1

Introduction

1.1 Historical Background

It is a well-known fact that the income distributions are commonly unimodal and skew with a heavy right tail. Therefore, different skew models such as the lognormal and the Pareto have been proposed as suitable descriptions of the income distribution, but they are usually applied in specific empirical situations.

For general studies, more wide-ranging tools have been considered. The most commonly used theory is based on the Lorenz curve. Lorenz (1905) developed it in order to analyse the distribution of income and wealth within populations. He described the Lorenz curve in the following way:

> *"Plot along one axis accumulated per cents of the population from poorest to richest, and along the other, wealth held by these per cents of the population".*

Consequently, the Lorenz curve $L(p)$ is defined as a function of the proportion p of the population. It is convex and satisfies the condition $L(p) \le p$ because the income share of the poor is less than their proportion of the population. A sketch of a Lorenz curve is given in Figure 1.1.1.

The theoretical Lorenz curve $L_X(p)$ for the income distribution $F_X(x)$ of a non-negative variable X can be described in the following way: Let $f_X(x)$ be the corresponding frequency distribution,

$$\mu_X = \int_0^\infty x\, f_X(x)dx \qquad (1.1.1)$$

be the mean of X and let x_p be the p quantile, that is $F_X(x_p) = p$. Then

$$L_X(p) = \frac{1}{\mu_X} \int_0^{x_p} x f_X(x) dx \,, \tag{1.1.2}$$

is the Lorenz curve. The Lorenz curve is not defined if the mean is zero or infinite.

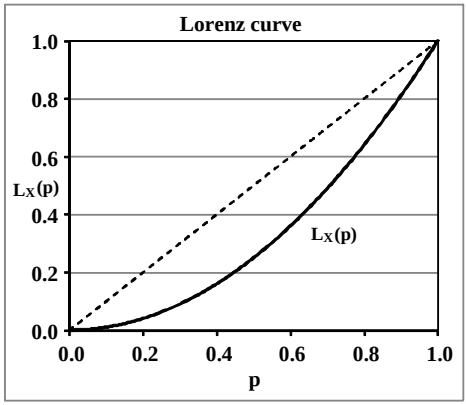

Figure 1.1.1 *A sketch of a Lorenz curve* $L_X(p)$.

Consider a transformed variable $Y = g(X)$, where $g(\cdot)$ is positive and monotone increasing. Define the inverse transformation $X = \gamma(Y)$. Then

$$F_Y(y) = P(Y \le y) = P(g(X) \le g(x)) = P(X \le x) = F_X(x) \,.$$

For the transformed variable Y the p quantile is $F_Y(y_p) = p$, that is $y_p = g(x_p)$.

Now

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(x)}{dx}\frac{dx}{dy} = f_X(x)\frac{dx}{dy} = f_X(x)\frac{d\gamma}{dy} \,. \tag{1.1.3}$$

Hence

$$\mu_Y = \int_0^{\infty} y f_Y(y)dy = \int_0^{\infty} g(x) f_X(x)dx \tag{1.1.4}$$

and

$$L_Y(p) = \frac{1}{\mu_Y} \int_0^{x_p} g(x) f_X(x)dx. \tag{1.1.5}$$

If the transformation is linear $g(x) = \theta x$, then $Y = \theta X$, $\mu_Y = \theta \mu_X$,

$$L_Y(p) = \frac{1}{\theta \mu_X} \int_0^{x_p} \theta x f_X(x)dx = L_X(p) \tag{1.1.6}$$

and consequently, the Lorenz curve is invariant under linear transformations.

A simple example of this property is that the Lorenz curve of the income distribution is independent of the currency used.

Consequently, the Lorenz curve satisfies the general rules:

To every distribution $F(x)$ corresponds a unique Lorenz curve, $L_X(p)$. The contrary does not hold because every Lorenz curve $L_X(p)$ is a common curve for a whole class of distributions $F(\theta x)$ where θ is an arbitrary positive constant.

A Lorenz curve always starts at $(0, 0)$ and ends at $(1, 1)$. The higher Lorenz curve the lesser is the inequality of the income distribution. The diagonal $L(p) = p$ is commonly interpreted as the Lorenz curve for complete equality between the income receivers, but according to Wang et al. (2011), $L(p) = p$ is strictly speaking not a Lorenz curve associated with complete inequality. They

discuss the possibility how to identify this Lorenz curve with the situation that all individuals receive the same income. Mathematically this result can be obtained as a limiting curve when the inequality of the income distribution converges towards zero. Increasing inequality lowers the Lorenz curve and theoretically, it can converge towards the lower right corner of the square.

Consider two variables X and Y, their distributions $F_X(x)$ and $F_y(y)$, and their Lorenz curves $L_X(p)$ and $L_Y(p)$. If $L_X(p) \geq L_Y(p)$ for all p, then measured by the Lorenz curves, the distribution $F_X(x)$ has lower inequality than the distribution $F_y(y)$ and $F_X(x)$ is said to *Lorenz dominate* $F_y(y)$. We denote this relation $F_X(x) \succ_L F_Y(y)$. An example of Lorenz dominance is given in Figure 1.1.2. This is the common definition of the Lorenz dominance although that some define the dominance in the opposite way.

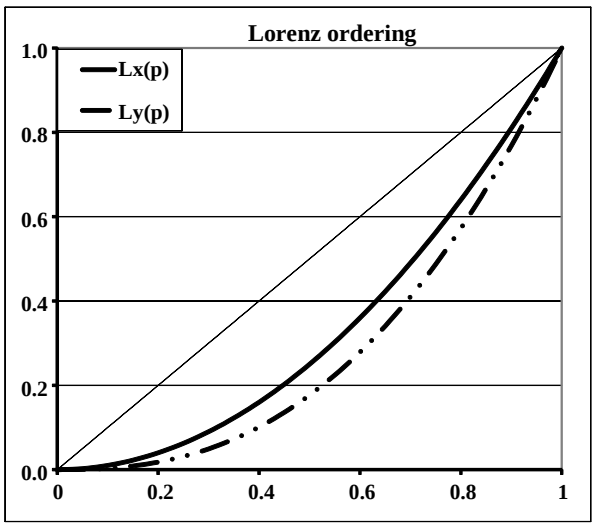

Figure 1.1.2 *Lorenz curves with Lorenz ordering, that is* $L_X(p) \succ_L L_Y(p)$.

Income inequalities can be of different type and the corresponding Lorenz curves may intersect and for these no Lorenz ordering can be identified (c.f. Figure 1.1.3). The Lorenz curve $L_2(p)$ corresponds to a population where the poor are relatively not so poor and the rich are relatively rich. On the other hand the Lorenz curve $L_1(p)$ corresponds to a population with very poor among the poor and the rich are not so rich.

For intersecting Lorenz curves alternative inequality measures have to be defined. The most frequently used is the Gini coefficient, *G* (Gini, 1914). Using the Lorenz curves, this coefficient is the ratio between the area between the diagonal and the Lorenz curve and the whole area under the diagonal. The formula is

$$G = 1 - 2\int_0^1 L(p)dp. \tag{1.1.7}$$

This definition yields Gini coefficients satisfying the inequalities $0 < G < 1$. The higher *G* value the stronger inequality. If $G_X < G_Y$, then the distribution $F_X(x)$, measured by the Gini coefficient, has lower inequality than the distribution $F_y(y)$ and we say that $F_X(x)$ *Gini dominates* $F_y(y)$. We denote this relation $F_X(x) \underset{G}{\succ} F_Y(y)$.

