Skewed and Extreme:
Useful distributions for economic heterogeneity [*]

Keith Head[†]

November 3, 2013

Abstract

To understand heterogeneous economic agents, one frequently models them with parametric statistical distributions. A variety of distributions have proven useful. This paper assembles information on these distributions from a number of sources. It also highlights the relationships between the distributions. Where possible, I also discuss the underlying processes that give rise to the different distributions. References to recent applications and debates are provided.

1 Introduction

Economic variables exhibiting heterogeneity—such as size, incomes, factor prices, productivity, and trade costs—share some common features. They are usually positive, continuous, and without upper limits. While heterogeneity has a long history in economics, most economists were trained to use models involving *representative* individuals and firms. There now appears to be a resurgence of interest in heterogeneity. Examples of variables where hetereogeneity is thought to be important include individual incomes (Pareto, 1896), consumer preferences (Anderson, de Palma and Thisse), urban populations (Zipf, Krugman, Gabaix, Eeckhout), firm sizes (Axtell, Cabral and Mata), and productivities (Kortum, Melitz). Attention to skewed distributions with long tails has even spilled over into books intende for popular audiences such *The Long Tail* and *Black Swan*.

Papers on heterogeneity often refer to specific distributions such as the Pareto, Frechet, and Log-normal. These references are, by necessity, usually quite brief. Authors wishing to know more about the relevant distributions and how they relate to each other must consult a variety of large and sometimes difficult-to-find volumes (or search through the fragmentary information on the Web).

This paper collects a variety of results about the main distributions used in the literature. Most references organize their coverage around the specific distributions. I depart from this practice and organize around particular topics: density functions, relationships between them,

[*] WARNING: This is a *preliminary version.* The reader should *verify* results before using them. I wish to thank Giordano Mion for helpful discussions.

[†] Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, BC, V6T1Z2, Canada. Tel: (604)822-8492, Fax: (604)822-8477, Email: keith.head@ubc.ca

Table 1: Symmetric Distributions

Uniform (Rectangular)
$f(x) = \frac{1}{\beta - \alpha}$ Range: $\alpha \leq x \leq \beta$
$F(x) = \frac{x-\alpha}{\beta-\alpha}$ Standardized form: $\alpha = 0$, $\beta = 1$

Normal (Gaussian)
$f(x) = \phi[(x-\mu)/\sigma]$ Range: $-\infty < x < \infty$
$F(x) = \Phi[(x-\mu)/\sigma]$ Standardized form: $\mu = 0$, $\sigma = 1$, where $\phi[]$ and $\Phi[]$ are PDF and CDF of the standard Normal(0,1) distribution.

Logistic
$F(x) = 1/(1 + \exp[-(x-\mu)/\sigma])$

moments. The motivation for this is that the standard treatment tends to make it difficult to see commonalities and differences. My imagined audience has not yet decided which distribution fits the data or which distribution to assume in a model but wants to scan through some plausible alternatives. A complete coverage requires a massive volume or *two*.[1] I have been very selective in what I include and tend to emphasize things that seem useful, interesting, or confusing. I am no expert on these matters and it is nearly certain that I remain confused on many issues. I hope for the reader's forbearance and request that he or she contact me with suggested corrections, clarifications, and additions.

2 The density functions

In this section I show the pdf, $f(x)$, and cdf, $F(x)$, of each density. Each reference I have employed uses different notation. Thus, for most distributions there is no standard set of parameters. To make it easier to see the relationships between distributions, I have tried to use parameters that correspond to each other. A parameter capturing a central tendency of the data is denoted μ. Lower and upper bounds are denoted α and β. The parameter measuring decay in the right tail is λ. It is inversely related to the spread of the distribution.

3 Origins and relations

Each distribution relates to others, sometimes because one is a limiting value of the order statistics of a large sample and sometimes through monotonic transformations. Most distributions seem to be linked to the Exponential in one way or the other.

[1] See Johnson, Kotz, and Balakrishnan (1994). Even this reference skimps on certain distributions, in particular the Frechet. On the other hand, Bury (1999) only mentions the Pareto briefly in connection with his thorough coverage of the Frechet.

Table 2: The Skewed Distributions

Exponential
$f(x) = \lambda \exp[-\lambda(x-\alpha)]$ Range: $0 \le \alpha \le x$
$F(x) = 1 - \exp[-\lambda(x-\alpha)]$ Standardized form: $\alpha = 0$, $\lambda = 1$

Pareto $f(x) = \lambda\alpha^{\lambda}x^{-\lambda-1}$ Range: $0 < \alpha \le x$
$F(x) = 1 - (x/\alpha)^{-\lambda}$ Standardized form: $\alpha = 1$, *Zipf* form: $\lambda = 1$.

Power
$f(x) = \lambda\beta^{-\lambda}x^{\lambda-1}$ Range: $0 < x < \beta$
$F(x) = (x/\beta)^{\lambda}$ Uniform(0,1): $\lambda = 1$, $\beta = 1$.

Gumbel (Type 1 Extreme Value, log-Weibull, double-exponential, Gompertz, Fisher-Tippet)
$f(x) = (1/\sigma)\exp\{-(x-\mu)/\sigma - \exp[-(x-\mu)/\sigma]\}$ Range: $-\infty < x < \infty$
$F(x) = \exp\{-\exp[-(x-\mu)/\sigma]\}$ Standardized form: $\sigma = 1$, $\mu = 0$.

