

Douglas R. White © 2006

Social scientists (Pareto 1896, Zipf 1949) have long been enamored with power laws, but rarely have understood what they mean (Krugman 1996), why they occur in the form that they do, how they should be explained or bounded by empirical limits, and when they should be dismissed as artifact. Only recently have social scientists begun to examine generalizations of power laws that involve functions that asymptote to power-law tails. It is useful to have a tutorial for these methods alongside a comparison with power laws. This tutorial draws on Borges (2004) and Thurner and Tsallis (2005).

The chief characteristic of a power law is linearity in log-log relationships. The linear slope is invariant if everything is multiplied by a constant. The intercept is in some sense irrelevant, as it is merely a scaling parameter to accommodate, say, sample size or other constraints. The power law is self-scaling or scale-free in the sense that the same or a similar relationship is implied at different magnitudes of scale. The fitting of a function $y(x) = Ax^\alpha$ to an empirical distribution is simple: log x , log y , plot, find the best (perhaps weighted) least-squares fit to a straight line, and extract the intercept A and slope α and the R squared goodness-of-fit.

A function $y(x)$ that asymptotes in the tail to a power law is in some sense a generalization that subsumes the power law if it entails stretching the power-law segment until it encompasses the whole distribution. Typically such a distribution can do so simply by reducing its scale parameter κ , which reflects in some way an inflection point at which the power-law segment begins.

1- and 2-parameter fits of a q -law

Let $P(\geq x)$ be the cumulative probability for empirical values of a variable binned in x . The q -logarithm of such a distribution is

$$Z_q(x) \equiv \ln_q[P(\leq x)] \equiv \frac{P(\leq x)^{1-q_c} - 1}{1 - q_c} \quad (1)$$

The fit is done by computing q -log distributions for different values of q_c . The function Z_q that is the best fit with a straight line determines the value of q_c . An example is shown in Figure 1 (Thurner and Tsallis 2005), the semi- q -log graph. The slope of this line is the value of $-\kappa$ (kappa) for the q -exponential distribution

$$e_q^x \equiv [1 - (1-q)x]^{1/(1-q)} \quad (2)$$

If we know the full distribution for $P(\leq x)$ then fitting of distribution involves optimizing on a single parameter, q , plus the normalizing assumption $P(\leq x_{\min}) = 1$. This would be the case, for example, for fitting the degree distribution of a network, where the number of nodes with zero degree are known (or can easily be extrapolated). For city size distributions, however, we need an additional parameter, $Y(0)$, an estimate of the total population in cities. In any

case, optimizing fit on 2-parameters is needed where $Y(0)$ must be estimated. Note that the slope in Figure 1 is $-\kappa (1 + (1-q) \ln_q Y(0))$.

