THE SIZE DISTRIBUTION OF CITIES IN A REGION:
AN EVALUATION OF PARETO, LOGNORMAL
AND PPS DISTRIBUTIONS

V. Christopher Amalraj¹ §, A. Subbarayan², S. Balamuralitharan³

¹Department of Community Medicine
S.R.M. Medical College and Research Centre
Kattankulathur, 603 203, INDIA

²Department of Computer Applications
S.R.M. University
Kattankulathur, 603 203, INDIA

³Faculty of Engineering and Technology
Department of Mathematics
S.R.M. University
Kattankulathur, 603202, Tamil Nadu, INDIA

Abstract: The Pareto-Positive Stable(PPS) distribution is introduced as a new model for describing city size data of a region in a country. The PPS distribution provides a flexible model for fitting the entire range of a set of city size data and the classical Pareto and Zipf distributions are included as a particular case. The new distribution is compared with two classical models: Pareto and lognormal distribution. In all the data sets considered, the Newtons forward and backward(equal intervals), Lagrange interpolation formula(unequal interval) outperforms the fits of Pareto and lognormal distributions.

AMS Subject Classification: 62H10, 62H12
Key Words: city-size distribution, Pareto distribution, Pareto Positive Stable law

Received: December 30, 2013

§Correspondence author
1. Introduction

In this work we have analyzed the city size distribution data using the methodology viz., PPS distribution developed by Sarabia and Prieto(2009).

Systems with measurable entities (which can be defined by their size) are characterized by particular properties of their distribution. There are extensive literature and case studies in this field that include work on population of countries, incomes of people in the same economy, frequency of words in languages etc. Scholars have been addressing the problem, regarding the size distribution of such systems; the first is finding a mathematical description for these distributions. The most popular suggestions are the lognormal distribution and the power law (known also as Zipfs law). Yet, there are other expressions that describe with equal success general observed distributions. The second problem is to develop model which explains the size distribution. Here also several models (either analytical or computer simulations) were proposed. These models can be divided into two classes: the first includes models with a limited number of parameters and the second class includes mostly economic models which are more complex and includes numerous parameters.

Pareto distribution was initially proposed Auerbach(1913) and followed by Zipf(1949) to fit City size data. Rosen and Resnick(1980) did a cross-country investigation of city sizes in 44 countries and found that Pareto exponent was in the interval $\alpha \in [0.81,1.96]$. They have also tried to explain the variations in the Pareto exponent, and showed that it is sensitive to city definition and city sample size. Based on 135 USA metropolitan areas in 1991, Krugman(1996) calculated the value of close to one. Using the same data set, Gabaix(1999a,1999b) derived a statistical explanation of Zipfs law for cities. Brakman et al.(1999) with Netherland data provided Pareto evidence over a wide range of time. Nitsch(2005) used meta analysis and concluded that Pareto distribution as an appropriate one to fit city size data. Zanette and Manrubia(1997) developed an intermittency model to large-scale city size distributions. David and Weinstein(2002) found that variation in Japanese regional population density, as well as the distribution of city sizes, obeyed a Pareto distribution, at all points in time. Soo(2005) updated values for the interval $[0.73,1.72]$ and tried to explain variations in the Pareto exponent. Moura and Riberio(2006) have showed that Pareto distribution was not valid for smaller cities.

Some probabilistic and economic models have been proposed by many researchers and the central idea among the above models is that Gibarts law (proportional growth) can lead to Pareto distribution. Simon(1955) has shown that a proportional growth can explain several different skew distributions, in-
including lognormal, Pareto and Yule. Anderson and Ge (2005) have shown the superiority of the lognormal distribution with respect to Pareto distribution, using size distribution of Chinese cities. Subbarayan (2009) extensively studied the size distribution of cities in Tamilnadu, Indian state for the period 1901-2001. Sarabia and Prieto (2009) have stated that the validity of the Pareto distribution disappears when all the population is fitted, including cities of medium and small size.

