NTC
Joint Research
Discussion Paper Series

Explaining Zipf's Law by Rapid Growth

Visiting Professor, National Tax College, Japan
Fellow, Research institute of Economy, Trade and Industry
Yoshiyuki Arata

Emeritus Professor, University of Tokyo
Honorary President, Policy Research Institute, Ministry of Finance
Hiroshi Yoshikawa

Assistant Professor, Research Dept., National Tax College, Japan
Shingo Okamoto

250600–02ST

税務大学校
National Tax College
https://www.nta.go.jp

Explaining Zipf's Law by Rapid Growth

Yoshiyuki Arata[*†]
Research Institute of Economy, Trade and Industry
National Tax College

Hiroshi Yoshikawa
University of Tokyo
Policy Research Institute, Ministry of Finance

Shingo Okamoto
National Tax College

Date: May 2025

Abstract

That the distributions of firm sales and individual incomes have a Pareto tail is one of the important statistical regularities, and numerous theoretical models have been proposed to explain it. However, recent studies have pointed out a difficulty with these existing models: they predict that the time required for firms to become giants or individuals to be super-rich is excessively long compared to what is observed in empirical data. Furthermore, our empirical data show that Zipf's law holds in the size distributions of younger firms and individuals, contradicting existing models that predict Zipf's law is primarily driven by older firms and individuals. This paper provides an alternative explanation for Zipf's law to address the inconsistencies observed in empirical data. We focus on the heavy-tailed nature of the distribution of growth rates for firm sales and individual incomes, showing that their growth dynamics are characterized by rapid growth over short periods. We show that the emergence of giant firms and the super-rich results from this rapid growth, leading to the formation of Zipf's law. Zipf's law reflects the common growth dynamics underlying firm sales and individual incomes.

Keywords: Zipf's law; Weibull-tail distributions; Principle of a single big jump
JEL codes: D22; D31; L11

---

[*]Corresponding author: y.arata0325@gmail.com

[†]This study is a part of "Joint Statistical Research Program" conducted in collaboration with the National Tax College (NTC), under the approval of the National Tax Agency (NTA), in accordance with "Guideline on the Utilization of National Tax Data in the Joint Statistical Research Program." This study is also a part of the project "Heterogeneity of Economic Agents and Challenges for the Japanese Economy" conducted at the Research Institute of Economy, Trade and Industry (RIETI). The views expressed herein are those of the authors and do not necessarily reflect the views of NTC, NTA, or RIETI. An earlier version of this paper circulated under the title, "Zipf's law without the stationarity assumption." Arata appreciates for the financial support by the Japan Society for the Promotion of Science (Grant Numbers: 21K01438). The authors are grateful to Kazuhiro Inaba for his excellent research assistance, which was invaluable in completing this paper. Conflict of interest: none.

1 Introduction

The Pareto-tail nature of the size distribution of economic agents, known as Zipf's law, is one of the most prominent stylized facts in economics. This statistical regularity has been observed across various fields, including firm size distributions, income and wealth distributions, and city size distributions (see, e.g., Axtell (2001); Gabaix (2009); Luttmer (2010)). Indeed, many economists have long studied the mechanisms underlying this regularity, with early contributions including Champernowne (1953), Wold and Whittle (1957), Simon and Bonini (1958), Ijiri and Simon (1964), and Ijiri and Simon (1977)). Recently, this question has regained attention, leading to the development of more sophisticated theoretical models to explain Zipf's law (e.g., Gabaix (1999); Reed (2001); Luttmer (2007); Luttmer (2011); Gabaix et al. (2016); Beare and Toda (2022)).

However, existing theoretical models exhibit discrepancies with empirical data. Recent studies have discussed the time required for economic agents to reach the tail region of the size distribution, where Zipf's law holds. For instance, regarding firm sales, Luttmer (2011) highlights that the time predicted by theoretical models for firms to grow into a giant firm is excessively long compared to empirical data. Similarly, Gabaix et al. (2016) show that in existing models, the time required for income distributions to converge to a stationary distribution is much longer than observed in empirical data, making it difficult for such models to explain observed fluctuations in income inequality. These inconsistencies with empirical data stem from the assumptions of existing models that the tail of the distribution is formed by older agents (e.g., firms or individuals). However, as shown in Section 2.2, Zipf's law for firm sizes and individual incomes is actually shaped by relatively younger agents. Thus, these empirical findings highlight the need to explore why relatively younger agents play a dominant role in shaping Zipf's law.

This paper aims to provide an alternative explanation for Zipf's law to resolve the inconsistencies with the empirical data. The core idea of our explanation is to analyze, in a probabilistic sense, the most likely patterns (or sample paths) that lead to the emergence of giant firms and super-rich individuals in the tail of the distribution. Unlike existing models, which assume that giant firms and super-rich individuals are the cumulative result of incremental growth in sales or income over a long period, our explanation implies that their emergence is driven by rapid, short-term growth, or *jumps*. These jumps enable agents to reach the tail region of the distribution within a short time, addressing the challenge of the long time required to become a giant firm or super-rich in existing models. Our explanation demonstrates that the characteristic tail shape of the size distribution described by Zipf's law reflects the nature of these jumps.[1]

This paper begins with recent empirical findings on the shape of growth rate distributions. Regarding the distribution of firm sales growth rates, it has been widely recognized since Stanley et al. (1996) that these

[1]In the supplementary paper to this study, Arata et al. (2023), the importance of jumps in the growth process is examined by combining an alternative firm-level dataset with information on the occurrence of mergers. The results confirm that even when samples involving mergers are excluded (i.e., focusing on internal growth), jumps remain significant in the growth process.

distributions deviate from a Gaussian distribution (for surveys, see Coad (2009); Dosi et al. (2017)).Specifically, growth rate distributions are characterized by high kurtosis and heavier tails, making them more closely approximated by a Laplace distribution than a Gaussian distribution (e.g., Bottazzi and Secchi (2006); Arata (2019)). Moreover, recent empirical studies such as Bottazzi et al. (2011) and Dosi et al. (2020) have pointed out that the tails of growth rate distributions are strictly heavier than those of a Laplace distribution, which follows an exponential function. Interestingly, these distinctive properties of growth rate distributions are not limited to firm sales but are also observed in the growth rate distributions of individual incomes. A pioneering study by Guvenen et al. (2021) demonstrated that the growth rate distribution of individual incomes in the United States deviates from a Gaussian distribution, exhibiting a high kurtosis and heavier tails. Moreover, similar shapes in growth rate distribution have been observed not only in U.S. data but also across various countries (see Guvenen et al. (2022)). In this paper, we demonstrate that the emergence of giant firms and super-rich individuals qualitatively differs depending on whether the growth rate distribution has a heavy tail. Specifically, we demonstrate that when the growth rate distribution has heavy tails, the presence of jumps plays a crucial role in the emergence of giant firms and super-rich individuals.

