Journal of Business Venturing Insights 9 (2018) 53–59

Contents lists available at ScienceDirect

Journal of Business Venturing Insights

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

Enhancing simulation-based theory development in entrepreneurship through statistical validation

Roman Breig, Maximilian Coblenz*, Michael Pelz

Department of Economics and Management, Karlsruhe Institute of Technology (KIT), Kaiserstrasse 12, 76131 Karlsruhe, Germany

ARTICLE INFO

Keywords:
Entrepreneurship
Power law
Generative mechanism
Simulation
Statistical validation
Theory development

ABSTRACT

Recent research has called for new theory development because variables in entrepreneurship have been shown to follow power laws. Usually, simulation is used to validate these new theories. However, validation has been insufficient because it fails to provide a quantitative comparison of distribution parameters. This neglect can cause misleading conclusions. To address this insufficiency, we contribute a four-step method: the *possible simulation parameter range* (PSPR). The fundamental advantage of the method is to compare distribution parameters of both empirical data and simulation results. We demonstrate the method's usefulness with an illustrative example.

1. Introduction

Power law (PL) distributions have received considerable attention in recent entrepreneurship studies because they challenge the fundamental assumption of normally distributed (i.e. Gaussian) variables in this research field (Crawford et al., 2015, 2014). PL distributions are a specific form of heavy-tailed distributions, but other forms exist, such as lognormal. Recent findings make clear that researchers have to consider heavy-tailed distributions, as they characterize many variables that are central to entrepreneurship theories, such as revenue growth, company size, or venture debt. Recently, researchers have called for development of new theories that "explain and predict the mechanisms that generate these distributions and the outliers therein" (Crawford et al., 2015, p. 696).

New theories should contain generative mechanisms, which cause PL and other heavy-tailed distributions. Prior research provides a comprehensive overview of generative mechanisms (Andriani and McKelvey, 2009; Mitzenmacher, 2004; Newman, 2005). Theory development can benefit substantially from simulation (Davis et al., 2007). Recent work has developed theory and implemented simulations in the context of entrepreneurship, including a simple model for the distribution of several entrepreneurial variables over time, which was implemented and validated in a simulation based on a multiplicative process as a generative mechanism (Shim, 2016). Further work developed a bibliometric method to generate agent-based simulation models (Shim et al., 2017).

Empirical data can provide the basis for using numerical simulation, which generates its own set of data. Researchers have described the conventional approach for conducting agent-based simulation (e.g. Shim et al., 2017; Shim and Bliemel, 2017). The comparison of empirical data and simulation results can be used to assess whether heavy-tailed distributions in entrepreneurial research, such as PL distributions, can be explained by a newly developed theory that compares the type of distribution from both sets of data (Shim, 2016; Shim et al., 2017). Having the same type of distribution is used to show that a developed theory can sufficiently explain the heavy-tailed distributed data.

However, as this paper demonstrates, this approach lacks a final step, namely a quantitative comparison of the distribution parameters of both the empirical data and simulation results. Even when both sets have the same type of distribution, qualitatively

* Corresponding author.
E-mail address: maximilian.coblenz@kit.edu (M. Coblenz).

https://doi.org/10.1016/j.jbvi.2018.02.003
Received 23 November 2017; Received in revised form 31 January 2018; Accepted 7 February 2018
2352-6734/

Table 1
Descriptive statistics and data analysis results for PSED II Wave F data.[a].

Data description	
N	99
Mean	653.650
Standard deviation	3.688.200
Skewness	9
Excess kurtosis	85
p-value test for normality	< 0.0001
Distribution fit	
PL fit	$\hat{\alpha} = 1.75$; $\hat{xmin} = 104{,}500$ $n_{tail}/n = 38.4\%$
KS statistic PL	0.08
p-value PL fit	0.45
Lognormal fit	$\hat{\mu} = 10.85$; $\hat{\sigma} = 2.03$
KS statistic lognormal	0.05
p-value lognormal fit	0.50
Test statistic Vuong's test	+0.09 (n.s.)

[a] Data are right-skewed and exhibit heavy tails. A test for normality is highly significant (i.e. normality is rejected). Both lognormal distribution and PL distribution fit the data well; *p*-values of lognormal and PL from the bootstrapping procedure show a good fit of both models. The Vuong model comparison test is not significant. Thus, both the PL distribution and the lognormal distribution are reasonable distribution choices and are included in the candidate collection.

the specific distribution parameters might still be significantly different. Only a more detailed comparison allows drawing inferences about whether a developed theory sufficiently explains heavy-tailed distributions in the empirical data.

A quantitative comparison of empirical and simulation data is a form of validation, and the literature introduces several categories of validation. One distinction is between micro-level (i.e. behavior of simulated agents) and macro-level validation (i.e. simulation output) (Takadama et al., 2008). A second distinction is that between conceptual (i.e. assumptions of simulation model) and operational validation (i.e. output) (Heath et al., 2009). Statistical and non-statistical validation have also been distinguished (Heath et al., 2009).

The contribution of this paper is twofold. First, we introduce a four-step method that enables researchers to easily compare distribution parameters from simulation and empirical data, allowing a more sophisticated and comprehensive validation of developed theory. The method is embedded in the overarching framework of simulation-based theory development and refers to that framework's last step, "*Validate with empirical data*" (Davis et al., 2007, p. 482, Table 1). As the method is macro-level, operational, and statistical, we respond to the call for statistical validation techniques (Heath et al., 2009). Second, we apply the method in an illustrative example of the entrepreneurial variable venture debt (VD), thereby showing why taking the distribution parameters into account is important. Both contributions reduce the gap in the current literature significantly.

2. The four-step method

We present the method in detail and offer an illustrative application in the next section. The method consists of the following four steps:

1. *Analyze the data* according to the method outlined in Clauset et al. (2009) and form a collection of candidate distributions that fit the data.
2. *Build a theory-guided simulation* that implements a generative mechanism corresponding to one of the candidate distributions.
3. *Run and assess the simulation*. Assess the results by explicitly comparing the *specific distribution parameters* of both simulation and empirical data.
4. *Validate the results and refine if necessary*. Validation is achieved as soon as the specific distribution parameters of Step 3 match. If parameters do not match, refine Step 2 and repeat Step 3 until the simulation (and thus the model) is validated.

2.1. Step 1 – analyze the data

Empirical data are analyzed according to the method in Clauset et al. (2009), who describe how to fit a PL to data. To assess the distribution fit, they suggest calculating the Kolmogorov-Smirnov (KS) test statistic. A low value (i.e. a value close to 0) indicates a good distribution fit. However, the distribution fit must be further supported by a semi-parametric bootstrapping procedure based on the KS test statistic. The method yields a *p*-value where values above 0.1 are considered to be a strong indicator for a PL (Clauset et al., 2009).

Following Shim (2016), the analysis is extended to other possible heavy-tailed distributions, such as lognormal and exponential. Each distribution is fit using the method described above, and every distribution that fits the data well is added to a collection of