

Sobre las mejores funciones para describir la distribución del tamaño
de las ciudades

Rafael González-Val

Arturo Ramos

Fernando Sanz

Departamento de Análisis Económico

Universidad de Zaragoza

Abstract

En este trabajo se analizan con detalle las prestaciones que una función prácticamente nueva en Economía Urbana, como es la q-exponencial, ofrece como descriptora de la distribución del tamaño de las ciudades. Destacamos dos aportaciones. Uno, se propone un procedimiento novedoso y sencillo para estimar sus parámetros. Y dos, más importante, se explican las características asociadas a dos métodos gráficos tradicionales (*Zipf plots* y *cumulative density functions*) para discriminar entre funciones y se aplican a la lognormal y q-exponencial, justificando que son las más adecuadas para explicar toda la distribución y que la relación entre ellas es de complementariedad.

Keywords: city size distribution, q-exponential distribution, lognormal distribution

JEL: C13, C16, R00.

Dirección:

Dpto. de Análisis Económico, Universidad de Zaragoza, Facultad de CC.
Económicas y Empresariales, Gran Vía, 2, 50005 Zaragoza (Spain)

E-mail: fsanz@unizar.es

1. Introducción

El estudio de la distribución del tamaño de las ciudades tiene una larga tradición en Economía Urbana. Véase, por ejemplo, y sin querer ser en absoluto exhaustivos, Rosen y Resnick (1980), Black y Henderson (2003), Sharma (2003), Ioannides y Overman (2003), Soo (2005), Anderson y Ge (2005), Bosker et al. (2008). Estas distribuciones tienen un interés más allá de lo puramente estadístico, por dos motivos esenciales, que se retroalimentan e influyen mutuamente. En primer lugar, porque la distribución de las dimensiones de las ciudades define el paisaje económico resultante – más concentrado, más disperso; sesgado o no hacia un número excesivo de núcleos grandes, o pequeños, con una o más modas; con ciudades muy iguales o muy desiguales en su tamaño– y todo ello tiene consecuencias directas sobre la distribución espacial de la renta, sobre la realización o no de inversiones públicas en infraestructuras de diverso tipo en ciertas áreas y, en definitiva, sobre los desequilibrios interterritoriales. ¿Y qué político va a atreverse a decir que no le interesan estos temas? Hablando de política, la distribución del tamaño de las ciudades también afecta, y mucho, al reparto final por áreas geográficas del número de escaños que va a uno u otro partido en unos comicios. Y en segundo lugar, porque esta distribución de los tamaños es susceptible de cambiar en el tiempo, de acuerdo a ciertos estímulos de índole esencialmente económica.

Históricamente, la distribución que ha generado un mayor número de trabajos y una mayor aceptación es la de Pareto, que es una ley potencial cuya función de densidad viene dada por:

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

donde a es una constante, $b > 0$ es el exponente de Pareto y x es el número de habitantes de cada núcleo urbano. Considerando el rango r (1 para el núcleo más

poblado, 2 para el segundo y así sucesivamente) de las N ciudades puede llegarse a la conocida expresión

$$\ln r = \text{const} - b \ln x, \quad (2)$$

que relaciona el logaritmo del rango con el logaritmo del tamaño de las ciudades si éstas siguen una distribución de Pareto. En el caso de que $b=1$ se obtiene la bien conocida ley de Zipf o *rank-size rule* (véanse al respecto los *surveys* sobre el tema de Cheshire, 1999, y Gabaix e Ioannides, 2004).

La ley de Zipf es relevante fundamentalmente por tres motivos. Uno, se aplica a otros fenómenos cuantificables de forma bastante ajustada, como es el caudal de los ríos, número de veces que aparece una misma palabra en un texto, o la intensidad de los terremotos (Zipf, 1949; Krugman, 1996a). Dos, si se consideran las 135 Metropolitan Statistical Areas existentes en USA en 1991, el exponente de Pareto es 1,005 (Krugman, 1996b, también recogido en Gabaix, 1999), que implica el cumplimiento prácticamente exacto de la citada ley. Y tres, existe cierto consenso acerca de que la estructura urbana que se deriva del cumplimiento de la ley define una jerarquía equilibrada, en la que las ciudades de todos los tamaños tienen una representación adecuada. En definitiva, la distribución de Pareto y un caso particular de la misma como es la ley de Zipf, son ciertamente útiles para explicar el comportamiento de los núcleos urbanos, especialmente de los más grandes (*upper tail*).

No obstante, la descripción de Pareto posee algunas deficiencias importantes. Por un lado, la función de densidad de Pareto tiene una asíntota vertical en $x=0$, lo que no es verosímil. Por otro, el Zipf plot (que relaciona en unos ejes cartesianos $\ln r$ con $\ln x$) se aleja de una línea recta de pendiente negativa como predice (2) cuando, de

nuevo, se toman todos los núcleos urbanos o un punto de corte de población suficientemente bajo.

En este orden de cosas, llega la aportación de Eeckhout (2004), que propone esencialmente tres ideas. Uno, cuando se toman todos los núcleos, sin ninguna restricción de tamaño, la distribución de Pareto decae y la mejor representación de los datos es una función lognormal. Dos, lo que constituye un resultado teórico: si la distribución subyacente es lognormal, lo que genera un Zipf plot cóncavo, el exponente de Pareto es decreciente con el tamaño muestral, lo cual implica que puede buscarse el tamaño muestral que consigue que la ley de Zipf se verifique exactamente. Estas dos primeras contribuciones ponen claramente de manifiesto la importancia de tomar todas las ciudades pues, en otro caso, se puede estar incurriendo en sesgos o llegar a conclusiones espúreas. Y tres, los datos de todos los núcleos USA en 1990 y 2000 apoyan la hipótesis de lognormalidad y el cumplimiento de la ley de Gibrat o del proportionate growth, algo que ya anticiparon desde un punto de vista teórico Gibrat (1931) y Kalecki (1945).

En resumen, las dos distribuciones que la literatura económica ha utilizado más profusamente son la paretiana y, recientemente, la lognormal¹. Sin embargo, en un medio de difusión relativamente extraño para un economista o un geógrafo puesto que se trata de una revista de física, Malacarne et al. (2001) demuestran que, cuando se toman todos los núcleos, la llamada distribución q -exponencial presenta un muy buen ajuste a los datos. Hasta donde conocemos, el único trabajo adicional que contrasta esta afirmación es el de Soo (2007), quien, al tomar las ciudades más grandes de Malasia (por encima de 10.000 habitantes) obtiene resultados negativos referente a las

¹ Aún más recientemente se han propuesto en la literatura distribuciones que son una combinación de ambas (Giesen et al., 2009 y Ioannides y Skouras, 2009).

prestaciones de la q -exponencial, lo que lleva a pensar, como ocurre con la lognormal, que es una distribución especialmente adecuada cuando no se define *truncation point* alguno.

Recientemente se han construido bases de datos mucho más completas, que permiten aproximarse con más información estadística de partida a la problemática que se plantea en este trabajo. En concreto, en González-Val (2010) se consideran todos los núcleos de población de USA durante todo el siglo veinte; en González-Val et al. (2010) se hace lo propio, además de para USA, para España e Italia. Si se emplean estos datos para representar el logaritmo del rango frente al logaritmo del tamaño se observa una clara desviación de la linealidad en todos los casos, lo que abre la puerta para que se tengan en cuenta distribuciones no paretianas. Lo que queremos destacar es que, a excepción de Eeckhout (2004), ninguno de los estudios previos considera la entera distribución de ciudades, ya que todos imponen un *truncation point*, ya sea explícito tomando las ciudades por encima de un umbral mínimo de población, o implícito al trabajar con MSAs². Generalmente esto es debido a una cuestión práctica de disponibilidad de datos.

En este contexto, los objetivos de este artículo son los siguientes. En primer lugar, contrastar para un período largo (cien años), para varias estructuras urbanas (las de España, Italia y USA) y considerando todos los núcleos (unos 8.000 para los países mediterráneos en todo el siglo y de 10.600 a más de 19.000 para USA según los años) las prestaciones de la q -exponencial para describir la distribución del tamaño de las ciudades. En segundo lugar, llevar a cabo el mismo ejercicio para la lognormal. Y en

² En Estados Unidos para obtener la calificación de MSA es necesario una ciudad de 50.000 o más habitantes, o la presencia de un área urbanizada de al menos 50.000 habitantes, y una población total metropolitana de cómo mínimo 100.000 (75.000 en New England) habitantes, según la definición oficial. En otros países se siguen criterios similares, si bien puede cambiar el umbral mínimo de población necesario para considerar un área como metropolitana.

tercer lugar, poner en relación las ventajas e inconvenientes de las dos distribuciones, a saber, q -exponencial y lognormal, y determinar si tienen un carácter sustitutivo o complementario y bajo qué circunstancias ocurre eso. El cumplimiento de estos tres objetivos básicos va a conducir, como se verá más adelante, a realizar una sencilla contribución en lo relativo al método de estimación y a presentar algunos desarrollos, nada complicados, de estadística teórica que justifican y explican las diferencias que surgen entre las distribuciones de la q -exponencial y la lognormal. En cualquier caso, hasta donde conocemos es la primera vez que se someten a contraste empírico estas cuestiones empleando unas bases de datos tan amplias.