More specifically, the key features of our theoretical explanation for Zipf's law are as follows: Assuming that the (log) growth rates in each period are independent and identically distributed (i.i.d.), the cumulative growth rate over n periods can be expressed as the sum of n i.i.d. random variables. In cases where the growth rate distribution has light tails, the large deviation of the sum of n i.i.d. random variables (i.e., a high growth rate over n periods) arises from the equal contributions of all n variables. In other words, each period's growth rate is of moderate magnitude, and the accumulation of these moderate growth rates over n periods leads to the large deviation in the sum. Existing models assume this growth process: since each growth rate is relatively small, becoming a giant firm or super-rich individual requires sustained moderate growth rates over a long period. Consequently, these models predict that it takes a significant amount of time to reach the tail region of the size distribution. By contrast, when the growth rate distribution has a heavy tail, the large deviation of the sum of n i.i.d. random variables is primarily driven by a single period in which the growth rate takes on an exceptionally large value. In other words, a single period of rapid growth (i.e., the occurrence of a jump) can immediately propel an agent into the tail region of the size distribution. A heavy-tailed growth rate distribution implies the existence of such jumps, demonstrating that reaching the tail region of the size distribution does not necessarily require a long time.

Another feature of our theoretical explanation is that it does not impose the stationarity assumption; that is, we do not assume that the distribution of firm sales or individual incomes reaches a stationary state in the limit $n \to \infty$. Rather than analyzing the limit as $n \to \infty$, we focus on how the distribution of the sum of n i.i.d. random variables evolves as n increases. Specifically, when n is not sufficiently large, the Gaussian approximation based on the central limit theorem, often employed in existing models, cannot be applied to the tail region. Instead, we show that the shape of the distribution in the tail region, where the

Gaussian approximation fails, provides the key to explaining Zipf's law. Our explanation of Zipf's law in the tail region for relatively small n aligns with the empirical observation that Zipf's law is most prominently observed among younger firms and individuals.

The closest study to our analysis in terms of providing a theoretical explanation for Zipf's law is Beare and Toda (2022), and we highlight the differences between their study and ours. Several recent studies have addressed the issue that the time required to become a giant firm or super-rich individual is excessively long (e.g., Luttmer (2011); Gabaix et al. (2016)). To resolve this, these studies propose models where agents are classified into multiple types, with certain types assumed to have persistently higher average growth rates than others, enabling agents of specific types to reach the tail of the distribution more quickly. Beare and Toda (2022) extend this approach within a more general framework, representing the state-of-the-art in this line of research. However, it should be noted that even in models that introduce multiple agent types, the growth processes of each type retain characteristics similar to those of existing models discussed above. Within each agent type, the growth process relies on moderate growth rates accumulated over time to reach the tail of the distribution. The size distribution of each type implies that the higher the tail is considered, the more it is dominated by older agents, which is the same as in existing models. Therefore, when aggregating across all agent types to obtain the overall distribution, the tail regions should predominantly consist of older agents. This, however, contradicts the empirical data (see Section 2.2). Additionally, the assumption of multiple agent types implies the existence of groups of firms or individuals with persistently higher average growth rates over long periods. Yet, recent empirical studies on high-growth firms suggest a different picture (e.g., [Schneck et al (2021, JBVI)]()). High growth is not an intrinsic characteristic of certain firms but rather reflects short-term periods of exceptional growth (i.e., high-growth episodes). Outside of these episodes, their growth processes are indistinguishable from other firms. Our theoretical framework aligns with this finding, suggesting that the growth process is indistinguishable from other agents before the occurrence of a jump. As we demonstrate below, this aspect is also supported by our empirical data.

The main contribution of this paper is to show that our alternative explanation for Zipf's law is consistent with empirical data. For firm sales data, we use the Tokyo Shoko Research (TSR) dataset, which covers millions of firms in Japan. For individual income data, we utilize tax return data provided by the National Tax College of Japan, which records income information for more than 20 million individuals annually. We use these two datasets to test the two key assumptions underlying our theoretical explanation. The first assumption is the random walk assumption, commonly used in existing models, which posits that agents' growth rates are independently and identically distributed. We examine the dependence between growth rates across consecutive periods, focusing not only on the overall dependence within the distribution but also on the tail dependence, which is more critical for explaining Zipf's law (i.e., whether jumps occur consecutively). The results show that while growth rates across consecutive periods are not entirely independent, the dependence in the tail region is weak enough to justify treating them as independent, as assumed in our theoretical

explanation. The second assumption is that the growth rate distributions have heavy tails, particularly tails heavier than the exponential function. To test this assumption, we employ density estimation, mean excess functions, and tail estimation methods proposed by Gardes et al. (2011) and El Methni et al. (2012). All these analyses consistently indicate that the growth rate distribution has heavier tails than the exponential, specifically resembling a Weibull tail. These two empirical findings confirm that the two assumptions underlying our explanation of Zipf's law are consistent with the data.

Finally, we demonstrate that the implications of our theoretical explanation for Zipf's law are consistent with the data. In particular, we provide the following three empirical findings that distinguish our theoretical explanation from existing models: (1) The tail exponent of the size distribution is determined by the tail exponent of the growth rate distribution (and the tail exponent of the initial size distribution). (2) Zipf's law holds not as a result of aggregating agents of different ages, but within the size distribution of agents at each age group (particularly among younger agents). (3) The growth process of agents achieving high growth rates over n periods is not characterized by persistently high average growth throughout the entire period but is instead driven by a single exceptional large jump in one period. Regarding (1), while both the distribution of firm sales and individual incomes exhibit Pareto tails, it is well-known that their tail exponents differ. We show that this difference in the tail exponents of size distributions corresponds to differences in those of growth rate distributions and initial size distributions. In existing models, no direct relationship exists between the tail exponent of the size distribution and that of the growth rate distribution, providing further evidence that our theoretical explanation is consistent with the data. Regarding (2), the size distribution by age highlight inconsistencies between existing models and the data. We demonstrate that this feature of the size distributions by age is consistent with our theoretical explanation. Specifically, for firm sales, we confirm that older groups tend to approximate a Gaussian distribution, a prediction also made by our theoretical explanation. Regarding (3), we examine the conditional probability of growth rates given that the sum of n i.i.d. random variables takes on a large deviation. According to our theoretical explanation, a large deviation in the sum of n i.i.d. random variables is achieved through a single jump among them. As a result, the $n-1$ smallest growth rates, excluding the jump, should follow the same distribution as the unconditional growth rate distribution. In contrast, existing models assume that the n growth rates, on average, achieve higher growth over the entire period. Thus, the distribution of the $n-1$ smallest growth rates should exhibit a positive drift compared to the unconditional distribution. Our empirical data supports the former scenario, providing evidence consistent with the predictions of our explanation.

The remainder of this paper is organized as follows. In Section 2, we explain the characteristics of existing models of Zipf's law and demonstrate their inconsistencies with empirical data. In Section 3, we propose an alternative explanation for Zipf's law. In Section 4, we empirically test the two assumptions underlying our theoretical explanation. In Section 5, we show that the predictions of our theoretical explanation are consistent with the data. In Section 6, we present our conclusions.

2 Inconsistencies with data

In this section, we review existing models of Zipf's law and demonstrate that their predictions are inconsistent with empirical data. In Section 2.1, utilizing Reed (2001)—which can be regarded as a simplified version of Beare and Toda (2022), where the agent type is single and the growth rate distribution is specified—we show that, in existing models, the tail of the size distribution is predominantly occupied by older agents. In Section 2.2, using our empirical data, we focus on the size distributions by age and demonstrate that the tail of the distributions of firm sales and individual incomes are inconsistent with the predictions of existing models.

