On the Approximation of Pareto Distribution to Exponential Distribution Using the Gini Coefficient of Inequality

W. B. Yahya[1]; M. K. Garba[2]; L. Amidu[3]; K. O. Olorede[4]; N. F. Gatta[5]; L. B. Amusa[6]

Department of Statistics,
University of Ilorin,
Ilorin, Nigeria.
e-mail: wbyahya@unilorin.edu.ng[1]; lateef_amidu@yahoo.com[3]

[4]Department of Statistics and Mathematical Sciences,
Kwara State University,
Malete, Nigeria.

Abstract **— Pareto proposed that income and wealth distribution obeys a universal power law valid for all times and countries, but subsequent studies have often disputed this position. Some even argued there is indeed no Pareto Law and that it should be entirely discarded in studies on distribution of wealth or resources. Many other probability distributions have been proposed such as log normal, exponential, gamma and two other forms by Pareto himself. Using data on imported goods from the National Bureau of Statistics as a case of distribution of wealth in Nigeria, we demonstrated that the distribution of money spent on importation in Nigeria also follow exponential distribution using the Gini coefficient which is a measure of inequality (degree of concentration) of a variable in the distribution of resources. Simulation studies were carried out at different sizes of items (or households) and varying values of the shape parameter and we compare how close the Gini coefficients of the exponential distribution approximate those obtained from the Pareto data as a credible alternative to Pareto distribution.**

Keywords-- *Pareto distribution, Gini coefficient, Lorenz curve, Exponential distribution.*

I. Introduction

The Pareto distribution named after the Italian civil engineer and economist Vilfredo Pareto is a power law distribution that is used in description of scientific, social, geographical, actuarial, and many other types of situations. There are various forms of Pareto distribution, one of which is the two-parameter Pareto distribution with two parameters, namely, the scale parameter (σ) and the shape parameter (α) and another one is the generalized Pareto distribution. They are one of the continuous probability distributions. This study focuses on the two-parameter Pareto distribution.

Various economic data which are positively skewed are assumed to follow the Pareto distribution and past researches have proposed that there are alternative distributions, for modeling wealth and income data, which are also positively skewed in nature.

Estimation of the shape parameter, α of the Pareto distribution characterizing the tail, with the scale parameter σ assumed known reduces it to a one parameter Pareto distribution following Bierlant, Teugels, and Vynckier (1996) and Klugman, Panjer, and Willmot (1998).

This study therefore examined the place of exponential distribution as a credible alternative to Pareto distribution to model economic or wealth data that originally follow Pareto distribution. In a Monte-Carlo study, this situation was investigated under different sample sizes of items (or households) and at varying values of the shape parameters of the two distributions using their Gini coefficients as assessment criteria. To further validate the results from the simulation study, real life data on expenditure or amount of money spent on imported items in Nigeria were employed to demonstrate the applications of the concepts treated here.

II. Materials and Methods

A. Simulation Design

The scheme used for the simulation studies are as follows: The simulations were done by setting a value for the shape parameter (α) and the value of the scale parameter is fixed at σ = 1(because it only implies the minimum value of any

unit of interest under the distribution of wealth). A sample of size n Pareto (1,α) data is selected and an iteration of 100000 is used to ensure stability in results, from which the Gini coefficient is calculated. The mean of those 100000 estimates is computed which will represent the estimates of G. The sample size n is made to vary over the points 20, 50,100,150,200,250,500,750, 1000, 2000, and 5000. This process was repeated for various values of α which ranges from 1.1 to 2.0 using the same estimation scheme.

B. Real life Data

In order to validate the results from simulation study, data on importation of items were collected from the Nigeria Bureau of Statistics website from year 2007 to 2010. Nigeria had imported goods from other countries in 2007 with an amount of ₦4127690.00 (in millions of naira) which is classified into 21 sections including agricultural products, non-agricultural products, mineral products, etc. It is noted that variations in earnings varies from as little as ₦5.60 to ₦911144.40 (in millions of naira) and similar classification has been done for the year 2008, 2009 and 2010.

