The accuracy of graphs to describe size distributions

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

Universidad de Zaragoza Institut d'Economia de Barcelona (IEB), Universidad de Zaragoza

July 2013

The accuracy of graphs to describe size distributions

Rafael González-Val[a]

Arturo Ramos[b]

Fernando Sanz-Gracia[b]

[a] Departamento de Análisis Económico (Universidad de Zaragoza) & Institut d'Economia de Barcelona (IEB). Facultad de Economía y Empresa. Gran Vía 2, 50005 Zaragoza (Spain). E-mail: rafaelg@unizar.es

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

Abstract

This paper analyses the performance of the graphs traditionally used to study size distributions: histograms, Zipf plots (double logarithmic graphs of rank compared to size) and plotted cumulative density functions. A lognormal distribution is fitted to urban data from three countries (the US, Spain and Italy) over all of the 20th century. We explain the advantages and disadvantages associated with these graphic methods and derive some statistical properties.

Keywords: city size distribution, Zipf plot, lognormal

JEL: C16, R00

1. Introduction

Size distributions are used in economics to study many economic entities (firms, mutual funds, stocks, cities, etc.). Most of the studies use graphical tools as an aproximation of the real behaviour of the distribution. In this paper, we examine the accuracy of the graphs traditionally used to describe size distributions: we study the performance of histograms, Zipf plots and plotted cumulative density functions.

In our empirical application we consider city size data from three countries: Spain, Italy and the United States. From the point of view of urban economics, the study of city size distribution has a long tradition and deep economic implications related to labour markets, income distribution or public expenditure. To illustrate the performance of the traditional graphs, we must fit a statistical distribution to the data. We choose the lognormal distribution, widely applied in urban economics (Eeckhout, 2004; Giesen et al., 2010; González-Val et al., 2013a) and in other fields of economics. Nevertheless, the discussion carried out in Section 4 is valid for any other distribution apart from the lognormal, which has the additional advantage of being easy to handle with it.

The paper is organised as follows. Section 2 introduces the databases we use, Section 3 describes the estimation method, Section 4 analyses the different graphical tools and their statistical properties and Section 5 concludes.

2. Data

We use the same dataset as González-Val et al. (2013b): this database includes the decennial census for each decade of the 20th century with un-truncated city population data from the three countries.[1]

The US database is created from the original documents of the annual census published by the US Census Bureau (www.census.gov) and consists of the available data on all incorporated places without any size restriction. The US Census Bureau uses the generic term *incorporated place* to refer to the governmental unit incorporated under state Law as a city, town, borough or village. Alaska, Hawaii and Puerto Rico are excluded because of data limitations. The number of cities considered by period is: 1900 (10,596 incorporated places), 1910 (14,135), 1920 (15,481), 1930 (16,475), 1940

[1] More information about the databases and comparisons between these countries can be found in González-Val et al. (2013b).

(16,729), 1950 (17,113), 1960 (18,051), 1970 (18,488), 1980 (18,923), 1990 (19,120) and 2000 (19,296).

For Spain and Italy, the geographical unit of reference is the *municipality*, and the data come from official statistical information services. In Italy, this is the *Istituto Nazionale di Statistica* (www.istat.it), while for Spain we have taken the census of the *Instituto Nacional de Estadística* (www.ine.es). For Italy, the number of cities by period is 7,711 municipalities in 1901 and 1911 and 8,100 municipalities from 1921 to 2001. For Spain, we consider the following years: 1900 (7,800 municipalities), 1910 (7,806), 1920 (7,812), 1930 (7,875), 1940 (7,896), 1950 (7,901), 1960 (7,910), 1970 (7,956), 1981 (8,034), 1991 (8,077) and 2001 (8,077).

We consider administratively defined cities (legal cities) in the three countries; thus their boundaries may not make economic sense and, in many cases, they may not correspond to a meaningful economic definition of a city. Although metropolitan areas are considered to be more natural economic units, some factors, such as human capital spillovers, are thought to operate at a local level, and there are important statistical reasons to consider an un-truncated city population dataset (Eeckhout, 2004).