2.1 Review of existing models

Let us begin with the notation used in the following. Economic entities such as firms or individuals are referred to as agents. The logarithmic value of an agent's size, such as sales or income, is referred to simply as its size and denoted by S.[2] Zipf's law implies that, when plotted on a logarithmic y-axis, the tail of the distribution is represented by a straight line with a slope of $-a$ (referred to as the tail exponent a), as follows:

$$\log \mathbb{P}(S > x) = -ax + b.$$

The theoretical model of Zipf's law given by Reed (2001) comprises two components: the size distribution by age and the proportion of agents of each age within the total population. Let S_n denote the size of an agent of age n. Reed (2001) assume that the size distribution by age, $\mathbb{P}(S_n > x \mid \text{age} = n)$, follows a Gaussian distribution with variance $n\sigma^2$ (for simplicity, we assume a mean of zero). [3] When x is sufficiently large, the tail probability of the Gaussian distribution can be approximated using Mills' ratio as follows:

$$\mathbb{P}(S_n > x \mid \text{age} = n) \approx \frac{\sigma\sqrt{n}}{x\sqrt{2\pi}} e^{-\frac{x^2}{2n\sigma^2}}.$$

It should be noted that the size distribution for each age group is not assumed to follow Zipf's law. By definition, given the size distributions by age, the overall size distribution $\mathbb{P}(S > x)$ can be obtained by summing the size distributions across all ages. Reed (2001) assumes that the proportion of agents in each age group within the total population follows a geometric distribution. This is based on the idea that agents born at a particular time exit each year with a constant probability, so that the survival probability at age n (denoted by p_n) is given by a geometric distribution with an exit probability of p. Thus,

$$\mathbb{P}(S > x) = \sum_n \mathbb{P}(S_n > x, \text{age} = n)$$

$$\mathbb{P}(S_n > x, \text{age} = n) = p_n \mathbb{P}(S_n > x \mid \text{age} = n) \approx p(1-p)^{n-1} \cdot \frac{\sigma\sqrt{n}}{x\sqrt{2\pi}} e^{-\frac{x^2}{2n\sigma^2}}$$

[2] Note that when using S as the size, it does not account for the age of agents.

[3] The rationale in existing models is that, by assuming a stationary distribution in the limit of $n \to \infty$, the probability $\mathbb{P}(S_n > x \mid \text{age} = n)$ can be approximated by a Gaussian distribution for sufficiently large n, based on the central limit theorem.

In this model, which age group of agents accounts for the tail probability $\mathbb{P}(S > x)$? Consider the ratio of the following tail probabilities:

$$\frac{\mathbb{P}(S_{n_2} > x, \text{age} = n_2)}{\mathbb{P}(S_{n_1} > x, \text{age} = n_1)} \approx (1-p)^{n_2 - n_1} \cdot \sqrt{\frac{n_2}{n_1}} \cdot e^{\frac{x^2(n_2 - n_1)}{2\sigma^2 n_1 n_2}}$$

where $n_2 > n_1$. The right-hand side is an increasing function of x. This implies that, as x becomes large, the distribution in the tail region becomes increasingly dominated by older agents. In other words, the model predicts that the Zipf's law observed in the tail region is primarily formed by older agents.

This characteristic of existing models can be explained by two effects as n increases. When birth and exit rates remain constant over time, the number of agents of a given age (i.e., p_n) decreases geometrically as n increases. At the same time, the variance in the size distribution for each age group grows with n, leading to an increase in the tail probability of S_n. For a large value of x, the latter effect becomes dominant, so as x grows, the tail of the distribution of S is dominated by older agents. This feature arises from a common assumption in existing models that Zipf's law can be explained by aggregating size distributions across different n, even though the distribution for each age group does not follow Zipf's law. Thus, the tendency for older agents to dominate the tail region is also a characteristic shared by existing models.[4] The next section examines whether this prediction holds in empirical data.

[4] Another mechanism employed in the previous literature to explain Zipf's law combines Brownian motion with a negative drift and a reflective boundary (e.g., Gabaix (1999)). Specifically, such a model assumes the existence of a lower barrier $S_{\min}$, where S_n behaves as a Brownian motion with negative drift when $S_n > S_{\min}$ and reflects back upon reaching $S_{\min}$. It is known that the stationary distribution of this model satisfies Zipf's law. The reflective boundary is interpreted as representing the balance in the stationary state between the probability of agents passing $S_{\min}$ from below and from above. Based on this interpretation, it can be shown that in this model as well, the tail probabilities are increasingly dominated by older agents as x becomes larger. Consider two agents with different ages n_1 and n_2, where $n_2 > n_1$, each following a normal distribution with means $\mu n_1, \mu n_2$ (where $\mu < 0$) and variances $n_1\sigma^2, n_2\sigma^2$, respectively. The derivative of the logarithm of the ratio of tail probabilities with respect to x is given by:

$$\frac{d}{dx} \log \left(\frac{\mathbb{P}(S_{n_2} > x, \text{age} = n_2)}{\mathbb{P}(S_{n_1} > x, \text{age} = n_1)} \right) \approx \frac{(n_2 - n_1)x}{\sigma^2 n_1 n_2} > 0$$

This implies that the ratio of tail probabilities is an increasing function of x, meaning that as x becomes larger, the tail probabilities are increasingly dominated by older agents.

2.2 Distributions by age

First, we examine the size distribution using Japanese firm-level data, defining firm size as the logarithm of sales.[5] Here, a firm's birth year is defined by its incorporation date. **Figure 1**(a) shows the density estimate of the distribution of firm size S (i.e., without considering age) on a logarithmic scale.[6] As seen in the figure, the right side of the distribution can be approximated by a straight line with a slope close to -1. This indicates that Zipf's law holds for the distribution of firm sales, consistent with the findings in the existing literature.

