ABSTRACT

Title of thesis: MODELING MEDIAN HOUSEHOLD INCOME DISTRIBUTION

Kevin M. Foster, Master of Arts, 2009

Thesis directed by: Professor Benjamin Kedem
Department of Mathematics

In this thesis we are going to use U.S. Census data to study median household income distribution for 13 U.S. counties and seven U.S. states. Over the years, researchers have fitted income data with various probability distributions. During our review of the literature, we saw that researchers do not agree on any one best distribution.

We will be looking at lognormal, gamma and Weibull, each of which has two parameters. We will also investigate the Singh-Maddala, which has three parameters. Finally, we will introduce the Generalized Beta II, which has four parameters.

These distributions will be tested using Mean Squared Error, Mean Absolute Error, Chi-square Goodness-Of-Fit, Akaike's Information Criterion and Bayesian Information Criterion. We also use the graphical technique of QQ Plots.

We discover that the Singh-Maddala most often provides the best fit model for our income data, and we make the recommendation that users choose the Singh-Maddala distribution as their model when studying median household income distribution.

MODELING MEDIAN HOUSEHOLD INCOME DISTRIBUTION

by

Kevin M. Foster

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Arts
2009

Advisory Committee:
Professor Benjamin Kedem, Chair/Advisor
Professor Eric Slud
Professor Paul Smith

Dedication

For my dear Maria.

Acknowledgments

I want to give thanks to the following people who helped me with along the way towards my degree.

First, thanks go to my advisor Professor Benjamin Kedem. His suggestions and ideas were always helpful. I consider Professor Kedem to be my teacher, and for that I am thankful. I am very grateful for his patience with me while finishing this degree.

Thanks are also due to Professor Eric Slud and Professor Paul Smith for agreeing to serve on my thesis committee and for sparing their invaluable time reviewing my thesis.

I must also mention my supervisors at the U.S. Census Bureau for their patience and understanding while I've been pursuing this degree. Steve Tourkin, Amy Newman-Smith, Nancy Swaim, and Bonnie Kegan deserve special thanks for allowing me to work part-time, change my work schedule around at a moment's notice, and to leave work early on occasion. Thank you all.

Special thanks go to Maria for all of her patience and encouragement along the way.

Lastly, I must acknowledge the Manufacturing & Construction Division of the U.S. Census Bureau for providing financial support during the final few semesters of my degree.

Table of Contents

List of Tables v

List of Figures vi

1 Introduction 1
- 1.1 Survey . . . 1
  - 1.1.1 Pareto . . . 2
  - 1.1.2 The Mid 20th Century . . . 4
- 1.2 An Example Of The Problems With Fitting Income Distributions . . . 5
  - 1.2.1 Weak Pareto Law . . . 6
  - 1.2.2 Majumder and Chakravarty . . . 7
  - 1.2.3 Parametric Models . . . 9
- 1.3 The Data . . . 9
  - 1.3.1 Assumptions About the Data . . . 12
- 1.4 The Counties and States . . . 13

2 Four Densities 22
- 2.1 The Lognormal Density . . . 23
  - 2.1.1 Choosing Between X and $\log X$ . . . 24
- 2.2 The Gamma Density . . . 28
- 2.3 The Weibull Density . . . 32
- 2.4 Distributions With More Than Two Parameters . . . 33
  - 2.4.1 Singh-Maddala . . . 34
  - 2.4.2 Generalized Beta Of The Second Kind . . . 35
  - 2.4.3 Relationship to other three distributions . . . 35
- 2.5 Estimating The Parameters . . . 36
- 2.6 QQ Plots . . . 41
- 2.7 Error Measures . . . 41
- 2.8 Measures of Goodness Of Fit and Distribution Selection . . . 46
  - 2.8.1 Chi Squared Test . . . 46
  - 2.8.2 Akaike Information Criterion . . . 46
  - 2.8.3 Bayesian Information Criterion . . . 49

3 Conclusions 51
- 3.1 New York City . . . 52
- 3.2 Future Research . . . 54

Bibliography 56

List of Tables

1.1 Number of truncated data per county or state. 14

2.1 Values of λ for Box Cox transform 27

2.2 Parameters for each county or state and each distribution 37

2.3 Parameters for each county or state and each distribution, continued 38

2.4 Standard errors for estimated parameters 39

2.5 Standard errors for estimated parameters, continued 40

2.6 Square root of the Mean Squared Error 43

2.7 Mean Absolute Errors 45

2.8 Chi-squared values for each county or state and distribution 47

2.9 AIC values for each county or state and distribution 48

2.10 BIC values for each county or state and distribution 50

3.1 Measures For All Five New York City Boroughs, $N = 5597$. 54

List of Figures

1.1 Pareto's Law 3

1.2 District of Columbia Census Block Groups 11

1.3 Three Counties Median Household Income Distribution 15

1.4 Three Counties Median Household Income Distribution, 2 16

1.5 Three Counties Median Household Income Distribution, 3 17

1.6 Three Counties Median Household Income Distribution, 4 18

1.7 One County, Two States Median Household Income Distribution, 5 . 19

1.8 Three States Median Household Income Distribution, 6 20

1.9 Two States Median Household Income Distribution, 7 21

2.1 The Lognormal Distribution for various values of σ^2. 25

2.2 When deciding whether to use normal or lognormal, we find the maximum log likelihood for λ. District of Columbia median household income per block group data. 28

2.3 Choosing between normal and lognormal models. Comparing the Q-Q plots of X vs. $\log(X)$, District of Columbia data 29

2.4 The Gamma Distribution for various values of α 30

2.5 Finding $\hat{\alpha}$ for the Gamma Distribution, District of Columbia data . . 31

2.6 The Weibull Distribution for various values of k 33

2.7 QQ Plots of 4 Distributions, Cook County 42

2.8 QQ Plots of 4 Distributions, Dallas County 44

3.1 Comparing New York County and New York City histogram and mean 53

3.2 Comparing New York County and New York City QQ plots using the Singh-Maddala 53

Chapter 1

Introduction

In this thesis we are going to use U.S. Census data to study median household income distribution for 13 U.S. counties and seven U.S. states. Over the years, researchers have fitted income data with various probability distributions. During our review of the literature, we saw that researchers do not agree on any one best distribution. This is a good thing for us, because it allows us to choose the distribution that best fits our data.

We will be looking at lognormal, gamma and Weibull, each of which has two parameters. We will also investigate the Singh-Maddala, which has three parameters. Finally, we will introduce the Generalized Beta II, which has four parameters.

These distributions will be tested using Mean Squared Error, Mean Absolute Error and Chi-square Goodness-Of-Fit. We explore QQ plots to make determinations as to which distribution best fits our observed data. We also use Akaike's Information Criterion, or AIC, and Bayesian Information Criterion, or BIC, for distribution selection.

1.1 Survey

In the following section, we survey some of the major results discovered in the study of income distributions.

1.1.1 Pareto

The study of income distributions began in earnest in the early 20th Century. H.L. Moore [14] argued for a statistical complement to the pure economics that had begun about 130 years earlier with the work of Adam Smith. It was around that time that new income and property tax laws in several countries began to provide statisticians and economists with rich sources of data.

Pareto [15] was one of the earliest researchers to take advantage of the wealth of new data. He plotted the cumulative distributions of income for several cities throughout Europe on double logarithmic paper, and he discovered that in each case the result was a straight line with about the same slope. Refering to Figure 1.1, we see the same effect with data from three U.S. counties.

Pareto argued that these distributions could be characterized by the curve

$$\log N = A - \alpha(\log x),$$

where N is the number of households with incomes greater than x, A is a parameter and α is the absolute value of the slope. The law says that for $x = x_1, x_2, \ldots$ that the logarithm of the number of incomes exceeding x is a linear function of $\log x$. Pareto's law is not obeyed for low incomes. In Figure 1.1, we did not include the lower incomes, as they do not form the nice downward sloping lines predicted by Pareto's equation. Pareto defined a minimum income $h = (A/P)^{1/\alpha}$, where P is total population, as the cutoff for including the lower incomes.

Pareto studied data from many different populations, as diverse as ancient Peru, the Cherokee Indians, and Prussia, and he noticed that they all had quite

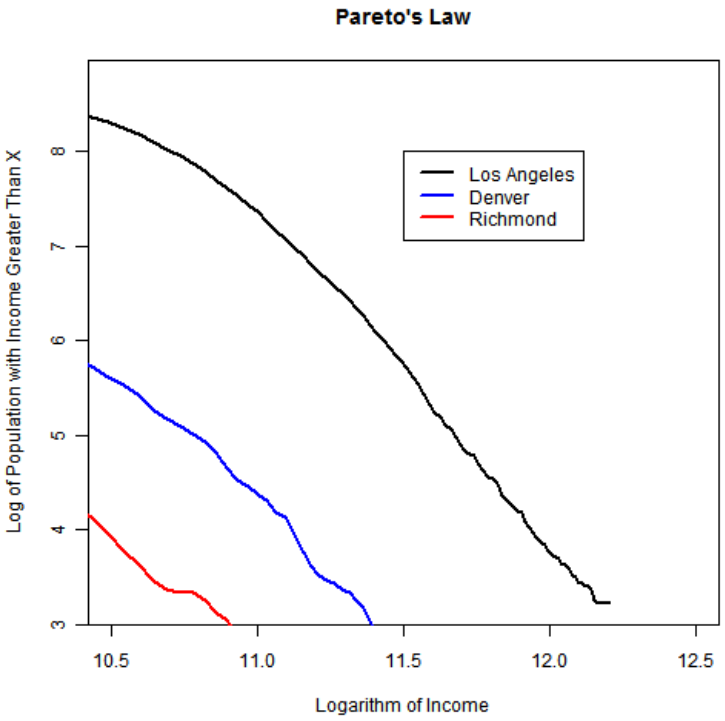

Figure 1.1: Pareto's Law

similar distributions of income. The same principle will be observed in the data used in this study from U.S. counties and states in the year 2000. In fact, Pareto claimed that this curve is the equilibrium position, and that wealth is always distributed in the same proportions across societies.

Pareto's claim of equilibrium convinces us that it is appropriate to use the measure of income provided by the Census, that is, the median household income by block group. We have this data per county and per state, and in both cases, we see the same results for the best fit.

1.1.2 The Mid 20th Century

Champernowne [2] suggests a cumulative distribution function of

$$F(t) = \frac{N}{\theta} \tan^{-1} \left\{ \frac{\sin \theta}{\cos \theta + (t/t_0)^{\alpha}} \right\}$$

and the probability density function is

$$f(t) = \frac{\alpha N \sin \theta}{\theta t \left\{ (t/t_0)^{\alpha} + 2 \cos \theta + (t/t_0)^{\alpha} \right\}}$$

where $F(t)$ is the number of incomes exceeding t, N is the value of $F(0)$, which can be interpreted as the total number of people with any income at all. The parameter t_0 is the median income, and α can be shown to equal the slope of the high-income asymptote to the curve $y = F(t)$ when plotted on log-log paper.

The problem is with the parameter θ. In [3], Champernowne suggests that no simple economic interpretation can be given to the fourth parameter θ. This is a problem, as we like our mathematical equations to be interpretable.

Champernowne says that this form is especially well suited for high incomes, and that it follows the form predicted by Pareto's Law. Champernowne [2] writes the following, which is an idea brought up by many authors who study income distribution:

> The forces determining the distribution of incomes in any community are so varied and complex, and interact and fluctuate so continuously, that any theoretical model must either be unrealistically simplified or hopelessly complicated.

Champernowne chooses the former, but then goes on to say in Section 3 of [2] that certain assumptions in his example model do not approach reality at all. Nevertheless, the author believes that his assumptions will lead to a better understanding of the actual distributions of income.

Rutherford [17] says that researchers want to produce an explanation of the factors determining the distribution of income. We want our parameters to have a clearly interpretable economic relevance. This means that we can't just have a model with many parameters, but one that fits the empirical data well, if we can't find a real world explanation for those parameters.

1.2 An Example Of The Problems With Fitting Income Distributions

Majumder and Chakravarty [8] state that in applied work, the models most frequently used are the Pareto, lognormal and gamma distributions. All three of these distributions are parsimonious with respect to the number of parameters, and those parameters are easily interpretable into real life meanings.

The authors describe four characteristics that income distributions should satisfy. They are (1) The *Weak Pareto Law* by Mandelbrot [9], (2) parametric parsimony, (3) generality, and (4) computational simplicity.

