Journal of Environmental Statistics
February 2013, Volume 4, Issue 6. http://www.jenvstat.org

Characterization Theorems Based on Conditional Quantiles with Applications

Ratan Dasgupta
Indian Statistical Institute

Abstract

We prove a characterization of Pareto variable based on quantiles when conditional distribution above a threshold is considered. A similar characterization for exponential distribution is also obtained. The results are extended to discretized random variables. For some well known distributions, effect of conditioning the variable crossing a threshold on quantiles is investigated. The results are further extended to bivariate exponential and bivariate Pareto type models which are relevant to explain lifestyle data. Applications of the results are made in estimation of conditional quantiles in environmental data. Pareto model for excess of flood-peaks of a river seems to be satisfactory with high threshold values. Applications are also made on Yam-yield, wind speed data of high energy due to extratropical cyclones in coastal regions and worldwide earth-tremor data.

Keywords: Pareto model, quantiles, Cauchy functional equation, lifestyle data, elephant foot yam, extratropical cyclones, peak gust wind.

1. Introduction

Pareto distribution and its variants have wide applications in modeling different branches of science, especially economics. Apart from explaining the distribution of wealth or income, this distribution may explain observed phenomena in sociology, anthropology, hydrology, meteorology, actuarial science, occupational health and safety etc.
See e.g., Cebrian, Denuit, and Lambert (2003), Dasgupta (2011), Jenkinson (1955), Klass, Biham, Levy, Malcai, and Soloman (2006), Krishnaji (1970), de Oliveira, Ebecken, de Oliveira, and Gilleland (2011), Morrow-Tlucak, Emhart, Sokol, Martier, and Ager (1989), Van Montfort and Witter (1985).

In this paper we prove a characterization of Pareto distribution based on quantiles when the

variable exceeds a threshold. Comparisons are made between the unrestricted quantiles of original variable, and restricted quantiles for the variable above a threshold. The relationship of constant ratio of unrestricted and restricted quantiles of variable beyond a threshold is seen to characterize the Pareto distribution. The proof involves solving functional equations, like Cauchy functional equation, over a restricted zone. A similar characterization based on constant shift of restricted quantiles from unrestricted quantiles is proved for exponential random variable. The results are generalized for discretized version of the random variables. Effect of conditioning the variable above a threshold on quantiles is discussed for some well known distributions including normal distribution. The case when underreporting of a variable is of exponential order that seems realistic in some specific situations is also studied. We further study the conditional quantiles to obtain relevant characterizations for bivariate exponential and bivariate Pareto type models those are useful in explaining lifestyle data. Applications of the results are made in different contexts including environmental data.

In section 2 we prove the results for Pareto distributions and exponential distributions. Similar characterizations for discrete random variables are proved in section 3. In section 4 we obtain results for bivariate exponential and bivariate Pareto type distributions. Section 5 discusses applications of the results in agricultural data of yam yield, environmental data on flood-peak, earth-tremor and peak gust (PGU) wind velocity. The value of median PGU exceeding the recorded maximum is estimated based on unrestricted median and recorded maximum peak gust, thus providing a glimpse of the scenario beyond observed range.

2. Characterization of Pareto and exponential distribution

We first prove the following.

Theorem 1. Let X be a random variable with support $(a, \infty), a > 0$, and distribution function F. Denote $c = c(p)$ to be the unrestricted p-th quantile of $X(> a)$, and consider p in a (small) dense neighborhood A_0 of origin (e.g., $p \in A_0 = (0, \epsilon) \cap Q, \epsilon > 0$, small and Q is the set of rational numbers). Then the p-th quantile of the distribution, $p \in A_0$, under the restriction $X > x_0(> a)$ is cx_0/a iff F is a Pareto distribution function.

Proof. Consider the distribution function of standardised Pareto variable with $a = 1$.

$$F(x) = 1 - x^{-\alpha},\ x > 1,\ \alpha > 0 \qquad ...(2.1)$$

The median of the distribution is at $2^{1/\alpha}$. Denote $\overline{F} = 1 - F,\ g(x) = \log \overline{F}(x) = -\alpha \log x \downarrow -\infty,\ x \uparrow \infty$. The c.d.f. of the variable, given that $x > x_o(> 1)$, then turns out to be $F(x)/\overline{F}(x_0)$, and one may write $P(X > x|X > x_0) = \frac{\overline{F}(x)}{\overline{F}(x_0)} = (\frac{x}{x_o})^{-\alpha}$. Equating this to 0.5 we obtain the new median of the random variable crossing the threshold x_0 as cx_0, where $c = 2^{1/\alpha}$ is the median of the random variable $X(> 1)$.
This property specifies the form of the distribution at the points $c, c^2, \cdots, c^m, \cdots$ as explained below.
For a general distribution function $F = F(x)$ of the random variable $X > 1$, denote $g(x) = \log \overline{F}(x) = \log(1 - F(x))$. Suppose that the new median of the random variable X under the

restriction $x > x_0$ is at cx_0, where c is independent of x_0. Indeed c is the median of original unrestricted random variable as seen by taking $x_0 \downarrow 1$. Next, write

$$e^{g(cx_0)-g(x_0)} = \frac{\overline{F}(cx_0)}{\overline{F}(x_0)} = 0.5 \qquad ...(2.2)$$

This provides,

$$g(cx_0) - g(x_0) = -k \qquad ...(2.3)$$

where, $k = \log 2$.
Thus $g(c^2) = g(c) - k = -2k,\ g(c^3) = -3k, \cdots, g(c^m) = -mk$. This implies the type of the distribution function is Pareto, $g(x) = \log \overline{F}(x) = -\alpha \log x$; where $\alpha = k/(\log c)$ at the points $x = c, c^2, \cdots, c^m, \cdots$

Note that a similar relation holds for the third quartile of the Pareto distribution (2.1) with $c = 4^{1/\alpha},\ k = \log 4$.

Thus equations (2.2)-(2.3) for third quartile of a general F imply Pareto distribution for some other points $x = c, c^2, \cdots, c^m, \cdots$ with a different choice of c.
Now assume that the above property of constant multiple factor of restricted and unrestricted quantiles holds for a dense set of quantiles corresponding to $p \in (0, 1)$, p rational. The form of difference equation then reduces to

$$g(cx_0) - g(x_0) = \log(1 - p) = -k \qquad ...(2.4)$$

$k = -\log(1 - p) > 0,\ c = (1 - p)^{-1/\alpha};\ p \in Q \cap (0, 1)$.

This specifies the distribution function F to be Pareto in a dense set $x = c, c^2, \cdots, c^m, \cdots$, of $(1, \infty)$. For an arbitrary real number $z > 1$, there exist integer m and $c = (1 - p)^{-1/\alpha};\ p \in Q \cap (0, 1)$ such that c^m is arbitrary close to the number z, where Q is the set of all rational numbers. Next from right continuity of distribution function, the form of F is Pareto at z, where $z > 1$ is arbitrary.
Finally, a dense choice of p in a small neighborhood of origin, e.g., $p \in A_0 = (0, \epsilon) \cap Q, \epsilon > 0$, small suffices for the Theorem to hold; as the resultant sequence $\{c^m : m = 1, 2, 3, \cdots\}$ still spans a dense support of the variable.

For the general case let the minimum possible value of X be $a > 0$. The Pareto distribution function F with minimum value a is then

$$F(x) = 1 - (x/a)^{-\alpha},\ x > a(> 0),\ \alpha > 0 \qquad ...(2.5)$$

One may then consider the transformed random variable $X/a(> 1)$. Proceeding as before the characterization of Theorem 1 holds.

Next we state a similar result for exponential variable.

Theorem 2. Let X be a random variable with support $(a, \infty), a \geq 0$, and distribution function F. Denote $c = c(p)$ to be the unrestricted p-th quantile of $X(> a)$, and consider p in

a (small) dense neighborhood A_0 of origin (e.g., $p \in A_0 = (0, \epsilon) \cap Q, \epsilon > 0$, small and Q is the set of rational numbers). Then the p-th quantile of the distribution, $p \in A_0$, under the restriction $X > x_0 (> a)$ is $c + x_0 - a$ iff F is an exponential distribution function.

