

Appendix B. Tutorial for using q -exponentials for city size distributions © Douglas R. White 2006

From:

*What do we know and not know about the Laws of City System Rise and Fall?
City-Size Distributions, Population, and Sociopolitical Instability in the case of China¹*

Douglas R. White^{1,4}, Laurent Tambayong¹, and Nataša Kejžar² © 2006

For: George Modelski, Tessaleno Devezas and William Thompson, eds.

Globalization as Evolutionary Process: Modeling, Simulating, and Forecasting Global Change

The purpose of this tutorial is to allow researchers to replicate a q -exponential analysis for city size (or other) distributions using Spss with data in the formats described, and to use the results productively.

To construct the Spss.sav data file, enter constant multiple city size bins 1-n in column 1, n being the largest city size, name the variable “bins.” In additional columns, one for each period, enter the (cumulative) frequency distributions for each period.² The mathematical model to be estimated is:

$$Y(x) = Y_0(1 - (1 - q)x/\kappa)^{1/(1-q)}$$

This function fits distributions that vary systematically as if “between” a power law and an exponential curve, as illustrated by our sample data from 900, 1500 and 1950 CE.

¹ This research benefited from generous contributions of data, commentary, methodology, and suggestions from Peter Spufford, Céline Rozenblat, Constantino Tsallis, Ernesto Borges, Peter Turchin, Chris Chase-Dunn, George Modelski, Hiroko Inoue, Daniel Pasciuti, Michael Batty, Denise Pumain, and many others. Our thanks to the Santa Fe Institute for support of this project at three different critical stages: a workshop what brought Turchin, White, and Spufford together; International funding that brought Nataša Kejžar to SFI to work with White and Tsallis, and core funding allowed White to work with Tambayong at SFI and to get feedback from Tsallis and Borges. White also thanks the ISCOM project funded by the European Union (PIs Lane, West and van der Leeuw) for support of his work on this project while in Europe.

² Fill zero frequencies below each highest size in each column with .0001 if power laws are also to be calculated. This will not affect calculation of q .

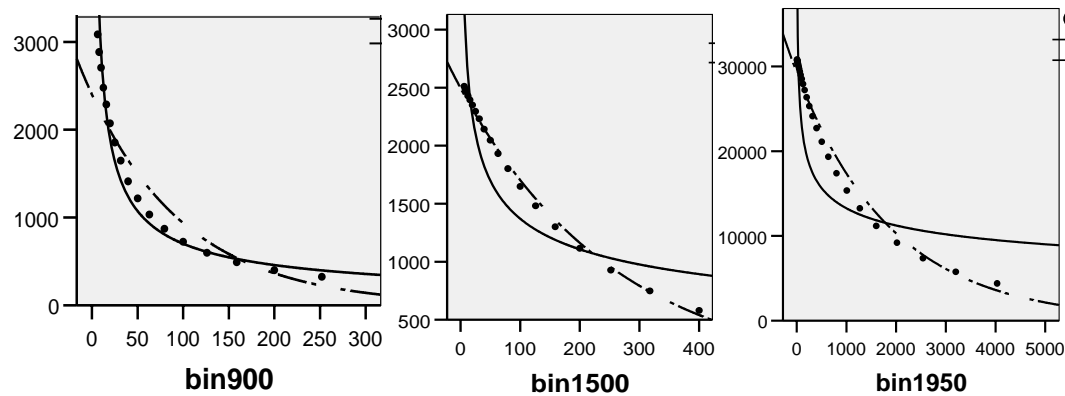


Figure B.1. Illustrative unlogged cumulative population plots for 900, 1500, 2000, fitted to power law and exponential, bins up to M , the largest expected city size.

Why not something simpler, like the rank-size rule?

Changes in city size distributions would seem an obvious and natural candidate for studying the rise and fall of city systems, but how best to characterize changes in city size distributions? What leads us to favor the particular construction of Figure 4?

