More on Hypothesis Testing:An example of some empirical research: An analysis of city size distributions in China introducing The Difference in Means and Associated Tests

Sometimes rather than be concerned about the particular value of the mean or variance of a population, or the nature of the underlying population distribution and we may wish to examine whether two population distributions differ, for example we may be interested in comparing the means or variances from two distinct populations. It turns out that with some minor modifications we perform very similar tests to the ones we have already discussed. To exemplify the tests discussed in the previous chapter and to introduce these new tests we will use some recent research on the nature and progress of city sizes in China over the period 1949-1999 (Anderson and Ge (2003)).

Some Background: A Theory of City Size Distributions.

There is a theory in urban economics (Gabaix (1999)), for which evidence has been found in many countries, which says that when measured by the number of inhabitants city sizes within a country are governed by Zipf's law Zipf (1949) (the logarithm of the rank of the city size ≈ - logarithm of the size relative to the minimum). This means that the distribution of city sizes at any point in time is a Pareto Distribution with a parameter $\theta = 1$. Letting x be the city size and x_{min} be the minimum possible city size, the pdf and cdf of this distribution are given by $f(x) = \theta (x_{min}/x^{\theta+1})$ and $F(x) = 1- (x_{min}/x)^{\theta}$ respectively.

The law derives from two ideas, the first is that individual cities start at some minimum size (call it Z_0) which is subject to a sequence of mutually independent multiplicative shocks (call the one in the i'th period $(1+X_i)$ where X_i is small relative to 1 and to Z_0). In essence the shock is related to the number of people that die in, the number of people that are born in, the number of people that immigrate to and the number of people that emigrate from the city in period i, all of which are random events. The logarithm of city size in period I may be written as:

$$\ln(Z_I) = \ln\left(Z_0 \left(\prod_{i=1}^{I} (1 + X_i) \right) \right) = \ln Z_0 + \sum_{i=1}^{I} e_i$$

where $e_i = \ln(1+X_i)$. For I sufficiently large (that is after a sufficiently long period of time), the logarithm of city sizes can be shown to be distributed $f(\ln x) = N(\ln Z_0 + (g - .5\sigma^2)I, I\sigma^2)$[1] where g is the long run growth component in e_i and σ^2 is its variance. This is known as Gibrat's law (Gibrat(1930)(1931)), note that if $g > .5\sigma^2$ this theory predicts that the mean and variance of the city size distribution will grow through time.

This is clearly not the same as Zipfs law and without the second idea this would be the city size distribution formula. The second idea is that city size is also subject to a lower reflective boundary below which it is not allowed to go so that if an X_i took Z_i below Z_0 the size would stay at Z_0 until a subsequent shock took it back up above Z_0. Gabaix was able to show that in this case the distribution of city sizes would become a Pareto Distribution with a paremter $\theta = 1$ as above. Thus whether or not the mean and variance of city sizes grow through time and whether or not city size distributions are log-normal or pareto speaks to which parts of the theory is most powerful. Anderson and Ge (2003) used these ideas using Data from the Peoples Republic of China and the USA (the USA is known to be a country where Zipfs law holds) tables of results drawn from their work are provided in the Appendix. In both countries the minimum city size is set at 100000. Tables 1 and 1a summarize the data.

Implementing Pearson Goodness of Fit Tests.

A natural thing to do is to employ Pearson Goodness of fit tests to check if the data on city sizes are Pareto or Log Normal. Tests based upon partitioning the range of the city size random variable into 10 equi-probable regions are reported in tables 2 and 2a for the Pareto distribution and in tables 3 and 3a for the Log-Normal distribution. It is instructive to see how these regions were determined. For 10 equi-probable intervals we need to

[1]As an aside it is interesting to note that this is really an application of the Central Limit Theorem we discussed in CN4.