Frechet (Type 2 Extreme Value, log-Gompertz, inverse-Weibull)
$f(x) = \sigma^{\lambda}\lambda x^{-\lambda-1}\exp\{-(x/\sigma)^{-\lambda}\}$ Range: $0 \le x$
$F(x) = \exp\{-(x/\sigma)^{-\lambda}\}$ Standardized form: $\sigma = 1$.

Weibull (Type 3 Extreme Value: minimum)
$f(x) = \sigma^{-\lambda}\lambda x^{\lambda-1}\exp\{-(x/\sigma)^{\lambda}\}$ Range: $0 \le x$, $0 < \sigma, \lambda$
$F(x) = 1 - \exp\{-(x/\sigma)^{\lambda}\}$ Standardized form: $\sigma = 1$. *Exponential* form: $\lambda = 1$
Rayleigh form: $\lambda = 2$, $\sigma_W \to \sigma_R\sqrt{2}$.

Rayleigh
$f(x) = [x/\sigma^2]\exp[-.5(x/\sigma)^2]$ Range: $0 \le x$, $0 < \sigma$
$F(x) = 1 - \exp[-.5(x/\sigma)^2]$ Standardized form: $\sigma = 1$

Log-Normal
$f(x) = \phi[(\ln x - \mu)/\sigma]/(x\sigma)$ Range: $0 < x$
$F(x) = \Phi[(\ln x - \mu)/\sigma]$ Standardized form: $\mu = 0$, $\sigma = 1$

Fisk (Log-Logistic)^2
$f(x) = (\lambda/\sigma)(x/\sigma)^{\lambda-1}[1 + (x/\sigma)^{\lambda}]^{-2}$ Range: $x > 0$
$F(x) = 1/(1 + [x/\sigma]^{-\lambda})$

3.1 Sums and products

Log-normality and the central limit theorem. An analogous derivation of Pareto (which may not have finite mean and variance therefore might violate the assumptions of the Central limit theorem)?

3.2 Extreme values

The Gumbel, Frechet and Weibull are called extreme value distributions (EVD). They are the only three forms that such distributions can take. Each has a "basin of attraction," i.e. a set of original distributions which, if one take a large number of draws from the resulting maximum, gives rise to the particular EVD. The name Weibull is conventionally used to denote the distribution of the *minimum*. To distinguish the the case of of the maximum, I propose that distribution be called Weibull-max.

Gumbels come from distributions that are not bounded above but do have a full set of finite moments. The *maximum* from a sample of Normal, Log-Normal, Exponential, Gamma, logistic, and Weibull distributions will be distributed Gumbel. The maximum of a finite set of n Gumbels is Gumbel with the same shape parameter σ but with μ replaced by $\mu + \sigma \ln(n)$.

Frechets arise when the maximum is unbounded and some of the moments are not finite. Two examples are the Cauchy and Pareto distribution. The maximum of Frechets would also be Frechet. The maximum of a finite set of n Frechets is Frechet with the same λ but with θ replaced by $\sigma n^{1/\lambda}$.

Weibulls come from distributions that have a bounded lower tail. The lowest draws from a sample of Log-Normals, Gammas, Betas, Frechets, and other Weibulls will also be Weibull. The minimum of a finite set of n Weibull is Weibull with the σ replaced by $\sigma n^{-1/\lambda}$.

The maximum of n Paretos is given by the distribution function $F(x) = [1 - (x/\alpha)^{-\lambda}]^n$. The minimum of n Paretos is also Pareto but with λ replaced with $n\lambda$. Thus the Pareto, like the Weibull, is "closed with respect to the minimum."[3]

3.3 Logarithmic, inverse, and power transformations

General rule for monotonic, continuous, invertible, differentiable functions[4]: If X has CDF $F_X(x)$ and $Y = g(X)$, then $F_Y(y) = F_X(g^{-1}(x))$ if $g'(X) > 0$ and $F_Y(y) = 1 - F_X(g^{-1}(x))$ for $g'(X) < 0$.

Two common transformations are applying the natural logarithm and raising the original variable to some power. As Pareto is an important distribution, I use it for two examples. Suppose X is Pareto with minimum $\underline{x}$ and shape λ. Then $Y = \ln(X)$ is exponentially distributed with location parameter $\ln \underline{x}$ and decay parameter λ.

$$\mathbb{P}(Y < y) = \mathbb{P}(\ln X < y) = \mathbb{P}(X < \exp(y)) = 1 - \left(\frac{\exp(y)}{\underline{x}}\right)^{-\lambda} = 1 - \exp[-\lambda(y - \ln \underline{x})].$$

Suppose instead $Y = aX^b$ in that case

$$\mathbb{P}(Y < y) = \mathbb{P}(aX^b < y) = \mathbb{P}(X < (y/a)^{1/b}) = 1 - \left(\frac{(y/a)^{1/b}}{\underline{x}}\right)^{-\lambda} = 1 - \left(\frac{y}{a\underline{x}^b}\right)^{-\lambda/b}.$$

[3] Kleiber and Kotz (2003, p. 72, 176)

[4] McCord and Moroney (1964).

Thus Y is also Pareto with new minimum parameter $a\underline{x}^b$ and shape λ/b.