The obvious measure is agreement or deviation from the rank-size rule $s=1/r$ or size s frequency f inversely proportional to size $f=1/s^{\alpha-1}$, proposed by Auerbach (1913) and popularized by Zipf (1949). We reject the study of rank-size distributions because they are not invariant under deletion of strings of elements in the tail or at the base of the distribution. Even a perfectly constructed rank-size Zipfian, where each city of rank r has a size $s=M/r$, where M is the largest city size, has this defect: the estimate of the power-law slope is unreliable. Binning methods do not suffer this defect.

A generalized function that gives rise to complex networks

The model of city distributions that we employ is neither abstract nor vague once it is specified in concrete terms: we model these distributions with a simple function that asymptotes to a power-law in the tail, above the crossover but also asymptotes to a total population as sizes become smaller, below the crossover. Further, one of the parameters of this model (q) takes a value of unity where the entire distribution asymptotes toward randomness. Below unity, as $q \rightarrow 0$, the entire distribution asymptotes to a simple linear function. Above unity, there is a value q_z where the entire distribution converges toward one that, above the crossover, asymptotes in the tail to a Zipfian. Further, above unity, as the crossover diminishes, the entire distribution converges to a simple power-law.

Given all these very specific characteristics and predictions, one can easily guess that this distribution is highly constrained. Its parameters cannot be manipulated to produce almost any curve. Only a very narrow range of curves are generated. Yet, almost all urban distributions that we have studied fits this curve very closely, except in the case of the largest cities.

The q -exponential functions for size distributions

This curve is related to what is known as the q -exponential family of curves, which, like the probabilistic $p(x)=x^{-\alpha}$ power law, has only one parameter: $e_q(x) = (1 - (1 - q)x)^{1/(1-q)}$, where, as $q \rightarrow 1$, $e_{q=1}(x) \equiv e^x$, the ordinary exponential. Fitting a raw data distribution requires an additional constant, as in $p(x)=Ax^{-\alpha}$ or $Y_0 e_q(x)$.³ The inverse of the e_q function is $\ln_q(x)=(x^{1-q} - 1)/(1-q)$. These are called q -exponentials and q logs, but the Tsallis q -exponential, $Y_q(x)\equiv(1-(1-q)x/\kappa)^{1/(1-q)}$, contains a third parameter, κ or kappa, that models the crossover phenomenon, but retains the scale-free property of power laws if only in the tail of the distribution.

A generalized model that encompasses power laws

Our contribution to the problem of measuring rise and fall in city size distributions is to take a more general approach, one that is 1) more sensitive to measuring change over time, 2) indifferent to largest-city effects, 3) reflective of the entire size distribution except for the largest city, 4) predictive of the expected norm for the largest city given the power-law trend in the tail, 5) predictive of the crossover region where the power-law tail of urban distributions gives way to an exponential curve for the body of the distribution, 6) predictive of the total urban population in the region, and, finally, 7) suggestive of the smallest size at which a settlement can no longer support the specialized division of labor needed for a city.

Estimating q parameters using Spss

Samples of the Spss syntax command are given below for q -fitting our data for periods 900-1100 (variable names c900, c1000, c1100) based on the model. The first and third commands are similar, but for c1000 the optimizing sweep over q began at 1.08 and worked downward, while k and y began lower for than the other periods. The idea is to have the optimizer sweep from one extreme to the other for each of the parameters (c1000: e.g., $q=4$ for $.01 \leq q \leq 4$) or to start with a middle value and sweep to either extreme (c1100: e.g., $q=1.56$ for $.01 \leq q \leq 4$). Use of a good sweep strategy may avoid local optima. The /bootstrap option computes estimates of standard errors, while r^2 fit is given automatically. Bounds may be adjusted for different years as the search ends at a boundary. Fits the Chandler data below $r^2=0.98$ should be rechecked for coding errors and for failure to reach a global optimum.⁴ For our four most recent periods with highest urban population Y0 again has to be set and constrained to 999999.

