The Open-Factor: Toward impact-aligned measures of open-access ebook usage

E. S. Hellman
Free Ebook Foundation

Abstract

A statistical analysis of usage data for open-access ebooks from two different publishers and from a free ebook distribution platform indicates that open-access ebook usage is distributed following log-normal statistics. Using a quantity related to the logarithm of download counts, dubbed the “open-factor”, will measure impact in alignment with the goals of open-access ebook publishing.

Introduction

Open access publishers need ways to measure their impact. Since the whole point of removing toll-access barriers is to increase access to information, open access publishers look to their usage logs for validation of their efforts and mission. Unit sales and profits do not align as well with the goals of open-access publishing. This requirement is not new to the digital world; libraries have always needed ways to measure the effectiveness of their collections. Books are for use,[1] and if a print collection doesn’t circulate, then readers aren’t having their needs met. In the absence of validated measures of impact, presentation of cherry-picked statistics will inevitably undermine advocacy around the impact of open-access content.

Counting downloads or circulations may seem easy to do, but there are pitfalls to avoid when interpreting the resulting data. Counting hits on a web server is particularly fraught with danger of misinterpretation. For example, a research item might be accessed multiple times in the course of a scholar's use, on multiple devices. In some cases, an item might require re-downloading because of a poor internet connection or a bad user interface so that an increased hit count is associated with a poor access rather than a successful access. Software agents such as indexing spiders routinely download materials; although many of these agents can be excluded from download counts, all downloads are ultimately done by software and it can be impossible to distinguish software serving an individual from software operating as a robot.

[1] The first of Ranganathan’s *Five laws of Library Science*.

As open-access ebooks become more common, the complexities of measuring their impact have become apparent. Compared to journal articles, ebooks are delivered and used diverse formats and modalities. They may be browsed, downloaded, and segmented. Ebooks cost more to produce than individual journal articles, so publishers have a bigger stake in justifying open access. As might be predicted, there have been exuberant press releases and white papers presenting gaudy download statistics without much context or statistical grounding.[2]

Supporters of open access publishers, such as libraries, also need impact validation. It's reasonable for a library administrator to ask "How many times have users from the library used items from Publisher X". In answering questions like these, publishers want to present their statistics in a favorable light. Initiatives such as COUNTER[3] have sprung up to help publishers and libraries produce and analyze statistics that purport to compare usage across publishers and across libraries. Physical circulation counts can be also be problematic, depending on how the data is used. Circulation of a particular item is affected by time off shelf, which may depend on many factors completely unrelated to the item's usage. Reshelving stats can also undercount because of "helpful" patrons. Any of these problems may lead to poor decisions if the data are used to inform collection development and management processes.

Several groups have begun tackling the problem of measuring impact for open access ebooks. One strand of activity has drawn from the field of "alt-metrics". Commercial services such as Altmetic.com[4] aim to provide data relevant to ebook impact. Projects such as *HIRMEOS*[5] have similarly begun to fill the data vacuum. A study of data for UCL Press open-access ebooks was a good first step towards providing much needed grounding and context.[6]

Objectives

Data on usage informs discussions about how open access resources are discovered and accessed. It's reasonable to ask questions about the effect on usage of license, format, subject area, and indexing. The author recently participated in such a study, focusing on open-access scholarly ebooks, funded by the Andrew W. Mellon Foundation. A joint effort of University of

[2] Some examples of the genre include "University College London Press Passes 1 Millionth Open Access Book Download", Porter Anderson, (*Publishing Perspectives*, May 24, 2018) https://publishingperspectives.com/2018/05/university-college-london-ucl-press-million-open-access-downloads/ and "The OA Effect: How Does Open Access Affect the Usage of Scholarly Books?", Christina Emery, Mithu Lucraft, Agata Morka, Ros Pyne, (Springer Nature, November 2017) https://media.springernature.com/full/springer-cms/rest/v1/content/15176744/data/v3

[3] "Project COUNTER." https://www.projectcounter.org/. Accessed 26 Nov. 2018.

[4] "Altmetric." https://www.altmetric.com/. Accessed 29 Mar. 2019.

[5] "Hirmeos Project – High Integration of Research Monographs in the European Open Science infrastructure" https://www.hirmeos.eu/

[6] "Getting the best out of data for open access monograph presses: A case study of UCL Press" Tama Leaver, Lucy Montgomery, Cameron Neylon , Alkim Ozaygen, Humanities Commons (2018) http://dx.doi.org/10.17613/M6HQ3RZ0T

Michigan Press, Open Book Publishers, and the Free Ebook Foundation[7] looked at a variety of data, including server log data, to determine how and where the ebooks are discovered and used. A full report of our conclusions can be found in Michigan's *Deep Blue* repository.[8] Cumulative download data on open access ebooks downloaded via Unglue.it[9] was also examined.