Yitzhaki (1983) proposed the generalized Gini coefficient

$$G(\nu) = 1 - \nu(1-\nu)\int_0^1 (1-p)^{\nu-2} L(p)dp, \tag{1.1.8}$$

where $\nu > 1$. Different ν's are used in order to identify different inequality properties. For low ν's greater weights are associated with the rich and for high

ν's greater weights are associated with the poor. Using the mean income (μ) and the Gini coefficient (G), Sen (1973) proposed a welfare index

$$W = \mu(1-G). \tag{1.1.9}$$

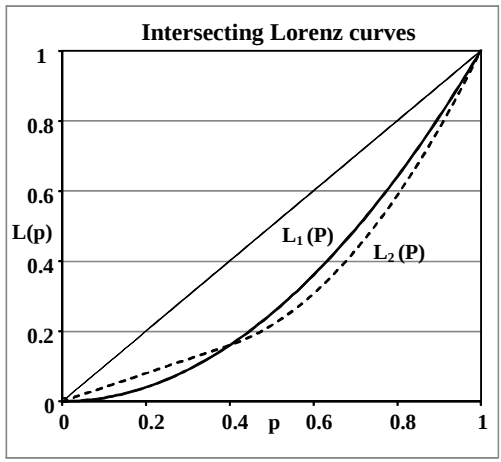

Figure 1.1.3 *Two intersecting Lorenz curves. Using the Gini coefficient presented in the text,* $L_1(p)$ *has less inequality (*$G_1 = 0.3333$*) than* $L_2(p)$ *(*$G_2 = 0.3600$*). The Pietra coefficients, presented below, are* $P_1 = 0.2500$ *and* $P_2 = 0.2940$.

Alternative inequality measures have been defined and such measures are discussed later in section 1.3.

1.2 Income Distributions

According to Aichison and Brown (1954) general description of an income distribution may be defined as a rule which gives for each value of income *x* the proportion $F(x)$ of persons in a given population who have an income not greater than *x*. Such a description is a useful analytical tool if it requires that $F(x)$ has to be given a precise mathematical expression involving known, or more frequently unknown, parameters. It is interesting to recall that Pareto (1897), when he first presented his law, emphasised its empirical basis, but on

the other hand the process of reasoning by Gibrat (1931) started from theory to observations.

Aichison and Brown (1954) gave four criteria on which the success of a particular description may be assessed.

- How closely the description approximate to the observed distribution of incomes when specific values are assigned to the parameters? These values will usually be estimated from the data.
- To what extent may the statistical description be shown to rest on assumptions which are consistent with our knowledge of the way in which incomes are generated?
- What facilities does the description provide in the statistical analysis of the data?
- What economic meaning or significance can be attached to the parameters of the description?

Furthermore, Aichison and Brown gave a thorough presentation of studies of income distributions presented during the first half of the 20^{th} century. They stated that it is well known that income distributions almost invariably possess a single mode and are positively skewed. Many statistical descriptions satisfying these rather general conditions have been proposed in the past as applicable to the distribution of incomes, among which one may note the frequency curves of Pareto (1897), Kapteyn (1903), Gibrat (1931) and Champernowne (1953).

Already Quensel (1944) stated that the lognormal curve agrees fairly well with the actual distribution of the lower incomes, although the Pareto curve often provides a more adequate description of the higher incomes.

Champernowne (1953) described an ingenious model which under realistic assumptions generates exactly or approximately a distribution of incomes obeying Pareto ́s law. Champernowne ́s model provides a basis for the comparison of processes of generating the Pareto and the lognormal descriptions of income distributions. Before Champernowne ́s article Rhodes (1944) and Castellani (1950) presented attempts to derive Pareto ́s distribution.

Furthermore, Aichison and Brown (1954) noted that the law of proportionate effect, postulated by models predicting lognormality, is less appropriate when we are considering a heterogeneous group of income receivers than if the population is divided in sectors, within each the postulate applies. Under the assumptions which are necessary for the application of the central limit theorem, the multiplicative form of the central limit theorem leads us to expect that the distribution of incomes will eventually be described by a lognormal curve. If the population is divided into a large number of sectors and that in each sector the basic postulate of proportionate effect may be assumed to apply, means that a lognormal description of incomes will be valid in each sector, though the parameters of the description may take on different numerical values in each sector.

Finally, Aichison and Brown (1954) stressed that it is useless to posit a statistical description of income distribution unless it is possible with the help of this description to derive analytical tools for any investigation that is likely to be required. To take an extreme example, there would be little point in giving $F(x)$ an explicit mathematical form involving unknown parameters if no method of estimating these parameters from data were available. It is, however, comforting in statistical work to be sure that one is not wasting any of the information available and this is always possible with the lognormal description.

An example of a skewed lognormal distribution can be seen in Figure 1.2.1. Note that the income receivers in this example are a homogeneous group from the upper part of the hierarchy (c.f. Aichison and Brown, 1954).

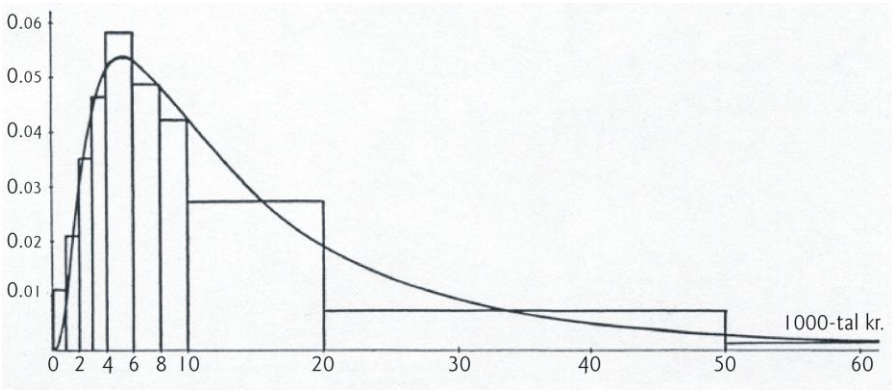

Figure 1.2.1 *Distribution of the income among 4103 industrial managers compared to a lognormal distribution (Cramér, 1949).*

McDonald and Ransom (1979) compared alternative income distribution models and applied them on US family income data. The interesting models were the lognormal, the gamma, the beta and the Sing-Maddala functions. They applied the models on family income for 1960 and 1969 through 1975 and compared the estimation methods: the method of scoring, the Pearson minimum chi-squared method and the least squares estimation. The estimation of the mean income and the Gini coefficient were directly obtained by substituting estimates of the parameters characterizing the associated distribution functions into the appropriate theoretical expressions of the coefficients. They noted that even though they observed situations in which parameter estimates change significantly from one time period to another, the associated population characteristics such as the mean and the Gini coefficients are much more stable. However, the estimated Gini coefficients associated with the scoring and the minimum chi-squared estimates of the lognormal density are much larger than for any other case considered. A general observation was that the scoring and

the minimum chi-squared results were very similar for the three parameter functions, with greater differences for the gamma, and still greater for the lognormal.

Summing up, McDonald and Ransom concluded that the gamma provided a better fit than the lognormal, regardless of the estimation technique used. The three parameter functions (beta and Singh-Maddala) provided a better fit to the data than did the gamma density function. This finding is obviously due to the number of distribution parameters.

Over time has come the realization that only the upper tail of the distribution is Pareto in form. Proceeding from the observation that the distribution has a Pareto tail for the top 15-20% of employees. Lydall (1968) advances a model of hierarchal earnings based on the notation that large organisations are organised on hierarchical principle.

Harrison (1981) noted that a number of observed earnings distributions were well described by the Pareto distribution

$$F(y)=\begin{cases}0 & y\le 1\\ 1-y^{-\alpha} & y>1\end{cases}, \tag{1.2.1}$$

where $\alpha>0$ and $y=Y/Y_L$, Y_L being the minimum income. For $\alpha>1$, the mean is $E(Y)=\frac{\alpha}{\alpha-1}$. Furthermore, the Lorenz curve is $L(p)=1-(1-p)^{\frac{\alpha-1}{\alpha}}$ and the Gini coefficient is $G=\frac{1}{2\alpha-1}$. It may perhaps be convenient to remark here that for commonly occurring values of the parameter α a second moment of the Pareto distribution does not exist unless $\alpha>2$. Furthermore, Harrison stressed that equally compelling reasons supporting the use of disaggregated data can be found in the case of the lognormal function.

