Power laws in the number of copies per title in libraries

Lane, Daniel J.

2022

Lane, D. J. (2022). Power laws in the number of copies per title in libraries. Library and Information Science Research E-Journal, 32(2), 96-108.
https://dx.doi.org/10.32655/LIBRES.2022.2.2

https://hdl.handle.net/10356/164564

https://doi.org/10.32655/LIBRES.2022.2.2

Libres volume 32, issue 2, pages 96-108 (2022). https://doi.org/10.32655/LIBRES.2022.2.2

Power Laws in the Number of Copies Per Title in Libraries

Daniel J. Lane
Department of Political Science,
Northeastern Illinois University, USA
djlane@neiu.edu

ABSTRACT

Background. Power law distributions have been found in many contexts, both natural and social. This study examines one way of applying the power law distribution to characterize the breadth and accessibility of library collections.
Objectives. The collections of three libraries were analyzed for evidence of power law distributions with regard to the number of copies per title.
Methods. Data gathered at the three libraries were analyzed to obtain the coefficient of determination (R2) between the actual distribution of the number of copies for each library, and the number of copies predicted by a power law model. The slope of the best line fit represents the "power" of the distribution, which determines the shape of the particular power law distribution
Results. The results indicate that the number of copies per title in all three collections do in fact fit the power law distribution. But not merely that: although each library collection strongly resembles a power law distribution, the slope, or "power," of the distribution differs in each case, in a direction that indicates that the underlying theory is sound. The evidence is seen by comparing the results of two of the libraries studied here. As a small academic library with a policy directed at expanding its number of titles at the expense of number of copies, Kenyon College has the "flattest" slope or power; while Oak Park Public Library, a large public collection that must balance breadth with accessibility, has a much "steeper" slope to its distribution of titles to copies.
Contributions. The results and methods presented here not only allow librarians to understand their collections in a new fashion, but may also have implications for anyone seeking to manage collections into the future: for example, baskets of securities in financial markets, and investments in weapon systems by militaries. Additionally, the results provide support for the "preferential attachment" explanation of the power law distribution development, an important theoretical issue with possible relevance to both the natural and social sciences, particularly economics.

INTRODUCTION

Librarians face a common problem in library collection management: the twin imperatives of width (or breadth) and depth (or accessibility). On the one hand, a collection should have many different types of objects (width), and on the other, many copies of each type (depth). In what follows, I present empirical evidence demonstrating that librarians over time may arrive at a similar solution to this problem, that is, the distribution of copies of titles in a library

Published by WKW School of Communication & Information & NTU Libraries, Nanyang Technological University

follow a mathematical pattern known as a power law distribution. This is also of interest to non-librarians because power law distributions and related phenomena, such as Zipf's law for the populations of cities and Benford's law for the distribution of digits, have been found in innumerable contexts since their discovery by Pareto (1869/2003) (Eliazar, 2020). Such distributions have been claimed "in systems as diverse as volcanic eruptions, rockfalls, landslides, forest fires, solar flares, pulsar glitches, or biological extinctions" (Corral & González, 2019). The contribution of this paper is two-fold, both practical and theoretical. Firstly, if the distribution in the number of copies per title in several libraries closely follow the power law pattern, then this allows librarians—and indeed anyone managing an inventory of objects over the long term—to describe their holdings with a single number that immediately quantifies both the extent and accessibility of their collections, permitting direct comparison along those two dimensions. Secondly, this finding lends credence to the hypothesis that power law distributions are the result of what is known as a "cumulative advantage" or "preferential attachment" process, that is, power law distributions are the result of multiplicative stochastic processes. If that is so, this insight will have wide-ranging implications for both the natural and social sciences.

These are large claims, particularly to anyone unfamiliar with the literature on power law distributions. It may seem strange to many to assert that the actions of librarians, unknown to one another, may over time conform to a particular mathematical or statistical pattern. Yet, such is in fact the case when it comes to the population of cities. For example, economist Krugman (1995) pointed out that the fact that urban populations in a wide variety of nations are distributed according to a power law distribution has been known for over a century, and is not only "one of the strongest statistical relationships we know," but additionally is one "lacking any clear basis in theory" (Krugman, 1995).

Moreover, it is not only city populations that distribute themselves in this fashion, but—as Batty (2015) observed—so do the buildings themselves within cities. He wrote, "the heights of … buildings all show much bigger numbers of small objects than large but with the largest objects themselves dominating the size distribution" (Batty, 2015). This fact may appear odd to some readers, who may question how buildings constructed in many different historical or cultural contexts nevertheless conform to a mathematical model. What Batty's data show is that, without the builders themselves being aware of the outcome, architects and construction firms have come to arrange the architectural fabric of cities according to a certain mathematical pattern.