In this paper we have considered the models evolved by Sarabia and Prieto (2009). The descriptive model evolved by them is called PPS distribution for city/town size data. It is interesting to note that more flexible models emerge from PPS under certain conditions. The classical Pareto and Zipf distributions are included as particular cases. The PPS distribution provides a flexible model for fitting the entire range of a set of city/town size data, when zero and uni-modality are possible. Therefore the probability density function always decreases or it has a local maximum.

2. Pareto Distribution

The linear relation between population of cities and their ranks on a log-log plot is found to be a power law, where the absolute value of this linear function is the exponent of the power law [4]. A power law is also known as a classical Pareto distribution with cumulative distribution function (cdf),

\[
F(x) = P_r(X \leq x) = 1 - \left(\frac{x}{\sigma}\right)^{-\alpha} \quad x \geq \sigma > 0
\]

and \( F(x) = 0 \) if \( x < \sigma \)

where \( \alpha > 0 \) is a shape parameter and \( \sigma \) is a scale parameter, which represents the population of the smallest city in the sample. The \( \alpha \) parameter is called the Pareto coefficient. The quantity \( \left(\frac{x}{\sigma}\right)^{-\alpha} \) represents the proportion of cities of large size than a given \( x \) value.

2.1. The PPS Distribution

Sarabia and Prieto (2009) have defined PPS distribution in terms of cdf.

If \( F(x) = P_r(X \leq x) \), then
\[ F(x) = 1 - \exp\left\{ \lambda \left[ \log \left( \frac{x}{\sigma} \right) \right]^\nu \right\} \] (2)

and \( F(x) = 0 \) if \( x < \sigma \), where, \( \lambda, \sigma, \nu > 0 \)

A random variable with cdf given by (2) will be denoted by \( X \sim PPS(\lambda, \sigma, \nu) \). It may be noted that \( \lambda \) and \( \nu \) are shape parameters and \( \sigma \) is a scale parameter.

Zipf distribution (\( \lambda = \nu = 1 \))

Classical Pareto (\( \nu = 1 \))

More flexible models emerge when (\( \nu > 1 \)).

### 2.2. PPS Based on Weibull Distribution

PPS distribution can also be obtained from a monotonic transformation of the Weibull distribution[5,7].

Let \( Z \) be a classical Weibull distribution with cdf

\[ F_z(Z) = 1 - \exp(-Z^\nu), Z > 0, where, \nu > 0 \] (3)

Then the random variable

\[ X = \sigma \exp\left[ \lambda \frac{1}{\nu} Z \right] \] (4)

where, \( \sigma, \lambda > 0 \) is distributed according to a PPS (\( \lambda, \sigma, \nu \)) distribution with cdf by (2). Using Eq.(4), if \( X \) is a PPS distribution with cdf given by Eq.(2), the random variable.

\[ Z = \lambda \frac{1}{\nu} \log \left[ \frac{x}{\sigma} \right] \]

is a Weibull random variable with cdf by (3) The pdf of PPS is given by

\[ f(x) = \frac{dF(x)}{dx} = \frac{\lambda \nu \left[ \log \left( \frac{x}{\sigma} \right) \right]^{\nu-1}}{x} \exp\left\{ - \lambda \left[ \log \left( \frac{x}{\sigma} \right) \right]^\nu \right\}, x \geq \sigma \] (5)

and \( f(x) = 0 \) if \( x < \sigma \)
If \( \nu > 1 \) the mode a local maximum of the pdf defined by Eq.(5) is at \( \sigma \exp(z_0) \), where \( z_0 \) is the unique solution of the equation in \( z \),

\[
\lambda \nu Z^\nu + Z - (\nu - 1) = 0
\]

### 2.3. The Three Parameter Lognormal Distribution

[3,12] Three-parameter lognormal Distribution

The pdf of the three-parameter lognormal distribution

\[
f(x, \mu, \sigma, \gamma) = \frac{1}{(x - \gamma)\sigma\sqrt{2\pi}} \exp\left\{ - \frac{[\ln(x - \gamma) - \mu]^2}{2\sigma^2} \right\}
\]

(6)

where, \( x > \gamma \geq 0, \alpha < \mu < \infty, \sigma > 0 \) and \( \gamma \) is the threshold parameter or location parameter that defines the point where the support set of the distribution begins; \( \mu \) is the scale parameter that stretch or shrink the distribution and \( \sigma \) is the shape parameter that affects the shape of the distribution.