Dagum (1977, 1980, 1987) has paid continuous attention to alternative income distributions.

A common technique for estimating the Pareto constant, α, is to linearize the survival function by taking logarithms, and apply ordinary least squares. The survival function is

$$S(y) = 1 - F(y) = \left(\frac{Y}{Y_L} \right)^{-\alpha}.$$

After taking natural logarithms one obtains the linear model

$$ln\big(S(y)\big) = -\alpha\, ln(Y) + \alpha\, ln(Y_L) = C - \alpha\, ln(Y).$$

This model indicates a linear, decreasing association between $\ln\big(S(y)\big)$ and $\ln(Y)$. A regression analysis gives an estimate of α and the coefficient of determination, R^2, measures the linearity in the model and the goodness of fit of the Pareto model.

We apply this analysis on annual taxable incomes in Finland for 2009 (http://pxweb2.stat.fi/Database/StatFin/tul/tvt/2009/2009_en.asp).

The data are presented in a grouped table (Table 1.2.1). We assume that the Pareto model may start from ca. $Y = 25000$€. For values equal to or greater than that we obtain the estimate $\hat{\alpha} = 2.637$ and in addition, the coefficient of determination is $R^2 = 0.99241$. For the income distribution for incomes greater than 25000 the Gini coefficient is $G = \dfrac{1}{2\alpha - 1} = 0.234$.

Table 1.2.1 *Taxable income receivers in Finland 2009.*

Classes of annual income (€)	Number of income recipients
- 1000	182281
1000 - 2000	96836
2000 - 3000	80056
3000 - 4000	65800
4000 - 5000	59595
5000 - 6000	62171
6000 - 7000	107558
7000 - 8000	146526
8000 - 9000	114602
9000 - 10000	121555
10000 - 12500	319042
12500 - 15000	329083
15000 - 17500	259979
17500 - 20000	243284
20000 - 25000	481753
25000 - 30000	487376
30000 - 35000	385672
35000 - 40000	266075
40000 - 50000	307810
50000 - 60000	152714
60000 - 80000	120327
80000 -	88488
All	4478583

In Figure 1.2.2 we sketch the result.

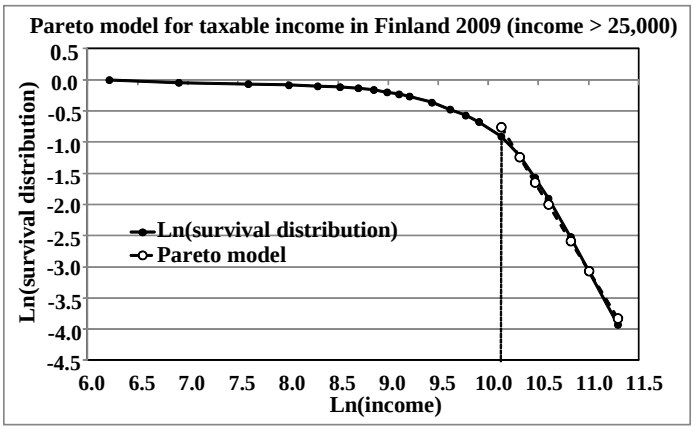

Figure 1.2.2 *Graphical sketch of the distribution of taxable income in Finland (2009) and a Pareto model for annual incomes greater than* $Y = 25000\,€$.

1.3 Lorenz Curves and Concentration of Incomes

A central topic in the analyses of income distributions is the concept of *concentration of incomes*, which is defined in the literature (Lorenz, 1905) in such a way as to be free of any particular hypothesis concerning the genesis of the description of the income distribution.

In Section 1.1 we introduced the Lorenz curve $L(p)$ defined by $L(p) = \frac{1}{\mu}\int_0^{x_p} x\, f(x)dx$, where $\mu_X = \int_0^{\infty} x\, f_X(x)dx$ is the mean and $F(x_p) = p$. Lorenz curves were presented in the Figures 1.1.1, 1.1.2 and 1.1.3.

The Lorenz curve has the following general properties:

i. $L(p)$ is monotone increasing.

ii. $L(p) \le p$.

iii. $L(p)$ is convex.

iv. $L(0)=0$ and $L(1)=1$.

The Lorenz curve $L(p)$ is convex because the income share of the poor is less than their proportion of the population. The higher Lorenz curve the lesser inequality in the income distribution (c.f. Section 1.1).

The Lorenz curve for a probability distribution is a continuous function. However, Lorenz curves representing discontinuous functions can be constructed as the limit of Lorenz curves of probability distributions, the line of perfect inequality being an example.

If the Lorenz curve is differentiable the derivatives have the following properties. Let $L_X(p)=\frac{1}{\mu_X}\int_0^{x_p} x\, f_X(x)dx$, $F_X(x_p)=p$ and the density function $f_X(x)$. When we differentiate the equation $F_X(x_p)=p$ we obtain

$$\frac{dF_X(x_p)}{dp}=\frac{dF_X(x_p)}{dx_p}\frac{dx_p}{dp}=1,$$

$$f_X(x_p)\frac{dx_p}{dp}=1$$

and

$$\frac{dx_p}{dp}=\frac{1}{f_X(x_p)}.$$

The derivation of $L_X(p)=\frac{1}{\mu_X}\int_0^{x_p} x\, f_X(x)dx$ yields

$$\frac{dL_X(p)}{dp} = \frac{1}{\mu_X} \frac{d\int_0^{x_p} x f_X(x)dx}{dx_p} \frac{dx_p}{dp} = \frac{1}{\mu_X} x_p f_X(x_p) \frac{dx_p}{dp} = \frac{x_p}{\mu_X}$$

and consequently,

$$\frac{dL_X(p)}{dp} = \frac{x_p}{\mu_X} \tag{1.3.1}$$

If the Lorenz curve is differentiable twice, then the second derivative is

$$\frac{d^2L_X(p)}{dp^2} = \frac{1}{\mu_X} \frac{dx_p}{dp} = \frac{1}{\mu_X} \frac{1}{f_X(x_p)}.$$

Hence,

$$\frac{d^2L(p)}{dp^2} = \frac{1}{\mu_X f_X(x_p)} \tag{1.3.2}$$

The difference between the diagonal and the Lorenz curve

$$D = p - L_X(p)$$

$$\frac{dD}{dp} = 1 - L'_X(p) = 1 - \frac{x_p}{\mu_X}$$

$$\frac{d^2D}{dp^2} = -L''_X(p) = -\frac{1}{\mu_X} \frac{dx_p}{dp} = -\frac{1}{\mu_X f_X(x)} < 0.$$

The maximum of D implies $1 - \frac{x_p}{\mu_X} = 0$, that is $x_p = \mu$.

For $x_p = \mu_X$, $L'_X(p) = \frac{\mu_X}{\mu_X} = 1$ and at the point $p_\mu = F_X(\mu_X)$ the tangent is parallel to the line of perfect equality. This is also the point at which the vertical

distance between the Lorenz curve and the egalitarian line attains its maximum $P = p_\mu - L_X(p_\mu)$. This maximum is defined as the Pietra index (Lee, 1999). According to this definition $0 < P < 1$. The lower bound is obtained when there is total income-equality that is the Lorenz curve coincides with the diagonal. The upper bound can be obtained when the Lorenz curve converges towards the lower right corner. The Pietra index can be interpreted as income of the rich that should be redistributed to the poor in order to obtain total income equality. Therefore, the index is sometimes named the Robin Hood index. Lee (1999) used the Lorenz curve and the summary measures based on it for diagnostic tests medical studies. He associated the Gini and the Pietra indices with the receiver operating characteristic curve (ROC). He also gave in his reference list additional papers where these summary statistics were applied.