With regard to the populations of cities within a given nation, that mathematical pattern has been known since the work of physicist Auerbach (1913), and has since been confirmed so many times that it has been said to be "so spectacular in its exactness and universality that it is positively spooky" by Krugman (1996a). As described by Singer (1936), the regularity is that "if we increase the population limit in any proportion, the number of towns above this limit will decrease in a given multiple of this proportion". Two decades later Madden (1956) described the same regularity as the "rank-size rule," which in his notation was signified as

$$S_R = A/R^n$$

where "A is approximately the size of the largest city in a given group of cities, R is the rank of a given city, and S_R is the size of the city of that rank". The larger point is that, within the sizes of cities, there is a clear regularity between the population of that city and its rank within the larger system of which it is a part, can be expressed as an exponent. That is, Madden's $S_R = A/R^n$ can be rewritten as "$A = S_R R^n$," such that "n" represents the exponent by which a city's rank must be calculated in order to generate the population of that city. It is this exponent, or "power," that gives the power law distributions their name.

As Krugman (1996a) intimated by calling the fact "spooky," it is often mystifying to find such a statistical regularity in what seems to be the disorder of the urban scene. Cities are by their nature far beyond the direction of any one human life, so to find cities and their buildings following a very regular mathematical pattern can be disorienting. Batty (2015) pointed out that any surprise is likely not warranted as "the distribution of income follows [a] power law [distribution] … as many from Pareto in the late nineteenth century onwards have observed," and that fact "must be reflected in the sizes of [buildings] that we build and can afford" (Pareto, 1869/2003). Nevertheless, it must be allowed that the source of disorientation isn't simply that there is a ratio between large incomes and small ones, or large cities and small ones, or even large buildings and small ones, but that the ratio is so often precisely delimited. As Krugman (1996b) remarked, the reason why "the power law on city sizes is a disturbing one" is due to the fact that "the exponent is very close to one". It isn't so much that there is a regularity between rank and size when it comes to cities, but that the regularity is so absurdly regular.

In the literature, the chief explanation for that regularity is due to the mathematical effects of a process variously called "preferential attachment," "cumulative advantage," or—more poetically—the "Matthew effect". As Blasius and Tönjes (2009) noted, "power law distributions … typically arise in preferential attachment schemes." The concept was pioneered by G. Udney Yule's 1925 argument regarding the distribution of species within genera. As a recent gloss restated Yule's point, mutations "within a genus are more likely to occur in a genus with more species" (Mitzenmacher, 2004), leading to a power law distribution of species within a genus. Similarly, Simon (1995), citing Yule, demonstrated that such a mechanism must, mathematically speaking, generate power law distributions, whatever the underlying substrate. As more recent scholarship has pointed out, however, these were hardly the only times the preferential attachment mechanism was connected to power law distributions.

Barabasi (2012)—himself a scientist with a strong connection to preferential attachment—asserted that the process "made its first appearance in 1923 in the celebrated urn model of the Hungarian mathematician György Pólya." Others however have dated the discovery to earlier work. In an article reviewing the many times and places the connection between preferential attachment and power law distributions has been made, Simkin and Roychowdhury (2011) not only noted how in 1999—unaware of their predecessors—Barabasi and Albert "proposed the Preferential Attachment model … in order to explain the power-law distribution of the connectivity in the World Wide Web and other networks," but also how, in 1916, "Smoluchowski considered a colloid suspension of particles, which collide with each other and stick together and resulting clusters in turn collide with each other forming larger clusters"(Simkin & Roychowdhury, 2011). Indeed, they argued not only that the basic point may be traced back to the work of pioneering statistician Francis Galton, but also that Bru, Jongmans, François and Seneta, (1992) have cited a work by Cournot from 1847 as the earliest instance. Regardless of the truth of the matter, in general the preferential attachment model describes "a positive feedback mechanism in which new elements of the system … are added with a probability proportional to the abundances of how many are already there," so that the "more there are, the more of that type are going to be added" (West, 2017). Furthermore, undoubtedly the literature demonstrating a strong link between the preferential attachment mechanism and power law distributions is an exceedingly lengthy one by scientific standards.

That is one reason why Batty's (2015) demonstration that skyscraper heights follow a power law distribution is such powerful evidence in favor of the preferential attachment model. His evidence was derived from The Skyscraper Center's Global Tall Building Database, which contains 4,325 buildings dating from 1909 to 2020, constructed in cities from New York to Shanghai. Given both the depth and breadth of this evidence, there can be little question of collusion between constructors so widely separated in time and space, even in the sense of historical or cultural or other connections, at least aside from the common technological ability to build tall structures. Instead, it seems far more plausible that the reason for the underlying commonality is due to preferential attachment, that is, cities with skyscrapers are likely to attract more, and taller, skyscrapers over time.