The authors admit some shortcomings of each distribution, such as the fact that the lognormal and gamma distributions do not satisfy the *Weak Pareto Law*, and that none of the three distributions accurately model the whole range of incomes.

1.2.1 Weak Pareto Law

To understand the *Weak Pareto Law*, we must first consider the *Strong Pareto Law*. Let $P(u)$ be the percentage of individuals with an income U exceeding some number u, where u is assumed to be continuous. The Strong Pareto Law says that

$$P(u) = \begin{cases} (u/u^0)^{-\alpha}, & \text{when } u > u^0 \\ 1, & \text{when } u < u^0 \end{cases}$$

Then the density $p(u) = -dP(u)/du$ satisfies

$$p(u) = \begin{cases} \alpha(u^0)^\alpha u^{-(\alpha+1)}, & \text{when } u > u^0 \\ 0, & \text{when } u < u^0 \end{cases}$$

This distribtuion is fully specified by two state variables, u^0, a scale factor, and α which is an index of inequality of income distribution. Graphically, this strong law tells us that the log-log graph of $y = \log P$ as a function of $v = \log u$ is a straight line.

The Strong Pareto Law is acknowledged to be unjustified by the empirical data, so Mandelbrot [9] suggests this Weak Pareto Law,

$$P(u) \text{ behaves like } (u/u^0)^{-\alpha}, \text{ as } u \to \infty.$$

or

$$\frac{P(u)}{(u/u^0)^{-\alpha}} \to 1, \text{ as } u \to \infty$$

Graphically, this means that the curve (log P, log u) is asymptotic to the straight line which represents the Strong Pareto Law.

1.2.2 Majumder and Chakravarty

The authors Majumder and Chakravarty [8] go on to propose the following model which is represented by the following probability density function

$$f(x;a,b,c,d) = \frac{bd^{a/b}c^{(b/d)-a}x^{(b/d)-a-1}}{B((1/d)-(a/b),a/b)}((cx)^b+d)^{-1/d},$$

for $x > 0$ and $b > ad$, and where $B(\cdot,\cdot)$ denotes the beta function, and the parameters a, b, c, d may be thought of as the Pareto constant for high income, the rate of convergence of the distribution to the Pareto distribution, the scale parameter, and the flexibility parameter respectively.

The authors Majumder and Chakravarty claim this model provides a better fit to some income data than the lognormal, gamma, Singh-Maddala, Dagum, and the generalized beta of the second kind ($GB2$) distributions, and they go on to back up their findings with empirical results.

This distribution proposed by Majumder and Chakravarty also contains the Singh-Maddala as a special case. The Singh-Maddala distribution can be obtained by setting

- $a = a_2a_3$
- $b = a_2$
- $c = (a_1/(a_3+1))^{1/a_2}$
- $d = 1/(a_3+1)$.

A few years after Majumder and Chakravarty claimed to have created a distribution which fits the data better than all other distributions, the authors McDonald

and Mantrala [11] gave dispute to those findings. They go on to show that the model proposed by Majumder and Chakravarty is actually a reprameterization of the Generalized Beta II, and that Majumder and Chakravarty's analysis of the data is flawed. The Generalized Beta II is defined by the probability density function

$$GB2(x; a, b, p, q) = \frac{|a|x^{ap-1}}{b^{ap}B(p,q)(1+(x/b)^a)^{p+q}}$$

where the parameters b, p, and q are positive. We will speak of the Generalized Beta II again in Section 2.4.2.

The authors McDonald and Mantrala show that the Majumder and Chakravarty model and the Generalized Beta II model can be related as follows:

$$M\&C \rightarrow GB2 : (|a|, b, p, q) = (\beta, \delta^{1/\beta}/\gamma, 1/\delta - \alpha/\beta, \alpha/\beta)$$

$$GB2 \rightarrow M\&C : (\alpha, \beta, \gamma, \delta) = (|a|q, |a|, (p+q)^{-1/|a|}/b, 1/(p+q))$$

The authors McDonald and Mantrala explain the flaws in Majumder and Chakravarty's analysis of data by saying that the different studies they compare against their findings are based on different intervals of income data, different groupings, different base years, and different estimation procedures.

This is just one example of how there is not much agreement on the best density to use to fit to income distribution. In many of the references in the bibliography of this thesis, we find that the authors admit that it is hard to agree on the best fit, and they go on to say that there is often a trade off between getting a good fit and hitting all four of the characteristics listed above.

1.2.3 Parametric Models

The authors Fesser and Ronchetti [22] have a few criticisms about parametric models. They state that many studies on income data have come to the conclusion that the model does not fit the data well. They mention several sources for deviations from the models, including outliers, grouping effects and other general misspecifications of the model. By grouping effects, the authors are referring to grouping the data into bins for histograms.

These deviations can affect the maximum likelihood estimators for the parameters of our models, and that can cause the model to become biased and inefficient. This could then affect income inequality measures used by economists, such as the Gini coefficient. The authors state that a few extreme observations could impact the inequality measure, causing the measure to no longer represent the inequality for the whole population.

1.3 The Data

The data we will use in the analysis of our distributions come from the 2000 Decennial Census. The United States Census Bureau conducted the census and published the data.

We downloaded our data from American FactFinder on the U.S. Census Bureau's website `www.census.gov` [1].

The data come from Table P53 of the SF3 file from the 2000 Decennial Census. The title of the table is *Median Household Income in 1999 (Dollars)*. Median

household income is defined as the amount which divides households into two equal groups, one having incomes above that amount and the other having incomes below that amount. The universe for our data are households, which the U.S. Census Bureau defines as all people who occupy a housing unit. So for our data, the value reported is that which divides all the households in a particular block group into two equal groups.

The U.S. Census Bureau conducts a full census every ten years in order to enumerate the people so that seats in the U.S. House of Representatives may be apportioned according to state population. The U.S. Constitution demands that a full census be taken every ten years in Article I Section 2.

The information contained in Table P53 is actually from survey data, not a full census. In the 2000 Census, the "long form" was used to collect extra data other than number of people in the household. This form was sent to approximately one out of every six households.

Since the long form is a sample survey and not a census, the data collected are not the same. For instance, in a survey, each household interviewed is given a weight greater than one, whereas in a census each household would have a weight of exactly one.

The data are reported in units as small as Census Block Groups and as large as the entire United States. Our data are reported as the median income X_i for each block group i in a county or state. According to `www.census.gov`, a block group is the smallest geographic unit for which the Census Bureau tabulates sample data.

In Figure 1.2 we see what block groups look like. There are 426 block groups

with valid data in the District of Columbia. By valid data we are speaking of data points which are not *NA*. We occasionally encounter block groups which are *NA*, an example of which is the block group that contains the National Mall in Washington, DC. Refer to Table 2.2 to see the number of block groups N in each data set.

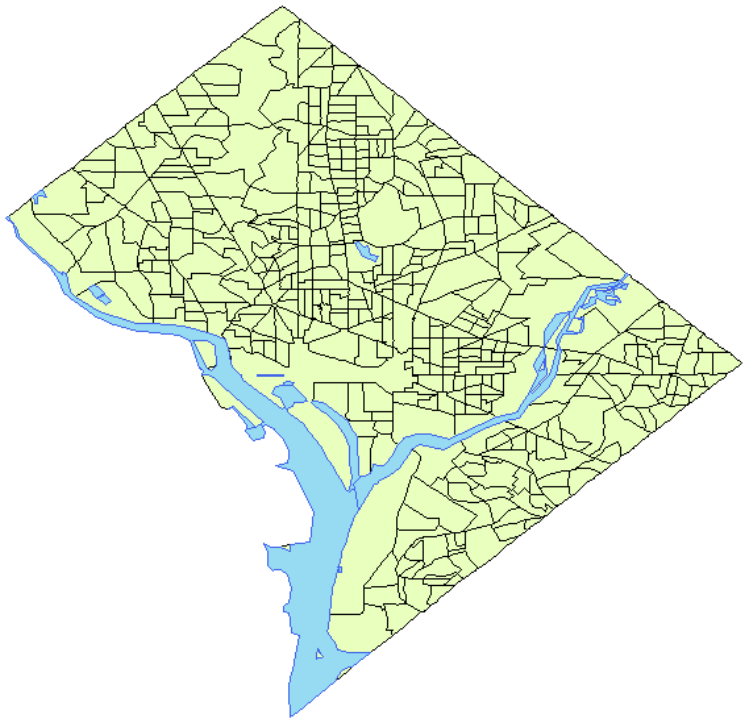

Figure 1.2: District of Columbia Census Block Groups

There is another thing to note about the data. For reasons related to data privacy and confidentiality, the very top incomes are masked in the reported data. If the median household income of a block group is above $200,000, then that block group is reported as $200,001. This practice is known as top-coding. The U.S. Census Bureau defines top-coding as method of disclosure limitation in which all cases in or above a certain percentage of the distribution are placed into a single

category.

This data masking affects the data in small ways. If there are many block groups with median household income greater than $200,000, then that causes a small bump in the histogram to appear near the end of the right hand tail of our data, which could affect the parameter estimates for our fitted models. To guard against this, we truncated the data to only include values less than $200,001.

1.3.1 Assumptions About the Data

We make two assumptions about the data. The data are top-coded, so we truncate the data. Our data are of the form $X|X < 200001$. Our first assumption is that we are using density functions of the form

$$\frac{f(x;\theta)}{P(X < 200001)}, \text{ where } 0 < x < 200001.$$

This cutting point for the truncation is very large. In every county or state we investigate, the amount of trucated data are less than 1% of our total amount of our total number of observations per county or state. In many cases, no data are truncated. The probability $P(X < 200001)$ is very close to 1 in every case. Therefore, for practical considerations, we are fitting data to density functions of the form

$$f(x;\theta), \text{ where } x > 0.$$

To see how many observations get truncated in each data set, refer to Table 1.1. The columns Pre N and Post N refer to the number of observations before truncation and after truncation, respectively. The column Percent refers to the percentage of

observations that were truncated.

We will also assume independence of the data, even though it is reasonable to believe that the data is not truly independent. Two neighboring block groups are more likely to have the same median household income than a block group on the other side of the county. Since we are not concerned with the spatial aspects of the data, this assumption seems reasonable.

We did not consider mixed distributions as they are not mentioned in the literature of income distributions. That is because we only considered positive income x where $x > 0$.

1.4 The Counties and States

Refer to Figures 1.3 through 1.9 to see the histograms for the data we are investigating. The vertical line in each figure shows the mean of our data for each histogram. Of special interest is the histogram for New York County, which is coextensive to the New York City Borough of Manhattan, in Figure 1.5. We will see later that the New York County data do not stand up against our conclusions for the best probability density model to use, and we suggest a way to make the data agree with our conclusions.

Table 1.1: Number of truncated data per county or state.