An alternative definition has also been given. The Pietra index can be defined as twice the area of the largest triangle inscribed in the area between the Lorenz curve and the diagonal line (Lee, 1999). In Figure 1.3.1 one observes that the triangle obtains its maximum when the corner lies on the Lorenz curve where the tangent is parallel to the diagonal. The height of the triangle is $h = \frac{P}{\sqrt{2}}$ and the base is the diagonal $b = \sqrt{2}$. The double of the area is

$$2\text{area} = 2\frac{h\sqrt{2}}{2} = 2\frac{P\sqrt{2}}{2\sqrt{2}} = P.$$

Compared to the Gini coefficient we obtain that $G > P$ (see, Lee, 1999).

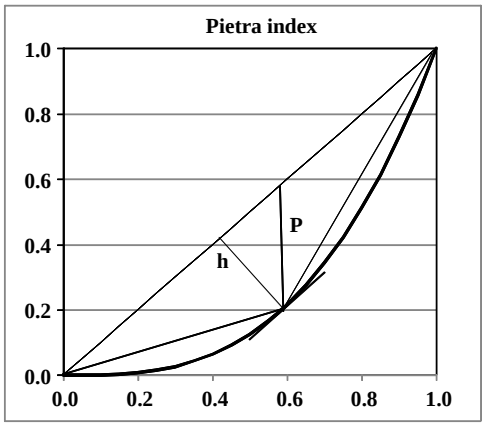

Figure 1.3.1 *The Lorenz curve and the geometric interpretations of the Pietra index.*

The definition yields Pietra coefficients satisfying the inequality $0 \le P \le 1$. If $P_X < P_Y$, then the distribution $F_X(x)$ measured by the Pietra index has lower inequality than the distribution $F_y(y)$ and we say that $F_X(x)$ *Pietra dominates* $F_y(y)$. We denote this relation $F_X(x) \underset{P}{\succ} F_Y(y)$. For the Lorenz curves in Figure 1.1.3, $P_1 \approx 0.2500$ and $P_2 \approx 0.2940$. According to the Pietra index, $L_1(p)$ is less unequal than $L_2(p)$.

In general, the Pietra and the Gini orderings are not identical. The following simple example supports this statement. Consider the situation described in Figure 1.3.2. There are two polygonal Lorenz curves, OABC ($L_1(p)$) and ODC ($L_2(p)$). [1] For $L_1(p)$ we obtain $P_1 < G_1$ and for $L_2(p)$ we obtain $P_2 = G_2$ because ODC is a triangle yielding identical indices. Furthermore if the point D

[1] The Lorenz curves in this example are not continuously differentiable, but slight modifications yield differentiable Lorenz curves. One has only to modify the edges to mini curves. If these modifications are minute, the inequalities given above still hold.

is close to the line AB, we observe that $G_1 > G_2$ and $P_1 < P_2$. Combining these inequalities we obtain $P_1 < P_2 = G_2 < G_1$. Consequently, $L_1(p)$ Pietra dominates $L_2(p)$, but $L_2(p)$ Gini dominates $L_1(p)$.

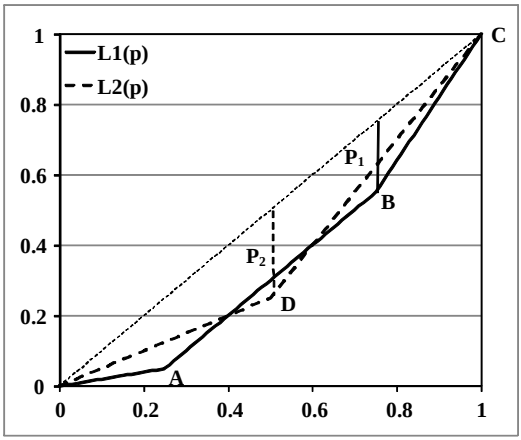

Figure 1.3.2 *Comparisons between Gini and Pietra indices. For the Lorenz curve* $L_1(p)$ *the Pietra index is* $P = 0.20$ *and the Gini coefficient is* $G = 0.30$ *and for Lorenz curve* $L_2(p)$ *the Pietra index is* $P = 0.25$ *and the Gini coefficient is* $P = 0.25$.

Above we obtained the inequality $0 < P < 1$. The limits in the inequalities can be obtained and this can be explained by the following example and Figure 1.3.3.

Consider the simplified RT model defined in (1.4.5)

$$L(p) = p^{\alpha} \quad \alpha \geq 1.$$

Examples of these Lorenz curves are sketched in Figure 1.3.3. The Gini coefficient is $G = \frac{\alpha - 1}{\alpha + 1}$. When $\alpha \to 1$ then $G \to 0$ and when $\alpha \to \infty$ then $G \to 1$. The Pietra index is $P = \left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}} - \left(\frac{1}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}$. We select a sequence of α

values $\alpha = 1 + \frac{1}{n}$, for $n = 1, 2, \ldots$. The P values are

$$P = \left(1 - \frac{1}{n+1}\right)^n - \left(\left(1 - \frac{1}{n+1}\right)^n\right)^{\left(1+\frac{1}{n}\right)}.$$

When $n \to \infty$, both terms converge towards e^{-1} and $P \to 0$. According to the definition of the P index

$$P = \left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}} - \left(\frac{1}{\alpha}\right)^{\frac{\alpha}{\alpha-1}} \geq p - p^{\alpha} \text{ for all } p < 1.$$

For increasing α values the supremum of $p - p^{\alpha}$ is one. This must also be the supremum of $P = \left(\frac{1}{\alpha}\right)^{\frac{1}{\alpha-1}} - \left(\frac{1}{\alpha}\right)^{\frac{\alpha}{\alpha-1}}$. Consequently, the interval $0 < P < 1$ cannot be shortened.

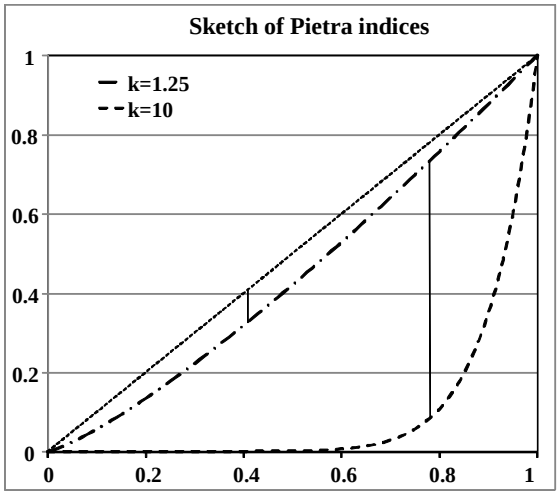

Figure 1.3.3 *Sketches of extreme Lorenz curves with corresponding P indices. For the Lorenz curve* $k = 1.25$ *the Pietra index is 0.0819 and for the Lorenz curve* $k = 10$ *the index is 0.6966.*

We can prove ($\lim_{p\uparrow 1}$ denotes limit from the left).

Theorem 1.3.1. If μ_X exists, then $\lim_{p\uparrow 1} L'(p)(1-p) = 0$.

