

WHITEPAPER

BIG AND SMALL – A STATISTICAL LOOK AT MARKET CAPITALIZATION BREAK POINTS

ANDREW CLARK

CHIEF INDEX STRATEGIST
THOMSON REUTERS INDICES

August 2009



THOMSON REUTERS

BIG AND SMALL – A STATISTICAL LOOK AT MARKET CAPITALIZATION BREAKPOINTS

INTRODUCTION

To most retail investors, intermediaries, and probably for most of the public at large, what makes a large capitalization (large cap) stock a large cap and what makes a small cap stock a small cap is something that is either taken for granted or assumed to be correct.

When it comes to the “pro’s” – CIO’s, performance analysts, mutual fund managers and the like – the lines that separate large from mid and mid from small and small from micro can vary quite a bit from firm to firm.¹ This variety of views comes about because some capitalization breakpoint methods are based on market lore and others are subjective assessments of varying accuracy.

In this brief paper, we will find that the common “70%” rule for large cap stocks is approximately correct but needs some “massaging” if a significant level of accuracy is needed. That micro-caps stocks have a clear breakpoint somewhere near common used levels. And small and mid-cap stocks are often very difficult to tell apart which lends some credence to the Salomon/Citigroup claim of “smid.”

The paper will first use statistical techniques to analyze global, U.S., U.K. and Japanese market capitalizations and make conjectures about where capitalization breakpoints may be. Then it will apply well-recognized liquidity proxies such as bid-ask spread and volume to propose reasons for those breakpoints being where they are.

STATISTICAL ANALYSIS

Before going into the details of the statistical analysis, readers should know that what follows is a light to moderately difficult write-up of the work done. Appendix B has the full mathematical description. Readers not inclined to read the

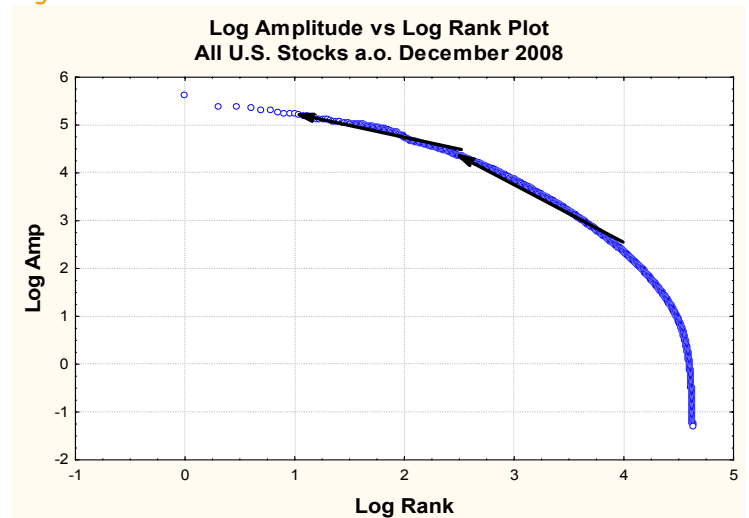
statistics can skip this section and read the next section - the economic analysis.

A few years back, in an unpublished working paper, the author used Zipf plots (log amplitude/log rank plots or log-log plots) to determine where capitalization breakpoints may be for U.S. stocks.

For the uninitiated, a Zipf plot is a way of determining if the data follows one particular type of distribution – such as a Gaussian or Normal – or is a mixture of distributions – such as the Truncated Levy.

To derive the values needed for the Zipf plots, the author ordered the market capitalization values of all U.S. stocks that trade on the NYSE and NASDAQ exchanges in descending order. The stocks were from 1 to more than 4000, with the stock with the largest capitalization having a rank of 1 and the smallest stock a rank greater than 4000. The log of the rank was computed (with 1 having the smallest log value and 4000 having one of the biggest). The log of the market capitalization was computed (the log of the market capitalization is the log amplitude – the higher the market capitalization the higher the log value). Figure 1 has the corresponding Zipf plot for data as of December 2008.

Figure 1



¹ See Appendix A for a sample of breakpoint values as calculated by major index providers.



