Centro de Estudios Económicos
www.colmex.mx
El Colegio de México, A.C.

Serie documentos de trabajo

LAS CIUDADES MEXICANAS NO SIGUEN LA LEY DE ZIPF

Carlos M. Urzúa

DOCUMENTO DE TRABAJO

Núm. VIII - 2000

Las ciudades mexicanas no siguen la ley de Zipf*

Carlos M. Urzúa**

El Colegio de México

Ultima revisión: julio de 2000

*Agradezco los comentarios de un dictaminador anónimo. Este trabajo forma parte del proyecto *Geografía y Desarrollo Económico* encargado en 1999 por el Banco Interamericano de Desarrollo a El Colegio de México.

**Profesor-investigador del Centro de Estudios Económicos de El Colegio de México.

RESUMEN

En este artículo se examina la hipótesis de que las ciudades mexicanas siguen la ley de Zipf, la cual establece que si se ordenan las ciudades de un país (o región) de acuerdo con su tamaño poblacional, entonces el rango de cada ciudad multiplicado por su tamaño produce siempre la misma constante. En este trabajo se afirma que tal ley no es aplicable al caso mexicano, y también se argüye que muchos estudios que pretenden mostrar que tal ley se aplica para otros países carecen de fundamentos estadísticos sólidos.

Introducción

La literatura sobre el tamaño de las ciudades abunda en referencias que pretenden mostrar, de una manera u otra, la validez de la llamada ley de Zipf (1949), así como algunas generalizaciones de ella. Dado que tal hipótesis tiene ya medio siglo de vida, es ciertamente notable el hecho de que haya permanecido vigente por tanto tiempo en la literatura [véanse, por ejemplo, Clark (1967), Rosen y Resnick (1980) y Krugman (1996)]. No obstante, uno de los dos propósitos de este artículo es el advertir que tal ley no se aplica al caso mexicano. El otro objetivo es el señalar que la mayoría de los estudios urbanos que pretenden haber encontrado ejemplos que validan la ley de Zipf no son robustos desde un punto de vista estadístico.

La ley de Zipf

¿Cuáles son las predicciones de dicha ley? Supongamos que ordenamos las n mayores ciudades (o, mejor aún, áreas metropolitanas) de México, usando como criterio el número de habitantes en cada una de ellas:

$$x_{(1)} \geq x_{(2)} \geq \ldots \geq x_{(r)} \geq \ldots \geq x_{(n)}$$

donde r denota el rango de la ciudad y $x_{(r)}$ denota su tamaño. La ley de Zipf asegura que si graficáramos el rango correspondiente a cada ciudad contra el tamaño de cada una de las n ciudades, entonces nos encontraríamos con una hipérbola rectangular perfecta. Es decir, la ley establece que el producto de r por $x_{(r)}$ sería igual a la misma constante en todos los casos. Así pues, por ejemplo, la ley de Zipf implica que si consideramos la dos ciudades más pobladas (rango igual a 1 y 2), entonces la ciudad más grande tiene que tener el doble de habitantes comparada con la segunda población.

Puesto de otra manera, la ley de Zipf afirma que si graficáramos el logaritmo natural del rango contra el logaritmo del tamaño de la ciudad, entonces lo que aparecería sería una línea recta con pendiente igual a –1; esto es porque de la ecuación $r\, x_{(r)} = c$ se sigue que $\ln(r) + \ln(x_{(r)}) = \ln(c)$. De hecho, las más de las veces los investigadores que afirman haber encontrado otro ejemplo donde dicha ley se cumple presentan como "prueba", aparte de una gráfica, los resultados de la siguiente regresión estimada por mínimos cuadrados ordinarios:

$$\ln r = b_1 + b_2 \ln x_{(r)} + e_r$$

Para después afirmar que el estimado de la pendiente es "cercano" a –1. Sin embargo, pocos han notado que tal procedimiento es muy ineficiente (y, por tanto, puede llevar fácilmente a inferencias erróneas). La razón es que la distribución del error en la regresión anterior no es ni lejanamente normal, puesto que la variable que está en el lado izquierdo, el logaritmo del rango, es una variable discreta.