If \( X \) is a random variable that has a three parameter log-normal probability distribution, then \( Y = \ln(X - \gamma) \) has a normal distribution with mean \( \mu \) and variance \( \sigma^2 \).

The cdf of the three-parameter lognormal Distribution is

\[
F_x(x : \mu, \sigma, \gamma) = \Phi\left[ \frac{\ln(x - \gamma) - \mu}{\sigma} \right]
\]

(7)

For the three parameter lognormal Distribution defined in equation (7), the value of \( \gamma \) is given by the minimum population size value.

### 3. Estimation

Let \( X_1, X_2, \ldots, X_n \) be a sample of size \( x \) drawn from a PPS distribution [2]. We assume that \( \sigma \) parameter is given and we obtain it using the population of the smallest city. We will use the Random Variable \( Z \) defined by \( Z = \log\left[\frac{X_i}{\sigma}\right] \) and its observed value by

\[
Z_i = \log\left(\frac{X_i}{\sigma}\right), \quad i = 1, 2, \ldots, n
\]
The log likelihood function is given by

\[ \log l(\lambda, \nu) = \sum_{i=1}^{n} \log f(x_i) \]

\[ = n \log \lambda + n \log \nu (\nu - 1) \sum_{i=1}^{n} \log Z_i - \lambda \sum_{i=1}^{n} Z_i - \sum_{i=1}^{n} \log X_i \]

where \( f(x) \) is pdf defined in (5)

### 3.1. Maximum Likelihood Estimate of \( \hat{\lambda} \) and \( \hat{\nu} \)

Taking partial derivatives with respect to \( \lambda \) and \( \nu \) and equating then to zero the following normal equations are obtained [17].

\[ \frac{\partial \log l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} Z_i = 0 \]  
\[ (8) \]

\[ \frac{\partial \log l}{\partial \nu} = \frac{n}{\nu} + \sum_{i=1}^{n} \log Z_i - \lambda \sum_{i=1}^{n} Z_i \log Z_i = 0 \]  
\[ (9) \]

If we eliminate \( \lambda \) in Equations. (8) and (9) the equation in \( \nu \) is obtained.

\[ \frac{1}{\nu} + \frac{1}{n} \sum_{i=1}^{n} \log Z_i - \frac{\sum_{i=1}^{n} Z_i^{\nu} \log Z_i}{\sum_{i=1}^{n} Z_i^{\nu}} = 0 \]  
\[ (10) \]

The above equation can be solved using Newton Raphson method. The \( \lambda \) estimator

\[ \hat{\lambda} = \left\{ \frac{1}{n} \sum Z_i \right\}^{-1} \]  
\[ (11) \]

We have already stated that more flexible models emerge when \( \nu > 1 \). The value of \( \hat{\nu} \) is considered with the range \( 2.0 \leq \hat{\nu} \leq 2.5 \).

### 3.2. Maximum Likelihood Estimation for Parameters \( \mu \) and \( \sigma \) for Three-Parameter Lognormal Distribution

The MLE for the parameters of \( \mu \) and \( \sigma \) are given by [17]

\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \log (X_i - \gamma) \]  
\[ (12) \]
\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \log(X_i - \gamma) - \hat{\mu} \right]^2 \]  

(13)

4. Numerical Results

4.1. For Equal Intervals

let \( Y = f(x) \) denote a function which takes the values \( y_0 + y_1 + \ldots + y_n \) corresponding to the values \( x_0 + x_1 + \ldots + x_n \) respectively of \( x \).

