Size Distributions for All Cities: Lognormal and q-exponential functions

González-Val, Rafael and Ramos, Arturo and Sanz-Gracia, Fernando

Universidad de Zaragoza

9 September 2010

Size Distributions for All Cities: Lognormal and q-exponential functions

Rafael González-Val[a]

Arturo Ramos-Gutiérrez[b]

Fernando Sanz-Gracia[b]

[a] Departamento de Economía Política y Hacienda Pública.Universitat de Barcelona & Institut d'Economia de Barcelona (IEB). Facultat d'Economia i Empresa. Av. Diagonal 690, 08034 Barcelona (Spain). E-mail: rafaelg@unizar.es

[b] Departamento de Análisis Económico. Universidad de Zaragoza. Facultad de Ciencias Económicas y Empresariales. Gran Vía 2, 50005 Zaragoza (Spain). E-mail: aramos@unizar.es, fsanz@unizar.es

Abstract

This paper analyses in detail the features offered by a function which is practically new to Urban Economics, the q-exponential, in describing city size distributions. We highlight two contributions. First, we propose a new and simple procedure for estimating their parameters. Second, and more importantly, we explain the characteristics associated with two traditional graphic methods (Zipf plots and cumulative density functions) for discriminating between functions. We apply them to the lognormal and q-exponential, justifying them as the best functions for explaining the entire distribution, and that the relationship between them is of complementarity. The empirical evidence relies on the analysis of urban data of three countries (USA, Spain and Italy) over all of the 20th century.

Keywords: city size distribution, q-exponential, lognormal

JEL: C13, C16, R00.

1. Introduction

The study of city size distribution has a long tradition in Urban Economics. To cite just a few examples, see Rosen and Resnick (1980), Black and Henderson (2003), Sharma (2003), Ioannides and Overman (2003), Soo (2005), Anderson and Ge (2005), Bosker et al. (2008). These distributions have an interest beyond the purely statistical, essentially for two reasons, which feed back to and influence each other. First, because city size distribution defines the resulting economic landscape – it may be more concentrated or dispersed, or biased towards an excessive number of large or small centres, with cities which are similar or very different in size – and all of this impacts directly on the spatial distribution of income, on public investment in infrastructure of various kinds in certain areas, and on imbalances between territories in general. And what politician would dare to say these subjects do not interest him? And second, because this size distribution is susceptible to change over time, according to certain, essentially economic, incentives.

Historically, the Pareto distribution has generated more works and greater acceptance; the density function of this power law is given by:

$$P(Size \geq x) = \frac{a}{x^b}, \qquad (1)$$

where a is a constant, $b > 0$ is the Pareto exponent and x is the number of inhabitants of each urban centre. Considering the rank r (1 for the most populous centre, 2 for the second, and so on) of the N cities we can obtain the well-known expression

$$\ln r = cons - b \ln x, \qquad (2)$$

which relates the logarithm of rank with the logarithm of the size of the cities if they follow a Pareto distribution. In the case that $b = 1$, we obtain the well-known Zipf's law

or rank-size rule (see surveys on this subject by Cheshire, 1999, and Gabaix and Ioannides, 2004).

Zipf's law is relevant fundamentally for three reasons. One, it is applied to other quantifiable phenomena with a fairly close fit, such as the flow rate of rivers, the number of times the same word appears in a text, or the intensity of earthquakes (Zipf, 1949; Krugman, 1996a). Two, if we consider the 135 Metropolitan Statistical Areas (MSAs) existing in the USA in 1991, the Pareto exponent is 1.005 (Krugman, 1996b, also shown in Gabaix, 1999), meaning that this law is fulfilled almost exactly. And three, there is a degree of consensus which holds that the urban structure arising from the fulfilment of the law defines a balanced hierarchy, in which cities of all sizes are well represented. In summary, the Pareto distribution, and a particular case of it such as Zipf's law, are certainly useful for explaining the behaviour of urban areas, especially the largest ones (upper tail distribution).

However, the description of Pareto's law has some substantial faults. On one hand, in a rank-size plot Pareto's distribution has a vertical asymptote in $x=0$. On the other hand, the Zipf plot (relating $\ln r$ with $\ln x$ on Cartesian axes) deviates from a straight line with a negative slope as predicted by (2) when, again, all urban areas are included, or the population cut-off is low enough. Much of the recent empirical work on the size distribution of cities has focused on deviations from Zipf's law (Soo, 2005; Garmestani et al., 2007; Garmestani et al., 2008; Rozenfeld et al., 2008; González-Val, 2010a).

In this order of things, the contribution of Eeckhout (2004) arrives, essentially proposing three ideas. One, that when all urban centres are taken, without any size restriction, Pareto's distribution falters and the best representation of the data is a lognormal function. Two, as a theoretical result: if the underlying distribution is

lognormal, which generates a concave Zipf plot, the Pareto exponent decreases with sample size, meaning that a sample size can be found which verifies Zipf's law exactly. These first two contributions clearly show the importance of taking all cities, as to do otherwise can lead to skewed or spurious results. And three, the data for all USA cities in 1990 and 2000 support the hypothesis of lognormality and the fulfilment of Gibrat's law, or the law of proportionate growth, something which was already anticipated from a theoretical viewpoint by Gibrat (1931) and Kalecki (1945).

Briefly, the two distributions most used in the economic literature have been Pareto's, and, more recently, the lognormal. Moreover, other statistical distributions have been proposed: q-exponential distribution (Malacarne et al., 2001; Soo, 2007), double Pareto lognormal distribution (Giesen et al., 2010). Ioannides and Skouras (2009) even proposed a new distribution function which switches between a lognormal and a power distribution. There is also an older literature that explores alternative functional forms: Hsing (1990), Kamecke (1990), or Cameron (1990). This paper is in line with all this literature.

In a physics journal, Malacarne et al. (2001) show that when all cities are taken, the so-called q-exponential distribution presents a very close fit to the data. As far as we know, the only other work to test this statement is that of Soo (2007), who, taking the largest cities of Malaysia (over 10,000 inhabitants) obtains negative results regarding the features of the q-exponential, leading us to think, as with the lognormal, that this distribution is especially suitable when no truncation point is defined.

Recently much more complete databases have been constructed, which enable us to bring more statistical information to bear on the problem dealt with in this work. Specifically, González-Val (2010b) considers all the population centres in the USA during the entire 20th century; González-Val et al. (2010) do the same for Spain and

Italy, as well as for the USA. If these data are used to represent the logarithm of the rank against the logarithm of city size, a clear deviation from linearity can be observed in all cases, opening the way for the consideration of non-Pareto distributions. What we want to emphasise is that, except for Eeckhout (2004), no previous studies consider the entire distribution of cities[1], as all of them impose a truncation point, either explicitly by taking cities above a minimum population threshold, or implicitly by working with MSAs[2]. This is usually due to a practical reason of data availability.

In this context, the aims of this article are as follows. First, to test the features of the q-exponential for describing city size distribution over a long period (a hundred years), for various urban structures (those of Spain, Italy and the USA) and considering all centres (about 8,000 for the Mediterranean countries for the whole century and from 10,600 to over 19,000 for the USA depending on the year). Second, to carry out the same exercise for the lognormal. And third, to weigh up the advantages and disadvantages of both distributions, q-exponential and lognormal, and to determine if they can be substituted for each other or if they are complementary in nature, and under what circumstances. As can be seen below, these three aims will lead to a simple contribution regarding estimation methods and to some developments of theoretical statistics, not at all complex, justifying and explaining the differences which arise between the q-exponential and lognormal distributions. In any case, as far as we know this is the first time that these matters have been subjected to empirical testing with such a large database.

[1] Michaels et al. (2010) use data from minor civil divisions (MCDs) to track the evolution of population across both rural and urban areas in the United States from 1880 to 2000.

[2] In the USA, classification as an MSA requires a city of at least 50,000 inhabitants or the presence of an urban area of at least 50,000 inhabitants and a total metropolitan population of a minimum of 100,000 inhabitants (75,000 in New England), according to the official definition. Other countries follow similar criteria, although the minimum population threshold needed to be considered a metropolitan area may vary.

This paper offers also three main contributions. First, it proposes a simple new way to estimate the parameters of the q-exponential distribution, improving on that hitherto used in the literature. Second, it explains the advantages and disadvantages associated with two traditional graphic methods (Zipf plots and cumulative density functions) in discriminating between density functions, and applies them to q-exponential and lognormal distributions. Last, it concludes that both distributions are suitable for describing city size distributions with precision, and that the relationship between them is basically of complementarity.

The article is organised as follows. The second section defines and characterises the q-exponential distribution; the third summarises and explains the databases used; the fourth is the longest, comparing and contrasting the q-exponential and lognormal as potentially valid functions for describing city size distribution; the fifth shows how both distributions are more complementary than interchangeable; finally, we end with the conclusions.

2. On the q-exponential distribution

The probability density function (pdf) of the q-exponential is given by:

$$f(x)=\frac{a}{q}\left(1+\frac{q-1}{q}ax\right)^{\frac{q}{1-q}}, \qquad (3)$$

where $a>0$ and $q>1$ are parameters and x denotes the population of urban centres. The mean and variance are $E[x]=\frac{q}{a(2-q)}\,(q<2)$ and $Var[x]=\frac{q^2}{a^2(2-q)^2(3-2q)}$ $(q<3/2)$.

The expression of the corresponding cumulative distribution function (cdf) is:

$$cdf(x) = 1 - \left(1 + \frac{q-1}{q} ax\right)^{\frac{1}{1-q}}, \qquad (4)$$

and that of the rank of cities according to population is

$$r(x) = r_0\left(1 - cdf(x)\right) = r_0\left(1 + \frac{q-1}{q} ax\right)^{\frac{1}{1-q}}, \qquad (5)$$

where $r_0 > 0$ is a new constant equivalent to the sample size. In the case that $q \to 1$, $f(x) \to ae^{-ax}$ and $r(x) \to r_0 e^{-ax}$, which justifies the name q-exponential.