Como se explica ampliamente en Rapoport (1978), por ejemplo, la única manera eficiente de probar la ley de Zipf es estableciéndola en términos probabilísticos: en lugar de establecer una relación rango-tamaño, lo que se requiere es encontrar una relación frecuencia-tamaño. Una vez hecho esto, uno puede mostrar (véase, por ejemplo, Urzúa,

2000) que la ley de Zipf, puesta en términos probabilísticos, simplemente afirma que el tamaño de los objetos sigue una distribución de Pareto con exponente igual a uno.

Es decir, la densidad implícita en la ley de Zipf está dada por:

$$f(x) = \frac{a}{m}\left(\frac{x}{m}\right)^{-(a+1)}$$

con el parámetro $\alpha = 1$ ($x \geq \mu$). Habiendo encontrado tal densidad, en Urzúa (2000) se propone un estadístico localmente óptimo (en un sentido explicado en ese documento) que puede ser usado para probar tal hipótesis de Zipf. El estadístico es el siguiente (el símbolo “LMZ” pretende denotar el hecho de que la prueba para la ley de Zipf está basado en el estadístico de multiplicadores de Lagrange):

$$LMZ = 4n[z_1^2 + 6z_1z_2 + 12z_2^2]$$

donde las dos variables dentro del paréntesis están dadas por:

$$z_1 = 1 - \frac{1}{n}\sum_{i=1}^{n} \ln \frac{x_i}{x_{(n)}}$$

$$z_2 = \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n}\frac{x_{(n)}}{x_i}$$

En las expresiones anteriores n es el tamaño de la muestra, x_i es el tamaño del i-ésimo objeto (es irrelevante que estén ordenados o no), y $x_{(n)}$ es el tamaño de la ciudad más pequeña en la muestra.

Bajo la hipótesis nula de que la ley de Zipf se cumple, el estadístico *LMZ* propuesto más arriba sigue, asintóticamente, una distribución ji-cuadrada con dos grados de libertad. Así pues, dada una muestra específica, para probar la hipótesis de Zipf basta seguir los siguientes tres pasos: Primero, hay que identificar el menor tamaño en la muestra ($x_{(n)}$). Segundo, hay que calcular el valor del estadístico *LMZ* haciendo uso de todos los datos. Y tercero, dado un cierto nivel de significancia, hay que comparar el valor de *LMZ* con el valor crítico correspondiente. Si aquél es mayor que éste, entonces uno debe rechazar la hipótesis nula (la hipótesis de que la ley de Zipf se cumple), de otra manera uno no puede rechazarla.

Podemos ser más específicos respecto a este último punto: si se usa un nivel de significancia del 10%, como es común en los estudios sobre distribuciones, entonces el valor crítico que debe usarse es aproximadamente 4.61. Este es el valor crítico de una ji-cuadrada con dos grados de libertad, hacia donde, como ya mencionamos, converge asintóticamente la distribución del estadístico. Para tener una mayor confiabilidad en la inferencia, puede usarse el cuadro 1 en Urzúa (2000), donde, por ejemplo, se establece que si la muestra es de un tamaño igual a 50, entonces el valor crítico es 4.49.

Aplicaciones al caso mexicano

Habiendo establecido un estadístico confiable para examinar la hipótesis de Zipf, tenemos ahora que decidir cuáles serán las muestras para el caso mexicano. Primero, como parecería natural emplear solamente los datos derivados de censos oficiales y confiables, hemos elegido entonces los cuatro censos levantados en 1960, 1970, 1980 y 1990. Segundo, en la selección de las ciudades (o, más bien, áreas metropolitanas) más importantes de México, utilizamos la selección de 81 áreas hecha por la Presidencia de la República (1994). Haciéndolo así, no podemos ser acusados de haber elegido de manera selectiva el número de ciudades para verificar o descalificar la hipótesis de Zipf. Las muestras resultantes son presentadas en el cuadro 1.

