ISSN 0819-2642
ISBN 978 0 7340 4000 8

THE UNIVERSITY OF MELBOURNE

DEPARTMENT OF ECONOMICS

RESEARCH PAPER NUMBER 1034

February 2008

Estimating Income Distributions Using a Mixture of Gamma Densities

by

Duangkamon Chotikapanich & William E Griffiths

Department of Economics
The University of Melbourne
Melbourne Victoria 3010
Australia.

Estimating Income Distributions Using a Mixture of Gamma Densities

Duangkamon Chotikapanich
Monash University

William E Griffiths
University of Melbourne

11 February, 2008

Abstract

The estimation of income distributions is important for assessing income inequality and poverty and for making comparisons of inequality and poverty over time, countries and regions, as well as before and after changes in taxation and transfer policies. Distributions have been estimated both parametrically and non-parametrically. Parametric estimation is convenient because it facilitates subsequent inferences about inequality and poverty measures and lends itself to further analysis such as the combining of regional distributions into a national distribution. Non-parametric estimation makes inferences more difficult, but it does not place what are sometimes unreasonable restrictions on the nature of the distribution. By estimating a mixture of gamma distributions, in this paper we attempt to benefit from the advantages of parametric estimation without suffering the disadvantage of inflexibility. Using a sample of Canadian income data, we use Bayesian inference to estimate gamma mixtures with two and three components. We describe how to obtain a predictive density and distribution function for income and illustrate the flexibility of the mixture. Posterior densities for Lorenz curve ordinates and the Gini coefficient are obtained.

1. Introduction

The estimation of income distributions has played a major role in economic analysis. Information from such estimations is used to measure welfare, inequality and poverty, to assess changes in these measures over time, and to compare measures across countries, over time and before and after specific policy changes, designed, for example, to alleviate poverty. Typical inequality measures are the Gini coefficient and Atkinson's inequality measure. Measures of poverty are based on the proportion of population below a threshold or the expected value of a function over that part of the income distribution below a threshold. See, for example, Kakwani (1999). Estimates of these quantities and the Lorenz curve, a fundamental tool for measuring inequality, depend on the income distribution and how it is estimated. Thus, the estimation of income distributions is of central importance for assessing many aspects of the well being of society. A convenient reference for accessing the literature on the various dimensions of inequality measurement, and how they relate to welfare in society is Silber (1999).

A large number of alternative distributions have been suggested in the literature for estimating income distributions. See Kleiber and Kotz (2003) for a review of many of them, one of which is the Dagum distribution, whose inventor is being honoured by this volume. Further reviews of alternative income distributions appear elsewhere in this volume. After an income distribution model has been selected and estimated, probability distributions are used to draw inferences about inequality and poverty measures. These probability distributions can be sampling distributions for estimators of inequality and poverty, or Bayesian posterior distributions for inequality and poverty measures. In each case the required

probability distributions are derived from corresponding probability distributions for the parameters (or their estimators) of the assumed income distribution. This parametric approach to the analysis of income distributions can be applied to a sample of individuals, typically obtained via household surveys, or to more limited grouped data which may be the only form available. An advantage of the parametric approach is the ease with which probability distributions for inferences about inequality and poverty can be derived from those for the income distribution parameters. Also, in the case of more limited grouped data, the parametric approach gives a complete picture of the income distribution by allowing for within-group inequality. For an example of where the latter advantage is utilized, see Chotikapanich, Griffiths and Rao (2007) who estimated generalized beta distributions from grouped data.

Assuming a particular parametric distribution also has disadvantages. Inferences about inequality can depend critically on what distribution is chosen. This was evident in the work of Chotikapanich and Griffiths (2006) who found the posterior probabilities for Lorenz and stochastic dominance were sensitive to the choice of a Singh-Maddala or Dagum income distribution. To avoid the sensitivity of inferences to choice of income distribution, nonparametric approaches are frequently used. See Cowell (1999) and Barrett and Donald (2003) for examples of nonparametric sampling theory approaches and Hasegawa and Kozumi (2003) for a Bayesian approach.

One way of attempting to capture the advantages but not the disadvantages of a parametric specification of an income distribution is to use a functional form that is relatively flexible. This paper represents an attempt in this direction. Mixtures of distributions can provide flexible specifications and, under certain conditions, can

approximate a distribution of any form. With these characteristics in mind, we consider a mixture of gamma distributions; the gamma density is convenient one and it has been widely used for estimating income distributions. Our approach is Bayesian. Using data on before-tax income for Canada in 1978, taken from the Canadian Family Expenditure Survey and kindly provided by Gary Barrett, we find (i) posterior densities for the parameters of a gamma mixture, (ii) an estimate of the income distribution and 95% probability limits on the distribution, (iii) the posterior density for the Gini coefficient and (iv) an estimate of the Lorenz curve and 95% probability limits on this curve.

In Section 2 we specify the Gamma mixture and describe the Markov chain Monte Carlo algorithm (MCMC) for drawing observations from the posterior density for the parameters of the mixture. The data set and our selection of prior parameters is given in Section 3. Section 4 contains the results and a summary of the expressions used to obtain those results. Goodness-of-fit comparisons with other functional forms for the income distribution are given in Section 5. Some concluding remarks appear in Section 6.

2. Estimating the Gamma Mixture Model

An income distribution that follows a gamma mixture with k components can be written as

$$f(x \mid \mathbf{w}, \boldsymbol{\mu}, \mathbf{v}) = \sum_{z=1}^{k} w_z G(x \mid v_z, v_z/\mu_z) \tag{1}$$

where x is a random draw of income from the probability density function (pdf) $f(x \mid \mathbf{w}, \boldsymbol{\mu}, \mathbf{v})$, with parameter vectors, $\mathbf{w} = (w_1, w_2, \ldots, w_k)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)'$, and

$\mathbf{v} = (v_1, v_2, \ldots, v_k)'$. The pdf $G(x \mid v_z, v_z/\mu_z)$ is a gamma density with mean $\mu_z > 0$ and shape parameter $v_z > 0$. That is,

$$G(x \mid v_z, v_z/\mu_z) = \frac{(v_z/\mu_z)^{v_z}}{\Gamma(v_z)} x^{v_z - 1} \exp\left(-\frac{v_z}{\mu_z} x\right) \tag{2}$$

Including the mean μ_z as one of the parameters in the pdf makes the parameterization in (2) different from the standard textbook one, but it is convenient for later analysis. The parameter w_z is the probability that the *i*-th observation comes from the *z*-th component in the mixture. To define it explicitly, let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a random sample from (1), and let $Z_1, Z_2, \ldots, Z_n$ be indicator variables such that $Z_i = z$ when the *i*-th observation comes from the *z*-th component in the mixture. Then,

$$P(Z_i = z \mid \mathbf{w}) = w_z \qquad \text{for } z = 1, 2, \ldots, k$$

with $w_z > 0$ and $\sum_{z=1}^{k} w_z = 1$. Also, conditional on $Z_i = z$, the distribution of x_i is $G\left(v_z, v_z/\mu_z\right)$.

