Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 41

5-1-2016

An Evaluation of Pareto, Lognormal and PPS Distributions: The Size Distribution of Cities in Kerala, India

Christopher A. Vallabados
S.R.M. Medical College and Research Centre, Kattankulathur, India, christopheramalraj@gmail.com

Subbarayan A. Arumugam
S.R.M. University, Kattankulathur, India, subbarayan1948@gmial.com

An Evaluation of Pareto, Lognormal and PPS Distributions: The Size Distribution of Cities in Kerala, India

Cover Page Footnote
V.Christopher Amalraj Lecturer in Bio-Statistics Department of Community Medicine SRM Medical College & Research Centre

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 825-835.

ISSN 1538 − 9472

An Evaluation of Pareto, Lognormal and PPS Distributions: The Size Distribution of Cities in Kerala, India

Christopher A. Vallabados
S.R.M. Medical College and Research Centre
Kattankulathur, India

Subbarayan A. Arumugam
S.R.M. University
Kattankulathur, India

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.

Keywords: City-size distribution, Pareto distribution, log normal distribution, Zipf's law, positive stable law

Introduction

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 Zipf's 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.

Dr. Vallabados is a Lecturer in the Department of Statistics. Email him at: christopheramalraj@gmail.com. Dr. Arumugam is Professor and Head of the Department of Computer Applications. Email him at: subbarayan1948@gmail.com.

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. A power law is also known as a classical Pareto distribution with cumulative distribution function (cdf),

$$F(x) = P_r(X \le x) = 1 - \left(\frac{x}{\sigma}\right)^{-\alpha}, x \ge \sigma > 0 \text{ and } F(x) = 0 \text{ if } x < \sigma, \qquad (1)$$

where $\alpha > 0$ is a shape parameter and σ is a scale parameter, which represents the population of the smallest city in the sample. The α 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.

A Select Review of City Size Distribution Models

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 $\alpha \varepsilon$ [0.81 to 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 α close to one. Using the same data set, Gabaix (1999a, 1999b) derived a statistical explanation of Zipf's law for cities. Brakman, Garretsen, Van Marrewijk, & Van Den Berg (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. Davis 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 α values for the internal [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 Gibart's law

(proportional growth) can lead to Pareto distribution. Simon (1955) has shown that a proportional growth can explain several different skew distributions, 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.

The models considered here evolved by Sarabia and Prieto (2009). The descriptive model evolved by them is called PPS distribution for city / town size data. 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-modelity are possible (i.e., the probability density function always decreases or it has a local maximum)

The PPS Distribution

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

$$\begin{aligned} &\text{If } F(x) = P_r(X \le x), \text{ then} \\ &F(x) = 1 - \exp\left\{-\lambda\left[\log(x/\sigma)\right]^{\upsilon}\right\}, x \ge \sigma \text{ and} \\ &F(x) = 0 \text{ if } x < \sigma, \text{ where } \lambda, \sigma, u > 0 \end{aligned} \tag{2}$$

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

- Zipf distribution ($\lambda = \upsilon = 1$)
- Classical Pareto ($\upsilon = 1$)

More flexible models emerge when $\upsilon > 1$.

PPS based on Weibull Distribution

PPS distribution can also be obtained from a monotonic transformation of the Weibull distribution.

Let Z be a classical Weibull distribution with cdf

$$F_z(z) = 1 - \exp\left(-z^{\upsilon}\right), z > 0, \text{ where } \upsilon > 0 \tag{3}$$

then the random variable

$$X = \sigma \exp\left[l^{-1/\upsilon} Z\right] \tag{4}$$

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

$$Z = l^{1/\upsilon} log\left(x/\sigma\right)$$

is a Weibull random variable with cdf by (3).

The pdf of PPS is given by

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

and $f(x) = 0$ if $x < \sigma$.

