



ELSEVIER

Contents lists available at [ScienceDirect](#)

Journal of Business Venturing Insights

journal homepage: www.elsevier.com/locate/jbvi

Toward a more nuanced understanding of long-tail distributions and their generative process in entrepreneurship



Jaehu Shim

Australian Centre for Entrepreneurship Research, Queensland University of Technology, Brisbane QLD 4000, Australia

ARTICLE INFO

Article history:

Received 1 June 2016

Received in revised form

8 August 2016

Accepted 10 August 2016

Keywords:

Long-tail distribution

Power-law distribution

Generative process

Fitting procedure

Simulation

Venturing process

ABSTRACT

Crawford et al.'s (2014, 2015) research on empirical distributions in entrepreneurship has shown that almost all input and outcome variables in entrepreneurship follow highly skewed long-tail distributions. They refer to these as power-law (PL) distributions based on a quantitative PL fitting procedure. However, the generative process of these distributions is still unclear. Building on their research, I cultivate a more nuanced understanding of the long-tail distributions and their plausible generative process in entrepreneurship. In this study, the fitting procedure is applied to new ventures' initial expectations and temporal outcome variables on employment and revenue, including comparisons of fitting results from alternative long-tail models. In conclusion, I find that ventures' less skewed early-stage outcome distributions change into more skewed PL distributions over time, while most expectation distributions do not fit a specific long-tail model. Using a simple simulation, I suggest that a multiplicative process may be a plausible generative mechanism for the transformation.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

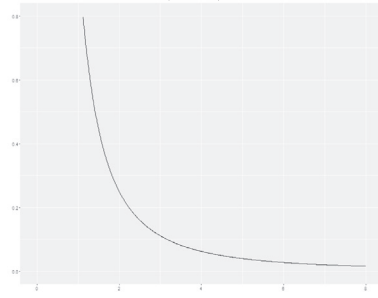
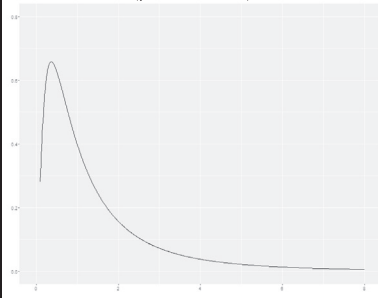
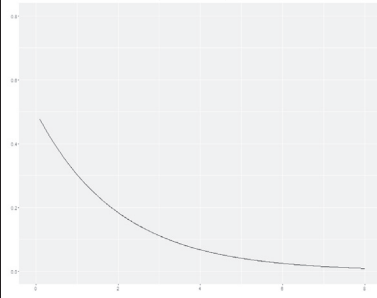
Crawford et al. (2014) have shown that firms' numbers of employees and amounts of annual revenue follow highly skewed long-tail distributions, which they refer to as power-law (PL) distributions. Furthermore, with additional co-authors Aguinis and Davidsson, they reported that almost all input and outcome variables in entrepreneurial processes follow PL distributions (Crawford et al., 2015). They arrived at this conclusion based on a quantitative PL fitting procedure, suggested by Clauset et al. (2009). Crawford et al.'s (2015) work may be regarded as seminal, as these findings challenge a common assumption of bell-shaped "normal" distributions in entrepreneurship studies, and call for new theories and methods in entrepreneurship research.

However, complementary fitting techniques, which are not reported in the Crawford et al.'s (2014, 2015) papers, may provide an even more nuanced understanding of how these long-tail distributions could emerge. For example, detailed fitting results for various distributions and their temporal changes can add insights into the generative process of the empirical distributions, which is still wrapped in a veil. Crawford and McKelvey (forthcoming) also provide more nuanced understanding of the phenomena by presenting more detailed methods and possible generative mechanisms.

Building on Crawford et al.'s (2014, 2015) work, I cultivate a more nuanced understanding of long-tail distributions in entrepreneurship. I applied the fitting procedure to ventures' expectations and outcome variables on employment and revenue, originating from the Panel Studies of Entrepreneurial Dynamics II (PSED II) (Reynolds and Curtin, 2008). In order to

E-mail address: jaehu.shim@qut.edu.au

Table 1
Probability density functions of power-law, log-normal, and exponential models.