³ Fortunately, if Y0 is the total urban population, cumulative distributions for city sizes do not fit the function $Y_0 e_q(x)$. If they did, cities might be considered to have multifractal geometries and their dynamics might be those characteristic of chaotic dynamical, where small differences in startup conditions would lead to very different outcomes, that is, subject to the butterfly effect.

⁴ One trick we discovered to make sure the optimization procedure has converged is to compare the κ input to the κ that is output, and keep entering the parameter k(output) and constraint new $k \geq$ old k until the output equals your last input. The reason is that with small κ there may be a relatively flat zone for optimization and accuracy is raised when κ increases. Also, once scaling is done with $q > 1.001$, it might be done again for $q < 1$, setting parameter $q(.999)$ and setting $q \leq .999$ and $q \geq .001$ as constraints. Usually this will not be needed if the R^2 is high.

* NonLinear Regression.

MODEL PROGRAM Y0 = 9999 q=1.56 k=8.

COMPUTE PRED_ = Y0*(1-(1-q)*binlogged/k)**(1/(1-q)).

CNLR C900

/BOOTSTRAP

/PRED PRED_

/BOUNDS Y0 >= 1000; Y0 <= 9999; q <= 4; q >= .001; k >=8; k<=2000

/CRITERIA ITER 200 STEPLIMIT 2 ISTEP 1E+20 .

* NonLinear Regression.

MODEL PROGRAM Y0 = 4000 q=1.08 k=25.

COMPUTE PRED_ = Y0*(1-(1-q)*bins/k)**(1/(1-q)).

CNLR C1000

/BOOTSTRAP

/PRED PRED_

/BOUNDS Y0 >= 1000; Y0 <= 99999; q <= 4; q >= .01; k >=5; k<=2000

/CRITERIA ITER 200 STEPLIMIT 2 ISTEP 1E+20.

EXECUTE.

* NonLinear Regression.

MODEL PROGRAM Y0 = 99999 q=1.56 k=8.

COMPUTE PRED_ = Y0*(1-(1-q)*bins/k)**(1/(1-q)).

CNLR C1100

/BOOTSTRAP

/PRED PRED_

/BOUNDS Y0 >= 1000; Y0 <= 99999; q <= 8; q >= .001; k >=8; k<=2000

/CRITERIA ITER 200 STEPLIMIT 2 ISTEP 1E+20.

EXECUTE.

To see the distribution generated by a fitted function $Y_q(x)$, a typical syntax command has the fitted parameters inserted as constants:

```
COMPUTE v1800= 7671*(1-(1-2.960)*bins/ 57.7)**(1/(1-2.960)). EXECUTE.
```

Important constraints on model fitting for city urban populations include $Y_0 \ll P$ where P is the total urban and nonurban population, and $\kappa/(q-1) < M$ if $q > 1$ where M is the maximal city size.

GENERAL ROBUSTNESS AND PRECISION

Robust mathematical properties of the q -exponential $Y_q(x)$ are easily demonstrated. Precision is added because the equations for derivatives of Y_q are known.

Invariance in q -fitting properties. What exponential- q provides is a model for curvature in the body of the city size distribution. To demonstrate invariance in output parameters under perturbation of the data sampled, we (1) fitted the exponential- q parameters to our data for each period, (2) generated the entire $Y_q(x)$ distribution from arbitrarily small to arbitrarily large bins, (3) selected a series of arbitrary slices of the data containing low, medium and high line segments, and segments with smaller or larger

values truncated, and (4) refit the exponential- q parameters to the segments of the $Yq(x)$ distribution. In each case, identical q parameters were returned by Spss. So long as the empirical data fit an exponential- q distribution, it does not matter what segment of the data is used. Truncating the largest cities does not change the parameter estimates. In this sense, exponential- q is highly robust. The reason is that when $\ln_q(x)$ of the real data series x is fitted against x , after q has been fitted with a high r^2 , the plot of $\ln_q(x)$ by x forms a straight line. The more the data fit this straight line, with small random error components, the more any attempt to refit a segment of the data will yield the same values of the parameters q , Y_0 and κ .