Table 3: Origins, transformations, results

$X \sim$	$Y = g(X)$	$Y \sim$	Parameters*
Uniform(0,1)	$Y = \ln(1/X)$	Exponential	$\lambda = 1$
Log-Normal	$Y = \ln(X)$	Normal(μ,σ^2)	
Log-Normal	$Y = aX^b$	Log-Normal	$\ln a + b\mu$, $b^2\sigma^2$
Fisk	$Y = \ln(X)$	Logistic	$\mu_L = \ln \sigma_F$, $\sigma_L = 1/\lambda$
Pareto	$Y = \ln(X)$	Exponential	$\alpha_E = \ln(\alpha_P)$
Pareto	$Y = 1/X$	Power	$1/\alpha$ in Pareto $= \beta$ in Power
Pareto	$Y = aX^b$	Pareto	$\alpha \to a\alpha^b$, $\lambda \to \lambda/b$
Frechet	$Y = \ln(X)$	Gumbel	$\mu_G = \ln \sigma_F$, $\sigma_G = 1/\lambda_F$
Frechet	$Y = (X/\sigma)^{-\lambda}$	Exponential	$\lambda_E = 1$
Frechet	$Y = aX^b$	Frechet	$\sigma \to a\sigma^b$, $\lambda \to \lambda/b$
Weibull	$Y = 1/X$	Frechet	
Weibull	$Y = \ln(1/X)$	Gumbel	$\mu = \ln \sigma_F$
Weibull	$Y = aX^b$	Weibull	$\sigma \to a\sigma^b$, $\lambda \to \lambda/b$
Weibull	$Y = X^\lambda$	Exponential	$\lambda_E = \sigma_W^{-\lambda_W}$
Rayleigh	$Y = aX^b$	Weibull	$\sigma_W = a(\sigma_R\sqrt{2})^b$, $\lambda = 2/b$

*: → means "is replaced with"

4 Moments, quantiles, etc.

The $F^{-1}()$ formula (inverse cumulative distribution) has several important uses. First, one can obtain medians by plugging in $q = .5$. Indeed it is called the quantile function because any other percentile of the data can be obtained by selecting q. Second, the inverse function is the basis for pseudo-random number generation. Substituting a Uniform(0,1) random number (available in almost all software) instead of q in the quantile formula gives a random variable distributed according the corresponding distribution.

If X is distributed log-normal, then its mean will be larger than its median. Indeed the mean to median ratio is given by $e^{\sigma^2/2}$. As $\sigma^2 \to 0$, the skewness of Log-Normal disappears and it approaches the Normal. There is a simple relationship between the mean to median relation and the coefficient of variation (cv):

$$\text{Log-normal} \qquad \frac{\text{mean}}{\text{median}} = \sqrt{1 + \text{cv}^2}.$$

The mean to median ratio for a Pareto variable is given by

$$\frac{\text{mean}}{\text{median}} = \frac{\lambda}{(\lambda - 1)2^{1/\lambda}}$$

The coefficient of variation is

$$cv = [\lambda(\lambda - 2)]^{-1/2}$$

Table 4: Describing the distributions

Distribution	Mean	Std. Dev.	Mode	$F^{-1}()$
Uniform	$(\alpha+\beta)/2$	$(\beta-\alpha)/\sqrt{12}$	NA	$\alpha+(\beta-\alpha)q$
Exponential	$\alpha+1/\lambda$	$1/\lambda$	α	$\alpha-(1/\lambda)\ln(1-q)$
Pareto	$\alpha\lambda/(\lambda-1)$	$\alpha G(\lambda)^{\star}$	α	$\alpha(1-q)^{-1/\lambda}$
Power	$\beta\lambda/(\lambda-1)$	$\beta H(\lambda)^{\bullet}$	β or 0	$\beta q^{1/\lambda}$
Gumbel	$\mu+.577\sigma$	$\pi\sigma/\sqrt{6}$	μ	$\mu-\sigma\ln(-\ln q)$
Frechet	$\sigma\Gamma(1-1/\lambda)$	$\sigma K(\lambda)^{\dagger}$	$\sigma[\lambda/(1+\lambda)]^{1/\lambda}$	$\sigma[\ln(1/q)]^{-1/\lambda}$
Weibull	$\sigma\Gamma(1+1/\lambda)$	$\sigma L(\lambda)^{\ddagger}$	$\sigma[(\lambda-1)/\lambda]^{1/\lambda}$	$\sigma[-\ln(1-q)]^{1/\lambda}$
Rayleigh	$\sigma\sqrt{\pi/2}$	$\sigma\sqrt{2-\pi/2}$	σ	$\sigma\sqrt{-2\ln(1-q)}$
Log-Normal	$e^{\mu+\sigma^2/2}$	$e^{\mu+\sigma^2/2}\sqrt{\exp(\sigma^2)-1}$	$\exp(\mu-\sigma^2)$	$\exp\{\mu+\sigma\Phi^{-1}[q]\}$
Fisk	$(\sigma/\lambda)\pi\csc(\pi/\lambda)$	$\sigma M(\lambda)^{*}$	$\sigma[(\lambda-1)/(\lambda+1)]^{1/\lambda}$	$\sigma[q/(1-q)]^{1/\lambda}$