PL: $x^{-\alpha}$ ($\alpha = 2$)	LN: $\frac{1}{x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$ ($\mu = 0, \sigma = 1$)	EXP: $e^{-\lambda x}$ ($\lambda = 0.5$)
		
$P_{(x>1)}: 1.00, P_{(x>4)}: 0.25, P_{(x>8)}: 0.13$	$P_{(x>1)}: 0.50, P_{(x>4)}: 0.08, P_{(x>8)}: 0.02$	$P_{(x>1)}: 0.60, P_{(x>4)}: 0.14, P_{(x>8)}: 0.02$

trace the temporal changes of the outcome distributions, I computed each venture's number of employees and amount of revenue by year (1st, 2nd, and 3rd year) from the venture emergence. I applied Clauset et al.'s (2009) fitting procedure, not only for the PL model, but also for alternative models, such as log-normal (LN) and exponential (EXP) distributions.

Findings from this study suggest that the PL is one of plausible models that explain outcome distributions in entrepreneurship, but most expectation variables do not fit the PL or other long-tail models. I also find that ventures' early-stage distributions of employment and revenue are more effectively described by the less skewed LN model, and then the distributions change into more skewed PL distributions over time. Through a simple simulation using 25,000 randomly generated values, I suggest that the multiplicative effect of each venture's numerous activities can be a plausible generative mechanism for the transformation. This study contributes to the entrepreneurship research by providing a more nuanced understanding of long-tail distributions and an insight into their generative process in entrepreneurship.

2. Long-tail distributions

Table 1 shows the probability density functions (PDFs) of the PL, LN, and EXP models. As Table 1 shows, each model's x value is defined as a positive number, and all three models similarly have highly skewed long tails. Thus, the LN and EXP models should be considered as alternative models for the PL, although the PL is a plausible model to describe an empirical distribution (Alstott et al., 2014; Clauset et al., 2009). Each model has its own features. For example, the mode (the most frequently occurred value) of the PL or the EXP is determined as the minimum value of the distribution, while the LN may have diverse modes according to its parameters.¹ Further, the models vary in terms of their tail lengths. The tail length in the PDF is related to each model's y decreasing rate for a one-unit increase in x . In the PL, the y decreasing rate is reduced by x increasing, while in the EXP, the rate is constant. Therefore, in general, the PL has a longer tail than the EXP, while the LN may have diverse tail lengths depending on its parameters.

Every model has its own scaling parameter(s), such as α for the PL model, μ or σ for the LN model, and λ for the EXP model. These parameters determine each distribution's detailed shape, and can be estimated by the maximum likelihood method. For most empirical distributions, only an upper part (i.e., a right part in the PDF) of the distribution follows a specific model. Thus, in this step, it is necessary to estimate a target model's x_{min} (minimal x) where the model's behavior starts, and n_{tail} denotes how many data points fit the model. The portion that fits a specific model can be calculated by (n_{tail}/n). If a fitted model's n_{tail} is relatively small, it means that the model describes only a small upper part of the empirical distribution.

3. Material and methods

3.1. Sample and variables

In this study, ten variables originating from the Panel Studies of Entrepreneurial Dynamics II (PSED II) were analyzed. The PSED II is a longitudinal dataset for 1214 emerging ventures in the United States. It conducted six yearly interviews (wave A to F) after a screening interview in 2005–2006 (Reynolds and Curtin, 2008). Half of the ten variables are employment-related (Employees-Expectation-Year1, Year5; Employees-Outcome-Year1, Year2, Year3), and the other half are revenue-

¹ The LN's mode is defined as $[\exp(\mu - \sigma^2)]$.

related (Revenue-Expectation-Year1, Year5; Revenue-Outcome-Year1, Year2, Year3). Among the variables, four variables concern nascent ventures' expectations (aspirations) in terms of their expected employees [or annual revenue], one year [or five years] after their venture emergence. Nascent ventures' expectations were measured at every wave, but their answers at first-wave (wave A) were analyzed in this study. Six variables concern new venture outcomes, measured by the number of employees and annual revenues.