If $\upsilon > 1$ 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,

$$\lambda \upsilon Z^{\upsilon} + Z - (u - 1) = 0$$

Three-parameter Lognormal Distribution

The pdf of the three-parameter lognormal distribution

$$f(x; \mu, \sigma, \gamma) = \frac{1}{(x-\gamma)\sigma\sqrt{2\pi}} \exp\left\{-\frac{\left[\ln(x-\gamma) - \mu\right]^2}{2\sigma^2}\right\} \tag{6}$$

where $x > \gamma \geq 0$, $-\infty < \mu < \infty$, $\sigma > 0$ and γ is the threshold parameter or location parameter that defines the point where the support set of the distribution begins; μ is the scale parameter that stretch or shrink the distribution and σ 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 μ and variance σ_2. The cdf of the three-parameter lognormal distribution is

$$F_x(x;\mu,\sigma,\gamma) = \Phi\left[\frac{\ln(x-\gamma)-\mu}{\sigma}\right] \tag{7}$$

For the three-parameter lognormal distribution defined in equation (7), the value of γ is given by the minimum population size value.

Estimation

Let $x_1, x_2, \ldots, x_n$ be a sample of size x drawn from a PPS distribution. We assume that σ – parameter is given and we obtain it using the population of the smallest city. We will use the random variable Z defined by $Z = \log[X/\sigma]$ and its observed value by

$$z_i = \log(x_i/\sigma) \quad i = 1,2,\ldots,n$$

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).

Maximum Likelihood Estimate of $\hat{\lambda}$ and $\hat{\upsilon}$

Taking partial derivatives with respect to λ and υ and equating then to zero the following normal equations are obtained.

$$\frac{\partial \log l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} z_i = 0 \tag{8}$$

$$\frac{\partial \log l}{\partial \nu} = \frac{n}{\nu} + \sum_{i=1}^{n} \log z_i - \lambda \sum_{i=1}^{n} z_i \log z_i = 0 \tag{9}$$

If λ is eliminated in Equations, (8) & (9) the equation in υ is obtained.

$$\frac{1}{\nu} + \frac{1}{n} \sum_{i=1}^{n} \log z_i - \frac{\sum_{i=1}^{n} z_i^{\nu} \log z_i}{\sum_{i=1}^{n} z_i^{\nu}} = 0 \tag{10}$$

The above equation can be solved using the Newton–Raphson method. The λ estimator

$$\hat{\lambda} = \left\{ \frac{1}{n} \sum z_i \right\}^{-1} \tag{11}$$

As already stated, more flexible models emerge when $\upsilon > 1$. The value of $\hat{\upsilon}$ is considered with the range $2.0 \leq \hat{\upsilon} \leq 2.5$.

Maximum Likelihood Estimation for Parameters μ and σ for Three-parameter Lognormal Distribution

The MLE for the parameters of μ and σ are given by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \log\left(x_i - \gamma\right) \tag{12}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left[\log\left(x_i - \gamma\right) - \hat{\mu} \right]^2 \tag{13}$$

Empirical Application to City Size

India has very rich source of information for urban studies. The census volumes, both at the national and state levels, provide a mine of information for rural and urban places for a period of 100 years. It is also main source of information for

the construction of city size distribution. The census periods covered are 1951, 1961, 1971, 1981, 1991 and 2001.

Urban population by size classification is based on the following:

Class-I	-	Population
I	-	Greater than 100,000
II	-	50,000 – 100,000
III	-	20,000 – 50,000
IV	-	10,000 – 20,000
V	-	5,000 – 10,000
VI	-	Less than 5,000

The number of cities / towns for each census year under six classes is given in the following Table 1.