To use Bayesian inference, we specify prior distributions on the unknown parameters $\mathbf{w}$, $\boldsymbol{\mu}$, and $\mathbf{v}$, and then combine these pdfs with the likelihood function defined by (1) to obtain a joint posterior pdf for the unknown parameters. This joint posterior pdf represents our post-sample knowledge about the parameters and is the source of inferences about them. However, as is typically the case in Bayesian inference, the joint posterior pdf is analytically intractable. This problem is solved by using MCMC techniques to draw observations from the joint posterior pdf and using these draws to estimate the quantities required for inference. Because we are interested in not just the parameters, but also the income distribution, the Gini

coefficient, and the Lorenz curve, the parameter draws are also used in further analysis to estimate posterior information about these quantities.

The MCMC algorithm used to draw observations from the posterior density for $(\boldsymbol{\mu}, \mathbf{v}, \mathbf{w})$ is taken from Wiper, Rios Insua and Ruggeri (2001). In the context of other problems, Wiper *et al*. consider estimation for both a known and an unknown *k*. We will assume a known value of *k* that is specified *a priori*. In our empirical work we considered $k = 3$ and $k = 2$ but settled on $k = 2$ as an adequate formulation. The MCMC algorithm is a Gibbs' sampling one where draws are taken sequentially and iteratively from the conditional posterior pdfs for each of the parameters. Because only the conditional posterior pdfs are involved in this process, it is not necessary to specify the complete joint posterior pdf. The relevant conditional posterior pdfs are sufficient; they are specified below after we introduce the prior pdfs.

Following Wiper *et al*. (2001), the prior distributions used for each of the parameters are

$$f(\mathbf{w}) = D(\boldsymbol{\varphi}) \propto w_1^{\phi_1 - 1} w_2^{\phi_2 - 1} \ldots w_k^{\phi_k - 1} \qquad \text{(Dirichlet)} \qquad (3)$$

$$f(v_z) \propto \exp\{-\theta v_z\} \qquad \text{(exponential)} \qquad (4)$$

$$f(\mu_z) = GI(\alpha_z, \beta_z) \propto \mu_z^{-(\alpha_z + 1)} \exp\left\{-\frac{\beta_z}{\mu_z}\right\} \qquad \text{(inverted gamma)} \qquad (5)$$

$$\text{for } z = 1, 2, \ldots, k$$

The Dirichlet distribution is the same as a beta distribution for $k = 2$ and a multivariate extension of the beta distribution for $k > 2$. Its parameters are $\boldsymbol{\varphi} = (\phi_1, \phi_2, \ldots, \phi_k)'$. To appreciate the relationship between the gamma and inverted

gamma pdfs, note that if $y \sim G(\alpha,\beta)$, then $q = (1/y) \sim GI(\alpha,\beta)$. The pdfs in (3), (4) and (5) are chosen because they combine nicely with the likelihood function for derivation of the conditional posterior pdfs, and because they are sufficiently flexible to represent vague prior information which can be dominated by the sample data. In addition to the above prior pdfs, the restriction $\mu_1 < \mu_2 < \cdots < \mu_k$ is imposed *a priori* to ensure identifiability of the posterior distribution. Settings for the prior parameters $(\boldsymbol{\varphi}, \theta, \alpha_z, \beta_z)$ are discussed in Section 3.

After completing the algebra necessary to combine the prior pdfs with the likelihood function in such a way that isolates the conditional posterior densities for use in a Gibbs' sampler, we obtain the following conditional posterior pdfs.

The posterior probability that the *i*-th observation comes from the *z*-th component in the mixture, conditional on the unknown parameters, is the discrete pdf

$$P(Z_i = z \mid \mathbf{x}, \mathbf{w}, \mathbf{v}, \boldsymbol{\mu}) = \frac{p_{iz}}{p_{i1} + p_{i2} + \cdots + p_{ik}} \tag{6}$$

where

$$p_{iz} = w_z \frac{(v_z/\mu_z)^{v_z}}{\Gamma(v_z)} x_i^{v_z - 1} \exp\left\{-\frac{v_z x_i}{\mu_z}\right\}$$

The posterior pdf for the mixture-component probabilities $\mathbf{w}$, conditional on the other parameters and on the realized components for each observation $\mathbf{z} = (z_1, z_2, \ldots, z_n)'$, is the Dirichlet pdf

$$f(\mathbf{w} \mid \mathbf{x}, \mathbf{z}, \mathbf{v}, \boldsymbol{\mu}) = D(\boldsymbol{\varphi} + \mathbf{n}) \tag{7}$$

where $\mathbf{n} = (n_1, n_2, \ldots, n_k)'$, with n_z being the number of observations for which $Z_i = z$. Thus, $\sum_{z=1}^{k} n_z = n$.

The posterior pdfs for the means of the component densities μ_z, conditional on the other parameters and on $\mathbf{z}$, are the inverted gamma pdfs

$$f(\mu_z \mid \mathbf{x}, \mathbf{z}, \mathbf{w}, \mathbf{v}) = GI(\alpha_z + n_z v_z,\ \beta_z + S_z v_z) \tag{8}$$

where $S_z = \sum_{i:Z_i = z} x_i$.

The form of the posterior pdfs for the scale parameters of the component densities v_k, conditional on the other parameters and on $\mathbf{z}$, is not a common recognizable one. It is given by

$$f(v_z \mid \mathbf{x}, \mathbf{z}, \mathbf{w}, \boldsymbol{\mu}) \propto \frac{v_z^{n_z v_z}}{\left[\Gamma(v_z)\right]^{n_z}} \exp\left\{-v_z\left(\theta + \frac{S_z}{\mu_z} + n_z \log \mu_z - \log P_z\right)\right\} \tag{9}$$

where $P_z = \prod_{i:Z_i = z} x_i$.