Como se observa en ese cuadro, *no* incluimos en nuestro examen a la Ciudad de México y sus zonas aledañas. La razón es que si la incluyéramos, entonces la ley de Zipf estaría condenada a ser rechazada de primera instancia, pues tal ley sostiene, en particular, que el tamaño de la segunda ciudad más grande debe ser aproximadamente igual a la mitad de la primera. Esto, sin embargo, está muy lejos de ser verdad: la Ciudad de México y sus zonas aledañas tenían en 1960, aproximadamente, una población seis veces mayor que la de Guadalajara, y aún hoy la relación es, aproximadamente, de cuatro a uno.

Antes de hacer un examen robusto de los datos en el cuadro 1, bien vale la pena preguntarse acerca de los resultados que se obtendrían si se usan los procedimientos de "prueba" de la ley de Zipf que han sido empleados hasta este momento en la literatura. A manera de ilustración, considere la muestra correspondiente a 1990. La gráfica 1 presenta la curva que resulta de graficar la población de cada ciudad contra su rango. ¿Es esa curva una hipérbola rectangular (es decir, simétrica) como lo establecería la ley de Zipf? Pues depende del cristal con que se mire.

Dada la ambigüedad presente en toda inspección ocular, es también común que los investigadores presenten los resultados de la regresión que se obtiene al correr el logaritmo natural del rango contra el logaritmo natural de la población (véase la sección anterior). Si

nosotros hacemos también esto para el año de 1990, los resultados que se obtienen son los siguientes:

$$\ln r = 14.93 - .95 \ln x_{(r)}$$

donde los errores estándares para los dos estimadores de los coeficientes son, respectivamente, .38 y .03, y donde la R^2 toma el valor de .92.

Si aceptáramos el procedimiento anterior como un método de inferencia confiable (lo cual, se insiste, no lo es), ¿qué podríamos inferir de esa regresión? Dado que la hipótesis nula, la ley de Zipf, establece que el coeficiente del logaritmo de la población debe ser igual a –1, entonces procederíamos a calcular el valor del correspondiente estadístico *t* el cual es en nuestro caso igual a –1.67 (= [-1 + .95]/.03). Así pues, no rechazaríamos la ley de Zipf con un nivel de significancia del 5%.

¿Qué pasaría en el caso de los años 1980, 1970 y 1960? Dado que los valores estimados para las pendientes serían, respectivamente, -.82, -.80 y -.71, con errores estándares cercanos a .04 en los tres casos, entonces para esas tres décadas la hipótesis de Zipf sería fácilmente rechazada (con un nivel de significancia del 5%).

Ahora procederemos a examinar, para cada una de esas cuatro muestras, la validez de la ley de Zipf mediante el estadístico *LMZ*. Como se comentó con anterioridad, esto nos permite efectuar un análisis estadístico más robusto. Considere, por ejemplo, el año de 1990. Como puede apreciarse en el cuadro 1, en ese momento la ciudad más pequeña de la muestra correspondía a Ciudad Constitución con una población de 34692 habitantes. Así pues, $x_{(81)}$ = 34692. Habiendo obtenido este valor, puede uno entonces calcular las dos expresiones para z_1 y z_2, para finalmente calcular el valor de *LMZ*.

Siguiendo ese procedimiento, los valores calculados usando el estadístico *LMZ* para 1960, 1970, 1980 y 1990 fueron, respectivamente: 545.54, 272.45, 304.25 y 51.76., todos los cuales son grandísimos comparados con 4.61 (el valor crítico aproximado cuando el nivel de significancia es del 10%). De hecho, necesitaríamos estar dispuestos a aceptar valores de significancia del orden de .0000000001% para no rechazar la ley de Zipf para el mejor de los valores de LMZ (51.76 para 1990). Esto querría decir que estaríamos dispuestos a aceptar valores muy cercanos al 100% en el caso de la probabilidad de cometer el error Tipo II (es decir, la probabilidad de que siendo la hipótesis de Zipf incorrecta, uno persista en afirmar que es cierta).

Nótese que aun cuando también habíamos detectado la invalidez de la ley de Zipf para 1960, 1970 y 1980 mediante el procedimiento de mínimos cuadrados ordinarios, éste nos había hecho concluir erróneamente que tal ley era posiblemente válida para 1990. Este ejemplo muestra de manera muy clara porqué tantos investigadores, al usar procedimientos estadísticos ineficientes, han acabado por concluir que la ley de Zipf es robusta en muchos países.