This distribution was used profusely by Tsallis (1988) and his group of collaborators in physics literature, arguing for its theoretical applicability to systems with long-range interactions, and the cited work by Malacarne et al (2001) can be included in this line of argument. However, the q-exponential is a particular case of the distribution known as generalised type II Pareto, which has been considered in various earlier works (for example, Hosking and Wallis, 1987, Grimshaw, 1993 and Choulakian and Stephens, 2001). Rank compared to size, according to (5), is a decreasing function (something which by definition should always happen), strictly convex, and for $x = 0$ reaching a finite value, r_0, in this aspect improving the behaviour of the Pareto distribution.

3. The databases

We use city population data from three countries: the USA, Spain and Italy[3]. We have taken the data corresponding to the census of each decade of the 20th century[4]. Table 1 presents the number of cities for each decade, and the descriptive statistics.

[3] We use data from "legal" cities. However, there are problems of international comparability, because the administrative definition of city changes from one country to another. Although the concept of municipality used in Spain and Italy is very similar.

[4] No census exists in Italy for 1941, due to its participation in the Second World War, so we have taken the data for 1936.

The data for the USA we are using are the same as those used by González-Val (2010b). Our base, created from the original documents of the annual census published by the US Census Bureau, www.census.gov, consists of the available data of all incorporated places without any size restriction, for each decade of the twentieth century. The US Census Bureau uses the generic term *incorporated place* to refer to the governmental unit incorporated under state Law as a city, town (except in the states of New England, New York and Wisconsin), borough (except in Alaska and New York), or village, and which has legally established limits, powers and functions.

Two details should be noted. First, that all the cities corresponding to Alaska, Hawaii, and Puerto Rico for each decade are excluded, as these states were annexed during the 20th century (Alaska and Hawaii in 1959, and the special case of Puerto Rico, which was annexed in 1952 as an associated free state), and data do not exist for all periods. Their inclusion would produce geographical inconsistency in the samples, which would not be homogenous in geographical terms and thus could not be compared. And second, for the same reason we also exclude all the unincorporated places (concentrations of population which do not form part of any incorporated place, but which are locally identified with a name), which began to be accounted after 1950. However, these settlements did exist earlier, so that their inclusion would again present a problem of inconsistency in the sample. Also, their elimination is not quantitatively important; in fact there were 1,430 unincorporated places in 1950, representing 2.36% of the total population of the USA, which by 2000 would be 5,366 places and 11.27%.

For Spain and Italy the geographical unit of reference is the municipality and the data comes from the official statistical information services. In Italy this is the *Servizio Biblioteca e Servizi all'utenza*, of the *Direzione Centrale per la Diffusione della Cultura e dell'informazione Statistica*, part of the Istituto *Nazionale di Statistica*, www.istat.it,

and for Spain we have taken the census of the *Instituto Nacional de Estadística*[5], INE, www.ine.es. The de facto resident population has been taken for each city.

The USA is an extremely interesting country in which to analyse the evolution of urban structure, as it is a relatively young country whose inhabitants are characterised by high mobility. On the other hand we have the European countries, with a much older urban structure and inhabitants who present greater resistance to movement; specifically, Cheshire and Magrini (2006) estimate mobility in the USA is fifteen times higher than in Europe.

Considering these two types of country gives us information about different urban behaviours, as while Spain and Italy have an already consolidated urban structure and new cities are rarely created (urban growth is produced by population increase in existing cities), in the USA urban growth has a double dimension: as well as increases in city size, the number of cities also increases, with potentially different effects on city size distribution. Thus, the population of cities (incorporated places) goes from representing less than half the total population of the USA in 1900 (46.99%) to 61.49% in 2000; at the same time the number of cities increases by 82.11%, from 10,596 in 1900 to 19,296 in 2000.

4. Lognormal versus q-exponential. The substitution approach

4.1. Estimation methods

4.1.1. Estimation of the lognormal

The probability density function (pdf) of the lognormal is given by:

[5] The official INE census have been improved in an alternative database, created by Azagra et al. (2006), reconstructing the population census for the twentieth century using territorially homogeneous criteria. We have repeated the analysis using this database and the results are not significantly different, so we have presented the results deduced from the official data.

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}} \quad , \quad x>0, \qquad (6)$$

where μ and σ are the mean and variance of $\ln x$, which in this case denotes the natural logarithm of the population of the urban centres. The expression of the corresponding cumulative distribution function (cdf) is:

$$cdf(x)=\frac{1}{2}+\frac{1}{2}erf\left(\frac{\ln x-\mu}{\sigma\sqrt{2}}\right), \qquad (7)$$

where *erf* denotes the error function associated with the normal distribution. The expression of the rank of cities according to population is

$$r(x)=r_0\left(1-cdf(x)\right)=r_0\left(\frac{1}{2}-\frac{1}{2}erf\left(\frac{\ln x-\mu}{\sigma\sqrt{2}}\right)\right). \qquad (8)$$

Maximum Likelihood (ML) is often used to estimate if data follow a lognormal distribution, although this method has not been applied as frequently for the population of urban centres as it has been in other fields[6]. ML estimators are expressed simply in terms of population data:

$$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}\ln x_i \quad , \quad \hat{\sigma}^2=\frac{1}{N}\left(\sum_{i=1}^{N}(\ln x_i)^2-\frac{1}{N}\left(\sum_{i=1}^{N}\ln x_i\right)^2\right), \qquad (9)$$

where N is the sample size. That is, the estimated mean and variance are exactly those of the data. Later, we estimate r_0 by OLS taking into account the estimated *cdf* and the equation (8).

The estimates of these parameters for our data are very significant in the three countries and for all years. This information is shown in Table 2. The estimations of $\hat{r}_0$

[6] We recall that Eeckhout (2004) is the first to propose lognormality for city size distribution.

are directly related to sample size, as we already know; those of $\hat{\mu}$ are very stable over time for all three countries, while as one would expect, the values of $\hat{\sigma}^2$ increase slightly over time for the three areas. R^2, corresponding to the OLS estimation of r_0 applying equation (8), shows that the degree of fit is very good.

4.1.2. Estimation of the q-exponential

In their original article, Malacarne et al. (2001) begin with what they call the q-logarithmic function $\ln_q(x)$, which in econometrics can be understood as a Box-Cox transformation. They show that if (and only if) the value of q used in this function coincides with the value of q of the expression of rank $r(x)$ given by (5), then $\ln_q(r(x))$ is linear in x and can be represented as a straight line. That is, the value of q which we want to find must be known beforehand, which is a problem both conceptually and operationally. Elsewhere, Soo (2007) estimates (5) giving values to q one decimal at a time, choosing the one which minimises the sum of the residues and, once this $\hat{q}$ is determined, the values of a and r_0 are obtained by non-linear estimation; this is a solution to the problem which could be improved upon, as it also involves proposing arbitrary initial values for the parameter which it seeks to estimate.

Meanwhile, the distribution of the q-exponential is estimated by ML in various works, or, more correctly, the generalised Pareto distribution (on this subject, see the references cited in the second section). In all of these the number of observations used is not high. In our case the sample size is more than 8,000 for Spain and Italy, and goes from 10,600 to nearly 20,000 in the USA, and this means that using ML presents serious difficulties. Indeed, to see their origins for large sample sizes, the first order conditions for maximising joint likelihood are:

$$\begin{cases} \sum_{i=1}^{N} \dfrac{x_i}{1+\dfrac{q-1}{q}ax_i} = \dfrac{N}{a}, \\ -\dfrac{N}{q} + \dfrac{1}{(q-1)^2}\sum_{i=1}^{N} \ln\left(1+\dfrac{q-1}{q}ax_i\right) + \dfrac{N}{q(1-q)} = 0. \end{cases} \tag{10}$$

It can be deduced at first sight that the second equation in (10) is especially difficult to resolve, as the root must be found to a very high degree equation.

Given the great difficulty of using ML on one hand, and the problems of estimating the parameters of interest presented in the preceding works on the other, we have opted to begin by estimating the parameters of (5) using non-linear methods, without setting the values of any of them beforehand. However, this non-linear estimation requires some initial values for the parameters, which will be seen to be fundamental for the algorithm to converge on a reasonable number of iterations.

We have developed a novel method for finding these initial values. The procedure is as follows. Let p be a new exogenous parameter which may take the value of q. Let $\varepsilon = \dfrac{q-p}{1-q}$. Then, from (5) we have:

$$r(x)^{1-p} = r_0^{1-p}\left(1+\frac{q-1}{q}ax\right)^{1+\varepsilon}. \tag{11}$$

In the case that $p = q$ we have $\varepsilon = 0$ and the expression of $r(x)^{1-p}$ becomes linear in x. We can then take a Taylor expansion of (11) of order two about $x = 0$

$$r(x)^{1-p} = \lambda_1 + \lambda_2 x + \lambda_3 x^2 + \dots, \tag{12}$$

where the values of the parameters can be found (by identification of the coefficients of the expansions of (11) and (12)). First

$$\varepsilon = \frac{2\lambda_1\lambda_3}{\lambda_2^2 - 2\lambda_1\lambda_3},$$

then, the value of q by way of $q = \frac{\varepsilon + p}{1+\varepsilon}$ (p is given exogenously) and

$$r_0 = \lambda_1^{1-p}, \qquad a = \frac{q}{q-1}\left(\frac{\lambda_2}{\lambda_1} - 2\frac{\lambda_3}{\lambda_2}\right).$$

The problem is thus reduced to estimating the values of $\lambda_1, \lambda_2, \lambda_3$. We have carried out linear regressions of r^{1-p} on the variables x, x^2 with our rank data for values of p which make the value of λ_3 very close to zero, and so that the values of the parameters obtained make sense. This procedure has been shown to be very effective for obtaining the initial values of the parameters. Later, we carried out a simultaneous estimation by ordinary nonlinear least-squares (Bates and Watts, 1988) of the three parameters r_0, q, a of specification (5), using the initial values obtained by the earlier procedure. The estimated parameters, shown in Table 3, are very significant in all the cases. The estimates of $\hat{q}$ present a clear tendency to increase over time in all three countries; those of $\hat{a}$ have this characteristic only in Spain and Italy, and after the mid-century point. The fit is also very good, and almost always very slightly better (we are talking about differences of thousandths or even ten thousandths for Italy) than that obtained in Table 2 with the lognormal distribution.