Las principales aportaciones que ofrece este trabajo son tres. En primer lugar, se propone un procedimiento novedoso y sencillo para estimar los parámetros de la distribución q -exponencial, que mejora lo hecho hasta ahora en la literatura. En segundo lugar, se explican las ventajas e inconvenientes asociadas a dos métodos gráficos tradicionales (*Zipf plots* y *cumulative density functions*) para discriminar entre funciones de densidad, y se aplican a las distribuciones q -exponencial y lognormal. Por último, se concluye que ambas distribuciones son adecuadas para describir con precisión las distribuciones del tamaño de las ciudades y que la relación entre ellas es fundamentalmente de complementariedad.

El artículo se organiza como sigue. En la Sección segunda se define y caracteriza la distribución q -exponencial; el apartado tercero da cuenta y explica las bases de datos empleadas; el epígrafe cuarto es el más largo del trabajo y en él se comparan y enfrentan la q -exponencial y la lognormal como funciones potencialmente válidas para describir la distribución del tamaño de las ciudades; la sección quinta demuestra cómo ambas distribuciones son más complementarias que sustitutivas; finalmente las conclusiones cierran el papel.

2. On the q -exponential distribution

La *probability density function* (pdf) de la q -exponencial está dada por:

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

donde $a > 0$ y $q > 1$ son parámetros y x denota, como ya sabemos, la población de los núcleos urbanos. La expresión de la *cumulative distribution function* (cdf) correspondiente es la siguiente:

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

y la del rango de las ciudades en función de la población es

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

donde $r_0 > 0$ es una nueva constante que equivale al tamaño muestral. En el caso de que $q \rightarrow 1$, $f(x) \rightarrow ae^{-ax}$ y $r(x) \rightarrow r_0 e^{-ax}$, lo cual justifica el nombre de q -exponencial.

Esta distribución ha sido usada profusamente por Tsallis (1988) y su grupo de colaboradores en la literatura de física, argumentando teóricamente su aplicabilidad a sistemas con interacciones de largo alcance, y el citado trabajo de Malacarne et al. (2001) puede encuadrarse en esta línea. No obstante, la q -exponencial es un caso particular de la distribución conocida con el nombre de Pareto generalizada tipo II, que se ha considerado en varios trabajos anteriores (por ejemplo, Hosking y Wallis, 1987, Grimshaw, 1993 y Choulakian y Stephens, 2001). El rango frente al tamaño, según (5), es una función decreciente (algo que, por definición, debe ocurrir siempre)

estrictamente convexa y que para $x = 0$ alcanza un valor finito, r_0 , mejorando en este aspecto el comportamiento de la distribución de Pareto.

3. Las bases de datos

Utilizamos datos de población de ciudades de tres países: Estados Unidos, España e Italia. Estados Unidos es un país sumamente interesante para realizar un análisis de la evolución de su estructura urbana, ya que se trata de un país relativamente joven cuyos habitantes se caracterizan por una elevada movilidad. En el otro lado tenemos los países europeos, con una estructura urbana mucho más antigua y con unos habitantes que presentan una mayor resistencia a moverse, en concreto Cheshire y Magrini (2006) evalúan que la movilidad en Estados Unidos es quince veces mayor que la europea.

El considerar estos dos tipos de países nos proporciona información de comportamientos urbanos distintos, ya que mientras que España e Italia tienen un tejido urbano ya consolidado y apenas se crean nuevas ciudades (el crecimiento urbano se produce por incremento de la población de las ya existentes), en Estados Unidos el crecimiento urbano tiene una doble dimensión, ya que además de aumentar la población de las ciudades también aumenta su número, con efectos potencialmente distintos sobre la distribución del tamaño de las ciudades. Así, se pasa de que la población de las ciudades (los *incorporated places*) represente menos de la mitad de la población total de Estados Unidos en 1900 (el 46,99%) hasta el 61,49% del año 2000, a la vez que el número de ciudades aumenta en un 82,11%, pasando de 10.596 en 1900 a 19.296 en el año 2000.

Los datos de Estados Unidos que utilizamos son los mismos que emplean González-Val (2010) y González-Val et al. (2010). Nuestra base, elaborada a partir de

los documentos originales de los censos anuales que publica el *US Bureau of the Census*, www.census.gov, está formada por los datos disponibles de todos los *incorporated places* sin ninguna restricción de tamaño para cada década del siglo veinte. El *US Census Bureau* utiliza el término genérico *incorporated place* para referirse a la unidad gubernamental incorporada bajo ley estatal como *city*, *town* (excepto en los estados de New England, New York y Wisconsin), *borough* (excepto en Alaska y New York), o *village*, y que tiene límites, poderes y funciones establecidos legalmente.

El número de ciudades (entre paréntesis) correspondiente para cada periodo es: 1900 (10.596 ciudades), 1910 (14.135), 1920 (15.481), 1930 (16.475), 1940 (16.729), 1950 (17.113), 1960 (18.051), 1970 (18.488), 1980 (18.923), 1990 (19.120) y 2000 (19.296).

Dos precisiones deben hacerse. Primero, se excluyen todas las ciudades correspondientes a Alaska, Hawaii y Puerto Rico para cada década, debido a que estos estados se anexionan a lo largo del siglo XX (Alaska y Hawaii en 1959, y el caso especial de Puerto Rico, que se anexiona en 1952 como estado libre asociado) y no existen datos para todos los periodos. Su inclusión provocaría inconsistencia geográfica de las muestras, que no serían homogéneas en términos geográficos y por tanto no se podrían comparar. Y segundo, por la misma razón excluimos también todos los *unincorporated places* (designan concentraciones de población que no forman parte de ningún *incorporated place*, pero que están localmente identificadas con un nombre), que comienzan a contabilizarse a partir de 1950. Sin embargo, estos asentamientos ya existían anteriormente, por lo que su inclusión nos plantearía de nuevo un problema de inconsistencia en la muestra. Además, su eliminación tampoco es importante cuantitativamente; en efecto, los *unincorporated places* son 1.430 en 1950 y representan

un 2,36% de la población total de Estados Unidos, cifras que pasan a ser 5.366 y un 11,27% en 2000.

Para España e Italia la unidad geográfica de referencia es el municipio y la fuente de los datos son los servicios oficiales de información estadística. En Italia es el Servizio Biblioteca e Servizi all'utenza de la Direzione Centrale per la Diffusione della Cultura e dell'informazione Statistica que se encuentra dentro del Istituto Nazionale di Statistica, www.istat.it, y para España se han tomado los censos que ofrece el Instituto Nacional de Estadística, INE, www.ine.es. De cada municipio se toma población de hecho (resident population).

Tomamos los datos correspondientes a los censos de cada década del siglo veinte. Para Italia se toman datos de los siguientes años, (entre paréntesis figura el número de municipios de cada año): 1901 (7.711), 1911 (7.711), 1921 (8.100), 1931 (8.100), 1936 (8.100), 1951 (8.100), 1961 (8.100), 1971 (8.100), 1981 (8.100), 1991 (8.100) y 2001 (8.100). No existe censo en Italia para el año 1940, debido a su participación en la Segunda Guerra Mundial, por lo que tomamos el dato del año 1936. Para España se han considerado los siguientes años: 1900 (7.800 municipios), 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) y 2001 (8.077).

4. Lognormal versus q -exponential. El enfoque sustitutivo

4.1. Métodos de estimación

4.1.1. Estimación de la lognormal

La *probability density function* (pdf) de la lognormal está dada por:

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

donde μ y σ son la media y la varianza de $\ln x$, que en este caso denota el logaritmo neperiano de la población de los núcleos urbanos. La expresión de la *cumulative distribution function* (cdf) correspondiente es:

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

donde erf denota la función de error asociada a la distribución normal. La expresión del rango de las ciudades en función de la población viene dada por:

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

Para estimar si unos datos siguen una distribución lognormal es habitual hacerlo por medio de *Maximum Likelihood* (ML), si bien este método no se ha aplicado tan profusamente como en otros campos para la población de núcleos urbanos³. Los estimadores de ML se expresan de forma sencilla en términos de los datos de población:

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

donde N es el tamaño muestral. Es decir, la media y varianza estimadas son justamente las de los datos. Posteriormente estimamos r_0 por OLS teniendo en cuenta la *cdf* estimada y la relación (8).

Las estimaciones de estos parámetros para nuestros datos son muy significativas en los tres países y para todos los años. Esta información se muestra en el Cuadro 1. Las estimaciones de \hat{r}_0 están directamente relacionadas con el tamaño muestral, como ya sabíamos, las de $\hat{\mu}$ son muy estables en el tiempo para los tres países, mientras que,

³ Recuérdese que Eeckhout (2004) es el primero en proponer la lognormal para la distribución del tamaño de las ciudades.

como es lógico, los valores de $\hat{\sigma}^2$ son ligeramente crecientes en el tiempo para las tres zonas. El R^2 , correspondiente a la estimación OLS de r_0 aplicando la ecuación (8), muestra que el grado de ajuste es muy bueno.