Proof. For exponential random variable Y with distribution function

$$G(y) = 1 - e^{-\lambda y}, y > 0 \qquad ...(2.6)$$

it is easy to see that the p-th quantile of the distribution under the restriction $Y > y_0 (> 0)$ is merely a shift of the unrestricted quantile by y_0.

$$P(Y > y | Y > y_0) = e^{-\lambda(y - y_0)} = 1 - p \Rightarrow y = y_0 - \frac{1}{\lambda} \log(1 - p) = y_0 + \xi_Y(p) \qquad ...(2.7)$$

where $\xi_Y(p) = -\frac{1}{\lambda} \log(1 - p)$ is the p-th quantile of $Y > 0$.

This property characterizes the exponential distribution.
To see this for a general random variable Y with distribution function G and $y > y_0 (> 0)$, assume that $(y_0 + c)$ to be the new p-th quantile; shifted from unrestricted p-th quantile c by y_0. Then write in a similar fashion as in (2.4),

$$e^{g(y_0+c)-g(y_0)} = \frac{1 - G(y_0 + c)}{1 - G(y_0)} = P(Y > y_0 + c | Y > y_0) = 1 - p = e^{\log(1-p)} = e^{-k} \qquad ...(2.8)$$

leading to the equation

$$g(y_0 + c) - g(y_0) = -k \qquad ...(2.9)$$

where $g(x) = \log \overline{G}(x), \overline{G} = 1 - G$.
One may solve (2.9) in a similar fashion as in (2.4), with the resultant solution of the form $g(x) = -\lambda x$, leading to the exponential distribution. To see this write $g(mc) = g((m-1)c) - k = \cdots = -mk$, and g is seen to be linear on the points $c, 2c, 3c, \cdots, mc, \cdots$ thus implying exponential distribution at those points. Theorem 2 is then immediate following similar steps of proof as in Theorem 1.
A dense choice of p in a small neighborhood of origin, $A_0 = (0, \epsilon) \cap Q, \epsilon > 0$, small suffices for the Theorem to hold; the resultant sequence $\{mc : m = 1, 2, 3, \cdots\}$; $c = c(p), p \in A_0$ still spans a dense support of the variable.

Equation (2.9) is seen to be a variant of equation (2.4). Write $f(x) = g(e^x)$, then from (2.4), $g(e^{\log c + \log x_0}) - g(e^{\log x_0}) = -k$. That is $f(x) = g(e^x)$, is of the form (2.9) in log scale as $f(\log x_0 + \log c) - f(\log x_0) = -k$.

Pareto and exponential distributions are inter related as follows. If X is Pareto-distributed with minimum a and index δ, then $Y = \log(X/a)$ is exponentially distributed with intensity δ. Equivalently, if Y is exponentially distributed with intensity δ, then ae^Y is Pareto-distributed with minimum a and index δ. This relationship is reflected in the similarity of equations (2.4) and (2.9).

Remark 1. Equations (2.4) and (2.9) are related to Cauchy functional equation. The constants in the r.h.s. of these two equations are $-k = \log(1-p) = g(c)$. Thus these two equations

can be rewritten in the form $g(cx_0) = g(c) + g(x_0)$ and $g(y_0 + c) = g(y_0) + g(c)$, respectively. As already mentioned (2.4) and (2.9) are reformulations of each other. Variation of x_0 is due to shift of threshold, the other coordinate c varies as $p \in (0, \epsilon) \cap Q$ varies.
Apart from some pathological examples, the solutions of Cauchy functional equation $g(x+y) = g(x) + g(y)$ over R or R^+ is of the form $g(x) = \lambda x$.
In the present case $g(x) = \log(1 - F(x))$ is a monotone function on R^+.

Remark 2. The change in the value of quantile is a result of conditioning the random variable towards the tail of the distribution. When the tail is moderately decaying like exponential then shift in quantile equals shift in the threshold of the random variable. However, for a thick tailed distribution like Pareto with polynomial decay, shift of quantile is high towards tail; it is a constant (> 1) multiple of original quantile. In this context it is worthwhile to examine some other distributions and the status of normal distribution in the scenario. For exponential distribution (2.6) note that

$$\frac{\overline{G}(y)}{\overline{G}(y_0)} = \frac{e^{-\lambda y}}{e^{-\lambda y_0}} = (\frac{x}{x_0})^{-\lambda} \qquad ...(2.10)$$

writing $e^y = x$. Equating the r.h.s to $(1 - p)$ the value of restricted quantiles are obtained.
The above also shows the interrelation of exponential distribution with Pareto distribution and the corresponding shifts of quantiles when crossing of threshold $x_0 = e^{y_0}$ is considered for the transformed variable $X = e^Y$. Tail probability of Pareto variable X decays at slower rate $O(x^{-\lambda})$ compared to exponential decay $O(e^{-\lambda y})$ for the exponential variable Y. As a result restricted quantile of Pareto is wide apart from unrestricted quantile, compared to that for exponential variable.

Below we check the effect of crossing a (large) threshold on the quantiles of some other distributions.

1. Normal distribution. For a standardized normal variable Z

$$P(Z > z)/P(Z > z_0) = \Phi(-z)/\Phi(-z_0) \sim (z/z_0)^{-1} e^{-(z^2 - z_0^2)/2} \qquad ...(2.11)$$

where $z > z_0 (> 0)$, and z_0 is large. Thus an approximate value of the restricted p-th quantile for standardized normal distribution having crossed a high threshold $z_0 (> 0)$ is given by the following

$$z \approx [-2\log(1 - p) + z_0^2]^{1/2} \sim z_0[1 - \frac{1}{z_0^2}\log(1 - p)] \qquad ...(2.12)$$

From r.h.s. of (2.12) it is seen that the restricted p-th quantile tends to z_0 for large value of the threshold z_0.
For exponential distribution with relatively thick tail the difference between the restricted and unrestricted quantiles, as we have seen earlier, is the amount of shift in threshold value, i.e., the difference between the restricted quantile and the threshold value is a constant, viz., the unrestricted p-th quantile; irrespective of the value of threshold. However, for normal distribution with relatively fast decaying tail the difference $z - z_0 \approx -\frac{1}{z_0}\log(1-p) \to 0$, as $z_0 \to \infty$.

2. Weibull distribution. The standard Weibull distribution have cumulative distribution function

$$H(v) = 1 - e^{-v^k}, v \geq 0,\ k > 0 \qquad ...(2.13)$$

Solving the equation

$$\frac{\overline{H}(v)}{\overline{H}(v_0)} = \frac{e^{-v^k}}{e^{-v_0^k}} = (1-p) \qquad ...(2.14)$$

for $v > v_0 (> 0)$, one gets the restricted p-th quantile for Weibull distribution having crossed the threshold v_0 as, $v = [v_0^k - \log(1-p)]^{1/k}$.
For $k > 1,\ v \approx v_0[1 - \frac{1}{kv_0^k}\log(1-p)]$, and conclusion similar to normal distribution holds in this case.
The distribution (2.13) for $0 < k < 1$ has a lower order decay of tail probability than exponential distribution ($k = 1$), and the above analysis indicates that the difference between restricted and unrestricted qualtiles is more than the shift in threshold, whereas with $k > 1$ tail probability decays faster than exponential distribution, and the difference between restricted and unrestricted quantiles shrinks towards zero as the value of the threshold increases towards infinity.

