Journal of Business Venturing Insights 6 (2016) 21–27

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.

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.

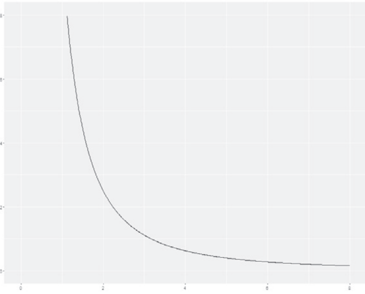
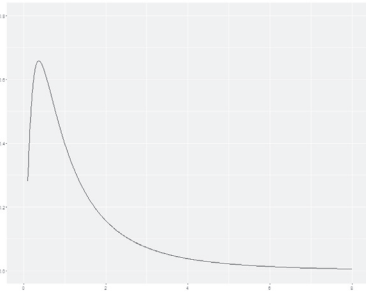
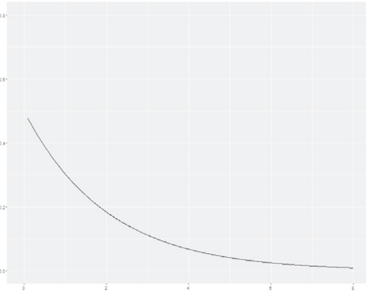
However, complementary fitting techniques, which are not reported in the Crawford et al.'s (2014, 2015) papers, may provide an even more nuanced understanding of how these long-tail distributions could emerge. For example, detailed fitting results for various distributions and their temporal changes 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

http://dx.doi.org/10.1016/j.jbvi.2016.08.001
2352-6734/

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

PL: $x^{-\alpha}$	**LN:** $\frac{1}{x} exp\left[-\frac{(ln\,x-\mu)^2}{2\sigma^2}\right]$	**EXP:** $e^{-\lambda x}$
($\alpha = 2$)	($\mu = 0, \sigma = 1$)	($\lambda = 0.5$)
		
$P_{(x>1)}$: 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(μ–σ^2)].