The ultimate significance of this point is far from trivial. It is one of the paradoxes of probability that seemingly "random"—that is, stochastic—outcomes are actually better capable of being determined than outcomes that, to the individual, appear more orderly. Thus, the great Soviet statistician Andrey Kolmogorov remarked that the "epistemological value of probability theory is based on the fact that chance phenomena, considered collectively and on a grand scale, create non-random regularity" (Gnedenko & Kolmogorov, 1954). By demonstrating that preferential attachment is a common process in not merely natural phenomena, but also in social phenomena, it is possible to calculate its effects and, over the long term, mitigate them in situations where, if unchecked, may foil the best-laid plans of people, and specifically librarians. Hence, if library collections can also be shown to conform to the power law distribution model, then that may also be supporting evidence for the "preferential attachment" explanatory mechanism—particularly since, as I will show, there is strong reason to believe that just such a mechanism underlies the development of library collections.

METHOD

To find the data for this project, I corresponded with many libraries, most of which declined to collect or provide the necessary data. I managed to obtain data concerning the number of copies per title owned by the libraries from three sources: the library of Northeastern Illinois University, a public commuter university in Chicago, Illinois; the library of Kenyon College, a private college in rural Ohio; and the public library of Oak Park, Illinois, a Chicago suburb best known as the home of novelist Ernest Hemingway and architect Frank Lloyd Wright. None of the three libraries are particularly known for having an extensive collection, and all three likely hold more titles pertaining to locally relevant subjects than other libraries. Oak Park's collection, for example, is noted for holding titles related to both Hemingway and Wright. Both of the college libraries were roughly similar in size, and both were dwarfed by the Oak Park collection.

Still, while each of these libraries differed in some ways from other libraries, there were no obvious substantial difference from other libraries of their respective types. The results of this analysis are likely to hold for these types of libraries elsewhere. If certain findings in the power law literature are correct, the results may hold for any library (Zipf, 1949). After all, librarians have always had to weigh the extent of their collections against the accessibility of those collections, even if they did not have a means of quantifying that dilemma.

This is not to say that individual libraries do not have additional factors that affect their collection development. It has been helpfully pointed out by a reviewer, for instance, that academic libraries may have additional pressures to maintain multiple copies than other libraries. Public libraries, in turn, may have multiple branches, again exerting a pressure towards multiple copies—a possibility I have guarded against here by specifically targeting only the main branch of the Oak Park library. Against that, however, a librarian at Kenyon College informed me that it “has been the library's policy for many years now not to purchase multiple copies of any titles for the collection, largely as a cost savings measure and for lack of demand”—a policy that, as we shall see, is apparently reflected in the collection. Aside from that statement, there is no evidence to suggest that these particular libraries have any unusual pressures to acquire or avoid acquiring multiple copies of titles.

Having assembled the data, I next proceeded to set a lower boundary for the number of copies per title I would consider. As remarked by Batty (2015), it is usual in the power law distribution literature to “simply deal with the heavy or upper tail of the distribution,” as that is both computationally easier and is where the focus of investigation lies. (“Strictly speaking,” Krugman (1996) noted, “it is only the upper tail that obeys” the power law distribution, because as urban centers become less populated they eventually stop being urban.) In this case, I limited the datasets from Northeastern Illinois and Kenyon College to titles with three or more copies, while I limited Oak Park’s collection—which again was much larger than the others—to titles with ten or more copies. Limiting the datasets in this way enabled me to grapple with what were still very large datasets with thousands of titles and even more copies.

I next entered the data into a common spreadsheet program (Numbers for Macintosh). There, I could make use of the program’s ability to fit data sets to distributional curves using the coefficient of determination function, often referred to as R^2 (or R squared). R is the correlation between the number of copies of each title in the collection, versus the number of copies predicted by a power law model (determined by the best fit line). R^2 can be interpreted as “the proportion of the variation [i.e., variance] in one variable [i.e. the number of copies] that is accounted for by the other” [variable, the predicted number of copies following the model] (Everitt & Skrondal, 2010). Admittedly, in some circles there has been a push towards using more precise statistical tools when attempting to demonstrate the presence of a power law distribution (Clauset, Shalizi & Newman, 2009; Cirillo, 2013; and Corral & Deluca, 2013). In general, the statistical literature appears to point towards using a Kolmogorov-Smirnov test as best practice. The present study then is designed as an initial foray into applying statistical knowledge to library collections; presumably, as practice evolves better techniques will become available.

RESULTS

The bar charts in Figures 1 to 3 show the rank-frequency distributions of titles for the three libraries. The x-axes represent the rank order of individual titles, with the title having the most number of copies assigned rank 1. The y-axes represent the number of copies. The curve labeled “Power” in the charts suggest how well the distributions fit the power law distribution.