Solving for M, the theoretical maximal city size consistent with fitted Yq

We can solve for $M=x=Yq(x)$ exactly, the theoretical maximal city size at point M according to the model represented in Figure 4, by optimizing the difference in the values of x and $Yq(x)$. Thus we set $M-Y_0/k [1-(1-q)M/k]^{(q/(1-q))}$ as a minimization objective. This is simple to minimize in Excel solver, changing x to minimize $(x-Yq(x))^2$. The exact solution for fitted $Yq(x)$ for the year 900 is $x=288$. A trace of these operations in the excel spreadsheet looks like this after optimization:

```
x= for cell a2 =(3909*(1-(1-2.010)*x/22.50)^(1/(1-2.010))) for b2
288.04 288.04 0.00 =(a2-b2)^2 minimize in c2
```

Derivatives

The following Spss syntax generates the derivative $Y'_q(x)=Y_0/k [1-(1-q)x/k]^{(q/(1-q))}$ of $Y_q(x)$ at each bin size x .⁵ The derivative for the c900 data in the figure at largest $M=288$, for example, is -0.92 and at a small city $x_{\min}=6.25$ it is -106. Note that because the derivatives are for Yq distributions, as in Figure B.1, they are monotonically decreasing because as x increases, $f(x)$ falls by decreasing amounts.

```
**[Y0 / k] [1-(1-q)x/k]^(q/(1-q)) Borges derivative
COMPUTE v900dEB= (3909/22.5)* (1-(1-2.01)*bins/22.5)**(2.01/(1-2.01)) .
EXECUTE .
```

Maximal Likelihood Estimates for q -exponentials

By normalizing our cumulative distribution by dividing by Y_0 once estimated, we could use maximal likelihood estimation (MLE) of the other two parameters in the model, κ and q . But since Y_0 in the case of cities has to be estimated itself, MLE would have an additional element of complexity. We would want to make sure that our Y_0 estimates are exactly on target before undertaking MLE.

⁵ Courtesy of Ernesto Borges (2004a). A formula used by Tsallis for the derivative $d Y(x) / dx = [Y_0 / k] [Y(x) / Y_0]^q$ did not work, perhaps due to arithmetic overflow or some other problem. In Spss:

```
**[Y0 / k] [Y(x) / Y0]^q Tsallis derivative
COMPUTE v900dCT= (3909/22.5)*(((3909*(1-(1-2.01)*bins/22.5))**(1/(1-2.01)))/3909)**2.01 .
EXECUTE .
```

Comparison to Power-Law Fit

One criticism of the q -exponential is has three parameters while the power-law has only two. The one extra parameter, however, actually buys not just one extra dimension that the power-law does not, but many: an estimate of the crossover at which the change is gradually taking place between an exponential and a power-law distribution, a derivative that translates to precise descriptions of demographic implications, a means of evaluating the deviations of primate cities from the general body of the distribution, and a whole series of other measures described below.

Comparisons of power-law coefficients across historical periods pose a number of problems. First, the power law relationship among cities sizes applies only to the largest cities, as was recognized by Zipf and all the subsequent researchers of size distributions. Second, the crossover at which a given distribution ceases to follow the power-law cannot be derived from first principles, and in any case, is difficult to assess. Third, the largest city in a size distribution is the most variable in its relation to others, and only adds unreliability and confusion to estimating a power-law slope, Largest cities may deviate from a power-law regression line because it is a capital city, a financial hub, a depressed economy, or an object of external attack, such as conquest of a capital.