$\star$: $G(\lambda)=\sqrt{\lambda/(\lambda-2)}/(\lambda-1)$
$\bullet$: $H(\lambda)=\sqrt{\lambda/(\lambda+2)}/(\lambda+1)$
$\dagger$: $K(\lambda)=\sqrt{\Gamma(1-2/\lambda)-\Gamma^2(1-1/\lambda)}$
$\ddagger$: $L(\lambda)=\sqrt{\Gamma(1+2/\lambda)-\Gamma^2(1+1/\lambda)}$
$*$: $M(\lambda)=\sqrt{(\pi/\lambda)[2\csc(2\pi/\lambda)-(\pi/\lambda)(csc(\pi/\lambda))^2]}$

Combining we obtain

$$\text{Pareto}\qquad \frac{\text{mean}}{\text{median}}=\frac{1+\sqrt{1+\text{cv}^{-2}}}{\sqrt{1+\text{cv}^{-2}}2^{1/(1+\sqrt{1+\text{cv}^{-2}})}}$$

As with Log-normal, the mean to median ratio goes to 1 as as cv goes to zero. But as cv becomes large the two distributions behave very differently. For log-normal the mean to median ratio converges on the cv (suppose the cv = 10, the square root of 101 is 10.05). On the other hand the Pareto distribution rises far less quickly and reaches a maximum of $\sqrt{2}=1.414$. Thus any mean to median ratio greater than 1.4 is inconsistent with a Pareto distribution with a finite coefficient of variation.

Taking logs of a log-normal variable, $Y=\ln X$ will have a mean equal to the median. Suppose however X is Pareto. Then the expected value will not even exist unless $\lambda>1$.

If we take logs of a Pareto variable, the expected value always exists as logged Pareto variables are distributed exponential (see Table 3). The expected value of the logged Pareto variable, $\ln(\alpha)+1/\lambda$—which equals the log of the geometric mean of the Pareto variable—should still be larger than its median, $\ln(\alpha)+(1/\lambda)\ln 2$. Indeed the difference should be $.31/\lambda$. This approach could provide a simple way to do an initial diagnosis of a variable that you think could be normal, log-normal, or Pareto. Another potentially useful fact is that the harmonic mean of the Pareto is given by $\alpha(1+1/\lambda)$. When λ is large the geometric and harmomic means should be about the same. But for small λ the geometric mean can be much larger. For the Zipf value of $\lambda=1$, the geometric mean is 36% larger.[5]

[5] The Stata command *means* provides arithmetic, geometric and harmonic means.

5 Truncated distributions

Truncated distributions are quite important in economic data. One cause of truncation is the data collecting and reporting process. One case is the size distribution of cities where populations are only reported for the top cities or cities above a certain threshold population. More on that below. A more interesting cause of truncation is where it is caused by economic forces. Melitz and Syverson consider the case where competition creates a threshold level of productivity below which firms cannot remain in the industry because the price is lower than costs.

Unfortunately, the literature on distributions gives scant coverage to truncation. Johnson et al (2000, p. 241) provide a general formula for the moments of the trunctated log-normal distribution.

$$\mathrm{E}[x \mid x > x_0] = \exp(\mu + 0.5\sigma^2)\frac{1 - \Phi(z_0 - \sigma)}{1 - \Phi(z_0)}, \tag{1}$$

where $z_0 \equiv (\ln x_0 - \mu)/\sigma$, x_0 is the truncation point, and μ and σ are parameters of the log-normal distribution.

Jawitz (2004) gives formulas that allow us to obtain the more general case of two-sided truncation. For $x \sim \text{log-}\mathcal{N}(\mu, \nu)$ truncated between lower limit ℓ and upper limit u the Jawitz formula can be expressed as

$$m_r = \exp(r\mu + r^2\nu^2/2)\left[\Phi\left(\frac{\ln u - \mu - r\nu^2}{\nu}\right) - \Phi\left(\frac{\ln \ell - \mu - r\nu^2}{\nu}\right)\right]$$

$$m_0 = \Phi\left(\frac{\ln u - \mu}{\nu}\right) - \Phi\left(\frac{\ln \ell - \mu}{\nu}\right)$$

The conditional expected value of x^r for $\ell < x < u$ is m_r/m_0.

The Pareto distribution is an interesting case because truncation does not change its shape but only the scale of measurement:

$$\mathrm{E}[x \mid x > x_0] = \frac{x_0\lambda}{\lambda - 1}. \tag{2}$$

In contrast, truncating a log-normal distribution can make it look very much like a Pareto. Mitzenmacher (2003) shows algebraically why this is the case and Eeckhout (2004) illustrates the point using the size distribution of populated places. I have also solve for the truncated expected value for the exponential and uniform distributions and I found the normal distribution. For the exponential with unconditional mean $1/\lambda$, the truncated mean is

$$\mathrm{E}[x \mid x > x_0] = x_0 + 1/\lambda. \tag{3}$$

For a uniform between α and β,

$$\mathrm{E}[x \mid x > x_0] = (x_0 + \beta)/2. \tag{4}$$

For the normal,

$$\mathrm{E}[x \mid x > x_0] = \mu + \frac{\sigma\phi(z_0)}{1 - \Phi(z_0)}, \tag{5}$$

where $z_0 \equiv (x_0 - \mu)/\sigma$.

Jawitz (2004) provides a general method and a table covering examples for other distributions including the gamma and log-gamma (LP3) distributions.