determine from the cumulative density function F(x) the values of x for which F(x) = 0.1, 0.2, 0.3, ….., 0.9 (recall that F(x) = P(X<x)). For the Pareto (θ=1) distribution this means given p and x_{min} = 100000 solving the formula:

$$F(x) = 1 - \frac{x_{\min}}{x} = p, \; for \; p = 01, 0.2, 0.3....., 0.9$$

$$\text{so that } x = \frac{x_{\min}}{1-p}, \; for \; p = 0.1, 0.2, ..., 0.9$$

For the normal distribution we do not have a closed form solution for F(x) so that we have to resort to the statistical tables. Given that x ~ N (μ,σ^2) and given values for μ and σ^2 recall that:

$$p = F(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le Z^*{}_p \left(= \frac{x-\mu}{\sigma}\right)\right)$$

so we simply look up in the standard normal tables the value Z_p* for which P($Z \le Z_p$*) = p and solve for x = $\mu + \sigma Z_p$* for p = 0.1, 0.2, ..., 0.9. When we do not know μ and σ^2 we have to use estimates and remember to adjust the degrees of freedom in the goodness of fit test accordingly.

Recalling that the Goodness of Fit test statistic is given by:

$$\sum_{k=1}^{K} \frac{\left(O_k - E_k\right)^2}{E_k}$$

O_k is simply obtained by counting up how many cities have populations in the k'th interval and E_k is simply the sample size x 0.1 in every case. The statistic is distributed as χ^2(K-1-m) where m is the number of parameters to be estimated so that when K = 10 and we estimate μ and σ^2 in the case of the normal we have a test with 7 degrees of freedom. In the case of the Pareto distribution Anderson and Ge actually estimated θ so that it became an 8 degree of freedom test rather than a 9 degree of freedom test.

As for the results the last columns of tables 2 and 2a indicate the Pareto distribution is not rejected for the USA (with the exception of 1950) but it is rejected for China (with the exception of 1949) whereas tables 3 and 3a show that the log-normal distribution is not rejected for China but it is for the USA. This suggests that the lower reflective

boundary operates in the USA but not in China so that China appears to obey Gibrat's rather than Zipf's law.

If the city size distribution in China obeys Gibrat's law the means and variances should be increasing over time. To examine this possibility we need to modify means and variances tests to enable us to make comparisons through time.

The difference in means test in the Chinese City Size Data example.

In the case of our data on Chinese city sizes the sample sizes are generally so large that the standard normal difference in means test would be appropriate. So for example the critical values for a two-sided difference in means test of size .05 for city sizes for 1985 and 1990 would be:

$$\pm\sqrt{\frac{76.5^2}{313}+\frac{73.4^2}{453}}\cdot\left(1.96\left(=Z_{0.975}\right)\right)=10.84$$

which, given the difference in sample means is 6.2 fails to reject the hypothesis of identical means.

References

Anderson and Ge (2003) The Size Distribution of Chinese Cities. Economics Department University of Toronto.

Gabaix, X. (1999) "Zipf's Law for Cities: An Explanation" *Quarterly Journal of Economics* 739-767.

Gibrat, R. (1930) "Une Loi Des Repartitions Economiques: L'effet Proportionelle" Bulletin de Statistique General, France, 19 p469.

Gibrat, R. (1931) Les Inegalites Economiques Paris: Libraire du Recueil Sirey.

Zipf, G. (1949) *Human Behavior and the Principle of Last Effort*, Cambridge MA: Addison Wesley.

Appendix

Table 1.
Summary Statistics of City Size (x10000) in China

	1949	1961	1970	1980	1985	1990	1994	1999
Mean	47.9	56.4	56.4	64	67.7	73.9	78.7	82.3
Std deviation	60.7	75.9	77.2	81.5	76.5	73.4	71.5	87.9
Median	28.4	35.3	29.9	36.7	43.6	57.1	63.2	64.1
Minimum	10.1	10.2	10.2	10.2	10.1	10.2	11.97	10
Maximum	418.9	641.2	580.2	601.3	698.3	783.5	953	1127
Number of cities	77	176	164	208	313	453	606	658

Table 1a.
Summary Statistics of City Size (x10000) in United States of America:

