Journal of Business Venturing Insights 1-2 (2014) 3–7

Contents lists available at ScienceDirect

Journal of Business Venturing Insights

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

The empirical reality of entrepreneurship: How power law distributed outcomes call for new theory and method

G. Christopher Crawford [a,*], Bill McKelvey [b], Benyamin B. Lichtenstein [c]

[a] Department of Management, College of Business, Ohio University, Athens, OH 45701, USA
[b] Kedge Business School, Domaine de Luminy, BP921, 13288 Marseille Cedex 9, France
[c] College of Management, University of Massachusetts, 100 Morrissey Blvd., Boston, MA 02125, USA

ARTICLE INFO

Available online 16 October 2014

Keywords:
Entrepreneurship
Entrepreneurial startups
Skew distributions
Power laws
Extreme outcomes

ABSTRACT

Entrepreneurship researchers typically assume that normal (i.e., Gaussian) distributions characterize the outcomes of interest. Our research challenges this assumption by examining a sample of 6,530 firms to uncover the shape of the distribution for two key variables in entrepreneurial firms: number of employees and revenues. Results show highly skewed power law distributions. Future researchers need to recognize the relevance of power laws and extreme outcomes and then search for the underlying generative causes.

1. Introduction

Why is it that research findings about entrepreneurship continue to be so disparate? Recent reviews, for example, reveal extreme variances between inputs and outcomes, conflicting empirical findings, inconsistent measures of growth, and weak research designs (Achtenhagen et al., 2010; Leitch et al., 2010). Although some of these challenges are due to the complexity of the phenomenon, the problem may also be due to the inaccuracy of certain unexplored assumptions that permeate the field. In particular, some scholars have argued that Gaussian distributions—the normal curve, which underlies all statistical methods used by the discipline—does not reflect the actual distribution of relevant data (Andriani and McKelvey, 2007, 2009). This one assumption could explain almost all of the disparate findings in the field.

In fact, some scholars propose that entrepreneurship data follow Pareto distributions, also known as power laws (Boisot and McKelvey, 2010; Crawford, 2012). Power law distributions (PLDs) are highly skewed, with "long tails" that identify extreme events, i.e. data which are outside the range of the normal curve. Whereas traditional statistics assumes these are *mistakes* that need to be removed from the data, in Pareto distributions they are *expected*, albeit rare; further, these extreme outcomes have a nonlinear influence in the system (Taleb, 2007). PLDs have been found to explain many systems—physical, natural, biological, social, and financial—as cited by Andriani and McKelvey (2009) and McKelvey and Salmador Sanchez (2011).

When graphed on linear scales, the PLD looks like Fig. 1a; graphed on log–log scales, it forms a straight line, as stylized in Fig. 1b. Outcomes in the long tail of the distribution, like the largest circle at the bottom right of Fig. 1b, are 'extreme'

* Corresponding author.
E-mail addresses: Crawforg@Ohio.edu (G. Christopher Crawford), mckelveybill1@gmail.com (B. McKelvey), benyamin.bml@gmail.com (B.B. Lichtenstein).

http://dx.doi.org/10.1016/j.jbvi.2014.09.001
2352-6734/

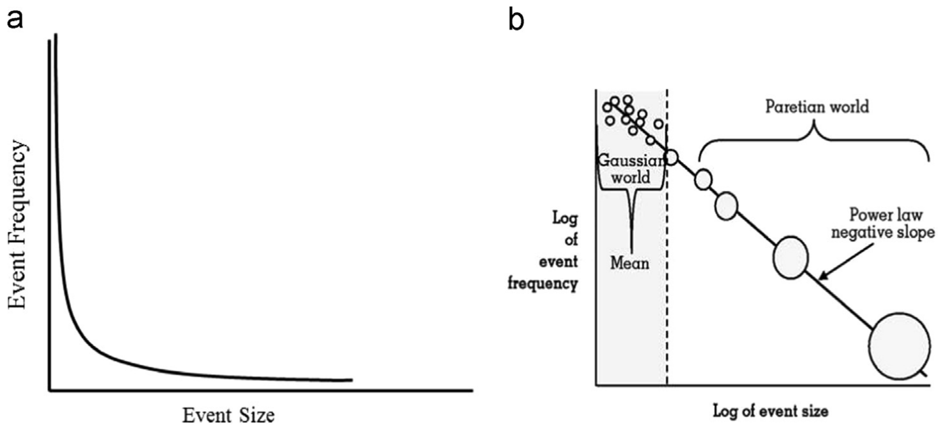

Fig. 1. (a) Rank/frequency distribution on linear scales and (b) distribution plotted on log–log scales.

compared to the others. In entrepreneurship, nonlinear outcomes might include a new firm that generates 1000 new jobs, \$100 M in revenue, or a 40,000% increase in growth. As we will identify, these extreme but uncommon events—think Amazon, Google, Facebook—emerge through the same underlying dynamics that produce $N=22{,}000{,}000$ small businesses in America. Unfortunately, linear methods, which underlie all statistical packages, are unable to include these outliers in an analysis, thus making them invisible. Outliers, then, have the potential to skew empirical analyses—and, thus, theory building and testing—across all entrepreneurship research.

We examine two of the most generalizable outcome measures in the domain—number of employees and annual revenue—at three levels of venture emergence: nascent ventures, young firms, and hyper-growth companies. Our goal is to determine whether these outcomes of entrepreneurship are actually PLD rather than conforming to a normal curve. After reporting the results, we draw from extant literature to identify some of the primary causal mechanisms that generate PLDs, and we link seminal entrepreneurship concepts to these mechanisms.

2. Method

We collect employee and revenue outcome data from three relevant datasets—the PSED, the Kauffman Firm Survey, and the INC 5000; the combined sample includes more than 11,000 firms. All three samples include the entire spectrum of business types and industries, and all data are collected from similar time periods to mitigate potential cohort affects.

2.1. Samples and procedures

Data collection in the Panel Study of Entrepreneurial Dynamics II (*PSED*) started in 2005 as a representative sample of 31,845 adults in the United States to assess the level of embryonic entrepreneurial activity (Reynolds, 2007). Information was collected from 1214 subjects who affirmatively answered the question, "Are you, alone or with others, currently trying to start a new business?" The Kauffman Firm Survey (*KFS*) started with a random sample of 32,469 businesses from a Dun & Bradstreet list which identified almost 250,000 firms that started operations in 2004. A start-up includes any independent business that was established by a single person or a team, or purchased as an existing business or new franchise (DesRoches et al., 2009). The Inc. 500® (*INC*) self-selected sample annually collects and publishes revenue, employee, and three-year growth-rate data on the fastest-growing, privately held, for-profit companies in the United States (Inc. Magazine, 2011; Markman and Gartner, 2002)—the top 500 are featured in the magazine's print version, and the top 5000 are listed on the website.

2.2. Outcome variables

Consistent with the most seminal outcome measures in the domain, we collect data on number of *Employees* and amount of annual firm *Revenue*. For the *PSED*, we use data from Wave F (Yr5, collected in 2010). *KFS* data are from the fourth follow-up (covering the business's fifth year, in 2008), and the *INC* data are self-reported for 2010, as published in the Inc. 500's 2011 issue.

2.3. Data analysis and results

To assess each sample's distribution characteristics, we use MATLAB software, version R2010a, and follow the protocol and techniques for calculating power law model fit, as described by Clauset et al. (2009). These authors suggest a three-step method: (1) estimate the parameters for the scaling exponent (α) and the minimum value in the distribution that exhibits