One set of objectives of this study was to compare attributes of the open access ebooks across the two participating publishers' catalogs, to learn what factors promoted sales and usage. For example, we wanted to know if usage was correlated to sales, and what effect the price had on sales. To answer these and other questions, we collated bibliographic, usage and sales data from the disparate systems used by the two publishers.[10] Google Analytics[11] was used to gather book download and webpage usage data for both publishers; Google Analytics data was validated by comparison to data from server logs. In particular, we were careful to aggregate and assign urls to specific books to account for the different web-page layouts used by the two publishers. Sales and usage data were taken from roughly equivalent time periods; a small burst of usage for each book appeared at its initial publication, but for most books this burst was eclipsed by usage over study's time span, and we did not attempt to correct for its effect. The python libraries NumPy[12], SciPy[13] and Pandas[14] were used in data analysis, curve fitting and regression analysis.

Results

Figure 1 shows a scatter plot of gross sales versus downloads, normalized by the length of the time period measured, for the 118 titles[15] we studied. The linear regression analysis, if we are to believe it (the standard error is 0.0176 for the slope of 0.0437), indicates a weak correlation between online traffic and sales. An unscientific analyst would be tempted to remove a few "outliers" at the right of the figure, resulting in a much stronger correlation of sales with online traffic.

[7] "Mapping the Free Ebook Supply Chain - Michigan Publishing." https://www.publishing.umich.edu/projects/mapping-the-free-ebook/. Accessed 26 Nov. 2018.
[8] "Mapping the Free Ebook Supply Chain: Final Report to the Andrew W ...." 16 Jun. 2017, https://deepblue.lib.umich.edu/handle/2027.42/137638. Accessed 26 Nov. 2018.
[9] "Unglue.it." https://unglue.it/. Accessed 28 Nov. 2018.
[10] The work required to assemble the data sets was greatly in excess of what was anticipated!
[11] "Google Analytics." https://analytics.google.com/analytics/web/. Accessed 26 Nov. 2018.
[12] "NumPy — NumPy." http://www.numpy.org/. Accessed 28 Nov. 2018.
[13] "SciPy.org." https://www.scipy.org/. Accessed 28 Nov. 2018.
[14] "Pandas." https://pandas.pydata.org/. Accessed 28 Nov. 2018.
[15] Data was collected for 138 titles, taken from the catalogs of University of Michigan Press, including its Open Humanities Press imprint, and of Open Book Publishers. 20 titles of the 138 were excluded from the present statistical analysis either because sales data or page view data was unavailable in a form that could be compared to the rest of the list.

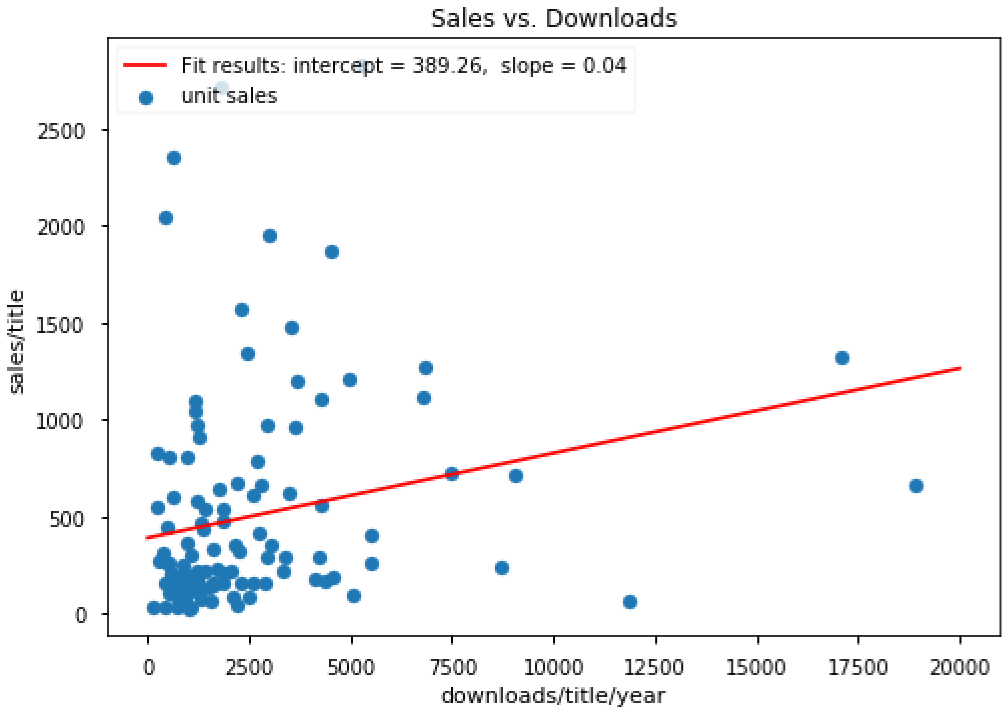

Figure 1. Sales vs. downloads for the 118 titles included in this study. The correlation coefficient is 0.224

Other statistical measures as well as common sense lend skepticism to the observed correlation. Statistical distributions can be characterized by their variance and kurtosis. Daily download statistics can be examined to determine if the are described by specific distributions. For example, normally distributed data has a variance of σ^2 and an excess kurtosis of 0. We computed an excess kurtosis of 100-500 in the daily download numbers for popular books in our data set. This is characteristic of “spiky’ data, not data that follows familiar statistical distributions.