3. Exponential underreporting and Pareto model. Underreporting to a high level of a variable may change the pattern of distribution of the variable of interest. The phenomenon of underreporting is present in many occasions like traffic injuries, HIV infection, credit card debt etc. In some cases it may be to the tune of 20 fold, e.g., gross underreported alcohol use in pregnancy, see de Oliveira *et al.* (2011). A model of exponential underreporting may then be more appropriate compared to underreporting to a multiplicative factor. Although in most of the present studies we consider income /wealth underreporting up to a multiplicative factor, it would be interesting to see how the model and relevant analysis change from traditional Pareto model (2.1), if we take into account the possibility of exponential underreporting for some specific cases as reported in de Oliveira *et al.* (2011). To this end consider the transformed random variable $U = e^X$ having a thicker tail than distribution $F(x) = 1 - x^{-\alpha},\ x > 1,\ \alpha > 0$ given in (2.1) of the reported Pareto variable X. Distribution of U has a slower decay of tail probability, viz., logarithmic decay $P(U > u) = (\log u)^{-\alpha}, u > e$; compared to polynomial decay in (2.1). The distribution has density of the form $f(u) = \alpha u^{-1}(\log u)^{-(\alpha+1)}, u > e$; which has a slower order decay than a Pareto density. As a result, the shift of restricted quantile under the condition of crossing a threshold is of higher magnitude than that for Pareto variable.
The shifted median under the restriction $U > u_0 (\geq e)$ is at $u_0^{2^{1/\alpha}}$, the shifted p-th quantile is at $u_0^{(1-p)^{-1/\alpha}},\ p \in (0,1)$.
The corresponding shifts for reported Pareto variable X mentioned (2.1) are $u_0 2^{1/\alpha}$ and $u_0(1-p)^{-1/\alpha},\ p \in (0,1),\ \alpha > 0$, these are of multiplicative order whereas that of the underlying unreported variable U are of power order.
Apart from the specific instance cited regarding alcohol consumption, the distribution of U may also be of interest to explain unaccounted gap between reported and unreported wealth. One may obtain Pareto type distributions from exponential distribution with random intensity following a beta prior, see Dasgupta (2011). A natural question arises whether it is possible to obtain from Bayesian consideration a distribution function with $P(U > u) = (\log u)^{-\alpha}, u > e$, having logarithmic decay; originating from Pareto distribution having polynomial decay. In

the following we answer the question in affirmative.
Such a representation provides a Bayesian insight into the situation when a traditional model fails in favor of an alternative model.

Proposition 1. Let the random variable X be Pareto distributed with density function $g(x|a) = ax^{-(a+1)}, x > 1, a > 0$. For a fixed $a > 0$, let $U = e^X|a$. Suppose a has a prior gamma density
$f_{\beta,p}(a) = \frac{\beta^p}{\Gamma(p)} e^{-a\beta} a^{p-1},\ \beta > 0,\ p > 0.$
Then the marginal density of X has similar decay as that of U, i.e., a monotonically decreasing density with decay lower than Pareto density,
$f(x) = p\beta^p x^{-1} (\log x + \beta)^{-(p+1)},\ x > 1.$

Proof. Follows from integrating the joint density $g(x|a)f_{\beta,p}(a)$ with respect to a.
The marginal density of X remains bounded at $x = 1$ for every fixed $\beta > 0$. However, this blows up at the rate $p\beta^{-1}$ as $\beta \downarrow 0$. Height of relative histogram near left end point 1 may provide an estimate of p/β, mode of the distribution.

Proposition 1 has following implication. A typical heavy tailed distribution of a phenomenon may follow a Pareto model. However, aggregate of different groups having random Pareto indices following e.g., a gamma density may result in a heavy tailed distribution that may be more realistic in some situations.

3. Characterization theorems for discrete random variables

Consider a random variable X with support either $\mathbf{N_0}$, the set of nonnegative integers; or set of positive integers $\mathbf{N_1} = \mathbf{N_0} - \{\mathbf{0}\}$. Let the cumulative distribution function of X be denoted by $F(x) = P(X \leq x)$, it is enough to define F at integer values. For $p \in (0, 1)$ the p-th quantile of F is defined as $F^{-1}(p) = \{\inf x : F(x) \geq p\}$.
The following two theorems are the counterparts of Theorem 1-2 stated for discrete random variables.

Theorem 3. For a random variable X with support $\mathbf{N_1}$ and distribution function $F(x) = P(X \leq x)$, let the p-th quantile of the distribution under the restriction $X \geq x_0 (\in \mathbf{N_1})$ be cx_0; where $c \in \mathbf{N_1}$ is the unrestricted p-th quantile of X. The above property holds for all p of the form $p = p_i = \sum_{j=1}^{i} P(X = j),\ i = 1, 2, 3, \cdots$ iff $F(x) = 1 - x^{-\alpha}$ for some $\alpha > 0$, where $x \in \mathbf{N_1}$.

Theorem 4. For a random variable X with support $\mathbf{N_0}$ and distribution function F, let the p-th quantile of the distribution under the restriction $X \geq x_0 (\in \mathbf{N_0})$ be $c + x_0$, where $c \in \mathbf{N_0}$ is the unrestricted p-th quantile of X. The above property holds for $p = p_1 = \sum_{j=0}^{1} P(X = j)$, iff F is a geometric distribution function on $\mathbf{N_0}$.

Proof. Theorems 3-4 follow similar lines as that of Theorems 1-2. One way implications of the Theorems are easy to see. Consider the 'only if' part.

In the case of Theorem 3, steps similar to (2.2)-(2.4) hold. The variable X has support $\mathbf{N_1}$. This set is same as the set $\{c, c^2, \cdots, c^m, \cdots\}$, where $c = c(p)$ is the p-th quantile of X, and p of the form $p = p_i = \sum_{j=1}^{i} P(X = j),\ i = 1, 2, 3, \cdots$. The p-th quantile is then an integer, as the jumps of F ocurr at integer points. For example when $F(x) = 1 - x^{-\alpha}$, the p-th quantile $c = c(p) = (1-p)^{-1/\alpha}$ is obtained as the solution i of the equation $p = p_i = \sum_{j=1}^{i} P(X = j) = 1 - i^{-\alpha}$.
Over the set $\mathbf{N_1}$, characterization for $g(x) = \log \overline{F}(x) = -\alpha \log x$ is seen to hold in a similar fashion like in Theorem 1.
The proof for 'only if' part of Theorem 4 is similar to Theorem 2 along the above lines. However, note that in this case the set $\mathbf{N_0} - \{\mathbf{0}\}$ is also spanned by $\{c, 2c, \cdots, mc, \cdots\}$, where $c = c(p)$ is the p-th quantile of X, $p = p_1 = \sum_{j=0}^{1} P(X = j)$, i.e., $c = c(p) = c(p_1) = 1$. Thus the characterization for $g(x) = \log \overline{F}(x)$ holds with the solution $g(x) = -\lambda x$ over $\mathbf{N_0} - \{\mathbf{0}\}$, on the condition of restricted quantile for $p = p_1$ only. Since the total probability is 1, the probability mass at origin is taken care of and Theorem 4 holds.

Modeling with above two discrete distributions depends on the tail behavior of observed frequency distributions. Distribution $F(x) = 1 - x^{-\alpha}$ for some $\alpha > 0$, where $x \in \mathbf{N_1}$ may be termed as Discrete Pareto distribution. This may be an appropriate model for grouped Pareto variable, grouped over class intervals of equal length.

4. Bivariate exponential model and related distributions

Consider a bivariate exponential distribution, with exponential marginal. The relation $Y = X + Z,\ Z \geq 0$, is a special case of a more general model

$$Y = aX + Z,\ a > 0,\ Z \geq 0, \qquad \ldots(4.1)$$

where Z is independent of X.
This distribution has application in lifestyle data to explain the number of future physical relationships for an individual, given the past in a social environment where tie from past is loose, see Dasgupta (2011).
The restriction that the marginal distributions of X and Y are exponential with respective intensities λ_x and λ_y requires that the distribution of Z is of the form $a\rho + (1-a\rho)(1-e^{-\lambda_y z}),\ \rho = \lambda_y/\lambda_x = \mu_x/\mu_y,\ z \geq 0$. This is distribution of a random variable that is a product of two independent random variables - a Bernoulli random variable with mean $(1-a\rho)$ and an exponential random variable with parameter λ_y. See Iyer, Manjunath, and Manibasakan (2002).
The Bernoulli variable takes the value zero with probability $a\rho$, thus $Z = 0$, with a positive probability $a\rho$. Under this exponential model, the correlation between the two random variables are $r_{x,\,y} = a\rho$.
Now, the restriction $X > x_0 (\geq 0)$, shifts the quantiles of X by the same magnitude from unrestricted quantiles, and this imposes a restriction on the exponential random variable Y as $Y > ax_0$, hence the restricted quantiles of the latter random variable is shifted by ax_0, from the corresponding unrestricted quantiles.