A Gibbs sampling algorithm that iterates sequentially and iteratively through the conditional posterior pdfs can proceed as follows:

1. Set $t = 0$ and initial values $\mathbf{w}^{(0)}, \boldsymbol{\mu}^{(0)}, \mathbf{v}^{(0)}$.
2. Generate $\left(\mathbf{z}^{(t+1)} \mid \mathbf{x}, \mathbf{w}^{(t)}, \mathbf{v}^{(t)}, \boldsymbol{\mu}^{(t)}\right)$ from (6).
3. Generate $\left(\mathbf{w}^{(t+1)} \mid \mathbf{x}, \mathbf{z}^{(t+1)}, \mathbf{v}^{(t)}, \boldsymbol{\mu}^{(t)}\right)$ from (7).
4. Generate $\left(\mu_z^{(t+1)} \mid \mathbf{x}, \mathbf{z}^{(t+1)}, \mathbf{v}^{(t)}, \mathbf{w}^{(t+1)}\right)$ from (8), for $z = 1, 2, \ldots, k$.

5. Generate $\left(v_z^{(t+1)} \mid \mathbf{x},\ \mathbf{z}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}, \mathbf{w}^{(t+1)}\right)$ from (9), for $z = 1, 2, \ldots, k$.

6. Order the elements for $\boldsymbol{\mu}^{(t+1)}$ such that $\mu_1 < \mu_2 < \ldots < \mu_k$ and sort $\mathbf{w}^{(t+1)}$ and $\mathbf{v}^{(t+1)}$ accordingly.

7. Set $t = t + 1$ and return to step 2.

To describe each of the generation steps in more detail, first consider (6). In this case we divide the interval (0,1) into *k* sub-intervals with the length of the *z*-th sub-interval equal to $P\left(Z_i = z \mid \mathbf{x}, \mathbf{w}, \mathbf{v}, \boldsymbol{\mu}\right)$. A uniform random number is generated from the (0,1) interval. The value assigned to Z_i is the sub-interval in which the uniform random number falls. To generate observations from the Dirichlet density in (7), we first generate *k* gamma random variables, say $\gamma_z,\ z = 1, 2, \ldots, k$ from $G(\phi_z + n_z, 1)$ densities, and then set $w_z = \gamma_z \Big/ \sum_{j=1}^{k} \gamma_j$. To generate μ_z from (8), we generate a random variable from a $G\left(\alpha_z + n_z v_z,\ \beta_z + S_z v_z\right)$ density and then invert it.

Generating v_z from equation (9) is more complicated, requiring a Metropolis step. We draw a candidate $\tilde{v}_z^{(t+1)}$ from a gamma density with mean equal to the previous draw $v_z^{(t)}$. That is, a candidate $\tilde{v}_z^{(t+1)}$ is generated from a $G\left(r,\ r/v_z^{(t)}\right)$ distribution and is accepted as $v_z^{(t+1)}$ with probability

$$\min\left\{1, \frac{f\left(\tilde{v}_z^{(t+1)} \mid \mathbf{x}, \mathbf{z}^{(t+1)}, \mathbf{w}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}\right)}{f\left(v_z^{(t)} \mid \mathbf{x}, \mathbf{z}^{(t+1)}, \mathbf{w}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}\right)} \frac{p\left(\tilde{v}_z^{(t+1)}, v_z^{(t)}\right)}{p\left(v_z^{(t)}, \tilde{v}_z^{(t+1)}\right)}\right\}$$

where $p\left(v_z^{(t)}, \tilde{v}_z^{(t+1)}\right)$ is the gamma density used to generate $\tilde{v}_z^{(t+1)}$. Non-acceptance of $\tilde{v}_z^{(t+1)}$ implies $v_z^{(t+1)} = v_z^{(t)}$. The value of r is chosen by experimentation to give an acceptance rate of approximately 0.4.

3. Data Characteristics and Prior Parameters.

Characteristics of the sample of incomes from the 1978 Canadian Family Expenditure Survey are presented in Figure 1. The units are thousands of Canadian dollars. There are 8526 observations with values ranging from 0.281 to 173.8. Sample mean income is 35.5 and the sample median income is 32.4. The histogram reveals two modes, one at approximately 23 and the other at approximately 32. The Gini coefficient computed from the sample is 0.3358.

[Insert Figure 1 near here]

In choosing values for the parameters of the prior densities, our objective was to have proper but relatively uninformative priors so that posterior densities would be dominated by the sample data. We initially tried a mixture of $k = 3$ components but encountered identification problems and then reduced the number of components to $k = 2$.

We set $\phi_z = 1$ for all *z*, thus implying a uniform prior for the weights on each component. For the exponential prior on the scale parameters v_z we set $\theta = 0.02$. A 95% probability interval for this prior is (0.5, 161) implying a large range of values are possible. For the μ_z we initially set $\alpha_z = 2.2$ for $z = 1, 2, 3$ and $\beta_1 = 24$, $\beta_2 = 54$, $\beta_3 = 120$. Then, when we proceeded with $k = 2$, we set $\beta_1 = 30$ and $\beta_2 = 95$. From this latter setting, and ignoring the truncation $\mu_1 < \mu_2$, 95% prior probability intervals

for μ_1 and μ_2 are, respectively, (5, 98) and (16, 306). In light of the sample mean of 35.5, these intervals suggest priors that are relatively uninformative.

4. Results

The algorithm described in Section 2 was used to generate 200,000 observations from the joint posterior density for the parameters $(\mathbf{w}, \boldsymbol{\mu}, \mathbf{v})$ and the first 100,000 were discarded as a burn in. In our first attempts with $k = 3$ there appeared to be an identification problem with the second and third components. For separate identification of these two components, we require $\mu_2 < \mu_3$. If $\mu_2 = \mu_3$, some other mechanism is required for identification (Wiper *et al.* 2001). The two-dimensional plot of the draws for μ_2 and μ_3 given in Figure 2 shows a large number of observations on the boundary where $\mu_2 = \mu_3$. Other evidence is the bimodal distributions for v_2 and v_3 (Figure 3), the very high correlation between w_2 and w_3 (Figure 4) and the fact that the marginal posterior densities for w_2 and w_3 were mirror images of each other.

[Insert Figures 2, 3 and 4 near here]

These issues led us to consider instead a model with two components ($k = 2$). In this case there was no apparent identification problem, and the Gibbs sampler showed evidence of converging. Summary statistics for the draws on the parameters are given in Table 1. There is relatively large weight (about 0.9) on the second component and a relatively small weight (about 0.1) on the first component. The posterior mean for the mean of the first component is relatively small (compared to

the sample mean) and, likely, serves to help capture the first mode of the income distribution.