4.2. Comparison in terms of Zipf plots

In this sub-section we compare both distributions in terms of Zipf plots, i.e., double logarithmic graphs of rank compared to population, which are used extensively in the specialised literature.

We will present these graphs in the comparison of the logarithm of the theoretical and empirical ranks for the two distributions studied, q-exponential and lognormal. In terms of Zipf plots, and in a visual analysis of them, there are various cases in which the fit by the q-exponential is apparently better than by the lognormal, or where it is not easy to discern the difference. We will show the most representative ones. Figure 1 displays the situation for the q-exponential, showing the Zipf plots of the actual data (in black) with the estimated or theoretical q-exponential (blue). Figure 2 presents the same for the lognormal. In general, the lognormal distribution is more curved than the q-exponential. This greater curvature means that, in most cases, the lognormal underestimates the empirical distribution at the upper tail of larger cities.

Beginning with the case of Italy in 1951, visually neither of the two distributions seems to give a good fit, while divergence with the theoretical lognormal appears only in the upper tail of the distribution. In other periods of the 20th century for Italy the fit of the two distributions is similar except in 2001 when the q-exponential apparently gives a better description.

However, in other cases the situation is more favourable to the q-exponential, for example in Spain in 1950. In this case the fit given by the q-exponential looks much better than that of the lognormal. For Spain, this happens for almost all of the 20th century, with 2001 being the year when they both give an apparently similar fit.

For the case of the USA, there is a gradual improvement in the description by the q-exponential until 1950 (when the fit appears to be very good) and then it worsens, while the lognormal shows a similar behaviour throughout the 20th century. The year

2000 in the USA, the same year[7] considered by Eeckhout (2004, 2009) and Levy (2009), is an example in which it is not clear which distribution is visually better: the q-exponential systematically overestimates the size of the largest cities, while the lognormal underestimates them.

In any case, it seems that the lognormal is slightly better suited to the rank of the smallest centres, confirming what was already observed in Tables 2 and 3 in reference to the estimations of r_0. Because of the lower curvature of the q-exponential, it provides a worse fit to the data at the lower tail of smaller cities.

Another general result is that, for either of the two distributions, it can be seen that discrepancies can be found in the Zipf plots between the data and the corresponding theoretical distribution, and that this tends to increase clearly and systematically with city size. It is not difficult to find a statistical explanation for this fact. Below, the quantities with overbar correspond to the empirical or sample distribution, and without overbar, to the theoretical distribution:

$$\bar{r}(x) = \bar{r}_0\left(1 - c\bar{df}(x)\right), \tag{13}$$

$$r(x) = r_0\left(1 - cdf(x)\right). \tag{14}$$

At origin both *cdf* s are null, thus $\bar{r}(0) = \bar{r}_0$ and $r(0) = r_0$. In turn, for an arbitrarily large value, infinite, of city population, both *cdf* s have to be equal to one, so that $r(\infty) = \bar{r}(\infty) = 0$.

If, as the Zipf plot demands, we take logarithms, which is important, and evaluate their difference, we obtain the following expression:

[7] Eeckhout (2004) includes the unincorporated places in his sample of cities; our results are robust to their inclusion. The appendix shows results using Eeckhout (2004)'s sample of cities (all incorporated and unincorporated places for the year 2000).

$$\ln\bar{r}(x)-\ln r(x)=\ln\bar{r}_0\left(1-c\bar{df}(x)\right)-\ln r_0\left(1-cdf(x)\right)=\ln\bar{r}_0-\ln r_0+\ln\left(1+\frac{cdf(x)-c\bar{df}(x)}{1-cdf(x)}\right) \quad (15)$$

where the last term in (15) is the fundamental one and deserves our attention. Indeed, when x is very large, the denominator $1-cdf(x)$ becomes very small, and when the discrepancy in the numerator is divided by it, $cdf(x)-c\bar{df}(x)$ is multiplied or amplified considerably, so that the contribution of the third term is due to the difference $cdf(x)-c\bar{df}(x)$ and to this multiplying effect. This is observed in most of the graphs in Figures 1 and 2, where the discrepancy increases as x does, and is an unavoidable effect, unless $cdf(x)-c\bar{df}(x)$ is practically null, something which happens for the q-exponential in the case of Spain in 1950 and USA in 1950. The previous observation, as far as we know, is new in the literature, and can contribute to clarifying questions recently posed in it (Levy, 2009; Eeckhout, 2009), which among other things, deal with the limitations of using only graphic methods to discriminate between distributions, and with the statistical implications of taking logarithms (see also Section 5 of this work).

4.3. Comparison in terms of the cumulative distribution functions

We devote this sub-section to comparing the distributions in terms of the associated *cdf* s. In principle, we would expect the results to be similar to those of the above section, but we will see that this is not exactly true. Figures 3 and 4 show the *cdf* s corresponding to the same cases in which we illustrated the Zipf plots.

It will be seen that generally, in the graphs shown, the fit in *cdf* s is apparently better for the lognormal than for the q-exponential, when, we recall, in terms of Zipf plot, the q-exponential often did better. To explain this apparent paradox it is useful to turn once again to the expressions (13) and (14). From these we deduce:

$$\overline{cdf}(x) - cdf(x) = \frac{r(x)}{r_0} - \frac{\bar{r}(x)}{\bar{r}_0}. \qquad (16)$$

We begin reasoning only for the q-exponential distribution. We know that its fit in ranks (see Section 4.2) is very good, except for the smallest cities, which means that $r(x) \cong \bar{r}(x)$ for practically all points, so that (16) is now:

$$\overline{cdf}(x) - cdf(x) = \frac{r(x)}{r_0}\left(\frac{\bar{r}_0 - r_0}{\bar{r}_0}\right) = (1 - cdf(x))\left(\frac{\bar{r}_0 - r_0}{\bar{r}_0}\right). \qquad (17)$$

It is worth studying equation (17) in detail, keeping in mind that it was obtained considering that the fit in ranks is almost perfect. The *cdf* s fit less well as the difference $\bar{r}_0 - r_0$ increases, and this gap is not negligible in the q-exponential, as can be confirmed by the information given in Table 3 (remember that $\bar{r}_0$ is identified with the sample size). Also, the discrepancy in *cdf* s increases with $1 - cdf(x)$, i.e., it increases as x decreases, and tends to disappear gradually as x increases. All this shows that in the q-exponential the discrepancy in *cdf* s is perfectly compatible with an almost perfect rank fit, except for the smallest cities; moreover, it is unavoidable if, as in reality, $\bar{r}_0 - r_0 \neq 0$. However, the fit by the q-exponential improves for the three countries as the 20th century goes on[8]. This is due to the reasons adduced and to the estimate of r_0 improving over time.

Is a similar situation produced for the lognormal? We will see it is not. We can deduce from Table 2 that now $r_0 \cong \bar{r}_0$, so that (16) is reduced to:

$$\overline{cdf}(x) - cdf(x) = \frac{1}{r_0}(r(x) - \bar{r}(x)). \qquad (18)$$

[8] More details available from the authors upon request.

So that based on (18) we derive that any lack of fit in ranks is directly transferred, in the lognormal, to a lack of fit in *cdf* s. This intuitive result is not found in the q-exponential for the reasons adduced above. Where the lognormal performs well in Zipf plots (at the lower tail) is more clearly reflected in the *cdf* s plots, whereas where it performs badly (at the upper tail) is less clearly reflected, since even the largest city is only one of all the cities (for example, New York is only one city in the sample of 20,000 cities in the USA).

4.4. Standard statistical tests

The parameters estimated in Section 4.1 for the lognormal and the q-exponential are worth an independent statistical test to verify the goodness of fit. In the above sub-sections we talked about visual criteria (Zipf plots and *cdf* s) in order to discriminate between distributions, which always involves a certain degree of subjectivity. Also, the graphic approximation using Zipf plots in Section 4.2 might not give us reliable information about the fit: problems with Zipf plots have been shown in the literature (Eeckhout, 2009). For this reason, in this sub-section we present statistical tests to compare distributions more objectively. For the case of the lognormal the Kolmogorov-Smirnov test is standard, but with a number of observations as high as in our data, this test tends systematically to reject the null hypothesis of lognormality unless the fit is perfect. On the other hand, in the case of the q-exponential or generalised Pareto a test of this type is not so standard. Consequently, we are looking for a test which does not tend to reject the null hypothesis merely because there is a high number of observations, and which is equally applicable to both distributions.

The Wilcoxon rank-sum test meets both requirements. It tests the hypothesis that two independent samples come from populations with the same distribution. Its result is

not so dependent on the high sample size, and it is the same test regardless of the underlying distribution. We carried out the test comparing the empirical and estimated ranks in each case. We show the results of these tests in Tables 4 and 5.

The results of the Wilcoxon test show that the null hypothesis of the q-exponential cannot be rejected with a 5% confidence level in any of the periods of the 20th century in Spain, Italy and the USA. Neither can lognormality be rejected with a 5% confidence level in any of the periods of the 20th century in Spain and Italy. In the USA a temporal evolution can be seen; the first decades reject the lognormal and the p-value decreases over time, but from 1930 the p-value begins to increase, until the lognormal distribution cannot be rejected at 5% from 1960 onwards. If we take a confidence level of 1%, instead of 5%, the null hypothesis would be rejected in the USA only for 1920 and 1930. These results are also obtained in González-Val (2010b) using kernels.

5. Lognormal and q-exponential. The complementarity approach

In the above section we have compared the suitability of q-exponential and lognormal distributions for our data, studying Zipf plots, *cdf* s and the Wilcoxon test. The main result which we can present now is that neither of the two is clearly better than the other, and that, taking into account the different criteria presented, either the lognormal is preferred (*cdf* s) or the q-exponential is (statistical tests and sometimes, Zipf plots). In any case, the focus of the above section was on comparing both distributions in different ways as substitutes for each other. In this section we change the focus and will try to show that rather than being rivals, they complement each other.

