Cent. Eur. J. Phys. • 6(1) • 2008 • 1-6
DOI: 10.2478/s11534-008-0008-2

VERSITA

Central European Journal of **Physics**

On the truncated Pareto distribution with applications

Research Article

Lorenzo Zaninetti[1]*, Mario Ferraro[2]†

1 Dipartimento di Fisica Generale, Università degli Studi di Torino via P.Giuria 1, 10125 Torino, Italy

2 Dipartimento di Fisica Sperimentale, Università degli Studi di Torino via P.Giuria 1, 10125 Torino, Italy

Received 21 Jun 2007; accepted 28 August 2007

Abstract: The Pareto probability distribution is widely applied in different fields such us finance, physics, hydrology, geology and astronomy. This note deals with an application of the Pareto distribution to astrophysics and more precisely to the statistical analysis of masses of stars and of diameters of asteroids. In particular a comparison between the usual Pareto distribution and its truncated version is presented. Finally, a possible physical mechanism that produces Pareto tails for the distribution of the masses of stars is presented.

PACS (2008): 95.75.-z ,96.30.Ys, 97.10.Xq

Keywords: statistical distributions • asteroids-meteoroids • stars: luminosity and mass functions

1. Introduction

The Pareto distribution [1, 2]. is a simple model for nonnegative data with a power law probability tail. In many practical applications, it is natural to consider an upper bound that truncates the tail [3–5]; the truncated Pareto distribution has a wide range of applications in several fields in data analysis [5, 6].

Power law distributions are often found in astrophysics: for instance in the range $1\mathcal{M}_\odot < \mathcal{M} < 10\mathcal{M}_\odot$, the mass of the stars (main sequence V), when expressed in terms of the solar mass $\mathcal{M}_\odot$, scales as $\psi(\mathcal{M}) \propto \mathcal{M}^{-\alpha}$ with $\alpha=$ 2.35, see [7], or $\alpha=$ 2.3 as suggested by a recent evaluation, see [8]. Other examples are the intensity of nonthermal emission from supernova remnants and extra-galactic radio-sources that scales as $\nu^{-\alpha}$, with numerical values of α ranging between 0.5 and 1, the observed differential spectrum of cosmic rays proportional to $E^{-2.75}$ in the interval 10^{10} eV 5.0 10^{15} eV [9, 10], and the gamma ray burst luminosity function that scales as L^{-2}, [11, 12]. Of course the Pareto distribution is not the only one to exhibit a power law tail, this behaviour being common to different distributions (e.g. the lognormal distribution); however, Pareto distributions are especially attractive for their simple analytical form.

In this paper we present in Section 2 a comparison between the Pareto and the truncated Pareto distributions. In Section 3 the theoretical results are applied to distributions of astrophysical data, namely the mass of stars and the radius of asteroids. A physical mechanism that produces a Pareto type distribution for the masses is suggested in Section 4.

*E-mail: zaninetti@ph.unito.it
†E-mail: ferraro@ph.unito.it

2. Preliminaries

Let X be a random variable taking values x in the interval $[a, \infty]$, $a > 0$. The probability density function (in the following PDF) known as the Pareto is defined by [2]

$$f(x; a, c) = c a^c x^{-(c+1)}, \tag{1}$$

$c > 0$, and the Pareto distribution functions are

$$F(x; a, c) = 1 - a^c x^{-c}. \tag{2}$$

An upper truncated Pareto random variable is defined in the interval $[a, b]$, the corresponding PDF is

$$f_T(x; a, b, c) = \frac{c a^c x^{-(c+1)}}{1 - \left(\frac{a}{b}\right)^c}, \tag{3}$$

[5] and the truncated Pareto distribution function is

$$F_T(x; a, b, c) = \frac{1 - (\frac{a}{x})^c}{1 - (\frac{a}{b})^c}. \tag{4}$$

Momenta of the truncated distributions exist for all $c > 0$. For instance, the mean $f_T(x; a, b, c)$ is, for $c \neq 1$ and $c = 1$, respectively,

$$\langle x \rangle = \frac{ca}{c-1} \frac{1 - (\frac{a}{b})^{c-1}}{1 - (\frac{a}{b})^c}, \qquad \langle x \rangle = \frac{c a^c}{1 - \left(\frac{a}{b}\right)^c} \ln \frac{b}{a} \tag{5}$$