CONSISTENCY CHECKS

Regularities for distributions people in cities must obey physical constraints and consistencies. We did the following consistency checks in relation to the each dataset under study. In general, when adjustments are required physically conservative estimates should be preferred.

The Real World Consistency Checks for Y_0

When the q -exponential spits out a value for parameter Y_0 , it is giving an estimate of the total urban population. One check is absolute: the *urban* population cannot exceed that of the region as a whole, as does the estimate of Y_0 from the curve of the misconstrued UN city distribution for 2015 relative to the likely total population of China, which just passed one billion.

The other consistencies expected of Y_0 are relative. First, relative to the total population P , the Y_0/P ratio, which determines the *percentage urban* or degree of urbanization, has to be (1) reasonable, (2) fairly consistent through time, and (3) consistent with general historical knowledge of the period and the adjacent periods.

Checking for Multiple Optima

A problem in optimization occurs if there are fewer than five bins with information, in which case the solution space may be flat near the optimum. For the year 1970 in the Chandler (1987) data, for example, the Y_0 and κ we obtained did not resemble those of adjacent time periods. We limited the constraint space in different ways to study how the

tradeoffs between Y_0 and a low κ made little difference to the tail or the fit of the model to the tail. Ultimately, we chose, among equally optimal solutions, the one with lowest Y_0 , which entailed the minimal assumptions about the total number of people living in cities.

Consistency and Variation in C, the crossover. When $q > 1$, the place where crossover occurs, as located to the left and right of point C in our Figure 4, is an important historical variable. For $q > 1$, this is computed (Borges 2004a:34-36) as

```
COMPUTE C_rossover=k/(q-1).
EXECUTE .
```

The crossover should be greater than Y_0 . If it is larger than M it may indicate, along with the r^2 , a poor fit.

Because $q > 1$ and κ is constrained to be positive, the ratio of κ/M must be positive, but κ will usually be less than M and always less than Y_0 . The ratio $C/M = \kappa/M(1-q)$ can a useful *flat-world ratio* of the urban distribution, referring to that part of the city residents who become fewer at smaller sizes and do not participate in the power-law scaling of populations in urban hubs.

Validating q for the body of the distribution

The use of q for a sample of largest cities is intended for the case, and works best, where there is actually a bend in the log-log cumulative distribution, such as we see for many of the curves in Figure 4. If the cities form a perfect line, or power-law in the cumulative distribution, q will be estimated but there is no crossover point C, no way to estimate κ , and no way to know the asymptote to Y_0 . Estimating Y_0 and the crossover is one of the great advantages of exponential- q . If the phenomena studied is q -exponential (which implies the bend and the crossover), then Y_0 should converge to the actual total urban population. The fact that so many of our distributions do have bends and crossovers from a power law and that so many of our cases fit the Y_0 parameter correctly – given what we know about the total population and the percentage urban – is evidence that exponential- q is an appropriate model for city size distributions.

Spotting Bad Forecasts from q -exponential inconsistency

The UN data that we analyzed for 1950 forward, in five-year intervals, included city distribution forecasts for 2005, 2010, and 2015. The fitted curve for 2015 was included in our Figure 4, with the comment “an unlikely (and poorly fitted) 3.342 billion Y_0 estimate for 2015 UN urban China projections.” This curve did not match those of the other empirical curves, either those fitted for the Chandler data or the UN’s own data.

Unusually high q values

Usually high q values, in the range 3-4, combined lower r^2 fit, may indicate a fit that should be checked as power-law throughout the distribution. The q -exponential has