6 Estimation and Detection

Typically you are confronted with data X_i and your goal is to detect which distribution fits it best and to estimate the corresponding free parameters.

6.1 Maximum-likelihood methods

The geometric mean is used to determine the maximum likelihood estimate (MLE) of λ in a Pareto distribution. The MLE of λ is $1/(\ln m^g - \ln \hat{\alpha})$, where m^g denotes the geometric mean given by $m^g = \exp\left[(1/n)\sum_{i=1}^{n} \ln X_i\right]$ and $\hat{\alpha}$ is either known in advance or estimated as the minimum of the observed X_i. Kleiber and Klotz (2003, pp. 86–89) point out that the MLE is heavily influenced by extreme observations, such as might arise from contaminated data. They describe the method of Brazauskas and Serfling (2001) that aims at "favorable tradeoff between efficiency and robustness." Since the log of a Pareto is exponential and the standard deviation of the exponential is $1/\lambda$, another possible estimate would be the inverse of the standard deviation of $\ln X$.

The geometric mean, m^g would also be used to estimate the parameters of log-normally distributed data. The MLE of μ is $\ln m^g$, or the mean of the logged data. The MLE of σ is just the standard deviation of the logged data.

Estimating parameters of Frechet or Weibull distributed data is not as straightforward, with the MLE requiring solution of two non-linear simultaneous equations.[6]

6.2 Method of moments

It is often desirable to make distributions as directly comparable to each other as possible. For two-parameter distributions, it is usually possible to choose parameters for each distribution such that they replicate the mean and variance of the data under consideration. Thus we use method-of-moments estimators based on an assumed mean and variance. Bury (1999) provides the estimators for Lognormal and Gamma. I derived analogous formulas for the Pareto and Uniform distributions. They are shown in Table 6 for the reader's reference, with m and v denoting mean and variance of x, the underlying raw performance measure (untruncated, untransformed).

Table 5: Method of moments estimators for mean = m and variance = v

Pareto(min=$\underline{x}$, shape=κ)	$\kappa = 1 + \sqrt{1 + m^2/v}$	$\underline{x} = m(\kappa - 1)/\kappa$
Log-normal(meanlog= μ, sdlog= σ)	$\sigma = \sqrt{\ln(v + m^2) - 2\ln m}$	$\mu = \ln m - \sigma^2/2$
Gamma(shape= γ_1, scale=γ_2)	$\gamma_2 = v/m$	$\gamma_1 = m/\gamma_2$
Uniform(min= $\underline{x}$, max=$\bar{x}$)	$\underline{x} = m - \sqrt{3v}$	$\bar{x} = 2m - \underline{x}$

For the Log-normal and Pareto things simplify considerably when we re-express the formulas in terms of the coefficient of variation. The formula for the log normal is $\sigma = \sqrt{\ln(1 + \mathrm{cv}^2)}$ and for the Pareto it is $\kappa = 1 + \sqrt{1 + 1/\mathrm{cv}^2}$. These formulas highlight the fact that σ is positively related to dispersion whereas κ is inversely related to dispersion. As cv→ ∞, we

[6] See Evans et al. (2000)

estimate $\kappa = 2$. I believe this is because $\kappa = 2$ is the smallest shape parameter for which the second moment is finite. Another interesting fact is that neither parameter relates to the scale of the variable in question (given cv, m does not matter).

If the DGP is log-normal it seems likely that the method-of-moments estimator of $\underline{x}$ will differ substantially from the maximum likelihood estimator. This is because the latter is equal to the minimum observed value and that goes to zero as the sample size increases under a log-normal DGP. In contrast the MoM estimator of $\underline{x}$ has to be large enough to be consistent with the observed mean of the data. The share of data below the MoM estimator of $\underline{x}$ ought to be a good indicator of the inadequacy of the Pareto distribution.

Rather than use the raw moments, a related approach that works for at least three interesting distributions is to match parameters to the moments of the logged data. The first and obvious case is the log-Normal distribution. There we know that μ is the mean $Y = \ln(X)$ and σ is the standard deviation of the Y. These estimates have the added advantage of being the MLE of μ and σ.

Extending the method to Pareto, recall that if X is Pareto, then $Y = \ln(X)$ is exponential. The Pareto shape parameter is estimated as $\hat{\lambda} = 1/\mathrm{SD}(Y)$. The Pareto scale (minimum value) parameter is estimated as $\underline{x} = \mathrm{MEAN}(Y) - \mathrm{SD}(Y)$.

Table 6: Method of moments estimators for sample mean $= M$ and variance $= V$

Pareto(min=$\underline{x}$, shape=θ)	$\theta = 1 + \sqrt{1 + M^2/V}$	$\underline{x} = M(\theta - 1)/\theta$
Log-normal(meanlog= μ, sdlog= ν)	$\nu = \sqrt{\ln(V + M^2) - 2\ln M}$	$\mu = \ln M - \nu_2^2/2$

Table 7: Method of moments estimators for sample mean $= m$ and standard deviation of $= s$ of the *log* of x

Pareto(min=$\underline{x}$, shape=θ)	$\theta = 1/s$	$\underline{x} = \exp(m - s)$
Log-normal(μ, ν^2)	$\nu = s$	$\mu = m$
Frechet(σ, λ)	$\lambda = 1.28/s$	$\sigma = \exp(m - 0.45s)$