To analyze statistically how well the distributions fit the power law distribution, the log (base n) of both the number-of-copies and the ranks were calculated, and the relation between *ln(number or copies)* versus *ln(rank)* were plotted (shown in Figures 4 to 6). A straight line curve indicates that the distribution fits a power law distribution. The correlation (Pearson r) between *ln(number or copies)* and *ln(rank)* when squared yields the coefficient of determination (R^2) that quantifies how well the distributions fit the power law. The coefficient of determination values for the three libraries are listed in Table 1.

Libres volume 32, issue 2, pages 96-108 (2022). https://doi.org/10.32655/LIBRES.2022.2.2

Linear regression was also carried out to determine the coefficients for the model:

ln(number or copies) = c + b * ln(rank)

The regression models for the three libraries are listed in Tables 2 to 4. The coefficient *b* represents the slope of the line, and can be converted to the "power" of the relationship. An example of this calculation is provided later.

In all three libraries, the spreadsheet program found that the power law distribution was in fact a better fit than such distributions as the exponential. Each library however did not match the power law distribution equally well (see Table1). Kenyon College's library, for example, had the worst fit, with a coefficient determination of 0.77—which is still a very high measure for social data. (As biostatistician Enders (2013) has put the point, an R^2 of 0.35 "might be a very high portion of variation to predict in a field such as the social sciences," but "in other fields, such as the physical sciences, one would expect R^2 to be much closer to 100 percent".) This is an interesting finding because, as mentioned above, it has long been the library's policy to avoid purchasing multiple copies: in short, the policy does appear to be reflected in the data.

Table 1. Libraries by Coefficient of Determination (R^2) value

Library	Coefficient of Determination (R^2)
Kenyon College Library	0.77
Northeastern Illinois University Library	0.92
Oak Park Public Library	0.98

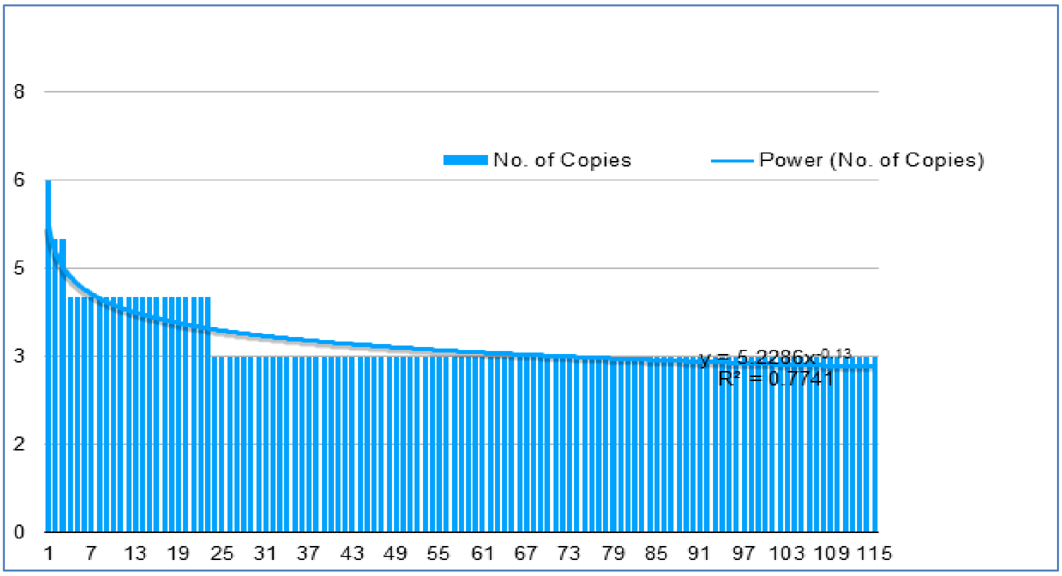

Figure 1. Kenyon College Library: rank-frequency distribution of titles (minimum 3 copies)

Source: Hudson, Chris. Director of Collection Services. Email. 2020.