County/State	Truncated	Pre N	Percent	Post N
Los Angeles	25	6268	0.399	6243
San Francisco	1	573	0.175	572
Denver	0	468	0.000	468
District of Columbia	1	427	0.234	426
Fulton	3	446	0.673	443
Cook	13	4184	0.311	4171
Suffolk	0	629	0.000	629
Baltimore City	0	701	0.000	701
New York County	5	850	0.588	845
Multnomah	0	508	0.000	508
Philadelphia	2	1775	0.113	1773
Dallas	10	1675	0.597	1665
Richmond City	1	163	0.613	162
Georgia	3	4775	0.063	4772
Massachusetts	8	5032	0.159	5024
Maryland	6	3648	0.164	3642
New Jersey	19	6447	0.295	6428
Virginia	5	4722	0.106	4717
Kansas	1	2288	0.044	2287
Utah	0	1472	0.000	1472

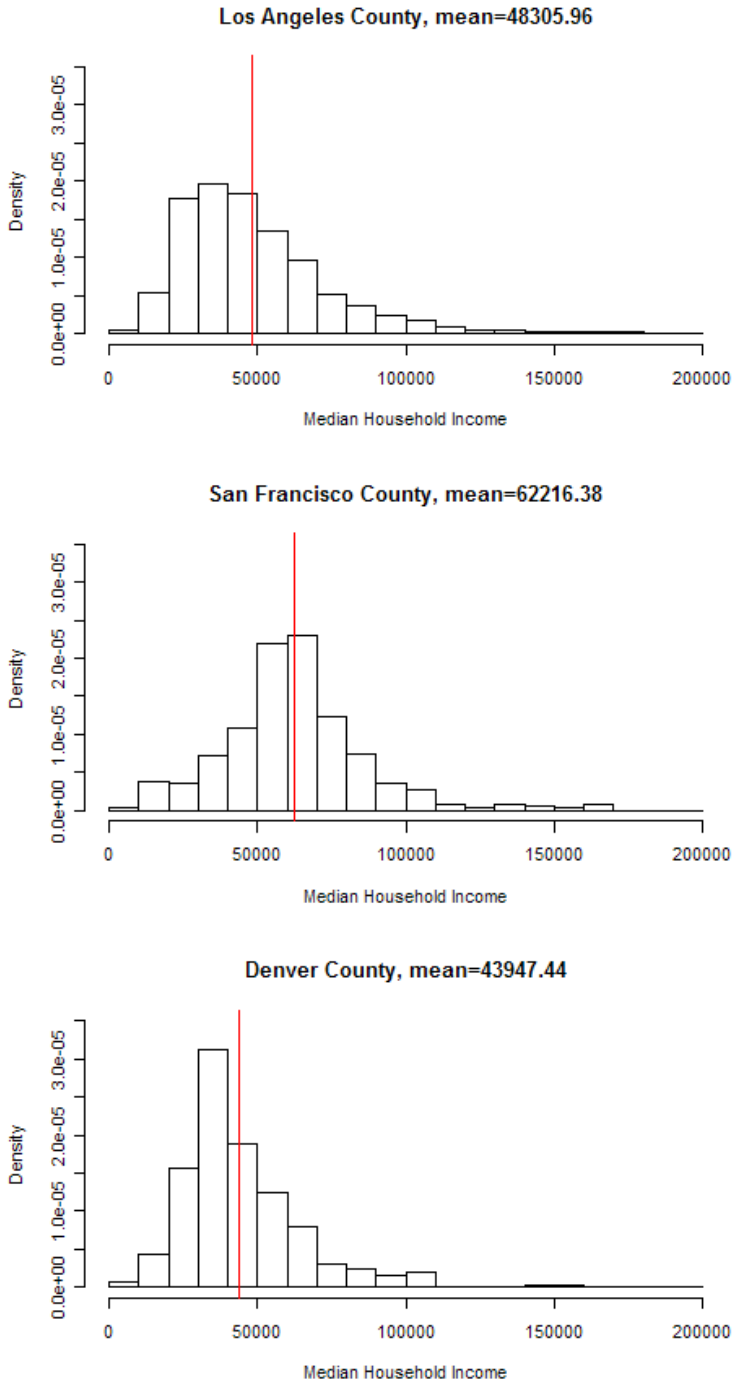

Figure 1.3: Three Counties Median Household Income Distribution

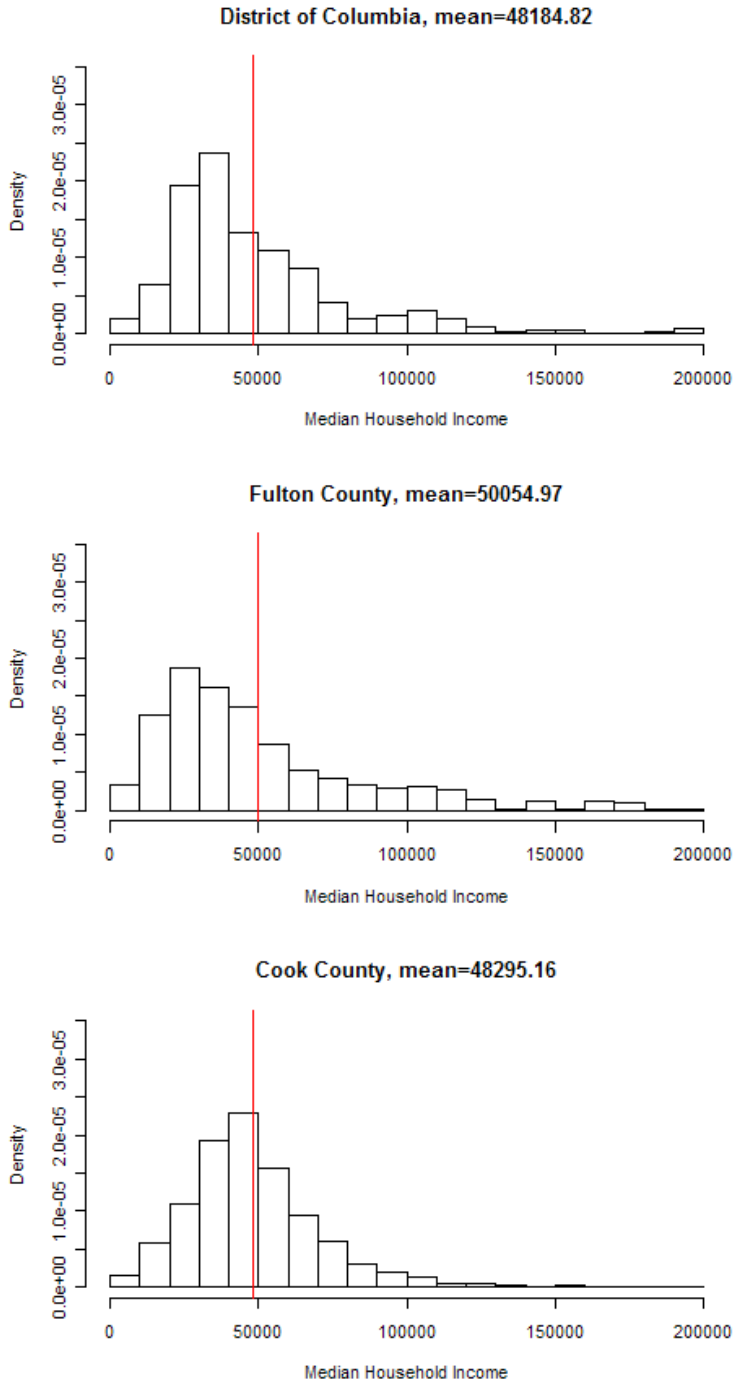

Figure 1.4: Three Counties Median Household Income Distribution, 2

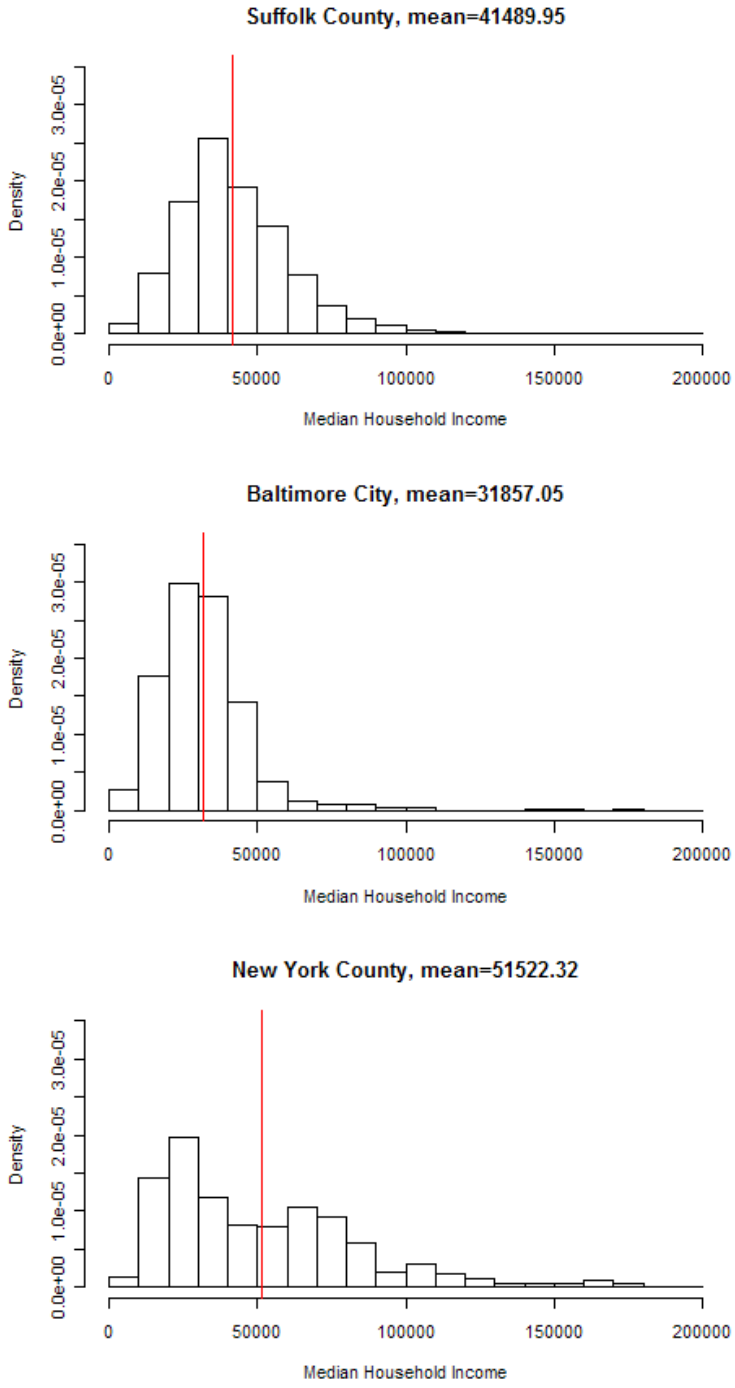

Figure 1.5: Three Counties Median Household Income Distribution, 3

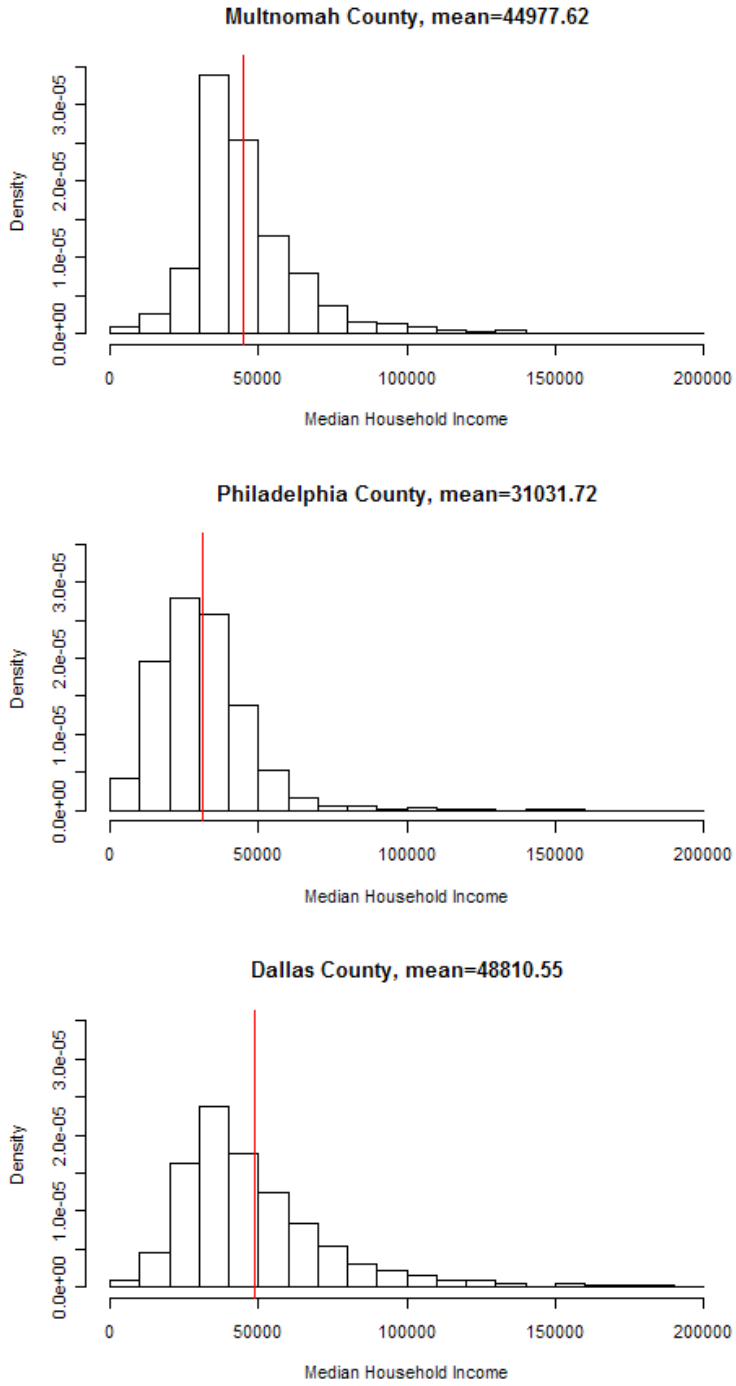

Figure 1.6: Three Counties Median Household Income Distribution, 4

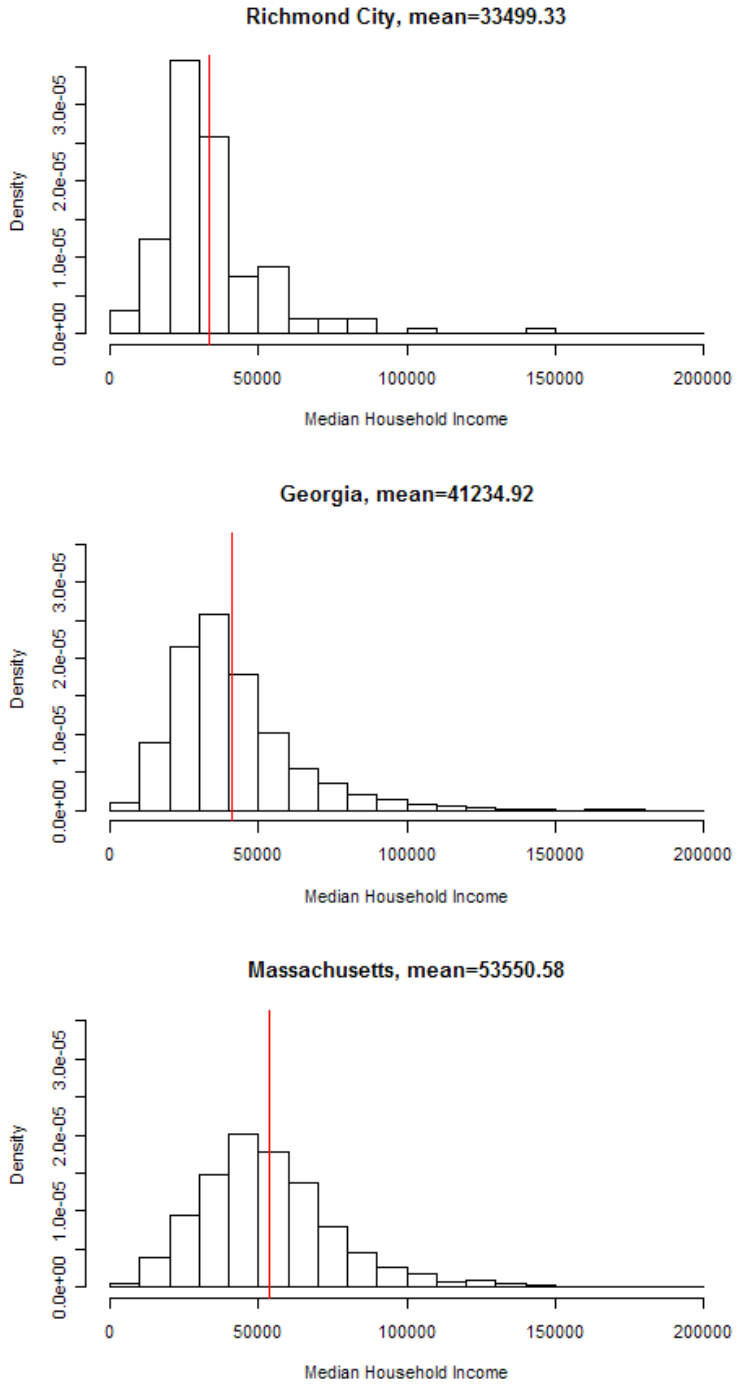

Figure 1.7: One County, Two States Median Household Income Distribution, 5

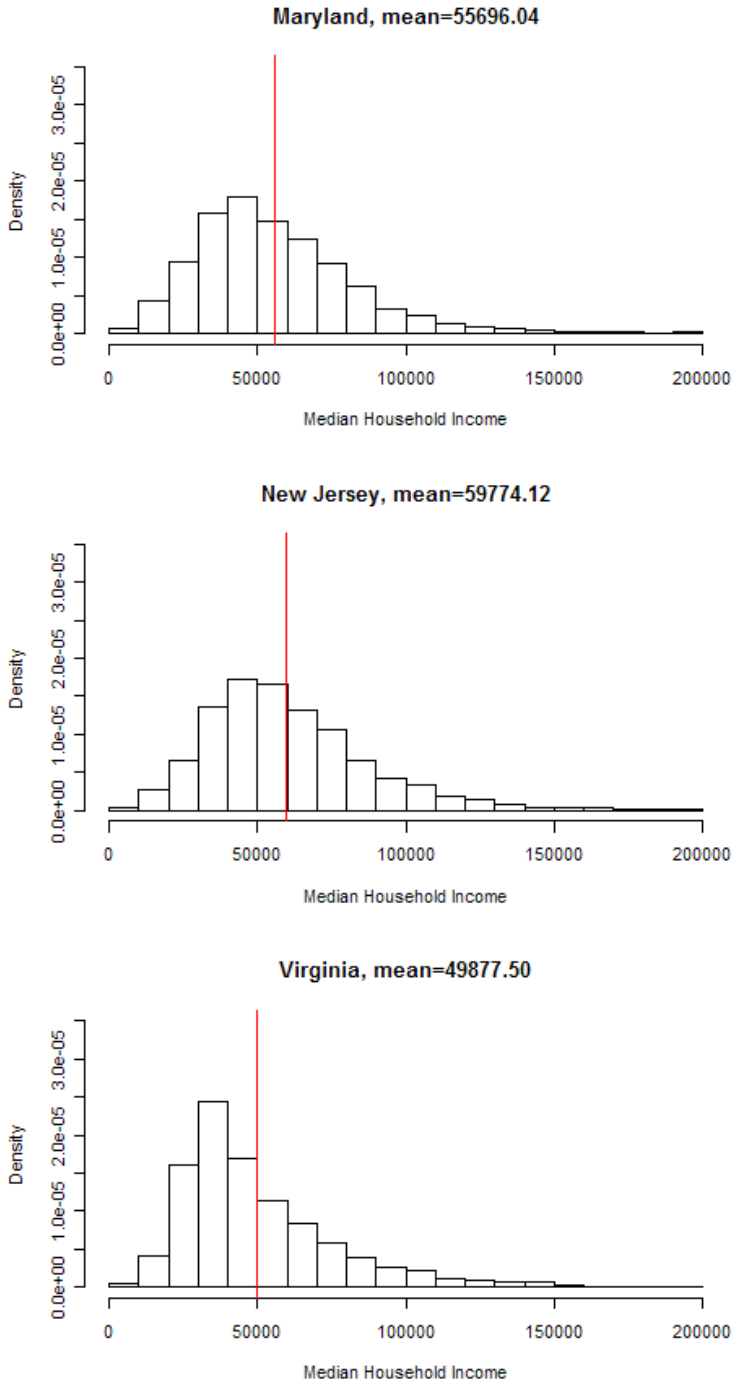

Figure 1.8: Three States Median Household Income Distribution, 6

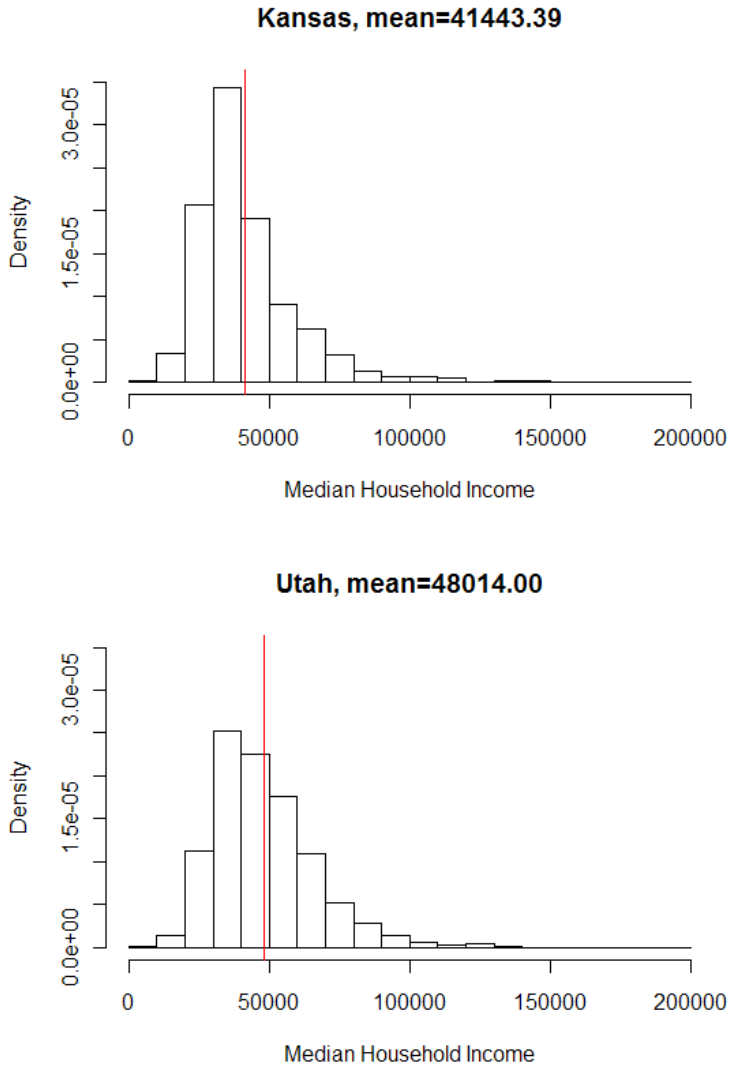

Figure 1.9: Two States Median Household Income Distribution, 7

Chapter 2

Four Densities

We will now investigate four probability density functions and try to find the best one that fits our data. Our data are median household incomes for block groups in several US counties and states. The data has been truncated so we have data x such that $x \leq 200,000$.

A household income distribution gives us a variability which can be interpreted as income inequality. This is a useful measure for economists and demographers. Fonseca and Tayman [6] state that there are three ways to estimate income distribution.

First, one can use sample surveys, such as the Current Population Survey (CPS). The CPS is a national level survey, and is not useful for substate areas such as counties or cities.

Next, one can analyze income tax data, which can be useful for smaller, substate areas, but income tax data also has both postive aspects and drawbacks. It represents a complete count and not a survey, but some states do not have an income tax, and therefore cannot be studied at the substate level.