Let us suppose that the values of \( x \) viz \( x_0, x_1, \ldots, x_n \) are equidistant.

That is \( x_i - x_{i-1} = h, \) for \( i = 1, 2, \ldots, n. \)

Therefore \( x_i = x_0 + h, \ x_2 = x_0 + 2h, \) etc.

\( x_i = x_0 + ih, \ i = 1, 2, \ldots, n. \)

Let \( P_n(x) \) be a polynomial of the nth degree in \( x \) such that

\[ y_i = f(x_i) = P_n(x_i), \ i = 0, 1, 2, \ldots, n. \]

Let us assume \( P_n(x) \) in the form given below

\[ P_n(x) = a_0 + a_1(x-x_0)(1) + a_2(x-x_0)^2 + \ldots + a_r(x-x_0)^r + \ldots + a_n(x-x_0)^n \ldots \]  

(4.1.1)

The \((n+1)\) unknowns \( a_0, a_1, a_2, \ldots, a_n \) can be found as follows.

\[ P_n(x_0) = y_0 = a_0 \] (setting \( x = x_0 \) in (4.1.1))

\[ \triangle^r P_n(x) = a_r, r!h^r + \text{terms involving} (x - x_0) \text{as a factor}. \ldots \]  

(4.1.2)

(since the first \( r \) terms vanish)

Setting \( x = x_0 \) in (4.1.2), \( \triangle^r P_n(x_0) = a_r r!h^r \) (since the other terms in (4.1.2) vanish)

\[ \text{Hence } a_r = \frac{1}{r!h^r} \triangle^r y_0, \ldots \]  

(4.1.3)
Putting \( r = 1, 2, 3, \ldots, n \) in (4.1.3), we get the values of \( a_1, a_2, \ldots, a_n \). Therefore,

\[
P_n(x) = y_0 + \frac{(x-x_0)}{h} \Delta y_0 + \frac{(x-x_0)^2}{2!h^2} \Delta^2 y_0 + \cdots + \frac{(x-x_0)^n}{n!h^n} \Delta^n y_0 \tag{4.1.4}
\]

Using in (4.1.4),

\[
P_n(x) = P_n(x_0 + uh) = y_0 + \frac{u(1)}{1!} \Delta y_0 + \frac{u(2)}{2!} \Delta^2 y_0 + \cdots + \frac{u(r)}{r!} \Delta^r y_0 + \cdots + \frac{u(n)}{n!} \Delta^n y_0 \tag{4.1.5}
\]

where \( u^r = u(u-1)(u-2)\cdots(u-r-1) \) (If \( x \) is given, \( u \) is found out)

Equation (4.1.5) is known as Gregory - Newton forward interpolation formula [4.1.2 and 4.1.3].

### 4.2. Equal Intervals for Interpolation

Suppose \( y = f(x) \) takes the values \( y_0, y_1, \ldots, y_n \) corresponding to the values \( x_0, x_1, \ldots, x_n \) of \( x \)

Let \( x_i - x_{i-1} = h \) for \( i = 1, 2, \ldots, n \) (equal intervals)

\[
\therefore \ x_i = x_0 + ih, \ i = 0, 1, 2, \ldots, n.
\]

Now, we want to find a collection polynomial \( P_n(x) \) of degree \( n \) in \( x \) such that
\[ P_n(x_i) = y_i, \ i = 0, 1, 2, \ldots, n. \]

Let \( P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_{n-1}) + \ldots + a_r(x - x_n)(x - x_{n-1})(x - x_{n-r+1}) + \ldots + a_n(x - x_n)(x - x_{n-1})(x - x_1) \) \( (4.2.1) \)

Since \( x_{n-1} = x_n - h, x_{n-2} = x_n - 2h \ldots x_{n-r+1} = x_n - (r - 1)h \)

\[ x_1 = x_n - (n - 1)h, \] \( we \) have

Let \( P_n(x) = a_0 + a_1(x - x_n) + a_2(x - x_n)(x - x_n + h) \)
\[ + a_3(x - x_n)(x - x_n + h)(x - x_n)(x - x_n + 2h) + \ldots \]
\[ + a_n(x - x_n)(x - x_n + h)(x - x_n + n - 1h) \]