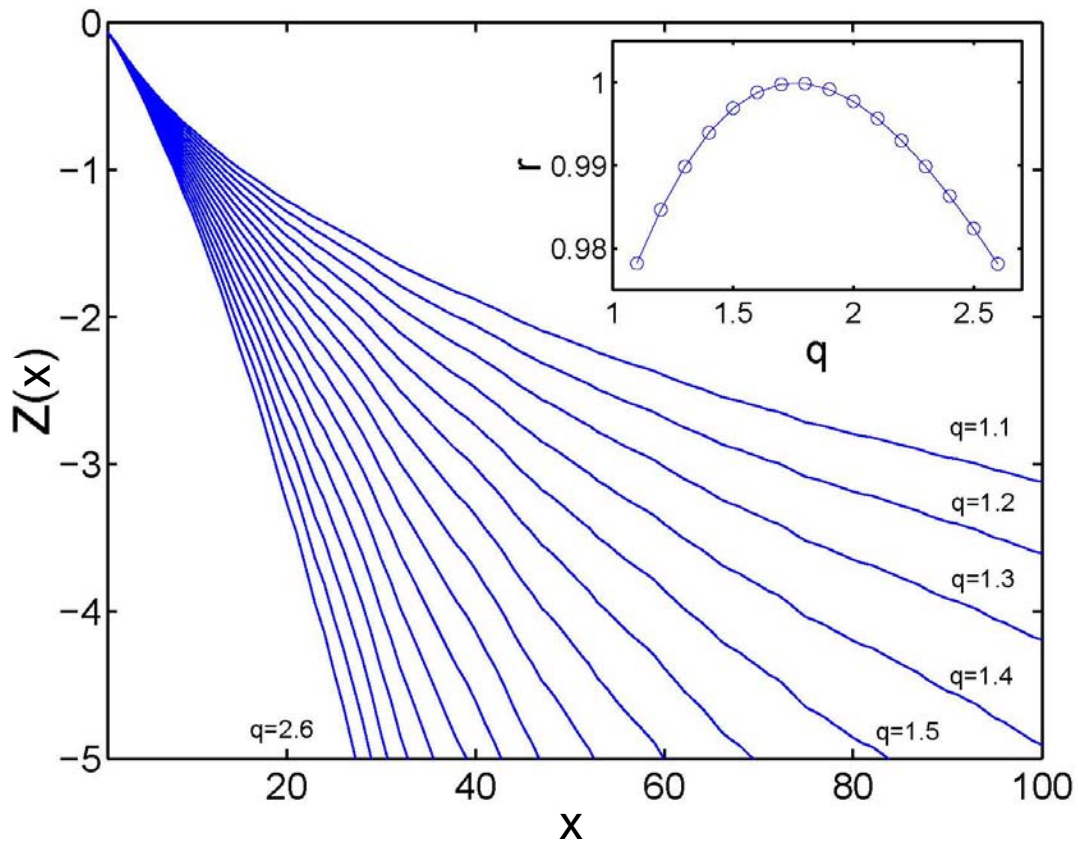


Figure 1: q -log fit (after Thurner and Tsallis 2005)

Figure 2 provides the interpretation for the relation between the three parameters that define the q -distribution.

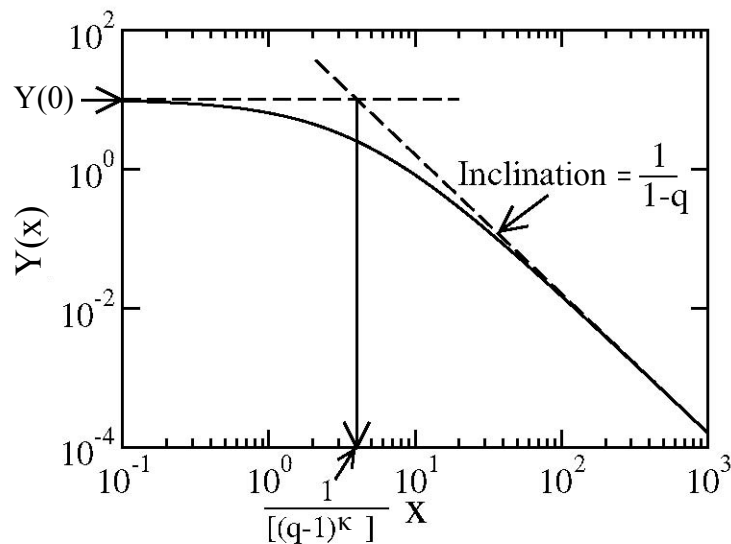


Figure 2: Interpretation of q , κ , and $Y(0)$ (after Borges 2004)

The full q -entropic distribution is

$$S_q \equiv Y(0)[1 + (1-q)x/\kappa]^{1/(1-q)}, \quad (3)$$

where x indexes the distributional variable and $Y(0)$ is the intercept for S_q^s on the ordinate for $x=0$. Here, $\alpha \equiv -1/(q-1) = 1/(1-q)$ is the asymptotic power-law slope to which the tail of the distributions converges, as in Figure 2 where $q \geq 1$.

Entropic q has been shown to fit degree distributions for various types of networks (Soares et al. 2004, Thurner and Tsallis 2005, White et al. 2006). It also provides a nearly perfect theoretical framework in which to scale the sizes of urban populations in which power laws dampen in the body of the size distribution, as recognized by Zipf (1949).

The theory of the q -entropic law is based on a generalization of the Boltzmann-Gibbs (BG) entropy measure (Tsallis 1988), which assumes statistical independence for all degrees of freedom. For social science, BG is equivalent to the usual null hypothesis given degrees of freedom. In the q -law where $q = 1$, $S_{q=1} \equiv S_{BG}$, where S_{BG} is the standard measure of BG entropy. To the extent that q varies from 1, either for $0 \leq q < 1$ (monotone increasing when κ is positive, decreasing when κ is negative) or for $q > 1$ (monotone increasing when κ is negative, decreasing when κ is positive), the power-law tail of the q -law grows steeper.

References

Borges, Ernesto P. 2004. *Manifestações Dinâmicas e Termodinâmicas de Sistemas Não-Extensivos*. Tese de Doutorado. Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro.

Chandler, Tertius. 1987. *Four Thousand Years of Urban Growth: An Historical Census*. Lewiston, N.Y.: Edwin Mellon Press.

Krugman, Paul. 1996. Confronting the mystery of urban hierarchy. *Journal of Political Economies* 10(4):399-418.

Pareto, Vilfredo. 1896. La courbe des revenus. *Le Monde économique*.

Soares, Danyel J. B., Constantino Tsallis, Ananias M. Mariz, and Luciano R. da Silva. 2004. Preferential attachment growth model and nonextensive statistical mechanics. arXiv:cond-mat/0410459. <http://arxiv.org/abs/cond-mat/0410459>.

Thurner, Stefan, and Constantino Tsallis. 2005. Nonextensive aspects of self-organized scale-free gas-like networks. *Europhysics Letters* 72 (2):197-203. SFI Working Papers, <http://www.santafe.edu/research/publications/wpabstract/200506026> DOI: SFI-WP 05-06-026.