4.1.2. Estimación de la q -exponencial

En su artículo original, Malacarne et al. (2001) parten de la allí denominada función q -logarítmica $\ln_q(x)$, que en el lenguaje econométrico puede entenderse como una transformación Box-Cox. Muestran que si (y sólo si) el valor de q usado en tal función coincide con el valor de q de la expresión del rango $r(x)$ dada por (5), entonces $\ln_q(r(x))$ es lineal en x y puede representarse como una línea recta. Es decir, se necesita conocer de antemano el valor de q que se pretende hallar, lo cual es un problema tanto desde un punto de vista conceptual como operativo. Por otro lado, Soo (2007) estima (5) dando valores a q de décima en décima, eligiendo aquél que minimiza la suma de los residuos y, una vez determinado éste \hat{q} , se obtienen los de a y r_0 por estimación no lineal; se trata de una solución al problema mejorable puesto que también implica proponer unos valores iniciales arbitrarios de aquel parámetro que precisamente se quiere estimar.

Por otra parte, la distribución de la q -exponencial se estima por ML en varios trabajos o, más precisamente, la distribución de Pareto generalizada (véanse, al respecto, las referencias sobre este tema citadas en la Sección segunda). En todos ellos el número de observaciones utilizadas no es elevado. En nuestro caso el tamaño muestral supera los 8.000 para España e Italia y va de los 10.600 a casi 20.000 en el caso de USA y esto hace que el usar ML presente serias dificultades. En efecto, para ver

el origen de las mismas para tamaños muestrales elevados, las condiciones de primer orden de maximización de la verosimilitud conjunta quedan:

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

Se deduce a simple vista que especialmente la segunda ecuación en (10) es difícil de resolver, puesto que ha de encontrarse la raíz de una ecuación de grado muy elevado.

Ante la gran dificultad de usar ML, por un lado, y los problemas para estimar los parámetros de interés que presentan los trabajos precedentes, por otro, nosotros hemos optado por estimar de partida dichos parámetros de (5) mediante métodos no lineales, sin fijar los valores de ninguno de antemano. Ahora bien, dicha estimación no lineal requiere unos valores iniciales de los parámetros, que se revelan fundamentales a la hora de que el algoritmo converja en un número razonable de iteraciones.

Hemos desarrollado un modo novedoso de hallar estos valores iniciales. El procedimiento es como sigue. Sea p un nuevo parámetro exógeno que eventualmente puede tomar el valor de q . Sea $\varepsilon = \frac{q-p}{1-q}$. Entonces, de (5) tenemos:

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

En el caso que $p = q$ tenemos que $\varepsilon = 0$ y la expresión de $r(x)^{1-p}$ se vuelve lineal en x . Procedemos entonces a tomar un desarrollo de Taylor de (11) hasta orden dos

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

de donde pueden hallarse los valores de los parámetros (por identificación de los coeficientes de los desarrollos de (11) y (12)). Primero

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

después, el valor de q por medio de $q = \frac{\varepsilon + p}{1 + \varepsilon}$ (p está dado exógenamente) y

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

El problema se reduce, por tanto, a estimar los valores de $\lambda_1, \lambda_2, \lambda_3$. Lo hacemos efectuando regresiones lineales de r^{1-p} en las variables x, x^2 con nuestros datos de rango para valores de p que hagan muy próximo a cero el valor de λ_3 , y de modo que los valores de los parámetros obtenidos tengan sentido. Este procedimiento se ha mostrado muy eficaz para obtener los valores iniciales de los parámetros.

Posteriormente, realizamos una estimación simultánea por mínimos cuadrados ordinarios no lineales (Bates y Watts, 1988) de los tres parámetros r_0, q, a de la especificación (5), suministrando los valores iniciales obtenidos por el procedimiento anterior. Los parámetros estimados son muy significativos en todos los casos y se muestran en el Cuadro 2. Las estimaciones de \hat{q} presentan una clara tendencia creciente en el tiempo en los tres países; las de \hat{a} sólo tienen esta característica en España e Italia y a partir de mitad de siglo. La bondad del ajuste es también muy buena y, casi siempre, ligerísimamente superior (estamos hablando de milésimas e, incluso, de diferencias en diezmilésimas para el país transalpino) a la obtenida en el Cuadro 1 con la distribución lognormal.

4.2. Comparación en términos de *Zipf plots*

Procedemos en este apartado a comparar ambas distribuciones en términos de *Zipf plots*, es decir, gráficos doble logarítmicos de la población frente al rango, que son profusamente utilizados en la literatura especializada.

Mostraremos tales gráficos en la comparación del logaritmo de los rangos teóricos y empíricos para las dos distribuciones estudiadas, q -exponencial y lognormal. En términos de *Zipf plots*, y en un análisis visual de los mismos, se dan varios casos en los que el ajuste por la q -exponencial es aparentemente mejor que por la lognormal o donde no es fácil discernir entre las dos. Ilustraremos los más representativos. En el Gráfico 1 mostramos la situación para la q -exponencial, mostrando los *Zipf plots* de los datos reales (negrita) con el de la q -exponencial estimada o teórica (azul). En el Gráfico 2 se muestra lo análogo para la lognormal.

Comenzando por el caso de Italia en 1951, visualmente ninguna de las dos distribuciones parece proporcionar un buen ajuste, si bien la divergencia con la lognormal teórica aparece solo en la cola superior de la distribución. En otros años del siglo XX para Italia el ajuste de las dos distribuciones es similar salvo en el año 2001 donde la q -exponencial proporciona aparentemente una mejor descripción.

Sin embargo, en otros casos la situación es más favorable a la q -exponencial, por ejemplo en España en el año 1950. En este caso el ajuste que proporciona la q -exponencial parece mucho mejor que el de la lognormal. Para España, esto ocurre a lo largo de casi todo el siglo XX, siendo 2001 el año donde las dos proporcionan aparentemente un ajuste similar.

Para el caso de USA, hay una mejora progresiva en la descripción por la q -exponencial hasta 1950 (cuando el ajuste parece muy bueno) y después un empeoramiento, mientras que la lognormal muestra un comportamiento similar a lo

largo del siglo XX. Precisamente, el año 2000 en USA, el mismo considerado en Eeckhout (2004, 2009) y Levy (2009), es un ejemplo en el que no está claro qué distribución es visualmente mejor: la q-exponencial sistemáticamente sobreestima el tamaño de las ciudades más grandes, mientras la lognormal los infraestima. En todo caso, parece ser que la lognormal se acomoda ligeramente mejor al rango de los núcleos más pequeños, confirmando lo ya observado en los Cuadros 1 y 2 referente a las estimaciones de r_0 .

Otro resultado general es que, para cualquiera de las dos distribuciones, se observa que puede haber discrepancias en los *Zipfs plots* entre los datos y la correspondiente distribución teórica y que ésta tiende a aumentar de forma clara y sistemática con el tamaño de las ciudades. No es difícil encontrar una explicación estadística a este hecho. En efecto, sean a partir de ahora, las cantidades con barra correspondientes a la distribución empírica o muestral, mientras que sin ella se refieren a la teórica:

$$\bar{r}(x) = \bar{r}_0(1 - \bar{cdf}(x)), \quad (13)$$

$$r(x) = r_0(1 - cdf(x)). \quad (14)$$

En el origen ambas *cdf*'s son nulas luego $\bar{r}(0) = \bar{r}_0$ y $r(0) = r_0$. A su vez, para un valor arbitrariamente grande, infinito, de la población de las ciudades las dos *cdf*'s han de valer uno, con lo cual $r(\infty) = \bar{r}(\infty) = 0$.

Si, como exige el *Zipf plot*, tomamos logaritmos, lo cual es importante, y evaluamos su diferencia se obtiene la siguiente expresión:

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

donde el último término en (15) es el fundamental y merece nuestra atención. En efecto, cuando x es muy grande, el denominador $1 - cdf(x)$ se vuelve muy pequeño y al dividir por él la discrepancia $cdf(x) - \bar{cdf}(x)$ se multiplica o amplifica de modo considerable, de modo que la contribución del tercer término se debe a la diferencia $cdf(x) - \bar{cdf}(x)$ y a dicho efecto multiplicador. Esto se observa en la mayoría de las figuras de los Gráficos 1 y 2, donde la discrepancia aumenta conforme lo hace x , y es un efecto inevitable a menos que $cdf(x) - \bar{cdf}(x)$ sea prácticamente nula. Esta observación, hasta donde sabemos, es nueva en la literatura, y puede contribuir a aclarar cuestiones suscitadas recientemente en la misma (Levy, 2009; Eeckhout, 2009), que tratan, entre otras cosas, de las limitaciones de utilizar únicamente métodos gráficos para discriminar entre distribuciones y de las implicaciones estadísticas que supone tomar logaritmos (véase también al respecto la Sección 5 de este trabajo).

4.3. Comparación en términos de *cumulative distribution functions*

Dedicamos esta Sección a comparar las distribuciones en términos de las cdf 's asociadas. En principio, cabría esperar que los resultados fueran similares a los del apartado anterior, pero veremos que esto no es exactamente cierto. En los Gráficos 3 y 4 mostramos las cdf 's correspondientes a los mismos casos en los que hemos ilustrado los *Zipf plots*.

