Notas y comentarios

Las ciudades mexicanas no siguen la ley de Zipf*

Carlos M. Urzúa**

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 arguye 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 (véase, por ejemplo, Clark, 1967; Rosen y Resnick, 1980; y Krugman, 1996). No obstante, uno de los dos propósitos de este artículo es advertir que tal ley no se aplica al caso mexicano. El otro objetivo es mostrar 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 de cada una de ellas:

*Agradezco los comentarios de los dictaminadores anónimos que revisaron el presente artículo. Este trabajo forma parte del proyecto Geografía y Desarrollo Económico que fue encargado por el Banco Interamericano de Desarrollo a El Colegio de México en 1999.

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

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

donde r denota el rango de la ciudad y $x_{(r)}$ denota su tamaño. La ley de Zipf asegura que si grafieáramos 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 las dos ciudades más pobladas (rango igual a 1 y 2), entonces la más grande debe 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:

$$\ln r = \beta_1 + \beta_2 \ln x_{(r)} + \varepsilon_r$$

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 explica ampliamente 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:

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

con el parámetro $\alpha = 1$ ($x \geq \mu$). Habiendo encontrado tal densidad, en "A Simple and Efficient Test for Zipf's Law" (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 χ^2 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 de 10%, como es común en los estudios sobre distribuciones, entonces el valor crítico que debe usarse es aproximadamente 4.61. Éste es el valor crítico de una χ^2 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 que se presenta 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 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 se presentan en el cuadro 1.

CUADRO 1
Áreas metropolitanas con mayor población en México*

	1960	*1970*	*1980*	*1990*
Guadalajara	907 511	1 533 485	2 323 380	2 987 194
Monterrey	740 732	1 278 780	2 040 521	2 603 709
Puebla	445 697	723 453	1 122 858	1 436 671
Ciudad Juárez	262 119	407 370	544 496	789 522
León	209 870	364 990	593 002	758 279
Tijuana	152 374	277 306	429 500	698 752
Torreón-Gómez Palacio-Lerdo	258 757	322 557	478 523	650 121
Mérida	170 834	212 097	400 142	523 422
Chihuahua	150 430	257 027	385 603	516 153
Acapulco	49 149	174 378	301 902	515 374
San Luis Potosí	159 980	230 039	362 371	489 238
Aguascalientes	126 617	181 277	293 152	440 225
Mexicali	174 540	263 498	341 559	438 377
Saltillo-Ramos Arizpe	102 764	167 319	294 310	437 743
Tampico-Ciudad Madero	176 163	270 414	400 401	433 021
Morelia	100 828	161 040	297 544	428 486
Culiacán	85 024	167 956	304 826	415 046
Hermosillo	95 978	176 596	297 175	406 417
Querétaro	67 674	112 993	215 976	385 503
Durango	97 305	150 541	257 915	348 036
Coatzacoalcos-Minatitlán	37 300	69 753	233 935	340 877
Reynosa-Río Bravo	74 140	176 401	249 929	332 755
Toluca	77 124	114 079	199 778	327 865
Veracruz	144 681	214 072	284 822	303 153
Tuxtla Gutiérrez	41 244	66 851	131 096	289 626