Similarly, if $c \neq 2$, the variance is given by

$$\sigma^2 = \frac{c a^2}{(c-2)} \frac{1 - \left(\frac{a}{b}\right)^{c-2}}{1 - \left(\frac{a}{b}\right)^c} - \langle x \rangle^2, \tag{6}$$

whereas for $c = 2$

$$\frac{c a^c}{1 - \left(\frac{a}{b}\right)^c} \ln \frac{b}{a} - \langle x \rangle^2. \tag{7}$$

In general the n-th central moment is

$$\begin{aligned} \int_a^b (x - \langle x \rangle)^n f_T(x) dx &= \left((a^c)^{-1} - (b^c)^{-1}\right)^{-1} \\ &\cdot \Big[(-\langle x \rangle)^n \, a^{-c} \; {}_2F_1(-c, -n; 1 - c; \frac{a}{\langle x \rangle}) \\ &- (-\langle x \rangle)^n \, b^{-c} \; {}_2F_1(-c, -n; 1 - c; \frac{b}{\langle x \rangle}) \Big], \end{aligned} \tag{8}$$

where ${}_2F_1(a, b; c; z)$ is a regularized hypergeometric function, see [13–15]. An analogous formula based on some of the properties of the incomplete beta function (see [16] and [17]) can be found in [18].

Parameters of the truncated Pareto PDF can be obtained from empirical data via the maximum likelihood method; explicit formulas for maximum likelihood estimators (MLE) are given in [3], and for the more general case in [5], whose results we report here for completeness.

Consider a random sample $\mathcal{X} = x_1, x_2, \ldots, x_n$ and let $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(n)}$ denote their order statistics so that $x_{(1)} = \max(x_1, x_2, \ldots, x_n)$, $x_{(n)} = \min(x_1, x_2, \ldots, x_n)$.

The MLE of the parameters a and b are

$$\tilde{a} = x_{(n)}, \qquad \tilde{b} = x_{(1)}, \tag{9}$$

respectively, and $\tilde{c}$ is the solution of the equation

$$\frac{n}{\tilde{c}} + \frac{n \left(\frac{x_{(n)}}{x_{(1)}}\right)^{\tilde{c}} \ln \left(\frac{x_{(n)}}{x_{(1)}}\right)}{1 - \left(\frac{x_{(n)}}{x_{(1)}}\right)^{\tilde{c}}} - \sum_{i=1}^{n} [\ln x_i - \ln x_{(n)}] = 0, \tag{10}$$

[5].

There exists a simple test to see whether a Pareto model is appropriate [5]: the null hypothesis $H_0 : \nu = \infty$ is rejected if and only if $x_{(1)} < [nC/(-\ln q)]^{1/c}$, $0 < q < 1$, where $C = a^c$. The approximate p-value of this test is given by $p = \exp\left\{-nCx_{(1)}^{-c}\right\}$, and a small value of p indicates that the Pareto model is not a good fit; of course this is not enough *per se* to demonstrate the goodness of a truncated Pareto distribution.

Given a set of data it is often difficult to decide if they agree more closely with f or f_T, in that, in the interval $[a, b]$, they differ only by a multiplicative factor $1 - (a/b)^c$, which if the interval $[a, b]$ is not too small approaches 1 even for relatively small values of c. For this reason, rather than f and f_T, the distributions $P(X > x)$ and $P_T(X > x)$ are used, often called survival functions, which are given respectively by

$$P(X > x) = S(x) = 1 - F(x; a, c) = a^c x^{-c} \tag{11}$$

and

$$P_T(X > x) = S_T(x) = 1 - F_T(x; a, b, c) = \frac{c a^c \left(x^{-c} - b^{-c}\right)}{1 - \left(\frac{a}{b}\right)^c}. \tag{12}$$

The probabilities P and P_T have qualitatively different trends that are better observed in a log-log plot. In this case P is obviously represented by a straight line, whereas P_T exhibits also a almost linear trend with a sharp drop when x tends to b. To illustrate this point we

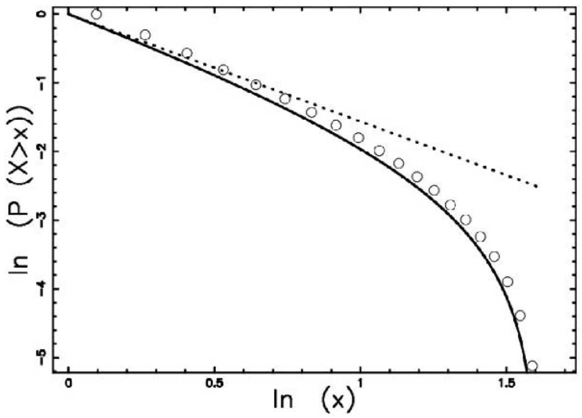

Figure 1. Log–log plot of the survival function: 10000 random data (empty circles), generated with Eq. (13), survival function of the truncated Pareto distribution (full line) and survival function of the Pareto distribution (dotted line).

