RAIS RESEARCH ASSOCIATION for INTERDISCIPLINARY STUDIES
APRIL 2019

DOI: 10.5281/zenodo.2650983

Emerging the U.S. Firm Size Distribution Using 4.2 Billion Individual Tax Records

Joseph A.E. Shaheen

Department of Computational & Data Science, George Mason University, Fairfax, VA, USA, jshaheen@gmu.edu

ABSTRACT: The firm size distribution describes important economic and labor properties of any economy. Government entities must expend enormous resources in data collection, cleaning, and analysis in order to construct this and other important distributions describing the aggregate properties large economies. In the U.S., this process can be cumbersome and relies on querying multiple databases and utilizing significant computational resources. I show that construction of the U.S. firm size distribution is plausible using only individual income tax records (W2s) drawn directly from Internal Revenue Service tax records (micro data) and that the emergent distribution is statistically identical to what is reported by the United States Census Bureau. The methodology represents an incremental advance for population-scale studies in economic analysis—specifically firm and labor analysis. Finally, this paper acts as a re-validation of earlier work in fitting the firm size distribution.

KEYWORDS: firm size, labor, taxation, data policy, economic analysis, data science

Background

Firm sizes (as measured by number of employees) is an important firm feature closely related to the health, growth, and success of modern economies. Sizes of U.S. firms have been well-studied and are generally considered critical to understanding the state of labor dynamics and aggregate economic properties.

One of the first works aimed at understanding properties of firm sizes argued that firm size at time *t* follows a random growth (Gaussian) process that is independent of size at time *t-1* (Gibrat, 1931) and produces firm size distributions that are power law or log-normally distributed. Formally, this rule is known as the Law of Proportional Effect or Gibrat's Law (shown in equation 1 as defined by Sutton (1997)). In historically active lines of inquiry aimed at understanding firm sizes, reliance on sample-based studies of firm properties dominated the literature (Evans, 1987a,b; Hall, 1986). Access to more powerful computational resources coupled with a more robust process of engagement between U.S. federal entities and The Academy has provided for more opportunities in population-scale analysis focused on labor and firm data.

$$x_t - x_{t-1} = \epsilon x_{t-1}, \text{ where } \epsilon \text{ denotes growth rate between t and t-1.} \quad (1)$$

For much of the period prior to the year 2000 and with reliance on sampling techniques, the firm size distribution was often thought to be log-normal (Equation 2) exhibiting a random multiplicative process. And, while off-the-shelf economic analysis and simulations based on Gibrat's Law (proportional growth) produce log-normal distributions under most variations, empirical analysis of the firm size distribution produced results favoring scaling behavior (power law distributions) (Stanley et al. 1996). Through reliance on sampling methodology Sutton (2002) examined the variance of firm growth rates through a re-analysis of Stanley et al. (1996) and found firm sizes to be power law, while others curbed their claim of skewness only to the log-normal distribution (Cabral and Mata 2003). Ultimately, the empirical U.S. firm size distribution was shown to follow a stationary power law (specifically Zipf) (Axtell 2001) when more complete population-scale data was treated analytically, relegating Gibrat's model to hypothetical uses.

$$f(x) = \frac{e^{-((\ln x)^2/2\sigma^2)}}{x\sigma\sqrt{2\pi}} \qquad x > 0; \sigma > \quad (2)$$

Power law distributions are a unique class of statistical distributions with rare properties. They are considered to be scale invariant, lack a well-defined mean under most parameter values, and exhibit undefined higher moments (Clauset et al. 2009) requiring a unique class of fitting techniques and methods of analysis (Virkar and Clauset 2014). Power law distributions are often synonymous with complex adaptive systems in that they signal the existence of some underlying process that governs the system as a whole, though dozens of explanations have been provided as to what those underlying processes may actually be (Reed 2001). This interesting class of statistical distributions have been found to describe the number of casualties in wars (Richardson 1948), the size of U.S. cities (Zipf 1949), and the degree distribution of complex networks (Barabasi and Albert 1999) as well as many other social, economic and natural phenomena. The standard functional form of a power law distribution is shown in Equation 3.