Finally, we have the method of mathematical modeling. Mathematical models are useful because they provide a way to describe income distribution without having to conduct a survey or collect data. Instead the model is formulated through a series

of assumptions and underlying processes and then tested against existing empirical data.

Different researchers have come up with different factors to relate to income distributions, such as age and educational level, but Fonseca and Tayman [6] state that there is no agreement about the underlying generation processes that lead to income distributions.

It makes sense to relate our income distributions to known densities, because that allows statisticians, economists and demographers to use all the knowledge about those densities in the analysis of our data.

Next we will review at some characteristics of four different densities that we will be looking at in the analysis of our data. Looking at a histogram of our data suggests that we study certain kinds of distributions, that is, distributions with a long right hand tail. We are going to study the Lognormal, Gamma, Weibull and Singh-Maddala. We will also introduce the Generalized Beta II distribution and relate it to the previous four distributions.

2.1 The Lognormal Density

A variable X has a lognormal distribution if $Y = \log X$ has a normal distribution. Suppose Y is $N(\mu, \sigma^2)$. Then we have

$$\mathrm{E}(X) = e^{\mu+\sigma^2/2},$$

$$\mathrm{Var}(X) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

The lognormal cumulative distribution function (cdf) is defined as

$$F(x) = \frac{1}{2} + \frac{1}{2}\text{erf}\left[\frac{\ln(x) - \mu}{\sigma\sqrt{2}}\right]$$

where erf is the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

The lognormal has a probability density function (pdf)

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln(x) - \mu)^2}{2\sigma^2}\right], \qquad x > 0.$$

It is interesting to note that even though all moments exist and are given by

$$\text{E}(X^s) = e^{s\mu + s^2\sigma^2/2},$$

there is no moment generating function for the lognormal distribution.

According to the Encyclopedia of Statistical Sciences by Kotz & Johnson [7], the lognormal distribution is often used as a model for the distribution of income and wealth, and it should be considered as a possible model any time a model with positive skewness is needed. Positive skewness means that the right hand tail is longer, and that is the case with our data.