Table 1 Posterior Summary Statistics for Parameters

Name	Mean	St.Dev	Min	Max
μ_1	9.6134	0.35688	7.8906	11.130
μ_2	38.704	0.42069	36.903	40.768
w_1	0.10896	0.012091	0.06080	0.15656
w_2	0.89104	0.012091	0.84344	0.93920
v_1	7.4761	0.80874	5.2314	12.653
v_2	3.3616	0.11977	2.9667	3.9985

Having obtained *M* MCMC-generated observations from the posterior density $f(\mathbf{w},\boldsymbol{\mu},\mathbf{v}\,|\,\mathbf{x})$, for a sample of observations $\mathbf{x}$ we can proceed to obtain estimates for the density and distribution functions for income and for the corresponding Lorenz curve as well as probability bands around these functions. Indexing an MCMC-generated observation by a superscript (*j*), an estimate for the density function at a given income *x* is given by

$$f(x\,|\,\mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M}\sum_{z=1}^{k} w_z^{(j)} G\left(x\,|\,v_z^{(j)}, v_z^{(j)}/\mu_z^{(j)}\right) \tag{10}$$

This function was calculated for 101 values of *x* from 0 to 200 such that the intervals between successive values of log *x* were equal. For each *x* 95% probability bands were found by sorting the *M* values of

$$f(x\,|\,\mathbf{w}^{(j)},\boldsymbol{\mu}^{(j)},\mathbf{v}^{(j)}) = \sum_{z=1}^{k} w_z^{(j)} G\left(x\,|\,v_z^{(j)}, v_z^{(j)}/\mu_z^{(j)}\right)$$

and taking the 0.025 and 0.975 percentiles of these values. The plots for the mean distribution and its probability bounds appear in Figure 5. The bimodal nature of the

distribution has been well captured, although, as one would expect, it is at the peaks of the distribution where the greatest uncertainty is exhibited through wider bounds.

[Insert Figure 5 near here]

An estimate of the distribution function and probability bounds on that distribution can be found in a similar way. In this case the value of the distribution function for a given value *x* is given by

$$\begin{aligned} F(x \mid \mathbf{x}) &= \frac{1}{M}\sum_{j=1}^{M}\sum_{z=1}^{k} w_z^{(j)} \int_0^x G\left(t \mid v_z^{(j)}, v_z^{(j)} / \mu_z^{(j)}\right) dt \\ &= \frac{1}{M}\sum_{j=1}^{M} F(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}) \end{aligned} \tag{11}$$

This function was evaluated for the same 101 values of *x*. To estimate the Lorenz curve we consider for each *x* the *M* points $F\left(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}\right)$ and the corresponding points for the first moment distribution which is given by

$$\begin{aligned} \eta\left(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}\right) &= \frac{1}{\mu^{(j)}} \int_0^x t\, f\left(t \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}\right) dt \\ &= \frac{1}{\mu^{(j)}} \sum_{z=1}^{k} w_z^{(j)} \int_0^x t\, G\left(t \mid v_z^{(j)}, v_z^{(j)} / \mu_z^{(j)}\right) dt \\ &= \frac{1}{\mu^{(j)}} \sum_{z=1}^{k} w_z^{(j)} \mu_z^{(j)} \int_0^x G\left(t \mid \left(v_z^{(j)} + 1\right), v_z^{(j)} / \mu_z^{(j)}\right) dt \end{aligned} \tag{12}$$

where $\mu^{(j)} = \sum_{z=1}^{k} w_z^{(j)} \mu_z^{(j)}$.

To see how to use these points to estimate a Lorenz curve and find its probability bounds it is instructive to examine a graph of the *M* points for $F(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)})$ and $\eta\left(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}\right)$ for a given value of *x*. Such a graph for

the point $x = 35$ is given in Figure 6. A graph like that in Figure 6 could be drawn for each of the 101 *x* points. To estimate the Lorenz curve and draw probability bounds around it, we need to "select" three points from each graph, an estimate of the Lorenz curve for each *x* and its corresponding upper and lower probability bounds. As an estimate of the Lorenz curve for a given *x* we can take the mean values of all the points in Figure 6. That is, the point $[\eta(x \mid \mathbf{x}), F(x \mid \mathbf{x})]$ where

$$\eta(x \mid \mathbf{x}) = \frac{1}{M}\sum_{j=1}^{M} \eta\left(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}\right) \tag{13}$$

and $F(x \mid \mathbf{x})$ is given in (11). Then an estimate of the complete Lorenz curve is obtained by joining these points for all *x*.

[Insert Figure 6 near here]

Finding 95% probability bounds for the Lorenz curve is more difficult than it is for the density and distribution functions because, for each *x*, we have a 2-dimensional space for $F(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)})$ and $\eta\left(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}\right)$ to consider. Two approaches were taken. In the first, for each *x*, we regressed the *M* values of $\eta\left(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)}\right)$ on the corresponding *M* values of $F(x \mid \mathbf{w}^{(j)}, \boldsymbol{\mu}^{(j)}, \mathbf{v}^{(j)})$ via a least squares regression. The residuals from this regression were ordered and the 0.025 and 0.975 percentiles of the residuals were noted. Denoting them by $\hat{e}_{.025}$ and $\hat{e}_{.975}$, the bounds at a given *x* were taken as the points

$$\left[F(x \mid \mathbf{x}),\ \eta(x \mid \mathbf{x}) + \hat{e}_{.025}\right] \quad \text{and} \quad \left[F(x \mid \mathbf{x}),\ \eta(x \mid \mathbf{x}) + \hat{e}_{.925}\right] \tag{14}$$

Note that $\hat{e}_{.025} < 0$, so we add it rather than subtract it from $\eta(x \mid \mathbf{x})$. To obtain the lower bound on the Lorenz curve, we computed $\left[F(x \mid \mathbf{x}),\ \eta(x \mid \mathbf{x}) + \hat{e}_{.025}\right]$ for each *x*,

and joined these points. Similarly, to obtain the upper Lorenz bound, we computed $\left[F(x\,|\,\mathbf{x}),\ \eta(x\,|\,\mathbf{x})+\hat{e}_{.925}\right]$ for each x and joined these points. These bounds and the estimated Lorenz curve are plotted in Figure 7. However, the bounds are so narrow that they are indistinguishable from the estimated curve. In Figure 8 we present a more distinct cross section of the plots for $0.4<F(x)<0.6$ and $0.2<\eta(x)<0.4$. Also, to give an idea of the width of the bounds, in Figure 9 we plot $\hat{e}_{.025}$ and $\hat{e}_{.975}$ against $F(x)$. The maximum width of the probability interval is less than 0.008, implying the Lorenz curve is accurately estimated.