$$f(x) \propto x^{-\alpha} \tag{3}$$

Since efforts herein focus on methodological issues of constructing large economic datasets (and specifically firm sizes) in more efficient ways and so as not to stray from that objective, I report a comparison of fit between a log-normal and power law for the year 1998 in 1 as a re-confirmation of the validity of a power law model put forth by Axtell (2001). The year 1998 was chosen since it is the closest year in our dataset that corresponds with his seminal analysis of the year 1997. Reporting this result here is advantageous as we will report a comparison of this model's parameters for Census Bureau and Internal Revenue Service datasets of the more appropriate power law model and omit a comparison of the log-normal fit.

Table 1. U.S. Firm Size Models

$ax^{-\alpha}$	$vs.$	$\frac{1}{(x\sigma\sqrt{2\pi})}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$
	power-law	log-normal
p	0.7	0.66
gof	0.341	0.425
α	1.11	
x_{min}	4	
μ		10.28
σ		4.91

Source: Calculated from IRS micro-data

Table 1: Firm size model comparison between log-normal and power law as implemented by Clauset et al. (2009) for the year 1998. *Note: $p > 0$ signifies rejection of the null hypothesis. $p \approx 0$ disallows rejection. Higher p values provide for increased confidence in the model.*

As shown in Table 1, while both the log-normal and power law can fit the data relatively well, with the log-normal model providing a closer fit, our confidence in the power law model is stronger. Axtell (2001) calculates the exponent parameter to be 1.059 while our re-analysis provides a value for the exponent of 1.11. Consequently, since we were able to, 1- establish that our data aggregation method can reproduce earlier results and, 2- that there exists sufficient evidence to use a power law model in order to conduct a comparison between reported public datasets and micro datasets, we will proceed onto describing our simple methodology.

Methodology

Internal Revenue Service Micro-data

In the previous section we provided, as background, a comparison of model fits based on the aggregation of roughly 4.2 Billion (over 16 years) Internal Revenue Service (IRS) records. Figure 1 displays the aggregate size distribution for all firms from 1998-2016 scaled logarithmically with exponential (log) binning assignments. This distribution and binning were the basis of said model.

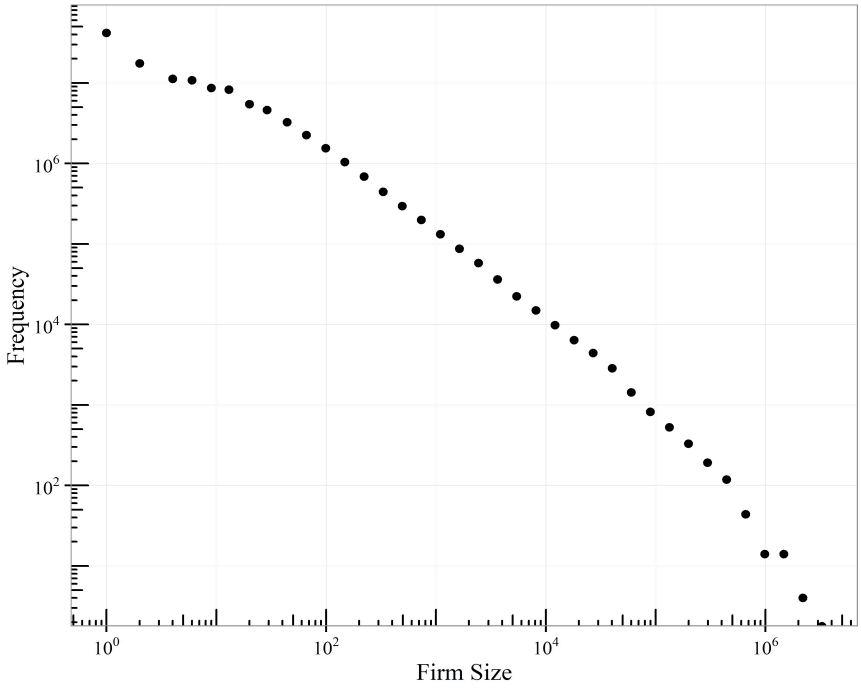

Figure 1. Aggregate Firm Size Distribution

Figure 1: Firm size distribution for all firms with at least one (1) employee in range 19982016. Frequency (y-axis) and size are plotted using a log-log scale with the data-set binned exponentially.