The plot's curving pattern has two fairly clear breakpoints, i.e., where the curve's slope begins to change and is signified by the black arrows. These breakpoints were the author's conjecture of where the large cap breakpoint is, at approximately 2.5 on the log rank axis, and the micro-cap breakpoint at approximately 4.0 on the same axis.

The author observed these same breakpoints values on the log rank axis over and over again, month after month, from December 2005 back to January 1996. At the time, he had no clear economic interpretation of why the breakpoints were occurring so there was some resistance, and understandably so, in the investment community because these breakpoints could be nothing more than statistical phenomenon – and possibly more artifact than fact.

The author returned to the topic recently and using the R package HyperbolicDist² fit hyperbolic distributions to global, U.S., U.K. and Japanese market capitalization data with much better results, both from an analytical and economic standpoint. In essence, he found that the Zipf plots like the one shown in Figure 1 were not a combination of distributions but one distribution alone – the hyperbolic.

The hyperbolic distribution was chosen because its goodness of fit versus other distributions (this was tested via the Cramer Von-Mises statistic). And because it has a shape like the normal distribution for the majority of the data (about 45% on either side of the mean) with its tails (the remaining 5% or less on either side) following an exponential distribution. Many quantitative market practitioners have used hyperbolic distributions to model financial data with significant success. Among the early users of hyperbolic data in finance were Barndorff-Nielsen, Blaesild and Eberlein, and Prause. This paper has benefited significantly, especially from the work of Barndorff-Nielsen. All of these early authors have shown the unusual effectiveness of working with hyperbolic distributions in finance.

² R is a freely available set of statistical and mathematical software packages which work with a base program (R base). More information can be found at <http://cran.r-project.org/>.

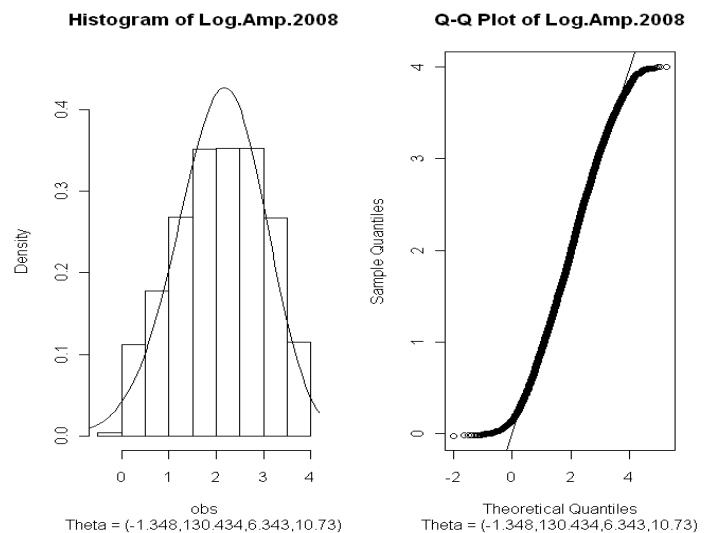
Hyperbolic distributions have also been used elsewhere to great effectiveness:

- In statistical physics , when they are accounting for relativistic effects in statistical mechanics
- In sedimentology
- In studies of turbulence

Note that the hyperbolic distribution is a random sample of various normal distributions and the limit case of the hyperbolic distribution is the normal distribution.

Figure 2 shows the fit of the hyperbolic distribution to the same data as Figure 1

Figure 2: U.S. Data Dec. 2008



The histogram on the left shows the fit of the log amplitude (the log of the ranked capitalization values) with the curve being the best fit hyperbolic distribution. Notice that the curve passes smoothly through the tops of all the histogram bins except for the last one on the left which corresponds to the smallest capitalizations (less than 100M USD) in the U.S. market.

The q-q or quantile – quantile plot confirms the greater than expected number of very small capitalizations. It also shows that the vast majority of the data fits the hyperbolic well. And the Cramer-Von Mises statistic (not shown) rules out the normal distribution as an equal or better fit than the hyperbolic. The “Theta = (...)” in both graphs are the parameters of the hyperbolic distribution fit.

The extreme tails, those log capitalizations that do not lie on the sloping line in the q-q plot are, the author will show, mostly mega caps on the top end and the many, many U.S. stocks that have very small capitalizations, in this case less than 100M USD.