Tsallis, Constantino. 1988. Possible Generalization of Boltzmann-Gibbs Statistics. *J. Stat. Phys.* 52:479-487. http://arxiv.org/PS_cache/cond-mat/pdf/0311/0311438.pdf

White, Douglas R., Nataša Kejžar, Constantino Tsallis, Doyne Farmer, and Scott White. 2006. A Generative Model for Feedback Networks. *Physical Review E* 73, 016119. SFI Working Papers, <http://www.santafe.edu/research/publications/wpabstract/200508034>. DOI: SFI-WP 05-08-034.

Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge MA: Addison Wesley.

Appendix 1: Using Excel for 2-parameter city size q-fits

Figure 3 shows in row 3 cumulative populations (in thousands) for China in the year 900 CE (Chandler 1987), according to the size bins in row 1, an estimate Y_0 of total population in cities in cell D3, and cumulative proportions of the Y_0 population (for equation 1, $P(\geq x)$), in row 2, columns E-F. Three possible values of q are shown in column A, cells 4-6. The row for each q value transforms the data in row 2 according to eqn. (1), e.g., $((E2^{(1-\$A4)})-1)/(1-\$A4)$. The zero is added in cell D1 to fix the data line to the origin when asking to fit a linear trendline with intercept set at 0 (thus the coefficient in front of the x is the negative κ and the R^2 is the goodness of fit to a straight line through the origin. These results are for an unweighted linear regression, but in our final analyses it is best in this case to use the weights provided by the cumulative number of cities in each bin, shown in row 7.

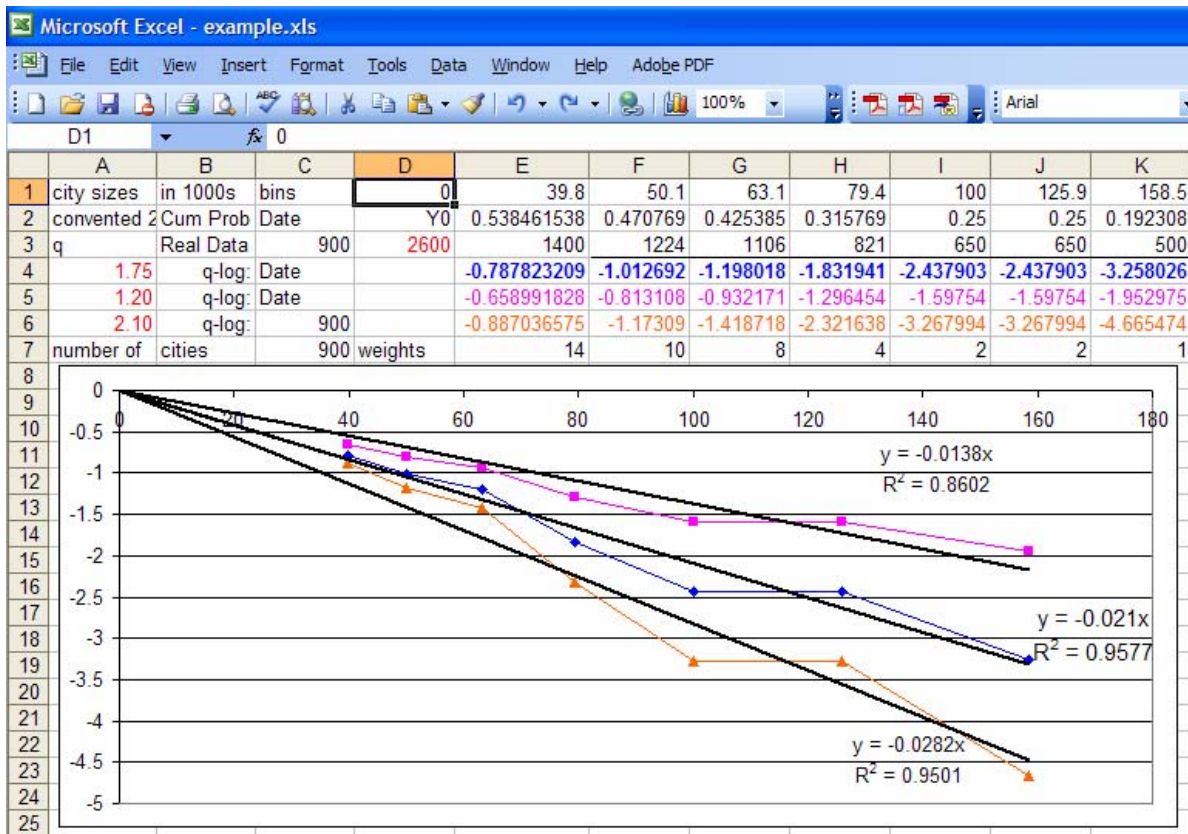


Figure 3: Spreadsheet showing Optimization of q and Y_0

Question for Ernesto: What is the exact formula for kappa, after finding optimal $Y(0)$ and q ?

$\kappa =$ what $f(Y(0), q)$? In your thesis you say the slope of monolog plot $\ln_q(x)$ by x is $-\beta_q(1+(1-q)\ln_q A)$.

Appendix 2: Using Excel for 1-parameter network degree q -fits