Statistical measures of both downloads and sales are dominated by titles with sales that are not representative of the collection as a whole: by selecting outliers carefully, we can make the numbers tell any story we want.

Is there anything that we can learn about the collections by looking at the downloads or sales numbers? Do the numbers MEAN anything? If one book has been downloaded more than another book, does it mean that the book was easier to find, or that the publishing press had done a better job of promotion or search engine optimization, or that it was more "viral"?

A better understanding of the download data is obtained by examining the distribution of downloads across the collections and their distribution over time. We looked at the daily download count for individual books and computed averages, variances, and kurtosis. Most of the books exhibited very large kurtosis in the daily download counts; only 2 books had kurtosis below 2 (kurtosis would be 0 for a normal statistical process). This indicated that comparing averages or variances was poorly justified. In other words, averages and variances would be typically dominated by the download counts on a very small number of days in the year.

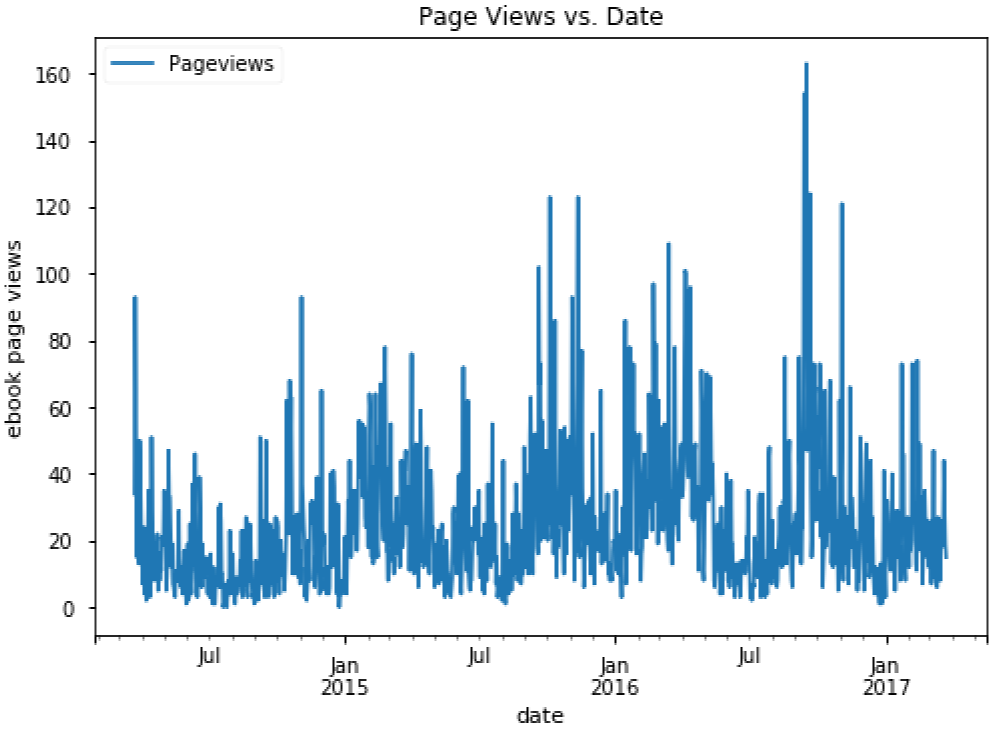

Figure 2a. Daily ebook page views for one OA ebook, measured by Google Analytics.

Figure 2a shows a plot of daily page views for a less “spiky” book in the study, exhibiting the irregular nature of the data even though the kurtosis for this book is still 6. Seasonal usage is evident, suggesting its use in courses tied to an academic calendar.