In order to trace the temporal changes of the outcome variable's distributions, I computed new variables according to each venture's timing of emergence. Thus, a venture's first year's number of employees [or revenue] may come from a different year's interview (Wave B to F) according to the venture's emergence timing (e.g., if a venture emerges between waves A and B, the venture's number of employees [or revenue] at the wave B interview was used for their first-year outcome). The original questions of the PSED II can be found in the code book (Curtin, 2012).

3.2. Data analysis approach

Initially, each variable's descriptive statistics were calculated, and each distribution's "normality" was tested by calculating the KS statistic and its p -value for "normal" distribution.² Thereafter, Clauset et al.'s (2009) fitting procedure was applied, not only for the PL model, but also for the LN and EXP models.³

3.2.1. Estimating parameters and x_{min}

As the first step of the fitting procedure, each target model's scaling parameter(s) and x_{min} were estimated. At the same time, each model's n_{tail} was determined by the x_{min} . The estimation procedures of scaling parameter(s) and x_{min} are based on the maximum likelihood method, and described in Clauset et al. (2009).

3.2.2. Calculating KS and p -value

As the second step, the goodness-of-fits between an empirical data and the theoretical models were measured by the Kolmogorov-Smirnov (KS) statistic, which is the maximum distance between empirical data and a theoretical model in the cumulative distribution function. However, a theoretical model's plausibility for an empirical distribution cannot be judged by the KS statistic alone. For judgment, a bootstrapping (resampling from the theoretical distribution) is suggested, which estimates the probability (p -value) that an empirical KS statistic can be obtained from the theoretical model due to sampling error (Clauset et al., 2009). In this study, the p -value for each model was calculated by 2500 times of bootstrapping. If the p -value from the bootstrapping is greater than 0.1, the model can be regarded as a plausible. This bootstrapping method can be applied not only to the PL model but also to alternative models (e.g., the LN or the EXP). However, it is not suggested for picking which model is more plausible by comparing the models' p -values, because each estimated model has its own x_{min} and n_{tail} , and sometimes a high p -value model may describe only a small part of an empirical distribution.

3.2.3. Comparing likelihood between power-law and alternative models

Even if a theoretical model is plausible for an empirical distribution, it is recommended to consider alternative models, by comparing likelihood between two models. For the model comparison, it is requested to set the same x_{min} to both models to estimate parameters. By doing so, the two models' fitting results for the same range can be compared fairly. For all variables in this study, the PL model's x_{min} was set to both models to compare, but if an alternative model (e.g., the LN model) had a lower x_{min} and described more data points, an additional model comparison was performed using the lower x_{min} . If a log-likelihood ratio obtained from the likelihood ratio test is less than 0 and the p -value from the test is less than 0.1, it is regarded that the alternative model is significantly better than the default model (Vuong, 1989).

4. Results

Table 2 shows the descriptive statistics for the variables. The statistics confirm that each distribution is highly skewed (Skewness > 7). The p -values close to 0 for "normal" distributions rule out each variable's possibility of "normal" distribution.

Table 3 shows each variable's fitting results for the PL, LN, and EXP models. Overall, the KS values for the PL and LN models are less (0.05–0.09) than the corresponding values for the EXP model (0.11–0.43). These results imply that the PL or LN model is preferable to the EXP model. Considering the p -values for the PL model (> 0.10, bold in Table 3), the distributions of all outcome variables and one expectation variable (Employees-Expected-Year5) fit the PL model. However, the other expectation variables (Employees-Expected-Year1, Revenue-Expected-Year1 and Year5) do not fit the PL model. In addition, for seven out of ten variables, the LN model describes wider ranges of the distributions than the corresponding PL model. If the LN model has a significantly better fit for the wider range of a distribution, the LN model should be regarded as a better model to describe the distribution.

² Using 'nortest' package (1.0–4) for R.

³ Using 'powerLaw' package (0.60.3) for R and 'plpva.m' function (1.0.8) for Matlab (<http://tuvalu.santafe.edu/~aaronc/powerlaws/>).