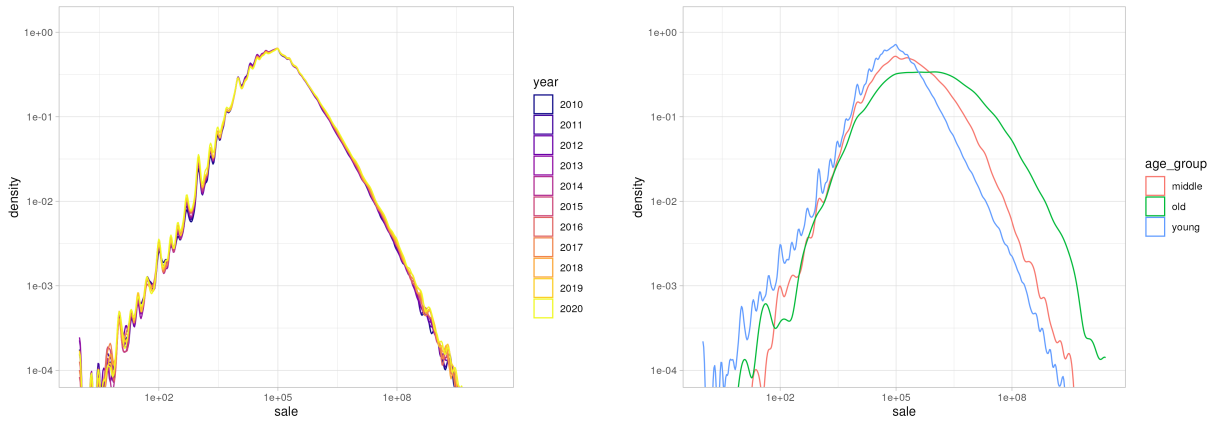
Which age group of firms accounts for Zipf's law observed in the aggregate distribution? **Figure 1**(b) compares firms categorized into three age groups: firms younger than 50 years, those between 50 and 70 years, and those older than 70 years. As illustrated, the distribution for younger firms exhibits a Pareto tail, while that of older firms deviates from the Pareto tail. To further examine the size distribution across different age groups, we divided the sample into 5-year age intervals, comparing the size distributions for each age group. **Figure 2**(a) shows that the distribution for younger firms display a Pareto tail across a wide range, with each size distribution exhibiting a similar slope in the tail region. In contrast, as shown in **Figure 2**(b), the size distribution for older firms deviates from the Pareto tail, instead resembling a bell shape similar to that of a Gaussian distribution. To examine this point further, **Figure 3**(a) presents QQ-plots of the size distributions for each age group. If the size distribution approximates a Gaussian distribution, it should align with the straight line in the plot. As shown in the figure, the tail of the size distribution for younger age groups deviates from the straight line, whereas it approaches the straight line as firm age increases. These results suggest that, contrary to the predictions of existing models which assume older firms form the Pareto tail, the observed Pareto tail in the aggregate distribution is primarily shaped by younger firms.

To clarify the inconsistency between the predictions of existing models discussed in Section 2.1 and the data, **Figure 3**(b) presents the proportions of firms from each age group in the tail probability $\mathbb{P}(S > x)$. The proportion of firms from each age group remains stable or even increases for younger firms as the size x increases. This result contrasts with existing models, which predict that the proportion of older firms should increase in the tail probability as x increases. These findings indicate that, to resolve the inconsistency with the data, an alternative explanation for Zipf's law is required.

Interestingly, the characteristics observed in the distribution of firm sales also appear in the distribution of individual incomes. We define the logarithm of individual income as the size and analyze the shape of its distribution. **Figure 4** displays both the aggregate distribution of size S and the size distributions divided into five-year age groups. Consistent with previous studies, a Pareto tail is evident in the right tail of the

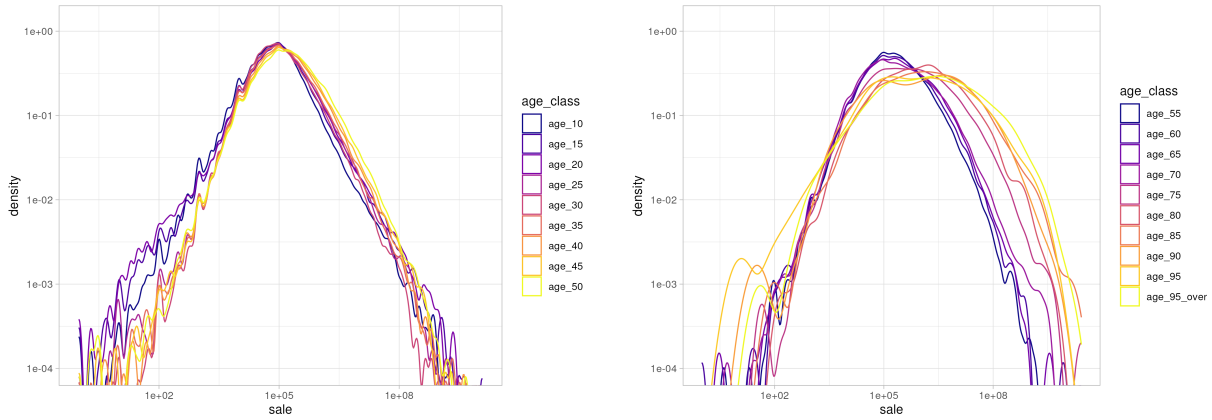
[5] Details about the data used are provided in Section 4.1.

[6] In our analysis, we utilize both the tail probability, $\mathbb{P}(S_n > x)$, and the density function, $\mathbb{P}(S_n \in dx)$, depending on the context. Note that when Zipf's law holds, both the tail probability and density function can be represented as straight lines on a log scale for the y-axis.



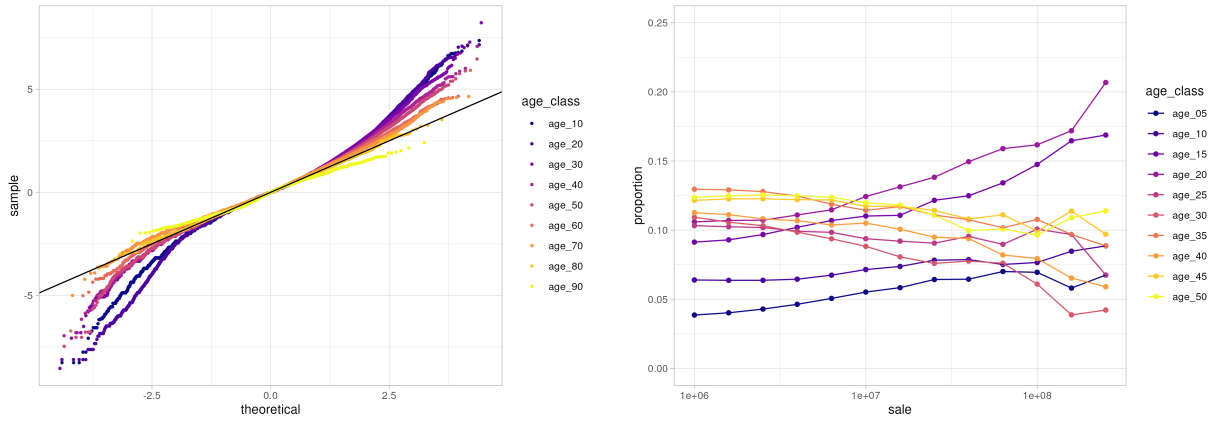
(a) Aggregate **(b)** Three age groups

Figure 1: Density estimates of the size distribution of firm sales. Using the logarithm of firm sales as size, the estimated density function of the aggregate distribution $\mathbb{P}(S > x)$ from 2010 to 2020 and the distributions $\mathbb{P}(S_n > x)$ for three age groups — young, middle, old — are provided in Panel (a) and Panel (b), respectively. The sample sizes for the three age groups used in the estimates in Panel (b) are $932,344$ for young, $152,512$ for middle, and $23,433$ for old. The y-axis is on a logarithmic scale.



(a) Ages 5 to 50 **(b)** Ages 50 and above

Figure 2: Size distributions by age group. The sample is divided into age groups in 5-year interval, with the density function of the size distribution estimated for each group. For instance, age_10 refers to firms between the ages of 5 and 10. Panel (a) shows age groups from 5 to 50, while Panel (b) shows age groups 50 and older. The y-axis is on a logarithmic scale.