As implied by the figure–certainly for large portions of the distribution, if not the entire range–the firm size distribution follows a power law. However, the major difference between power law and log-normal fits are usually apparent in the tails, not in the main component of the distribution, hence requiring a formal statistical test (Table 1). Clauset et al. (2009)'s well-accepted method for fitting power law distributions was utilized.

Through a joint research partnership with the IRS, the distribution of firm sizes was constructed using a unique identifier for each firm as reported by both employers and employees on W-2 tax records. This construction was carried out using primary source electronic databases and records at IRS facilities under IRS subject matter experts' direct supervision and while maintaining anonymity of all tax records to ensure maximum privacy for U.S. taxpayers and to secure our high ethical standards.

W-2 tax records are documents that are (usually) transmitted electronically to the U.S. IRS by employers once every calendar year and are included on individual tax returns by employees once every calendar year (IRS form 1040). Thus, they form a highly reliable employer-employee matched population-scale collection of records. W-2 records include unique identifiers for employers–the Employer Identification Number (EIN) and for the employee–the Social Security Number (SSN), providing a natural defense against duplication and mismatching. They also include additional information such as yearly wages, income taxes paid, and deduction information.

Using these natural constraints to our advantage, we issued database queries that removed duplicate records, re-filed records, and amended records, and issued a count conditional on firms' EINs. This was implemented on the aggregate (all years combined) and temporally for the years 1998-2014. Exclusion of more recent years (2015-2019) was to ensure that late tax filings would not play a role in our analytical construction. By using this platform to construct firm sizes we ensure maximum accuracy, precision and can immediately eliminate anomalies from our data cleansing processes.

Construction of the firm size distribution from individual tax records is an example of bottom-up analysis–a notion commonly accepted by scholars of complexity science and the agent-based modeling community (Axtell 2000), but likely feels unintuitive to traditionalists that span economics, statistics and the new area of data science. Perhaps this is why a comparison of this method with top-down approaches that rely on construction through an amalgamation of firm tax records, to my knowledge, has never been attempted. Let us review one such method used by the U.S. Census Bureau.

Census Bureau Methodology

The United States Census Bureau (U.S. Census) is the government agency responsible for collecting national data of interest to the U.S. as well as for the analysis of such data. Much of what is gathered by U.S. Census is shared publicly, which includes economic information such as the firm size distribution. Occasionally and as needed, U.S. Census may or may not rely on datasets from other agencies, though little is disclosed as to where primary source data originates–only that the data is gathered into a "product"–which references a data warehouse– and that these datasets are then made available to the public in part or in whole.

Specifically, aggregate data on firm properties in the U.S. economy, such as age, wage, size and labor of firms (Shaheen, 2019) are reported within the Business Dynamics Statistics product (BDS) (United States Census Bureau, 2019). U.S. Census reports that this product was constructed by the Center for Economic Studies compiled from the Longitudinal Business Database using "annual snapshots" from the Census Bureau's Business Register using "probabilistic name and address matching" at the establishment level (many establishments can be part of a single firm). It is clear that this dataset has been used as the foundations by which many studies and analysis have been conducted–that it is important.

Once firm-level data is gathered by U.S. Census, firms are divided into sector classifications, including Mining, Construction, Retail Trade and other categories based on the Standard Industrial Classification (SIC). Additionally, U.S. Census carries out the exclusion of certain types of employees, such as the self-employed, domestic service workers, railroad workers, and the majority of government employees. A full accounting of industry classifications and excluded employees can be learned from this body's most previous citation.

In order to build such a cumbersome collection of linkages between tables and databases many resources must be expended–computational, human, and financial. There are undoubtedly good reasons to do so: One such reason would be the construction of the firm at the establishment level. An establishment is simply a physical location for a firm in part or in whole. Firms can be comprised of many establishments or a single establishment. In popular terms, establishments can be equated with branches of a bank, offices of a start-up, or even stations of a transit service–a less obvious case. Indeed, the single most reliable record of employment–the W-2–does not report establishment unique identifiers (to my knowledge no such identifier currently exists), but only location information. Consequently, U.S. Census must first identify firms and then use location information to identify establishments, since identifying firms first is key to identifying their establishments–an expensive top-down approach, from firm to worker rather than from worker to firm.

