3305	1.0904
	83	LogLaplace	25.7421	0.1210	12.7686	1.7223	0.3804	2.2728
		Loglogistic	14820.1000	1.9368	104962.0000	2.3438	0.2668	2.0283
		Lognormal	3.3061	0.0649	0.9599	0.8167	0.3325	1.0411
	82	LogLaplace	25.4201	0.1596	20.1466	1.7243	0.3140	1.9227
		Loglogistic	5358.6100	1.2896	23073.5000	2.3268	0.2266	1.9135
		Lognormal	3.2913	0.1146	1.8807	0.8236	0.3811	1.3958

The goodness-of-fit has been measured by

$$Mean\ \ Absolute\ \ Error(MAE):\ \frac{\sum_{i=1}^{h}\left|q_i^* - q_i\right|}{h} \tag{28}$$

$$Largest\ \ Absolute\ \ Error(LAE):\ \max_{1\le i\le h}\left\{\left|q_i^* - q_i\right|\right\} \tag{29}$$

where q_i is an estimated income share and q_i^* is the actual income share. Both indices lie in the range from zero to one. Table III below displays the values of (28) and (29) for the 18 situations we have explored

Table III: values of the goodness-of-fit indices

Country		Procedures					
		LogLaplace		Loglogistic		Lognormal	
	Year	MAE	LAE	MAE	LAE	MAE	LAE
Italy							
	84	0.00114	0.00211	0.00084	0.00202	0.00072	0.00186
	83	0.00170	0.00531	0.00218	0.00601	0.00245	0.00634
	82	0.00274	0.00502	0.00340	0.00575	0.00361	0.00594
USA	84	0.00076	0.00199	0.00067	0.00217	0.00061	0.00245
	83	0.00083	0.00161	0.00074	0.00167	0.00066	0.00180
	82	0.00063	0.00147	0.00070	0.00151	0.00079	0.00157

Index (28) suggests the LogLaplace for Italy 82, Italy 83, USA 82 and the Lognormal for the others. Index (29) confirms the Lognormal for Italy 84 and indicates the LogLaplace for the others. However, on the basis of the two indices, none of the procedures is convincingly better than the others.

Gastwirth (1972) derived upper and lower bounds for the Gini index with grouped data (see also Giorgi and Pallini, 1987); the same bounds were suggested by Gastwirth and Smith (1972) for testing the fit of a distribution to grouped data.

Mehran (1975b), using a geometric approach, obtained bounds slightly larger, but requiring scarcer data then Gastwirth's. The bounds proposed by Mehran are

$$\begin{cases} GL = \sum_{i=1}^{h} \left(p_{i-1} q_i - p_i q_{i-1} \right); \quad D = \sum_{i=1}^{h} \dfrac{\left(p_i - q_{i-1} \right)^2 \left(\alpha_i^* - \alpha_i \right) \left(\alpha_{i-1} - \alpha_{i-1}^* \right)}{\left(\alpha_i^* - \alpha_{i-1}^* \right)} \\ GU = GL + D \end{cases} \tag{30}$$

The symbol α_i denotes the slope of the line R_i joining (p_{i-1},q_{i-1}) and (p_i,q_i); α_i^* is given by

$$\alpha_i^* = \begin{cases} \alpha_i & if \quad b_i^* < \alpha_i \\ b_i^* & if \quad \alpha_i \le b_i^* < \alpha_{i+1} \\ \alpha_{i+1} & if \quad b_i^* > \alpha_{i+1} \end{cases} \tag{31}$$

b_i^* being the slope of the line joining $\left(p_{i-1}^*, q_{i-1}^* \right)$ and $\left(p_i^*, q_i^* \right)$. The point $\left(p_i^*, q_i^* \right)$ is the intersection of R_i and R_{i+2}.

In table IV are reported the values of the lower bound GL and of the grouping factor D both for the observed income shares and for those interpolated using the three procedures.

Table IV: bounds on the Gini index.

Country	Year	Observed		LogLaplace		Loglogistic		Lognormal	
		GL	GD	GL	GD	GL	GD	GL	GD
Italy	84	0.31840	0.00750	0.31570	0.00740	0.31650	0.00743	0.31700	0.00746
	83	0.31000	0.00522	0.31110	0.00576	0.31230	0.00579	0.31300	0.00582
	82	0.32210	0.00511	0.32640	0.00529	0.32770	0.00532	0.32820	0.00533
USA									
	84	0.38110	0.00346	0.37950	0.00337	0.37970	0.00336	0.37990	0.00334
	83	0.37950	0.00256	0.37780	0.00253	0.37810	0.00252	0.37820	0.00251
	82	0.36010	0.00174	0.36130	0.00174	0.36110	0.00172	0.36130	0.00171

As it can be seen from the table, the estimates of the lower bound are always adequate and the bias associated with all of the procedures has the same sign.