Proof. Consider the integral $\int_x^\infty t\, f_X(t)dt$. If μ_X exists, then $\int_0^\infty t\, f_X(t)dt = \mu_X$ and for every $\varepsilon > 0$ there exists an x' such that $\int_x^\infty t\, f_X(t)dt < \varepsilon$ if $x > x'$. Choose p so that $x_p > x'$, then

$$\varepsilon > \int_{x_p}^\infty t\, f_X(t)dt \geq x_p \int_{x_p}^\infty f_X(t)dt = x_p(1-p). \tag{1.3.3}$$

As a consequence of (1.3.3),

$$\lim_{p\uparrow 1} L'_X(p)(1-p) = \lim_{p\uparrow 1} \frac{x_p}{\mu_X}(1-p) = \frac{1}{\mu_X} \lim_{p\uparrow 1} x_p(1-p) = 0.$$

Consider an one-parametric class of cumulative distribution functions $F(x,\theta)$, defined on the positive x-axis. If we assume that $F(x,\theta) = F(\theta x)$, i.e. it depends only on the product θx, then the following theorem holds:

Theorem 1.3.2. Let $F(x,\theta)$ be a one-parametric class of distributions with the properties:

i. $F(x,\theta) = F(\theta x)$.

ii. $F(\theta x)$ is defined on the positive x-axis.

iii. $F(\theta x)$ and its derivative are continuous.

iv. $\mu_X = E(X)$ exists.

Let $T = \theta X$, then

$$x_p(\theta) = \frac{t_p}{\theta} \tag{1.3.3}$$

and

$$\mu_X(\theta) = \frac{c}{\theta}, \tag{1.3.5}$$

where t_p and c are independent of θ.

Proof. Let θ be an arbitrary, positive parameter. Then the quantile $x_p(\theta)$ is defined by the equation $F(\theta x_p) = p$. If we define t_p by the equation $F(t_p) = p$ then t_p does not depend on θ and $\theta x_p(\theta) = t_p$ and (1.3.3) is proved. The formula (1.3.5) and the statement that $L(p) = \frac{1}{\mu(\theta)} \int_0^{x_p(\theta)} x\, dF(\theta x)$ is independent of θ is proved by using the substitution $t = \theta x$ in the integrals $E(X) = \int_0^{\infty} x\, dF(\theta x)$ and $L(p) = \frac{1}{\mu(\theta)} \int_0^{x_p(\theta)} x\, dF(\theta x)$.

Furthermore, we can prove:

Theorem 1.3.3. Consider a function $L(p)$ defined on the interval [0, 1] with the properties:

i. $L(p)$ is monotone increasing and convex.

ii. $L(0) = 0$ and $L(1) = 1$.

iii. $L(p)$ is differentiable twice.

iv. $\lim_{p\uparrow 1} L'(p)(1-p)=0$.

then $L(p)$ is a Lorenz curve of a distribution with finite mean.

Proof. If we denote the unknown distribution $F(x)$ and its derivative $f(x)$, then necessarily $L'(p)=\frac{x_p}{\mu}$. The derivative $L'(p)$ is a monotone-increasing function. If its inverse is denoted $M(p)$, we get the necessary relation

$$F(x_p)=p=M(\frac{x_p}{\mu}).$$

If $\theta=\frac{1}{\mu}$, then $F(x)=M(\theta x)$. Now we shall prove the sufficiency, that is, that $M(\theta x)$ is a distribution, whose mean is $\mu=\frac{1}{\theta}$ and whose Lorenz curve is $L(p)$. We denote $M(\theta x)=F(x)$ then $f(x)=F'(x)=\theta M'(\theta x)$. After observing that the property (iv) indicates that $L'(p)$ is integrable from *0* to *1*, we introduce the variable transformation

$$y=M(\theta x)$$

$$dy=\theta M'(\theta x)dx$$

$$x=\frac{1}{\theta}L'(y)$$

We obtain

$$\mu=\lim_{t\to\infty}\int_0^t x\theta M'(\theta x)dx=\lim_{p\uparrow 1}\int_0^p \frac{1}{\theta}L'(y)\,dy=\frac{1}{\theta}\lim_{p\uparrow 1}\int_0^p L'(y)\,dy=\frac{1}{\theta}$$

The given function $L'(p)$ has a monotone-increasing inverse function, $M(\theta x)$ giving a corresponding distribution function $F(x) = M(\theta x)$ whose mean is μ.

Using the same transformation we obtain that the Lorenz curve $\tilde{L}(p)$ of $F(x) = M(\theta x)$ is

$$\tilde{L}(p) = \theta \int_0^{x_p} x\theta M'(\theta x)dx = \int_0^p L'(v)\,dv = \int_0^p L'(v)\,dv$$

and the theorem is proved.

These results have been collected in the following theorem (Fellman, 1976, 1980).

Theorem 1.3.4. Consider a given function $L(p)$ with the properties:

i. $L(p)$ is monotone increasing and convex to the p-axis.

ii. $L(0) = 0$ and $L(1) = 1$.

iii. $L(p)$ is differentiable.

iv. $\lim_{p \uparrow 1} L'(p)(1-p) = 0$.

Then $L(p)$ is the Lorenz curve of a whole class of distribution functions $F(\theta x)$, where θ is an arbitrary positive constant and the function $F(\cdot)$ is the inverse function to $L'(p)$.

In Fellman (1976) the result was presented and later Fellman (1980) presented the following theorem.

Theorem 1.3.5. A class of continuous distributions $F(x,\theta)$ with finite mean has a common Lorenz curve if and only if $F(x,\theta) = F(\theta x)$.

The formula for Gini coefficient G is given in (1.1.7) and G satisfies the inequality $0 \le G \le 1$. The higher G value the stronger inequality in the income distribution. Later, alternative inequality indices have been defined and introduced. The generalized Gini coefficient $G(\nu) = 1 - \nu(1-\nu)\int_0^1 (1-p)^{\nu-2} L(p)dp$, where $\nu > 1$, is given in (1.1.3) and has been proposed by Yitzhaki (1983) in order to identify different distribution properties. The welfare index $W = \mu(1-G)$, given in (1.1.8) and proposed by Sen (1973) is based on the mean income (μ) and the Gini coefficient (G).

Kleiber and Kotz (2001, 2002) have outlined how the income distributions can be characterised by their Lorenz curves:

1.4 Modelling Lorenz Curves

As an alternative to income distributions some scientists have built models for the Lorenz curve. Among these we may list the following studies: Kakwani & Podder (1973, 1976), Kakwani (1980), Rasche et al. (1980), Gupta (1984), Rao & Tam (1987), Chotikabanich (1993), Ogwang & Rao (2000), Cheong (2002), Rohde (2009) and Fellman (2012). The theoretical step from Lorenz curve to distribution function is more difficult than that from distribution function to Lorenz curve. Fellman (2012) noted that there is a difference between advanced and simple Lorenz models. Advanced Lorenz models yield a better fit to data, but are difficult to exactly connect to income distributions. Simple one-parameter models can more easily be associated with the corresponding income distribution, but when statistical analyses are performed the goodness of fit is often poor.

Rao and Tam (1987) compared five different models. The first was the three-parameter Kakwani & Podder (KP) model (1973),

$$\eta = a\pi^b\left(\sqrt{2}-\pi\right)^c \quad \begin{matrix} a>0 \\ 0\le b\le 1 \\ 0\le c\le 1 \end{matrix}, \tag{1.4.1}$$

where $\pi = \dfrac{L+p}{\sqrt{2}}$ and $\eta = \dfrac{L-p}{\sqrt{2}}$.