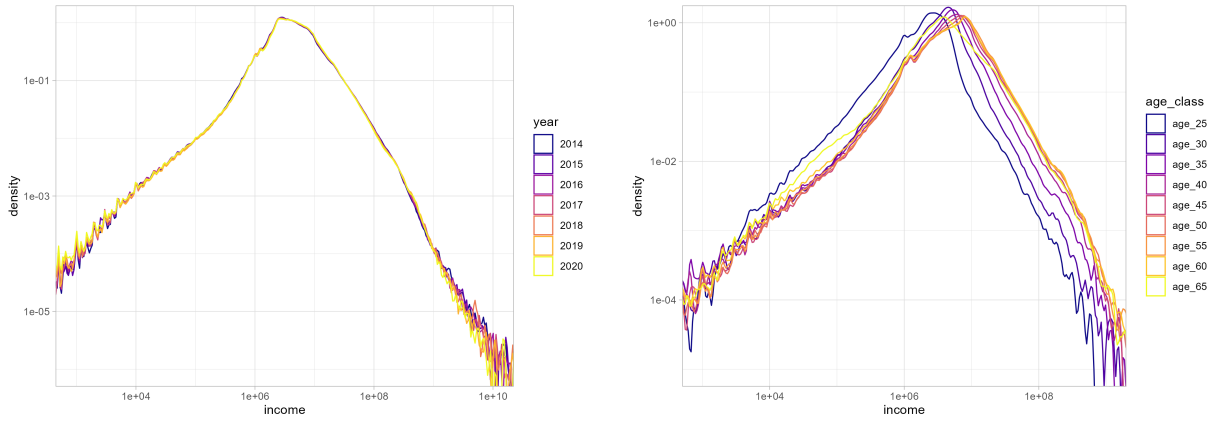
(a) QQ-plot

(b) Proportion in the tail probability

Figure 3: QQ-plots of size distributions by age group and the proportion of firms from each age group in the tail probability. In Panel (a), the straight line corresponds to a Gaussian distribution. For simplicity, only a subset of the age groups in five-year intervals is shown. Panel (b) displays the proportion of firms from each age group within the tail probability $\mathbb{P}(S > x)$ as a function of x.

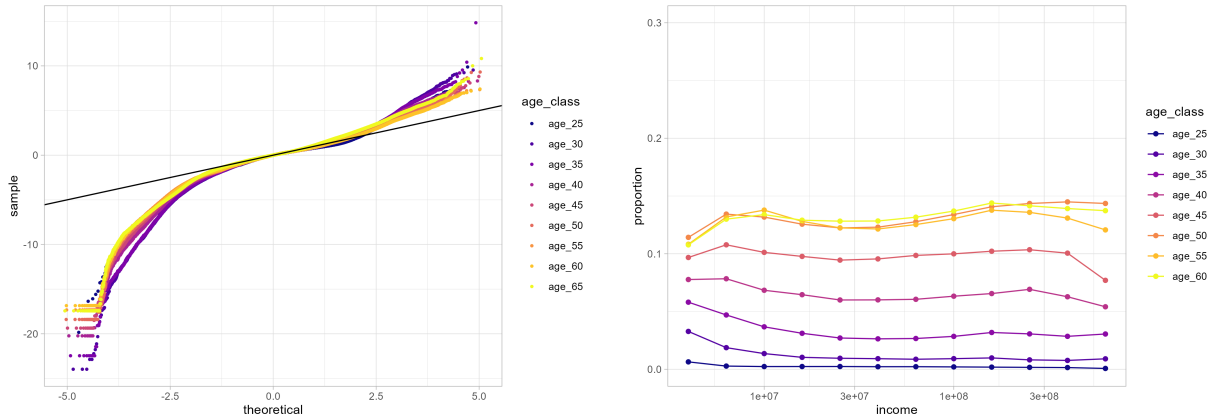
aggregate distribution of individual income, demonstrating that Zipf's law holds. Furthermore, this Pareto tail is observed in the size distributions for each age group, with all size distributions sharing a common slope in the tail region. In addition, **Figure 5**(a) presents QQ-plots for the size distributions by age group. Although convergence toward a Gaussian distribution is less pronounced that in the case of firm sales, the deviation from a Gaussian distribution is clear especially in the younger age groups.

In **Figure 5**(b), the proportion of individuals from each age group in the tail probability, $\mathbb{P}(S > x)$, is shown. As illustrated, this proportion remains stable across different sizes, x, with no trend of increasing proportions of older individuals as x increases. These results suggest that Zipf's law already holds within the size distribution of younger individuals, contradicting the predictions of existing models. Zipf's law is not formed as a result of aggregating across different age groups. In the following section, we provide an alternative explanation of Zipf's law that aligns with these empirical findings observed in both firm sales and individual incomes.



(a) Aggregate **(b)** Each age group

Figure 4: Density estimates of the size distribution of individual incomes. Using the logarithmic value of individual income as size, Panels (a) and (b) present density estimates for the aggregate distribution $\mathbb{P}(S > x)$ from 2014 to 2020, and for the distribution $\mathbb{P}(S_n > x)$ for the 2020 sample divided by five-year age groups, respectively. For example, age_30 refers to the group of individuals aged 25 to 30 years. The y-axis is on a logarithmic scale.



(a) QQ-plot **(b)** Proportion in the tail probability

Figure 5: QQ-plots of size distributions by age group and the proportion of firms from each age group in the tail probability. In Panel (a), the straight line corresponds to a Gaussian distribution. Panel (b) displays the proportion of individuals from each age group in the tail probability $\mathbb{P}(S > x)$ as a function of x.

3 Alternative explanation

In this section, we provide an alternative explanation for Zipf's law. In Section 3.1, we explain that the tail of the distribution of the sum of n i.i.d. random variables cannot be approximated by a Gaussian distribution. In Section 3.2, we focus on how the sum of n i.i.d. random variables changes as n increases and discuss what determines the tail behavior. In Section 3.3, we examine the impact of the initial size on the size distribution. In Section 3.4, we analyze how large deviations in the sum of n i.i.d. random variables occur.

3.1 Setup

Let us introduce our notations. As in the previous section, let the size at the initial point (i.e., the logarithm of a firm's sales or an individual's income) be S_0, and define the growth rate at period k (i.e., the difference in logarithmic values) as X_k. Thus, the growth rate over n periods (denoted by $\widetilde{S}_n$) is expressed as the sum of n individual growth rates, and the size at time n (denoted by S_n) is expressed as the initial size plus $\widetilde{S}_n$:

$$\widetilde{S}_n := X_1 + \cdots + X_n, \quad S_n := S_0 + \widetilde{S}_n$$

Here, we assume that S_n follows a random walk as follows.

Assumption 3.1. S_n follows a random walk with an initial condition S_0, i.e., n random variables $X_1, ..., X_n$ are independent and identically distributed with mean 0 and variance σ^2.