3. Estimation

We fit a lognormal distribution to our city size data. The probability density function (pdf) of the lognormal is:

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

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

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

where erf denotes the error function associated with the normal distribution. A relationship between rank (1 for the most populous centre, 2 for the second, and so on) and cdf can be easily found (Eeckhout, 2004; Stanley et al., 1995). The expression of the rank of cities $r(x)$ according to population is

$$r(x) = r_0(1 - cdf(x)) = r_0\left(\frac{1}{2} - \frac{1}{2}erf\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right)\right). \qquad (3)$$

where $r_0 > 0$ is a new constant equivalent to the sample size. We use the Maximum Likelihood estimators, and later we estimate r_0 by OLS taking into account the estimated cdf and Equation (3). The estimates of these parameters are very significant in the three countries and for all years. The estimations of $\hat{r}_0$ are directly related to sample size; those of $\hat{\mu}$ are very stable over time for all three countries, while 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 (3), shows that the degree of fit is very good.[2]

4. The accuracy of traditional graphs

The first graphical tool we consider is the histogram. 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 (Graph (a) in Figure 1, data from Spain in 1900). A histogram values the frequencies associated with intervals of a constant width on the x-axis. However, in a histogram of the population logarithm (Graph (b) in Figure 1, 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 common ratio e^{δ}. This reasoning is valid for any numerical variable which is measured alternatively in levels or in natural logarithms (populations, sales or employees).

This fact explains why taking logarithms gives a bell-shaped curve: the first intervals are very narrow; 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

[2] The results, not shown for size restrictions, are available from the authors on request.

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 a bell-shaped curve is obtained. Therefore, the same population data can be well fitted by different statistical distributions, depending on the scale of the variable (levels or logarithms).

The second tool we examine the performance of is Zipf plots, i.e., double logarithmic graphs of rank compared to population, which are used extensively in the specialised literature (Stanley et al., 1995). Panel (a) in Figure 2 shows the most representative ones.[3] These graphs represent the actual data (black dots) with the estimated lognormal distribution (blue line). In general, the lognormal distribution is a good description of the overall city size distribution, but, in most cases, the lognormal underestimates the empirical distribution at the upper tail of larger cities. The discrepancies between the data and the estimated theoretical distribution tend to increase clearly and systematically with city size.

We can demonstrate that these discrepancies are augmented in the Zipf plot for a statistical reason. Below, the quantities with overbar correspond to the empirical or sample distribution and those without overbar to the estimated theoretical distribution (lognormal). From Equation (3):

$$\bar{r}(x) = \bar{r}_0\left(1 - \overline{cdf}(x)\right), \tag{4}$$

$$r(x) = r_0\left(1 - cdf(x)\right). \tag{5}$$

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 and evaluate their difference, we obtain:

$$\ln\bar{r}(x) - \ln r(x) = \ln\bar{r}_0\left(1 - \overline{cdf}(x)\right) - \ln r_0\left(1 - cdf(x)\right) = \ln\bar{r}_0 - \ln r_0 + \ln\left(1 + \frac{cdf(x) - \overline{cdf}(x)}{1 - cdf(x)}\right) \tag{6}$$

[3] The results for the decades not shown are available from the authors on request.