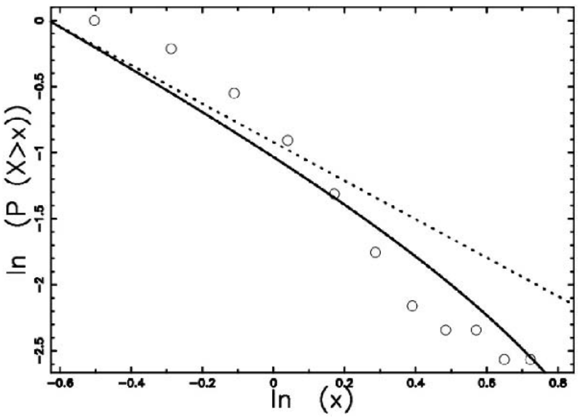

Figure 2. Log–log plot of the survival function of the mass distribution of the stars: data (empty circles), survival function of the truncated Pareto PDF (full line) and survival function of the Pareto PDF (dotted line). A complete sample (main sequence V) is considered with parameters as in Table 1.

have generated a set of $n = 10000$ random points drawn from a truncated Pareto distribution, via the formula

$$X : a, b, c \sim a\left(1 - R\left(1 - \left(\frac{a}{b}\right)^{c}\right)\right)^{-\frac{1}{c}}, \qquad (13)$$

where R is the unit rectangular variate, and we have fitted them with S and S_T respectively, see Figure 1.

3. Applications

3.1. Mass of stars

The sample of stellar masses has been obtained from the Hipparcos data as a function of the absolute magnitude and (B-V) [19].

Results of the fitting with P and P_T are presented in Table 1 where a, b, c and n, the number of sample elements, are reported and in Figure 2 that shows the data with the fit.

In this case in the range $3.44\mathcal{M}_{\odot} > \mathcal{M} \geq 0.53\mathcal{M}_{\odot}$, see Table 1 , the coefficient $\alpha = c + 1$ =2.45 is in agreement with modern estimates [8]. In this case, the power of the Pareto test results to be $p = 0.032$, indicating that the Pareto distribution is not a good fit, as can also be seen from Figure 2 .

Table 2 therefore reports the χ^2 of the fit of the star masses when using the Pareto and the truncated Pareto, respectively.

Table 1. Coefficients of mass distribution of the stars in the first 10 pc, of a complete sample (main sequence V). The parameter c is derived trough MLE and $p = 0.032$.

a $[\mathcal{M}_{\odot}]$	b $[\mathcal{M}_{\odot}]$	c	n	$P(X > x)$
0.53	3.44	1.45	52	Truncated Pareto
0.53	∞	1.77	52	Pareto

Table 2. χ^2 of different distributions when the number of bins is 5 for the stars in the first 10 pc.

Distribution	χ^2
Pareto	7.1
Truncated Pareto	5.26

3.2. Distribution of asteroid size

Supposing that not just the masses of stars but also those of other astrophysical objects have a power law tail, then is not difficult to prove also that their linear dimension, radii or diameters, must follow a power law. We have tested this hypothesis by considering the diameters of different families of asteroids, namely, Koronis, Eos and Themis.

In the following the sample parameter of the families are reported in Table 3, Table 4 and Table 5 , whereas Figure 3, Figure 4, Figure 5 report the graphical display of data and the fitting distributions.

In case of the Koronis family P_T fits the data better than P and indeed $p = 0.039$ is correspondingly small, whereas for the Eos family, P performs slightly better than P_T

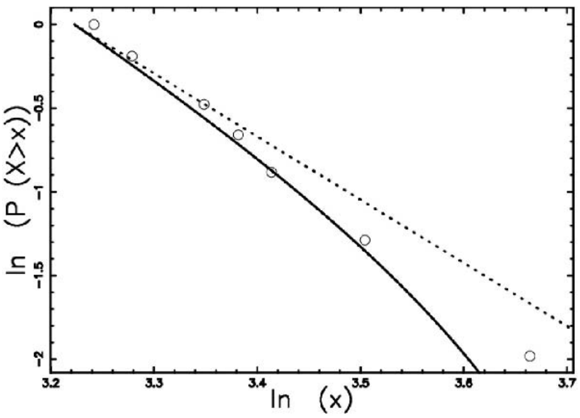

Figure 3. Ln–ln plot of the survival function of the diameter distribution of the Koronis Family: data (empty circles), survival function of the truncated Pareto PDF (full line) and survival function of the Pareto PDF (dotted line). A complete sample is considered with parameters as in Table 3.