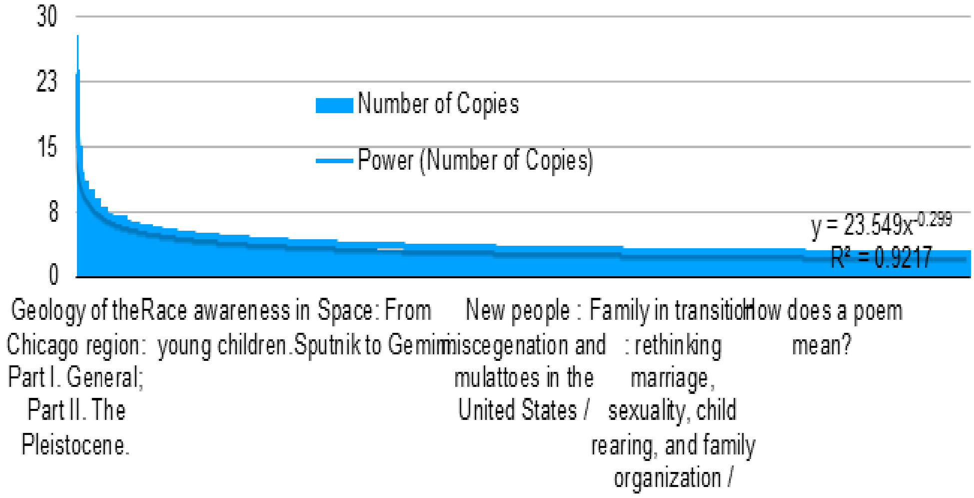

Figure 2. Northeastern Illinois University Library: rank-frequency distribution of titles (minimum 3 copies)

Source: Marszalik, Elizabeth. Director of Collections & Technology, Oak Park Public Library. Email. 2020.

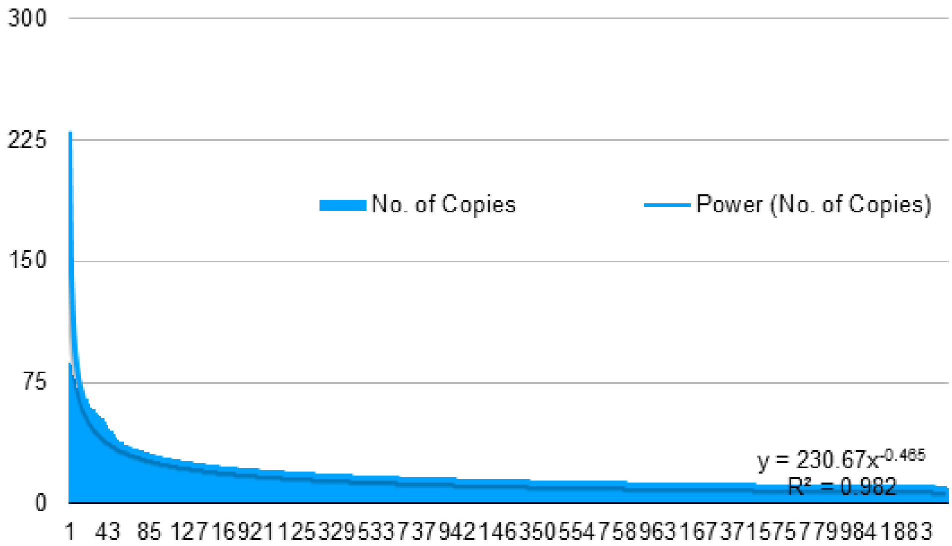

Figure 3. Oak Park Public Library: rank-frequency distribution of titles (minimum 10 copies)

Source: Marszalik, Elizabeth. Director of Collections & Technology, Oak Park Public Library. Email. 2020

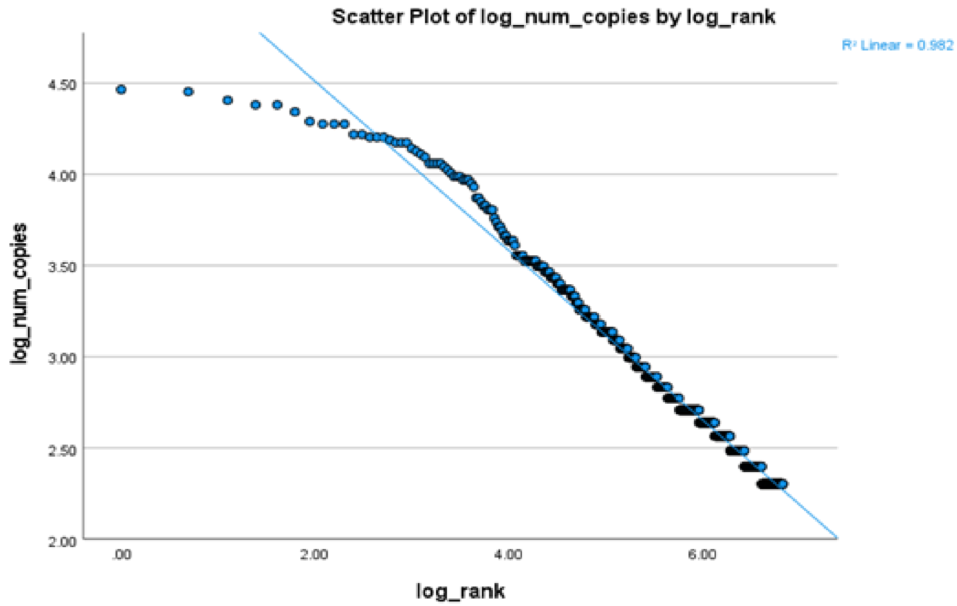

Figure 4. Scatter plot for Oak Park

Table 2. Linear regression model for Oak Park

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	5.44	.012		442.78	<.001
	ln(rank)	-.47	.002	-.99	-223.55	<.001

a. Dependent Variable: ln(num_copies)