Therefore, we will now examine this complementarity. First, we see that one of the main results of Eeckhout (2004) on the concavity of the lognormal Zipf plot can also

be applied in the case of the q-exponential. Indeed, in this case the expression of rank regarding population is given by (5). To express variables in logarithms we call $y = \ln x$, so that $x = e^y$. Thus:

$$\ln r(y) = \ln r_0 \left(1 + \frac{q-1}{q} a\, e^y\right)^{\frac{1}{1-q}} = \ln r_0 + \frac{1}{1-q} \ln\left(1 + \frac{q-1}{q} a\, e^y\right). \quad (19)$$

So we have:

$$\frac{d}{dy} \ln r(y) = -\frac{a\, e^y}{a\, e^y (q-1) + q} = -h(y), \quad (20)$$

which is the negative of the hazard rate $h(y)$. Obtaining the second derivative, we have:

$$\frac{d^2}{dy^2} \ln r(y) = -\frac{a\, e^y q}{\left(a\, e^y (q-1) + q\right)^2} = -h'(y), \quad (21)$$

which is strictly negative for all y, so that $\ln r(y)$ is strictly concave for all the values of the variable and the corresponding Zipf plot is concave in turn[9]. In short, this is a theoretical result which affirms that if the process generating the underlying data follows a q-exponential (or a lognormal as shown by Eeckhout, 2004) the Pareto exponent decreases with the sample size. In other words, it is the first proof of the non-rivalry or equivalence, at least to this respect, of both distributions.

A key point which arises when studying the q-exponential and the lognormal is that if we consider city populations in absolute levels, the empirical distribution of probability is a decreasing and convex curve like the density function of the q-exponential, while if the logarithm of population is taken, the empirical distribution of

[9] Similar expressions to (19), (20), (21) for the case of the lognormal can be found on page 1442 of Eeckhout (2004).

probability is a bell curve, like the density function of the lognormal. This result is somewhat counterintuitive and merits closer examination.

Let us suppose that we order the urban centres from our data from smaller to greater populations. A histogram of these creates a decreasing graph as the population rises (Figure 5, data from Spain in 1900). As is well known, a histogram values the frequencies associated with intervals of a constant width on the x-axis. However, in a histogram of the population logarithm (Figure 6, same data) these are also counted in frequencies according to intervals of constant width but now in logarithms; but what does this mean in levels? Let δ be this constant width, and the lower and upper ends of one of these intervals be $\ln x_j$ and $\ln x_{j+1}$ respectively. By definition, $\ln x_{j+1} - \ln x_j = \delta$ or, to put it another way, $x_{j+1} = x_j e^{\delta}$. Generalising, $x_{j+1} = x_j e^{\delta} = x_{j-1} e^{2\delta} = x_1 e^{j\delta}$, where x_1 is the lower end of the first interval, which cannot be zero. This indicates that the upper ends of the intervals, in levels, follow a geometric progression of e^{δ}. It should be underlined that this reasoning is valid for any numerical variable which is used alternatively in levels and in natural logarithms.

This fact explains why taking logarithms gives a bell curve: the first intervals are very narrow and there are also very few cases included in them; then, as the intervals widen according to the geometric progression, the number of cases in each interval grows considerably, and the graph rises. There will come a moment when, although the intervals are very wide, the number of cases will be very small, for obvious reasons (for example, very large cities, of more than, let us say, 500,000 inhabitants), so that the graph decreases. The process has arrived at a maximum and it is obtained a bell curve.

These two results show that the q-exponential and the lognormal are complementary, rather than being substitutes for each other. Both give a remarkably

close fit to the data, and both improve the description of the Pareto distribution when all population centres are considered.

6. Conclusions

City size distribution has been the subject of numerous empirical investigations by urban economists, statistical physicists, and urban geographers. From the point of view of urban economics, the study of city size distribution has deep economic implications. For example, an urban structure of cities of very similar population invites an egalitarian treatment by the public bodies in charge of investment in transport infrastructure, education or healthcare. However, large differences in size require policies which tend towards convergence and strive for territorial cohesion.

This work has minutely examined a density function which as far as we know is new to Urban Economics, the q-exponential. A first contribution involves the estimation method of its key parameters. Indeed, rather than taking predetermined values for them, they are approximated by a Taylor series expansion enabling us to obtain initial values for later estimation, without restrictions or a priori assumptions, by nonlinear least-squares.

Elsewhere, since the pioneering work of Eeckhout (2004) the risks have been demonstrated of considering only the largest centres, i.e., only the upper tail. In turn, if the data permit, the analysis of city size distribution should be done as a long-term analysis. With both considerations as premises, this article uses census data for the entire 20th century, in decades, and all the cities of three countries: the USA (from 10,600 to 19,300 centres, according to year), Spain (about 8,000 centres) and Italy (about 8,000 centres). Using such large databases and such a vast temporal horizon undoubtedly adds robustness to the results.

What are the main results? All of them relate the distribution which since Eeckhout (2004) has been postulated as the best for studying city size distributions without a truncation point, the lognormal, with the one we present in this paper, the q-exponential. There are basically three.

First, the fit of both (lognormal and q-exponential) in terms of ranks in a Zipf plot is extremely good, although very slightly better for the latter. However, it is statistically demonstrated that this better fit in ranks of the q-exponential unavoidably means that, in terms of cumulative density functions, the lognormal is better.

Second, the Wilcoxon test shows that the null hypothesis of the q-exponential cannot be rejected with a 5% confidence level in any of the periods of the 20th century in Spain, Italy and the USA. Neither can lognormality be rejected at 5% in any of the periods of the 20th century in Spain and Italy; in the USA, only from 1960 onwards. Then, in standard statistical tests, for our sample, both distributions work well, especially the q-exponential.

And third, the relevant question which might be asked by an urban planner, economist or geographer studying the complete city size distribution of a given area or country is: does this work recommend using the lognormal or the q-exponential? The answer is simple. From our point of view they are complementary: if working with populations in levels the q-exponential is better; if, as is usual, the size logarithm is taken, both are suitable, and the q-exponential may be very slightly better; but it is also fair to say that the lognormal offers good features and two additional advantages: historically it has been much more used in the literature (not only in Urban Economics but in any discipline) and the method of estimating its parameters is simpler, conceptually and computationally, than the q-exponential. In any case, it is a question

which each researcher must answer for himself and on which this work has tried to shed some light.

Finally, we could not end this work without a brief reflection on the theoretical growth process underlying the q-exponential distribution. The links between the lognormal and the Pareto distribution and Gibrat's law are well known, and there is even a generalised version, the double Pareto lognormal (Reed, 2002). What is behind the q-exponential? This is an important question which needs to be answered, and thus constitutes an excellent field for future research.

Acknowledgements

The authors would like to thank the Spanish Ministerio de Educación y Ciencia (SEJ2006-04893/ECON and ECO2009-09332 projects), the DGA (ADETRE research group) and FEDER for their financial support. An earlier version of this paper has been presented at the XIII Encuentro de Economía Aplicada (Seville, 2010) with all the comments made by the participants being highly appreciated.

References

Anderson, G. and Y. Ge (2005). "The size distribution of Chinese cities," Regional Science and Urban Economics 35, 756-776.

Bates, D. M. and D. G. Watts (1988). "Nonlinear regression analysis and its applications," New York: Wiley.

Black, D. and J. V. Henderson (2003). "Urban evolution in the USA," Journal of Economic Geography 3, 343-372.

Bosker, M., S. Brakman, H. Garretsen and M. Schramm (2008). "A century of shocks: the evolution of the German city size distribution 1925-1999," Regional Science and Urban Economics 38, 330-347.

Cameron, T. A. (1990). "One-stage structural models to explain city size," Journal of Urban Economics, 27(3): 294-207.

Cheshire, P. (1999). "Trends in sizes and structure of urban areas," in Handbook of Regional and Urban Economics, Vol. 3, P. Cheshire and E. S. Mills, (eds.) Amsterdam: Elsevier Science, Chapter 35, 1339-1373.

Cheshire, P. C. and S. Magrini (2006). "Population growth in European cities: weather matters-but only nationally," Regional Studies 40(1), 23-37.

Choulakian, V. and M. A. Stephens (2001). "Goodness-of-fit tests for the generalized Pareto distribution," Technometrics 43, 478-484.

Eeckhout, J. (2004). "Gibrat's Law for (all) cities," American Economic Review 94(5), 1429-1451.

Eeckhout, J. (2009). "Gibrat's Law for (all) cities: reply," American Economic Review 99(4), 1676-1683.

Gabaix, X. (1999). "Zipf's law for cities: An explanation," Quarterly Journal of Economics, 114(3):739-767.

Gabaix, X. and Y. M. Ioannides (2004). "The evolution of city size distributions," in Handbook of urban and regional economics, Vol. 4, J. V. Henderson and J. F. Thisse, (eds.) Amsterdam: Elsevier Science, North-Holland.

Garmestani, A. S., C. R. Allen and C. M. Gallagher (2008). "Power laws, discontinuities and regional city size distributions," Journal of Economic Behavior & Organization, 68: 209–216.

Garmestani, A. S., C. R. Allen, C. M. Gallagher and J. D. Mittelstaedt (2007). "Departures from Gibrat's law, discontinuities and city size distributions," Urban Studies, 44(10): 1997–2007.

Gibrat, R. (1931). "Les inégalités économiques," Paris: Librairie du recueil Sirey.

Giesen, K., A. Zimmermann and J. Suedekum (2010). "The size distribution across all cities – double Pareto lognormal strikes," Journal of Urban Economics, 68: 129-137.