[Insert Figures 7, 8 and 9 near here]

To introduce our second approach for finding probability bounds on the Lorenz curve, first note that, in the first approach, the bounds do not correspond to one set of parameter values for all x. The upper and lower extreme 2.5% of parameter values is likely to be different for each x setting. While this is not necessarily a bad thing – it is also a characteristic of the estimated density function for income – it is interesting to examine an alternative method of obtaining bounds that "discards" the same parameter values for each x. One way to use a unique set of upper and lower 2.5% of parameter values is to order Lorenz curves on the basis of their Gini coefficients. Denoting the 101 x points as $x_1, x_2, \ldots, x_{101}$, the Gini coefficient for the j-th set of parameters can be approximated by

$$\begin{aligned} Gini\left(\mathbf{w}^{(j)},\boldsymbol{\mu}^{(j)},\mathbf{v}^{(j)}\right) = &\sum_{m=1}^{100}\eta\left(x_{m+1}\,|\,\mathbf{w}^{(j)},\boldsymbol{\mu}^{(j)},\mathbf{v}^{(j)}\right)F\left(x_m\,|\,\mathbf{w}^{(j)},\boldsymbol{\mu}^{(j)},\mathbf{v}^{(j)}\right) \\ &-\sum_{m=1}^{100}\eta\left(x_m\,|\,\mathbf{w}^{(j)},\boldsymbol{\mu}^{(j)},\mathbf{v}^{(j)}\right)F\left(x_{m+1}\,|\,\mathbf{w}^{(j)},\boldsymbol{\mu}^{(j)},\mathbf{v}^{(j)}\right) \end{aligned} \tag{15}$$

In this approach the probability bounds of the Lorenz curve were taken as the Lorenz curves corresponding to the parameter values that yield the 0.025 and 0.975 percentiles for the Gini coefficient. Thus, the bounds on the Lorenz curve are found by using the area under the Lorenz curve to determine a parameter ordering. Specifically, if the parameter values corresponding to the 0.025 and 0.975 percentiles of the Gini coefficient are $(\mathbf{w}_{.025}, \boldsymbol{\mu}_{.025}, \mathbf{v}_{.025})$ and $(\mathbf{w}_{.975}, \boldsymbol{\mu}_{.975}, \mathbf{v}_{.975})$, then the upper bound is the curve joining the points $[\eta(x \mid \mathbf{w}_{.975}, \boldsymbol{\mu}_{.975}, \mathbf{v}_{.975}), F(x \mid \mathbf{w}_{.975}, \boldsymbol{\mu}_{.975}, \mathbf{v}_{.975})]$ for each *x*, and the lower bound is the curve joining the points $[\eta(x \mid \mathbf{w}_{.025}, \boldsymbol{\mu}_{.025}, \mathbf{v}_{.025}),$ $F(x \mid \mathbf{w}_{.025}, \boldsymbol{\mu}_{.025}, \mathbf{v}_{.025})]$ for each *x*.

While it is straightforward to draw the bounds in this way, it is not obvious how one might define the "errors" between the estimated Lorenz curve and its 95% probability bounds if one is interested in these values. In the regression approach, where η was treated as the "dependent" variable and *F* was treated as the "explanatory" variable, it was natural to define the errors as the vertical distances as specified in (14). In this case, however, there is no reason why they should be vertical or horizontal distances. To solve this dilemma, we define the errors as the orthogonal distances from the Lorenz curve

$$\hat{d}_U(x) = \sqrt{\left(F(x \mid \mathbf{w}_{.975}, \boldsymbol{\mu}_{.975}, \mathbf{v}_{.975}) - F(x \mid \mathbf{x})\right)^2 + \left(\eta(x \mid \mathbf{w}_{.975}, \boldsymbol{\mu}_{.975}, \mathbf{v}_{.975}) - \eta(x \mid \mathbf{x})\right)^2}$$

$$\hat{d}_L(x) = \sqrt{\left(F(x \mid \mathbf{w}_{.025}, \boldsymbol{\mu}_{.025}, \mathbf{v}_{.025}) - F(x \mid \mathbf{x})\right)^2 + \left(\eta(x \mid \mathbf{w}_{.025}, \boldsymbol{\mu}_{.025}, \mathbf{v}_{.025}) - \eta(x \mid \mathbf{x})\right)^2}$$

Once again, it turned out that the Lorenz curve is estimated very accurately with the probability bounds not discernible from the mean Lorenz curve. Rather than present another figure that appears identical to Figure 7, in this case we simply plot

the errors $\hat{d}_U(x)$ and $\hat{d}_L(x)$ that appear in Figure 10. The pattern of these differences is a strange one, and, as expected, they are larger than those obtained using the regression method. Larger differences are expected because the regression method minimizes the "error" for each *x*. Nevertheless, the largest error is still relatively small, being less than 0.016.

[Insert Figure 10 near here]

Also, of interest is the Gini coefficient. Its posterior density, estimated from the 100,000 points defined by equation (15), is plotted in Figure 11. The posterior mean is 0.337 and 95% probability bounds for the Gini coefficient are 0.333 and 0.342.

[Insert Figure 11 near here]

5. Goodness of Fit

Given our objective was to specify a gamma mixture as a flexible parametric model for an income distribution, it is useful to assess its goodness of fit against those of some common income distributions. To do so we compare the estimated distribution function $F(x \mid \mathbf{x})$ with the empirical distribution function $F_0(x_j) = j/n$ where *j* refers to the *j*-th observation after ordering them from lowest to highest and *n* is the sample size. We compute goodness of fit using the root mean squared error

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(F\left(x_j \mid \mathbf{x}\right) - F_0(x_j)\right)^2}$$

In addition we perform a Kolomogorov-Smirnov test which is based on the largest difference between $F\left(x_j \mid \mathbf{x}\right)$ and $F_0(x_j)$. Table 2 contains the results for the

Bayesian-estimated gamma mixture and for maximum likelihood estimates of the log-normal, beta2, Singh-Maddala and Dagum distributions. Clearly, the gamma mixture is far superior to other models in terms of goodness of fit.