Under this assumption, $\widetilde{S}_n$ can be viewed as the sum of n iid random variables. The question we need to address here is: what is the distribution of $\widetilde{S}_n$? According to the central limit theorem, the normalized sum of n iid random variables converges to the standard Gaussian distribution as n goes to infinity (see Chapter 5 of Petrov (1995)). More precisely, using the following notation,

$$Z_n = \sigma^{-1} n^{-1/2} \widetilde{S}_n, \quad F_n(x) = \mathbb{P}(Z_n < x)$$

the central limit theorem implies that for any fixed x,

$$\frac{1 - F_n(x)}{1 - \Phi(x)} \to 1, \ \frac{F_n(-x)}{\Phi(-x)} \to 1 \quad \text{as } n \to \infty \tag{1}$$

Here, Φ is the standard Gaussian distribution. From this theorem, one might think that if n is sufficiently large, the distribution of $\widetilde{S}_n$ (or S_t) can be well approximated by a Gaussian distribution across all regions. Indeed, many previous studies (e.g., Reed (2001)) assume that the distribution of firms or individuals in the same generation (i.e., the same age) can be approximated by a normal distribution due to the central limit theorem. However, this view is generally incorrect because the convergence to a Gaussian distribution occurs only for *fixed* x.

Now, what if x depends on n? Regarding this problem, the most widely used result in probability theory is the following theorem proved by Harald Cramér:

Theorem 3.1 (Theorem 5.23 in Petrov (1995)). *Suppose that Cramer's condition holds.*[7] *Then for* $x \geq 0, x = o(n^{1/2})$,

$$\begin{aligned} \frac{1-F_n(x)}{1-\Phi(x)} &= \exp\left\{\frac{x^3}{\sqrt{n}}\lambda\left(\frac{x}{\sqrt{n}}\right)\right\}\left[1+O\left(\frac{x+1}{\sqrt{n}}\right)\right], \\ \frac{F_n(-x)}{\Phi(-x)} &= \exp\left\{-\frac{x^3}{\sqrt{n}}\lambda\left(-\frac{x}{\sqrt{n}}\right)\right\}\left[1+O\left(\frac{x+1}{\sqrt{n}}\right)\right] \end{aligned} \tag{2}$$

where $\lambda(t) = \sum_{k=0}^{\infty} c_k t^k$ *is called Cramer's series, which is power series with coefficients that depend only on the cumulant of random variable* X_1.

The right-hand side of Eq.(2) is known as the Cramer correction, which captures the deviation from the Gaussian distribution. As indicated by Eq.(2), by imposing the additional condition $x = o(n^{1/6})$—which focuses on a narrower region around $x = 0$—we can recover Eq.(1). This implies that normal convergence under the central limit theorem begins in the vicinity of $x = 0$ and gradually extends as n increases. However, beyond the zones of $o(n^{1/6})$ or $x = o(n^{1/2})$, normal convergence is generally not guaranteed. This poses a significant issue for our analysis, particularly when considering the tail behavior of distributions, such as those described by Zipf's law. In real data, the value of n is finite, while our focus often lies on large values of x, which correspond to the distribution's tail. Therefore, explaining Zipf's law requires accounting for regions where normal convergence does not hold. As will be discussed in detail below, the core of our explanation of Zipf's law concerns the tail behavior of the distribution, where normal convergence does not occur.

Before discussing the general setting, we examine two cases where X_k follows specific distributions and observe how the distribution of the sum $\widetilde{S}_n$ changes as n increases. The first case is when X_k follows a Laplace distribution, with its probability density given by:

$$\mathbb{P}(dx) = \frac{1}{2}\exp(-|x|)dx$$

The second case involves a distribution with a Weibull tail given by:

$$\mathbb{P}(X_k > x) = \frac{1}{2}\exp(-|x|^{\alpha}), \quad 0 < \alpha < 1$$

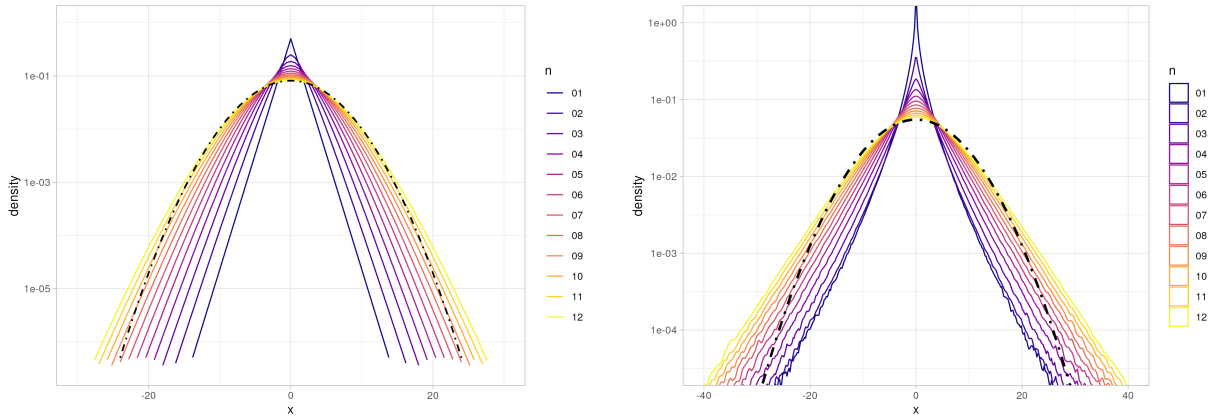
In the former case, we can derive an explicit expression for the distribution of $\widetilde{S}_n$.[8] In the latter case, while an explicit expression for the distribution is not obtainable, we generate pseudo-samples through simulation and estimate the density function using these samples (with α set to 0.7).

[7] For Cramér's condition, refer to the next section. As discussed in the next section, the convergence to a Gaussian distribution around $x = 0$ holds even when Cramér's condition is not satisfied.

[8] When X_k follows a Laplace distribution, the probability density function of $\widetilde{S}_n$ is given by the following expression (cf. Chapter 2.3.1 in Kotz et al. (2001)):

$$\mathbb{P}_{\widetilde{S}_n}(dx) = \frac{e^{-|x|}}{(n-1)!2^n}\sum_{j=0}^{n-1}\frac{(n-1+j)!}{(n-1-j)!j!}\frac{|x|^{n-1-j}}{2^j}$$

Using this, the probability density function of $\widetilde{S}_n$ is depicted in **Figure 6**(a).



(a) Laplace distribution

(b) Weibull tail distribution

Figure 6: Probability density functions of the sum $\widetilde{S}_n$ for various values of n. In both panels, the probability density functions of the sum $\widetilde{S}_n$, for $n = 1, \ldots, 12$, are presented. For reference, a Gaussian distribution is also included, with σ set to match the standard deviation for $n = 12$.

Figure 6 illustrates the density functions of the distribution of $\widetilde{S}_n$ on a logarithmic scale for both the Laplace distribution and the Weibull tail distribution. As is evident from the figure, for small values of x (i.e., around $x = 0$), the central peak of the density function flattens and approaches a bell shape as n increases. This behavior reflects the normal convergence discussed above, where the distribution approaches a Gaussian distribution in the central region as n grows. On the other hand, in the large x region (i.e., the tail region), the deviation from the Gaussian distribution becomes significant. In both figures, the density function of $\widetilde{S}_n$ in the tail region, when plotted on a logarithmic scale, shifts upward parallel to the probability density function of X_k as n increases. This indicates that, even as n increases, the tail behavior of the distribution of the sum $\widetilde{S}_n$ is determined by the tail probabilities of its individual components, $\mathbb{P}(X_k > x)$. A rigorous explanation of this phenomenon will be provided in the next section.