Figure 3 shows a similar plot for global data and Figure 4 shows similar plots for global data from year end 1997.

Figure 3: Global Data Dec. 2008

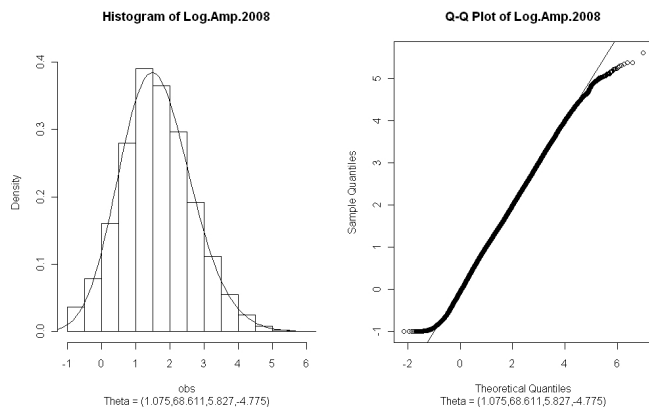
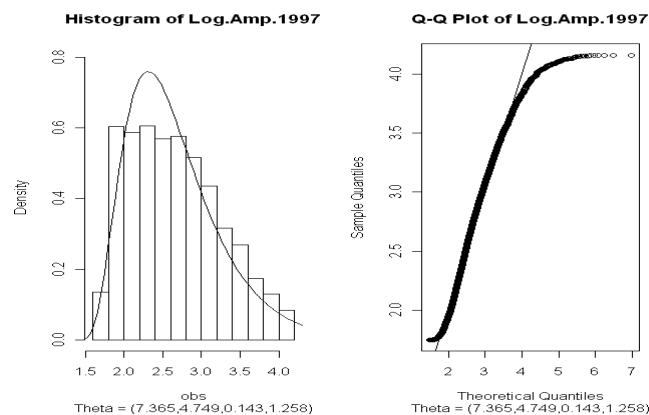


Figure 4: Global Data Dec. 1997



In all cases, the significance of the hyperbolic fit as measured by the Cramer-Von Mises test is better than 99% and the probability of the data being fit in most cases with a normal was at least less than 75%. Note however that in the global q-q plot it is the very large number of very large capitalization stocks that are not completely covered by the exponential tails of the hyperbolic.

Now, to determine the breakpoints for small caps and large caps, we will gradually reduce the lowest capitalization (largest ranking) stocks in steps of 10^x until the Cramer Von-Mises test fails, i.e., the

hyperbolic distribution is no longer the best fit. And 10^x steps because we are dealing with log values of capitalization and rank. We will also use Cramer's function $S(u)$ to verify the breakpoint range and to help narrow that range where needed. The mathematical logic for using both of these methods in order to find breakpoints is contained in Appendix B.

Using 2008 global capitalization as our first case, neither 10^1 or 10^2 tripped the Cramer Von-Mises test. This is confirmed by the calculation of $S(u)$ as well. However at 10^3 the test is tripped so somewhere between 10^2 and 10^3 or \$100M and \$1B is the small cap floor breakpoint. The values of $S(u)$ confirm that the breakpoint is in this range. No further testing or rigor will be used at this point because as is often the case in finance, financial models are only approximate so the application of greater rigor would only mean the use of various assumptions that are often manifestly violated by the real world.

Similarly, tests at 10^6 and 10^5 do not trip Cramer Von-Mises but 10^4 does so the large cap breakpoint is somewhere between \$100B and \$10B. Now, this is a big range, but as will be shown in the next section, the common use of 70% of an exchange's total market cap as the cut-off for large caps is well justified.

Furthermore, identical market capitalization tests were performed on TRX U.K., Japan, and Global index constituents. And for each of the year-ends from 2008 back to 1997, similar market capitalization breakpoints were found for both country indices as well as the global index.

ECONOMIC ANALYSIS

It is commonly accepted in financial research that measures of liquidity can help tell small cap stocks from large cap stocks. In this section, we will show that this is certainly true for large caps and micro caps but possibly not for small and mid caps.