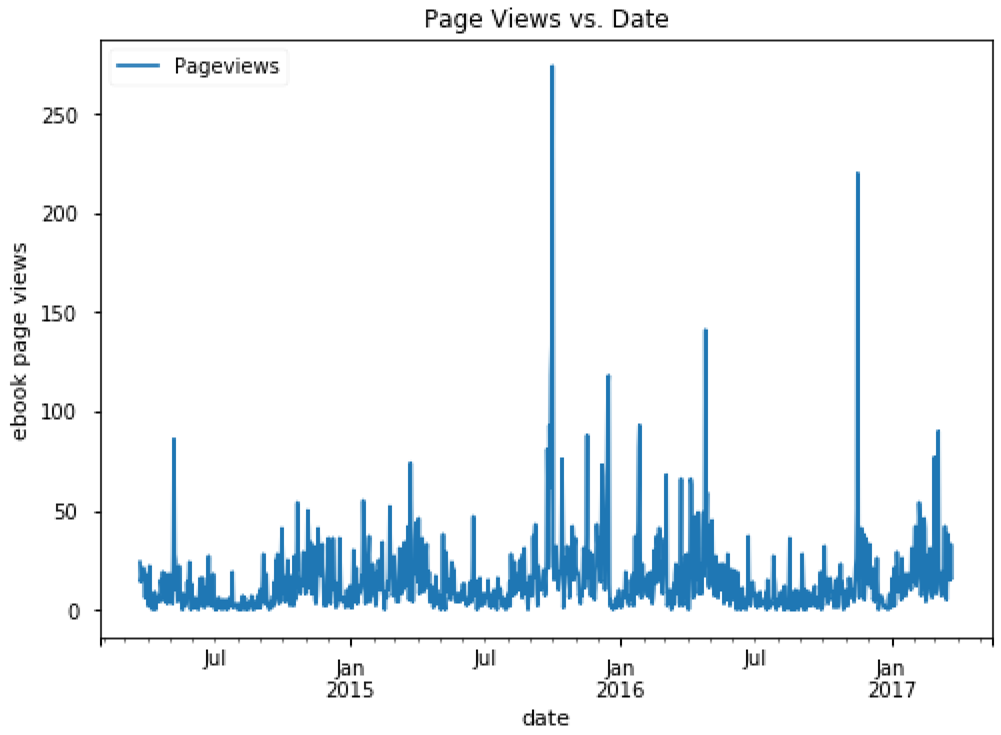

Figure 2b. Daily ebook page views for a more typical OA ebook in the study.

Figure 2b shows page view data for a more typical ebook in the study, with kurtosis=53.8. Seasonal variation is less apparent; spikes of usage occur, apparently at random times.

Analysis of the page load data indicated that the data could be most usefully characterized by log-normal distributions. Figure 3a and 3b show the distribution of daily page loads for the ebooks examined in figures 2a and 2b. It can be seen that all the "spikiness" seen in figure 2 is well described by a normal distribution in the natural logarithm of the daily page views. There is a departure from the log-normal fit at small counts where the continuous log-normal distribution is a poor approximation of the discrete measurements.[16]

[16] You can't take the logarithm of zero! For a review of discrete analogs of the log-normal distribution, see "*Generating discrete analogues of continuous probability distributions-A survey of methods and constructions*", Subrata Chakraborty, *Journal of Statistical Distributions and Applications* **2**(6), 2015. https://doi.org/10.1186/s40488-015-0028-6

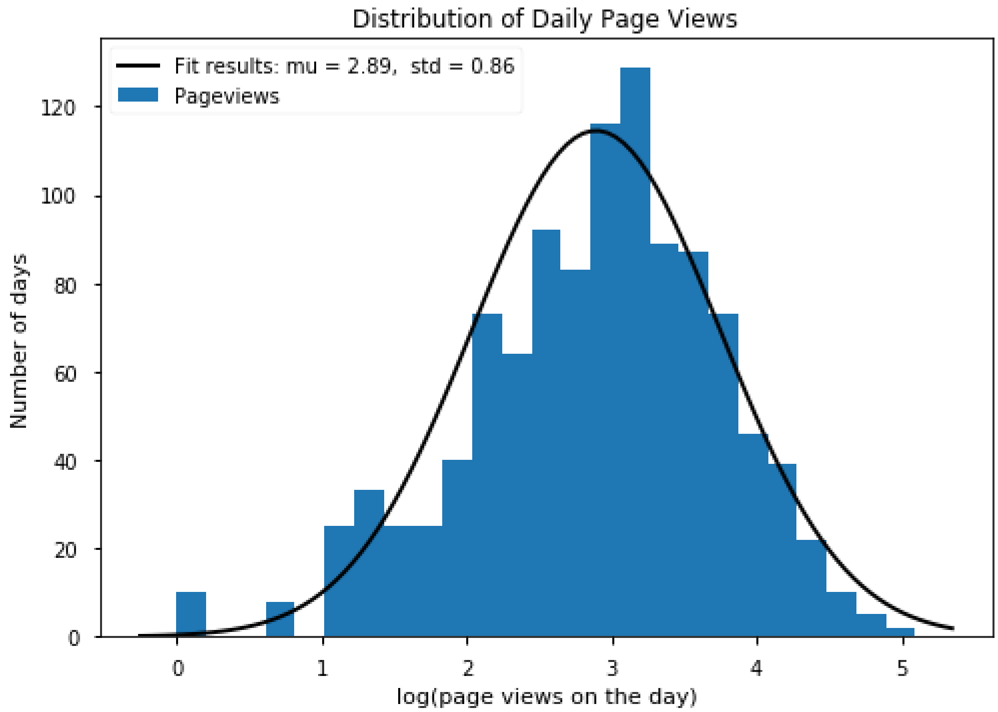

Figure 3a. Distribution of daily page views for the book from figure 2a.