Let \( P_n(x) = a_0 + a_1(x - x_n)^{(1)} + a_2(x - x_n + h)^{(2)} + a_3(x - x_n)(x - x_n + 2h)^{(3)} + \ldots + a_r(x - x_n + r - 1h)^{(r)} + \ldots + a_n(x - x_n + r - 1h)^{(n)} \) \( (4.2.2) \)

We shall find \( a_0, a_1, \ldots, a_n \) such that \( P_n(x_i) = y_i \)

Since \( \triangle = E^{-1} \triangle \)

\[ \nabla^r(x - a)^m = E^{-r} \nabla^r(x - a)^m \]
\[ = E^{-r} [m(m - 1)(m - 2)(m - r + 1)h^r(x - a)^{(m-r)}] \]
\[ = m(m - 1)(m - 2)(m - r + 1)h^r(x - rh - a)^{(m-r)} \text{ if } r \leq m \] \( (4.2.3) \)

\( x - x_n \) is a factor in all terms of RHS of \( (4.2.2) \) except in \( a_0 \)

Putting \( x = x_n \) in \( (4.2.2) \)

\[ P_n(x_n) = y_n = a_0 \]

Operating \( (4.2.2) \) by \( \nabla^r \) using \( (4.2.3) \)

\[ \nabla^r P_n(x) = 0 + 0 + \ldots + 0 + a_r r! h^r + (r + 1).r.(r - 1)\ldots 2.h^r a_{r+1}(x - x_n)^{(1)} \]
+ terms involving \((x - x_n)\) as a factor.

Setting \(x = x_n\) in this,

\[
\nabla^r P_n(x_n) = \nabla^r y_n = a_r r! h^r. \text{ since other terms vanish.}
\]

\[
\therefore a_r = \frac{1}{r! h^r} \nabla^r y_n, \text{ where } r = 1, 2, \ldots, n. \quad (4.2.4)
\]

Putting the values of \(a_0, a_1, \ldots, a_n\) in (4.2.2), we get

\[
P_n(x) = y_n + \frac{(x - x_n)^{(1)}}{1! h} \nabla y_n + \frac{(x - x_n + h)^{(2)}}{2! h^2} \nabla^2 y_n + \ldots + \frac{(x - x_n + r - 1 h)^{(r)}}{r! h^r} \nabla^r y_n + \ldots + \frac{(x - x_n + n - 1 h)^{(n)}}{n! h_n} \nabla^n y_n \quad (4.2.5)
\]

Let \(\frac{x-x_n}{h} = v\); hence \(x = x_n + vh\)

Then \(\frac{(x-x_n+r-1h)^{(r)}}{h^r} = \frac{(vh+r-1h)^{(r)}}{h^r} = (v + r - 1)^{(r)}\)

Therefore, (4.2.5) becomes,

\[
P_n(x) = P_n(x_n + vh) = y_n + \frac{v^{(1)}}{1!} \nabla y_n + \frac{(v + 1)^{(2)}}{2!} \nabla^2 y_n + \ldots + \frac{(v + r - 1)^{(r)}}{r!} \nabla^r y_n + \ldots + \frac{(v + n - 1)^{(n)}}{n!} \nabla^n y_n
\]

\[
\therefore P_n(x) = P_n(x_n + vh) = y_n + \frac{v}{1!} \nabla y_n + \frac{v(v+1)}{2!} \nabla^2 y_n + \frac{v(v+1)(v+2)}{3!} \nabla^3 y_n + \ldots + \frac{v(v+1)(v+2)(v+n-1)}{n!} \nabla^n y_n \quad (4.2.6)
\]

### 4.3. Example 1

Find the values of \(y\) at \(x = 21\) and \(x = 28\) from the following data.