We focus on the last term. The discrepancy $cdf(x)-\overline{cdf}(x)$ is small (and gets smaller as x becomes very large) but it is nonzero and it is quite bigger than the quantity $1-cdf(x)$, which indeed unequivocally tends to zero as x becomes very large. Figure 3 shows these two elements for the example of the upper tail city size distribution of the US in 1950. Thus, the quotient $\dfrac{cdf(x)-\overline{cdf}(x)}{1-cdf(x)}$ is a quantity much bigger than the discrepancy $cdf(x)-\overline{cdf}(x)$. Adding the unity to the quotient and taking the natural logarithm has the effect of reducing the quotient considerably, but the resulting quantity is still much bigger than the original discrepancy $cdf(x)-\overline{cdf}(x)$. Figure 4 plots the elements $\dfrac{cdf(x)-\overline{cdf}(x)}{1-cdf(x)}$ and $\ln\left(1+\dfrac{cdf(x)-\overline{cdf}(x)}{1-cdf(x)}\right)$ for the same case as in Figure 3, namely, the upper tail city size distribution of the US in 1950. The graph of the last quantity is equivalent, up to the terms $\ln\overline{r}_0-\ln r_0$, to the discrepancy at the upper tail in the Zipf plot of Figure 2, panel (a), USA in 1950. In short, the discrepancy $cdf(x)-\overline{cdf}(x)$ has been amplified in the upper tail by taking logarithms of the ranks. This observation is not in contradiction with common wisdom about Zipf plots but rather reinforces and qualifies it: Zipf plots are adequate to see whether there are deviations between theoretical and empirical cumulative distribution functions at the upper tail, but bearing in mind that the possible discrepancies are automatically amplified. Thus, if it happens that there is absence of differences between empirical and theoretical Zipf plots at the upper tail, then we can assure that the fit is extremely good. Moreover, this observation can contribute to the clarification of recent questions raised in the literature (Levy, 2009; Eeckhout, 2009). In particular, this is why the confidence bands in Zipf plots fan out as population increases in the upper tail of the distribution.

Finally, we study the graphical representation of the cumulative distribution functions (Eeckhout, 2004; Giesen et al., 2010). Panel (b) in Figure 2 shows the cdf s corresponding to the same cases in which we illustrated the Zipf plots. The black dots represent the empirical cdf and the blue line is the estimated cdf corresponding to the lognormal distribution. In principle, we would expect the results to be similar to those of the Zipf plots, but we can see that this is not exactly true. Surprisingly, the fit in the lower tail is not as good as it seemed in the Zipf plots, while the fit in the upper tail seems almost perfect. To explain this apparent paradox it is useful to turn again to Equations (4) and (5). From these, we deduce:

$$\bar{cdf}(x) - cdf(x) = \frac{r(x)}{r_0} - \frac{\bar{r}(x)}{\bar{r}_0}. \tag{7}$$

Imagine first that the fit in ranks for the estimated distribution was very good except for the smallest cities, which would mean that $r(x) \cong \bar{r}(x)$ for practically all points, so that Equation (7) would be:

$$\bar{cdf}(x) - cdf(x) = \frac{r(x)}{r_0}\left(\frac{\bar{r}_0 - r_0}{\bar{r}_0}\right) = (1 - cdf(x))\left(\frac{\bar{r}_0 - r_0}{\bar{r}_0}\right). \tag{8}$$

Equation (8) is obtained assuming that the fit in ranks is almost perfect ($r(x) \cong \bar{r}(x)$ except for the smallest cities). The cdf s fit less well as the difference $\bar{r}_0 - r_0$ increases. 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.[4] Thus, the discrepancy in cdf s could be perfectly compatible with an almost perfect rank fit, except for the smallest cities. Furthermore, it is unavoidable if $\bar{r}_0 - r_0 \neq 0$. This happens for the cdf in Spain in 1950.

Second, however, in most cases of our estimated lognormal distribution it happens that $r_0 \cong \bar{r}_0$ (remember that $\bar{r}_0$ is identified with the sample size), so that Equation (7) is actually reduced to:

$$\bar{cdf}(x) - cdf(x) = \frac{1}{r_0}(r(x) - \bar{r}(x)). \tag{9}$$

Thus, we derive that when $r_0 \cong \bar{r}_0$, any lack of fit in ranks (*not logarithms* of ranks) is directly transferred, in most of our estimations with the lognormal, to a lack of fit in cdf s. These (rather small) discrepancies are shown in the cdf s plotted for Italy in 1951 and for the US in 1950 and 2000.

5. Conclusions

[4] See Figure 2(b). The divergence between $\bar{cdf}(x)$ and $cdf(x)$ is noticeable for $\ln(x)$ (in the horizontal axis) lower than, say, 7, and from that value the differences are negligible.

In this paper we show some limitations of the traditional graphs used to study size distributions in economics: histograms, Zipf plots and plotted cumulative density functions. We fit a lognormal distribution to un-truncated city population data from three countries: the US, Spain and Italy. We obtain some statistical properties to explain the graphical behaviours at the lower and upper tail distribution. This evidence suggests that the appropriate tools to test statistical size distributions properly are standard statistical tests and information criteria (see Giesen et al., 2010; González-Val et al., 2013a), rather than these graphical tools.