The converse is also true, we have the following Proposition.

Proposition 2. Consider the model $Y = aX + Z,\ a > 0,\ Z \geq 0$, where Z is independent of X. Under the restriction $X > x_0 (\geq 0)$, let the quantiles of X be shifted by x_0. Let the resultant restriction $Y > ax_0$, shifts the quantiles of Y by ax_0. Then both the variables X and Y are exponential, and the distribution of Z is of the form $a\rho + (1 - a\rho)(1 - e^{-\lambda_y z}),\ \rho = \lambda_y/\lambda_x = \mu_x/\mu_y,\ z \geq 0$.

Proof. From the characterization of exponential distribution via conditional quantiles, it follows that both X and Y have exponential marginal. The result then follows from the assumed relation $Y = aX + Z,\ a > 0,\ Z \geq 0$, where Z is independent of X.

Next we investigate a bivariate Pareto model in terms of conditional quantiles. It is possible to obtain Pareto type distributions from exponential distribution with random intensity following a beta prior. The following result is proved in Dasgupta (2011).

Theorem A. Let the random variable X be exponentially distributed with density function $g(x|\theta) = (-\log\theta)\theta^x,\ x > 0, 0 < \theta < 1$, where θ has a prior beta density
$f_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ \alpha > 0,\ \beta > 0.$
Then the marginal distribution of X is approximately Pareto with monotonically decreasing density having polynomial decay
$f(x) = O_e((x+\alpha)^{-(\beta+1)}),\ x > 0.$

It may not be out of place to mention that in view of Theorem A along with Proposition 1, starting from exponential density it is possible to obtain a monotonically decreasing density with decay lower than Pareto density viz., $f(x) = O_e(x^{-1}(\log x + \beta)^{-(p+1)}), \beta > 0, p > 0,\ x > 1$; via a two stage prior of beta density and gamma density, as mentioned in Theorem A and Proposition 1. This step wise reduction provides a Bayesian insight when a candidate exponential model is replaced, from data viewpoint, by a heavy tailed distribution. Possible fluctuation of parameters over heterogeneous groups/items in a population, governed by beta and gamma distributions may explain such phenomena.
Consider the model (4.1). The intensities of Y and aX are λ_y and λ_x/a respectively. As in Dasgupta (2011) associate a beta prior $f_{\alpha,\beta}(\theta)$ of Theorem A, on $\theta = e^{-\lambda_y}$. In the r.h.s. of (4.1), this induces a prior on $e^{-\lambda_x} = e^{-\lambda_y/\rho}$, where $\rho = \lambda_y/\lambda_x$ is considered to be a constant. Integrating both sides of (4.1) with respect to the prior probability on θ, we then have the relationship, see Dasgupta (2011);

$$Y^* = aX^* + Z^*,\ a > 0,\ Z^* \geq 0 \qquad ...(4.2)$$

where the transformed variables $X^*,\ Y^*$ have polynomially decaying densities as given in Theorem A.
For a Pareto variable the conditional quantile of the variable crossing a threshold is a constant multiple of the shift. Thus the conditional quantile of X^* under the restriction $X^* > x_0^*$ is approximately a constant multiple of unrestricted quantile, and this restriction on X^* imposes the restriction $Y^* > ax_0^*$ on Y^*, which is approximately a Pareto variable having polynomially decaying density. Thus the conditional quantiles of Y^* is also approximately a constant multiple of unrestricted quantiles. The variable Z^* is the product of two independent random variables - a Bernoulli random variable with mean $(1 - a\rho)$ and an approximately Pareto

random variable with same parameter as that of Y^*. Unlike the earlier case Z^* may not be independent of X^*, as conditional independence and marginal independence are not related in general. The parameter $\beta(>0)$ quantifies the dispersed nature of the transformed variables obtained from original exponential distribution. Smaller the value of β, more dispersed is the transformed variable with heavy tail caused by diversity of individual intensities under consideration.

Such bivariate models are useful when value of one random variable is necessarily bounded below by the other, e.g., maximum diameter vs. minimum diameter of an approximate oval object in industrial production, number of relationships / physical encounters of an individual up to two successive time points from a common start, see e.g., Dasgupta (2011).
Observed frequency distributions may reveal sharp fall like exponential or, these may have relatively thick tails with approximate polynomial decay suggesting Pareto model. One may study simultaneous behavior of the conditional quantiles of two variables under the restriction of crossing thresholds, to search for an appropriate bivariate exponential or Pareto model.

In univariate case, there are situations when one is interested in studying the large values of the random variable with distribution having a thick tail, i.e., the behavior of the variable near the thick tail is of interest. In some cases Pareto model may provide a reasonable fit when the value of the variable exceeds some high threshold value. The Pareto fit is equivalent to constant multiplicative factor of restricted and unrestricted quantiles, former may then be computed in terms of the latter, thus providing magnitude of restricted quantiles indicating how large the variable can be near the tail.

5. Some examples

It is well known that Pareto distribution may explain the uneven distribution of wealth and income. Therefore the above mentioned property of constant multiplicative factor of restricted and unrestricted quantiles holds in such situations. We may examine the validity of such assumption in other situations. The above property regarding constant multiplicative factor of quantiles may hold for variable beyond a large threshold value. In such a situation Pareto distribution is appropriate above that threshold value.
The characterization provides the magnitude of shift in quantiles due to shift of threshold. In real life situations one may check the stability of shifted quantiles observed over several repetitions. Such stability of conditional quantiles of multiplicative form cx_0 in empirical distributions, crossing a threshold x_0 may indicate a Pareto model. The same may be said about exponential model by examining the stability of conditional quantiles of the form $c+x_0$ in empirical distributions, crossing a threshold.
In the following we check for Pareto model fit near the tail via R^2 of regression.

Example 1. *High tide water level at Arabian Sea*
High tide at sea causes tidal bore, a high tidal wave experienced in a narrow river or estuary that may cause substantial damage to lives and properties of inhabitants in nearby localities. Very high water levels are of concern.

The following data in feet, relates to high tidal range at Arabian Sea, west coast of India near Alang ship cycling yards. Each of these 170 observations was taken as maximum of two observations at different tide times, viz. at early hours and evening/night hours in a day. Thus, the observations represent the maximum height of sea water level in a 24 hour cycle. The data is spread over first six months in a year. Observation for a day is not taken into consideration, if any one of the two tide readings is missing in that day. The recorded observations are as follows.

36.39,36.65,36.46,35.70,34.39,32.45,30.32,30.22,30.62,31.44,32.32,33.01,33.40,
33.60,33.60,33.47,33.21,32.75,31.99,30.91,29.43,28.09,28.22,28.94,30.42,32.19,
33.96,35.50,36.59,37.08,36.95,36.10,34.49,32.16,30.88,30.09,30.06,30.58,31.27,
31.90,32.39,32.75,32.95,32.91,32.55,31.86,30.81,30.25,29.80,29.20,29.01,29.89,
31.60,33.54,35.21,36.39,36.82,36.46,35.31,34.36,32.95,31.14,29.47,28.71,28.94,
29.66,30.48,31.21,31.73,32.03,31.99,32.32,32.49,32.39,31.96,31.21,30.16,29.40,
29.83,31.27,32.98,34.42,35.28,36.36,36.72,36.26,35.08,33.31,31.24,29.24,27.99,
27.86,28.35,29.07,29.99,31.37,32.52,33.31,33.80,33.93,33.77,33.21,32.26,31.11,
30.25,30.22,30.98,32.39,34.62,36.23,37.01,37.01,36.29,34.95,33.31,31.44,29.60,
28.19,27.47,27.47,29.14,30.78,32.22,33.40,34.26,34.78,34.95,34.75,34.23,33.37,
32.35,31.37,30.68,31.44,33.50,35.18,36.23,36.59,36.36,35.60,34.52,33.24,31.86,
30.45,29.04,27.79,26.81,29.47,31.04,32.55,33.80,33.80,35.41,35.70,35.70,35.37,
34.68,33.60,32.19,30.55,32.22,33.60,34.65,35.28,35.41,35.21,34.72,34.06,33.31,
32.39

Figure 1 of $\log x$ vs. $-\log(1-F(x))$ suggest that Pareto model may be appropriate beyond a large threshold value rather than the whole data set. In Figure 2 the same is plotted for $\log x > 3.55$ with 37 observations. The fit now seems better with squared value correlation as $R^2 = 0.8138$, and estimated value of $\alpha = 44.5849$.
With a further increase of the threshold value to $\log x > 3.58$ the Pareto fit (2.5) with $a = e^{3.58}$ to 20 observations seems more appropriate; providing the value of $\alpha = 102.6164$ from least square regression fit with a high value of $R^2 = 0.9207$.