González-Val, R. (2010a). "Deviations from Zipf's law for American cities: an empirical examination," Urban Studies, forthcoming. DOI: 10.1177/0042098010371394

González-Val, R. (2010b). "The evolution of US city size distribution from a long term perspective (1900-2000)," Journal of Regional Science, forthcoming. DOI 10.1111/j.1467-9787.2010.00685.x

González-Val, R., L. Lanaspa and F. Sanz (2010). "Gibrat's Law for cities revisited," Mimeo, Universidad de Zaragoza.

Grimshaw, S. D. (1993). "Computing maximum likelihood estimates for the generalized Pareto distribution," Technometrics 35, 185-191.

Hosking, J. R. M. and J. R. Wallis (1987). "Parameter and quantile estimation for the generalized Pareto distribution," Technometrics 29, 339-349.

Hsing, Y. (1990). "A note on functional forms and the urban size distribution," Journal of Urban Economics, 27(1): 73-79.

Ioannides, Y. M. and H. G. Overman (2003). "Zipf's law for cities: an empirical examination," Regional Science and Urban Economics 33, 127-137.

Ioannides, Y. M. and S. Skouras (2009). "Gibrat's Law for (all) cities: a rejoinder," Economics Department Working Paper, Tufts University.

Kalecki, M. (1945). “On the Gibrat distribution,” Econometrica 13(2), 161-170.

Kamecke, U. (1990). “Testing the rank size rule hypothesis with an efficient estimator,” Journal of Urban Economics, 27(2): 222-231.

Krugman, P. R. (1996a). “Confronting the mystery of urban hierarchy,” Journal of the Japanese and the International Economies 10, 399-418.

Krugman, P. R. (1996b). “The self-organizing economy,” Oxford: Blackwell.

Levy, M. (2009). “Gibrat’s Law for (all) cities: a comment,” American Economic Review 99(4), 1672-1675.

Malacarne, L. C., R. S. Mendes and E. K. Lenzi (2001). “q-exponential distribution in urban agglomeration,” Physical Review E 65, (017106) 1-3.

Michaels, G., F. Rauch and S. J. Redding (2010). “Urbanization and Structural Transformation,” unpublished manuscript, London School of Economics.

Reed, W. (2002). “On the rank-size distribution for human settlements,” Journal of Regional Science 42, 1-17.

Rosen, K. and M. Resnick (1980). “The size distribution of cities: an examination of the Pareto law and primacy,” Journal of Urban Economics 8, 165-186.

Rozenfeld, H. D., D. Rybski, J. S. Andrade, Jr., M. Batty, H. E. Stanley and H. A. Maksea (2008). “Laws of population growth,” Proceedings of the National Academy of Sciences, 105(48): 18702–18707.

Sharma, S. (2003). “Persistence and stability in city growth,” Journal of Urban Economics 53, 300-320.

Soo, K. T. (2005). “Zipf’s Law for cities: a cross-country investigation,” Regional Science and Urban Economics 35, 239-263.

Soo, K. T. (2007). “Zipf's Law and urban growth in Malaysia,” Urban Studies 44(1), 1-14.

Tsallis, C. (1988). “Possible generalization of Boltzmann-Gibbs statistics,” Journal of Statistical Physics 52, 479-487.

Zipf, G. K. (1949). “Human Behaviour and the Principle of Least Effort,” Cambridge, MA:Addison-Wesley.

Table 1. Number of cities and descriptive statistics

US					
Year	Cities	Mean	Standard deviation	Minimum	Maximum
1900	10,596	3,376.04	42,323.90	7	3,437,202
1910	14,135	3,560.92	49,351.24	4	4,766,883
1920	15,481	4,014.81	56,781.65	3	5,620,048
1930	16,475	4,642.02	67,853.65	1	6,930,446
1940	16,729	4,975.67	71,299.37	1	7,454,995
1950	17,113	5,613.42	76,064.40	1	7,891,957
1960	18,051	6,408.75	74,737.62	1	7,781,984
1970	18,488	7,094.29	75,319.59	3	7,894,862
1980	18,923	7,395.64	69,167.91	2	7,071,639
1990	19,120	7,977.63	71,873.91	2	7,322,564
2000	19,296	8,968.44	78,014.75	1	8,008,278
SPAIN					
Year	Cities	Mean	Standard deviation	Minimum	Maximum
1900	7,800	2,282.40	10,177.75	78	539,835
1910	7,806	2,452.01	11,217.02	92	599,807
1920	7,812	2,621.92	13,501.02	82	750,896
1930	7,875	2,892.18	17,513.90	79	1,005,565
1940	7,896	3,180.65	20,099.96	11	1,088,647
1950	7,901	3,479.86	26,033.29	64	1,618,435
1960	7,910	3,801.71	33,652.11	51	2,259,931
1970	7,956	4,240.98	43,971.93	10	3,146,071
1981	8,034	4,701.40	45,995.35	5	3,188,297
1991	8,077	4,882.27	45,219.85	2	3,084,673
2001	8,077	5,039.37	43,079.46	7	2,938,723
ITALY					
Year	Cities	Mean	Standard deviation	Minimum	Maximum
1901	7,711	4,274.84	14,424.61	56	621,213
1911	7,711	4,648.11	17,392.98	58	751,211
1921	8,100	4,863.80	20,031.61	58	859,629
1931	8,100	5,067.10	22,559.85	93	960,660
1936	8,100	5,234.38	25,274.48	116	1,150,338
1951	8,100	5,866.12	31,137.52	74	1,651,393
1961	8,100	6,249.82	39,130.55	90	2,187,682
1971	8,100	6,683.52	45,581.66	51	2,781,385
1981	8,100	6,982.33	45,329.33	32	2,839,638
1991	8,100	7,009.63	42,450.26	31	2,775,250
2001	8,100	7,021.20	39,325.47	33	2,546,804

Table 2. Values of $\hat{r}_0$, $\hat{\mu}$ and $\hat{\sigma}^2$ for USA. Standard errors in parenthesis. Lognormal

USA-lognormal	$\hat{r}_0$	$\hat{\mu}$	$\hat{\sigma}^2$	R^2
1900	10374.27	6.648714	1.261147	0.9954
	(6.871518)	(0.0122517)	(0.0086632)	
1910	13805.79	6.64682	1.292809	0.9953
	(8.005956)	(0.0108739)	(0.007689)	
1920	15126.01	6.674667	1.318693	0.9951
	(8.510406)	(0.0105985)	(0.0074943)	
1930	16104.56	6.692269	1.401552	0.9949
	(8.988392)	(0.0109193)	(0.0077211)	
1940	16347.44	6.775817	1.431982	0.9954
	(8.587507)	(0.0110714)	(0.0078287)	
1950	16771.11	6.837732	1.501686	0.9968
	(7.3058)	(0.0114793)	(0.0081171)	
1960	17698.55	6.923707	1.605794	0.9975
	(6.640407)	(0.011952)	(0.0084513)	
1970	18153.87	7.004047	1.666934	0.9979
	(6.081665)	(0.0122595)	(0.0086688)	
1980	18576.96	7.114369	1.661872	0.9981
	(5.860968)	(0.012081)	(0.0085425)	
1990	18799.24	7.0984	1.742746	0.9985
	(5.308348)	(0.0126035)	(0.008912)	
2000	18968.83	7.18272	1.782151	0.9986
	(5.17997)	(0.0128295)	(0.0090718)	

Table 2 (continued). Values of $\hat{r}_0, \hat{\mu}$ and $\hat{\sigma}^2$ for Spain. Standard errors in parenthesis. Lognormal

Spain-lognormal	$\hat{r}_0$	$\hat{\mu}$	$\hat{\sigma}^2$	R^2
1900	7610.979	6.96552	1.062761	0.9975
	(4.281377)	(0.0120334)	(0.0085089)	
1910	7614.829	7.012887	1.079165	0.9975
	(4.280091)	(0.0122144)	(0.0086369)	
1920	7621.96	7.025287	1.107306	0.9977
	(4.174949)	(0.0125281)	(0.0088587)	
1930	7684.351	7.05515	1.142478	0.9977
	(4.13726)	(0.0128743)	(0.0091035)	
1940	7706.974	7.062808	1.181922	0.9975
	(4.320335)	(0.013301)	(0.0094052)	
1950	7711.78	7.086039	1.202717	0.9977
	(4.173386)	(0.0135308)	(0.0095677)	
1960	7717.208	7.033353	1.272214	0.9976
	(4.219623)	(0.0143045)	(0.0101148)	
1970	7791.69	6.82857	1.442019	0.9982
	(3.71842)	(0.0161668)	(0.0114316)	
1981	7889.339	6.631256	1.623441	0.9985
	(3.379101)	(0.0181122)	(0.0128072)	
1991	7916.954	6.534098	1.714716	0.9985
	(3.454906)	(0.0190795)	(0.0134913)	
2001	7892.812	6.540983	1.754564	0.9982
	(3.772936)	(0.0195229)	(0.0138048)	

Table 2 (continued). Values of $\hat{r}_0, \hat{\mu}$ and $\hat{\sigma}^2$ for Italy. Standard errors in parenthesis. Lognormal

Italy-lognormal	$\hat{r}_0$	$\hat{\mu}$	$\hat{\sigma}^2$	R^2
1901	7676.142	7.78953	0.9154127	0.999
	(2.753235)	(0.0104247)	(0.0073713)	
1911	7672.715	7.843163	0.9313805	0.9991
	(2.648972)	(0.0106065)	(0.0074999)	
1921	8050.636	7.835906	0.9628311	0.9992
	(2.598348)	(0.0106981)	(0.0075647)	
1931	8054.081	7.838977	0.9912791	0.9993
	(2.38133)	(0.0110142)	(0.0077882)	
1936	8062.379	7.84206	1.009614	0.9994
	(2.237987)	(0.0112179)	(0.0079323)	
1951	8067.352	7.894767	1.049305	0.9994
	(2.147735)	(0.0116589)	(0.0082441)	
1961	8056.921	7.84784	1.100507	0.9992
	(2.463067)	(0.0122279)	(0.0086464)	
1971	8041.767	7.788053	1.185047	0.9992
	(2.563667)	(0.0131672)	(0.0093106)	
1981	8045.297	7.792515	1.245428	0.9994
	(2.102519)	(0.0138381)	(0.009785)	
1991	8057.702	7.795891	1.282116	0.9996
	(1.721827)	(0.0142457)	(0.0100733)	
2001	8068.104	7.803148	1.306614	0.9997
	(1.473446)	(0.0145179)	(0.0102657)	