References

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.

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.

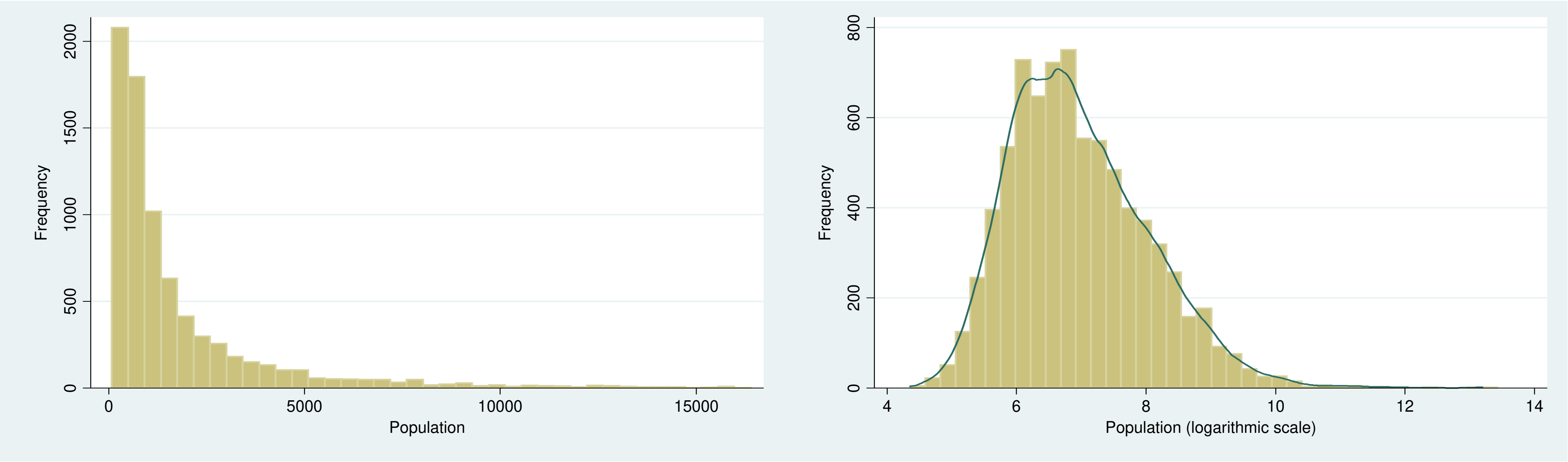
González-Val, R., L. Lanaspa and F. Sanz (2013b). “New evidence on Gibrat’s law for cities,” Urban Studies, forthcoming.

González-Val, R., A. Ramos, F. Sanz, and M. Vera-Cabello (2013a). “Size distributions for all cities: which one is best?,” Papers in Regional Science, forthcoming.

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

Stanley, M. H. R., S. V. Buldyrev, S. Havlin, R. N. Mantegna, M. A. Salinger and H. E. Stanley, (1995). “Zipf plots and the size distribution of firms,” Economics Letters, 49: 453-457.