The import of items from different sectors is very important for a country like Nigeria due to the limited amount of goods manufactured within the country. The expenditure on imported items is an important factor that influences the Balance of Payments (BOP) of a country. A good import policy provides growth in industries, agricultural production, livestock, etc. and provides employment to people. In this study, we examined the distribution of expenditure on imported items in Nigeria for the year 2007 to 2010 by applying theoretical concepts and applications.

It was found that the variable of interest here: expenditure on imported goods (X, in Naira) follows an exponential distribution which is a highly positive skewed distribution. The exponential distribution gives a good, even though, not perfect description of the import expenditure data. We also examined associated measures of inequalities, Lorenz curve and Gini's coefficient of inequality in context of wealth inequality. The exponential distribution is a decay model, and shows high inequality, which is a good indication for import policy

Following Bierlant, Teugels, and Vynckier (1996) and Klugman, Panjer, and Willmot (1998), estimation of the shape parameter, α of the Pareto distribution characterizing the tail with the scale parameter σ assumed known, reduces it to a one parameter Pareto distribution which is employed in this work. An estimator of the scale parameter, σ using the maximum likelihood estimation is the minimum value of the sample observations.

i.) The Pareto Distribution

If X follows the Pareto(α,σ) distribution, then the probability density function is:

$$f(x) = \frac{\alpha\sigma^{\alpha}}{x^{\alpha+1}}, x > \sigma, \sigma > 0, \alpha > 0 \quad (1)$$

The maximum likelihood estimate of α is:

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n}(Inx_i - In\sigma)} \quad (2)$$

while the maximum likelihood estimator of the scale parameter σ of the Pareto distribution is the minimum value of all the observations/data in the selected sample.

ii.) The Exponential Distribution

If a random variable X follows the exponential distribution, the probability density function of X is given by:

$$f(x) = \frac{1}{\lambda}e^{-\frac{x}{\lambda}}, x > 0 \quad (3)$$

This is a positively skewed distribution with the cumulative distribution function given by:

$$F(x) = 1 - e^{-\frac{x}{\lambda}} \quad (4)$$

iii.) The Lorenz Curve

The Lorenz Curve is a graphical device used to demonstrate the equity of distribution of a given variable such as income, wealth, assets etc. A Lorenz curve provides complete information on the whole distribution of resources (e.g. income, expenditure, wealth, etc.) relative to mean. The horizontal axis of the Lorenz curve, F represents the cumulative fraction of Population with amount spent below r, and the vertical axis L (F) represents the fraction of expenditure this population accounts for.

Let $F = \int_0^{\infty} xf(x)dx$ and $L(F) = \frac{1}{\mu}\int_0^{x(F)} xf(x)dx$ The Lorenz curve for the exponential distribution is:

$$L(F) = (1-F)\,In(1-F) - (1-F) + 1$$

$$L(F) = F + (1-F)\,In(1-F) \quad , 0 \le F \le 1.$$

The Lorenz curve for the Pareto density function is:

$$L(F) = 1 - (1-F)^{1-\frac{1}{\alpha}} \quad (5)$$

iv.) The Gini Coefficient

The Gini coefficient (also known Lorenz concentration ratio) is a measure of inequality of a variable in a distribution of its elements and it has a scale of 0 to 1.

Geometrically, it is one minus twice the area between the Lorenz curve and the identity function (equal distribution). A value of G = 0.0 means that there is no inequality (total equality, i.e. everyone has same amount) and when G = 1.0 it shows that there is complete inequality (no equality, i.e. one person has everything).

A consistent estimator of the Gini index for a population with values y_i, where i ranges from 1 to n, that are indexed in non-decreasing order ($y_i \leq y_{i+1}$) is given as:

$$G = \frac{1}{n}(n+1) - 2\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i} \qquad (6)$$

According to Francis (2010), the algebraic formula for the Gini index obtained by linear interpolation of data as used by the CIA world fact book in calculating the income inequality of various countries in 2006 is given as:

$$G = \left[\frac{1}{T}\sum_{j=1}^{n} x_j(p_j + p_{j-1})h_j\right] - 1 \qquad (7)$$

where n is the size of the dataset or population and T is the sum or total of the whole observations and we formulate a table such that there are h_j units (sections or items) that, on the average, have an amount x_j of our resource of interest. If the table has n rows, then j ranges from 1 to n and the order of the table means that $x_j < x_k$ when $j < k$,

$$n = \sum_{i=1}^{n} h_i \text{ and } T = \sum_{i=1}^{n} x_i h_i \qquad (8)$$

Using this notation, the mean amount owned is T/n, which we call X. The number

$$P_j = \sum_{i=1}^{n} hi \,/\, n \qquad (9)$$

gives us n (not necessarily equally spaced) points along the x-axis between 0 and 1. For convenience, we define $p_0 = 0$..