Year	1930	1940	1950	1960	1970	1980	1990	2000
mean	39.2	47	47.3	43.6	40.5	33.5	32.6	30.3
s.d.	81.1	89.5	89.9	81.6	76.5	64.8	63.3	60.2
median	16.6	19.3	18.7	19.4	17.7	16.9	17.2	17.3
minimum	10	10.1	10.2	10	10	10	10	10
maximum	693	745.5	789.2	778.2	789.6	707.2	732.3	800.8
observations	94	98	112	136	161	168	192	238

Table 2
The Rank Order (Single Pareto) Distribution Model (China)

Year	Sample size	p_{minML}	θ_{ML}	$Var(\theta_{MLE})$	(θ_{OLS})	$Var(\theta_{OLS})$	$\chi^2(8)$ GF
1949	77	10.0083	0.8649	0.0097	0.9138	0.0004	12.4805*
1961	176	10.1719	0.7762	0.0034	0.8591	0.0003	36.8409
1970	164	10.1875	0.8027	0.0039	0.8747	0.0002	21.0000
1980	208	10.1410	0.7072	0.0024	0.7989	0.0002	60.7500
1985	313	10.0378	0.6377	0.0013	0.7374	0.0003	187.9265
1990	453	10.1675	0.5830	0.0008	0.6793	0.0002	446.8896
1994	606	11.9502	0.6025	0.0006	0.7021	0.0002	686.5743
1999	658	9.9848	0.5367	0.0004	0.6259	0.0002	822.2128

Table 2a
The Rank Order (Single Pareto) Distribution Model (US)

Year	Sample size	p_{minML}	θ_{ML}	$Var(\theta_{ML})$	θ_{OLS}	$Var(\theta_{OLS})$	$\chi^2(8)$GF	$1-F(\chi^2)$
1950	112	10.0624	1.0463	0.0098	1.0166	0.0004	16.2142	0.0394
1960	136	9.9612	1.0729	0.0085	1.0600	0.0004	9.2941	0.3181
1970	161	9.9414	1.1314	0.0080	1.1110	0.0003	9.2484	0.3218
1980	192	9.9458	1.3014	0.0101	1.2637	0.0003	10.6905	0.2199
1990	168	9.9695	1.3213	0.0091	1.2945	0.0002	11.7500	0.1627
2000	238	.9844	1.3920	0.0081	1.3679	0.0003	6.1176	0.6341

Legend

p_{minML}	Unbiased transformation of Maximum Likelihood Estimate of the Lower bound (see footnote 5).
θ_{ML}	Maximum likelihood estimate of Zipf parameter.
$Var(\theta_{ML})$	Variance of maximum likelihood estimate.

θ_{OLS}	Restricted Least Squares Estimate of Parameter.
Var(θ_{OLS})	Variance of Restricted Least Squares Estimate.
$\chi^2(8)$GF	Pearson Goodness of Fit Test (based upon 10 equiprobable cells).
1-F(χ^2)	Upper tail probability of Pearson Goodness of Fit Test

Table 3

Year	Sample size	Mean	Std Dev	Std Error of mean	$\chi^2(7)$ GF	1-F(χ^2)
1949	77	3.4596	0.8312	0.0947	7.8052	0.3501
1961	176	3.6079	0.8435	0.0636	8.8864	0.2609
1970	164	3.5670	0.8771	0.0685	16.0000	0.0251
1980	208	3.7307	0.8532	0.0592	20.8462	0.0040
1985	313	3.8744	0.7792	0.0440	11.3131	0.1255
1990	453	4.0344	0.7022	0.0330	17.6181	0.0138
1994	606	4.1406	0.6479	0.0263	5.0231	0.6571
1999	658	4.1641	0.6606	0.0258	11.3921	0.1224

Log-normal Model (China)

Table 3a

Log-normal Model (USA)

Year	Sample size	Mean	Std Dev	Std Error of mean	$\chi^2(7)$ GF*
1950	112	12.4749	0.9126	0.0862	40.1429
1960	136	12.4411	0.8706	0.0747	51.6471
1970	161	12.3909	0.8454	0.0666	64.4037
1980	168	12.2756	0.7547	0.0582	69.2619
1990	192	12.2667	0.7356	0.0531	70.1875
2000	238	12.2298	0.7003	0.0454	90.8235

***** Upper tail probabilities not reported since they are all substantially less than .01