De los gráficos mostrados se observa que, en general, el ajuste en cdf 's es aparentemente mejor para la lognormal que para la q -exponencial, cuando, recordemos, en términos de *Zipf plot*, a menudo lo hacía mejor la q -exponencial. Para explicar esta aparente paradoja es útil recurrir de nuevo a las expresiones (13) y (14). De ellas se deduce:

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

Empezamos razonando sólo para la distribución q -exponencial. Sabemos que su ajuste en rangos (véase la Sección 4.2) es muy bueno, salvo para las ciudades más pequeñas, lo cual quiere decir que $r(x) \cong \bar{r}(x)$ para prácticamente todos los puntos, con lo que (16) queda ahora:

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

Merece la pena detenerse en la ecuación (17), manteniendo en mente que se ha obtenido considerando que el ajuste en rangos es casi perfecto. El desajuste en cdf 's es tanto mayor cuanto mayor sea la diferencia $\bar{r}_0 - r_0$, y esta brecha es significativa en la q -exponencial, tal y como puede comprobarse de la información que se desprende del Cuadro 2 (recuérdese que \bar{r}_0 se identifica con el tamaño muestral). Además, la discrepancia en cdf 's crece con $1 - cdf(x)$, esto es, es tanto mayor cuanto más pequeño es x y tiende a desaparecer gradualmente conforme x aumenta. Todo esto muestra que en la q -exponencial la discrepancia en cdf 's es perfectamente compatible con un ajuste en rangos, salvo para las ciudades más pequeñas, perfecto; es más, es inevitable si, como ocurre en la realidad, $\bar{r}_0 - r_0 \neq 0$. No obstante, el ajuste por la q -exponencial mejora para los tres países conforme avanza el siglo XX⁴. Esto se debe a las razones aducidas y a que mejora la estimación de r_0 a lo largo del tiempo.

Para la lognormal, ¿se produce una situación similar? Veamos que no. Del Cuadro 1 se puede deducir que ahora $r_0 \cong \bar{r}_0$, por lo que (16) se reduce a:

⁴ Más detalles disponibles de los autores bajo petición.

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

Con lo cual a partir de (18) se deriva que cualquier desajuste en rangos se traslada, en la lognormal, de forma directa a un desajuste en cdf 's. Este resultado intuitivo no se da en la q -exponencial por las razones antes aducidas.

4.4. Contrastes estadísticos estándares

Los parámetros estimados en la Sección 4.1 para la lognormal y la q -exponencial merecen un contraste estadístico independiente para verificar la bondad del ajuste. En las subsecciones previas hemos hablado de criterios visuales (Zipf plots y cdf 's) para discriminar entre distribuciones, lo cual conlleva siempre un cierto grado de subjetividad. Además, la aproximación gráfica por medio de Zipf plots de la Sección 4.2 puede no darnos una información fidedigna acerca de la bondad del ajuste: en la literatura se conocen inconvenientes de los Zipf plots (Eeckhout, 2009). Por ello, en este subapartado presentamos contrastes estadísticos para comparar distribuciones de una forma más objetiva. Para el caso de la lognormal un contraste estándar es el de Kolmogorov-Smirnov, pero con un número de observaciones tan elevado como los de nuestros datos tal contraste tiende a rechazar sistemáticamente la hipótesis nula de lognormalidad a menos que el ajuste sea perfecto. Por otro lado, en el caso de la q -exponencial o Pareto generalizada no es tan estándar un contraste del citado tipo. En consecuencia, estamos buscando un contraste que no tienda a rechazar la hipótesis nula por el mero hecho de tener un alto número de observaciones y que sea aplicable a las dos distribuciones por igual.

El contraste de Wilcoxon *rank-sum test* satisface ambos requerimientos. Este test contrasta la hipótesis de que dos muestras independientes provienen de poblaciones con

una misma distribución. Su resultado no depende tanto del elevado tamaño muestral y es el mismo test independientemente de la distribución subyacente. Realizamos el test comparando los rangos empírico y estimado en cada caso. Mostramos los resultados de estos tests en los Cuadros 3 y 4.

Los resultados del contraste de Wilcoxon muestran que la hipótesis nula de la q -exponencial no puede rechazarse con un nivel de confianza del 5% en ninguno de los periodos del siglo veinte en España, Italia y USA. La lognormalidad tampoco puede rechazarse con un nivel de confianza del 5% en ninguno de los periodos del siglo veinte en España e Italia. En USA se observa una evolución temporal; las primeras décadas se rechaza la lognormal y el p-valor decrece con el tiempo, pero a partir de 1930 el p-valor comienza a crecer hasta que no puede rechazarse al 5% la distribución lognormal de 1960 en adelante. Si en lugar del 5% tomamos un nivel de confianza del 1% la hipótesis nula sólo se rechazaría para USA en 1920 y 1930. Estos últimos resultados se obtienen también en González-Val (2010) usando kernels.

5. Lognormal y q -exponencial. El enfoque de complementariedad

En la Sección anterior hemos comparado la adecuación de las distribuciones q -exponencial y lognormal a nuestros datos estudiando los *Zipf plots*, las *cdf*'s y el contraste de Wilcoxon. El resultado principal que podemos adelantar ahora ya es que ninguna de las dos es claramente mejor que la otra y que, atendiendo a los distintos criterios presentados, la lognormal es preferida (*cdf*'s) o lo es la q -exponencial (contrastos estadísticos y en ocasiones, *Zipf plots*). En cualquier caso, el espíritu de todo el apartado anterior intentaba enfrentar según diversos enfoques a ambas distribuciones, haciéndolas sustitutivas. En esta Sección el enfoque es bien diferente y vamos a intentar demostrar que lejos de ser rivales más bien se complementan entre sí.

En consecuencia, vamos a incidir en dicha complementariedad. Primero veamos que uno de los resultados principales de Eeckhout (2004) sobre la concavidad del *Zipf plot* de la lognormal, también se aplica al caso de la q -exponencial. En efecto, en tal caso la expresión del rango respecto de la población está dada por (5). Para expresar las variables en logaritmos llamamos $y = \ln x$, con lo que $x = e^y$. De este modo:

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

Así, tenemos

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

que es la negativa de la *hazard rate* $h(y)$. Obteniendo la segunda derivada, se tiene:

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

que es estrictamente negativa para todo y , con lo que $\ln r(y)$ es estrictamente cóncava para todos los valores de la variable y y el correspondiente *Zipf plot* es a su vez cóncavo. En definitiva, se trata de un resultado teórico que sostiene que si el proceso generador de datos subyacente sigue una q -exponencial (o una lognormal como demostró Eeckhout, 2004) el exponente de Pareto decrece con el tamaño muestral. En otras palabras, es una primera prueba de la no rivalidad o equivalencia, al menos en este punto, de ambas distribuciones.

Una cuestión clave que surge al estudiar la q -exponencial y la lognormal, es que si se considera la población de las ciudades en niveles absolutos, la distribución empírica de probabilidad es una curva decreciente y convexa como es la función de

densidad de la q -exponencial, mientras que si se toma el logaritmo de la población, la distribución empírica de probabilidad tiene forma de campana como la función de densidad de la lognormal. Este resultado es algo contraintuitivo y merece la pena que nos detengamos en ello.

Supongamos que ordenamos de menor a mayor población los núcleos urbanos de nuestros datos. Si hacemos un histograma de los mismos resulta una gráfica decreciente conforme la población aumenta (Gráfico 5, datos de España en 1900). Como es bien sabido, un histograma valora las frecuencias asociadas a intervalos de anchura constante en el eje de abscisas. En cambio, si construimos el histograma del logaritmo de la población (Gráfico 6, mismos datos), también estos se cuentan en frecuencias respecto de intervalos de anchura constante, pero ahora en logaritmos; sin embargo, ¿qué significa esto en niveles? Sea δ esta anchura constante, y los extremos inferior y superior de uno de dichos intervalos $\ln x_j$ y $\ln x_{j+1}$ respectivamente. Por definición, $\ln x_{j+1} - \ln x_j = \delta$ o, lo que es lo mismo, $x_{j+1} = x_j e^\delta$. Generalizando, $x_{j+1} = x_j e^\delta = x_{j-1} e^{2\delta} = x_1 e^{j\delta}$, siendo x_1 el extremo inferior del primer intervalo, que no puede ser cero. Esto indica que los extremos superiores de los intervalos, en niveles, siguen una progresión geométrica de razón e^δ . Debe destacarse que este razonamiento es válido para cualquier variable numérica que se use alternativamente en niveles y en logaritmos neperianos.

Este hecho explica por qué al tomar logaritmos resulta una forma de campana: los primeros intervalos son muy estrechos y además habrá muy pocos casos incluidos en ellos; luego, conforme los intervalos se ensanchan como consecuencia de la progresión geométrica, el número de casos en cada intervalo crece de forma considerable y el gráfico aumenta. Llegará un momento en el que, aunque los intervalos sean muy

anchos, el número de casos sea, por motivos obvios (por ejemplo, ciudades muy grandes, digamos, con más de 500.000 habitantes), muy pequeño, con lo que el gráfico disminuye. En el proceso se ha llegado a un máximo, y en conjunto se obtiene una forma de campana.