The Gini coefficient for the exponential distribution is derived as follows:

$$G = 1 - 2\int_0^1 L(F)\,dF$$

$$G = 1 - 2\int_0^1 (F + (1-F)In(1-F))\,dF$$

$$G = 1 - 2\left(\frac{1}{4}\right)$$

$$G = 1 - \frac{1}{2}$$

$$G = \frac{1}{2}$$

This is independent of the parameter λ of the exponential distribution.

The Gini coefficient for the Pareto distribution is:

$$G = 1 - 2\int_0^1 L(F)\,dF$$

$$G = 1 - 2\left[F + \frac{(1-F)^{2-\frac{1}{\alpha}}}{2-\frac{1}{\alpha}}\right]_0^1$$

$$G = 1 - 2\left[1 - \frac{1}{2-\frac{1}{\alpha}}\right]$$

$$G = \frac{1}{2\alpha - 1} \qquad (10)$$

Mathematically, it can be seen that when α =1.5, the Gini coefficient for the Pareto distribution will be equal to that of the exponential distribution. But in real life data, this mathematical equivalence may not be true as a result of varying number of items or categories. We therefore examine the deviation of calculated G from the simulated Pareto data to that of the exponential distribution which is independent of the parameter (that is, G = 0.5) for values of α revolving around 1.5 and at different number of items or objects (sample sizes).

III. Analysis

The scheme used for the simulation studies are as follows: The simulations were done by setting a value for the shape parameter (α) and the value of the scale parameter is fixed at σ = 1 (because it only implies the minimum value of any unit of interest under the distribution of wealth). A sample of size n Pareto (1,α) data is selected and an iteration of 100000 is used to ensure stability in results, from which the Gini coefficient is calculated. The mean of those 100000 estimates is computed which will represent the estimates of G. The sample size n is made to vary over the points 20, 50,100,150,200,250,500,750, 1000, 2000, and 5000. This process was repeated for various values of α which ranges from 1.1 to 2.0 using the same estimation scheme.

In order to validate the results, data on importation of items were collected from the Nigeria Bureau of Statistics website from year 2007 to 2010. Nigeria had imported goods from other countries in 2007 with an amount of ₦4127690.00 (in millions of naira) which is classified into 21 sections including agricultural products, non-agricultural products, mineral products, etc. It is noted that variations in earnings varies from as little as ₦5.60 to ₦911144.40 (in millions of naira) and similar classification has been done for the year 2008, 2009 and 2010.

IV. Results

Table 3 presents the results from the simulation studies carried out in this research work at various sample sizes and values of the shape parameter. After that, we test to see if the import data are exponentially distributed using the Kolmogorov-Smirnov test and the results is displayed in Table 1.

Table 1: Kolmogorov-Smirnov Test

K-S Test	*YEAR*			
Kolmogorov Smirnov Z	**2007**	**2008**	**2009**	**2010**
P-Value	0.099	0.189	0.153	0.259

Table 2: Summary table of Gini coefficients for amount spent on imported items.