An interesting case is where one works with a logged variable (sales, city size, wages) that is truncated from below, sometimes because of data availability but other times because an economic *selection* mechanism at work (e.g. fixed costs, test scores). Let L denote the log of the raw truncation point. Let μ and ν denote the expectation and standard deviation of the logged variables in the absence of truncation. The expected value of the logged truncated variable is

$$\mathbb{E}[\ln x \mid \ln x > L] = \mu + \nu h,$$

where $h = \phi(t_\ell)/[1-\Phi(t_\ell)]$ and $t_\ell = (L-\mu)/\nu$. The standard deviation of the logged truncated variable is

$$\nu\sqrt{1 + t_l h - h^2}.$$

6.3 Graph and regression based methods

Going back to Vilfredo Pareto's original work, one method to detect a Pareto variable is to plot the log of the number cases of where X_i is larger than some number against the log of that number. If the distribution is Pareto then the rank-size figure should exhibit a linear relationship in logs. A frequently used estimate (at least by economists) of the Pareto parameter λ is to regress the logged rank data on $\ln X_i$. The coefficient on $\ln X_i$ is $-\hat{\lambda}$. A similar procedure, called a "probability plot", Chambers (1983), is used more broadly in exploratory data analysis.[7] In that approach the horizontal axis consists of the ordered statistical medians for a given distribution and the vertical axis is the actual corresponding data values. Thus probability plots reverse the Pareto plots. I suggest an approach that follows the same orientation as the Pareto graph but generalizes it to incorporate a wide variety of possible distributions.

The graphical method starts with the data in levels, X_i or logs, $\ln X_i$. The data should be sorted in ascending order so that $i = 1$ is the minimum and $i = n$ is the maximum. Let $\hat{F}_i = (i - 0.3)/(n + 0.4)$. The vertical axis is a transformation, $g()$ of $\hat{F}_i$. The slope and intercept—denoted a and b—in the relationship between $g(\hat{F}_i)$ and X_i (in levels or logs) correspond to the parameters of the model:

$$g(\hat{F}_i) = a + bh(x), \tag{6}$$

where $h()$ is either the natural logarithm or the identity function. One could estimate them by regressing the $g(\hat{F}_i)$ on $h(X_i)$. This almost equivalent to the rank-size regression since rank is given by $1 + (n - i)$.

Table 8: Generalized linear-in-parameters rank-size relationships

Ordinate (y-axis)	Abscissa (x-axis)					
$g(\hat{F}_i)$	X_i	a	b	$\ln X_i$	a	b
$\hat{F}_i$	Uniform	$-\alpha/(\beta - \alpha)$	$1/(\beta - \alpha))$	?		
$\ln \hat{F}_i$	?	n/a	n/a	Power	$-\lambda \ln \beta$	λ
$\ln(1 - \hat{F}_i)$	Exponential	α	$-\lambda$	Pareto	$\lambda \ln \alpha$	$-\lambda$
$\ln(-\ln \hat{F}_i)$	Gumbel	$\lambda\mu$	$-\lambda$	Frechet	$\lambda \ln \sigma$	$-\lambda$
$\ln(-\ln[1 - \hat{F}_i])$	?	n/a	n/a	Weibull	$-\lambda \ln \sigma$	λ
$\Phi^{-1}[\hat{F}_i]$	Normal	$-\mu/\sigma$	$1/\sigma$	Log-Normal	$-\mu/\sigma$	$1/\sigma$
$\ln[\hat{F}_i/(1 - \hat{F}_i)]$	Logistic	$-\mu/\sigma$	$1/\sigma$	Fisk	$-\lambda \ln \sigma$	λ

My conjecture: As $F \rightarrow 1$, the Fisk resembles Pareto. As $F \rightarrow 1$, the Fisk resembles the Power distribution.

The correlation coefficient measures the strength of a linear relationship between two variables. Hence the correlation of $g(\hat{F}_i)$ and $h(X_i)$ might be a good way to measure how well each distribution fits the data.

[7] http://www.itl.nist.gov/div898/handbook/eda/section3/probplot.htm

7 Applications

There are two typical uses of the information here. The first is a theory involving heterogeneity. Applied theories may want to choose a particular distribution that yields analytical solutions of their model. Although tractability may be an over-riding concern (hence the use of the uniform which is almost never observed in actual data), ideally the applied theorist selects a distribution that "fits" the application. The second type of users are empiricists who want to see which distribution fits the data. One reason to do this is the hope that one can map the distribution to the underlying generating process.

7.1 Engineering and physics analogies

Extreme value distributions are very important for engineers. Gumbels and Frechets are used for floods. Weibulls are used for models of breaking strength. The Weibull is sometimes interpreted with a chain metaphor. Suppose chains consist of a large number of links, each of varying strength. Then the strength of the chain depends on the strength of the *weakest link*. There could be economic implications for Leontieff productions functions comprising a large number of processes where the performance of the whole is constrained by the worst performer.

The distribution of energies is exponential. Boltzmann. Statistical mechanics. Maxwell speed distribution.