Comparing IRS to Census Data

As we are authorized to report binned IRS data only and possessed no access to primary source U.S. Census datasets our comparisons are conducted on binned data. Our original binned dataset conforms to

exponential bins (bins of exponentially increasing size). Exponential bins are often used when the hypothesized distribution of a given dataset is power law so as to ensure that enough data points exist in each bin and to take advantage of the smoothness of the resulting curve. A growing body of literature on proper power law analysis exists (Virkar and Clauset 2014).

Publicly reported U.S. Census datasets use much lower resolution bins than the exponential bins used to aggregate IRS micro data, especially in the lower firm size bound. As a comparison of our collected dataset and the dataset reported publicly by U.S. Census, the aggregate IRS-based firm size data is shown in Table 2 using exponential bins, and to follow, the corresponding Census dataset over the same time-period is shown in Table 3. We will avoid listing firm size data for all years due to length requirement.

Table 2. Firm Size Frequency Distribution

Lower Bound	Upper Bound	Frequency	LB (cont)	UB (cont)	Freq. (cont)
0	1	41865506	1480	2321	67112
1	2	17504300	2321	3640	39694
2	4	11232023	3640	5710	23351
4	6	13450753	5710	8955	14733
6	10	10286684	8955	14044	9300
10	16	7452135	14044	22026	5703
16	25	5706500	22026	34544	3782
25	40	3863039	34544	54176	1839
40	63	2572038	54176	84965	966
63	99	1706751	84965	133252	569
99	156	1079534	133252	208981	334
156	244	677493	208981	327747	201
244	383	419631	327747	514011	97
383	601	266169	514011	806129	16
601	943	170437	806129	1264263	16
943	1480	106274	1264263	1982759	10

Source: Internal Revenue Service

Table 2: Shown is the firm size distribution of U.S. firms aggregated between 1998-2015 using micro-data collected over 4.2 Billion IRS W-2 records. Firm sizes are binned logarithmically. The data is compiled from a primary source database of all IRS W2s. Upper bins are inclusive.

Since reported Census bins are arbitrary, while our reported tables use exponential binning designed to maximize power law model-building statistical power, in order to issue a one-to-one comparison we could either re-bin IRS data to conform to Census frequency breakpoints and choose some arbitrary bin reduction method (e.g. IRS dataset bin 40-63 crosses the 20-49 and 50-99 census bins and so we could proportionally assign the frequency of this bin to both Census bins) or we could bin the original dataset into census breakpoints ensuring a direct comparison while giving up statistical validity in power law model building. In the background section, we reported a comparison between power law and log-normal models using exponential bins observed from IRS micro-data, but for the remainder of this analysis and to ensure a direct comparison we re-binned original IRS data into Census bin breakpoints even when constructing a comparison of the power law model parameters for the years 1998-2015. Doing so will reduce our confidence in the power law model since the binning method is not ideal.

Table 3. Firm Size Frequency Distribution

Lower Bound	Upper Bound	Frequency
1	4	46730304
5	9	17692470
10	19	10419317
20	49	6445387
50	99	2020866
100	249	1094316
250	499	318863
500	999	151029
1000	2499	95040
2500	4999	36569
5000	9999	21027
10000	2500000	22576

Source: U.S. Bureau of the Census

Table 3: Shown is the firm size distribution of U.S. firms aggregated between 1998-2015 using U.S. Census publicly reported data. Firm sizes are binned in ranges chosen by U.S. Census. The data is compiled from public sets available on census.gov. All bins are inclusive. Upper bin (2.5M) assumed.

In order to compare our (IRS) construction to U.S. Census data we used three (3) methods: Firstly, a visual inspection of the frequency distributions (exploratory); secondly, a comparison of the parameter value of a power law fit (parametric); and thirdly, utilizing the non-parametric Kolmogorov-Smirnoff (K-S test) for all years in our dataset. I report the results, henceforth.