What happens in the absence of a specific policy against acquiring multiple copies may be reflected in the R^2 of the other two libraries. These appear much closer to what Enders (2013) said is the standard for the physical sciences: for Northeastern the R^2 was 0.92, while for Oak Park it was 0.98. These are very strong findings, perhaps indicating the presence of what the physical sciences call an "attractor": "an attractor is what the behavior of a system settles down to, or is attracted to" (Crutchfield, Doyne, Norman & Robert, 1986). As I hope to have shown in the discussion above, there are good theoretical reasons for supposing that libraries should display such behavior because it is in line with findings regarding other cumulative processes. Nevertheless, the power of the statistical signal here is difficult to overstate, particularly since it concerns a social, not a natural, process: it is difficult to believe that any librarian or collections manager could have been aware of the statistical effects of acquisition policies over time, so it is surely not the case that these figures were deliberately targeted. I address this point more fully in the "Conclusion" section below.

Once I established that there is reason to think that the collections of at least these libraries strongly resemble a power law distribution, I proceeded to establish the "power" of the respective power law distributions within each library. These are reproduced below.

For Oak Park

ln(num copies) = 5.44 – 0.47 ln(rank)
= ln(5.44 * $rank^{-0.47}$)
num copies = 5.44 * $rank^{-0.47}$
= 5.44 / rank $^{0.47}$
Power = 0.47

For Northeastern Illinois

Power = 0.30

For Kenyon College

Power = 0.13

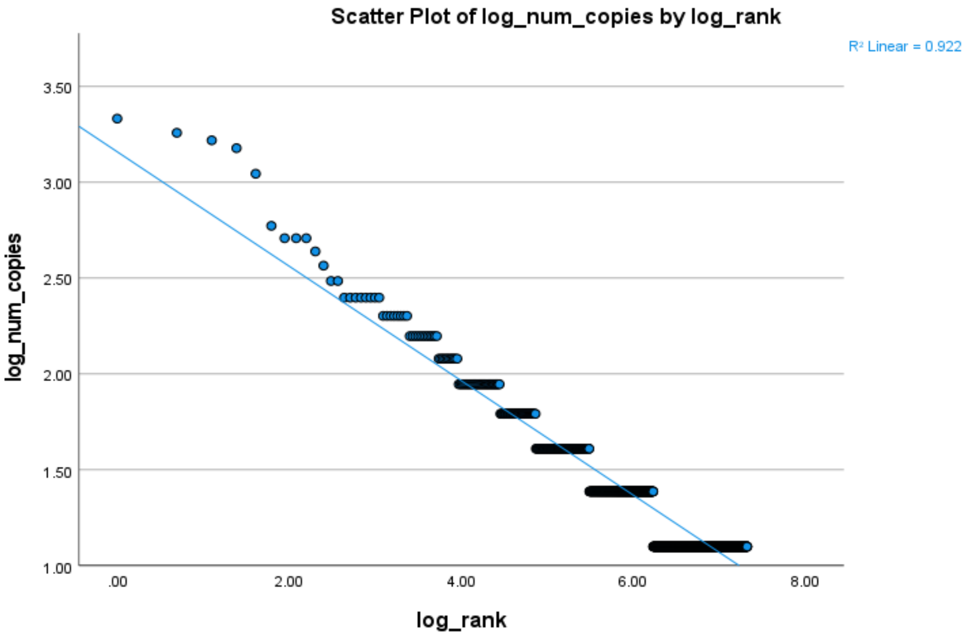

Figure 5. Scatter plot for Northeastern Illinois

Table 3. Linear regression model for Northeastern Illinois

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	3.16	.014		220.14	<.001
	log_rank	-.30	.002	-.96	-133.23	<.001

a. Dependent Variable: ln(num_copies)