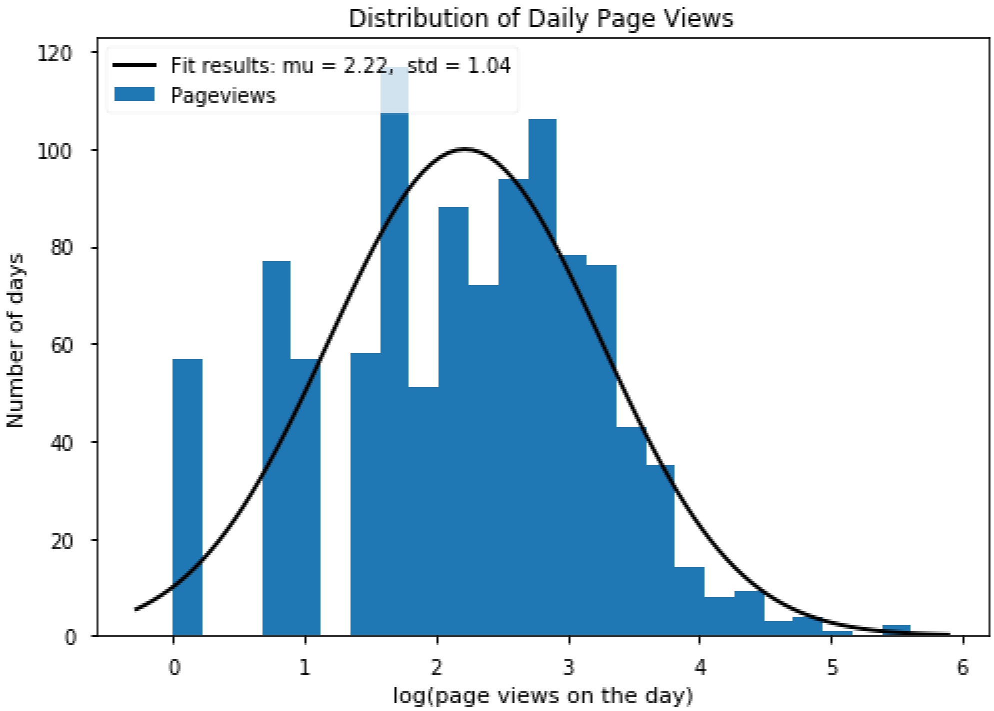

Figure 3b. Distribution of daily page views for the book from figure 2b.

Building on the insight that ebook page views, downloads, and related quantities can be described by log-normal distributions, we can return to our analysis of downloads and unit sales across the 150 books in the study. Figure 4 shows the histogram of the *logarithm* of the download count per title. It is well fit by the classic bell curve of a normal distribution. Figure 6 shows a similar analysis for unit sales. A histogram of the logarithm of unit sales per year also looks like a normal distribution.