Table 3. Values of $\hat{r}_0, \hat{q}$ and $\hat{a}$ for USA. Standard errors in parenthesis. q-exponential

USA q-exponential	$\hat{r}_0$	$\hat{q}$	$\hat{a}$	R^2
1900	12658.84	1.603419	0.0027559	0.999
	(12.37365)	(0.0032688)	(0.0000141)	
1910	16939	1.694559	0.0031086	0.9991
	(14.08978)	(0.0027492)	(0.0000133)	
1920	18457.28	1.709492	0.0030609	0.9991
	(14.15269)	(0.0025757)	(0.0000122)	
1930	19340.12	1.818757	0.003332	0.999
	(14.61102)	(0.0026694)	(0.0000133)	
1940	19454.64	1.8933	0.0032611	0.9991
	(13.47561)	(0.0025121)	(0.0000121)	
1950	19241.77	1.975214	0.003091	0.9995
	(9.194967)	(0.0019319)	(8.51e-06)	
1960	19785.46	2.144436	0.0031838	0.9997
	(6.949107)	(0.0015544)	(6.79e-06)	
1970	19911.43	2.221329	0.0030247	0.9998
	(5.661598)	(0.0013473)	(5.45e-06)	
1980	20338.61	2.228124	0.0027042	0.9998
	(5.206677)	(0.00122)	(4.41e-06)	
1990	20191.54	2.325746	0.0028769	0.9999
	(4.104436)	(0.001043)	(3.98e-06)	
2000	20264.44	2.393507	0.0027802	0.9999
	(3.904862)	(0.0010135)	(3.62e-06)	

Table 3 (continued). Values of $\hat{r}_0, \hat{q}$ and $\hat{a}$ for Spain. Standard errors in parenthesis. q-exponential

Spain q-exponential	$\hat{r}_0$	$\hat{q}$	$\hat{a}$	R^2
1900	9768.342	1.539363	0.0019318	0.9995
	(8.919785)	(0.0025706)	(8.32e-06)	
1910	9737.746	1.569519	0.0019022	0.9995
	(8.536555)	(0.0024854)	(7.89e-06)	
1920	9642.943	1.606816	0.0019348	0.9996
	(7.541392)	(0.0022763)	(7.27e-06)	
1930	9599.185	1.658583	0.0019577	0.9996
	(7.225424)	(0.0022592)	(7.20e-06)	
1940	9545.302	1.698713	0.0020182	0.9996
	(7.722284)	(0.0024931)	(8.09e-06)	
1950	9479.403	1.729926	0.0020183	0.9997
	(6.729436)	(0.0022309)	(7.19e-06)	
1960	9344.968	1.834201	0.0023419	0.9997
	(6.30277)	(0.0022176)	(8.12e-06)	
1970	8881.374	1.99013	0.0030433	0.9998
	(4.450282)	(0.0020147)	(8.84e-06)	
1981	8627.245	2.192375	0.0041819	0.9998
	(3.438239)	(0.0018711)	(0.0000106)	
1991	8607.263	2.359453	0.0053741	0.9998
	(3.54808)	(0.002003)	(0.0000144)	
2001	8644.453	2.478272	0.0061204	0.9997
	(4.134219)	(0.0022863)	(0.0000188)	

Table 3 (continued). Values of $\hat{r}_0, \hat{q}$ and $\hat{a}$ for Italy. Standard errors in parenthesis. q-exponential

Italy q-exponential	$\hat{r}_0$	$\hat{q}$	$\hat{a}$	R^2
1901	9217.565	0.9955913	0.0003635	0.9989
	(8.965359)	(0.003305)	(2.19e-06)	
1911	9128.744	1.007638	0.0003498	0.9991
	(8.35108)	(0.0031358)	(1.98e-06)	
1921	9522.924	1.068256	0.0003819	0.9993
	(7.643256)	(0.0027875)	(1.88e-06)	
1931	9404.778	1.108075	0.0003946	0.9994
	(6.948147)	(0.0026432)	(1.80e-06)	
1936	9313.638	1.118255	0.0003923	0.9994
	(6.394332)	(0.0025241)	(1.69e-06)	
1951	9174.556	1.52508	0.0003806	0.9995
	(5.479968)	(0.0022996)	(1.46e-06)	
1961	9114.941	1.224953	0.0004378	0.9995
	(5.332345)	(0.0023121)	(1.64e-06)	
1971	9008.062	1.369771	0.000551	0.9996
	(4.735081)	(0.0021592)	(1.84e-06)	
1981	8863.086	1.470554	0.0005982	0.9998
	(3.715982)	(0.0018076)	(1.62e-06)	
1991	8751.159	1.515978	0.000609	0.9998
	(3.110422)	(0.0015959)	(1.43e-06)	
2001	8683.71	1.548455	0.0006143	0.9998
	(2.853292)	(0.0015113)	(1.34e-06)	

Note: By definition $\hat{q} > 1$. The point estimate for Italy in 1901 does not satisfy this restriction by thousandths, although the condition would be met if considering the estimation by interval.

Table 4. Wilcoxon (Rank-sum test) test for the q-exponential

USA											
Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
p-value	0.4662	0.4519	0.4360	0.4132	0.4718	0.619	0.7635	0.845	0.8948	0.9829	0.9620
SPAIN											
Year	1900	1910	1920	1930	1940	1950	1960	1970	1981	1991	2001
p-value	0.7673	0.7845	0.8252	0.8588	0.8372	0.8817	0.9123	0.9448	0.9973	0.9289	0.8912
ITALY											
Year	1901	1911	1921	1931	1936	1951	1961	1971	1981	1991	2001
p-value	0.9077	0.529	0.5615	0.6093	0.6242	0.6619	0.6587	0.7049	0.7869	0.8245	0.8536

Ho: The city distribution follows a q-exponential

Table 5. Wilcoxon (Rank-sum test) test for the lognormal

USA											
Year	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
p-value	0.0252	0.017	0.0078	0.0088	0.0208	0.0464	0.1281	0.1836	0.2538	0.323	0.4168
SPAIN											
Year	1900	1910	1920	1930	1940	1950	1960	1970	1981	1991	2001
p-value	0.5953	0.6144	0.6233	0.6525	0.4909	0.5792	0.6049	0.522	0.5176	0.622	0.7212
ITALY											
Year	1901	1911	1921	1931	1936	1951	1961	1971	1981	1991	2001
p-value	0.2081	0.2205	0.2352	0.291	0.2864	0.3118	0.2589	0.272	0.382	0.4671	0.5287

Ho: The city distribution follows a lognormal

Figure 1. Empirical and theoretical Zipf plots with the q-exponential.

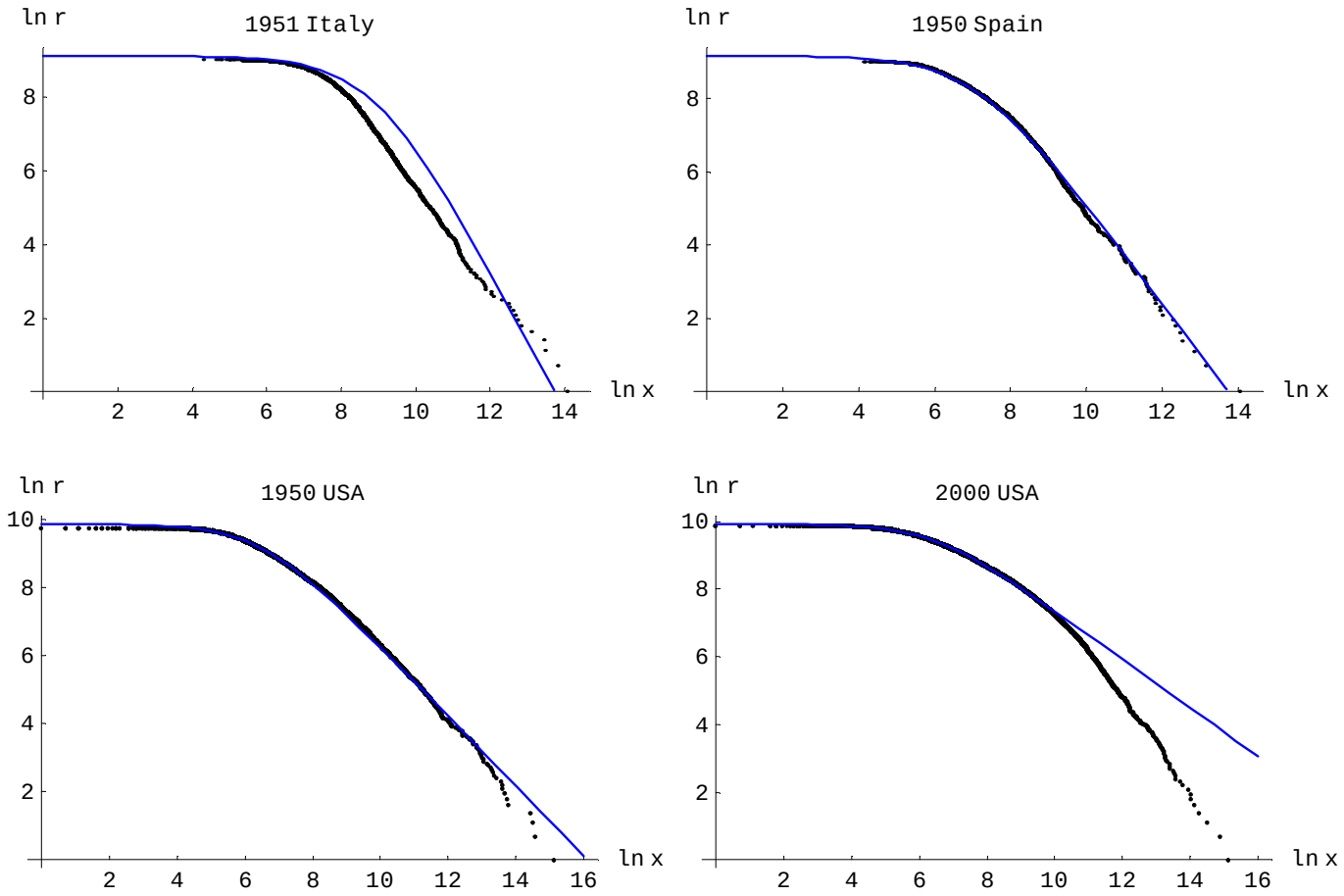


Figure 2. Empirical and theoretical Zipf plots with the lognormal.