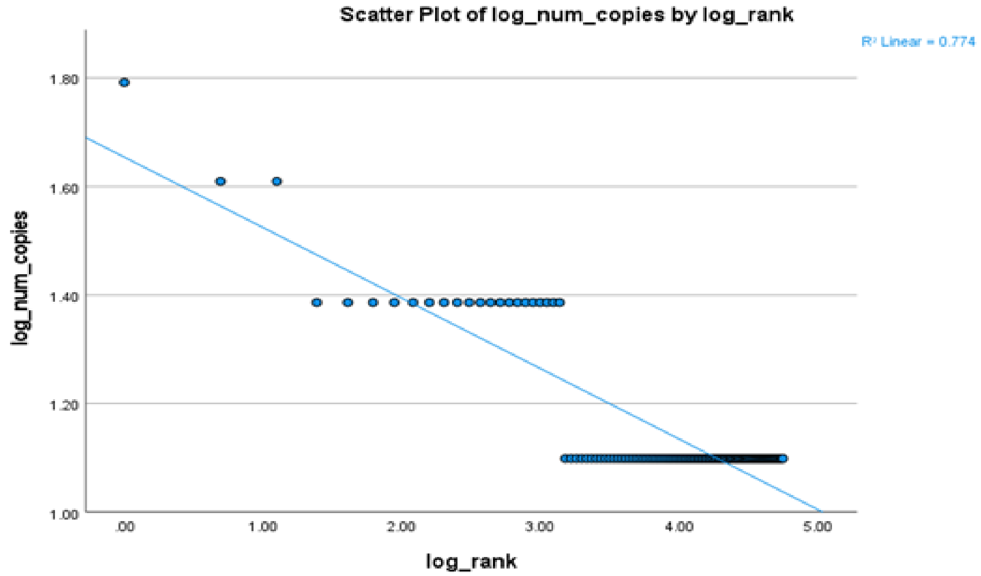

Figure 6. Scatter plot for Kenyon College

Table 4. Linear regression model for Kenyon College

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	1.65	.026		64.43	<.001
	ln(rank)	-.13	.007	-.88	-19.68	<.001

a. Dependent Variable: ln(num_copies)

As can be seen, Kenyon College had the "flattest" slope, which again is in line with the anecdotal evidence regarding its collections policy, while the Oak Park Public Library had the "steepest" slope, which may be explicable due to the fact that it is a large publicly-accessible collection. In other words, Kenyon's policy was explicitly driven by breadth of coverage, or number of titles—unsurprising as an academic institution—while Oak Park, as a public library, is presumably also concerned with accessibility (or number of copies). Hence, it may be possible to employ the "power" in power law distributions to characterize libraries in terms of both their breadth of coverage and accessibility.

CONCLUSION

While not conclusive given the small sample size, the results demonstrate that there is reason to investigate whether there is a relationship between the number of titles and the number of copies libraries possess for at least two reasons. The first is a practical matter concerning library operations: calculating these numbers will allow librarians to understand their own collections better, and perhaps thereby enable them to navigate between the two imperatives of breadth (the number of titles) and accessibility (the number of copies). The other reason is more theoretical, because it is the same reason that there is a basis for investigating the relationship between the sizes of cities or the sizes of buildings: by understanding how the ratio of copies to titles develops over time, it may be possible to gain insight into how collections in general—whether of city populations, or anything else—may develop.

Still, the theoretical possibilities raised by this research should not distract us from the main result, which is that by measuring a "copy to title ratio," say by the power of the power law in the collection (assuming that there is one), librarians can understand their collections better by comparison with other libraries. The pragmatic consequence for librarians—and anyone else charged with deploying resources over a long time horizon, like financial managers or defense planners—is that this paper provides the seed of a practical method for evaluating present resources and acquiring more in the future. One way to think of this is that it provides a kind of "batting average" for libraries: a shorthand statistical method of evaluating the diversity of a collection as against its utility for users. In other words, this method provides a single number—the "power" in the power law—that will quantify a collection's balance between its coverage of all possible library materials and the ability of a patron to access those materials. Surely that is a hugely significant piece of knowledge for any librarian.

That consequence, however, may go further than libraries: one way to conceive of libraries, after all, is as collections of resources, or capital. In that sense then librarians may be thought of as analogous to financial traders: the diversity of a library collection may be thought of as the diversity of, say, a trading operation's risk, while having resources tied up in multiple copies may be thought of as having capital tied up in single investments. Hence, although developed first for libraries this method could quite easily be generalized for any situation in which the degree of exposure to capital—that is, how many titles—must be balanced against the concentration of capital—that is, how many copies. This more abstract conception of the method thereby leads to more theoretical considerations.

The first concerns the very basis of the discovery of power law distributions by Pareto (2003) in economic data. For some decades economists have noted increasing economic inequality in many developed countries: as one not-so-recent paper has put the point, the "share of total income earned by the top 1% of families was less than 10% in the late 1970s, but now exceeds 20% as of the end of 2012" (Zucman & Saez, 2014). Over time, in other words, the distribution of income and other economic measures has become more and more sharply distinguishable: in mathematical terms, the slope or power of the distribution has become steeper. Understanding power law distributions thereby enhances understanding of the economy in a general sense.