3.2 Three zones of the distribution of $\widetilde{S}_n$

This section discusses how the probability distribution of $\widetilde{S}_n$ is characterized depending on n and x. As will be discussed below, the behavior of $\widetilde{S}_n$ varies significantly depending on whether the tail of the distribution of its components X_k is heavier or lighter than the exponential function. More precisely, we refer to the distribution of X_k as light-tailed if it satisfies the following Cramér condition: for some $\lambda > 0$,

$$Ee^{\lambda X_k} < \infty$$

For example, the Gaussian and Laplace distributions are examples of light-tailed distributions. On the other hand, if the distribution is not light-tailed (i.e., if $Ee^{\lambda X_k}$ is not finite for any $\lambda > 0$), it is referred to as a heavy-tailed distribution. In particular, heavy-tailed distributions that satisfy (very weak) regularity conditions[9] are called subexponential distributions. Distributions with heavy tails, such as the Weibull and Pareto distributions, belong to the class of subexponential distributions.

One of the most important properties of subexponential distributions is that when $X_1, \ldots, X_n$ are i.i.d., the following approximation holds for the tail probability of the sum $\widetilde{S}_n$:

$$\mathbb{P}(\widetilde{S}_n > x) \sim n\mathbb{P}(X_k > x) \quad \text{as } x \to \infty$$

In particular, recall that the term $n\mathbb{P}(X_k > x)$ on the right-hand side represents the probability that the maximum of the elements, $\max\{X_1, \ldots, X_n\}$, exceeds x.[10] This property indicates that as $x \to \infty$, the tail probability of the sum is asymptotically equivalent to the tail probability of the maximum element. In other words, the large deviation of the sum is driven by the large deviation of the largest individual element.

In this analysis, we assume that the growth rates $X_1, X_2, \ldots, X_n$ are random variables following a common subexponential distribution. Subexponential distributions can generally be divided into two categories: those with Pareto tails and those with Weibull tails. We specifically assume that the growth rate distribution belongs to the class with Weibull tails. More precisely, this assumption is formalized as follows (the empirical validity of this assumption will be discussed in Section 4.3):

---

[9] The regularity condition for a heavy-tailed distribution to be subexponential is as follows: Let F denote the distribution of the random variable $X_k^+ := \max\{0, X_k\}$ on the positive real half-line $\mathbb{R}^+$, and let $\overline{F}(x) := F[x, \infty)$. The following limit exists:

$$\lim_{x \to \infty} \frac{\overline{F * F}(x)}{\overline{F}(x)}$$

where $F * F(x)$ represents the convolution of F. Empirically, heavy-tailed distributions often used in applied analysis (e.g., Weibull tail, Pareto tail) satisfy this condition, so in practice, subexponential distributions can be regarded as equivalent to heavy-tailed distributions. For further details, refer to Chapter 3 in Foss et al. (2011).

[10] This can be proven as follows. Letting F be the distribution function of X_k (i.e., $F(x) = \mathbb{P}(X_k \leq x)$), the tail probability of the maximum $\max\{X_1, \ldots, X_n\}$ can be expressed as:

$$\mathbb{P}(\max\{X_1, ..., X_n\} > x) = 1 - F^n(x) = (1 - F(x)) \sum_{k=0}^{n-1} F^k(x) \sim n(1 - F(x)), \quad \text{as } x \to \infty$$

Assumption 3.2. The growth rate distribution is of the form of the following Weibull-like distribution:

$$\mathbb{P}(X_k > x) = e^{-\ell(x)}, \quad \ell(x) := x^\alpha L(x),\ 0 < \alpha < 1$$

where $L(x)$ is a slowly varying function at infinity.

One characteristic of a Weibull tail is that, for sufficiently large x, the change in the slope of the tail on a log scale (i.e., the second-order derivative of $\ell(x)$) becomes small. In other words, over a wide range of the tail, the distribution appears to follow a straight line. This property is utilized in Section 4.3 and Section 5.1.

What shape does the distribution of the sum $\widetilde{S}_n$ take under the above assumptions? As can be inferred from the discussion in Section 3.1, the central limit theorem suggests that the distribution of $\widetilde{S}_n$ can be approximated by a Gaussian distribution around $x = 0$, and this region expands as n increases. On the other hand, due to the properties of subexponential distributions, the tail of the distribution of $\widetilde{S}_n$ should be approximated by the tail probability of the growth rates X_k, multiplied by n. Thus, the characteristics of the distribution of $\widetilde{S}_n$ depend on both n and x. A rigorous proof of this behavior is provided below.

Theorem 3.2 (Theorem 5.4.1 in Borovkov and Borovkov (2008))**.** *For $x \le \sigma_1(n)$,*

$$\mathbb{P}(\widetilde{S}_n \geqslant x) = \left[1 - \Phi\left(\frac{x}{\sqrt{n}}\right)\right] e^{-n\Lambda_\kappa^0(x/n)}(1 + o(1))$$

Here, $\Lambda_\kappa^0(x/n) := \Lambda_\kappa(x/n) - \frac{x^2}{2n^2}$, where $\Lambda_\kappa(x/n)$ is the truncated Cramér series. For $x \gg \sigma_1(n)$,

$$\mathbb{P}(\widetilde{S}_n \geqslant x) = ne^{-M(x,n)}(1 + \varepsilon(x, n))$$

In particular, for $x \gg \sigma_2(n)$,

$$\mathbb{P}(\widetilde{S}_n \geq x) = n\mathbb{P}(X_k > x)(1 + o(1))$$

Here, boundaries $\sigma_1(n)$ and $\sigma_2(n)$ are given by $\sigma_1(n) := n^{1/(2-\alpha)}L_1(n)$ and $\sigma_2(n) := n^{1/(2-2\alpha)}L_2(n)$ with L_1 and L_2 being some slowly varying functions, respectively.

According to this theorem, the distribution of $\widetilde{S}_n$ can be divided into three regions depending on x and n: (i) the Cramér approximation region, (ii) the intermediate deviation region, and (iii) the extreme deviation region. First, in region (i) (i.e., $x \le \sigma_1(n)$), similar to the result discussed in Section 3.1, the Cramér approximation holds near $x = 0$. Specifically, in a narrower region around $x = 0$, normal convergence applies, and the distribution can be approximated by a Gaussian distribution. On the other hand, region (iii) corresponds to the domain of the principle of a single large jump, where the distribution is determined by the tail probability of the individual elements X_k. In the intermediate region (ii), $\sigma_1(n) \ll x \ll \sigma_2(n)$, the distribution of $\widetilde{S}_n$ exhibits a mixture of the properties of both the Cramér and extreme deviation regions, making it generally difficult to have a simple expression. However, when considering the distribution on a logarithmic scale (i.e., $\log \mathbb{P}(\widetilde{S}_n > x)$), the following approximation holds by using the relation $M = \ell(x)(1 + o(1))$: For $x \gg \sigma_1(n)$,

$$\log \mathbb{P}(\widetilde{S}_n > x) = (1 + o(1)) \log n\mathbb{P}(X_k > x)$$

This implies that if we are interested in the shape of the distribution on a logarithmic scale, as in Zipf's law, then for $x \gg \sigma_1(n)$, the distribution can be approximated by $\log n\mathbb{P}(X_k > x)$, in both the intermediate and extreme deviation regions. In the next section, we analyze the shape of the distribution of S_n, which includes the initial size S_0, depending on n and x.