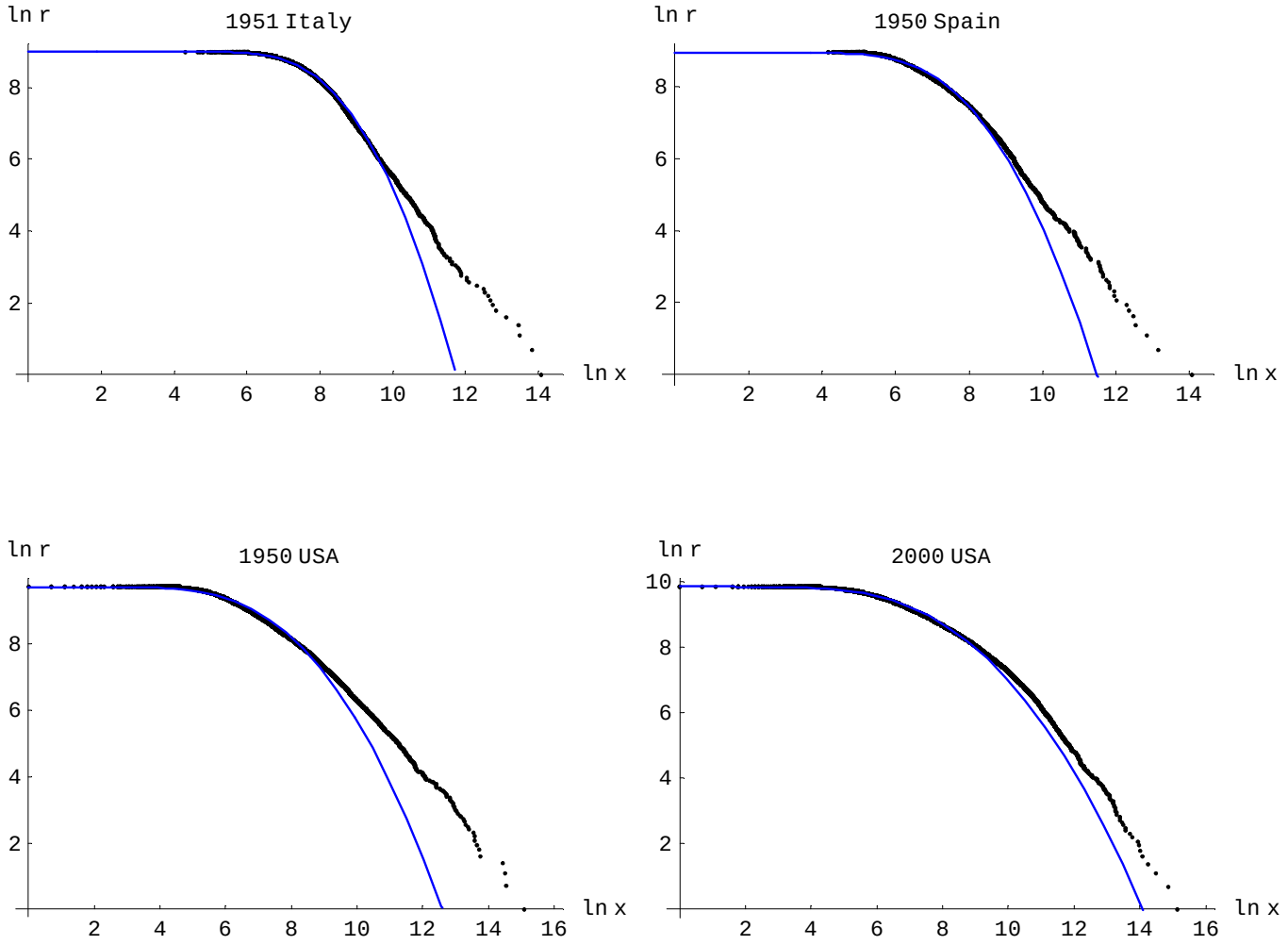


Figure 3. Empirical and theoretical *cdf* plots with the q-exponential.

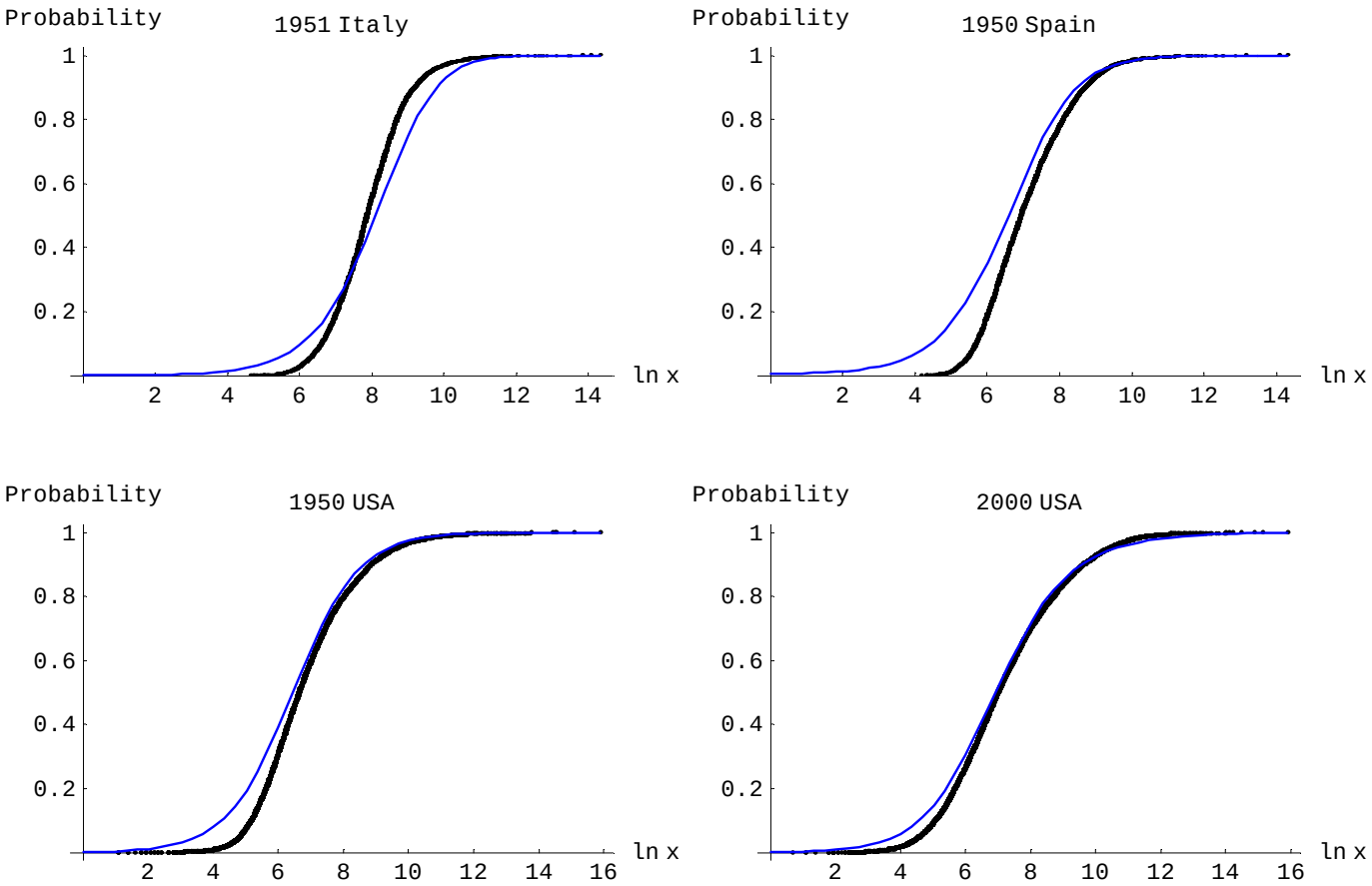


Figure 4. Empirical and theoretical *cdf* plots with the lognormal.