Estos dos resultados muestran que la q -exponencial y la lognormal son complementarias más que una sustitutiva de la otra. Ambas proporcionan una aproximación a los datos muy notable y las dos mejoran la descripción de la distribución de Pareto al considerar todos los núcleos de población.

6. Conclusiones

Cualquier fenómeno o, mejor, variable cuantificable, esto es, que adopta valores numéricos de forma en principio continua, es susceptible de ser estudiada desde un punto de vista estadístico. Este punto de vista a que hacemos referencia tiene que ver con el análisis de la distribución que sigue la variable en cuestión. Y esto se puede hacer con cualquier tipo de datos sea cual sea su disciplina, siempre y cuando cumplan los requisitos mencionados cuatro líneas más arriba. La distribución del tamaño de las ciudades no es ajena a este hecho y cuenta con una amplia tradición en la literatura de Economía Urbana. Este trabajo se enmarca en esta corriente.

Pero no sólo se trata de una mera curiosidad estadística. También hay Economía o, mejor dicho, también genera profundas implicaciones económicas el estudio de la distribución del tamaño de las ciudades. Por ejemplo, una estructura urbana de ciudades muy iguales en población invita a un tratamiento a su vez igualitario por parte del organismo público pertinente en lo relativo a inversiones en infraestructuras de transporte, educación o sanidad. Por el contrario, grandes disparidades en tamaño

requieren de políticas que tiendan a la convergencia y a la consecución de la cohesión territorial.

En este trabajo se ha llevado a cabo, hasta donde conocemos, un examen minucioso de una función de densidad novedosa en Economía Urbana, como es la denominada q-exponencial. Una primera aportación tiene que ver con el método de estimación de sus parámetros clave. En efecto, en lugar de suponer valores predeterminados para ellos, se aproximan por un desarrollo en serie de Taylor que nos permite obtener unos valores iniciales para la posterior estimación, sin restricciones ni apriorismos, por mínimos cuadrados no lineales.

Por otra parte, desde el trabajo pionero de Eeckhout (2004) se pusieron de manifiesto los peligros de considerar sólo los núcleos más grandes, esto es, exclusivamente la cola superior de la distribución. A su vez, si los datos lo permiten, el análisis de la distribución del tamaño de las ciudades debe realizarse como un análisis de largo plazo. Con ambas consideraciones como premisas, en este artículo se utilizan datos censales para todo el siglo veinte, de diez en diez años, y todas las ciudades de tres países: USA (de 10.600 a 19.300 núcleos, según el año), España (unos 8.000 núcleos) e Italia (unos 8.000 núcleos). Emplear bases de datos tan amplias y con un horizonte temporal tan vasto otorga, sin duda, robustez a los resultados.

¿Y cuáles son los resultados principales? Todos ellos ponen en relación la distribución que desde Eeckhout (2004) se postula como la mejor para estudiar la distribución del tamaño de las ciudades sin *truncation point*, la lognormal, con la que presentamos en este papel, la q-exponencial. Son fundamentalmente tres.

En primer lugar, el ajuste de ambas (lognormal y q-exponencial) en términos de rangos en un *Zipf plot* es extremadamente bueno; si acaso ligerísimamente mejor el de

la última. Sin embargo, se demuestra estadísticamente que este mejor ajuste en rangos de la q -exponencial genera inevitablemente que, ahora en términos de *cumulative density functions*, la lognormal sea superior.

En segundo lugar, el contraste de Wilcoxon muestra que la hipótesis nula de la q -exponencial no puede rechazarse con un nivel de confianza del 5% en ninguno de los periodos del siglo veinte en España, Italia y USA. La lognormalidad tampoco puede rechazarse al 5% en ninguno de los periodos del siglo veinte en España e Italia; en USA solo de 1960 en adelante. Luego, en el capítulo de los contrastes estadísticos estándares, para nuestra muestra, las dos distribuciones funcionan bien, especialmente la q -exponencial.

Y tres, la pregunta relevante que se puede plantear un planificador urbano, un economista o geógrafo que aborda el estudio de la completa distribución del tamaño de las ciudades de una determinada zona o país es la siguiente: ¿este trabajo recomienda que se debe usar la lognormal o la q -exponencial? La respuesta es simple. Desde nuestro punto de vista ambas son complementarias: si se trabaja con poblaciones en niveles es mejor la q -exponencial; si, como es habitual, se toma el logaritmo del tamaño, las dos son adecuadas, quizá ligerísimamente mejor la q -exponencial; pero también es de justicia admitir que la lognormal ofrece buenas prestaciones y presenta dos ventajas adicionales: históricamente ha sido mucho más utilizada en la literatura (no sólo de Economía Urbana sino de cualquier disciplina) y el método de estimación de sus parámetros es más sencillo conceptual y computacionalmente que el de la q -exponencial. En cualquier caso, es un interrogante que debe responder cada investigador y sobre el cual este trabajo ha pretendido aportar alguna luz.

Por último, no queremos acabar este trabajo sin hacer una breve reflexión sobre el proceso teórico de crecimiento que subyace en la distribución q -exponencial. Son

bien conocidos los vínculos entre la lognormal y la distribución de Pareto con la ley de Gibrat, incluso, con una versión generalizada de la misma, la doble Pareto lognormal (Redd, 2002; véase la nota al pie 1). ¿Qué hay detrás de la q-exponencial? Se trata de una cuestión importante que es necesario responder y que constituye, por tanto, un excelente campo de investigación futura.

Referencias

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

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

Black, D. y J. V. Henderson (2003), “Urban evolution in the USA”, *Journal of Economic Geography* 3, 343-372.

Bosker, M., S. Brakman, H. Garretsen y 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.

Cheshire, P. (1999). “Trends in sizes and structure of urban areas”, *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. y S. Magrini (2006). “Population growth in European cities: weather matters-but only nationally”, *Regional Studies* 40(1), 23-37.

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

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

Giesen, K., A. Zimmermann, y J. Suedekum (2009). "The size distribution across all cities - Zipf, lognormal, double Pareto, or what?" Mimeo.

González-Val, R. (2010), "The evolution of US city size distribution from a long term perspective (1900-2000)", *Journal of Regional Science*, forthcoming.

González-Val, R., L. Lanasa y 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. y J. R. Wallis (1987). "Parameter and quantile estimation for the generalized Pareto distribution", *Technometrics* 29, 339-349.

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

Ioannides, Y. M. Y 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.

- 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 y E. K. Lenzi (2001). "q-exponential distribution in urban agglomeration", *Physical Review E* 65, (017106) 1-3.
- Reed, W. (2002). "On the rank-size distribution for human settlements", *Journal of Regional Science* 42, 1-17.
- Rosen, K. y M. Resnick (1980). "The size distribution of cities: an examination of the Pareto law and primacy", *Journal of Urban Economics* 8, 165-186.
- 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.

Cuadro 1. Valores de \hat{r}_0 , $\hat{\mu}$ y $\hat{\sigma}^2$ para USA. Desviaciones típicas estimadas entre paréntesis

USA- lognormal	\hat{r}_0	$\hat{\mu}$	$\hat{\sigma}^2$	R^2
1900	10374,27 (6,871518)	6,648714 (0,0122517)	1,261147 (0,0086632)	0,9954
1910	13805,79 (8,005956)	6,64682 (0,0108739)	1,292809 (0,007689)	0,9953
1920	15126,01 (8,510406)	6,674667 (0,0105985)	1,318693 (0,0074943)	0,9951
1930	16104,56 (8,988392)	6,692269 (0,0109193)	1,401552 (0,0077211)	0,9949
1940	16347,44 (8,587507)	6,775817 (0,0110714)	1,431982 (0,0078287)	0,9954
1950	16771,11 (7,3058)	6,837732 (0,0114793)	1,501686 (0,0081171)	0,9968
1960	17698,55 (6,640407)	6,923707 (0,011952)	1,605794 (0,0084513)	0,9975
1970	18153,87 (6,081665)	7,004047 (0,0122595)	1,666934 (0,0086688)	0,9979
1980	18576,96 (5,860968)	7,114369 (0,012081)	1,661872 (0,0085425)	0,9981
1990	18799,24 (5,308348)	7,0984 (0,0126035)	1,742746 (0,008912)	0,9985
2000	18968,83 (5,17997)	7,18272 (0,0128295)	1,782151 (0,0090718)	0,9986

Cuadro 1 (contin.). Valores de \hat{r}_0 , $\hat{\mu}$ y $\hat{\sigma}^2$ para España. Desviaciones típicas estimadas entre paréntesis