Results

A visual inspection of Figure 2 reveals the identical visual nature of the firm size distribution constructed from IRS data (blue cross) when compared to the firm size distribution reported by U.S Census (red circle) in the time-period 1998-2014 using census bin break-points.

Moreover, the distributions are indistinguishable under a two-sample Kolmogorov-Smirnoff (Smirnov) test providing a distance (D) value of 0.167, corresponding to a p-value of 0.996 for every year, suggesting that the two methods–at least where firm sizes are concerned–not only correspond statistically, but that U.S. Census aggregation of firm sizes post-construction of firm and establishment identifiers is likely partially based on W2 records. In statistical terms, we fail to reject the null hypothesis (the distributions are identical). There are limitations to this test: first, the K-S test is more suited for continuous data, and second, it is more suited for a larger number of observations (Koziol, 1980) than our specific implementation. However, this test still represents an agreeable piece of evidence towards determining the efficacy of the proposed construction method.

Provided in Table 4 is a direct comparison between IRS and U.S. Census data constructions based on a power law model. Estimates of the α parameter for both models appear to be relatively close. Variations of the goodness-of-fit and p-value (p-value $>>$ 0 supports the validity of the power law model (Clauset et al. 2009)) are expected since data construction of IRS data contained far more resolution. The resulting similarity between the estimated parameters supports the hypothesis that there is no statistically distinguishable difference in the data construction method.

Moreover, using a paired t-test for difference I compare the parameter values of each year in both model constructions. Table 5 lists both standardized t-scores and p-values. In all but one case (the year 2000) we can be highly confident that there is no real significant difference between the power law model's parameter–confirming our larger hypothesis.