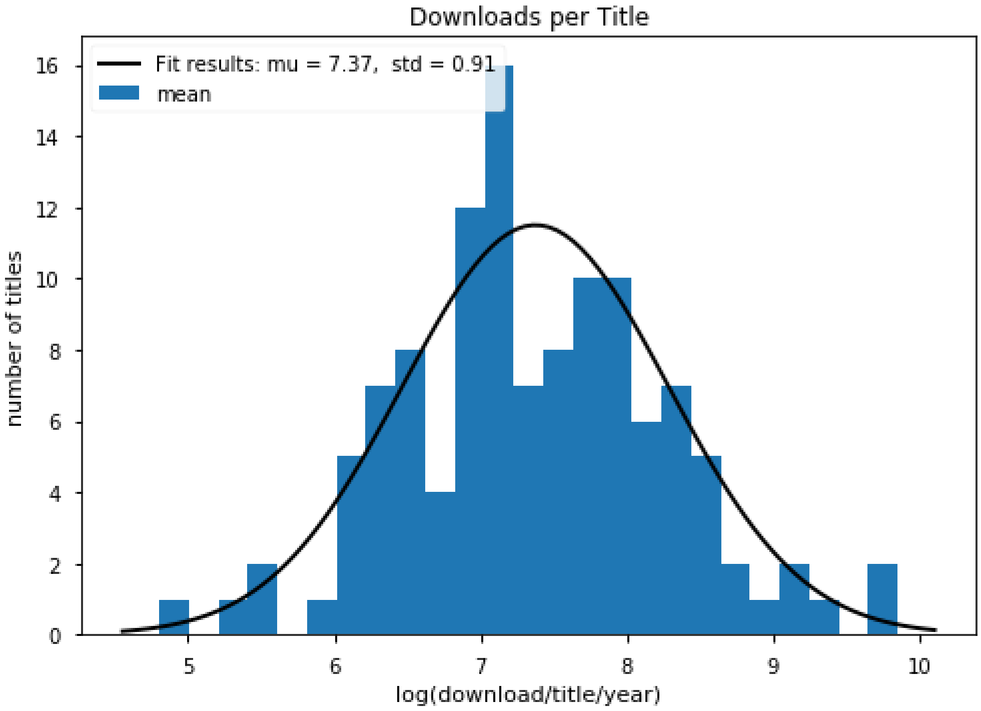

Figure 4. Distribution of book download rates per title

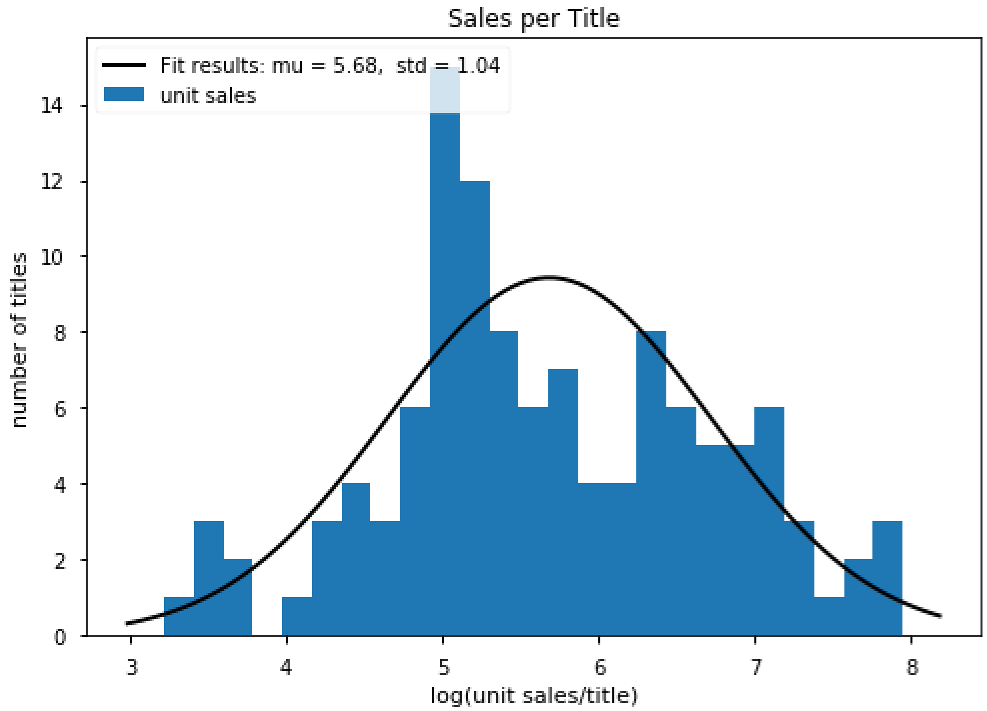

Figure 5. Distribution of sales across titles.

Repeating the regression analyses, in figure 6 it is seen that there’s little correlation between sales and pageviews. The slope obtained, 0.35 ± 0.10, corresponds to a rather weak power law, and there is a large amount of scatter, leading to a correlation coefficient of 0.306. Apparently, the qualities that attract sales are only weakly related to the qualities that attract page views. There's no statistically significant evidence in this dataset that downloads drive sales or that downloads suppress sales.

With a methodology grounded in log-normal statistics, it’s possible to answer a variety of questions. For example, the lists from the two publishers in the study, University of Michigan Press and Open Book Publishers were statistically quite similar. Unit sales for a Michigan title had a logarithmic mean of 7.27, and standard deviation of 1.056, while an OBP title register logarithmic mean sales of 7.45 with standard deviation of 0.77. We wondered whether characteristics of a book’s title affected the page views. The answer: apparently not. Measures such as title length and title entropy were not significantly correlated with ebook page views. We wondered if there was a relationship between a book’s sales and its price. The answer: not in this collection. We found no single book characteristic that strongly correlated to downloads or sales. The book publishing industry has an aphorism that explains this: "every book is different".

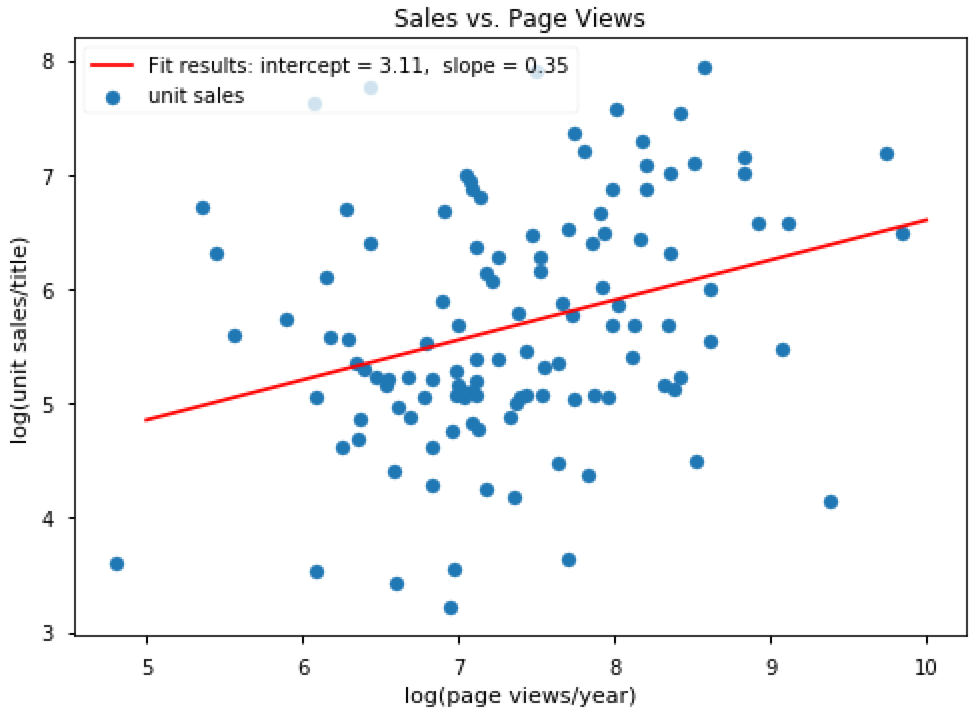

Figure 6. Log-Log plot of page views and unit sales. The correlation coefficient is 0.306