Table 2
Descriptive statistics of employees- and revenue-related variables in entrepreneurship.

Variables	n	Min	Max	Med	Mean	s.d.	Skew	Normal D.KS (p)
Employees								
Expectation-Year1	1202	0	30,000	1	36	895	31.8	0.48 (0.00)
Expectation-Year5	1173	0	8500	3	34	335	19.2	0.46 (0.00)
Outcome-Year1	254	0	200	0	3	13	13.6	0.42 (0.00)
Outcome-Year2	167	0	100	0	3	9	7.7	0.37 (0.00)
Outcome-Year3	127	0	1500	0	17	134	10.8	0.45 (0.00)
Revenue (\$000)								
Expectation-Year1	1110	0.1	100,000	47	575	4746	14.5	0.45 (0.00)
Expectation-Year5	1106	0.2	1,000,000	100	4001	39,249	18.5	0.46 (0.00)
Outcome-Year1	242	0.4	12,000	50	296	1191	7.8	0.40 (0.00)
Outcome-Year2	157	1.0	18,000	60	311	1517	10.4	0.42 (0.00)
Outcome-Year3	122	0.3	15,000	48	370	1774	7.1	0.42 (0.00)

The maximum value of the Revenue-Expectation-Year5 is extraordinarily large (\$999,999,995); but the value was included for the analysis, because this value is not a special code, such as DK (Don't Know) or NA (No Answer).

Table 3
Parameter estimates and their goodness-of-fits (KS) for power-law, log-normal, and exponential models.

Variables (n)	n	Power-law			Log-normal				Exponential		
		x_{\min} (n_{tail})	α	KS (p)	x_{\min} (n_{tail})	μ	σ	KS	x_{\min} (n_{tail})	λ	KS
Employees											
Expectation-Year1	1202	3(357)	1.95	0.05(0.00)	2(496)	-5.45	3.10	0.04	1(617)	0.42	0.17
Expectation-Year5	1173	26(96)	1.75	0.07(0.12)	3(649)	-3.86	3.16	0.06	1(856)	0.28	0.27
Outcome-Year1	254	4(51)	2.41	0.05(0.80)	1(105)	1.12	1.04	0.05	1(105)	0.16	0.17
Outcome-Year2	167	2(60)	1.91	0.08(0.15)	2(60)	-0.80	1.86	0.05	8(15)	0.07	0.18
Outcome-Year3	127	2(47)	1.76	0.06(0.74)	7(13)	-47.2	8.93	0.09	20(6)	3E-3	0.43
Revenue (\$000)											
Expectation-Year1	1110	110(245)	1.72	0.07(0.00)	110(245)	-23.8	6.64	0.07	600(73)	7E-4	0.24
Expectation-Year5	1106	175(443)	1.64	0.08(0.00)	40(848)	0.37	3.62	0.06	320(278)	9E-4	0.17
Outcome-Year1	242	55(116)	1.78	0.06(0.27)	23(168)	2.62	2.26	0.06	800(14)	6E-4	0.17
Outcome-Year2	157	62(75)	1.86	0.06(0.50)	1(157)	3.79	1.85	0.06	300(19)	1E-3	0.11
Outcome-Year3	122	57(58)	1.85	0.06(0.56)	52(59)	-32.1	6.98	0.07	700(8)	6E-4	0.34

Reported p -values are calculated by 2500 times of bootstrapping; plausible p -values (> 0.10) are in bold.

Table 4 shows the results from model comparison through likelihood ratio tests. This table includes log-likelihood ratios and their p -values between the PL and alternative models. Overall, the PL model has a better fit than the corresponding EXP model ($p \leq 0.10$ for seven variables, bold in Table 4). The differences between the PL and LN models are not significant from each PL model's x_{\min} with one exception (Employees-Expected-Year1). However, the LN model has a significantly better fit than the corresponding PL model for three early-stage outcome distributions (Employees-Outcome-Year1, Revenue-Outcome-Year1 and Year2) from each LN model's lower x_{\min} .