Example 2. *Growth model for Elephant foot yam*
The following data relates to weights in kilogram of 100 yams from a growth experiment conducted in the year 2010 at Indian Statistical Institute, Giridih farm. We check the appropriateness of Pareto fit, especially beyond a threshold value.

4.50, 3.20, 2.60, 3.15, 2.05, 2.10, 2.65, 0.80, 1.70, 1.15, 2.90, 3.50, 4.35, 3.85, 3.60, 1.30, 2.20,
1.70, 3.70, 2.50, 3.40, 3.10, 4.45, 5.60, 4.15, 1.50, 1.90, 2.00, 3.10, 3.00, 3.10, 2.25, 2.65, 2.90,
3.60, 1.50, 1.20, 0.70, 2.80, 2.70, 3.75, 2.05, 1.60, 1.50, 3.60, 2.20, 1.40, 1.20, 0.00, 2.40, 2.50,
1.45, 1.05, 0.70, 0.00, 2.25, 2.00, 2.45, 1.55, 0.90, 0.75, 2.65, 2.25, 1.20, 2.25, 2.00, 3.80, 3.00,
3.00, 2.35, 1.05, 0.80, 3.80, 2.30, 3.80, 1.60, 0.00, 3.60, 1.60, 4.00, 3.00, 1.95, 2.00, 3.65, 3.60,
1.40, 1.40, 1.30, 3.90, 3.60, 5.50, 2.90, 2.60, 1.70, 2.80, 1.90, 1.70, 1.80, 1.10, 2.80.

In Figure 4 with 97 nonzero yam data we plot $\log x$ vs. $-\log(1-F(x))$ and observe that Pareto model for Yam yield may be appropriate beyond a large threshold value, much like

the earlier data on sea tide. Figure 5 plots the same for $\log x > 1$ with 38 observations. The fit of Pareto model (2.5) with $a = e$ now seems better as $R^2 = 0.9460$, estimated value of $\alpha = 5.7038$.
If the threshold value is increased slightly further to $\log x > 1.2$ as shown in Figure 6, we have 23 observations and a further increase in $R^2 = 0.9637$, providing a value of $\alpha = 7.0045$ for the model (2.5) with $a = e^{1.2}$.
One may compare the Pareto indices α_1, α_2 over two production scenarios, the smaller value of α signifies a better production; for in such case the (right) tail of the corresponding distribution is thicker compared to that with higher value. The ratio α_1/α_2 may serve as an index of production performance of situation 2 with respect to situation 1.

Example 3. *Peak gust wind velocities (PGU) in Florida, USA*
The following 156 observations relates to Peak gust wind velocities (PGU) for coastal city Florida, USA in miles per hour (mph) over 12 months recordings for several years during 1930-96. When peak gust wind velocities are not available, 5-second winds velocity preceding PGU are given. Wind types may be combined to reflect the highest reported wind velocity, see http://www.ncdc.noaa.gov/oa/mpp/wind1996.pdf for details.

41, 49, 41, 43, 61, 38, 41, 68, 68, 44, 85, 47, 52, 58, 77, 49, 69, 67, 67, 68, 48, 56, 47, 43, 52, 58, 48, 52, 62, 97, 69, 69, 74, 74, 68, 49, 40, 39, 46, 39, 40, 46, 45, 44, 92, 45, 31, 35, 55, 62, 66, 67, 56, 58, 69, 61, 55, 47, 46, 45, 58, 52, 75, 63, 52, 51, 51, 56, 58, 67, 69, 48, 45, 61, 59, 55, 46, 58, 56, 115, 62, 47, 49, 46, 48, 51, 62, 53, 68, 62, 74, 62, 56, 40, 41, 43, 54, 60, 59, 63, 60, 69, 64, 78, 79, 49, 69, 53, 35, 35, 35, 35, 32, 32, 35, 35, 53, 35, 35, 34, 44, 51, 53, 48, 41, 76, 67, 64, 83, 58, 68, 36, 44, 46, 58, 49, 51, 61, 60, 48, 45, 53, 60, 37, 46, 40, 43, 38, 39, 53, 32, 41, 52, 45, 46, 48.

In Figure 7 with 156 PGU data we plot $\log x$ vs. $-\log(1 - F(x))$ and observe that Pareto model for wind gust may fit well beyond a large threshold value, much like the earlier data on sea tide and yam-yield. Figure 8 plots the same for $\log x > 4.2$ with 31 observations. The fit of Pareto model (2.5) with $a = e^{4.2}$ seems reasonable as $R^2 = 0.9819$, estimated value of $\alpha = 8.0530$.
In Figure 9 we see that Pareto model to peak wind gust fits better for $\log x > 4.3$ with 13 observations and a high value of $R^2 = 0.9860$, with estimated value of $\alpha = 8.0128$ when $a = e^{4.3}$. The values of α seem to stabilize around 8, indicating stability of the model towards higher values of wind gust.
Taking $\alpha = 8$, the median wind gust exceeding the value $a = e^{4.3} = 73.70$ is $73.70 \times e^{\frac{1}{8}\log 2} = 73.70 \times 1.090508 = 80.37$ mph.
In a similar manner the median wind gust exceeding 115, the largest observation recorded in above PGU data, is $115 \times 1.090508 = 125.41$ mph.
This provides an idea about the magnitude of the variable *beyond* the reported records.

Example 4. *Worldwide earthquake data.*
The following data relates to earthquake measurements during 30 September - 1 October 2011 on Richter scale recorded worldwide,
see http://earthquake.usgs.gov/earthquakes/catalogs/eqs7day-M1.txt for details.
The webpage is continuously updated, and the following segment of data was collected some-

times on 1 October 2011.

1.7,4.8,2.2,2.8,1.8,2.3,1.2,2.1,1.4,3.0,4.8,3.6,1.6,1.3,2.4,3.8,1.3,1.7,4.7,2.8,5.4,5.2,3.1,
1.3,1.6,1.7,2.9,1.2,2.0,1.2,2.9,1.8,1.8,2.2,1.7,1.4,1.8,1.2,1.5,1.8,1.6,1.7,2.5,2.0,1.6,2.1,
1.0,1.3,3.3,1.7,4.4,2.9,1.6,2.5,2.8,2.6,1.3,1.9,2.0,1.6,1.1,2.5,1.3,1.1,1.2,1.2,1.1,1.4,2.4,
4.7,1.4,2.0,1.6,2.2,1.9,1.6,4.8,1.2,1.2,1.6,1.4,1.7,2.5,1.0,1.4,1.3,1.6,1.1,1.6,5.1,2.0,2.4,
4.5,2.5,2.8,1.7,1.1,3.2,1.4,1.5,2.4,1.2,4.8,1.4,2.0,4.8,1.3,1.8,4.6,1.0,1.8,1.3,2.5,1.1,1.9,
4.2,3.2,1.1,1.7,1.3,1.2,1.2,1.4,1.8,1.1,1.2,2.3,4.4,1.9,2.5,1.0,1.3,1.1,5.0,2.2,5.0,1.1,1.7,
1.3,2.9,1.2,1.0,1.9,1.8,2.1,1.8,1.6,1.6,2.7,1.1,1.5,1.8,1.1,1.4,4.6,1.6,2.0,2.5,1.4,1.8,1.1,
1.4,2.2.