Table 2: Goodness of Fit Comparisons

	RMSE	Max Dif (δ_n)	$\delta_n\sqrt{n}$	*p*-value
Gamma Mix	0.0064	0.01449	1.33795	0.055738
LogNormal	0.0414	0.07449	6.87813	0.000000
Beta2	0.0310	0.05523	5.09974	0.000000
Singh-Maddala	0.0122	0.02757	2.54571	0.000005
Dagum	0.0135	0.03146	2.90490	0.000000

6. Concluding Remarks

A mixture of gamma densities has been suggested as a model for income distributions. Mixtures have the advantage of providing a relatively flexible functional form and at the same time they retain the advantages of parametric forms that are amenable to inference. We have demonstrated how a Bayesian framework can be utilized to estimate the gamma mixture and related quantities relevant for income distributions. In addition to showing how the income distribution estimate and its 95% probability bounds can be calculated, we considered the distribution function, the Lorenz curve and the Gini coefficient. Two ways of computing 95% probability bounds for the Lorenz curve were explored. Goodness-of-fit comparisons showed the gamma mixture fits well compared to a number of commonly used income distributions.

An attempt to estimate a mixture with 3 components was not successful leading us to opt for a model with 2 components. The results for 3 components

suggested a lack of identification between the second and third components. Most likely, the empirical characteristics of the distribution are well captured by 2 components, making it hard for the data to discriminate when 3 are specified. This outcome does not necessarily imply 2 will always be adequate. There could be other distributions where more components improve the specification. Also, the number of components can be treated as an unknown parameter which, in a Gibbs sampling algorithm, can vary from iteration to iteration.

Further research will focus on the use of estimated gamma mixtures in the measurement of inequality and poverty and in methodology for examining stochastic and Lorenz dominance for income distributions. Expressing uncertainty about such quantities in terms of posterior densities facilitates making inferences and probability statements about relative welfare scenarios.

References

Barrett, G.F. and S.G. Donald (2003), “Consistent Tests for Stochastic Dominance”, *Econometrica*, 71(1), 71-104.

Chotikapanich, D. and W. E. Griffiths (2006), "Bayesian Assessment of Lorenz and Stochastic Dominance in Income Distributions", in J. Creedy and G. Kalb, editors, *Research on Economic Inequality***,** Volume 13: *Dynamics of Inequality and Poverty*, Amsterdam: Elsevier, 297-321.

Chotikapanich, D., W.E. Griffiths, and D.S.P. Rao (2007), “Estimating and Combining National Income Distributions using Limited Data”, *Journal of Business and Economic Statistics*, 25, 97-109

Cowell, F.A. (1999), "Estimation of Inequality Indices", in J. Silber, editor, *Handbook on Income Inequality Measurement*, London: Kluwer Academic Publishers.

Hasegawa, H and H. Kozumi (2003), "Estimation of Lorenz Curves: A Bayesian Nonparametric Approach", *Journal of Econometrics*, 115, 277-291.

Kakwani, N. (1999), "Inequality, Welfare and Poverty: Three Interrelated Phenomena", in J. Silber, ed., *Handbook on Income Inequality Measurement*, London: Kluwer Academic Publishers.

Kleiber C. and S. Kotz (2003), *Statistical Size Distributions in Economics and Acturial Sciences*, New Jersey: Wiley & Sons.

Silber, J. (1999), *Handbook on Income Inequality Measurement*, London: Kluwer Academic Publishers.

Wiper, M., D. Rios Insua and F. Ruggeri (2001), “Mixtures of Gamma Distributions with Applications”, *Journal of Computational and Graphical Statistics*, 10, 440-454.