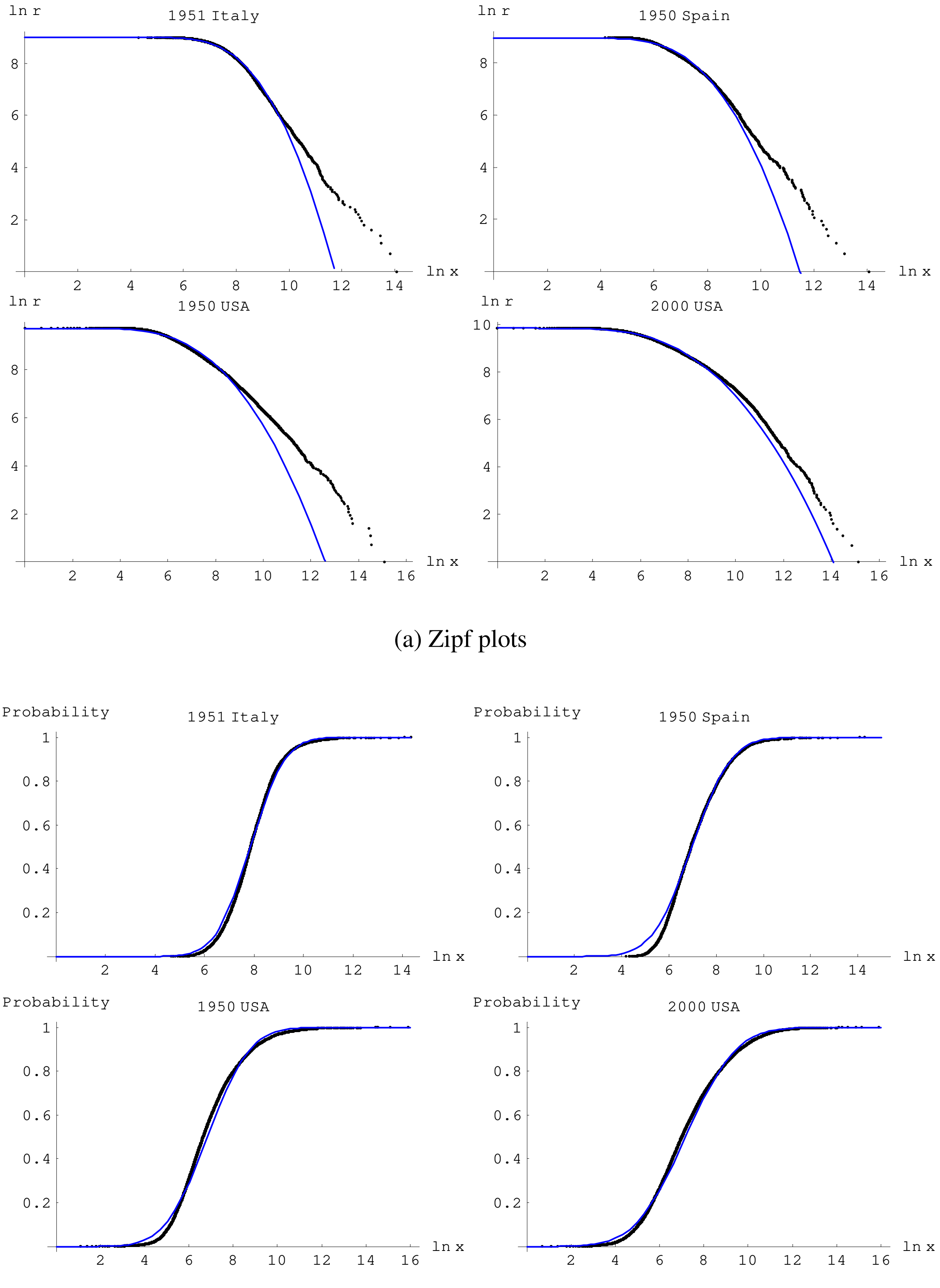
Figure 1. Histogram of Spanish cities in 1900



(a) Population in levels

(b) Population in logarithm

Figure 2. Zipf and *cdf* plots



(a) Zipf plots

(b) *cdf* plots

Figure 3. Plot of $cdf(x) - c\bar{d}f(x)$ (red) and $1 - cdf(x)$ (blue) for the upper tail city size distribution in the US (1950)