Year	MLE	Gini Estimation	Gini using Interpolation	Gini for Exponential distribution
2007	196556.6667	0.6178	0.56	0.5
2008	157099.8386	0.6551	0.54	0.5
2009	240372.9048	0.6671	0.54	0.5
2010	316596.471	0.6464	0.56	0.5

V. Discusion

From the simulation studies results in Table 3, n is used to represent the number of items or objects or sample size and α denotes the shape parameter of the Pareto distribution and it was observed that:

- For $n < 50$ and $1.0 \leq \alpha \leq 1.3$, the exponential distribution gives a good, though not perfect approximation to the Pareto distribution.
- For $50 \leq n \leq 1000$ and $1.35 \leq \alpha \leq 1.45$, the exponential distribution gives a very good approximation to the Pareto distribution.
- At approximately $\alpha = 1.5$, it requires a very large number of items or objects (say at least 750 or 1000) before the exponential distribution can provide a perfect approximation to the Pareto distribution.
- For $\alpha \geq 1.6$, irrespective of the value of n, the exponential distribution does not provide a good fit to a Pareto data.

After collecting data for validation of results, we test to see if the import data are exponentially distributed using the Kolmogorov-Smirnov test and the results is displayed in Table 1. Since all the p-values are greater than 0.05, we conclude that the data are exponentially distributed and we further proceed to compute the Gini coefficients of the data for the four years in study. The results are displayed in Table 2.

It can be observed from Table 2 that the Gini Coefficients of inequalities obtained from the data gives a good (albeit not perfect) approximation of the exponential distribution since the computed values of G are close to 0.5 which is the Gini coefficient for the exponential distribution. This is an indication that economic data can also be assumed to follow an exponential distribution as presented in this study.

The Line graphsin Figure 1,2 and 3 show the estimated Gini coefficients of Pareto distribution in Table 3 against the various sample sizes at some values of alpha for better comparison to that of exponential distribution (0.5).

Table 3: Table of estimated Gini Coefficients for Simulated Economic Data for Pareto distribution at different values of the shape parameters (α) and at various sample sizes

SAMPLE SIZES	**Value of the shape parameter (α) of Pareto distribution**									
	1.1	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2**
	Gini coefficient (α)									
20	0.52	0.49	0.45	0.42	0.39	0.37	0.34	0.32	0.31	0.29
50	0.59	0.55	0.50	0.47	0.43	0.40	0.37	0.35	0.33	0.31
100	0.63	0.58	0.53	0.49	0.45	0.42	0.39	0.36	0.34	0.32
150	0.65	0.59	0.55	0.50	0.46	0.42	0.40	0.37	0.34	0.32

SAMPLE SIZES	Value of the shape parameter (α) of Pareto distribution									
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
	Gini coefficient (α)									
200	0.66	0.60	0.56	0.51	0.47	0.43	0.40	0.37	0.35	0.33
250	0.67	0.61	0.56	0.51	0.47	0.44	0.40	0.37	0.35	0.33
500	0.69	0.63	0.57	0.52	0.48	0.44	0.41	0.38	0.35	0.33
750	0.70	0.64	0.58	0.53	0.48	0.44	0.41	0.38	0.35	0.33
1000	0.71	0.64	0.58	0.53	0.49	0.45	0.41	0.38	0.35	0.33
2000	0.72	0.66	0.59	0.54	0.49	0.45	0.41	0.38	0.36	0.33
5000	0.74	0.69	0.61	0.55	0.49	0.45	0.41	0.38	0.36	0.33

.