The second is the two-parameter generalised Pareto model (GP) analysed by Rasche et al. (1980)

$$L_{GP} = \left(1-(1-p)^a\right)^{1/b} \qquad \begin{matrix} 0\le a\le 1 \\ 0\le b\le 1 \end{matrix}, \tag{1.4.2}$$

and the third is the one-parameter Gupta (G) model (1984)

$$L_G = p\beta^{p-1}, \quad \beta>1. \tag{1.4.3}$$

In addition, Rao and Tam constructed a generalized two-parameter Gupta model (RT)

$$L_{RT} = p^a\beta^{p-1}, \quad a,\beta>1. \tag{1.4.4}$$

Finally, they introduced a simplified one-parameter version (S) of the RT model ($\beta=1$)

$$L_S = p^a \quad \alpha>1 \tag{1.4.5}$$

Chotikabanich (1993) defined an alternative one-parameter Lorenz curve (C):

$$L_C(p) = \frac{e^{kp}-1}{e^k-1} \qquad k>0. \tag{1.4.6}$$

The models G, S and C contain only one parameter. They are so simple that it is impossible to distinguish between the estimated length of the range for the income distribution function and the Gini coefficient. If one of these properties is estimated the other is fixed. Therefore, Fellman (2012) paid these models special attention and analysed them in more detail.

In general, the step from the Lorenz curve to the income distribution starts from the formula

$$L'(p)=\frac{x_p}{\mu}, \tag{1.4.7}$$

where x_p is the p-percentile and μ is the mean of the corresponding distribution $F(x)$. We define $M(\cdot)$ as the inverse function of $L'(\cdot)$. From (1.4.7) we obtain

$$p = M\left(\frac{x_p}{\mu}\right). \tag{1.4.8}$$

Equation (1.4.8) indicates that $M(\cdot)$ is the income distribution function corresponding to the given Lorenz curve, that is, $F(x) = M\left(\frac{x}{\mu}\right)$. This connection between the Lorenz curve and the distribution function is easily defined, but for most of the exact Lorenz curves it is difficult or even impossible to obtain the income distribution mathematically.

The Gupta model. Examples of Lorenz curves for the Gupta model (1.4.3) are given in Figure 1.4.3.

Following Gupta, we observe that

$$L'_G(p) = p\beta^{p-1}\log\beta + \beta^{p-1} = \frac{x_p}{\mu}. \tag{1.4.9}$$

Consequently,

$$\lim_{p\to 0^+} x_p = \mu \lim_{p\to 0^+} L'_G = \frac{\mu}{\beta} \tag{1.4.10}$$

and

$$\lim_{p\to 1^-} x_p = \mu \lim_{p\to 1^-} L'_G = \mu(1+\log\beta) \tag{1.4.11}$$

From this it follows that Gupta ś model corresponds to distributions defined on a finite interval $\left(\mu\beta^{-1}, \mu(1+\log\beta)\right)$. In spite of the fact that the Gupta model is relatively simple, the corresponding income distribution is not attainable. The equation (1.4.9) cannot be solved exactly with respect to variable *p* because the variable *p* can be found both as a factor and in the exponent.

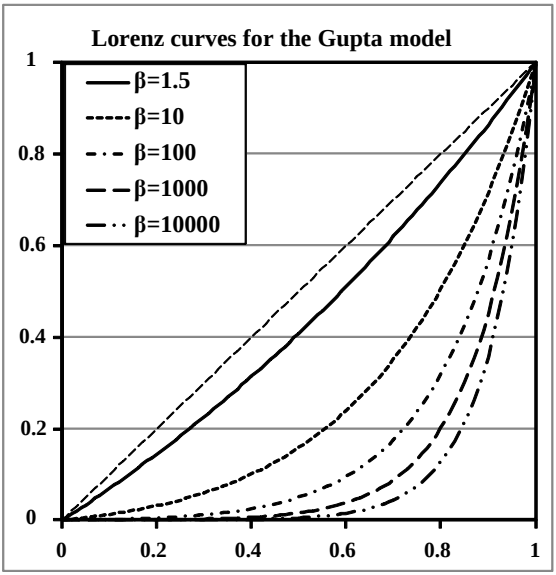

Figure 1.4.3 *The Lorenz curves for the Gupta model for various β values (Fellman, 2012).*

For the Gupta model, the Gini coefficient is

$$G_G = 1 - 2\int_0^1 p\beta^{p-1}dp = 1 - \frac{2}{\log\beta}\left(1 - \frac{\beta - 1}{\beta \log\beta}\right). \tag{1.4.12}$$

Figure 1.4.3 shows that the Gini coefficient tends towards 1 when $\beta \to \infty$.[2]

Following Gupta, the variable $log\,\beta$ can be estimated by using the logarithm of the model in (1.4.4), that is, from the equation $\log\left(\frac{L}{p}\right) = (p-1)\log(\beta)$.

The generalized Gupta model (RT). For the generalized Gupta model, we obtain.

$$\frac{x_p}{\mu} = L_G'(p) = p^{\alpha-1}\left(p\beta^{p-1}\log\beta + \alpha\beta^{p-1}\right). \tag{1.4.13}$$

The income distribution is defined on the interval $\left(0, \mu(\alpha + \log\beta)\right)$. It can be observed that if $\beta \to \infty$ the range of the income distribution then tends towards $(0, \infty)$ for both the Gupta and the generalized Gupta models.

Following Gradsheteyn and Ryshnik (1965), Rao and Tam give for the generalised Gupta model the Gini coefficient

$$G_{RT} = 1 - 2e^{-\frac{\log\beta}{(1+\alpha)}}\,{}_1F_1(1+\alpha; 2+\alpha; \log\beta), \tag{1.4.14}$$

where ${}_1F_1$ denotes the confluent hyper-geometric function with the parameters indicated in the parentheses.

[2]In Rao and Tam (1987), the formula for the Gini coefficient based on the Gupta model contains a misprint, but a numerical check of the Rao and Tam results indicates that the authors have used the correct formula in their calculations.

The simplified RT model (S). The simplified RT model is obtained for $\beta = 1$ and is given in (1.4.5). The Lorenz curves for various α values are given in Figure 1.4.4.

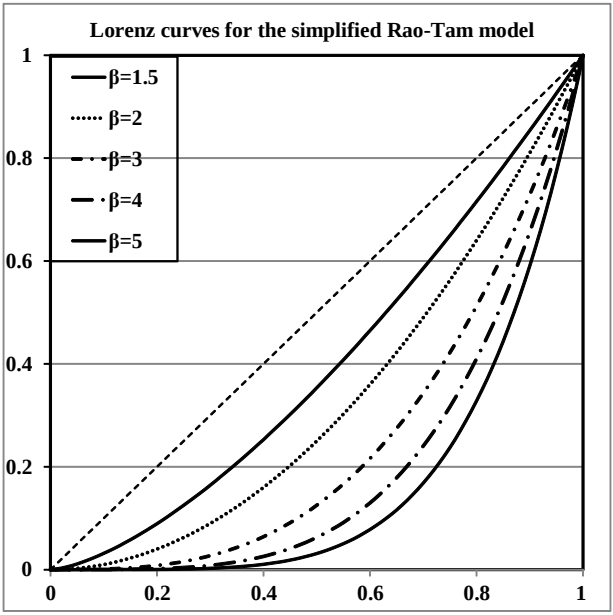

Figure 1.4.4 *Rao-Tam simplified Lorenz curves (Fellman, 2012).*

The Gini coefficient is $G_S = \dfrac{\alpha - 1}{\alpha + 1}$. The income distribution corresponding to the S model can be found. The derivative of $L_S(p) = p^{\alpha}$ is $L_S'(p) = \alpha p^{\alpha-1}$.

We obtain $\dfrac{x_p}{\mu} = L_S'(p) = \alpha p^{\alpha-1}$, $p^{\alpha-1} = \dfrac{x_p}{\alpha\mu}$ and $p = \left(\dfrac{x_p}{\alpha\mu}\right)^{\frac{1}{\alpha-1}}$.

Hence, the income distribution is $F(x) = \left(\dfrac{x}{\alpha\mu}\right)^{\frac{1}{\alpha-1}}$ defined on the interval $(0, \alpha\mu)$. Income distributions are given in Figure 1.4.5 for various α values.