2.1.1 Choosing Between X and $\log X$

How do we know we should use lognormal instead of the more common and more familiar normal density? Let's look at some of the literature. Sclove [19] compares the maximum likelihood method of Box and Cox to the method of using correlation coefficients to make a choice between $Y = \alpha + \beta X$ and $\log Y = \alpha + \beta X$.

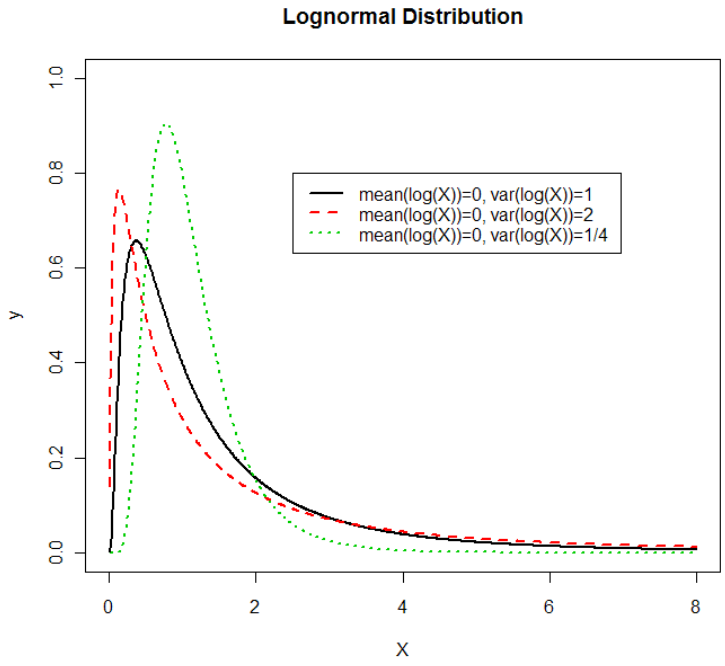

Figure 2.1: The Lognormal Distribution for various values of σ^2.

Sclove goes on to show that the likelihood method ends up being preferable since it it based on a probability model which lets us give a confidence statement about the choice of the model.

In our case, Y represents the median household income in 1999, the full year prior to the 2000 Census. The Box and Cox transformation is defined for all real λ by

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0 \\ \log y, & \text{if } \lambda = 0 \end{cases}$$

for $y > 0$. To find the correct λ to use in the Box-Cox transformation, we need to maximize the log likelihood function,

$$l(\lambda) = -\frac{n}{2} \log \hat{\sigma}^2(\lambda) + (\lambda - 1) \sum_{i=1}^{n} \log y_i$$

with respect to λ. The maximum likelihood estimate of λ is the value $\hat{\lambda}$ which maximizes $l(\lambda)$. The $100(1-\alpha)\%$ confidence interval for λ is

$$\left\{\lambda : l(\lambda) > l(\hat{\lambda}) - \frac{1}{2}\chi^2_{1-\alpha}(1)\right\}$$

where $\chi^2_p(f)$ denotes the $100 \cdot p$-th percentile of the chi-square distribution with f degrees of freedom.

We can use the software package `R` [16] to graph λ so we can find the maximum. We graph λ versus $l(\lambda)$ using the `R` command `boxcox` [21]. The command `boxcox` computes the log-likelihoods for the parameter of the Box-Cox power transformation within the framework of a normal error model. We see in Figure 2.2 that the 95% confidence interval contains 0, so we will choose $\lambda = 0$ and we will log transform our data Y. Therefore, it is correct to use the lognormal density instead of the normal distribution for the District of Columbia data. Refer to Table 2.1 to find out if it is appropriate to use lognormal for each of our data sets. As we see in Table 2.1, it is not always appropriate to consider each of our data sets to be lognormal.

Another way of discovering whether to use Y or $\log(Y)$ is to look at the Q-Q plot. We notice in Figure 2.3 that the left figure is not a good fit and that the right figure shows a better fit. We can see that the log transform renders the distribution closer to normal, although the fit is only really good in the middle. The fit in the tails is a little bad, but that was expected from reading the literature.

Table 2.1: Values of λ for Box Cox transform

County/State	λ	Does 95%CI contain 0?
Los Angeles County	0.07	Yes
San Francisco County	0.58	No
Denver County	0.14	Yes
District of Columbia	0.11	Yes
Fulton County	0.12	Yes
Cook County	0.38	No
Suffolk County	0.49	No
Baltimore City	0.12	Yes
New York County	0.11	Yes
Multnomah County	0.18	Yes
Philadelphia County	0.29	No
Dallas County	0.08	Yes
Richmond City	0.02	Yes
Georgia	0.12	No
Massachusetts	0.37	No
Maryland	0.30	No
New Jersey	0.33	No
Virginia	0.02	Yes
Kansas	-0.17	No
Utah	0.09	Yes

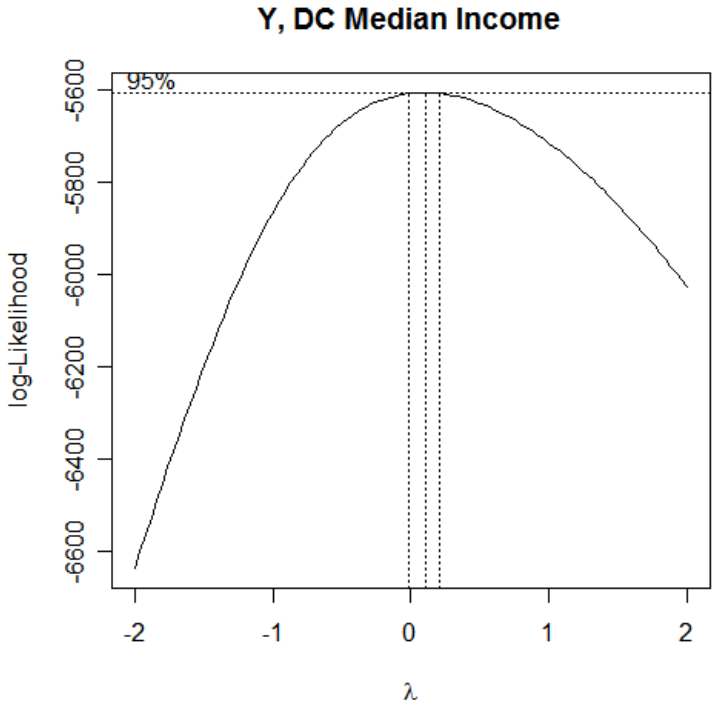

Figure 2.2: When deciding whether to use normal or lognormal, we find the maximum log likelihood for λ. District of Columbia median household income per block group data.

2.2 The Gamma Density

The gamma cdf is defined as

$$F(x;k,\theta) = \int_0^x f(u;k,\theta)\,du = \frac{\gamma(k,x/\theta)}{\Gamma(k)}$$

where $\gamma(k,x/\theta)$ is called the lower incomplete gamma function and is defined as

$$\gamma(a,x) = \int_0^x t^{a-1}\,e^{-t}\,dt.$$

The gamma distribution has a probability density function of

$$f(x;\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \text{ for } x > 0 \text{ and } \alpha,\lambda > 0$$

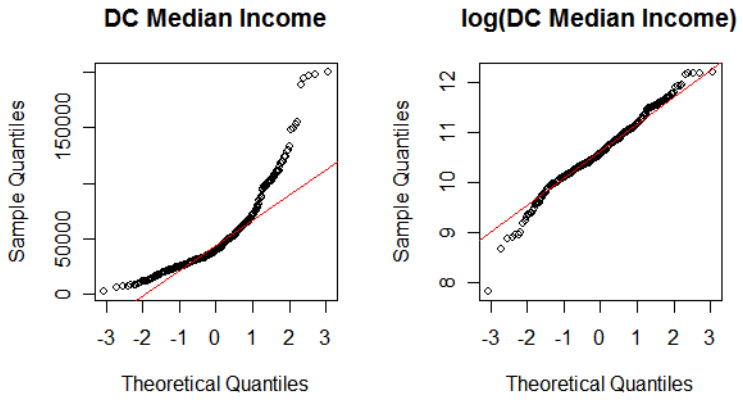

Figure 2.3: Choosing between normal and lognormal models. Comparing the Q-Q plots of X vs. $\log(X)$, District of Columbia data

where α and λ are the shape parameter and scale parameter, respectively, and

$$\Gamma(\alpha) = \int_0^\infty e^{-u} u^{\alpha-1} du$$

is the gamma function. The mean is given by α/λ and the variance is given by α/λ^2. Note that when $\alpha = 1$, the gamma distribution reduces to the exponential distribution. To see this and other values of α refer to Figure 2.4.

The moments for the gamma distribution exist and are given by

$$\mathrm{E}(X^s) = \frac{\Gamma(\alpha + s)}{\lambda^s \Gamma(\alpha)}$$

and the moment generating function is given by

$$\left(\frac{\lambda}{\lambda - t}\right)^r$$

for $t < \lambda$.

The maximum likelihood estimators of λ and α may be derived by solving the

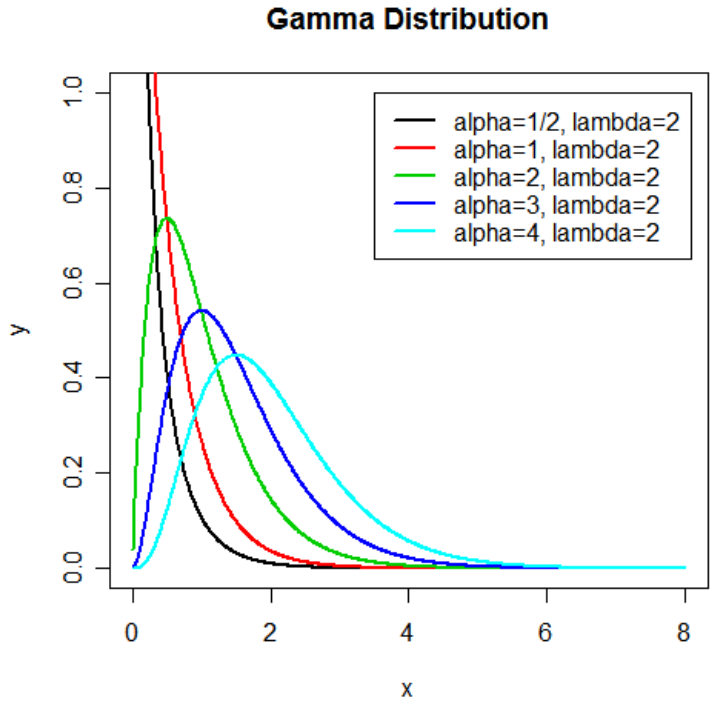

Figure 2.4: The Gamma Distribution for various values of α

following two equations,

$$\hat{\lambda} = \hat{\alpha}/\bar{X},$$

$$\log\hat{\alpha} - \psi(\hat{\alpha}) = \log\bar{X} - \log\tilde{X} = \log\bar{X}/\tilde{X},$$

where $\bar{X} = \sum_i X_i/N$ is the arithmetic mean, $\tilde{X} = (\prod_i X_i)^{1/N}$ is the geometric mean, and $\psi(\alpha)$ is the digamma function, which is available in `R`. The digamma function is defined as $\psi(x) = \Gamma'(x)/\Gamma(x)$. We want to estimate $\hat{\lambda}$ and $\hat{\alpha}$ so we can use them as initial values for the `R` function `fitdistr()` which we will use in our analysis of the data.

Note that $\log\hat{\alpha} - \psi(\hat{\alpha})$ is a strictly decreasing function of $\hat{\alpha}$. So, given $\log\hat{\alpha} - \psi(\hat{\alpha})$, we can find $\hat{\alpha}$ visually by looking at Figure 2.5.