España- lognormal	\hat{r}_0	$\hat{\mu}$	$\hat{\sigma}^2$	R^2
1900	7610,979 (4,281377)	6,96552 (0,0120334)	1,062761 (0,0085089)	0,9975
1910	7614,829 (4,280091)	7,012887 (0,0122144)	1,079165 (0,0086369)	0,9975
1920	7621,96 (4,174949)	7,025287 (0,0125281)	1,107306 (0,0088587)	0,9977
1930	7684,351 (4,13726)	7,05515 (0,0128743)	1,142478 (0,0091035)	0,9977
1940	7706,974 (4,320335)	7,062808 (0,013301)	1,181922 (0,0094052)	0,9975
1950	7711,78 (4,173386)	7,086039 (0,0135308)	1,202717 (0,0095677)	0,9977
1960	7717,208 (4,219623)	7,033353 (0,0143045)	1,272214 (0,0101148)	0,9976
1970	7791,69 (3,71842)	6,82857 (0,0161668)	1,442019 (0,0114316)	0,9982
1981	7889,339 (3,379101)	6,631256 (0,0181122)	1,623441 (0,0128072)	0,9985
1991	7916,954 (3,454906)	6,534098 (0,0190795)	1,714716 (0,0134913)	0,9985
2001	7892,812 (3,772936)	6,540983 (0,0195229)	1,754564 (0,0138048)	0,9982

Cuadro 1 (contin.). Valores de \hat{r}_0 , $\hat{\mu}$ y $\hat{\sigma}^2$ para Italia. Desviaciones típicas estimadas entre paréntesis

Italia- lognormal	\hat{r}_0	$\hat{\mu}$	$\hat{\sigma}^2$	R^2
1901	7676,142 (2,753235)	7,78953 (0,0104247)	0,9154127 (0,0073713)	0,999
1911	7672,715 (2,648972)	7,843163 (0,0106065)	0,9313805 (0,0074999)	0,9991
1921	8050,636 (2,598348)	7,835906 (0,0106981)	0,9628311 (0,0075647)	0,9992
1931	8054,081 (2,38133)	7,838977 (0,0110142)	0,9912791 (0,0077882)	0,9993
1936	8062,379 (2,237987)	7,84206 (0,0112179)	1,009614 (0,0079323)	0,9994
1951	8067,352 (2,147735)	7,894767 (0,0116589)	1,049305 (0,0082441)	0,9994
1961	8056,921 (2,463067)	7,84784 (0,0122279)	1,100507 (0,0086464)	0,9992
1971	8041,767 (2,563667)	7,788053 (0,0131672)	1,185047 (0,0093106)	0,9992
1981	8045,297 (2,102519)	7,792515 (0,0138381)	1,245428 (0,009785)	0,9994
1991	8057,702 (1,721827)	7,795891 (0,0142457)	1,282116 (0,0100733)	0,9996
2001	8068,104 (1,473446)	7,803148 (0,0145179)	1,306614 (0,0102657)	0,9997

Cuadro 2. Valores de \hat{r}_0 , \hat{q} y \hat{a} para USA. Desviaciones típicas estimadas entre paréntesis

USA q - exponencial	\hat{r}_0	\hat{q}	\hat{a}	R^2
1900	12658,84 (12,37365)	1,603419 (0,0032688)	0,0027559 (0,0000141)	0,999
1910	16939 (14,08978)	1,694559 (0,0027492)	0,0031086 (0,0000133)	0,9991
1920	18457,28 (14,15269)	1,709492 (0,0025757)	0,0030609 (0,0000122)	0,9991
1930	19340,12 (14,61102)	1,818757 (0,0026694)	0,003332 (0,0000133)	0,999
1940	19454,64 (13,47561)	1,8933 (0,0025121)	0,0032611 (0,0000121)	0,9991
1950	19241,77 (9,194967)	1,975214 (0,0019319)	0,003091 (8,51e-06)	0,9995
1960	19785,46 (6,949107)	2,144436 (0,0015544)	0,0031838 (6,79e-06)	0,9997
1970	19911,43 (5,661598)	2,221329 (0,0013473)	0,0030247 (5,45e-06)	0,9998
1980	20338,61 (5,206677)	2,228124 (0,00122)	0,0027042 (4,41e-06)	0,9998
1990	20191,54 (4,104436)	2,325746 (0,001043)	0,0028769 (3,98e-06)	0,9999
2000	20264,44 (3,904862)	2,393507 (0,0010135)	0,0027802 (3,62e-06)	0,9999

Cuadro 2 (contin.). Valores de \hat{r}_0 , \hat{q} y \hat{a} para España. Desviaciones típicas estimadas entre paréntesis

España q - exponential	\hat{r}_0	\hat{q}	\hat{a}	R^2
1900	9768,342 (8,919785)	1,539363 (0,0025706)	0,0019318 (8,32e-06)	0,9995
1910	9737,746 (8,536555)	1,569519 (0,0024854)	0,0019022 (7,89e-06)	0,9995
1920	9642,943 (7,541392)	1,606816 (0,0022763)	0,0019348 (7,27e-06)	0,9996
1930	9599,185 (7,225424)	1,658583 (0,0022592)	0,0019577 (7,20e-06)	0,9996
1940	9545,302 (7,722284)	1,698713 (0,0024931)	0,0020182 (8,09e-06)	0,9996
1950	9479,403 (6,729436)	1,729926 (0,0022309)	0,0020183 (7,19e-06)	0,9997
1960	9344,968 (6,30277)	1,834201 (0,0022176)	0,0023419 (8,12e-06)	0,9997
1970	8881,374 (4,450282)	1,99013 (0,0020147)	0,0030433 (8,84e-06)	0,9998
1981	8627,245 (3,438239)	2,192375 (0,0018711)	0,0041819 (0,0000106)	0,9998
1991	8607,263 (3,54808)	2,359453 (0,002003)	0,0053741 (0,0000144)	0,9998
2001	8644,453 (4,134219)	2,478272 (0,0022863)	0,0061204 (0,0000188)	0,9997

Cuadro 2 (contin.). Valores de \hat{r}_0 , \hat{q} y \hat{a} para Italia. Desviaciones típicas estimadas entre paréntesis

Italia q - exponential	\hat{r}_0	\hat{q}	\hat{a}	R^2
1901	9217,565 (8,965359)	0,9955913 (0,003305)	0,0003635 (2,19e-06)	0,9989
1911	9128,744 (8,35108)	1,007638 (0,0031358)	0,0003498 (1,98e-06)	0,9991
1921	9522,924 (7,643256)	1,068256 (0,0027875)	0,0003819 (1,88e-06)	0,9993
1931	9404,778 (6,948147)	1,108075 (0,0026432)	0,0003946 (1,80e-06)	0,9994
1936	9313,638 (6,394332)	1,118255 (0,0025241)	0,0003923 (1,69e-06)	0,9994
1951	9174,556 (5,479968)	1,52508 (0,0022996)	0,0003806 (1,46e-06)	0,9995
1961	9114,941 (5,332345)	1,224953 (0,0023121)	0,0004378 (1,64e-06)	0,9995
1971	9008,062 (4,735081)	1,369771 (0,0021592)	0,000551 (1,84e-06)	0,9996
1981	8863,086 (3,715982)	1,470554 (0,0018076)	0,0005982 (1,62e-06)	0,9998
1991	8751,159 (3,110422)	1,515978 (0,0015959)	0,000609 (1,43e-06)	0,9998
2001	8683,71 (2,853292)	1,548455 (0,0015113)	0,0006143 (1,34e-06)	0,9998

Nota: Por definición $\hat{q} > 1$. La estimación puntual para Italia en el año 1901 no satisface esta restricción por milésimas, aunque la condición se cumpliría al considerar la estimación por intervalo.

Cuadro 3. Wilcoxon (*Rank-sum test*) test para la q -exponential

USA											
Año	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
p-valor	0,4662	0,4519	0,4360	0,4132	0,4718	0,619	0,7635	0,845	0,8948	0,9829	0,9620
ESPAÑA											
Año	1900	1910	1920	1930	1940	1950	1960	1970	1981	1991	2001
p-valor	0,7673	0,7845	0,8252	0,8588	0,8372	0,8817	0,9123	0,9448	0,9973	0,9289	0,8912
ITALIA											
Año	1901	1911	1921	1931	1936	1951	1961	1971	1981	1991	2001
p-valor	0,9077	0,529	0,5615	0,6093	0,6242	0,6619	0,6587	0,7049	0,7869	0,8245	0,8536

Ho: La distribución sigue una q -exponencial

Cuadro 4. Wilcoxon (*Rank-sum test*) test para la lognormal

USA											
Año	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
p-valor	0,0252	0,017	0,0078	0,0088	0,0208	0,0464	0,1281	0,1836	0,2538	0,323	0,4168
ESPAÑA											
Año	1900	1910	1920	1930	1940	1950	1960	1970	1981	1991	2001
p-valor	0,5953	0,6144	0,6233	0,6525	0,4909	0,5792	0,6049	0,522	0,5176	0,622	0,7212
ITALIA											
Año	1901	1911	1921	1931	1936	1951	1961	1971	1981	1991	2001
p-valor	0,2081	0,2205	0,2352	0,291	0,2864	0,3118	0,2589	0,272	0,382	0,4671	0,5287

Ho: La distribución sigue una lognormal

Gráfico 1. *Zipf plots*, empírico y teórico con la q -exponencial.