difficulty fitting a straight line in a log-log plot. To demonstrate and calibrate this possibility we generate a Pareto distribution for number of cities for each b as $c(b)=Y_0 b^{-\alpha}$. Then the number of people is $n(b)=b Y_0 b^{-\alpha}=Y_0 b^{1-\alpha}$. For $\alpha=1$ this entails an identical number of people in each city size bin. If $\alpha<1$ the number of people in lower bins is decreasing, and if $\alpha>1$ the number in lower bins is increasing. Our historical data typically show the former. In this type of cumulative distribution, as in Figure 4, the urban population in bin j is $U_j=Y_0 \sum_{i=j}^B b_i^{1-\alpha}$. Figure B.2 shows simulated data summarized in cumulative plots. Here, α is varied for .6, .7, .8, and .9, Pareto coefficients for decreasing people in lower binds. When we fit exponential- q to these curves we get the values shown in Table A1. These recover the actual cumulative urban population plus $5\% \pm 1\%$ (i.e., within 5% accuracy) and the Pareto α coefficients within 2% accuracy using $\alpha=-(q+1)/5$. For a power-law generator of the number of cities per bin the exponent is recovered from $\alpha=(q/4.1)^{.7}$, with the inverse function $q \sim 4\alpha^3$ which for $\alpha=1$ is $q=4$. Pareto power law or quasi-Zipfian city distributions are thus a special case of exponential- q .

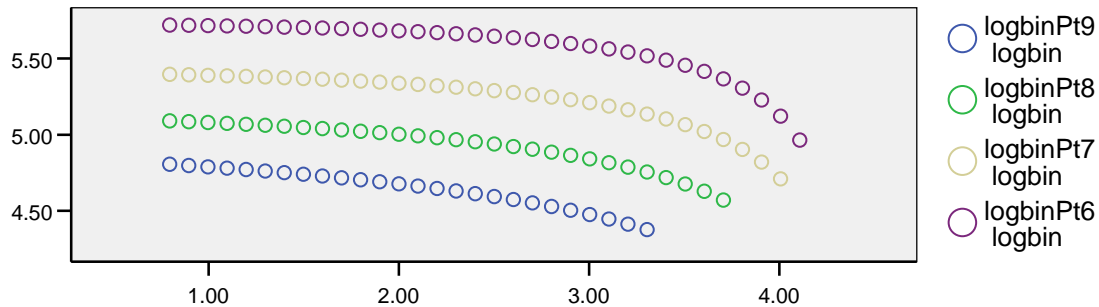


Figure B.2: Four simulated cumulative distributions, x axis the log of city size and the y axis the log of cumulative urban population at those bin sizes

Table A1 Asymptotic tip of the distribution

* α *	kappa	q	Y_0	R^2	Tip β
0.49		1.50			-2.00
0.50	5800	1.56	1107354	0.989	-1.80
0.60	3628	2.02	502347	0.985	-1.04
0.70	2000	2.52	236046	0.960	-0.66
0.75			$\frac{3}{4}$ Zipfian		-.55?
0.80	960	3.02	116138	0.980	-0.44
0.90	408	3.44	60655	0.980	-0.33
1.00		4.00			

A Pareto $\alpha \sim 1$ has been ascribed as the normal state of urban distributions. That would make $q = -5\alpha - 1 = 4$ the “Zipfian” power-law coefficient for q . All of our observed variations are for $q < 4$ and hence $-5\alpha - 1 < 4$, thus $\alpha < 1$ for historical q .

INCLINES: FROM BODY TO TAIL, CITIES TO POPULATIONS IN CITIES

Slope Measurements

Although the cumulative populations of cities and of populations in cities can each be extrapolated from the other, we created the cumulative population by city size distribution for three reasons. One is for purposes of comparative visualization, as with Figure 4 – to see the variation in population curves. Second, by using averaging, cumulative curves give highly accurate estimates. Averaged number of cities of different sizes and averaging numbers of people in cities of different sizes, however, are vastly different.

Demographics and Levels

The cumulative distribution of population rather than number of cities is directly translatable into demographics. Since these two sets of curves take different shapes, we wanted to see if there are laws of cities at one level that we fail to see or comprehend at the other. As we have shown in Appendix A, for example, the scaling results methods of q scaling of the bodies of cumulative urban population distributions, when compared to power law models, give very different results, especially as for differentiating historical periods. Our evidence also shows that cumulative population distributions are more likely to give more accurate and reliable measures than cumulative numbers of cities.