<table>
<thead>
<tr>
<th>(x)</th>
<th>20</th>
<th>23</th>
<th>26</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>0.3420</td>
<td>0.3907</td>
<td>0.4384</td>
<td>0.4848</td>
</tr>
</tbody>
</table>
Solution:

Since $x = 21$ is nearer to the beginning of the table, we use Newton’s forward interpolation formula

$$y(21) = P_3(x)$$

$$\therefore y(21) = 0.3583$$

Since $x = 28$ is nearer to end value, we use Newton’s backward interpolation formula

$$y(28) = P_3(x)$$

$$\therefore y(28) = 0.4695$$

4.4. Numerical Results for Unequal Interval

The Lagrangian interpolation formula can also be written as

$$f(x) = \sum_{i=0}^{n} \frac{\pi_n(x)}{(x-x_i)\pi_n(x_i)} y_i, \text{ where}$$

$$\pi_n(x) = (x-x_0)(x-x_1)...(x-x_n) \text{ and } \pi_n(x) = \frac{d}{dx}[\pi_n(x)]$$

$$\pi_n(x) = (x-x_0)(x-x_1)...(x-x_n)$$

Differentiating this and substituting $x = x_i$ we get

$$\pi_n(x_i) = (x_i-x_0)(x_i-x_1)...(x_i-x_{i-1})(x_i-x_{i+1})...(x-x_n)$$

substituting this equation in (4.3.1)

$$y = f(x) = \frac{(x-x_1)(x-x_2)...(x-x_n)}{(x_0-x_1)(x_0-x_2)...(x_0-x_n)} y_0 + \frac{(x-x_0)(x-x_2)...(x-x_n)}{(x_1-x_0)(x_1-x_2)...(x_1-x_n)} y_1 + ...$$

$$+ \frac{(x-x_0)(x-x_1)...(x-x_{i-1})(x-x_{i+1})...(x-x_n)}{(x_i-x_0)(x_i-x_1)...(x_i-x_{i-1})(x_{i+1})...(x_i-x_n)} y_i + ...$$

$$+ \frac{(x-x_0)(x-x_1)...(x-x_{i-1})(x-x_{i-1})}{(x_n-x_0)(x_n-x_1)...(x_n-x_{i-1})(x_{n-1})} y_n \quad (4.3.1)$$
\( f(x) = \frac{\pi_n(x)}{x - x_0} \cdot \frac{y_0}{\pi_n(x_0)} + \frac{\pi_n(x)}{x - x_1} \cdot \frac{y_1}{\pi_n(x_1)} + \ldots + \frac{\pi_n(x)}{x - x_n} \cdot \frac{y_n}{\pi_n(x_n)} \)

\[ = \sum_{i=0}^{n} \frac{\pi_n(x)}{\pi_n(x_i)} \cdot \frac{y_i}{x - x_i} \]

4.5. Example 2

Using Lagrange’s formula of interpolation find \( y(2011) \) given

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>No.of.Cities (y)</td>
<td>45</td>
<td>52</td>
<td>60</td>
<td>88</td>
<td>140</td>
<td>159</td>
</tr>
</tbody>
</table>

\[ y = f(x) \]

\[ f(2011) = 643 \]

\[ \therefore \] in 2011 Kerala have 643 census towns

5. Conclusion

In this work we have analyzed the city size distribution data using the methodology viz., PPS distribution developed by Sarabia and Prieto(2009). The PPS distribution provides a comparative flexible model for all range of a set of city size for 6 census periods. For purposes we have considered lognormal distribution and Pareto distribution for fitting City Size Distribution data. These distributions are frequently used by urban researchers for fitting City Size Distribution data sets. We have used maximum likelihood estimate method for the estimation of the parameters of lognormal Pareto, and PPS. By using AIC we have compared the results and it is noted that PPS distribution outperforms the fit provided by Pareto and lognormal distribution. This clearly indicates that PPS is considered to be a good fit not only for country data but also for regional city size data.

References