In Figure 10 with 163 earthquake data we plot $\log x$ vs. $-\log(1-F(x))$ and observe that Pareto model for earth-tremor may fit well as a mixture of two Pareto distributions. Figure 11 plots the same quantities for $\log x > 1.5$ with 15 observations. The fit of Pareto model (2.5) with $a = e^{1.5}$ seems reasonable as $R^2 = 0.9668$, estimated value of $\alpha = 18.0267$.
Pareto model to earth-tremor fit for $\log x < 1.5$ with 148 observations, is shown in Figure 12, $R^2 = 0.9378$, with estimated value of $\alpha = 3.0470$ when $a = 1.0$. The Pareto fit in lower range of the earthquake data is also satisfactory.
There seems to be a change in the parameter of distribution for tremor exceeding 4.2 in Richter scale. The physical interpretation of this is worth investigating.

References

Cebrian A, Denuit M, Lambert P (2003). "Generalized Pareto Fit to the Society of Actuaries Large Claims Database." *North American Actuarial Journal*, **7**, 18–36.

Dasgupta R (2011). "Discrete distributions with application to lifestyle data." *International Conference on Productivity, Quality, Reliability, Optimization and Modeling Proceedings, Allied Publishers, New Delhi*, **1**, 502–520.

de Oliveira MMF, Ebecken NFF, de Oliveira JLF, Gilleland E (2011). "Generalized extreme wind speed distributions in South America over the Atlantic Ocean region." *Theory of Applied Climatology*, **104**, 377–385.

Iyer S, Manjunath D, Manibasakan R (2002). "Bivariate exponential distributions using linear structures." *Sankhyā A*, pp. 156–166.

Jenkinson A (1955). "The Frequency Distribution of the Annual Maximum (or Minimum) of Meteorological Elements." *Quarterly Journal of the Royal Meteorological Society*, **81**, 158–171.

Klass O, Biham O, Levy M, Malcai O, Soloman S (2006). "The Forbes 400 and the Pareto wealth distribution." *Economics Letters*, **90**, 290–295.

Krishnaji N (1970). "Characterization of the Pareto Distribution Through a Model of Under-reported Incomes." *Econometrica*, **38**, 251–255.

Morrow-Tlucak M, Emhart C, Sokol R, Martier S, Ager J (1989). "Underreporting of Alcohol Use in Pregnancy: Relationship to Alcohol Problem History." *Alcoholism: Clinical and Experimental Research*, **13**, 399–401. Doi: 10.1111/j.1530-0277.1989.tb00343.x.

Van Montfort M, Witter J (1985). "Testing Exponentiality Against Generalized Pareto Distribution." *Journal of Hydrology*, **78**, 305–315.