Table 1. Size Distribution of Cities and Towns in Kerala (1951-2001)

Census Year	> 100,000	50,000 – 100,000	20,000 – 50,000	10,000 – 20,000	5,000 – 10,000	< 5,000	Total
1951	4	3	10	21	6	1	45
1961	4	4	22	17	4	1	52
1971	5	8	32	11	3	1	60
1981	6	8	55	14	4	1	88
1991	9	17	69	34	10	1	140
2001	10	24	72	37	15	1	159

Data for Model Fitting

Some relevant information about the data sets used appears in following Table 2. For each census year, the third column shows the size (number of people) of the smallest town we have considered. The fourth column shows the number of cities and towns fitted. The fifth column represents the percentage of the total Kerala cities / towns which have been considered. The sixth column shows the number of people who live in the cities and towns fitted and finally the seventh column the percentage of the total Kerala population that the number of inhabitants represents. For example, in 1951 we have considered 44 cities and towns with at least 3,098 people, which correspond to 97.78% of cities and towns and 99.79 of the total population of Kerala.

Table 2. Some relevant information about Kerala city size data sets used.

Census Year	Minimum town Size Considered	Town Considered		Population Considered	
		Number	% of Total	Number	% of Total
1951	3,098	44	97.78	14,85,347	99.79
1961	2,859	51	98.08	21,06,197	99.86
1971	4,750	59	98.33	30,68,436	99.84
1981	4,489	87	98.86	43,95,172	99.89
1991	4,820	139	99.28	72,57,261	99.94
2001	4,699	158	99.37	82,62,226	99.94

Fitted Models and Results

Three models were fitted and compared: classical Pareto distribution, three-parameter lognormal distribution and PPS distribution. The Pareto distribution was included for comparison purposes and it is known that this distribution is used to fit the upper tail of the distribution. The lognormal distribution was a classical distribution to fit a set of city size data. The PPS distribution was adjusted according to maximum likelihood method discussed in the Estimation section.

Akaike Information Criterion (AIC) for Pareto, Lognormal and PPS

For model identification Akaike (1974) suggested Akaike Information Criterion and the same is given by $AIC = 2\log l - 2d$ where $\log l$ is the likelihood of the model evaluated at the maximum likelihood estimates and d is the number of parameters.

The AIC is a measure of the goodness of fit of an estimated statistical model and a useful tool for model selection. In view of this we have to choose a model among the three models fitted which has the highest AIC. Parameter estimates and value of AIC statistics are given in the following Table 3 for Pareto and lognormal distribution.

Table 3. Parameter estimates and value of AIC obtained from the fitting of the lognormal ($\hat{\mu}$ and $\hat{\sigma}$ parameters) and Pareto distribution ($\hat{\alpha}$ parameter) to the city size data in Kerala by maximum likelihood.

Census Year	N	$\hat{\mu}$	$\hat{\sigma}$	AIC - Lognormal	$\hat{\alpha}$	AIC - Pareto
1951	45	9.493	5.425	-1100.104	1.87	-1054.742
1961	52	3.557	1.005	-2783.805	2.147	-523.918
1971	60	4.304	0.948	-4960.471	1.829	-1048.969
1981	88	4.358	0.616	-1308.733	1.931	-2152.896
1991	140	2.546	1.266	-1271.844	0.829	-2521.164
2001	159	2.208	1.386	-1329.925	0.691	-2746.167

For all the data sets, the lognormal distribution presents a higher value of AIC statistics than the Pareto distribution. For example, in 2001 the value of AIC statistics is -1329.925 for the lognormal and -2746.167 for the Pareto distribution. In consequence, with these data sets the lognormal distribution is preferable to the Pareto distribution. This conclusion is consistent with the results obtained by Anderson and Ge (2005).

The results of PPS distribution appear in Table 4 for $2.0 < \hat{\upsilon} < 2.5$. In all the six considered census years, the distribution of PPS presents the highest values of the AIC statistics, in comparison with other two models. For example, in 2001 the AIC value is -688.802, higher than lognormal and Pareto AIC values. We can conclude that the distribution outperforms the classical Pareto and lognormal distribution in all the 6 data sets considered for the regional city size distribution.

Table 4. Parameter estimates and value of AIC obtained from the fitting of the PPS distribution ($\hat{\lambda}$ and $\hat{\upsilon}$ parameters) to the city size data in Kerala by maximum likelihood.