The grouping factor is also approximated fairly well and the results are very close for all of the distributions considered.

Such findings corroborate the impression that each of the three techniques produces reasonable surrogates of the unknown income quantiles.

6. Conclusions.

This paper attempted to identify a satisfactory method for estimating the income quantiles when the class means are unknown It has been shown that piecewise interpolation is a viable and flexible technique to solve such a problem. None of the three special cases considered systematically dominates the others. Persuasive approximations can be obtained using a spline function of Loglogistic densities also in the case that, in addition to the frequency count, only the total mean income is reported (the LogLaplace would require additional information on the median income). It is important to note that when the first class is open-ended the only applicable model is the Loglogistic.

References

Aigner D.J., 1968, A Linear Approximation for the Class Marks of a Grouped Frequency Distribution, With Especial Reference to the Unequal Interval Case. *Technometrics*, 10, 793-809

Bartels C.P., Vries O.M., 1977, Succinct Analytical Descriptions of Income Distributions Using Transformation Functions. *Èconomie Appliquée*, 30, 369-390.

Beasley J.B., Springer S.G., 1977, Algorithm AS111: the Percentages Points of the Normal Distribution. *Applied Statistics*, 26, 118-121

Bellman R., 1969, Fitting by Segmented Straight Lines., *Journal of the American Statistical Association*, 64, 1079-1084

Budd E.C., 1970, Postwar Changes in the Size Distribution of Income. *American Economic Review*, 40, 247-260

Bury V.K., 1975, Statistical Models in Applied Science, John Wiley & Sons, New York

Campa G., Visco V., 1973 La distribuzione dei redditi, Franco Angeli, Milano.

Chipman J.S., 1985, The Theory and Measurement of Income Distribution. *Advances in Econometrics*, 4, 135-165

Cowell F.A., Mehta F.,1982, *The Estimation and Interpolation of Inequality Measures.* Review of Economic Studies, 49, 273-290

Dagum C., 1984, Income Distribution Models, The Encyclopedia of Statistical Science, 4, 27-34.

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

Gastwirth J.L., Smith J.T.,1972, A New Goodness-of-fit Test,. Proceedings of the American Statistical Association, 320-322

Giorgi G.M., Pallini A., 1987, About a General Method for the Lower and Upper Distribution-free Bounds on Gini's Concentration Ratio from Grouped Data., *Statistica*, XLVII, 171-184

Haitovsky Y., 1983, Grouped Data, The Encyclopedia of Statistical Science, 2, 527-536.

ISTAT, 1985, La distribuzione quantitativa del reddito in Italia nelle indagini sui bilanci di famiglia. Anno 1984. Supplemento al Bollettino Mensile di Statistica, N.7

Kendall M., Stuart A., 1977, The Advanced Theory of Statistics, Vol.1,4th ed. Charles Griffin & Co.,London.

Krieger A.M., 1983, Bounding moments from Grouped Data and the Importance of Group Means. *Sankhya, Series B*, 45,309-319

Leibenberg M., Kaitz H., 1951, An Income Size Distribution from Income Tax and Survey Data 1944. *Studies in Income and Wealth*, 13, 441-444

Majumder K.L., Bhattacharjee G.P.,1973, Algorithm AS63: the Incomplete Beta Integral. *Applied Statistics*, 22, 409-411

Mandelbrot B., 1960, The Pareto-Lévy Law and the Distribution of Income. In*ternational Economic Review*, 1, 79-106.

McDonald J.B., and Ransom M.R., 1979, Alternative Parameter Estimators Based upon Grouped Data, Communications in Statistics, part A, Theory and Methods, 8, 899-917.

Mehran F. 1975a, Dealing with Grouped Income Distribution Data, Working Paper n. 20, International Labour Office, Geneva.

Mehran F. 1975b, Bounds on Gini Index Based on Observed Points of the Lorenz Curve. *Journal of the American Statistical Association*, 70, 64-66

Needleman L., 1978, On the Approximation of the Gini Coefficient of Concentration. *The Manchester School*, 46, 105-122

Petersen H.G., 1979, Effects of Growing Incomes on Classified Income Distributions, the Derived Lorenz Curves and Gini Indices, *Econometrica*. 47, 183-198

Seiver D.A., 1979, A Note on the Measurement of Income Inequality with Interval Data., *The Review of Income and Wealth*, 25, 229-233

Spiers F., 1978, Estimation of Summary Measures of Income Size Distribution from Grouped data, Proceedings of the American Statistical Association, Washington US Bureau of the Census, Current population reports, consumer income, Series P-60.

Vianelli S., 1982, Sulle lognormali di ordine r quali famiglie di distribuzioni di errori di proporzione. *Statistica*, 42, 155-176

Wold S. 1974, Spline Functions in Data Analysis. *Technometrics*, 16, 1-11.