The website Unglue.it provides free hosting and links for thousands of free-licensed and public-domain ebooks. The larger number of titles and diverse content should make it easier to see patterns in the usage distribution. Download data for 10,109 ebooks available for at least 9 months via Unglue.it was analyzed. As seen in figure 7, the distribution of usage appears to result from at least two factors. The frequently downloaded distribution marked in green are books that have been featured on the Unglue.it homepage. These books are promoted in the Unglue.it twitter feed, on the Unglue.it Facebook page, and are given a higher weight in the website's sitemap. The distribution with the orange histogram corresponds to ebooks that are available in EPUB and MOBI formats. It's clear from the data that both of these two factors are important determinants of usage on the Unglue.it website.

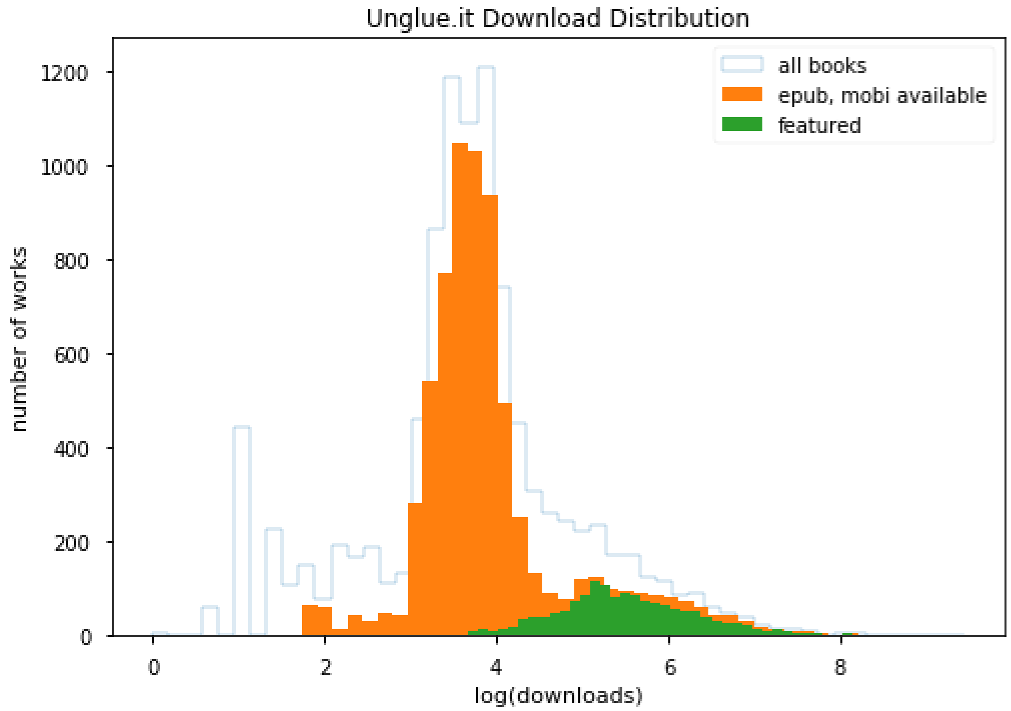

Figure 7. Distribution of download counts for ebooks on Unglue.it

Discussion

Distributions that look like a bell curve after application of a logarithm are called log-normal distributions. These are well known across science but are periodically rediscovered in new fields of inquiry. Log-normal distributions are typically found in systems exhibiting exponential growth (or in some fields “preferential attachment”).[17] The size distribution of raindrops are a good example: drops grow in size at a rate proportional to their size. It should not be surprising to see this distribution in ebook downloads or page views; the publishing industry understands the process as "word-of-mouth". Sales of books or downloads are driven by favorable comment; the number of these comments is proportional to the number of books that have been sold or downloaded.[18]

[17] “A Brief History of Generative Models for Power Law and Lognormal Distributions”, Michael Mitzenmacher, *Internet Mathematics* ***1*** (2), 226-251.

[18] In calculus, growth of a quantity in proportion to the quantity is the defining characteristic of the exponential; the logarithm is the inverse of the exponential.

In fact, it has been noted that the long tails in data sets of book sales and website hits can be well fit by log-normal distributions.[19] Analysis of circulation data from University of Huddersfield library was very well fit over several orders of magnitude by a log-normal distribution.[20] In the field of "alt-metrics", patterns of article citation are often characterized by log-normal distributions.[21]

Having observed that open access ebook downloads and library book circulation exhibit log-normal statistics, well-grounded statistical analyses are possible. To make collection-level statistical comparisons of download counts, unit sales, or circulation counts, we first compute the logarithm of the count.