5. Discussion and implications

The results from model comparison by the likelihood ratio tests indicate that the LN model is one of plausible models that describe the long-tail distributions in entrepreneurship. Furthermore, for three early-stage outcome variables, the LN model describes wider ranges of the distributions significantly better than the PL model. Thus, it is reasonable to conclude that ventures' early-stage outcome variables are more effectively described by the less-skewed LN model, and two or three years after their venture emergence, the outcome variables show more-skewed PL distributions. This result implies temporal changes of the outcome distributions in entrepreneurship from the LN model to the PL model. Table 5 (upper-side) illustrates the temporal change of an outcome variable.

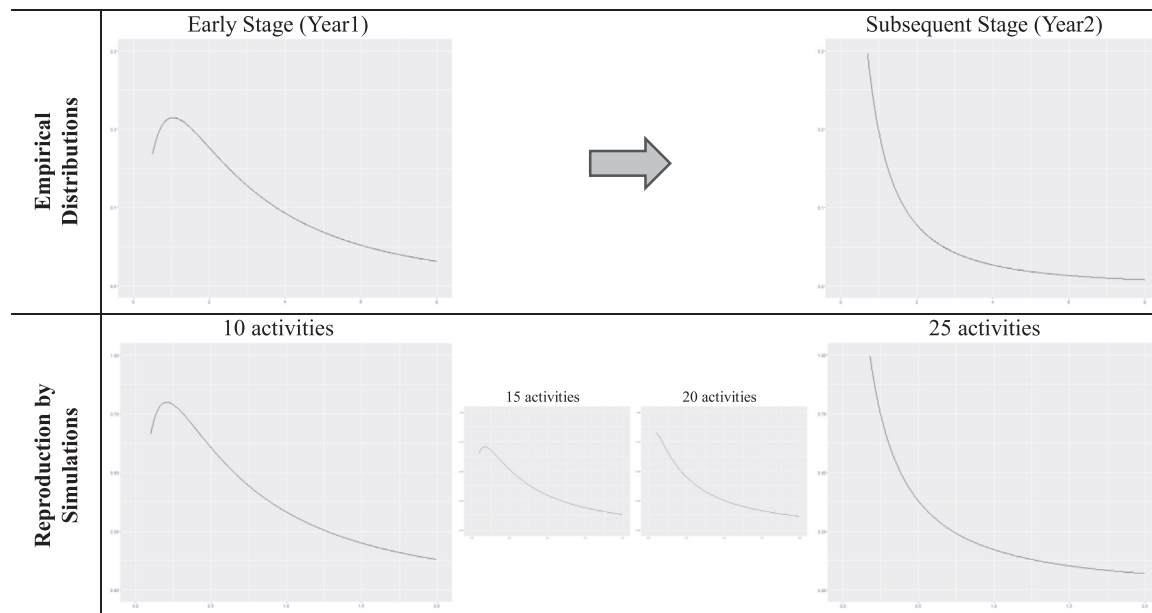
However, most expectation variables do not fit the PL or other long-tail distributions. Three expectation variables' near-zero p -values for the PL model rule out their plausibility of PL distributions. Moreover, neither the LN model nor the EXP model shows a significantly better fit for the expectation variables (Table 4). Among the expectation variables, only Employees-Expected-Year5 fits the PL model ($p=0.12$, in Table 3), but the PL model describes only upper 8% of the empirical distribution (96/1173). This result implies that the PL model does not fully describe the empirical distribution (e.g., Clauset, 2009).

Table 4
Comparing likelihood between power-law and alternative models.