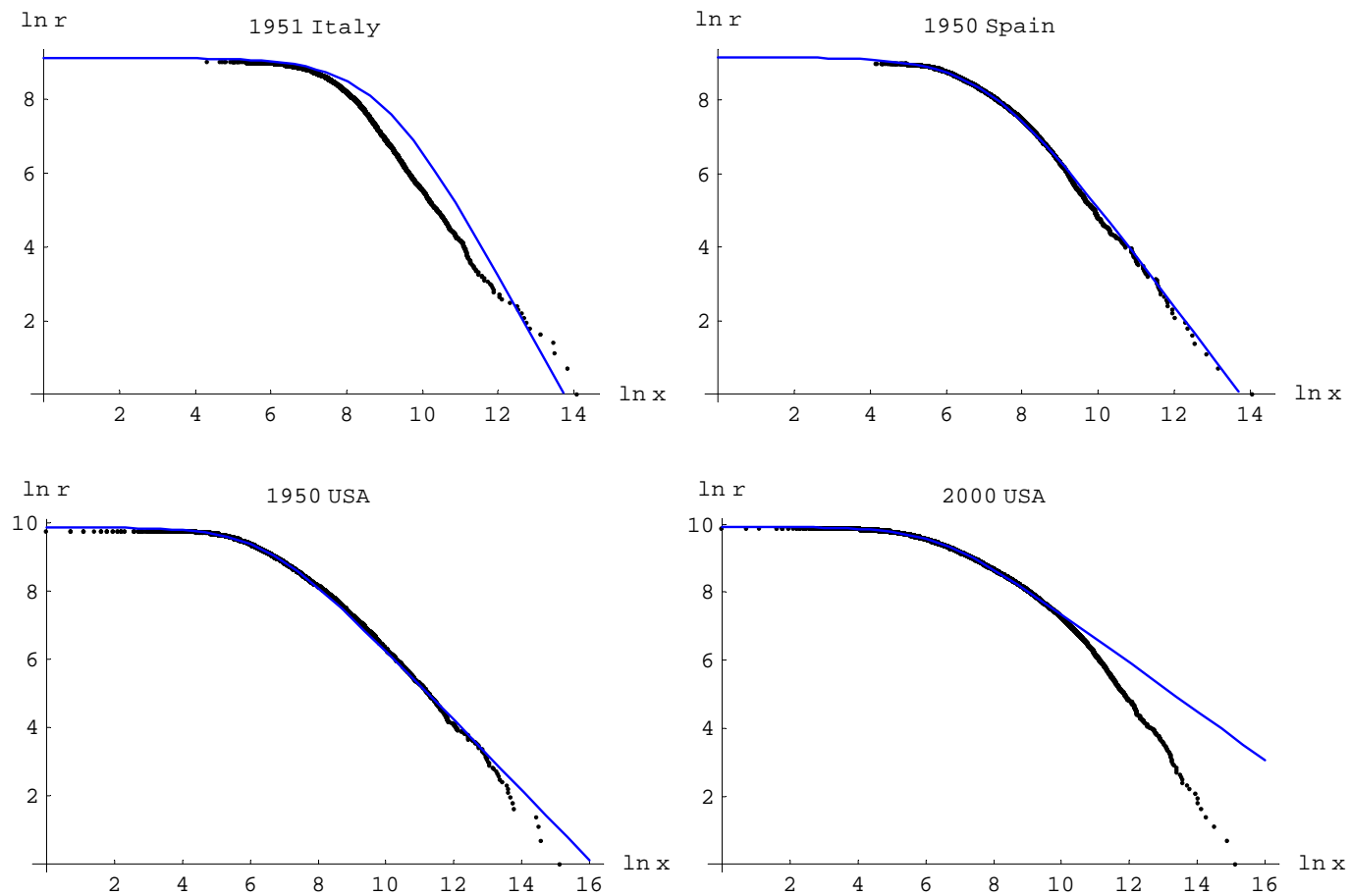


Gráfico 2. *Zipf plots*, empírico y teórico con la lognormal.