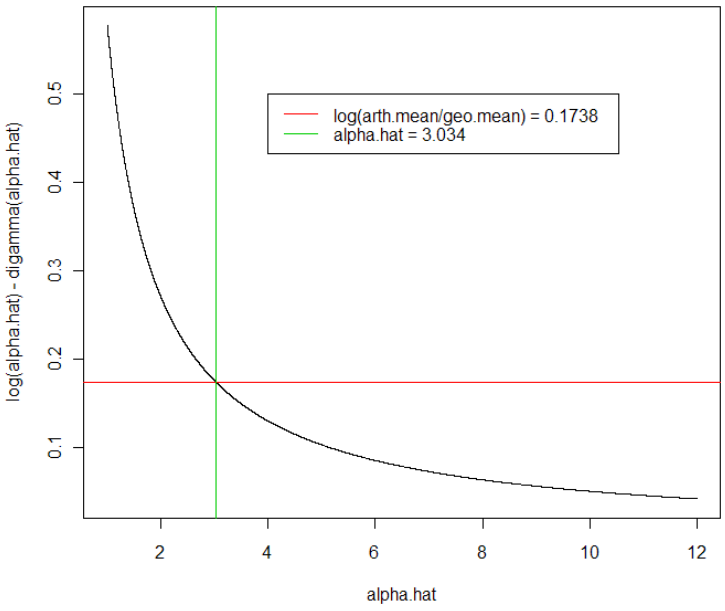

Figure 2.5: Finding $\hat{\alpha}$ for the Gamma Distribution, District of Columbia data

For the District of Columbia median household income data, $\log \bar{X}/\tilde{X} = 0.1738$, so we can look at Figure 2.5 and see that $\hat{\alpha} = 3.034$ approximately. That gives us $\hat{\lambda} = 6.2956 \times 10^{-05}$.

Salem and Mount [18] assert that income distribution density functions should provide reasonably close approximation to the true distribution and have parameters that are easy to estimate and to interpret in an economically meaningful way. In practice, these criteria often compete with each other. For example, the lognormal distribution has parameters that are easy to estimate and to relate to inequality measures, but it doesn't fit the full range of income data very well.

According to Salem and Mount [18], the shape parameter can be associated with the concept of proportionate growth in income, while the scale parameter can be

thought of as a measure of income inequality. The two parameters of the lognormal distribution can be interpreted in the same way, in fact, the authors assert that lognormal is the closest alternative to gamma in terms of the ease of interpretation of the parameters. However, the authors claim that the gamma fits income data better than lognormal.

2.3 The Weibull Density

Our final two parameter probability density is the Weibull distribution.

The Weibull distribution has a cdf defined as

$$F(x; k, \lambda) = 1 - e^{-(x/\lambda)^k}$$

for $x \geq 0$.

The Weibull distribution has a pdf of

$$f(x; k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, \qquad x \geq 0$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. Note that when $k = 1$ the Weibull reduces to the exponential, and when $k = 3.4$ the Weibull appears similar to the normal distribution. The mean and variance of the Weibull are given by the following

$$E(X) = \lambda\Gamma(1 + 1/k)$$

$$Var(X) = \lambda^2[\Gamma(1 + 2/k) - \Gamma^2(1 + 1/k)]\,.$$

The Weibull distribution has moments of

$$\mathrm{E}(X^s) = (1/\lambda^k)^{-s/k}\Gamma\left(1 + \frac{s}{k}\right).$$

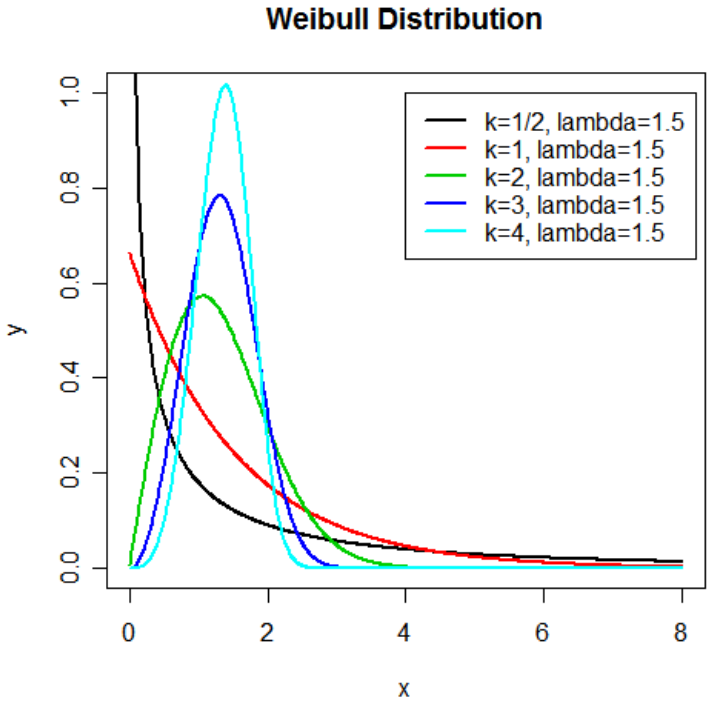

Figure 2.6: The Weibull Distribution for various values of k

The authors of the Weibull article in [7] state that the Weibull density is generally compared to the lognormal, gamma and exponential distributions. This means it is appropriate for us to include Weibull in our investigation of the proper model to fit to income distributions.

2.4 Distributions With More Than Two Parameters

Now we can look at other distributions that have more than two parameters. In some cases, it might not be appropriate to compare the fit of these distributions with our previous three distributions. Particularly, see the comment on the Singh Maddala paper by Cramer [4]. Cramer points out that Singh and Maddala report

the Sum of Squared Errors between observed and calculated frequencies, and they compare their results to Salem and Mount [18], who favor the Gamma density when fitting income distributions. Cramer criticizes Singh and Maddala for failing to point out that they are estimating three parameters while Salem and Mount use two parameters.

2.4.1 Singh-Maddala

The authors Singh and Maddala [20] suggest a cdf for the Singh-Maddala function given by

$$F(x; a, b, c) = 1 - \frac{1}{(1 + ay^b)^c}$$

The pdf is then given as

$$f(x; a, b, c) = \frac{abcx^{b-1}}{(1 + ax^b)^{c+1}}$$

and the authors McDonald and Ransom [13] derived the following mean

$$E(X) = \frac{\Gamma(1/b + 1)\Gamma(c - 1/b)}{(a^{1/b})\Gamma(c)}$$

The help files for the software *Stata* give the variance for the Singh-Maddala density as

$$Var(X) = \frac{\Gamma(2/b + 1)\Gamma(c - 2/b)}{(a^{2/b})\Gamma(c)} - (E(X))^2.$$

Singh and Maddala compare their distribution to the work done by Salem and Mount [18] on the gamma distribution, and the authors find that their distribution fits Salem and Mount's data better than either lognormal or gamma.

Note that the Weibull distribution is a special case of the Singh-Maddala distribution where we take the limit as $c \to \infty$ and let $a = 1/\beta^b c$. Therefore we should expect the Singh-Maddala to fit our data at least as well as the Weibull in every case.

2.4.2 Generalized Beta Of The Second Kind

In this section, we will introduce the Generalized Beta of the Second Kind, or $GB2$. We noted that the Majumder and Chakravarty [8] model was actually shown to be a re-parameterization of the $GB2$. McDonald and Mantrala [11] define the four parameter $GB2$ pdf by the expression

$$f(x; a, b, p, q) = \frac{a x^{ap-1}}{b^{ap} B(p, q)(1 + (x/b)^a)^{p+q}}$$

for $x \geq 0$ and $B(\cdot, \cdot)$ is the Beta function.

According to the authors of the `VGAM` library [23] for the software `R` [16], the $GB2$ has a mean of

$$E(X) = \frac{b\Gamma(p + 1/a)\Gamma(q - 1/a)}{\Gamma(p)\Gamma(q)}$$

when $-ap < 1 < aq$.

2.4.3 Relationship to other three distributions

McDonald [12] has a figure that shows how all our distributions are actually special cases of the $GB2$. Starting with the four paramter $GB2$ distribution, we can get the three parameter Singh-Maddala distribution by setting $p = 1$.

We can derive our other two parameter distributions as well. We must first introduce an intermediate distribution called Generalized Gamma (GG) which can be found by letting $q \to \infty$ and $b = q^{1/a}\beta$ in the GB2. Thus, the pdf for the GG is

$$f(x; a, \beta, p) = \frac{a x^{ap-1} e^{-(x/\beta)^a}}{\beta^{ap}\Gamma(p)}$$

for $x \geq 0$. From this equation for GG, we can get the lognormal distribution by letting $\beta^a = \sigma^2 a^2$ and $p = (a\mu + 1)\beta^{-a}$. To get the gamma distribution, we let $a = 1$, and to get the Weibull distribution we let $p = 1$.

Since the three distributions lognormal, gamma, and Weibull are all special cases of the GB2, we would expect GB2 to always fit an income size distribution at least as well as those. But consideration must be given to the difference between having four parameters and the simpler, more parsimonious distributions which have only two parameters.

2.5 Estimating The Parameters

We use the `R` function `fitdistr()` to fit the Lognormal, Gamma and Weibull density models to our data. To fit the Singh-Maddala model, we use the `VGAM` library in `R`. The values for the estimated parameters for each distribution are in Table 2.2 and Table 2.3. We can look at the standard errors for the estimates of the parameters of each distribution in Table 2.4 and Table 2.5.

Table 2.2: Parameters for each county or state and each distribution

County/State	N	ln.mean	ln.sd	g.scl	g.shp
Los Angeles	6243	10.66	0.49	11277	4.28
San Francisco	572	10.95	0.45	10676	5.83
Denver	468	10.60	0.44	8154	5.39
District of Columbia	426	10.61	0.60	15884	3.03
Fulton	443	10.58	0.72	22838	2.19
Cook	4171	10.67	0.51	10908	4.43
Suffolk	629	10.53	0.48	8209	5.05
Baltimore City	701	10.26	0.47	6808	4.58
New York	845	10.63	0.68	20771	2.48
Multnomah	508	10.64	0.39	6399	7.03
Philadelphia	1173	10.22	0.51	7252	4.28
Dallas	1665	10.67	0.51	12165	4.01
Richmond City	162	10.29	0.51	8328	4.02
Georgia	4772	10.51	0.50	9512	4.34
Massachusetts	5024	10.79	0.47	10519	5.09
Maryland	3642	10.81	0.51	12697	4.39
New Jersey	6428	10.89	0.48	12136	4.93
Virginia	4717	10.69	0.50	12188	4.09
Kansas	2287	10.55	0.39	6476	6.40
Utah	1472	10.71	0.37	6441	7.45

Table 2.3: Parameters for each county or state and each distribution, continued

County/State	w.shp	w.scl	sm.log(a)	sm.log(scl)	sm.log(q)
Los Angeles	2.05	54205	1.20	10.75	0.19
San Francisco	2.62	70671	1.27	11.30	0.88
Denver	2.27	49322	1.44	10.58	-0.04
District of Columbia	1.70	53879	1.12	10.60	-0.02
Fulton	1.48	55273	0.80	10.78	0.30
Cook	2.18	54812	1.12	11.01	0.72
Suffolk	2.46	47087	1.09	11.06	1.15
Baltimore City	2.03	35760	1.22	10.45	0.43
New York	1.62	57347	0.67	11.41	1.13
Multnomah	2.55	50278	1.63	10.61	-0.11
Philadelphia	2.09	35093	1.07	10.61	0.80
Dallas	1.93	54717	1.29	10.63	-0.08
Richmond City	1.90	37507	1.27	10.31	0.03
Georgia	2.04	46325	1.24	10.58	0.16
Massachusetts	2.36	60542	1.17	11.12	0.75
Maryland	2.17	62976	1.08	11.17	0.73
New Jersey	2.32	67548	1.16	11.19	0.65
Virginia	1.99	55868	1.31	10.62	-0.15
Kansas	2.28	46085	1.73	10.41	-0.45

Table 2.4: Standard errors for estimated parameters

County/State	ln.mean	ln.sd	g.scl	g.shp
Los Angeles	0.0063	0.0044	203	0.0729
San Francisco	0.0189	0.0134	631	0.3302
Denver	0.0204	0.0145	522	0.3302
District of Columbia	0.0290	0.0205	1221	0.2119
Fulton	0.0343	0.0242	2937	0.2320
Cook	0.0079	0.0056	264	0.1002
Suffolk	0.0191	0.0135	477	0.2791
Baltimore City	0.0179	0.0127	367	0.2393
New York	0.0232	0.0164	1301	0.1356
Multnomah	0.0171	0.0121	411	0.4350
Philadelphia	0.0122	0.0086	246	0.1372
Dallas	0.0125	0.0089	419	0.1302
Richmond City	0.0398	0.0282	909	0.4143
Georgia	0.0072	0.0051	197	0.0848
Massachusetts	0.0066	0.0047	206	0.0952
Maryland	0.0084	0.0059	317	0.1030
New Jersey	0.0059	0.0042	242	0.0926
Virginia	0.0073	0.0052	264	0.0831
Kansas	0.0082	0.0058	193	0.1838
Utah	0.0097	0.0069	246	0.2746