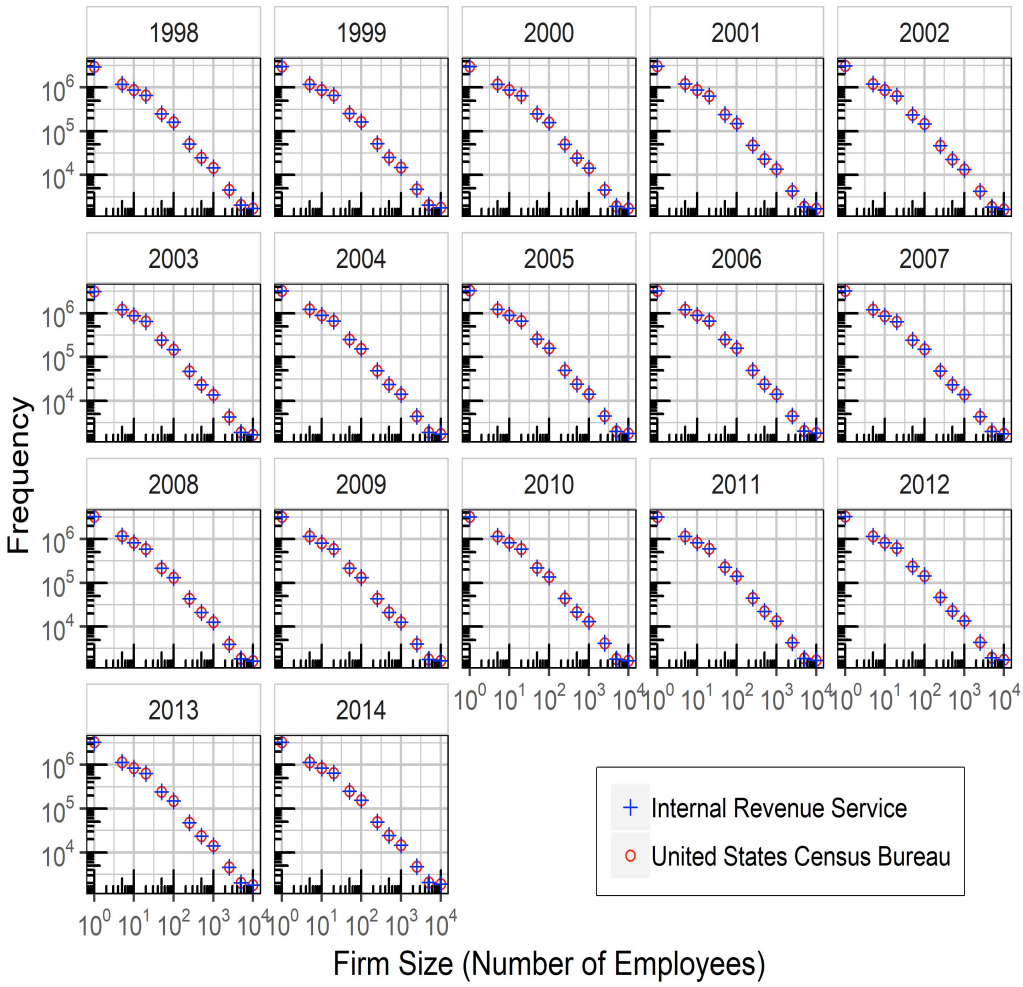

Figure 2. Calculated Temporal Firm Size Distribution Using Census Bins

Figure 2: Yearly firm size distributions for all firms with at least one (1) employee in range 1998-2014 including comparison between U.S. reported data and collected IRS data. Frequency (y-axis) and size are plotted using a log-log scale with the data-set binned exponentially.

Conclusions

Methodology-as-science is often frowned upon in The Academy, but in an age where computational resources often determine the availability of certain types of analysis or when those resources can even constrain the types of questions pursued by scholars, it is inevitable that one must embark on methodological questions. In this work, we embarked on precisely that noble pursuit.

The construction and consequent emergence of the U.S. firm size distribution–an important economic property–through the use of a bottom-up approach instead of a top-down analysis was a computationally efficient method resulting in an almost identical result to what is reported by the internal analytical groups of the U.S. Census Bureau. Statistically, there was little difference–as we have shown through exploratory, parametric and non-parametric validation methods. In the process of developing this construction method, new questions about the age, wage, size, labor, and generally–life and death of firms–are more feasible for independent scholars seeking access to micro-data. And, while we used this opportunity to re-confirm historical marquis results, perhaps the most important incremental advance here is the incorporation of bottom-up, agentized analysis in the data scientist's toolkit. In future works, I will report further analysis as a direct consequence of this data construction method.

Table 4. Power Law Model Comparisons

Dataset	IRS			Census		
Year	**parameter**	**gof**	**p-value**	**parameter**	**gof**	**p-value**
1998	1.2661	0.4046	0.57	1.2889	0.3214	0.7
1999	1.2685	0.4049	0.58	1.2913	0.3225	0.71
2000	1.2681	0.4053	0.57	1.2942	0.321	0.71
2001	1.2679	0.4082	0.56	1.2896	0.3208	0.71
2002	1.2666	0.4098	0.55	1.2907	0.321	0.71
2003	1.2673	0.4096	0.55	1.2849	0.3203	0.71
2004	1.2678	0.409	0.57	1.2866	0.3187	0.72
2005	1.269	0.4089	0.58	1.2848	0.3185	0.72
2006	1.2702	0.4091	0.58	1.2844	0.3186	0.72
2007	1.2697	0.4101	0.56	1.2874	0.3202	0.72
2008	1.2698	0.4134	0.52	1.2891	0.3205	0.72
2009	1.2711	0.4128	0.53	1.2823	0.3233	0.72
2010	1.2707	0.4119	0.54	1.2857	0.3194	0.72
2011	1.2703	0.4117	0.55	1.284	0.3206	0.71
2012	1.2704	0.4109	0.55	1.2841	0.3222	0.71
2013	1.2714	0.4106	0.56	1.2838	0.3231	0.71
2014	1.273	0.4103	0.56	1.289	0.3222	0.71

Table 4: Shown are power law fits for both IRS and Census datasets. Parameter values (α), goodness-of-fit, and the corresponding p-values are listed for each year in our dataset. Parameter values range from 1.26 to 1.29, in line with expectations. The power law model is a strong model for both datasets. Disparities in the goodness-of-fit and p-values are likely due to the higher resolution of IRS dataset. *Note: $p > 0$ signifies rejection of the null hypothesis. $p \approx 0$ disallows rejection. Higher p values provide for increased confidence in the model.*