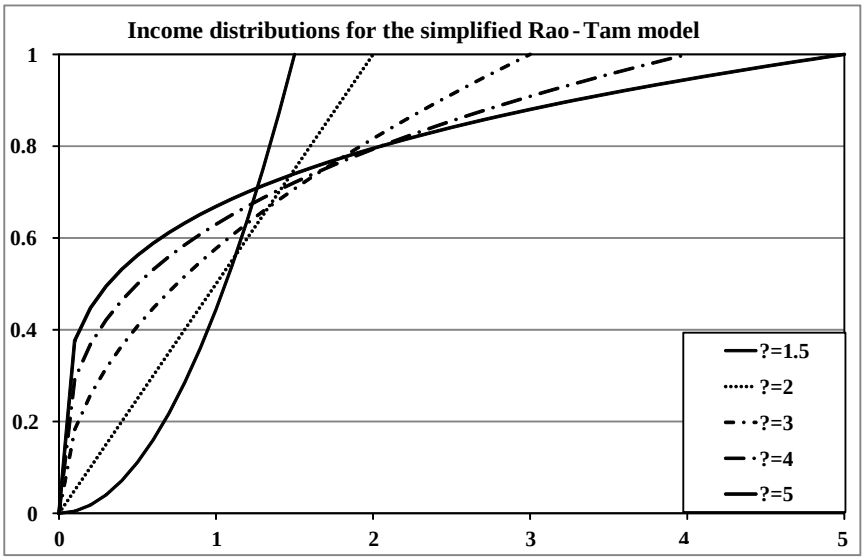

Figure 1.4.5 *Income distributions corresponding to the Rao-Tam simplified Lorenz curve (Fellman, 2012).*

The Chotikabanich model. Chotikabanich (1993) introduced an alternative one-parameter Lorenz curve (cf. 1.4.6)

$$L_C(p) = \frac{e^{kp} - 1}{e^k - 1} \qquad k > 0 .$$

It is easily found that

$$L_C(0) = 0 ,\ L_C(1) = 1 ,\ \frac{dL_C(p)}{dp} = \frac{ke^{kp}}{e^k - 1} > 0$$

and

$$\frac{d^2 L_C(p)}{dp^2} = \frac{k^2 e^{kp}}{e^k - 1} > 0 .$$

The second derivative is positive and hence the Lorenz curve is convex. Consequently, the first derivative is increasing from the minimum $\frac{dL_C(0)}{dp} = \frac{k}{e^k - 1} > 0$ to $\frac{dL_C(1)}{dp} = \frac{ke^k}{e^k - 1}$.

If we consider an income distribution with the mean μ, then income is distributed over the interval ($\frac{\mu k}{e^k - 1}$, $\frac{\mu k e^k}{e^k - 1}$). When $k \to \infty$, this interval converges towards $(0, \infty)$

Lorenz curves as functions of parameter *k* are given in Figure 1.4.6.

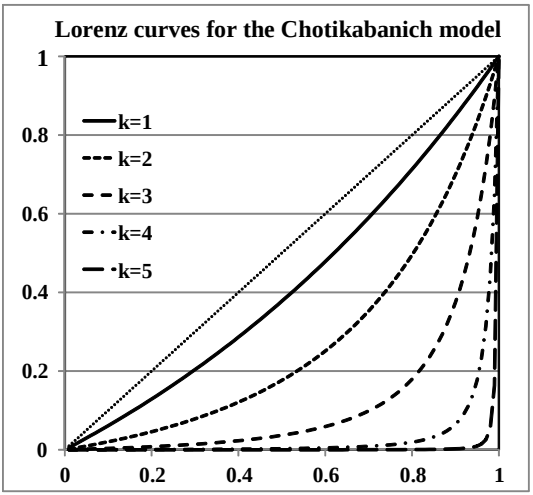

Figure 1.4.6 *Lorenz curves for the Chotikabanich models (c.f. Fellman, 2012).*

The Gini coefficient is.

$$G_C = 1 - 2\int_0^1 L_C(p)dp = 1 - 2\int_0^1 \frac{e^{kp} - 1}{e^k - 1} dp$$

$$= 1 - 2\left(\left(\frac{\frac{1}{k}e^{kp} - 1}{e^k - 1} \right)_{p=1} - \left(\frac{\frac{1}{k}e^{kp} - 1}{e^k - 1} \right)_{p=0} \right)$$

$$= 1 - 2\left(\left(\frac{\frac{1}{k}e^k - 1}{e^k - 1} \right) - \left(\frac{\frac{1}{k} - 1}{e^k - 1} \right) \right) = 1 - 2\left(\frac{e^k - 1}{k(e^k - 1)} \right)$$

$$= \frac{k(e^k - 1) - 2e^k + 2}{k(e^k - 1)} = \frac{(k-2)e^k + k + 2}{k(e^k - 1)}$$

The Gini coefficient increases toward *1* when $k \to \infty$.

If we assume an arbitrary μ , then $x_p = \frac{\mu dL_C(p)}{dp} = \frac{\mu k e^{kp}}{e^k - 1}$ and we get $\frac{\mu k e^{kp}}{e^k - 1} = x_p$. Hence, $F(x) = \frac{1}{k} log\left(\frac{x(e^k - 1)}{\mu k} \right)$ and the theoretical income distribution is obtained.

Figure 1.4.7 presents income distributions for various *k* values.

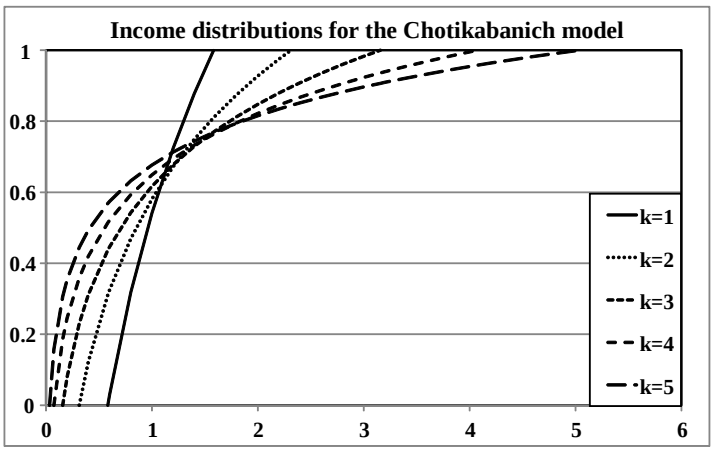

Figure 1.4.7 *Income distributions for the Chotikabanich models (Fellman, 2012).*

Kakwani and Podder (1976) applied their Lorenz model to Australian data, comparing four alternatives, of which all resulted in accurate estimates. The estimates varied between 0.3195 and 0.3208 when the actual value was 0.3196. Rao and Tam (1987) applied the Kakwani-Podder, the generalised Pareto, the RT, the Gupta and the simplified RT models to the same data. Their comparison of the models indicates that the Kakwani-Podder, the generalised Pareto and the RT model yielded the best estimates. The G and the S models resulted in estimates with the largest errors. For the Gupta model, the estimate was too high (0.3691) and for the simplified RT model it was too low (0.2508). The magnitudes of these errors were comparable. These findings support the criticism of the estimation based on simple one-parameter Lorenz models.

Fellman (2012) applied the Chotikabanich model and obtained the following results. He considered $\min_k \sum \left(f_{obs} - f(k)\right)^2$ and estimated the parameter k and performed the minimization by using $f = L$ and $f = log(L)$. Fellman fitted the model to the Kakwani & Podder data obtained, $k = 0.2095$ and $G = 0.3262$, and $k = 0.2097$ and $G = 0.3263$, respectively. He observed that the one-parameter Chotikabanich model yields slightly better but still less exact results. As a comparison, he presented Lorenz models fitted to the Australian data graphically in his Figure 6, which we reprint in Figure 1.4.8. One observes that the Chotikabanich model is closest to the empirical curve. The simplified RT and the Gupta models show larger but comparable discrepancies. These findings support the results obtained by Rao and Tam. In Figure 1.4.8, we also observe that Gupta model yields too high an estimate of G and the simplified model too low an estimate.