Table 2.5: Standard errors for estimated parameters, continued

County/State	w.shp	w.scl	sm.log(a)	sm.log(scl)	sm.log(q)
Los Angeles	0.0184	320	0.0187	0.0232	0.500
San Francisco	0.0789	1269	0.0570	0.0972	0.2405
Denver	0.0733	993	0.0710	0.0623	0.1661
District of Columbia	0.0575	1261	0.0741	0.0904	0.1754
Fulton	0.0514	1551	0.0690	0.1365	0.1975
Cook	0.0245	458	0.0214	0.0388	0.0807
Suffolk	0.0727	727	0.0532	0.1291	0.2770
Baltimore City	0.0510	706	0.0540	0.0745	0.1671
New York	0.0415	1099	0.0460	0.1686	0.2359
Multnomah	0.0780	912	0.0690	0.0484	0.1557
Philadelphia	0.0347	389	0.0326	0.0648	0.1302
Dallas	0.0372	1250	0.0379	0.0377	0.0869
Richmond City	0.1031	1704	0.1192	0.1282	0.2899
Georgia	0.0208	355	0.0215	0.0252	0.0563
Massachusetts	0.0248	463	0.0195	0.0340	0.0748
Maryland	0.0265	561	0.0229	0.0432	0.0871
New Jersey	0.0212	939	0.0174	0.0289	0.0624
Virginia	0.0210	463	0.0229	0.0215	0.0503
Kansas	0.0319	424	0.0350	0.0191	0.0668
Utah	0.0510	563	0.0386	0.0354	0.1019

2.6 QQ Plots

The QQ plot allows us to see how closely our median household income data fits our distribution models. We use the estimated parameters from Table 2.2 and compare each of the four distribution models to the Census household income data.

Let us look at the QQ plots of each distribution in Figure 2.7 and Figure 2.8. This is one of the first diagnostics we use to figure out which distribution we want to use in our models. We pick the distribution that follows the red line $y = x$ the closest.

In Figure 2.7, we can choose the Singh-Maddala distribution. In Figure 2.8, our choice is a little harder. We can eliminate the Gamma and Weibull distributions, and we recommend more investigation before choosing between Lognormal and Singh-Maddala.

2.7 Error Measures

Now we must also make use of some statistics, since relying on graphs and other pictures isn't always the most accurate way of determining a good fit. First let us look at the statistic Mean Squared Error, where MSE is defined as

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

and Y_i are our observed quantiles and $\hat{Y}_i$ are the corresponding estimated values, obtained using the parameters from Tables 2.2 and 2.3. The term $(Y_i - \hat{Y}_i)$ is considered our error. Refer to Table 2.6 to see the results. We have taken the square root of our mean squared error in order to improve readability. We see that

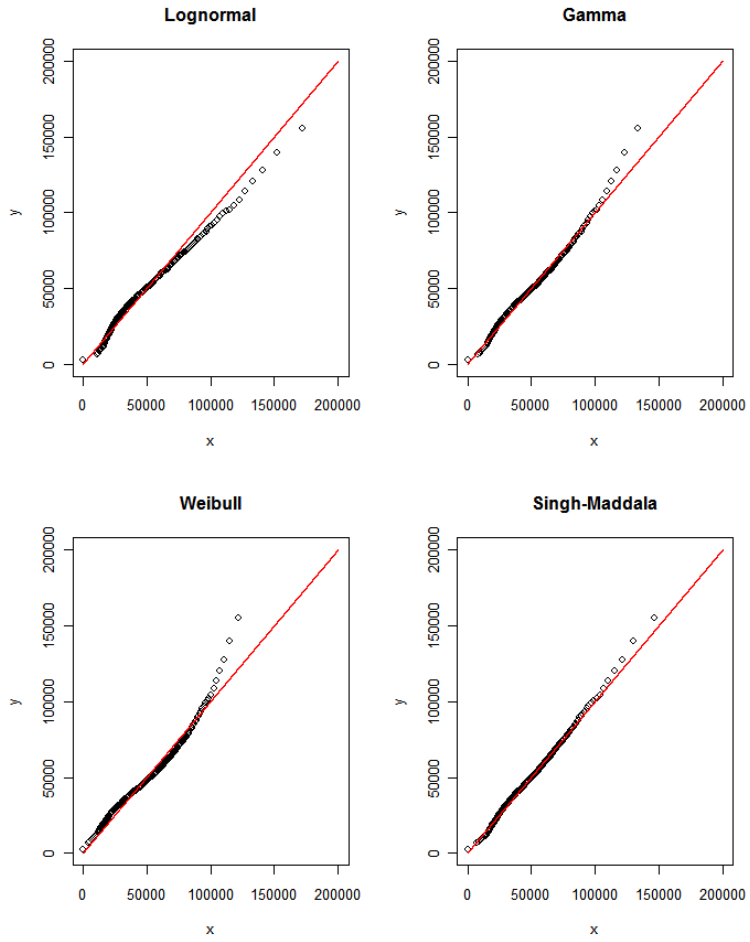

Figure 2.7: QQ Plots of 4 Distributions, Cook County

Singh-Maddala is the best model to choose most of the time, with 10 counties or states having Singh-Maddala as the best fit model. Lognormal is the next best with 7, and Gamma is chosen 3 times. Weibull is never the best fitting model.

Also, we can look at the Mean Absolute Errors. They are defined by the formula

$$\frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|,$$

where Y_i are our observed quantiles and $\hat{Y}_i$ are the corresponding estimated values, obtained using the parameters from Tables 2.2 and 2.3. Look at Table 2.7 to see

Table 2.6: Square root of the Mean Squared Error

County/State	Lognormal	Gamma	Weibull	Singh-Maddala
Los Angeles	1295.79	3299.68	4985.75	5046.90
San Francisco	7025.52	4796.05	6097.75	3924.31
Denver	2099.19	3841.60	5278.01	2465.58
District of Columbia	2777.29	6596.33	7615.75	6140.72
Fulton	7344.58	5315.55	6022.01	10804.72
Cook	4351.72	2989.59	4852.30	1548.53
Suffolk	4218.00	1174.32	2143.22	981.89
Baltimore City	3937.69	5216.36	5999.05	3605.28
New York	8637.42	3331.11	4132.38	4205.14
Multnomah	2506.75	3880.01	5508.92	1719.07
Philadelphia	2652.03	2871.53	3820.77	1872.08
Dallas	2151.36	5449.13	6863.01	6100.50
Richmond City	4001.82	5684.01	6278.84	2757.36
Georgia	1209.30	3732.73	5152.00	2848.62
Massachusetts	3902.16	1868.62	3993.05	1028.03
Maryland	4188.71	2174.44	4444.04	1467.73
New Jersey	4299.70	1496.85	4142.29	2327.27
Virginia	2301.33	4087.44	5583.02	11283.97
Kansas	3553.73	5047.00	6355.15	3037.27
Utah	669.98	1867.53	3783.88	2118.59

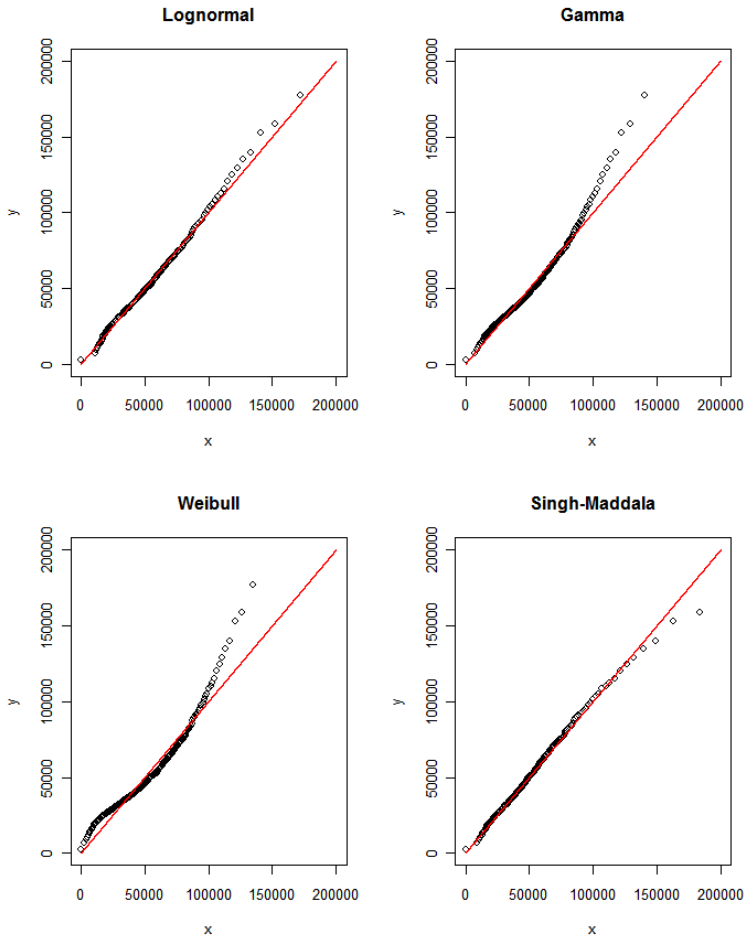

Figure 2.8: QQ Plots of 4 Distributions, Dallas County

the values of Mean Absolute Error. Once again, Singh-Maddala usually is the best fit for our data, having the least error 14 of our 20 counties or states.

Table 2.7: Mean Absolute Errors

County/State	Lognormal	Gamma	Weibull	Singh-Maddala
Los Angeles	394	1840	3347	1101
San Francisco	5799	4090	4652	2766
Denver	1469	2272	3952	1203
District of Columbia	1854	3981	5328	1923
Fulton	1992	3702	4428	2870
Cook	3202	1943	3345	997
Suffolk	2563	930	1712	711
Baltimore City	1780	1637	3231	1029
New York	4081	2628	2846	2786
Multnomah	1889	2445	4244	1039
Philadelphia	1805	1085	1995	707
Dallas	1470	3133	4880	1207
Richmond City	1574	2371	3641	1150
Georgia	947	1949	3524	646
Massachusetts	2627	1292	2840	547
Maryland	2485	980	2942	432
New Jersey	2250	1023	3111	614
Virginia	1274	2879	4334	2207
Kansas	1801	2737	4521	786
Utah	398	1080	2701	852

2.8 Measures of Goodness Of Fit and Distribution Selection

2.8.1 Chi Squared Test

Now let us look at how the distributions compare using Chi Squared values.

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i are our observed counts and E_i represent our expected counts, which we computed using our estimated parameters from Table 2.2. The value k is the number of bins we use in our histogram. In this case, we let $k = 6$, and the breaks are at 0, 20000, 40000, 60000, 80000, 100000 and 200000.

The smaller the chi squared value, the better the model fit. In Table 2.8, we see that Singh-Maddala is the best fitting distribution model 16 out of 20 times.

2.8.2 Akaike Information Criterion

Now let us look at the measure Akaike Information Criterion, or AIC. The AIC is a measure that is used for distribution selection. The distribution with the lowest AIC is the distribution we want to select. The AIC is defined by

$$AIC = -2\ln(L) + 2k$$

where k is the number of parameters in our model, and $\ln(L)$ is the maximized log-likelihood value.

Refer to Table 2.9 to see the AIC values. AIC helps us select the Singh-Maddala distribution 15 times, and lognormal is selected the other five times.