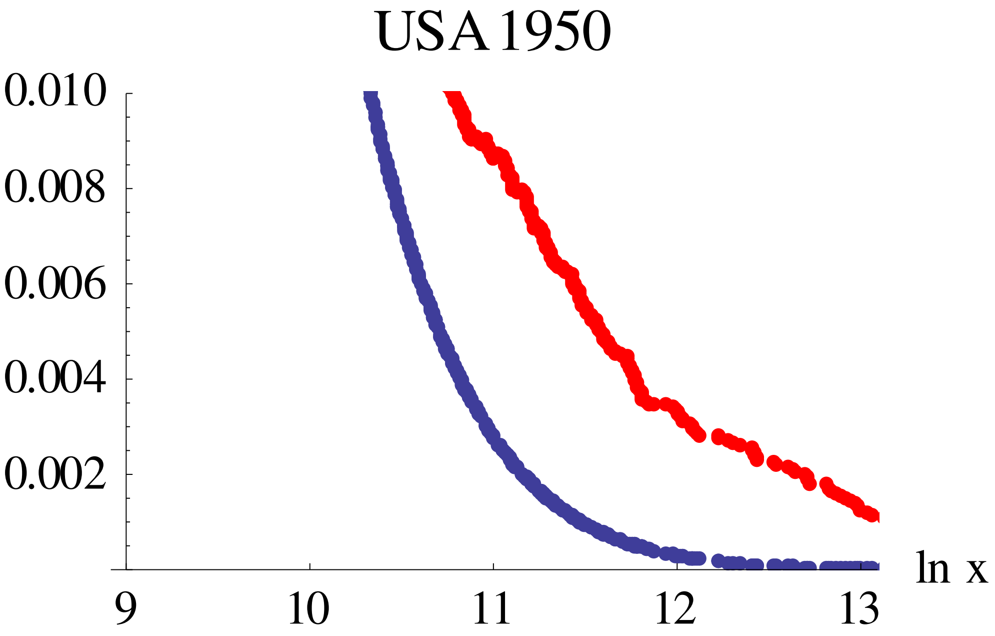
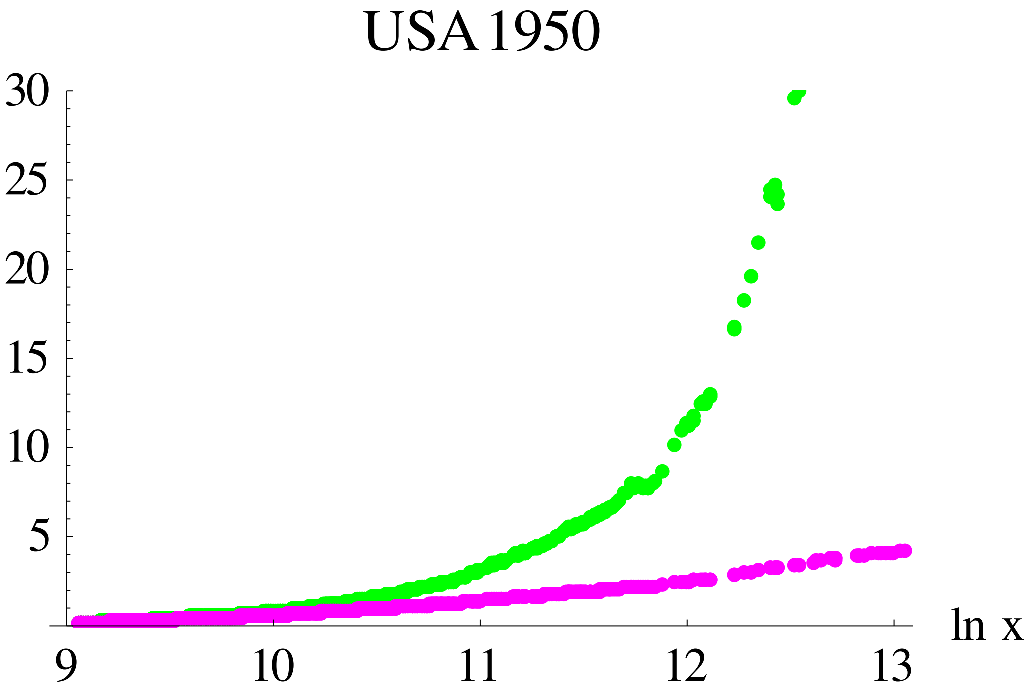


Figure 4. Plot of $\frac{cdf(x) - c\bar{d}f(x)}{1 - cdf(x)}$ (green) and $\ln\left(1 + \frac{cdf(x) - c\bar{d}f(x)}{1 - cdf(x)}\right)$ (magenta) for the upper tail city size distribution in the US (1950)


USA 1950
ln x