We will use the bid-ask spread, the 52 week average volume, and the last 3 months average volume as proxies for liquidity. We will focus exclusively on U.S. stocks because that is where the 3 proxies have the best historical coverage.

Using year-end 2008 data as an example, the bid-ask spread between large cap and mid cap stocks is not statistically significant as measured by a t-test. Small and micro cap spreads however are indeed distinct and at a variety of small cap floor cutoffs between about 750M USD and 100M USD, small caps and micro caps can be separated. This helps to confirm the hypothesis in the statistical section above that the small cap floor is somewhere between 100 USD and 1B USD. The $S(u)$ values confirm that the small cap floor is more than likely not much smaller than 500M USD.

The same tests were repeated for year ends going back to 1997 and while the lowest floor varied substantially – from a low of approximately 300M USD to 750M USD – at no time did floor pierce 1B USD.

A second bid-ask test was done to see if mid cap stocks could be separated from small cap stocks. The tests were inconclusive. There were years when t-tests showed no significant difference between the two market capitalizations. And others when the bid-ask spreads were indeed different but the range of market cap breakpoints were not as tight as those seen for the small cap floor.

T-tests show both volume values can distinguish large, mid, small and micro cap from each other. The large cap volume values starts breaking way from mid cap values at levels very close to the aforementioned 70% rule. And repeated year end sampling showed little variation in this breakpoint. The breakpoints vary between 75% and 65%, with the majority of values close to 70%. All these ranges are confirmed by the $S(u)$ values.

Small and mid caps also have a good clean break but it changes and can change significantly (more than \$1B or \$2B) from year to year. This result, teamed with the q-q plots where small and mid cap stock can't be separated and the bid-ask results, makes the author put forth the hypothesis that the breakpoint between small and mid cap stocks is a market construct and is not substantiated empirically. This is true whether one uses q-q plots or liquidity proxies such as bid-ask spreads and volume.

Small and micro cap stocks can be nicely separated using volume, like they could with their respective bid-ask spreads. Again a range of small cap floor values were found, all were within the range of 100B USD and 1B USD.

The same separations exist in the sample of global, U.K. and Japanese year-end capitalizations examined.

CONCLUSIONS

The hypothesis that the use of hyperbolic distributions, the Cramer Von-Mises test and the Cramer function can be used to determine capitalization breakpoints is found to be true. The breakpoints found statistically are confirmed by commonly used liquidity proxies that can tell large caps, mid caps, etc. from each other. This is especially true for large cap and micro cap stocks.

The finding that there may not be a breakpoint between small and mid cap stocks may surprise some but the author's earlier study came to the same conclusion as have a small number of articles in the literature. The author is not saying a breakpoint does not *definitely* exist but rather the breakpoint may be a market construct with little empirical evidence to back it up.



APPENDIX A

Major Index Providers Market Capitalization Guidelines

Russell Global Index membership

When the total universe has been screened as described in Section 2, and after securities have been allocated to their home countries as described in Section 3, Russell determines index membership. Russell includes the top 98% of U.S. market capitalization, the Russell 3000®; and the top 98% of the rest of the world's market capitalization. This index design preserves global equity market integrity and effectively relieves the overrepresentation of U.S. from the global perspective. Additionally, this design assures consistency between the Russell Global Index and its U.S. sub-indexes as components.

The broad building blocks capturing 98%-plus of the investable market, enable thousands of modular subindexes, including country, region, sector, market capitalization and style segments. Each division of the parent index provides a set of sub-indexes with no gaps and no overlaps. Additionally, each sub-index, as a stand-alone index, provides comprehensive representation of a particular subgroup of the global investment opportunity set.

Global large cap and small cap index construction Research summary

The need for cap-size indexes is based on a well-documented phenomenon known as the cap-size effect. Stated simply, it means that large stocks tend to behave like other large stocks, and small stocks tend to behave like other small stocks. Russell observed this effect in the U.S. more than 20 years ago, and the effect has been seen to prevail in global markets as well. Much research has been focused on determining an appropriate dividing point between large and small stocks, but Russell's research has demonstrated that there is none. Instead, the division between large and small stocks should be established as a range or band around which representative large cap and small cap indexes can be created.