Table 2.8: Chi-squared values for each county or state and distribution

County/State	Lognormal	Gamma	Weibull	Singh-Maddala
Los Angeles	12.44	178.97	788.40	16.02
San Francisco	5144.18	64.62	93.95	46.40
Denver	22.07	44.32	419.41	0.91
District of Columbia	17.31	43.91	78.36	12.69
Fulton	6.83	22.02	34.26	11.43
Cook	33193.07	90.69	432.68	28.14
Suffolk	3009.26	0.99	13.48	1.48
Baltimore City	47.20	2019.22	102295.90	17.67
New York	62.01	17.99	25.27	19.62
Multnomah	136.97	19.06	69.21	4.73
Philadelphia	3241.18	292.99	167219.30	16.64
Dallas	17.28	106.25	275.25	8.35
Richmond City	1357.98	138.07	1222.92	4.88
Georgia	131478.80	230.26	1353.51	10.98
Massachusetts	215.72	71.17	329.15	18.62
Maryland	425.36	19.93	182.57	2.67
New Jersey	2391.94	28.37	300.07	10.33
Virginia	492.37	320.07	795.62	76.18
Kansas	578.07	287.79	596.65	18.48
Utah	290.79	20.27	137.29	15.79

Table 2.9: AIC values for each county or state and distribution

County/State	Lognormal	Gamma	Weibull	Singh-Maddala
Los Angeles	142118.8	142274.4	142961.6	142175.0
San Francisco	13245.89	13178.63	13187.00	13139.97
Denver	10485.45	10489.74	10551.92	10467.90
District of Columbia	9814.028	9825.554	9867.802	9802.96
Fulton	10342.05	10350.16	10373.95	10346.08
Cook	95236.22	94939.83	95181.16	94808.38
Suffolk	14112.36	14059.34	14068.10	14051.84
Baltimore City	15329.56	15343.59	15456.39	15303.40
New York	19711.07	19719.68	19756.10	19739.38
Multnomah	11290.10	11290.44	11371.53	11250.44
Philadelphia	38910.44	38840.75	38983.82	38801.58
Dallas	38015.62	38071.12	38286.66	37968.00
Richmond City	3577.563	3585.53	3609.473	3575.632
Georgia	107123.5	107194.3	107769.2	107007.7
Massachusetts	115065.5	114800.6	115088.1	114724.8
Maryland	84110.1	83965.78	84193.1	83940.35
New Jersey	148763.9	148476.2	148868.9	148435.2
Virginia	107788.7	107977.4	108529.2	107798.7
Kansas	50460.9	50632.4	51148.1	50346.7
Utah	32804.4	32822.3	33009.3	32819.2

2.8.3 Bayesian Information Criterion

Finally, we can look at the measure Bayesian Information Criterion, or BIC. The BIC is another measure that is used for distribution selection. The BIC is similar to the AIC, but it has an extra penalty for the number of observations in the dataset being studied. The distribution with the lowest BIC is the distribution we want to select. The BIC is defined by

$$BIC = -2\ln(L) + k\ln(n)$$

where k is the number of parameters in our model, $\ln(L)$ is the log-likelihood value, and n is the number of observations in our dataset.

Refer to Table 2.10 to see the BIC values. The BIC helps us select the Singh-Maddala distribution 14 times, and lognormal is selected the other six times.

Table 2.10: BIC values for each county or state and distribution

County/State	Lognormal	Gamma	Weibull	Singh-Maddala
Los Angeles	142132.3	142287.9	142975.1	142195.2
San Francisco	13254.59	13187.33	13195.70	13153.02
Denver	10493.75	10498.03	10560.22	10480.34
District of Columbia	9822.136	9833.663	9875.911	9815.123
Fulton	10350.23	10358.35	10382.14	10358.36
Cook	95248.9	94952.5	95193.84	94827.39
Suffolk	14121.25	14068.22	14076.99	14065.17
Baltimore City	15338.67	15352.69	15465.49	15317.05
New York	19720.55	19729.16	19765.58	19753.59
Multnomah	11298.57	11298.90	11379.99	11263.13
Philadelphia	38921.4	38851.71	38994.78	38818.02
Dallas	38026.45	38081.96	38297.50	37984.25
Richmond City	3583.738	3591.705	3615.648	3584.895
Georgia	107136.5	107207.2	107782.2	107027.1
Massachusetts	115078.5	114813.7	115101.2	114744.4
Maryland	84122.5	83978.18	84205.5	83958.95
New Jersey	148777.5	148489.7	148882.4	148455.5
Virginia	107801.6	107990.4	108542.1	107818.0
Kansas	50472.4	50643.9	51159.6	50363.9
Utah	32814.9	32832.9	33019.8	32835.1

Chapter 3

Conclusions

We have studied four different density functions, Lognormal, Gamma, Weibull and the Singh-Maddala. We used a graphical technique called QQ Plots in order to choose between our distributions. We also studied several measures of fit and distribution selection:

- Mean Squared Errors
- Mean Absolute Errors
- Chi Squared Values
- AIC
- BIC

There does not exist a clear winner among the three distributions with two parameters. When we include the three parameter Singh-Maddala, there is a clear winner. In each of these measures, almost every county or state takes the Singh-Maddala as the best fit. The Singh-Maddala is a good density to use for several reasons. It only has three parameters, one more than our other densities, but not so many more that we would be accused of over-fitting.

Also, there is readily available software to deal with the Singh-Maddala distribution. In the literature, especially by McDonald [12], it is asserted that the

Generalized Beta II is an even better model for income data. This would be expected, since Singh-Maddala has been shown to be a special case of the $GB2$. The only problem with using $GB2$ to estimate a distribution for our data is that there is not much functionality in software packages like `R` to deal with the $GB2$. This is definitely something to consider when choosing the proper density function to fit to our data.

3.1 New York City

We notice that New York County never fits well against the Singh-Maddala density. We see in Figure 1.5 that the histogram for New York County does not look like all the other county and state histograms. New York County is coextensive with the Borough of Manhattan. New York City itself is composed of five boroughs: Manhattan, Bronx, Queens, Brooklyn, and Staten Island.

In all the other counties that we tested, each county contained one large United States city. In the case of New York County, we are actually only seeing one subdivision of the much larger New York City.

When we only tested New York County, we noticed that Singh-Maddala was not the best fitting distribution. This might be because Manhattan is so expensive to live in compared to the rest of the counties we considered, and therefore median household income per block group is higher. When we take into account all five boroughs, the data are more in line with what we expect, which is that the Singh-Maddala gives the best fit. Figure 3.1 compares the histogram and mean of our

data for New York County only versus all of New York City.

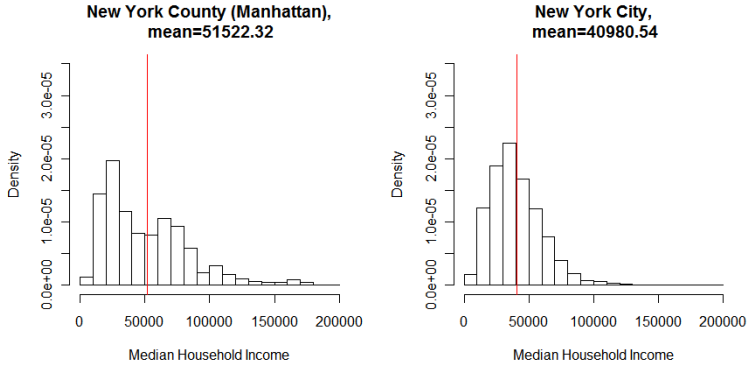

Figure 3.1: Comparing New York County and New York City histogram and mean

We can look at the QQ plots of New York County and all of New York City when fitting the Singh-Maddala model using estimated parameters. In Figure 3.2, we can see that the Singh-Maddala fits the data better for the whole of New York City.

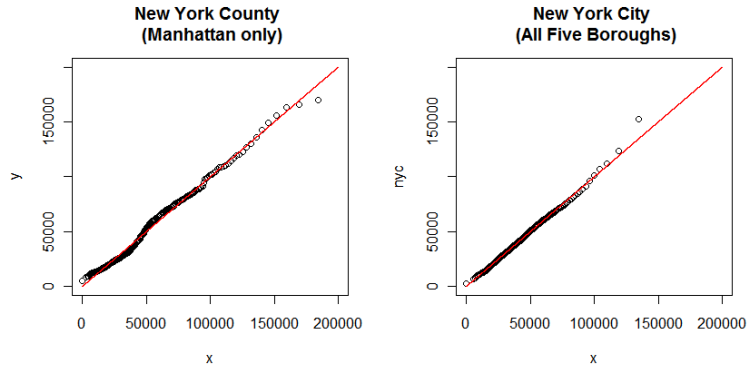

Figure 3.2: Comparing New York County and New York City QQ plots using the Singh-Maddala

Table 3.1 has a summary of our measures now that we have included all five boroughs of New York City. As in Table 2.6, we have taken the square root of the

Table 3.1: Measures For All Five New York City Boroughs, $N = 5597$.

Measure	Lognormal	Gamma	Weibull	Singh-Maddala
Mean Square Error	3435.5	2883.0	4316.0	1393.7
Mean Absolute Error	2109	931	2299	506
Chi Squared	107.23	31.22	177.40	20.44
AIC	126359.1	126162.1	126550.5	126098.6
BIC	126372.4	126175.3	126563.7	126118.5

mean squared error in order to improve readability.

McDonald [12] writes that "it is interesting to note that the Singh-Maddala distribution function provides a better fit to the data than any of the distribution functions except for the $GB2$ regardless of the criterion used for comparison". While Singh-Maddala was not the best fit for every one of our counties or states, we agree with this McDonald's statement.

For the counties and states for which we have investigated median household income by block group, and under our assumptions, including the truncating of our data at the level $x \leq 200000$, we recommend that Singh-Maddala is the best choice to fit to the income data.

3.2 Future Research

It would be interesting to perform likelihood ratio tests using the Generalized Beta II or other high parameter distributions. This has the potential of reducing

the distribution to a distribution with fewer parameters.

McDonald and Butler [10] suggest using mixture distributions to fit models to income distribution. This seems appropriate as some of our histograms appear to have two modes. Also of interest is the work of Drăgulescu and Yakovenko [5], who use ideas from physics and thermodynamics to argue that the distribution of income follows a two part regime, with exponential distribution for the low income values and a power-law tail for high income values.

Bibliography

[1] U.S. Census Bureau. *2000 Census of Population and Housing, Summary File 3: Technical Documentation*, 2002.

[2] D. G. Champernowne. A model of income distribution. *The Economic Journal*, 63(250):318–351, 1953.

[3] D.G. Champernowne. The graduation of income distributions. *Econometrica*, 20(4):591–615, 1952.

[4] J. S. Cramer. A function for size distribution of incomes: Comment. *Econometrica*, 46(2):459–460, 1978.

[5] A. Drăgulescu and V. M. Yakovenko. Exponential and power-law probability distributions of wealth and income in the united kingdom and the united states. *Physica A*, 299:213–221, 2001.

[6] L. Fonseca and J. Tayman. Postcensal estimates of household income distributions. *Demography*, 26(1):149–159, 1989.

[7] S. Kotz and N. L. Johnson, editors. *Encyclopedia of Statistical Sciences*, volume 5. Wiley, New York, 1981.

[8] A. Majumder and S. R. Chakravarty. Distribution of personal income: Development of a new model and its application to U.S. income data. *Journal of Applied Econometrics*, 5(2):189–196, 1990.

[9] B. Mandelbrot. The pareto-levy law and the distribution of income. *International Economic Review*, 1(2):79–106, 1960.

[10] J. B. McDonald and R. J. Butler. Some generalized mixture distributions with an application to unemployment duration. *The Review of Economics and Statistics*, 69(2):232–240, 1987.

[11] J. B. McDonald and A. Mantrala. The distribution of personal income: Revisited. *Journal of Applied Econometrics*, 10(2):201–204, 1995.

[12] J.B. McDonald. Some generalized functions for the size distribution of income. *Econometrica*, 52(3):647–665, 1984.

[13] J.B. McDonald and M.R. Ransom. Functional forms, estimation techniques and the distribution of income. *Econometrica*, 47(6):1513–1525, 1976.

[14] H. L. Moore. The statistical complement of pure economics. *Quarterly Journal of Economics*, 23:1–33, 1908.

[15] J. Persky. Retrospectives: Pareto's law. *The Journal of Economic Perspectives*, 6(2):181–192, 1992.

[16] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[17] R.S.G. Rutherford. Income distributions: A new model. *Econometrica*, 23(3):277–294, 1955.

[18] A. B. Z. Salem and T. D. Mount. A convenient descriptive model of income distribution: The gamma density. *Econometrica*, 42(6):1115–1127, 1974.

[19] Stanley L. Sclove. (Y vs. X) or (log Y vs. X)? *Technometrics*, 14(2):391–403, 1972.

[20] S.K. Singh and G.S. Maddala. A function for size distribution of incomes. *Econometrica*, 44(5):963–970, 1976.

[21] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[22] M. Victoria-Feser and E. Ronchetti. Robust methods for personal-income distribution models. *The Canadian Journal of Statistics*, 22(2):247–258, 1994.

[23] Thomas W. Yee. *VGAM: Vector Generalized Linear and Additive Models*, 2008. R package version 0.7-7.