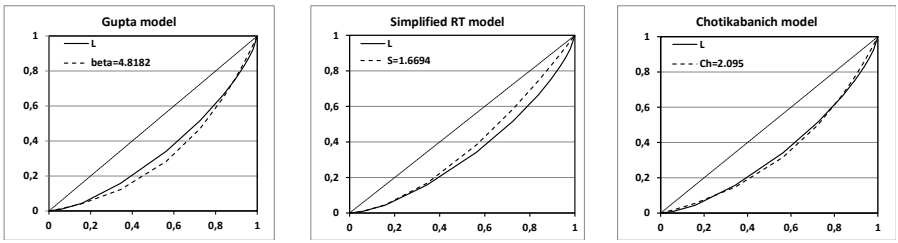

Figure 1.4.8 *Graphical presentation of the goodness of fit obtained by the Gupta, RT and Chotikabanich models. Note that the Chotikabanich gives the best fit (Fellman, 2012).*

Fellman (2012) studied the numerical estimation of the Gini coefficient based on Lorenz curves and this is discussed more in detail in Section 2.4. The methods were the trapezium rule, Simpson ́s rule, a modified version of Golden ́s method (2008) and the Lagrange method. In Fellman (2012) the Simpson rule could not be performed because it demands equidistant points. In general, the trapezium rule yields Gini coefficients which are too low. For the Australian data, the trapezium rule yielded the result *0.3134*, which is slightly below the correct value. Since the Lagrange method demands an even number of sub-intervals, Fellman (2012) had to modify the method slightly. He applied the Lagrange method for the ten last sub-intervals and added a small (triangular) correction from the first sub-interval. The estimate obtained was *0.3199*, a result which is closest to the correct value. Fellman (2012) presented a modified version of Golden ́s method. When he applied this method to the Australian data, he obtained the estimate of *0.3075*. This is too low, but still greater than the extremely low value obtained by the simplified RT model. Summing up, one has to choose the Lorenz model with due consideration. This is especially important if the selection should be performed among simple one-parameter models.

References

[1] Aaberge, R. (2000). Characterizations of Lorenz curves and income distributions. *Social choice and welfare 17*:639-653.

[2] Aichison, J., Brown, J. A. C. (1954). On criteria for description of income distribution. *Metroeconomica 6*:88-107.

[3] Castellani, M. (1950). On Multinomial Distributions with Limited Freedom: A Stochastic Genesis of Pareto's and Pearson's Curves. *Ann. Math. Statist. 21*:289-293.

[4] Cheong, K. S. (2002). An empirical comparison of alternative functional forms for the Lorenz curve. *Applied Economics Letters 9*:171-176

[5] Champernowne, (1953). A model of income distribution. *The Economic Journal 63*:318-351.

[6] Chotikabanich, D. (1993). A comparison of alternative functional forms for the Lorenz curve. *Economics Letters 41*:129-138.

[7] Cramér, H. (1949). *Sannolikhetskalkylen och några av dess användningar*. Almqvist & Wiksell, Uppsala. 255pp.

[8] Dagum, C. (1977). A new model of personal income distribution: specification and estimation. *Economie Appliquee 30*:413-436.

[9] Dagum, C. (1980). Inequality measures between income distributions with applications. *Econometrica 48*:1791-1803.

[10] Dagum, C. (1987). Measuring the economic affluence between populations of income receivers. *Journal of Business & Economic Statistic, 5*:5-12.

[11] Fellman, J. (1976). The effect of transformations on Lorenz curves. *Econometrica 44*:823-824.

[12] Fellman, J. (1980). *Transformations and Lorenz curves.* Swedish School of Economics and Business Administration Working Papers: 48, 18 pp.

[13] Fellman. J. (2012). Modelling Lorenz curves. *Journal of Statistical and Econometric Methods*, 1(3):53-62.

[14] Gibrat, (1931). *Les Inégalités Economiques*. Librairie du Recueil Sirey, Paris.

[15] Gini, C. (1914). Sulla misura della concentrazione e della variabilita dei caratteri. Atti del R. Instituto veneto 73c:1203-1248.

[16] 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 47*:171-184.

[17] Golden, J. (2008). A simple geometric approach to approximating the Gini coefficient. *J. Economic Education 39(1)*:68-77.

[18] Gradshteyn, I. S. & Ryshnik, I. M. (1965). *Tables of Integral Series and Products*. New York Academic press. 1086 pp.

[19] Gupta, M. R. (1984). Functional form for estimating the Lorenz curve. *Econometrica 52*:1313-1314.

[20] Harrison, A. (1981). Earnings by size: A tale of two distributions. *Review of Economic Studies 48*:621-631.

[21] Kakwani, N. (1980). On a Class of Poverty Measures *Econometrica 4*: 437-446.

[22] Kakwani N. C. & Podder N. (1973). On the Estimation of Lorenz Curves from Grouped Observations. *International Economic Review 14*: 278-292.

[23] Kakwani N. C. & Podder N. (1976). Efficient estimation of the Lorenz curve and the associated inequality measures from grouped observations. *Econometrica 44*:137-148.

[24] Kapteyn J. C. (1903). *Skew Frequency Curves in Biology and Statistics*. Vol. 1-2. 45 pp.

[25] Kleiber, Ch., Kotz, S. (2001). Characterizations of income distributions and the moment problem of order statistics. The 53rd Session of the International Statistical Institute in Seoul, Korea, Aug 22-29, Contributed Papers.

[26] Kleiber, Ch., Kotz, S. (2002). A characterization of income distributions in terms of generalized Gini coefficients. *Social Choice and Welfare 19*:789-794.

[27] Lee, W.-C. (1999). Probabilistic analysis of global performances of diagnostic tests: Interpreting the Lorenz curve based summary measures. *Statistics in Medicine 18*:455-471.

[28] Lorenz, M. O. (1905). Methods for measuring concentration of wealth. *J. Amer. Statist. Assoc. New Series, No. 70*: 209-219.

[29] Lydall, H. F. (1968). *The Structure of Earnings*. (London: Oxford University press).

[30] McDonald, J. B., Ransom, M. R. (1979). Functional Forms, estimation techniques and the distribution of income. *Econometrica 47*:1513-1525.

[31] Ogwang, T. & Rao, U. L. G, (2000). Hybrid models of the Lorenz curve, *Economics Letters, 69*:39-44.

[32] Pareto, V. (1897). *Cours d'Economie Politique*. Lausanne, Suisse.

[33] Quensel, C. E. (1944). *Inkomstfördelning och skattetryck*. Sveriges industriförbund, Lund.

[34] Rao, U. L. G. & Tam, A. Y.-P. (1987). An empirical study of selection and estimation of alternative models of the Lorenz curve. *J. of Applied Statistics 14*:275-280.

[35] Rasche, R. H., Gaffney, J., Koo A. Y. C. & Obst, N. (1980). Functional Forms for Estimating the Lorenz Curve. *Econometrica 48*:1061-1062.

[36] Rhodes, E. C. (1944). The Pareto distribution of incomes. *Economica New Series, XI 41*:1-11.

[37] Rohde, N. (2009). An alternative functional form for estimating the Lorenz curve. *Economics Letters 105*:61-63.

[38] Sen, A. (1973). *On Economic Inequality*. Clarendon Press, Oxford.

[39] Yitzhaki, S. (1983). On an extension of the Gini index. *International Economic Review 24*:617-628.