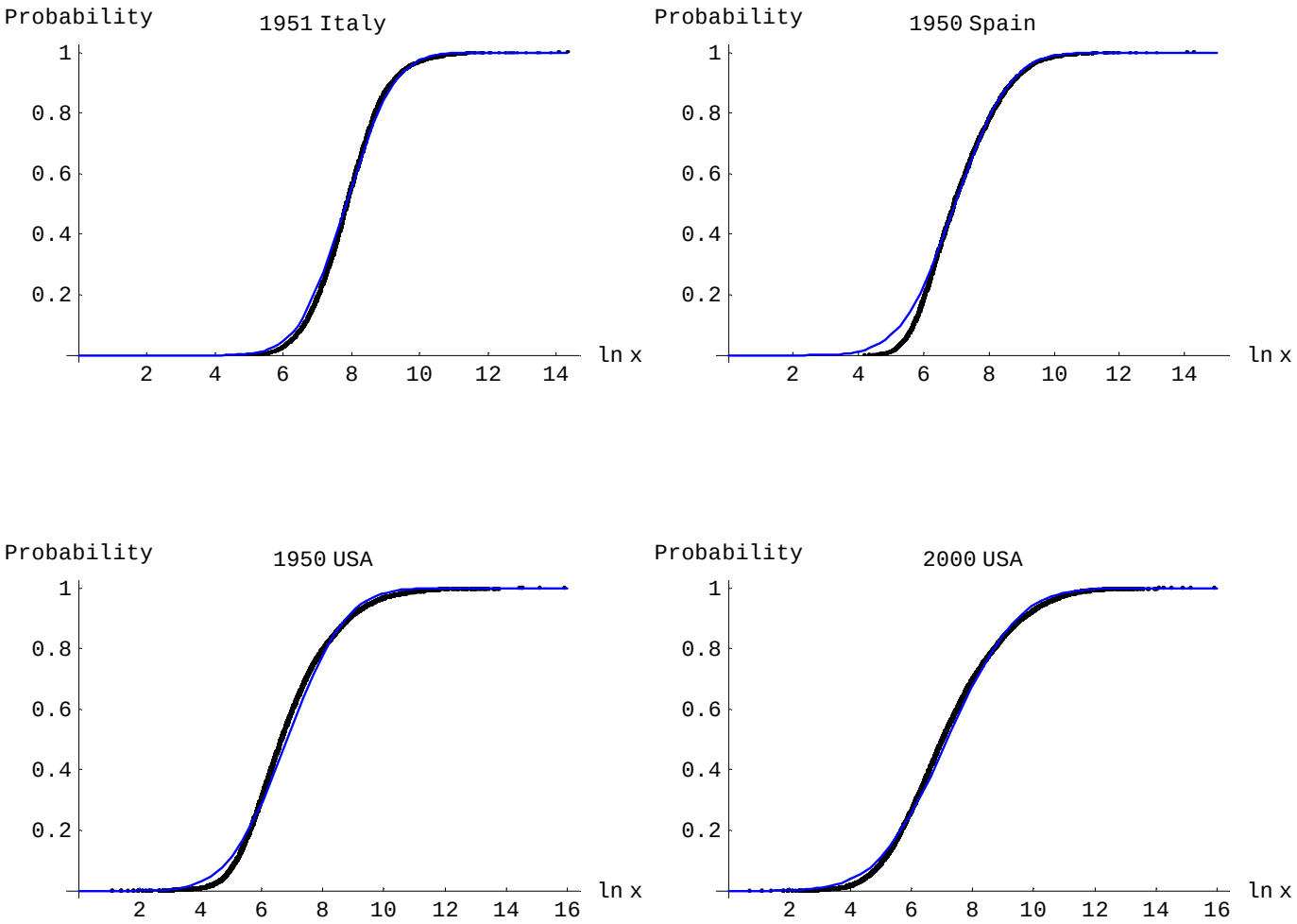


Figure 5. Histogram of Spanish cities in 1900. Population in levels

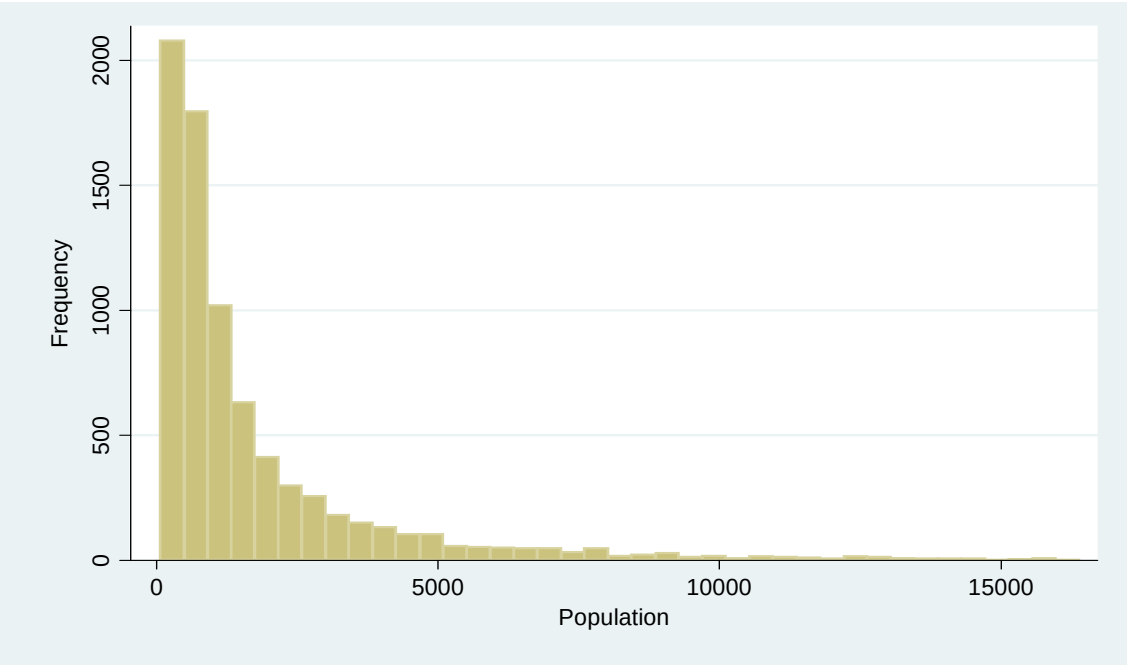
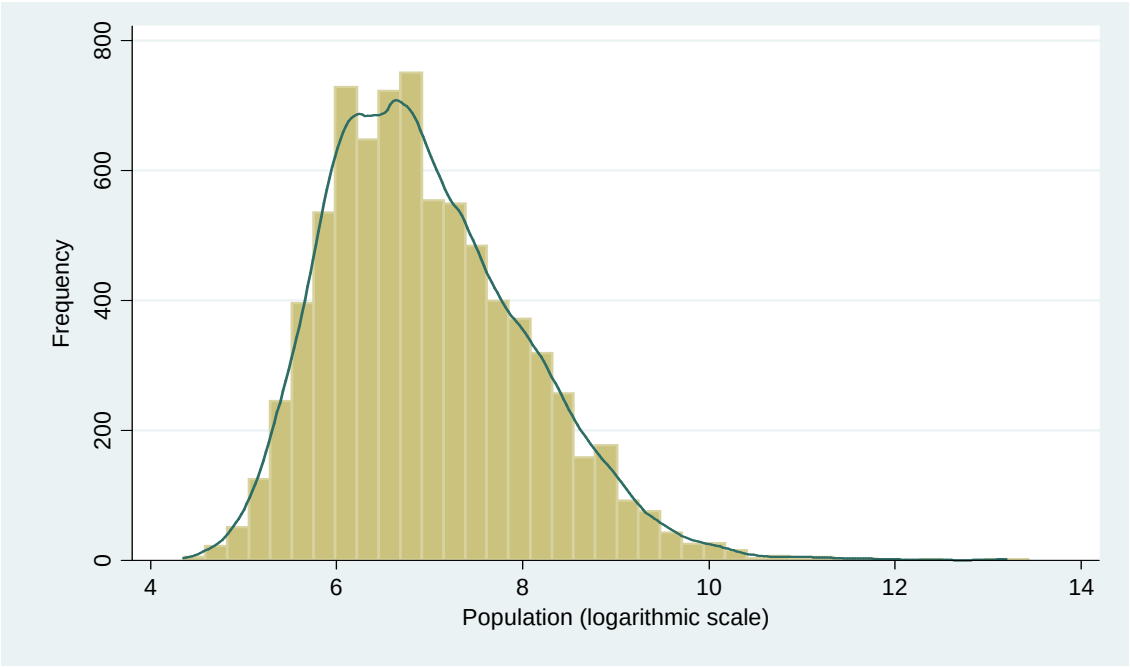
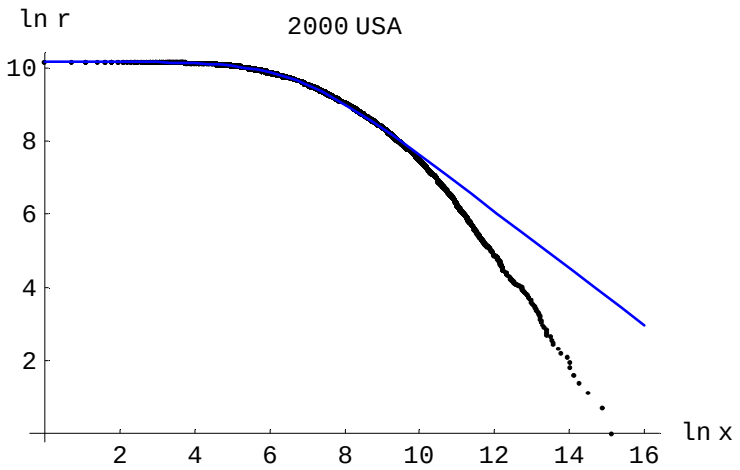


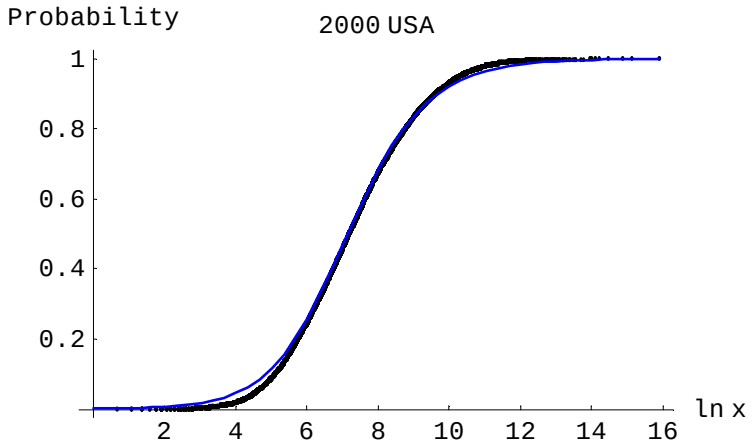
Figure 6. Histogram of Spanish cities in 1900. Population in logarithms



Appendix. Results using Eeckhout (2004)'s sample of cities (25,359 places, all incorporated and unincorporated places for the year 2000).

q -exponential





lognormal