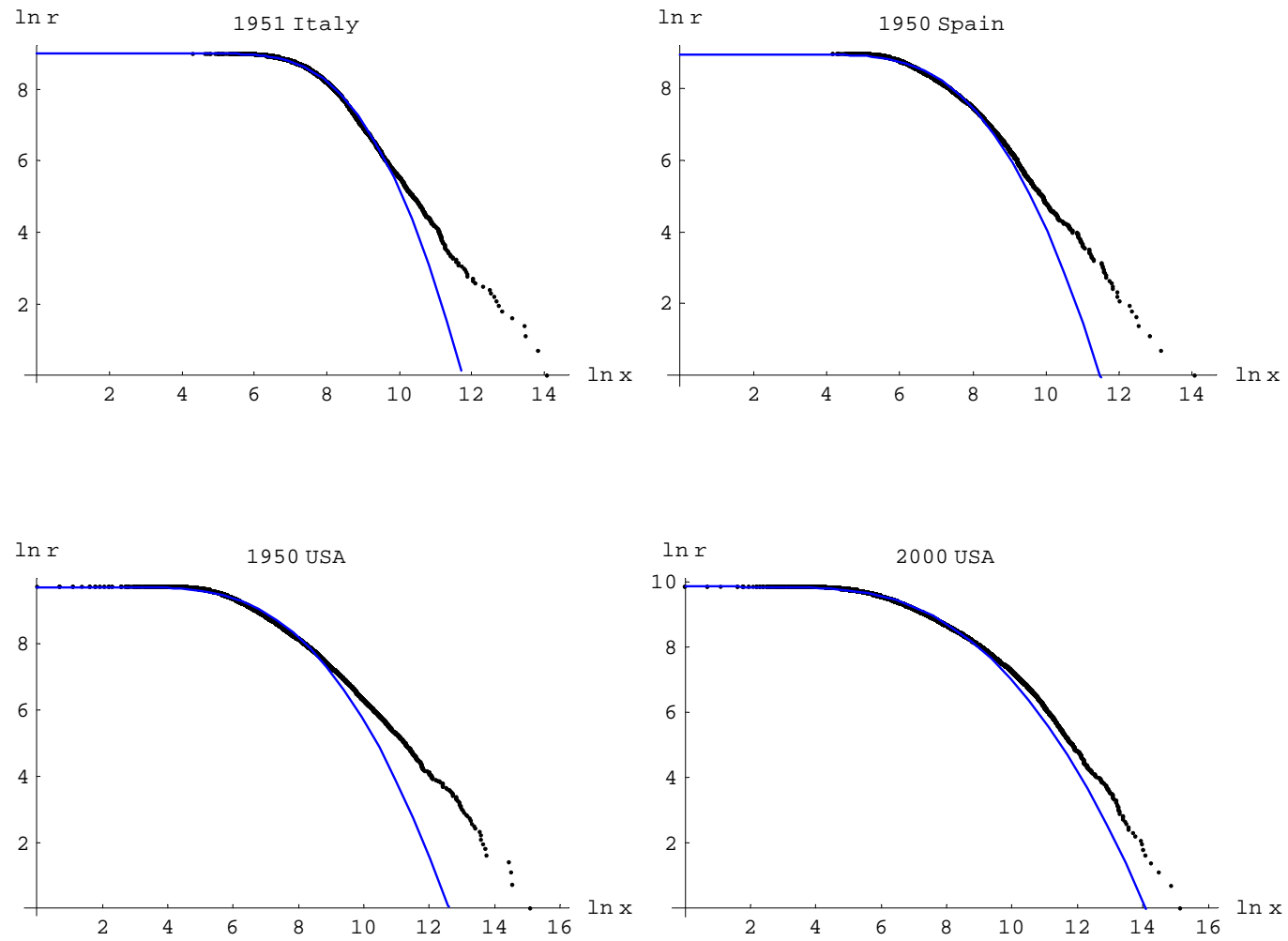


Gráfico 3. *cdf plots*, empírico y teórico con la q -exponencial

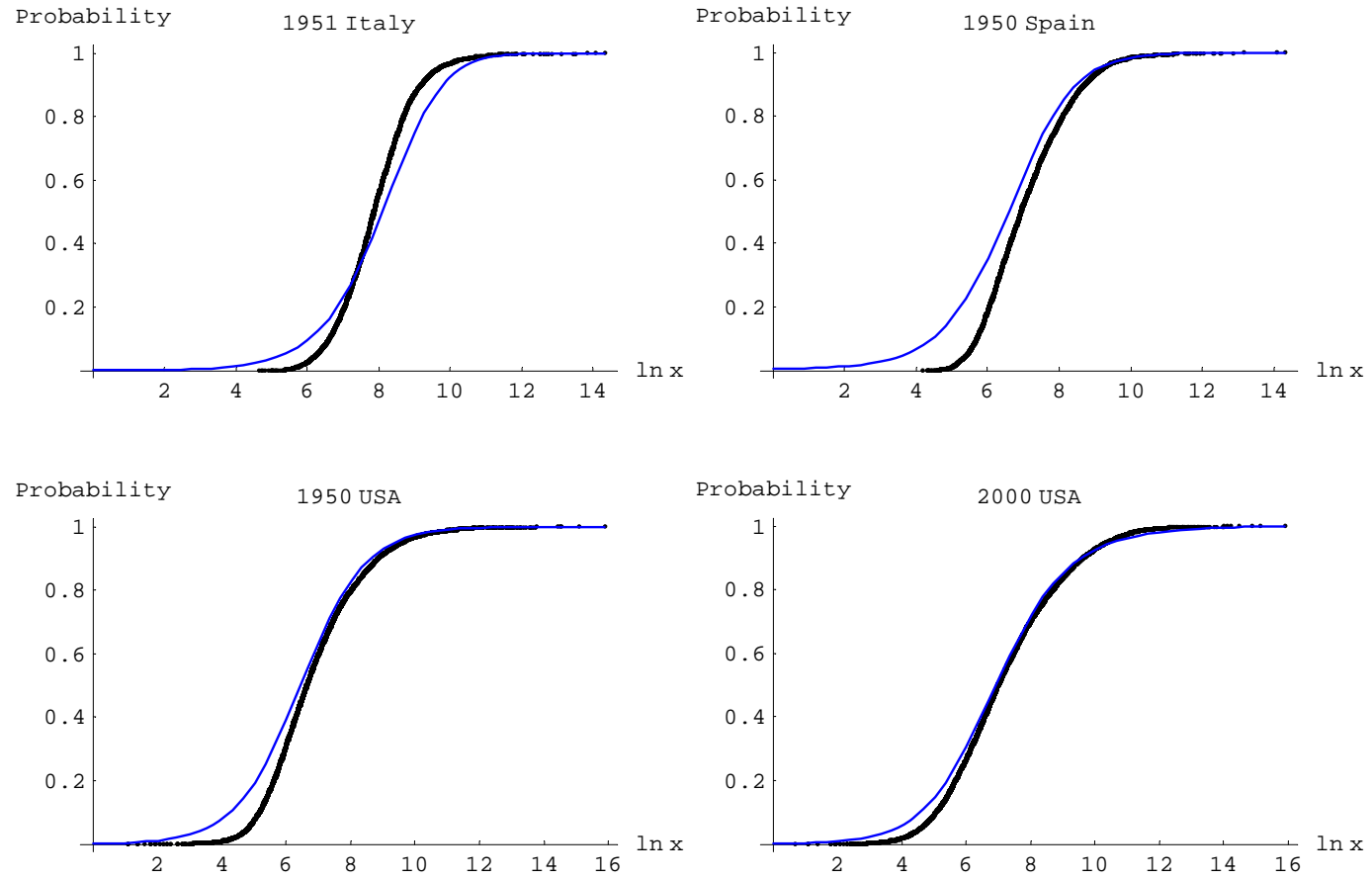


Gráfico 4. *cdf plots*, empírico y teórico con la lognormal

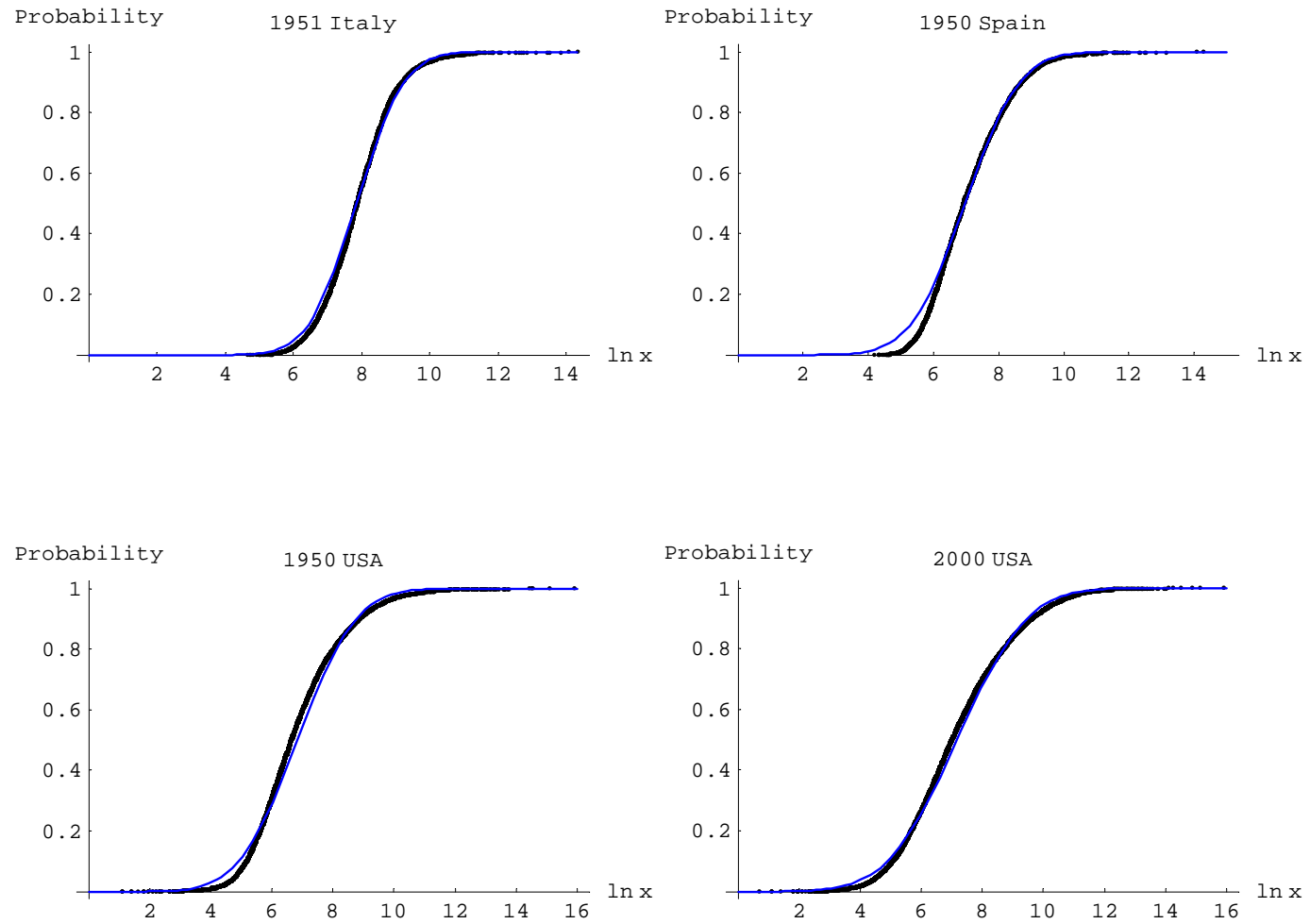


Gráfico 5. Histograma de ciudades de España en 1900. Población en niveles

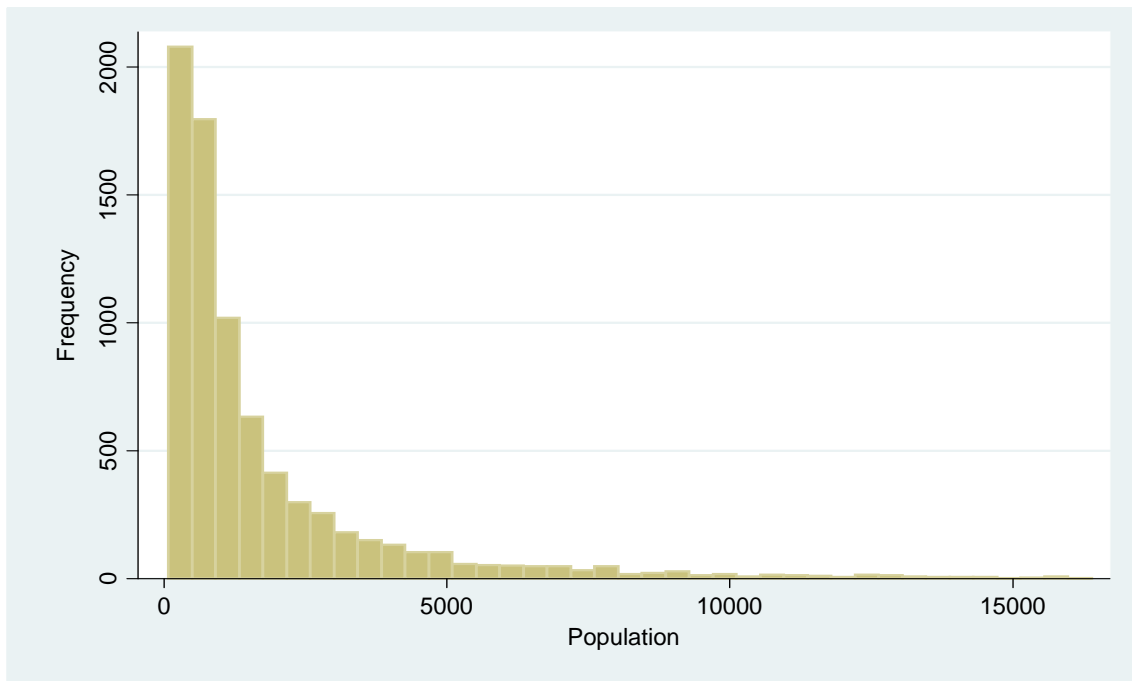


Gráfico 6. Histograma de ciudades de España en 1900. Población en logaritmos