The Rayleigh distribution occurs in formulas that sum the squares of two orthogonal Normal(0,σ^2) random variables. Suppose we have locations given as $\{X, Y\}$ where the horizontal and vertical position are normally distributed with standard deviations, σ. Then the distance, d from $\{X, Y\}$ to the origin ($d = \sqrt{X^2 + Y^2}$) is distributed Rayleigh with parameter σ. Now suppose that there is an interaction (trade flow, externality, etc.) given by ad^b. The interaction should be distributed Weibull.[8] The maximum likelihood estimate of σ is $\sqrt{(1/2n)\sum_i d_i^2} = (1/\sqrt{2})$RMSD.[9] We can use this to calculate the average distance to the center of a city if the population has a bell-shaped distribution around the center.

7.2 Cities, Countries, Firms, Incomes

Many economic variables are thought to follow Pareto distributions. Pareto's original application was incomes. Zipf, Simon, Krugman, and Gabaix drew attention to the fact that the size distribution of the largest American metro areas can be fit well by the inverse rank-size rule, which is equivalent to a Pareto distribution with $\lambda = 1$. Eeckhout (2004) countered that in fact the size distribution of the full set of populated places in the US is distributed log-normal. The appearance of Pareto comes from a focus on the right tail of the distribution. As discussed in section 5, the right tail of a log-normal closely resembles a Pareto.

Rose (2005) examines distributions of country populations and concludes they, like city populations, follow Zipf's law. However, his rank-size figures suffer from the truncation problem since they are based on the 50 largest countries. When Rose graphs the PDF of country size (as a histogram), it is clear that it is not Pareto and hence not globally Zipf.

[8] Problem: the transformation rule in Table 3 only works when $b > 0$. With $b < 0$, I think the result might be Frechet instead.

[9] RMSD = Root mean squared distance. See `www.mathworks.com/access/helpdesk/help/toolbox/stats/prob_d43.html`.

A simple static model of lognormal city size: Let there be a single output y whose production function is given by $y = g(L)h(R)$ where L is labour and R are immobile natural resources or amenities. The $g()$ function may have upward sloping regions but the interior solution occurs where $g''() < 0$. In the region of the interior solution let $g(L) = L^{\alpha}$ where $\alpha < 1$. Then suppose worker migration sets marginal products equal to a common wage, and set units so that wage is one. Then $L = (\alpha h(R))^{1/(1-\alpha)}$. Finally suppose $h(R) = \prod_{i=1}^{N} x_i^{\theta_i}$. That is aggregate resources of a location are a Cobb-Douglas function of a large number of individual resources, denoted x_i. If the θ_i are constants and $\ln x_i$ are independently distributed with finite first, second, and third moments then the central limit implies a normal distribution for $\log L$ (as the number of amenities or resources, N becomes large), and hence log-normality for L.

7.3 Search models

Extreme value distributions are useful in theoretical models of the search process. The idea is that if you take draws and retain the best draw obtained so far your technology is distributed as the maximum (or minimum if draws are interpreted as unit-input requirements instead of productivity). For example, Kortum (1997) shows a model where research leads to draws from a Pareto distribution, causing the technological frontier to be distributed Frechet. He shows that this can explain several stylized facts about the relationship between research efforts, patents, and productivity.

Muth (1986) used a search model to provide microfoundations for the learning curve. As mentioned in section 3, the minimum of n draws from a Weibull is also Weibull with a parameter that is a power function of n. This implies (via the moments reported in section 4, that the expected value of the minimum of n draws from a Weibull is given by $(n^{-1/\lambda}/\theta)\Gamma[1 + 1/\lambda]$. Taking logs we see a linear relationship, with slope $-1/\lambda$ between log costs and the log number of draws. If the latter are proportional to cumulative production experience, then we have the linear-in-logs relationship between costs and cumulative output that has been observed in hundreds of empirical studies.

7.4 Discrete choice models

Anderson, de Palma, and Thisse (1992) and Eaton and Kortum (2002) draw on the properties of extreme value distributions to specify models where consumers select from a set of possible suppliers the one that yields them the maximum utility. Anderson et al follow McFadden's approach of putting the heterogeneity in a random term in the consumer's utility function.

Suppose $U_i = v_i + \epsilon_i$, where v is a function of observables and ϵ_i is seen only by the chooser, not the econometrician. Take the case of two choices. The Probability of choosing 1 is

$$\Pr(U_1 > U_2) = \Pr(U_1 > U_2) = \Pr(v_1 + \epsilon_1 > v_2 + \epsilon_2) = \Pr(\epsilon_2 - \epsilon_1 < v_1 - v_2) = F(v_1 - v_2),$$

where $F(\cdot)$ is the cumulative distribution for $\epsilon_2 - \epsilon_1$. The difference between to random variables does not generally offer a closed form for $F()$. Two cases that do have this useful property are ϵ distributed uniform or ϵ distributed Gumbel. In the uniform ϵ case, $\epsilon_2 - \epsilon_1$ is also uniform.[10] In the $\epsilon \sim$ Gumbel(μ,σ) case $\epsilon_2 - \epsilon_1$ is logistic:

$$F(v_1 - v_2) = 1/(1 + \exp[-(v_1 - v_2)/\sigma]).$$

[10] Needs to be verified. Also consider how this result could be reinterpreted as a Hotelling line model.