The more specific question regarding power law distributions concerns whether the claims of a number of scientists and statisticians are correct regarding how such distributions are connected, if at all, with preferential attachment mechanisms. In this connection the importance of the Kenyon College Library's acquisition policy is hard to overstate: recall that the library's explicit policy was to focus on breadth at the expense of depth, and that is reflected in the data. Since the preferential attachment model stipulates that power law distributions result from multiplicative, rather than additive, processes, then the school's acquisition policy of adding single copies rather than multiple copies appears to provide an excellent empirical example of how the preferential attachment model could work in practice, indicating that libraries may be excellent illustrations of the preferential attachment theory of how power law distributions develop in reality. Because, as Krugman (1996a) mentioned, the appearance of such distributions in data is largely a mystery, any such evidence is significant.

In that sense, libraries supply terrific laboratories for understanding how collections—that is, capital—accumulate over time. Libraries contain knowledge in more than their books. I invite libraries to collaborate with me to explore power law and other distributions in their collections.

REFERENCES

Auerbach, F. (1913). The Law of Population Concentration. *Scientific American Supplement, LXXVI*. Munn & Co., New York.
Barabási, A. (2012). *Luck or Reason*. Nature, 489, 507–508.
Batty, M. (2015). Competition in the Built Environment: Scaling Laws for Cities, Neighbourhoods and Buildings. *Nexus Network Journal: Architecture and Mathematics, 17*, 831–850. https://doi.org/10.1007/s00004-015-0268-2
Blasius, B., & Tönjes, R. (2009.) Zipf's Law in the Popularity Distribution of Chess Openings. *Physical Review Letter, 103*, 218701.
Bru, B.; Jongmans, F.; & Seneta, Ee. (1992). I.J. Bienaymé: Family Information and Proof of the Criticality Theorem. *International Statistical Review, 60*(2), 177–183. https://doi.org/10.2307/1403648
Cirillo, P. (2013). Are your data really Pareto distributed? *Physica A: Statistical Mechanics and its Applications, 392*, 5947-5962.
Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev. 51*, 661–703.
Corral, A., & Deluca, A. (2013). Fitting and Goodness-of-Fit Test of Non-Truncated and Truncated Power-Law Distributions. *Acta Geophysica, 61*, 1351-1394. https://doi.org/10.2478/s11600-013-0154-9
Corral, Á.,González, Á. (2019). Power law size distributions in geoscience revisited. *Earth and Space Science, 6*, 673– 697. https://doi.org/10.1029/2018ea000479
Crutchfield, J. P., Doyne, F., Norman, H. P., & Robert S. S. (1986). Chaos. *Scientific American, 255*, 46-57.
Eliazar, I. (2020). *Power laws: A statistical trek*. Springer International Publishing.
Enders, F. B. (2013). Coefficient of Determination. *Britannica*. https://www.britannica.com/science/coefficient-of-determination.
Everitt, B.S., & Skrondal, A. (2010). *The Cambridge Dictionary of Statistics*. 4th Ed. Cambridge University Press, London.
Gnedenko, B., & Kolmogorov, A. (1954). *Limit distributions for sums of independent random variable*s. New York, Addison-Wesley Pub. Co.
Krugman, P. (1995). *Development, Geography, and Economic Theory*. Massachusetts Institute of Technology, Boston MA.
Krugman, P. (1996a). *The Self-Organizing Economy*. London, Wiley-Blackwell.
Krugman, P. (1996b). Confronting the mystery of urban hierarchy. *J. of the Japan and Int. Econ..10*, 399-418.
Madden, C. H. (1956). On Some Indications of Stability in the Growth of Cities in the United States. *Economic Development and Cultural Change, 4*(3), 236-252. www.jstor.org/stable/1151775
Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. *Int Math, 1*, 226–51.
Pareto, V. (2003). *On the distribution of wealth and income: The economics of poverty and inequality*. (F. A. Cowell, Trans.). Houndmills. (Original work published 1896)

Simkin, M.V., & Roychowdhury, V.P. (2011). Re-Inventing Willis. *Physics Reports, 502*, 1-35. https://doi.org/10.1016/j.physrep.2010.12.004

Simon, H. A. (1955). On a Class of Skew Distribution Functions. *Biometrika, 42*(3/4), 425-440.

Singer, H. W. (1936). The 'Courbe des Populations': A Parallel to Pareto's Law. *The Economic Journal, 46*(182), 254-263. https://doi.org/10.2307/2225228

West, G. B. (2017). *Scale: The Universal Laws of Growth, Innovation, Sustainability and the Pace of Life in Organisms, Cities, Economies, and Companies*. New York, Penguin Press.

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

Zucman, G., & Saez, E. (2014). Exploding wealth inequality in the United States. *VoxEU*. https://voxeu.org/article/exploding-wealth-inequality-united-states