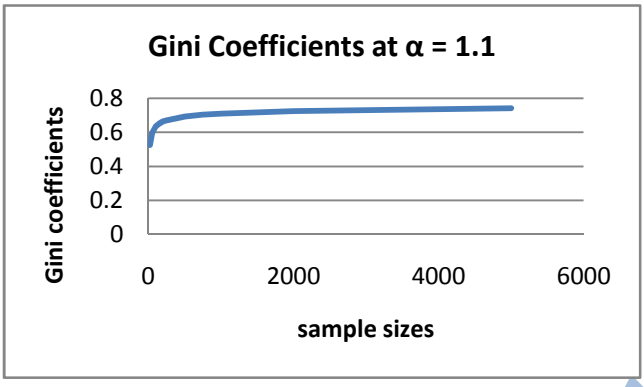

Figure 1: Line Graph of Gini coefficients when α=1.1

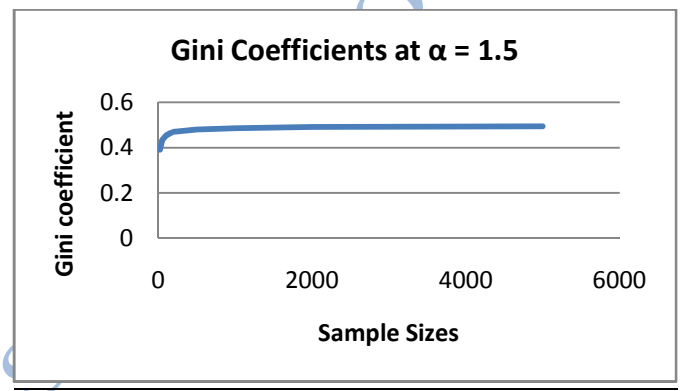

Figure 2: Line Graph of Gini coefficients when α=1.5

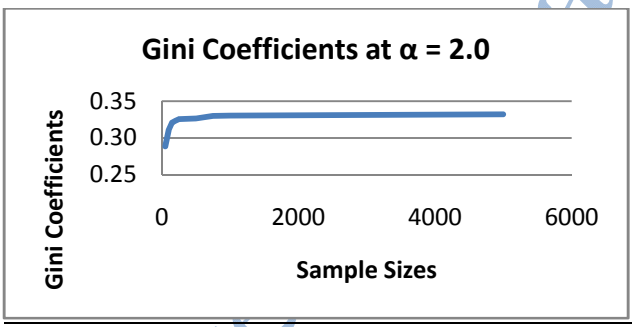

Figure 3: Line Graph of Gini coefficients when α=2.0

VI. Conclusion

From the analysis performed in the previous sections, it can be concluded theoretically based on mathematical proofs that when α = 1.5, the Gini coefficient of the Pareto distribution will equal that of the exponential distribution. However, in real-life situations, this may not be true due to some unobservable and/or hidden factors. As can be observed from the simulation studies, the approximation of Pareto distribution to exponential distribution is dependent on the number of items/categories of the dataset in the economic data.

Generally, it can be concluded that the exponential distribution provides a good fit and credible alternative to Pareto distribution to model wealth distribution data.

It was also demonstrated in this work that importation of goods in Nigeria can be assumed to follow an exponential distribution based on the estimated Gini coefficient of inequality. This is a good development that would aid proper modelling of such data which would in turn assist the import policy makers in Nigeria as it affect the balance of payment positively.

.

References

[1] Roussas, G. G., A Course in Mathematical Statistics, 2nd ed., California: Academic Press. 1997.

[2] Gaswirth, J. L. A.,"The Estimation of the Lorenz Curve and Gini Index." Review of Economics and Statistics, 1972, pp. 306-316

[3] Douglas C. Montgomery and George C. Runger, Applied statistics for Sciences and Engineers, 6th ed., New York: Wiley, 2014.
[4] S.A. Klugman, H. H. Panjer, and G.E. Willmot, "Loss Models: From Data to Decisions," Wiley, New York. 1998.
[5] E. L. Lehmann, "Theory of Point Estimation,".New York: Wiley,1983.
[6] J. Bierlant, J.F. Teugels, and P. Vynckier. "Practical Analysis of Extreme Values". Lenven: Leuven University Press, Belgium, 1996.
[7] N.L. Johnson, S. Kotz, N. Balakrishnan, "Continuous univariate distributions". New York: Wiley, 1994.
[8] Pareto distribution.Wikipedia (2012),The free encyclopedia:http://en.m.wikipedia.org/wiki/statistical/Paretodistribution.Accessed on 20th January, 2017
[9] Samprit Chatterjee, and Ali S. Hadi, Regression by Example, 4th ed., New York: Wiley, 2006.
[10] Gini coefficient. Wikipedia;The free encyclopedia. http://en.wikipedia.org/wiki/Gini_coefficient
[11] Lorenzo G. Bellu and Paolo Liberati, "Inequality Analysis: The Gini Index. EASYPol On-line resource materials for policy making. Module 040, 2006.
[12] Robert T. Jantzen and Klaus Volpert, "On the Mathematics of Income inequality: Splitting the Gini Index into two." Available at http://dx.doi.org/10.4169/amer.math.monthly.119.10.824. MSC: Primary 97M40. December, 2012.