Replacing $-(v_1 - v_2)/\sigma$ with $\mathbf{X}\beta$ gives Logit regressions.[11]

The Gumbel ϵ assumption becomes most valuable for the case of $N > 2$ choices. In that case a result usually attributed to McFadden is that

$$\mathrm{Pr}_i = \exp(v_i/\sigma)/V,$$

where $V \equiv \sum_j \exp(v_j/\sigma)$ is sometimes called the inclusive value. Anderson, de Palma, and Thisse (1992) show that the Expected value of the maximum utility is

$$\mathrm{E}\max\{U_1, ...U_n\} = \mu + \sigma(.577 + \ln V),$$

where μ and σ are the parameters from the Gumbel distribution defined above.

An important case is $v_i = K - \eta \ln p_i$, where K is everything that does not vary across choices and p_i is the price of option i.[12] In that case, we have

$$\mathrm{Pr}_i = p_i^{-\eta/\sigma}/V,$$

where $V = \sum_j p_j^{-\eta/\sigma}$ is now sometimes considered "price" index. This probability function is observationally equivalent to a deterministic case with a representative consumer with demand for variety, maximizing a CES utility function with an elasticity of substitution between varieties given by $\rho = 1 + \eta/\mu$.

Eaton and Kortum put the heterogeneity into prices rather than unobservable shocks to utility. The randomness of prices follows from underlying variation in supplier productivity that itself could be linked back to a search process *a la* Kortum (1997). This approach could also be applied to the migration decisions of workers.

7.5 Firms with heterogeneous productivity

Melitz, etc.

Relative outputs and revenues, equation (6)

Equation 7. Productivity index "completely summarizes the information in the distribution of productivity levels $\mu(\psi)$ relevant for all aggregate variables"

Footnote 8: $\tilde{\psi}$ is a harmonic weighted mean.

Equation 12. Free entry (FE) and zero-cutoff profit (ZCP) conditions.

Chaney

References

Anderson, S., A. De Palma and J.-F. Thisse (1992), *Discrete Choice Theory of Product Differentiation* (MIT Press, Cambridge).

Axtell, Robert L., 2001, "Zipf Distribution of U.S. Firm Sizes," *Science* 293, 1818-1820.

Beirlant, Jan, Yuri Goegebear, John Segers, and Jozef Tengels, *Statistics of Extremes: Theory and Applications*, Wiley, 2004.

[11] Note that the presence of the heterogeneity parameter σ implies estimated coefficients, $\hat{\beta}$, from the Logit results are related to the underlying v parameters by an unknown scalar.

[12] Anderson, de Palma, and Thisse make assumptions (which?) that lead to $\eta = 1$.

Brazauskas, V. and Serfling, R., 2001, "Robust estimation of tail parameters for two-parameter pareto and exponential models via generalized quantile statistics," *Extremes*, 3, 231–249.

Bury, Karl, 1999, *Statistical Distributions in Engineering*, Cambridge University Press.

Cabral, Luis M. and José Mata, 200?, "On the Evolution of the Firm Size Distribution: Facts and Theory," *American Economic Review*

Chaney T., 2005, "Distorted gravity: Heterogeneous Firms, Market Structure and the Geography of International Trade," University of Chicago, Mimeo.

Eaton, Jonathan and Samuel Kortum, 2002, "Technology, Geography, and Trade," *Econometrica* 70(5), 1741–1779.

Eeckhout, 2004, *American Economic Review.*

Evans, Merran, Nicholas Hastings, and Brian Peacock, 2000, *Statistical Distributions, Third Edition*, Wiley.

Fisk, P.R. 1961, "The Graduation of Income Distributions," *Econometrica* 29 (2): 171-185, doi:10.2307/1909287

Gabaix, Xavier, 1999, "Zipf's Law for cities: An explanation," *Quarterly Journal of Economics.*

Galambos, János, 1978, *The asymptotic theory of extreme order statistics* New York: Wiley.

Helpman, Melitz, and Yeaple, 200?, *American Economic Review*

Jawitz, James W., 2004, "Moments of truncated continuous univariate distributions," *Advances in Water Resources* 27, 269–281.

Johnson, Norman L., Samuel Kotz, and N. Balakrishnan, 2000, *Continuous Univariate Distributions*, ew York: Wiley

Kleiber, Christian and Samuel Kotz, 2003, *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley, Hoboken, NJ.

Kortum, Samuel, 1997, "Research, Patenting, and Technological Change," *Econometrica* 65(6), 1389-1419.

Kotz, Samuel and Saralees Nadarajah, 2000, *Extreme Value Distributions* Imperial College Press: London.

MacGregor, D.H., 1936, "Pareto's Law," *Economic Journal* 46, 80–87.

McCord, James R. and Richard M. Moroney, 1964, *Introduction to Probability Theory*, New York: MacMillan.

McLaughlin, Michael P., 1999, *Regress+ Compendium of Common Probability Distributions.*

Melitz, Marc J., 2003, "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," *Econometrica* 71(6), 1695–1725.

Mitzenmacher, Michael, 2003, “A Brief History of Generative Models for Power Law and Lognormal Distributions” *Internet Mathematics* Volume 1(2), 226–251.

NIST/SEMATECH, *e-Handbook of Statistical Methods*, `http://www.itl.nist.gov/div898/handbook/eda/section3/eda366g.htm`

Pareto, Vilfredo, 1896, *Course d’Economie Politique*

Ramsden, J.J., and Gy Kiss-Haypal, 2000, “Company size distribution in different countries.” *Physica A* 277, 220-227.

Rose, Andrew, 2005, “Cities and Countries.”

Xycoon Scientific Resources, `www.xycoon.com`