Census Year	N	$\hat{\lambda}$	$\hat{\upsilon}$	AIC -PPS
1951	45	0.813	2.1	-4749.968
1961	52	1.063	2.2	-535.435
1971	60	0.803	2.3	-580.813
1981	88	0.838	2.4	-1095.619
1991	140	2.675	2.5	-122.848
2001	159	2.332	2.6	-688.802

Conclusion

City size distribution data were analyzed using the PPS distribution developed by Sarabia and Prieto (2009). It provided a comparative flexible model for all range of a set of city size for six census periods. The lognormal distribution and Pareto distribution were also included comparison purpose because they are frequently used by urban researchers. The maximum likelihood estimate was the method used for the estimation of the parameters of lognormal Pareto, and PPS. Via AIC, it was noted that PPS distribution outperforms the fit provided by Pareto and lognormal distribution. This indicates that PPS is considered to be a good fit not only for country data but also for regional city size data.

References

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716-723. doi:10.1109/TAC.1974.1100705

Anderson, G., & Ge, Y. (2005). The size distribution of Chinese cities. *Regional Science and Urban Economics, 35*(6), 756-766. doi:10.1016/j.regsciurbeco.2005.01.003

Auerbach, F. (1913). Das gesetz der bevölkerungskoncentration. *Petermanns Geograpische Mitteilungen, 59*(1), 74-76.

Brakman, S., Garretsen, H., Van Marrewijk, C., & Van Den Berg, M. (1999). The return of Zipf: Towards a further understanding of the rank-size distribution. *Journal of Regional Science, 39*(1), 182-213. doi:10.1111/1467-9787.00129

Davis, D. R., & Weinstein, D. E. (2002). Bones, bombs, and break points: The geography of economic activity. *The American Economic Review, 92*(5), 1269-1289. doi:10.1257/000282802762024502

Gabaix, X. (1999a). Zipf's law and the growth of cities. *The American Economic Review, 89*(2), 129-132. doi:10.1257/aer.89.2.129

Gabaix, X. (1999b). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics, 114*(3), 739-767. doi:10.1162/003355399556133

Krugman, P. (1996). *The Self-Organizing Economy*. Cambridge, MA: Blackwell Publishing.

Moura Jr., N. J., & Ribeiro, M. B. (2006). Zipf law for Brazilian cities. *Physica A: Statistical Mechanics and its Applications, 367*, 441-448. doi:10.1016/j.physa.2005.11.038

Nitsch, V. (2005). Zipf zipped. *Journal of Urban Economics, 57*(1), 86-100. doi:10.1016/j.jue.2004.09.002

Rosen, K. T., & Resnick, M. (1980). The size distribution cities: An examination of the Pareto law and primacy. *Journal of Urban Economics, 8*(2), 165-186. doi:10.1016/0094-1190(80)90043-1

Sarabia, J. M., & Prieto, F. (2009). The Pareto-positive stable distribution: A new descriptive model for city size data. *Physica A: Statistical Mechanics and its Applications, 388*(19), 4179-4191. doi:10.1016/j.physa.2009.06.047

Simon, H. A. (1955). On a class of skewed distribution functions. *Biometrika, 42*(3/4), 425-440. doi:10.2307/2333389

Soo, K. T. (2005). Zipf's law for cities: A cross country investigation. *Regional Science and Urban Economics, 35*(3), 239-263. doi:10.1016/j.regsciurbeco.2004.04.004

Subbarayan, A. (2009). The size distribution of cities in Tamilnadu (1901-2001). *International Journal of Agricultural and Statistical Sciences, 5*(2), 373-382.

Zanette, D. H., & Manrubia, S. C. (1997). Role of intermittency in urban development: A model of large-scale city formation. *Physical Review Letters, 79*(3), 523-526. doi:10.1103/PhysRevLett.79.523

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Reading, MA: Addision–Wesley.