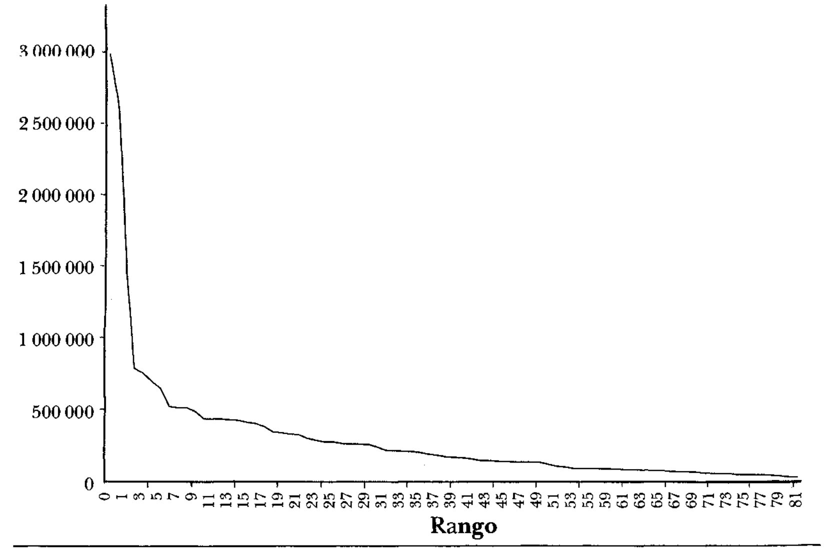
Xalapa	66 269	122 377	204 594	279 451
Cuernavaca	37 144	134 117	192 770	279 187
Matamoros	92 327	137 749	188 745	266 055
Irapuato	83 768	116 651	170 138	265 042
Mazatlán	75 751	119 553	199 830	262 705
Villahermosa	52 262	99 565	158 216	261 231
Córdoba-Orizaba	117 157	171 012	214 820	244 912
Ciudad Obregón	67 956	114 407	165 572	219 980
Nuevo Laredo	92 627	148 867	201 731	218 413
Celaya	58 851	79 977	141 675	214 856
Oaxaca	72 370	99 535	154 223	212 818
Tepic	54 069	87 540	145 741	206 967
Ciudad Victoria	50 797	83 897	140 161	194 996
Uruapan	45 727	82 677	122 828	187 623
Monclova	43 077	78 134	115 786	177 792
Pachuca	64 571	83 892	110 351	174 013
Ensenada	42 561	77 687	120 483	169 426
Los Mochis	38 307	67 953	122 531	162 659
Poza Rica	19 564	120 462	166 799	151 739
Campeche	43 874	69 506	128 434	150 518
Zacatecas-Guadalupe	39 589	63 497	105 483	146 484
Zamora-Jacona	47 673	80 489	116 953	145 597
Colima	43 518	58 450	86 044	142 844
Tehuacán	31 897	47 497	79 547	139 450
Tapachula	41 578	60 620	85 766	138 858
La Paz	24 253	46 011	91 453	137 641
Salamanca	32 663	61 039	96 703	123 190
Cuautla	12 427	13 946	24 153	110 242
Nogales	37 657	52 108	65 603	105 873
Chilpancingo	18 022	36 193	67 498	97 165
Piedras Negras	44 992	41 033	67 455	96 178
San Luis Río Colorado	28 545	49 990	76 684	95 461
Chetumal	12 855	23 685	56 709	94 158
Puerto Vallarta	7 484	24 155	38 645	93 503
Ciudad Valles	23 823	47 587	65 609	91 402
Hidalgo del Parral	41 474	57 619	75 590	88 197
Guaymas	34 865	57 492	54 826	87 484
Tlaxcala-Chiautempan	18 841	22 299	31 641	85 984
Ciudad del Carmen	21 164	40 855	72 489	83 806
Iguala	26 845	45 355	66 005	83 412
Apatzingán	19 568	44 849	55 522	76 643
Tulancingo	26 839	35 799	53 400	75 477
Fresnillo	35 582	44 475	56 066	75 118
Guanajuato	28 212	36 809	48 981	73 108
Manzanillo	19 950	20 777	39 088	67 697
San Juan Bautista Tuxtepec	8 471	17 700	29 060	62 788
Salina Cruz	14 897	22 004	40 010	61 656
San Juan del Río	11 177	15 422	27 204	61 652
Matehuala	19 927	28 799	41 550	54 713
Tepatitlán de Morelos	19 835	29 292	41 813	54 036
Juchitán de Zaragoza	19 797	30 218	38 801	53 666
Lázaro Cárdenas	1 906	4 766	26 217	53 581
Allende	14 891	24 286	30 003	48 935
Apizaco	15 705	21 189	30 498	43 663
Zihuatanejo	1 619	4 874	6 887	37 328
Ciudad Constitución	1 706	15 968	23 557	34 692

*Exceptuando la Ciudad de México, ordenadas de acuerdo con su población en 1990.
Fuente: *Anexo al 6o. Informe de Gobierno,* Presidencia de la República, 1994, p. 517.

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 contenidos en el cuadro 1, bien vale la pena preguntarse acerca de los resultados que se obtendrían al aplicar los procedimientos de "prueba" de la ley de Zipf que han sido empleados hasta este momento en la literatura. A manera de ilustración consideramos 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.

GRÁFICA 1
Población contra rango para 81 áreas metropolitanas de México (excluyendo la capital), 1990



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 hacemos lo mismo 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 es 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*, que en nuestro caso es igual a −1.67 (= [-1 + .95]/.03). Así pues, no rechazaríamos la ley de Zipf con un nivel de significancia del 5 por ciento.

¿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 de 5 por ciento).

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. Consideremos, por ejemplo, el año de 1990. Como puede apreciarse en el cuadro 1, en ese momento la ciudad más pequeña de la muestra era Ciudad Constitución, con una población de 34 692 habitantes. Así pues, $x_{(81)}$ = 34692. Habiendo obtenido este valor, es posible 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 de 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 a 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, se 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 por qué 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 (entre paréntesis se anota 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 de 5%. Tales resultados tampoco parecen muy alentadores para los seguidores de esa hipótesis.

En busca de mejores resultados decidimos reducir aun más las muestras. Si solamente consideramos las ciudades cuyo tamaño es 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: 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, aun 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 más probable es que los estudios que pretenden justificar la ley de Zipf en otros países (o en regiones aun mayores) sean muy poco confiables desde un punto de vista estadístico (véase también la discusión en Rapoport, 1978). De hecho, en el trabajo de Urzúa (2000) se muestra que en el caso estadunidense tampoco se cumple la ley de Zipf, excepto para muestras muy bien seleccionadas. Esto no debe sorprendernos porque, como también se prueba en dicho trabajo, "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. (1967), *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,* México, 1994.

Rapoport, A. (1978), "Rank-Size Relations", en H. Kruskal y J. M. Tanur (comps.), *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.