Variables	X _{min}	Power-law vs. Log-normal			Power-law vs. Exponential		
		LR	p	Meaning	LR	p	Meaning
Employees							
Expectation-Year1	3	3.25	(0.00)	PL > LN	1.27	(0.20)	–
	2	–0.39	(0.70)	–			
Expectation-Year5	26	0.16	(0.87)	–	4.11	(0.00)	PL > EXP
	3	–1.46	(0.14)	–			
Outcome-Year1	4	0.67	(0.50)	–	1.55	(0.11)	–
	1	–3.28	(0.00)	LN > PL			
Outcome-Year2	2	–0.97	(0.33)	–	1.50	(0.13)	–
	2	0.30	(0.38)	–			
Outcome-Year3	2	0.30	(0.38)	–	3.07	(0.00)	PL > EXP
	2	0.30	(0.38)	–			
Revenue (\$000)							
Expectation-Year1	110	–0.09	(0.93)	–	4.45	(0.00)	PL > EXP
Expectation-Year5	175	0.11	(0.91)	–	2.95	(0.00)	PL > EXP
	40	–1.59	(0.11)	–			
Outcome-Year1	55	–0.53	(0.59)	–	3.57	(0.00)	PL > EXP
	23	–1.91	(0.06)	LN > PL			
Outcome-Year2	62	0.09	(0.93)	–	2.56	(0.01)	PL > EXP
	1	–5.40	(0.00)	LN > PL			
Outcome-Year3	57	0.07	(0.94)	–	3.49	(0.00)	PL > EXP
	52	0.06	(0.95)	–			

Reported p-values are calculated by two-sided likelihood ratio tests; significant p-values (≤ 0.10) are in bold. 'M1 > M2' denotes M1 is significantly better model than M2; '-' denotes no significant difference between two models.

Table 5
Temporal change of empirical distributions and its reproduction by simulation.



The elusiveness of the expectation distributions may be related to the mixed nature of the measurement for expectation variables. Even if ventures' real expectations (aspirations) follow a long-tail distribution, the respondents should consider the feasibility of their expectations when a specific timeframe is given (e.g., in one year). The feasibility consideration may distort the expectation distributions, and the distortion may be bigger when a shorter timeframe is given. This reasoning explains why any one-year expectation variable analyzed in this study does not fit the long-tail distributions. In addition, a venture's expectation may change during the venturing process, and the distribution of their expectation may vary according to their venturing stages. Therefore, it may be meaningful to discern the expectation distributions by the subsets of the data, like the outcome variables analyzed in this study.

Model comparisons by the likelihood ratio tests are beneficial in order to find a preferable model. In general, the likelihood ratio test is more time-effective than the bootstrapping procedure. The likelihood ratios can be obtained through a

Table 6

Fitting results of log-normal and power-law models for simulated distributions of temporal variables.

Variables	n	Log-normal				Power-law			LN vs. PL	
		$x_{\min} (n_{\text{tail}})$	μ	σ	KS (p_1)	$x_{\min} (n_{\text{tail}})$	α	KS	LR (p_2)	Meaning
Multiplication of:										
10 activities	1000	0.63(693)	-0.16	1.19	0.02(0.65)	3.42(149)	2.81	0.04	5.69(0.00)	LN > PL
15 activities	1000	0.02(999)	0.03	1.34	0.02(0.13)	1.94(323)	2.17	0.04	36.6(0.00)	LN > PL
20 activities	1000	0.00(999)	0.06	1.58	0.02(0.01)	16.9(48)	2.77	0.05	38.6(0.00)	LN > PL
25 activities	1000	0.85(547)	-0.94	2.10	0.02(0.68)	6.90(139)	2.09	0.05	3.43(0.00)	LN > PL

Reported p_1 -values are calculated by 2500 times of bootstrapping; plausible p -values (> 0.10) are in bold.

Reported p_2 -values are calculated by two-sided likelihood ratio tests; significant p -values (≤ 0.10) are in bold.

'LN > PL' denotes LN is significantly better model than PL.

relatively simple computation, while the bootstrapping procedure relies on complex computations involving more than 1000 times of resampling and estimating, so it may take several hours or days. Furthermore, the bootstrapping procedure may show that all of the assessed models are, or none of them is, plausible. In this situation, the likelihood ratio test can tell a better model (Alstott et al., 2014).