Figure 1. Pareto fit for Tide data

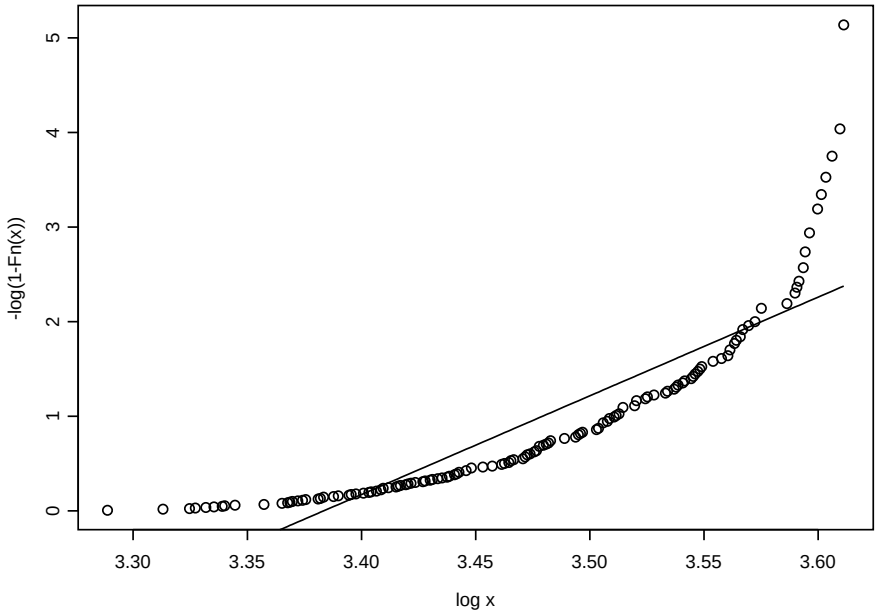


Figure 2. Pareto fit for Tide data: log X > 3.55

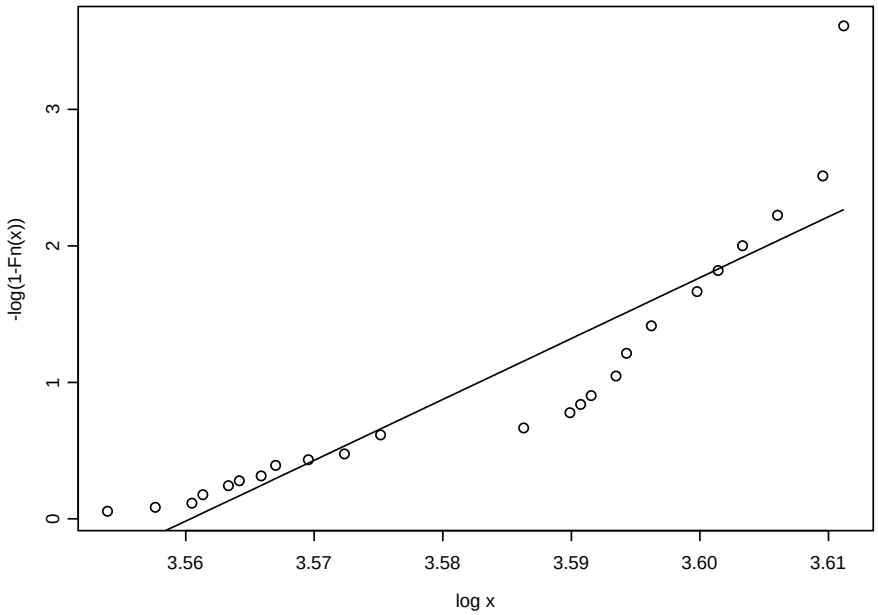


Figure 3. Pareto fit for Tide data: log X > 3.58

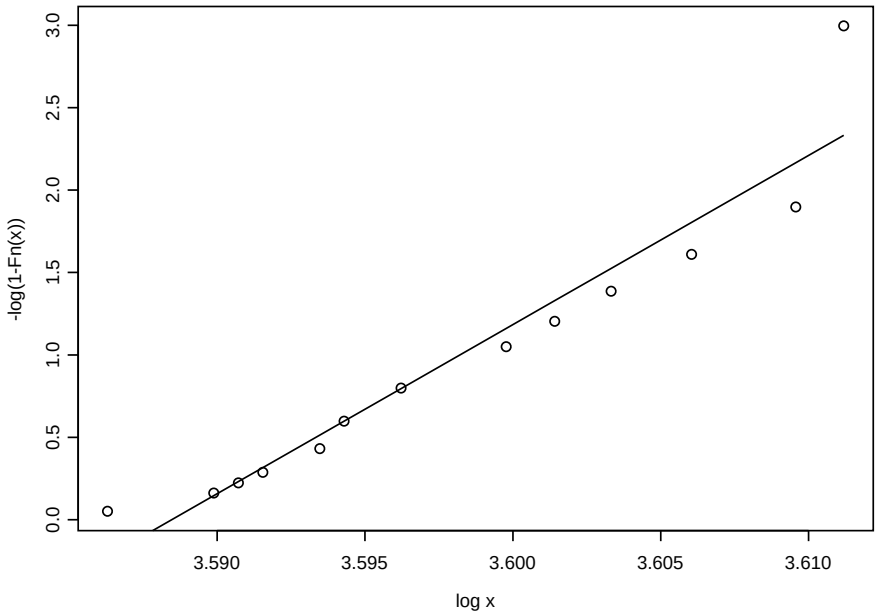


Figure 4. Pareto fit for Yam data

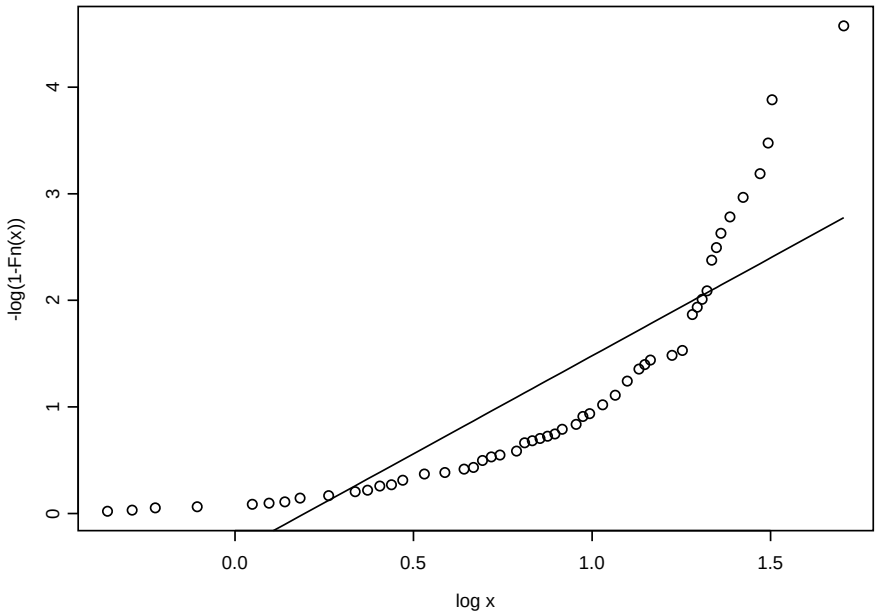


Figure 5. Pareto fit for Yam data: log x > 1

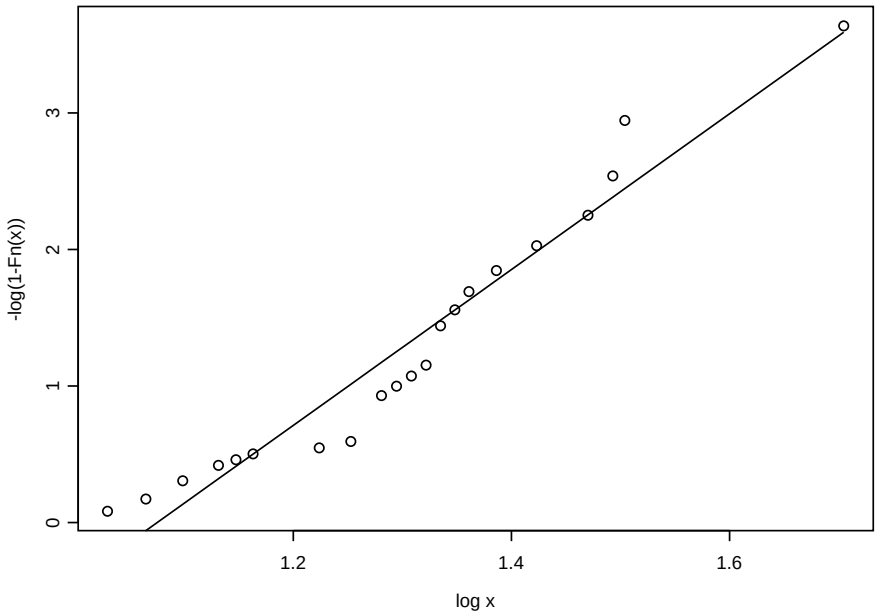


Figure 6. Pareto fit for Yam data: log x > 1.2

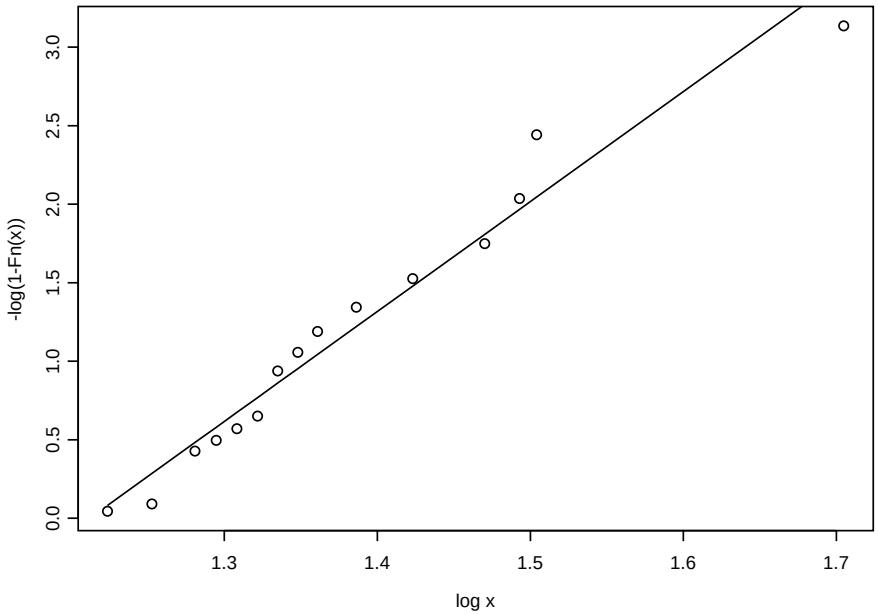


Figure 7. Pareto fit for peak wind gust data

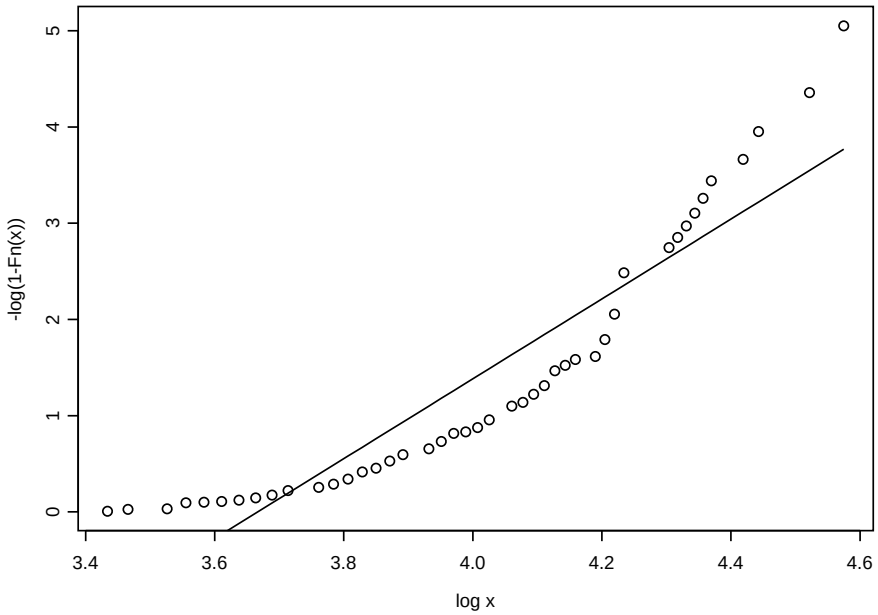


Figure 8. Pareto fit for peak wind gust data: log x > 4.2