In addition, Russell research has demonstrated that the cap-size effect exists across regional boundaries; that is, companies of similar size tend to behave similarly regardless of geographic location. While this relationship is not equally strong across all regions (particularly in emerging markets) it does appear to be increasing as markets continue to globalize.

As a result of its research into the global cap-size effect, Russell implemented a global-relative methodology with banding when constructing the Global Large Cap, Global Mid Cap and Global Small Cap indexes, beginning with the June 2007 reconstitution. This approach differs fundamentally from the current industry practice of determining cap size on a country-by-country basis, where companies with very different market capitalizations may be classified in the same cap-size index, or, alternatively, where companies with similar market capitalizations may be classified in different

cap-size indexes simply because they are in different countries or regions. Cap-size indexes constructed by use of country-relative distinctions (whether banded or not) can generate substantial overlap when combined into broader indexes, and this reduces an index's usefulness in accurately representing what it was intended to measure.

Construction rules

At reconstitution, all companies in the Global Index (ex-U.S.) are ranked by their total market capitalization in descending order, and the cumulative total market capitalization percentile for each company will be calculated.

To determine the Russell Global Large Cap and Russell Global Small Cap indexes, all companies that rank below the 90th percentile will be classified as small cap, and all companies that rank above the 85th percentile will be classified as large cap. Companies that rank within the capitalization band between the 85th and 90th percentiles and that are members of the current index will retain their existing classification (i.e., if a member of the existing Russell Global Small Cap Index is within the 85th-90th percentile band at reconstitution, it will remain classified as small cap). New companies being added to the Russell Global Index will be classified relative to the midpoint of the range (i.e., new companies ranking above 87.5 will be classified as large cap, and new companies ranking below 87.5 will be classified as small cap).

To determine the Global Mid Cap Index, which is a sub-component of Global Large Cap, all companies that rank below the 60th percentile will be classified as mid cap, and all companies that rank above the 55th percentile will be classified as mega cap. Companies that rank within the capitalization band between the 55th and 60th percentiles and that are members of the current index will retain their existing classification (i.e., if a member of the existing Global Mid Cap Index is within the 55th-60th percentile band at reconstitution, it will remain classified as mid cap). New companies being added to the Global Index will be classified relative to the midpoint of the range (i.e., new companies ranking above 57.5 will be classified as mega cap, and new companies ranking below 57.5 will be classified as mid cap).

Using a global-relative 5% band has been shown to create portfolios that are robust representations of large and small stock behavior and to provide consistently better tracking results when tested against global and non-U.S. cap-tier mandated managers. Use of the banding approach also has the associated benefit of dramatically reducing turnover at reconstitution. Russell's research shows that a 5% band provides an optimal balance between representing asset-class return behavior and reducing turnover, which ultimately benefits investors who are using the indexes as passive vehicles or active portfolio benchmarks



Index name Upper range (percentiles) Lower range (percentiles)

Russell Global Mega Cap NA 55%.60%

Russell Global Mid Cap 55%.60% 85%.90%

Russell Global Small Cap 85%.90% NA

Percentiles are based on descending total market capitalization. Large Cap = Mega Cap + Mid Cap

Regional and country cap-size indexes

After every security in the Global Index has been assigned a cap-size classification, all other regional and country cap-size indexes will use these classifications in their construction. As a result, any combination of regional or country indexes across cap-size portfolios will be based on a consistent methodology.

Comparison to U.S. methodology

The new Global Index cap-size methodology is an extension of the existing U.S. indexes cap-size methodology, with some key enhancements. Both U.S. and Global methods first rank companies according to their total market capitalization. In the U.S. indexes, a fixed number of securities are chosen to establish the respective cap-size indexes; for example, the largest 1,000 stocks in the Russell 3000 determine the large cap index, and the smallest 2,000 stocks determine the small cap index. In the Global Index, a fixed percentile based on total market capitalization is used. While a fixed number of securities have worked very well in the U.S., Russell's research has shown that in the construction of global cap-size indexes, the relationship between large and small stocks is better represented by use of a fixed-percentile approach. Coinciding with the reconstitution of the global indexes in June 2007, the U.S. indexes methodology incorporated a banding model similar to that of the Global Index to determine the U.S. large cap, mid cap and small cap series.