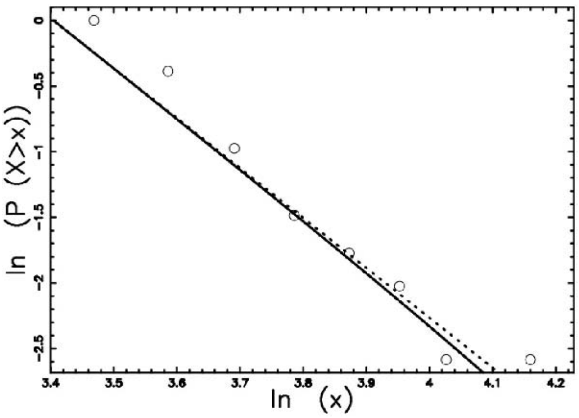

Figure 4. Ln–ln plot of the survival function of the diameter distribution of the Eos Family: data (empty circles), survival function of the truncated Pareto PDF (full line) and survival function of the Pareto PDF (dotted line). A complete sample is considered with parameters as in Table 4.

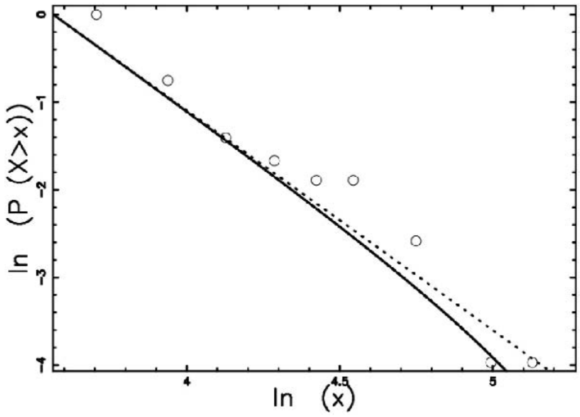

Figure 5. Ln–ln plot of the survival function of the diameter distribution of the Themis Family: data (empty circles), survival function of the truncated Pareto PDF (full line) and survival function of the Pareto PDF (dotted line). A complete sample is considered with parameters as in Table 5.

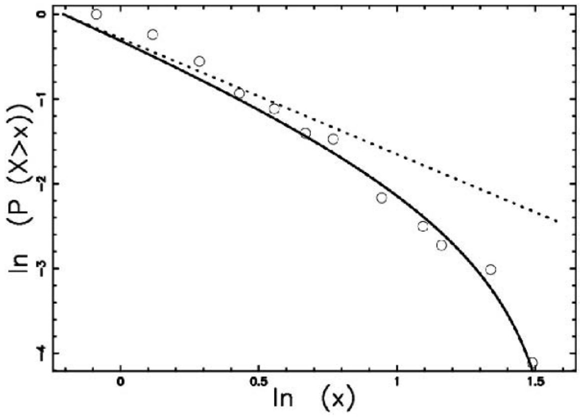

Figure 6. Log–log plot of the survival function of the mass distribution for the primeval nebula when $m \geq 0.5\mathcal{M}_{\odot}$ are considered. The truncated Pareto parameters are $c = 1.36$ and $p = 0.0058$.

(p=0.68), and the estimated of c are very closed in both cases. Finally in the third case, the Themis family, the two distributions are the same, due to the fact that the ratio $a/b = 0.14$ is small.

4. Generating Pareto tails

As a simple example of how a distribution with power can be generated, consider the growth of a primeval nebula via accretion, which is the process by which nebulae “capture” mass. We start by considering a uniform PDF for the initial mass of N primeval nebulae, m, in a range $m_{min} < m \leq m_{max}$. At each interaction the i-th nebula has a probability λ_i to increase its mass m_i that is given by

$$\lambda_i = (1 - \exp(-akm_i)), \tag{14}$$

where ak is a parameter of the simulation; thus more “massive” nebulae are more likely to grow via accretion. The quantity by which the primeval nebula can grow varies with time, in order to take into account that the total mass available is limited,

$$\delta m(t) = \delta m(0) \exp(-t/\tau), \tag{15}$$

Table 3. Coefficients of the diameter distribution of the Koronis family. The parameter c is derived through MLE and $p = 0.033$.

a [km]	b [km]	c	n	$P(X > x)$
25.1	44.3	3.77	29	truncated Pareto
25.1	∞	5.04	29	Pareto