Ahora bien, alguien puede aún argüir que en el caso mexicano la selección gubernamental incluye ciudades demasiado pequeñas, lo cual hace que la ley de Zipf no sea satisfecha *a priori*. Por ello, en un segundo examen se restringieron las muestras en cada una de las cuatro décadas a ciudades cuyo tamaño fuese de al menos 50,000 personas. Los valores entonces obtenidos para *LMZ* en 1960, 1970, 1980 y 1990 son, respectivamente (en paréntesis ponemos el tamaño de cada muestra): 1.91 (n=35), 4.64 (n=53), 10.74 (n=65) y

17.73 (*n*=77). Así pues, para 1970, 1980 y 1990 rechazamos de nueva cuenta la hipótesis de Zipf, pero para 1960 no la podemos rechazar usando un nivel de significancia del 5%. Unos resultados tampoco muy alentadores para los seguidores de esa hipótesis.

En busca de mejores resultados, decidimos reducir aún más las muestras. Si solamente consideramos las ciudades cuyo tamaño fuese de al menos 100,000 personas en cada una de las cuatro décadas, los nuevos valores de *LMZ* en 1960, 1970, 1980 y 1990 son entonces (en paréntesis seguimos poniendo el tamaño de cada muestra): 1.49 (*n*=17), 5.14 (*n*=32), 2.86 (*n*=47) y 9.52 (*n*=54). Así pues, para 1960 y 1980 ya no podemos rechazar la hipótesis nula, pero para 1970 y 1990 sí la rechazamos. De hecho, para 1990, aún en esta muestra tan selecta, la hipótesis de Zipf es rechazada sobradamente.

Conclusiones

Los resultados obtenidos en la sección anterior son muy desalentadores para los seguidores de Zipf. ¿Quiere esto decir que México constituye una excepción a esa supuesta "ley natural"? No. Lo que es más probable es que los estudios que pretenden justificar la ley de Zipf en otros países (o en regiones aún mayores) sean muy poco confiables desde un punto de vista estadístico (véase también la discusión en Rapoport, 1978). De hecho, en Urzúa (2000) se muestra que en el caso estadounidense tampoco se cumple la ley de Zipf, excepto para muestras muy bien seleccionadas. Esto no debe sorprendernos porque, como también se prueba en Urzúa (2000), "estrictamente hablando, la ley de Zipf no puede ser cierta excepto para un determinado tamaño de la muestra, y esto si acaso".

Bibliografía

Clark, C. (1996), *Population Growth and Land Use*, Nueva York, Macmillan.

Gabaix, X. (1999), “Zipf’s Law for Cities: An Explanation”, por aparecer en *Quarterly Journal of Economics*.

Krugman, P. (1996), *The Self-Organizing Economy*, Oxford, Blackwell.

Presidencia de la República (1994), *Anexo al 6º. Informe de Gobierno*, Ciudad de México, 1994.

Rapoport, A. (1978), “Rank-Size Relations”, en H. Kruskal, y J. M. Tanur, compiladores, *International Encyclopedia of Statistics*, vol. 2., Nueva York, Free Press, pp. 847-854.

Rosen, K.T. y M. Resnick (1980), “The Size Distribution of Cities: An Examination of the Pareto Law and Primacy”, *Journal of Urban Economics*, vol. 8, pp. 165-186.

Urzúa, C..M. (2000), “A Simple and Efficient Test for Zipf’s Law”, *Economics Letters*, vol. 66, núm. 3, pp. 257-260.

Zipf, G. (1949), *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Reading, Mass., Addison-Wesley.