Flatland Indices: Flatland Cities versus Hubs

We can directly integrate the number of urban people who live in the power-law tail of that distribution versus outside the tail, translating the *flatness ratio* of the distribution into a *flatland population percentage* of urban residents according to the number of people in flatlands and urban hubs. We can also separate, above and below the crossover, the *flatland cities* from the *urban hubs* according to the characteristic crossover population size for each period.

Smallest Cities Bounds and Invariance

In the q -exponential model of cities as we have formulated it, Y_0 is always less than P , the total population size, and in our analyses averages about 7% of the total Chinese population until very recently. The bound of what is a “city” is thus established by the model itself! What this entails is that Y_0 marks a convergence to some total urban population, and that we move down the scale of sizes, while there may be increasing numbers of settlements at smaller sizes, *the proportion that are urban* is decreasing, if only according to the processes specified in the q -exponential model.

This is extremely helpful with respect to comparisons with what constitutes a city in China’s contemporary census definitions, which classifies all settlements down to size 2,000 as “urban.” If we compared the number of cities at that bin size (2,000-2,520) estimated from our q -scaling with the number in the Chinese census of cities, about 1% of the latter would be consistent with the q -exponential estimate.

Archaeologically and historically, what constitutes a “city” is relatively well defined, and includes a complex and specialized division of labor capable of producing multiple and

shifting produces for export, and showing signs of the importation of luxury as well as staple goods. An informal survey of archaeologists places the limiting minimal city size at 5,000.

The derivatives of $Y_q(x)$ at arbitrarily small sizes (like our hypothetical x_{\min} value) can help determine the diminishing marginal return in population from smaller cities to check if there is some relatively uniform x_{\min} size over all time periods at which this return is vanishingly small, like 3-4% of the total population in settlements at a given small size. The idea here is that *not all settlements of 5,000 or above are cities*, but that those that are can easily be identified among the settlements of a given size class. At smaller sizes, a diminishing fraction will be true cities, and the rest rural settlements. Thus we consider that the q -exponential model gives basically the right intuitions for cities.

The q -exponential function behaves very differently than the lognormal distribution as a model for city populations posed by Gibrat's principle. It is an unresolved issue in rank size scaling as to whether the lognormal, which often approximates the rank size, is taken to fit all population clusters down to the smallest level and therefore does not necessarily involve a lower bound other than single individuals.⁶

To go further with his analysis of percent cities among all settlements of smaller size classes, we would need total population numbers at different settlement sizes. We will not take up this problem here because at present we do not have the China city data broken down by numbers of cities. Whether further investigation would find some invariance principles at this level is difficult to say, but there is no evidence from the q -modeling results that the cutoff size for "smallest" cities changes with historical evolution of the cumulative urban population, the total population, or the percentage "urbanization," which the q -model seems uniquely able to specify in an asymptotic manner that is unbiased by arbitrary settlement size cutoffs.

Maximum Likelihood Estimation (ML)

Formulation of ML methods for q exponential estimation is beyond the scope of this tutorial but will involve a probability model where $P_q(S \geq x) \equiv Y_0 (1 - (1-q)x/\kappa)^{1/(1-q)}/T$, where T is a normalizing factor that is also an unbiased total urban population ML parameter converging with Y_0 .

⁶ Thanks for this comment to an anonymous reviewer. This reviewer also asked: Do changes in the Gibrat law lead to the different q -exponentials? Can this be rephrased, since the size of units and their growth percentage statistics are statistically independent under Gibrat's law, to say: if one generated different sizes of units and their growth percentage statistics independently, are all such distributions consistent with the q -exponential? The answer should be no, the q -exponential is much more tightly constrained. This might be a good subject for a simulation, following the simulation of the Gibrat law in Batty (2005).