Some scholars suggest that the multiplicative process may be a plausible generative mechanism for the long-tail distributions (e.g., Mitzenmacher, 2004; Nirei and Souma, 2007). Thus, I performed a simple simulation using R software (3.2.2) to discern whether long-tail distributions can emerge by the multiplicative process in entrepreneurship. In this simulation, I assumed 1000 virtual ventures, and initially generated 10,000 random values (R) that follows a "normal" distribution ($\mu=0$, $\sigma=0.5$). After this, a new set of 10,000 values was generated by calculating 2^r , where $r \in R$. The geometric mean of the new set is 1.0, and its 95% values range from 1/2 to 2.0. Each generated value connotes the outcome of one venturing activity, which can halve or double the previous outcome. From the new set of values, I assigned 10 values per venture. The product of each venture's 10 values was regarded as the integrated outcome of each venture's initial 10 activities. Following the same procedure, I generated 5000 additional random values (five per venture) at every stage, and updated each venture's outcome value by the product of the current value and the additional values. The updated values are regarded as each venture's temporal outcomes after their 15, 20, and 25 activities.

Table 5 (lower-side) illustrates the distributions of the simulated outcomes by the number of activities, which can be regarded as venturing stages. This table shows that the changing patterns of empirical distributions and their reproduction by simulation are quite similar to each other. Table 6 shows that each LN model's σ parameters are growing by the number of activities (venturing stages), and this pattern is similar in the empirical fitting results (Table 3). This means that the later outcome distributions have longer tails in general, as the σ parameter is the standard deviation of the variable's natural logarithm.

6. Conclusion

I showed that ventures' less-skewed outcome distributions change into more skewed over time. The simulation results suggest that the random multiplicative process may be a plausible generative mechanism for the transformation. However, in the simulation results, the LN model has better fit than the PL model at every stages (10–25 activities), unlike the empirical findings. This result implies that a more complicated mechanism, in addition to the random multiplicative process, may exist behind the transformation. More sophisticated agent-based modeling and simulations with plausible assumptions will be useful to discern the generative process.

Acknowledgments

I appreciate Per Davidsson for his guidance and encouragement. I thank Aaron Clauset and other contributors to the implementations of the fitting procedure. I also thank Dimo Dimov and the reviewer(s) for their constructive feedback.

References

- Alstott, J., Bullmore, E., Plenz, D., 2014. Powerlaw: a python package for analysis of heavy-tailed distributions. *PLoS One* 9 (1) <http://dx.doi.org/10.1371/journal.pone.0085777>.
- Clauset, A., 2009. Comment on Yu et al., "High quality binary protein interaction map of the yeast interactome network". *Science* 322 (2008), 104. arXiv preprint arXiv:0901.0530.
- Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power-law distributions in empirical data. *SIAM Rev.* 51 (4), 661–703.
- Crawford, G.C., McKelvey, B., 2016. Using maximum likelihood estimation methods and complexity science concepts to research power law-distributed phenomena. In: Mitleton-Kelly, E., Paraskevas, A., Day, C. (Eds.), *The Edward Elgar Handbook of Research Methods in Complexity Science and Their*

Application (forthcoming).

- Crawford, C.G., Aguinis, H., Lichtenstein, B., Davidsson, P., McKelvey, B., 2015. Power law distributions in entrepreneurship: implications for theory and research. *J. Bus. Ventur.* 30 (5), 696–713, <http://dx.doi.org/10.1016/j.jbusvent.2015.01.001>.
- Crawford, C.G., McKelvey, B., Lichtenstein, B.B., 2014. The empirical reality of entrepreneurship: how power law distributed outcomes call for new theory and method. *J. Bus. Ventur. Insights* 1, 3–7, <http://dx.doi.org/10.1016/j.jbvi.2014.09.001>.
- Curtin, R., 2012. *Panel Study of Entrepreneurial Dynamics II: Codebook*. University of Michigan, Ann Arbor, MI.
- Mitzenmacher, M., 2004. A brief history of generative models for power law and lognormal distribution. *Internet Math.* 1 (2), 226–251, <http://dx.doi.org/10.1080/15427951.2004.10129088>.
- Nirei, M., Souma, W., 2007. A two factor model of income distribution dynamics. *Rev. Income Wealth* 53 (3), 440–459.
- Reynolds, P.D., Curtin, R.T., 2008. Business creation in the United States: panel study of entrepreneurial dynamics II initial assessment. *Found. Trends Entrep.* 4 (3), 155–307.
- Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. *Économ.: J. Econom. Soc.* 57 (2), 307–333.