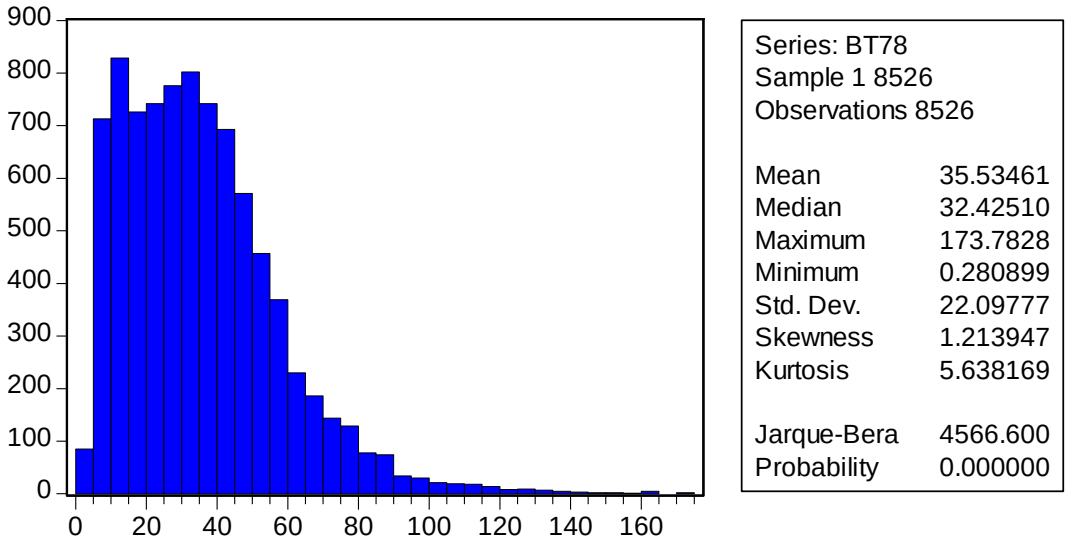

Figure 1: Characteristics of Canadian income data

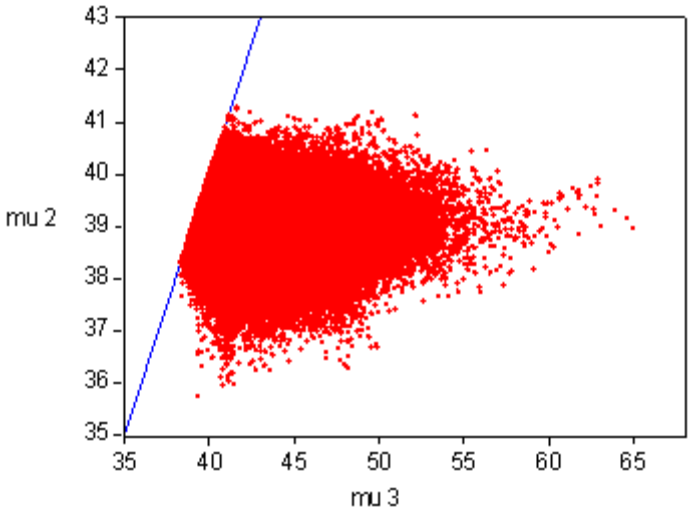

Figure 2: Posterior observations on μ_2 and μ_3 for $k = 3$

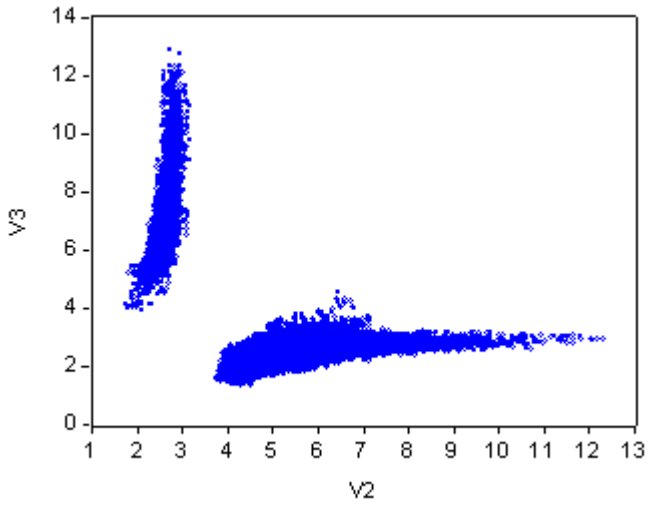

Figure 3: Posterior observations on v_2 and v_3 for $k = 3$

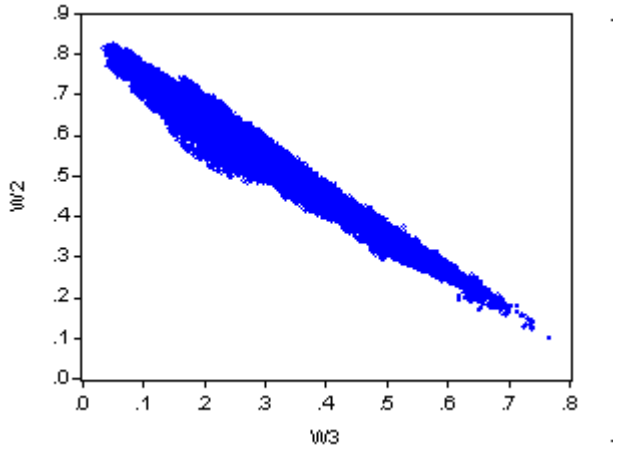

Figure 4**:** Posterior observations on w_2 and w_3 for $k = 3$

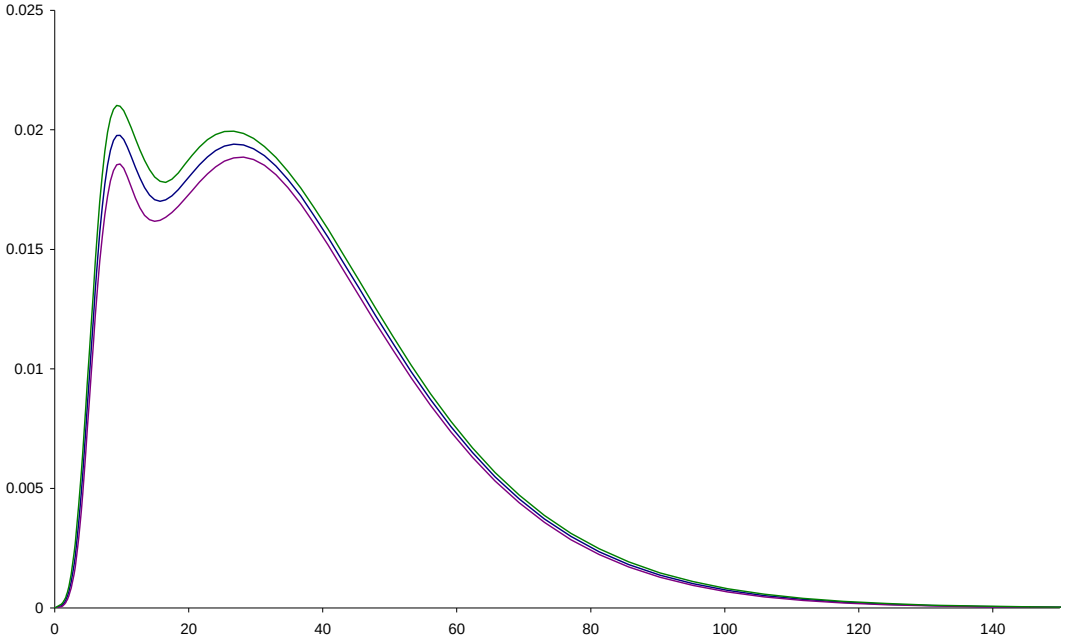

Figure 5: Mean and 95% probability bounds for the predictive density for income

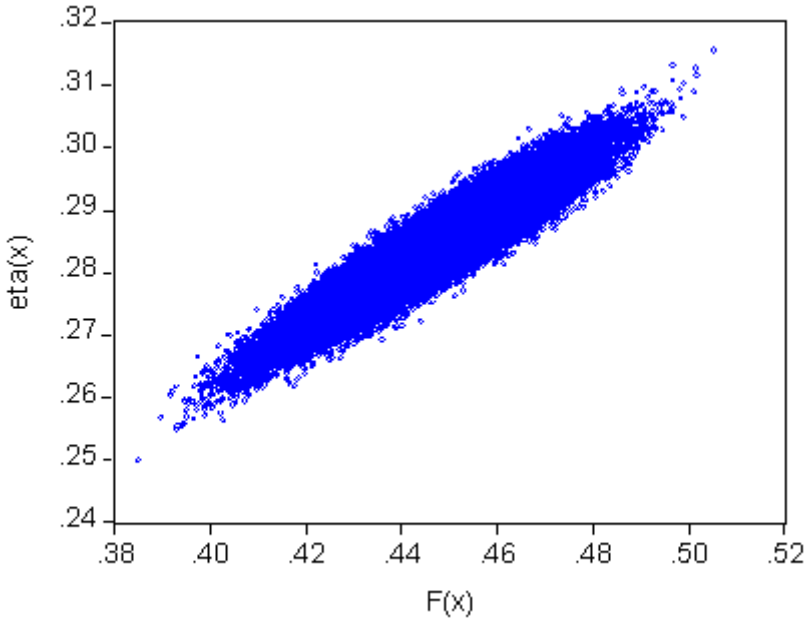