Table 5. Parmeter Value Comparisons

Year	**t-score**	**p-value**
1998	-1.121	0.279
1999	-1.111	0.283
2000	-1.881	0.078
2001	-0.874	0.395
2002	-1.422	0.174
2003	0.048	0.962
2004	-0.233	0.819
2005	0.463	0.65
2006	0.813	0.428
2007	0.019	0.985
2008	-0.351	0.73
2009	1.504	0.152
2010	0.628	0.539
2011	0.942	0.36
2012	0.937	0.362
2013	1.246	0.231
2014	0.393	0.7

Table 5: Paired t-test for no difference results for all model parameter values. IRS models was compared to Census models. In almost every comparison there is no valid statistically significant different between IRS data models and Census data models. No single observation exceeds two (2) standard deviations in difference. *Note: $p > 0$ signifies failure to reject the null hypothesis (i.e. there is no difference between the values). $p < 0$ disallows rejection and forces us to conclude that there is a difference.*

References

Axtell, R. 2000. "Why agents?: on the varied motivations for agent computing in the social sciences." *Center on Social and Economics Dynamics - The Brookings Institution 47*(17): 1– 23.

Axtell, R. 2001. "Zipf Distribution of U.S. Firm Sizes." *Science 293*(5536): 1818–1820.

Baraba´si, A.-L. and R. Albert. 1999. "Emergence of Scaling in Random Networks." *Science 286*(5439): 509–512.

Cabral, L. M. and J. Mata. 2003. "On the Evolution of the Firm Size Distribution: Facts and Theory." *The American Economic Review 93*(4): 1075–1090.

Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review 51*(4): 661–703.

Evans, D. S. 1987a. Tests of Alternative Theories of Firm Growth. *Journal of Political Economy 95*(4): 657–674.

Evans, D. S. 1987b. The Relationship Between Firm Growth, Size, and Age: Estimates for 100 Manufacturing Industries. *The Journal of Industrial Economics 35*(4): 567–581.

Gibrat, R. 1931. Les in´egalit´es ´economiques: applications: aux in´egalitˆes des richesses, `a la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc: d'une loi nouvelle: la loi de l'effet proportionnel. Librairie du Recueil Sirey.

Hall, B. H. 1986. "The Relationship Between Firm Size and Firm Growth in the U.S. Manufacturing Center." Available at https://www.nber.org/papers/w1965.pdf.

Koziol, J. A. 1980. "Percentage points of the asymptotic distributions of one and two sample kuiper statistics for truncated or censored data." *Technometrics 22*(3): 437–442.

Reed, W. J. 2001. "The Pareto, Zipf and other power laws." *Economics Letters 74*(1): 15–19.

Richardson, L. F. 1948. "Variation of the Frequency of Fatal Quarrels with Magnitude." *Journal of the American Statistical Association 43*(244): 523–546.

Shaheen, J. A. E. 2019. Data Explorations in Firm Dynamics: Firm Birth, Life, & Death Through Age, Wage, Size & Labor. Ph. D. thesis, George Mason University.

Stanley, M. H. R., L. a. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. a. Salinger, and H. E. Stanley. 1996. "Scaling behaviour in the growth of companies." *Nature 379*(6568): 804–806.

Sutton, J. 1997. "Gibrat's Legacy." *Journal of Economic Literature 35*(1): 40–59.

Sutton, J. 2002. "The Variance of Firm Growth Rates: The 'Scaling' Puzzle." *Physica 312*: 577–590.

United States Census Bureau 2019. "Census Methodology." Available at https://www.census.gov/ces/dataproducts/bds/methodology.html. *Accessed 01/2019.*

Virkar, Y. and A. Clauset. 2014. "Power-law distributions in binned empirical data." *Annals of Applied Statistics 8*(1): 89–119.

Zipf, G. K. 1949. Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology. Cambridge, Mass.: Addison-Wesley Press.