Historically, the following methodology was used to build the Russell Global cap-tier indexes. The large/small breakpoint was made by using the corresponding breakpoints for the years 1996 to 2006 in the Russell U.S. indexes. These breakpoints generally correspond to the 90th percentile, on the basis of cumulative float-adjusted market capitalization of the global universe ranked in descending order by total market capitalization, including the U.S. Japan was calculated using the Russell/Nomura Total Market Index and their corresponding breakpoints. Russell/Nomura Total Market was used as the Japan portion from 96-08.

The mega cap/mid cap breakpoint was made by using the corresponding breakpoints for the years 1996 to 2006 in the Russell U.S. indexes. These breakpoints generally correspond to the 60th percentile, on the basis of cumulative float-adjusted market capitalization of the global universe ranked in descending order by total market capitalization, including the U.S. No banding was used in the historical construction.

DOW JONES INDICES

Size-segment categories are established as follows:

- a. The selected companies for each country are sorted by full market capitalization to determine market-

cap cutoffs for the large-cap, small-cap and mid-cap indexes.

- b. The market cap of the company that brought the cumulative market cap to 85% becomes the bottom cutoff target for the large-cap Index.
- c. Countries without micro-cap Indexes have no bottom target for the small-cap index. (As of April 1, 2009, only the U.S. market had a micro-cap index.) The market representation thresholds at that date were 98% of each country's float-adjusted market capitalization for developed markets ex-U.S. and 95% for emerging markets.
- d. Each mid-cap Index overlaps its country's respective large and small-cap Indexes. The market cap of the company that brought the cumulative market cap to 80% became the top cutoff target for the mid-cap index. The market cap of the company that brought the cumulative market cap to 90% became the bottom cutoff target for the mid-cap Index.

Stocks selected for each country index in step 1 whose full market capitalizations are equal to or larger than the 85% capitalization target are assigned to the large-cap index. All other stocks are assigned to the small-cap index. Stocks whose market caps fell between the 80% and 90% targets also are assigned to the mid-cap index.

FTSE

Large cap – 68% - 72%

Mid-cap – 86% - 92%

Small-cap – 97% - 99%

MSCI

Defining the Market Coverage Target Range for Each Size Segment

To define the Size Segment Indices for a market, the following free float-adjusted market capitalization Market Coverage Target Ranges are applied to the Market Investable Equity Universe:

Large Cap Index: 70% ± 5%.

Standard Index: 85% ± 5%.

Investable Market Index: 99%+1% or -0.5%.

The Mid Cap Index market coverage in each market is derived as the difference between the market coverage of the Standard Index and the Large Cap Index in that market.

The Small Cap Index market coverage in each market is derived as the difference between the free float-adjusted market coverage of the Investable Market Index and the Standard Index in that market.



Appendix B Mathematical Appendix

This section will be in two parts, first a brief description of the hyperbolic distribution and then a look at the same's extreme events .

Hyperbolic Distribution

The hyperbolic distribution interpolates between a Gaussian "body" and exponential tails:

$$P_H(x) \equiv \frac{1}{2x_0 K_1(\alpha x_0)} \exp\left(-\alpha \sqrt{x_0^2 + x^2}\right)$$

where the normalization $K_1(\alpha x_0)$ is a modified Bessel function of the second order. For small x compared to x_0 , $P_H(x)$ behaves as a Gaussian although its asymptotic behavior for $x \gg x_0$ is fatter and reads $\exp(-\alpha|x|)$.

From the characteristic function

$$\hat{P}_H(z) = \frac{\alpha x_0 K_1(x_0 \sqrt{1 + \alpha z})}{K_1(\alpha x_0 \sqrt{1 + \alpha z})}$$

we can compute the variance

$$\sigma^2 = \frac{x_0 K_2(\alpha x_0)}{\alpha K_1(\alpha x_0)}$$

and kurtosis

$$\kappa = 3 \left(\frac{K_2(\alpha x_0)}{K_1(\alpha x_0)} \right)^2 + \frac{12}{(\alpha x_0)} \frac{K_2(\alpha x_0)}{K_1(\alpha x_0)} - 3$$