Figure 6: Plots of 100,000 pairs of $F(x)$ and $\eta(x)$ for $x = 35$

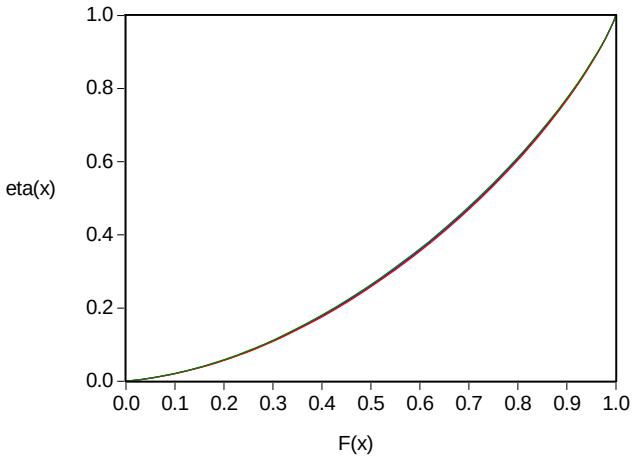

Figure 7: Entire Lorenz curve and 95% probability bounds

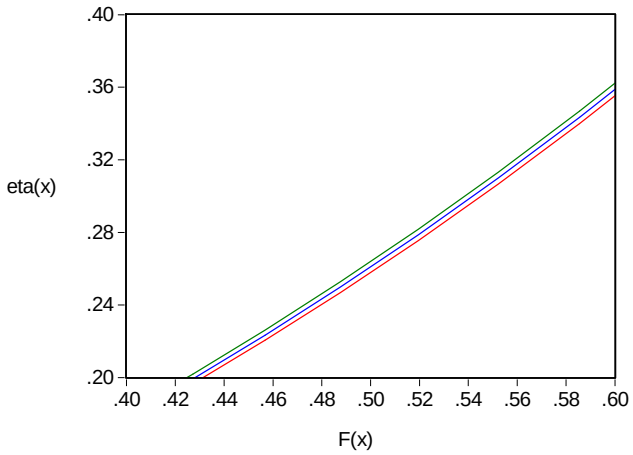

Figure 8: Close up of Lorenz curve and 95% probability bounds

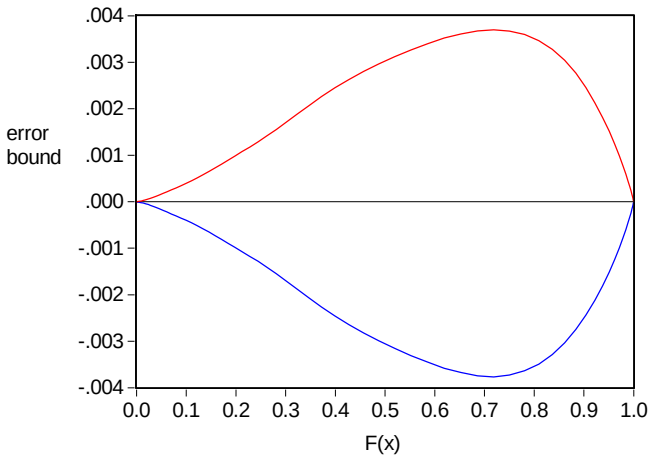

Figure 9: Plots of the differences between
the estimated Lorenz curve and 95% probability bounds

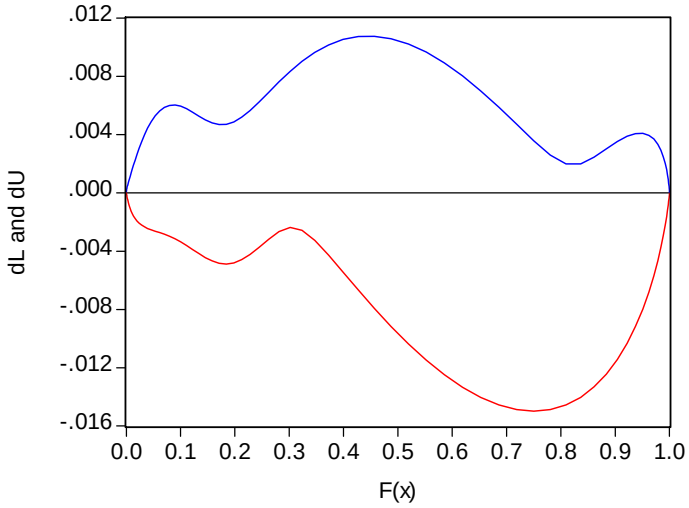

Figure 10: Orthogonal differences between Lorenz curve
and 95% probability bounds.

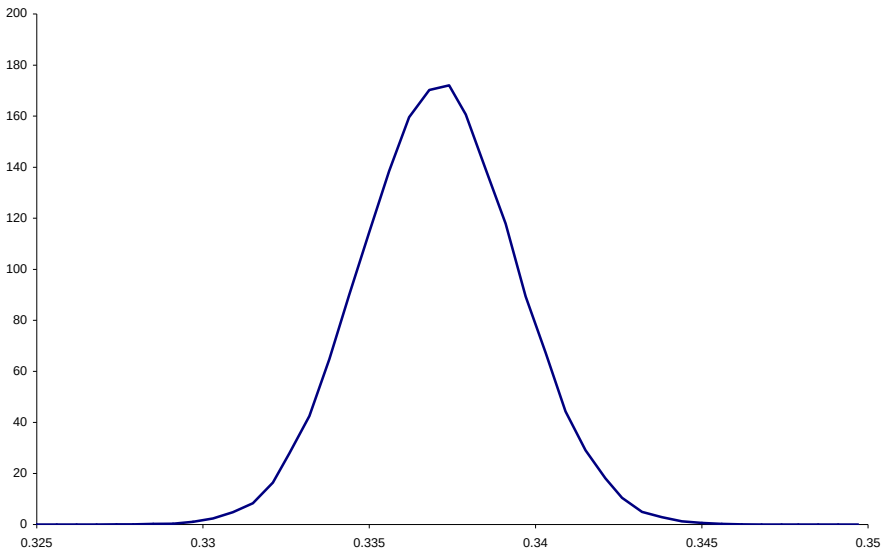

Figure 11: Posterior density for the Gini coefficient.