Table 4. Coefficients of the diameter distribution of the Eos family . The parameter c is derived through MLE and $p = 0.681$.

a [km]	b [km]	c	n	$P(X > x)$
30.1	110	3.80	53	truncated Pareto
30.1	∞	3.94	53	Pareto

Table 5. Coefficients of the diameter distribution of the Themis family. The parameter c is derived through MLE and $p = 0.67$.

a [km]	b [km]	c	n	$P(X > x)$
35.3	249	2.5	53	truncated Pareto
35.3	∞	2.6	53	Pareto

where $\delta m(0)$ represents the maximum mass of exchange and τ the scaling time of the phenomena. The simulation proceeds as follows: a number r, is randomly chosen in the interval $[0, 1]$ for each nebula, and, if $r < \lambda_i$, the mass m_i is increased by $\delta m(t)$, where t denotes the iteration of the process. The processes proceed in parallel: at each temporal iteration all the primeval nebulae are considered. Results of the simulations have been fitted with both Pareto survival distributions. see Figure 6.

Due to a photometric effect [19] the sample of observed stars is complete only for $m \geq 0.5\mathcal{M}_{\odot}$. We therefore have set the lower boundary of the masses to $0.5\mathcal{M}_{\odot}$, and the resulting subset has been fitted with the Pareto and truncated Pareto survival distributions, Figure 6. It should be noted that the results of the simulation give $c = 1.36$, that is $\alpha = 2.36$ in agreement with the experimental estimate.

5. Conclusions

Results of the analysis presented here show that the truncated Pareto distribution provides a good fit for the distribution and performs better than the usual Pareto distribution. When the asteroid diameters are considered the situation is not so clear in that it depends on the family one considers. It is also clear the there can be cases, such as with the Themis family, in which the ratio between the minimum and the maximum value of the sample is so small that there no real difference between the two distributions. Finally we have shown that Pareto distributions can result from a simple growth process, in which the increase of the state variable (here mass) depends on the values taken in the previous state; furthermore results of the simulations agree well with the experimental data.

As remarked earlier, distributions are not the only statistics with a power law tail; in astrophysics alternatives have also been proposed for the statistics of asteroid diameters (e.g. [20]). However, Pareto distributions are particularly simple; for instance note that they have just a free parameter c, the others, a and b, being determined by the minimum and maximum values of the sample, respectively.

References

[1] V. Pareto, Cours d' economie politique (Rouge, Lausanne, 1896)

[2] M. Evans, N. Hastings, P. Peacock, Statistical Distributions, 3rd edition (John Wiley & Sons Inc, New York, 2000)

[3] A. Cohen, B. Whitten, Parameter Estimation in reliability and Life Span Models (Marcel Dekker, New York, 1988)

[4] D. Devoto, S. Martnez, Math. Geol. 30, 661 (1998)

[5] I. Aban, M. Meerschaert, A. Panorska, J. Am. Stat. Assoc. 101, 270 (2006)

[6] K. Rehfeldt, J.M. Boggs, L.W. Gelhar, Water Resour. Res. 28 , 3309 (1992)

[7] E.E. Salpeter, Astrophys. J. 121, 161 (1955)

[8] P. Kroupa, Monthly Notices of the RAS 322, 231 (2001)

[9] K.R. Lang, Astrophysical formulae (Springer, New York, 1999)

[10] R. Schlickeiser, Cosmic ray astrophysics (Springer, Berlin, 2002)

[11] E.M. Rossi, Nuovo Cimento C 28, 387 (2005)

[12] J.S. Bloom, D.A. Frail, R. Sari, Astron. J. 121, 2879 (2001)

[13] M. Abramowitz, I.A. Stegun, Handbook of mathematical functions with formulas, graphs, and mathematical tables (Dover, New York, 1965)

[14] D. von Seggern, CRC Standard Curves and Surfaces (CRC, New York, 1992)

[15] W.J. Thompson, Atlas for computing mathematical functions (Wiley-Interscience, New York, 1997)

[16] I. Gradshteyn, I. Ryzhik, Table of Integrals, Series, and Products (Academic Press, San Diego, 2000)

[17] A. Prudnikov, O. Brychkov, Y. Marichev, Integrals and Series (Gordon and Breach Science Publishers, Amsterdam, 1986)

[18] M. Masoom Ali, S. Nadarajah, Comput. Commun. 30,

1 (2006)

[19] L. Zaninetti, Astronomische Nachrichten 326, 754 (2005)

[20] L. Zaninetti, A. Cellino, V. Zappala, Astronom. Astrophys. 294, 270 (1995)