Note that the kurtosis of the distribution is always between zero and three. In the case where $x_0 = 0$, the distribution changes to the symmetrical exponential

$$P_E(x) = \frac{\alpha}{2} \exp(-\alpha|x|)$$

with even moments $m_{2n} = \text{fac}(2n) \alpha^{-2n}$, which gives

$\sigma^2 = 2\alpha^{-2}$ and $\bullet = 3$. Its characteristic function reads

$$P_E(z) = \frac{\alpha^2}{(\alpha^2 + z^2)}$$

Zipf Plots and the Statistics of Extremes

One can rank a set of random variables x_i in decreasing order and estimate the n th encountered value which will be noted as $\Lambda[n]$ and $\Lambda[1] = x_{\max}$. The distribution P_n of $\Lambda[n]$ is the following full generality

$$P_n(\Lambda[n]) = N C_{N-1}^{n-1} P(x = \Lambda[n]) (P(x > \Lambda[n])^{n-1}) (P(x < \Lambda[n])^{N-n})$$

The above distribution means one has to first choose $\Lambda[n]$ among N variables (N ways), $n-1$ variables among the $N-1$ remaining as the $n-1$ largest ones C_{N-1}^{n-1} ways and then assign the corresponding probabilities to the configuration where $n-1$ of them are larger than $\Lambda[n]$ and $N-n$ are smaller than $\Lambda[n]$.

One can study the position $\Lambda^*[n]$ of the maximum of P_n and also the width of P_n via the second derivative of $\log P_n$ calculated at $\Lambda^*[n]$. This calculation simplifies in the limit where $N \rightarrow \infty, n \rightarrow \infty$ when the ratio n/N is fixed. At the limit, the following generalization can be found

$$P_\succ(\Lambda^*[n]) = n/N$$

The width w_n of the distribution is

$$w_n = \frac{1}{\sqrt{N}} \frac{\sqrt{1 - (n - N)^2}}{P(x = \Lambda^*[n])}$$

which shows that at the limit $N \rightarrow \infty$, the value of the n th variable is more and more sharply peaked around its most probably value $\Lambda^*[n]$ as given in the limit generalization above.

In the case of an exponential tail (for example the tails of a hyperbolic distribution), one finds $\Lambda^*[n] \cong \log(N/n)/\alpha$. On a log-log plot (a Zipf plot), exponentially distributed variables approximately follow a straight line, but with an effective slope which varies with the total number of points N : the slope is less and less as N/n grows larger. Note that inverting the log-log plot – amplitude on the x -axis and rank along the y – one obtains an estimate of the cumulative distribution P_\succ .

Via Cramer's function, when N is large, one can write the distribution of the sum of the N iid random variables as

$$P(x, N) \underset{N \rightarrow \infty}{\cong} \exp\left[-NS\left(\frac{x}{N}\right)\right]$$

where S is Cramer's function which gives some information about the probability of X even outside the 'central' region.



When the variance is finite, S grows as $S(u) \propto u^2$ for small u 's, which leads to a Gaussian central region. For finite u , S can be computed using Laplace's saddle point method

$$P(x, N) = \frac{1}{2\pi} \int \exp N \left(-iz \frac{x}{N} + \log \left[\hat{P}(z) \right] \right) dz$$

When N is large, the saddle point method is dominated by the neighborhood of the point z^* where the term in the exponential is stationary. The above saddle point results can be written as

$$P(x, N) \cong \exp \left[-NS \left(\frac{x}{N} \right) \right]$$

with $S(u)$ given by

$$\frac{d \log \left[\hat{P}(z) \right]}{dz} \Bigg|_{z=z^*} = iu$$

$$S(u) = iz^* u + \log \left[\hat{P} z^* \right]$$

which allows one to estimate $P(x, N)$ even outside the central region.



FOR MORE INFORMATION

Thomson Reuters Indices

Andrew Clark

Chief Index Strategist

Tel: +1 303.357.0557

andrew.clark@thomsonreuters.com

© 2009 Thomson Reuters. All rights reserved.
Republication or redistribution of Thomson Reuters content,
including by framing or similar means, is prohibited without the
prior written consent of Thomson Reuters. 'Thomson Reuters' and
the Thomson Reuters logo are registered trademarks and
trademarks of Thomson Reuters and its affiliated companies.