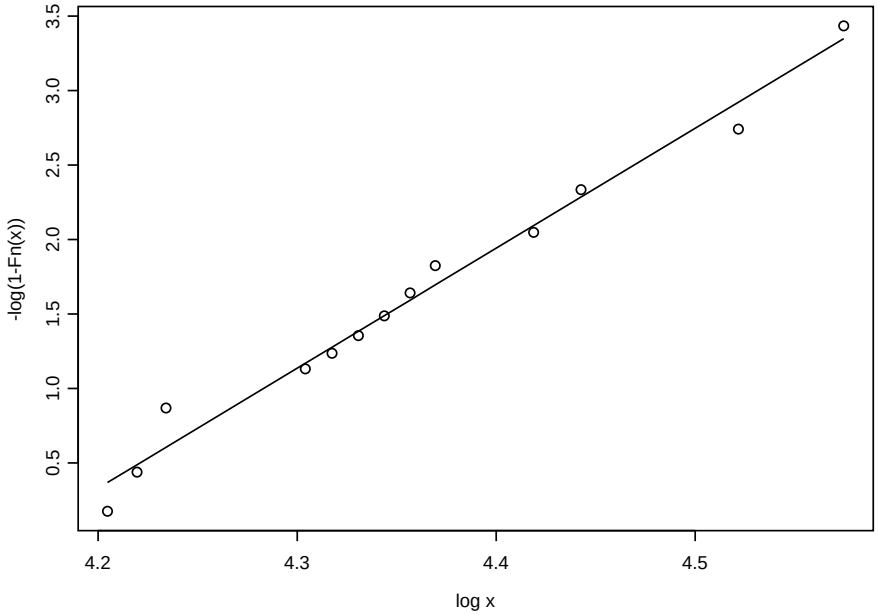


Figure 9. Pareto fit for peak wind gust data: log x > 4.3

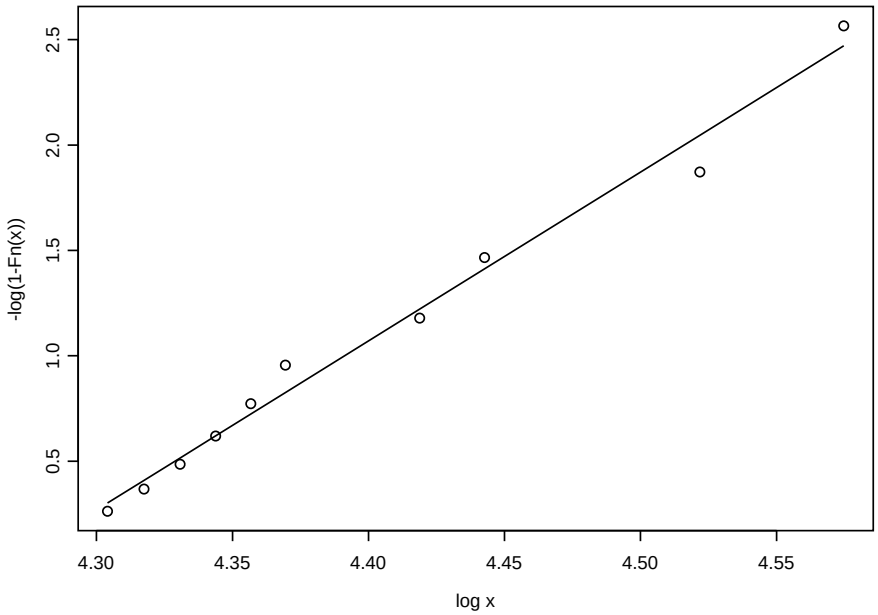


Figure 10. Pareto fit for earthquake data

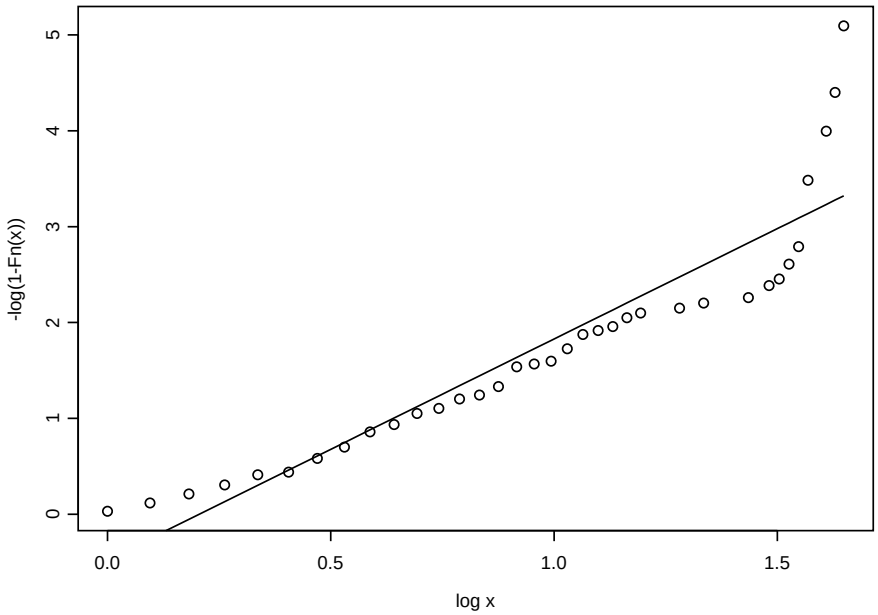


Figure 11. Pareto fit for earthquake data: log x > 1.5

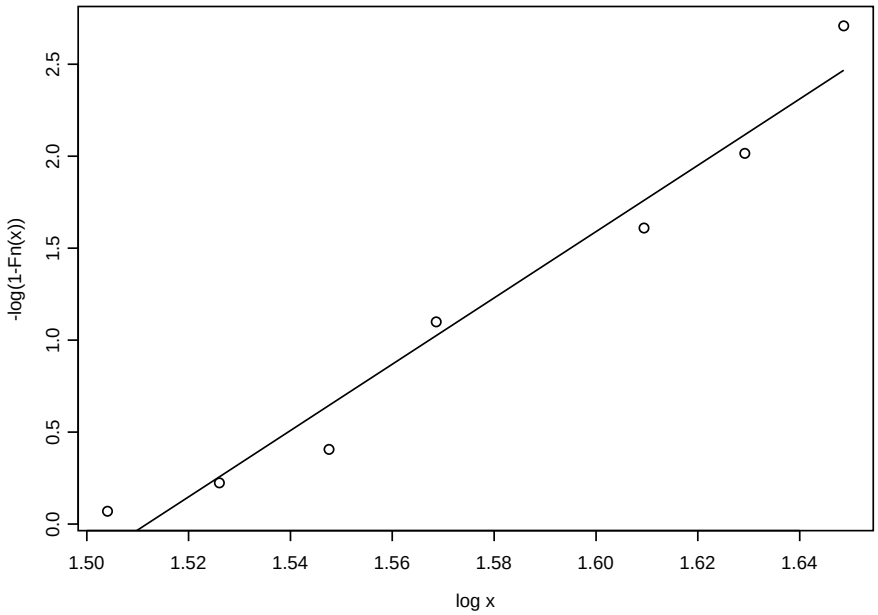
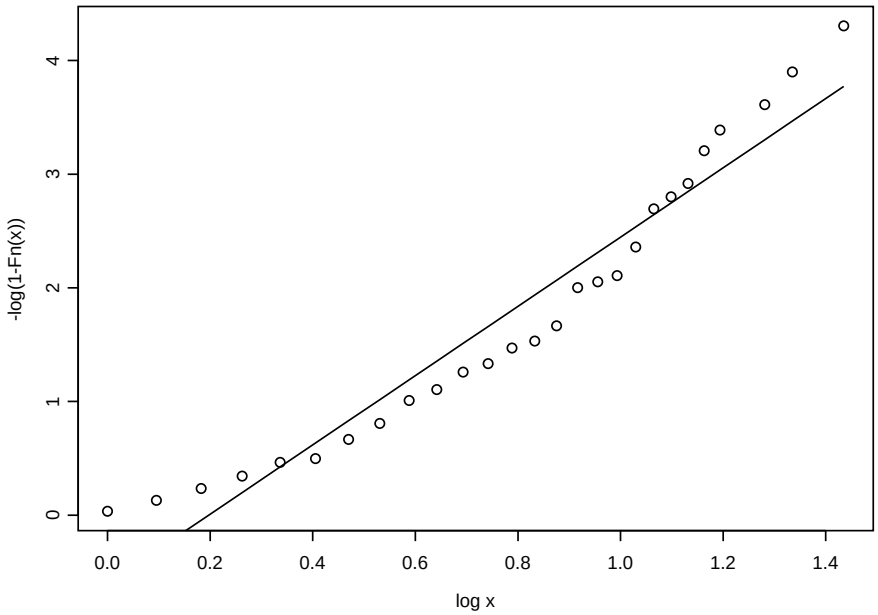


Figure 12. Pareto fit for earthquake data: log x < 1.5



Affiliation:

Ratan Dasgupta
Indian Statistical Institute
Theoretical Statistics and Mathematics Unit
Calcutta 700 108, INDIA
E-mail: ratandasgupta@gmail.com
rdgupta@isical.ac.in

http://www.jenvstat.org
Submitted: 2012-02-01
Accepted: 2012-10-10