CUADRO 1

Areas metropolitanas con mayor población en México

(Exceptuando Ciudad de México; ordenamiento de acuerdo con la población en 1990)

	1960	1970	1980	1990
Guadalajara	907511	1533485	2323380	2987194
Monterrey	740732	1278780	2040521	2603709
Puebla	445697	723453	1122858	1436671
Ciudad Juárez	262119	407370	544496	789522
León	209870	364990	593002	758279
Tijuana	152374	277306	429500	698752
Torreón-Gómez Palacio-Lerdo	258757	322557	478523	650121
Mérida	170834	212097	400142	523422
Chihuahua	150430	257027	385603	516153
Acapulco	49149	174378	301902	515374
San Luis Potosí	159980	230039	362371	489238
Aguascalientes	126617	181277	293152	440225
Mexicali	174540	263498	341559	438377
Saltillo-Ramos Arizpe	102764	167319	294310	437743
Tampico-Ciudad Madero	176163	270414	400401	433021
Morelia	100828	161040	297544	428486
Culiacán	85024	167956	304826	415046
Hermosillo	95978	176596	297175	406417
Querétaro	67674	112993	215976	385503
Durango	97305	150541	257915	348036
Coatzacoalcos-Minatitlán	37300	69753	233935	340877
Reynosa-Río Bravo	74140	176401	249929	332755
Toluca	77124	114079	199778	327865
Veracruz	144681	214072	284822	303153
Tuxtla Gutiérrez	41244	66851	131096	289626
Xalapa	66269	122377	204594	279451
Cuernavaca	37144	134117	192770	279187
Matamoros	92327	137749	188745	266055
Irapuato	83768	116651	170138	265042
Mazatlán	75751	119553	199830	262705
Villahermosa	52262	99565	158216	261231
Córdoba-Orizaba	117157	171012	214820	244912
Ciudad Obregón	67956	114407	165572	219980
Nuevo Laredo	92627	148867	201731	218413
Celaya	58851	79977	141675	214856
Oaxaca	72370	99535	154223	212818
Tepic	54069	87540	145741	206967
Ciudad Victoria	50797	83897	140161	194996
Uruapan	45727	82677	122828	187623
Monclova	43077	78134	115786	177792
Pachuca	64571	83892	110351	174013
Ensenada	42561	77687	120483	169426
Los Mochis	38307	67953	122531	162659

Poza Rica	19564	120462	166799	151739
Campeche	43874	69506	128434	150518
Zacatecas-Guadalupe	39589	63497	105483	146484
Zamora-Jacona	47673	80489	116953	145597
Colima	43518	58450	86044	142844
Tehuacán	31897	47497	79547	139450
Tapachula	41578	60620	85766	138858
La Paz	24253	46011	91453	137641
Salamanca	32663	61039	96703	123190
Cuautla	12427	13946	24153	110242
Nogales	37657	52108	65603	105873
Chilpancingo	18022	36193	67498	97165
Piedras Negras	44992	41033	67455	96178
San Luis Río Colorado	28545	49990	76684	95461
Chetumal	12855	23685	56709	94158
Puerto Vallarta	7484	24155	38645	93503
Ciudad Valles	23823	47587	65609	91402
Hidalgo del Parral	41474	57619	75590	88197
Guaymas	34865	57492	54826	87484
Tlaxcala-Chiautempan	18841	22299	31641	85984
Ciudad del Carmen	21164	40855	72489	83806
Iguala	26845	45355	66005	83412
Apatzingán	19568	44849	55522	76643
Tulancingo	26839	35799	53400	75477
Fresnillo	35582	44475	56066	75118
Guanajuato	28212	36809	48981	73108
Manzanillo	19950	20777	39088	67697
San Juan Bautista Tuxtepec	8471	17700	29060	62788
Salina Cruz	14897	22004	40010	61656
San Juan del Río	11177	15422	27204	61652
Matehuala	19927	28799	41550	54713
Tepatitlán de Morelos	19835	29292	41813	54036
Juchitán de Zaragoza	19797	30218	38801	53666
Lázaro Cárdenas	1906	4766	26217	53581
Allende	14891	24286	30003	48935
Apizaco	15705	21189	30498	43663
Zihuatanejo	1619	4874	6887	37328
Ciudad Constitución	1706	15968	23557	34692

Fuente: *Anexo al 6o. Informe de Gobierno*, Presidencia de la República, 1994, p. 517.