When we compute an average or a variance, we're implicitly making an assumption about the quantity we're averaging, i.e. that is has an average value, and that if we use a larger sample, we'll get a better measure of that average. In statistics, this is called the Central Limit Theorem. If the distribution of the measured quantity is "Normal" or "Gaussian", we can use measured variance to compute the probable value of a subsequent measurement. Further, we can use familiar tools and criteria to decide if a result is "statistically significant". If we try to average measurements of a quantity that follows log-normal statistics, we'll need to make an exponentially large number of measurements before the averages converge.

An important message of this work is that statistical averages are frequently misleading. It's obvious in the extreme case: if averages are used in a sales analysis of novels published in 2011, the analysis will erroneously conclude that books with "grey" in the title[22] will tend to outsell other books with sorts of titles. If, as is likely, book sales generally follow log-normal statistics, a more useful analysis will result if the quantity analyzed is the logarithm of the sales.

More specifically, the present results are relevant to anyone trying to use download counts to measure or understand open access ebook usage. For example Emery and coworkers[23] reported ebook usage data purporting to show that OA increases average downloads per book compared to non-OA books. This effect is large and supported by the data, but the authors go on to say that "Engineering, mathematics and computer science OA books perform much better than the average number of downloads for OA books across all subject areas." but declined to

[19] "Power-law distributions in empirical data", Aaron Clauset, Cosma Rohilla Shalizi, M. E. J. Newman. arXiv:0706.1062 [physics.data-an] (Submitted on 7 Jun 2007)

[20] "The Distribution of Library Book Circulation Is Not a Power Law, or, Gauss and Man at Huddersfield" , Cosma Rohilla Shalizi, (16 Mar. 2011), http://bactra.org/weblog/744.html. Accessed 28 Nov. 2018.

[21] "Three practical field normalised alternative indicator formulae for research evaluation", Mike Thelwall, *Journal of Informetrics*, ***11***(1),128-151 (2017). https://doi.org/10.1016/j.joi.2016.12.002

[22] "Fifty Shades of Grey - Wikipedia." https://en.wikipedia.org/wiki/Fifty_Shades_of_Grey. Accessed 29 Mar. 2019.

[23] "The OA Effect: How Does Open Access Affect the Usage of Scholarly Books?", Christina Emery, Mithu Lucraft, Agata Morka, Ros Pyne, (Springer Nature, November 2017) https://media.springernature.com/full/springer-cms/rest/v1/content/15176744/data/v3

release any actual download data or any statistical characterizations of the data.[24] Since this conclusion is based on average download counts for a collection of similar size to our study, it's as likely as not that the reported subject area differences will disappear in an analysis using log-normal statistics.

More generally, the present results suggest a rethinking of how the OA ebook community measures “usage”. Even the term “usage” is wrong, in that it implies that a resource is used up or depleted. Open access ebooks provide value in many ways - communities read them, digest the information they contain, recommend them to colleagues, cite them, repurpose them. OA ebooks are NOT used up or depleted. Each of these activities will leave a signature in the download data. Because the downloads appear to follow log-normal statistics, it’s the logarithm of the download counts that usefully measure a normally distributed quantity. Let’s call this quantity the “*open-factor*” (explained below). For a commercial publisher, optimizing sales (and thus profits) is the goal. For an impact-driven publisher, whether university press or library-publisher, or scholarly non-profit, perhaps the figure of merit should be something more like the open-factor, not download counts, its exponential. The *open-factor* isn’t so tricky to measure (because it allows use of familiar statistical methods) and it aligns better with organizational impact than raw downloads.

What is this mysterious “open-factor”? We call it that because it’s a quantification of what causes the usage or impact of an open ebook to grow. Mathematically, it’s the probability that a usage will generate another usage. Practically, it’s a combination of quality and accessibility; the quality of a book makes people want to spread its use it, open accessibility enables them to do so. If our data is download counts, simply taking the logarithm of the counts will give us a quantity roughly proportional to the open-factor.[25] If instead of ebooks we were measuring internet memes, we would call this quantity the “repost-factor” of the meme. A large repost-factor blows up exponentially resulting in a viral meme. If social media feeds valued repost-factor rather than its exponential, our social media feeds would probably fill up with reasoned discourse instead of cats and angry babies.

Conclusion

This article describes a study of open-access ebook usage data and finds that it is log-normally distributed across the titles studies. It suggests that the *open-factor*, a quantity characterizing the log-normally distributed data, is a useful measure of an open access ebook’s impact.

[24] "Handle with Care: pitfalls in analysing book usage data | Dr. Rupert Gatti." 11 Dec. 2017, https://rupertgatti.wordpress.com/2017/12/11/handle-with-care-pitfalls-in-analysing-book-usage-data/. Accessed 29 Mar. 2019.

[25] It’s more complicated than just taking a logarithm, because real measurements sum over a distribution of log-normally distributed phenomena.

Organizations that optimize this *open-factor* rather than downloads will avoid being slaves to virality and the radicalization that results from using a measurement (download count) that values the extremes that result from exponentiation of a book’s qualities. They’ll value a breadth of quality; we don’t want open access to be dominated by *50 shades of scholarly grey*.

Acknowledgements

This work was sponsored in part by a grant from the Mellon Foundation ”Mapping the Free Ebook Supply Chain: Grant 11600118 (03/01/16 to 02/28/17).