3.3 Initial size and the distribution of S_n

In this section, we consider the distribution of S_n when the distribution of S_0 is also taken into account and analyze how the shape of S_n varies with x and n. Note that S_n becomes a combination of two random variables, $\widetilde{S}_n$ and S_0. Here, we assume that S_0 follows another subexponential distribution (different from that of X_k), particularly a Weibull tail distribution; i.e.,

$$\mathbb{P}(S_0 > x) = e^{-\ell_0(x)}, \quad \ell_0(x) := x^{\alpha_0} L_0(x),\ 0 < \alpha_0 < 1$$

where $L_0(x)$ is a slowly varying function.[11] Theorem 3.2 can be extended as follows.

Theorem 3.3 (Theorem 11.3.1(iii) in Borovkov and Borovkov (2008))**.** *For $x \gg \sigma_1(n)$ and $x \gg \sigma_1^0(n)$,*

$$\mathbb{P}(S_n > x) \sim e^{-M_0(x,n)} + ne^{-M(x,n)}$$

In particular, for $x \gg \sigma_2(n)$ and $x \gg \sigma_2^0(n)$,

$$\mathbb{P}(S_n > x) \sim \mathbb{P}(S_0 > x) + n\mathbb{P}(X_k > x)$$

Here, $\sigma_1(n), \sigma_2(n)$, and $M(x,n)$ are the same as given in Theorem 3.2. The functions $\sigma_1^0(n)$ and $\sigma_2^0(n)$ take the form of $\sigma_1^0(n) = n^{1/(2-\alpha_0)} L_3(n)$ and $\sigma_2^0(n) = n^{1/(2-2\alpha_0)} L_4(n)$, where L_3 and L_4 are slowly varying functions. For $x \gg \sigma_1^0(n)$, $M_0(x,n)$ takes the form of $M_0 = \ell_0(x)(1 + o(1))$.

Based on this theorem, let us consider the slope of the tail of the distribution of S_n in the logarithmic scale. First, in the extreme deviation region, the theorem indicates that the tail of the distribution of S_n is determined by the tail probabilities of S_0 and X_k. Note that the slopes of the tails of the distributions of S_0 and X_k in the logarithmic scale are given by $\ell_0'(x)$ and $\ell'(x)$, respectively. Therefore, by considering the logarithmic scale and differentiating the right-hand side, we obtain the following equation:

$$\frac{d}{dx} \log \mathbb{P}(S_n > x) = -w_0 \ell_0'(x) - w_n \ell'(x),$$
$$w_0 := \frac{\mathbb{P}(S_0 > x)}{\mathbb{P}(S_0 > x) + n\mathbb{P}(X_k > x)},\ w_n := \frac{n\mathbb{P}(X_k > x)}{\mathbb{P}(S_0 > x) + n\mathbb{P}(X_k > x)}$$

To analyze the tail exponent of Zipf's law, we consider the case where x is sufficiently large and $\ell_0(x)$ and $\ell(x)$ can be approximated as follows:

$$\ell_0(x) \approx a_0 x + b_0, \quad \ell(x) \approx a_1 x + b_1$$

In particular, when the slopes of these two approximations are equal ($a_0 = a_1$), the slope of the tail of $\log \mathbb{P}(S_n > x)$ also matches this common slope. For different values of n, the slope of the tail of

[11] The assumptions made here are not as crucial to our analysis as those in Assumption 3.1 and Assumption 3.2. This is because if the tail of the distribution of S_0 is substantially lighter than that of the growth rates X_k (or $\widetilde{S}_n$), the influence of S_0 on S_n is negligible, and the distribution of S_n can be regarded as identical to that of $\widetilde{S}_n$. On the other hand, if the distribution of S_0 is substantially heavier than that of the growth rates X_k, S_n will be dominated by S_0, and the distribution of S_n will be the same as that of S_0. Therefore, aside from these trivial cases, our interest here lies in situations where the shape of S_0 is similar to that of the growth rates X_k (or $\widetilde{S}_n$), which is why we assume that the distribution of S_0 takes the form $e^{-\ell_0(x)}$. It should be noted that ℓ_0, α_0, and L_0 may generally differ from the corresponding parameters ℓ, α, and L of the growth rate distribution, making this a weak assumption.

$\log \mathbb{P}(S_n > x)$ remains the same, with only the intercept increasing as n grows. When the slopes of these two approximations differ ($a_0 \neq a_1$), the slope of the tail of $\log \mathbb{P}(S_n > x)$ is expressed as a weighted average of a_0 and a_1, with weights w_0 and w_n. In particular, when $a_0 < a_1$ (i.e., the distribution of S_0 has a heavier tail than that of X_k), w_0 is an increasing function of x. As a result, especially for small n, the slope of the tail of $\log \mathbb{P}(S_n > x)$ tends to be closer to that of the distribution of S_0. As n increases, the weight w_n grows, causing the slope to approach that of the distribution of X_k.[12]

A similar argument can be made regarding the tail of the distribution of S_n in the intermediate deviation region. Using the fact that the functions M_0 and M can be approximated in this region by $\ell_0(x)(1 + o(1))$ and $\ell(x)(1 + o(1))$, respectively, the slope of $\log \mathbb{P}(S_n > x)$ is given by the following expression:

$$\frac{d}{dx} \log \mathbb{P}(S_n > x) = (1 + o(1))(-w_0 \ell_0'(x) - w_n \ell'(x))$$
$$w_0 := \frac{e^{-M_0(x,n)}}{e^{-M_0(x,n)} + ne^{-M(x,n)}}, w_n = \frac{ne^{-M(x,n)}}{e^{-M_0(x,n)} + ne^{-M(x,n)}}$$

As in the case of the extreme deviation zone, the slope of the tail of $\log \mathbb{P}(S_n > x)$ is expressed as a weighted average of a_0 and a_1. Thus, our analysis demonstrates that the slope of the tail of S_n is directly related to the slopes of the tails of S_0 and X_k.

[12] We consider the case $a_0 = a_1$ as the scenario where Zipf's law holds more strictly.

3.4 Patterns driving the large deviations in $\widetilde{S}_n$

Here, we return to $\widetilde{S}_n$ and examine how large deviations in $\widetilde{S}_n$ are generated. Specifically, we focus on the most likely events given $\widetilde{S}_n > x$, where x is a large value. As demonstrated below, the way in which $\widetilde{S}_n > x$ occurs differs depending on whether the growth rate distribution is light-tailed or heavy-tailed.

As an illustrative example, let X_k be a non-negative random variable with a probability density function denoted by f (i.e., $F(dx) = f(x)dx$).[13] Assume that $f(x)$ can be written as follows:

$$f(x) = e^{-h(x)}, \quad x \geq 0$$

For instance, if $f(x)$ is the density of an exponential distribution, then $h(x) = x$. When n iid random variables follow the distribution F, the probability that their sum equals x is given by

$$\mathbb{P}(\widetilde{S}_n = x) = \int_{\widetilde{S}_n = x} \exp\left(-\sum_{k=1}^{n} h(X_k)\right) dX_1 \dots dX_n$$

Given the sum equals u, which combination of $X_1, ..., X_n$ is most likely to occur? This question is equivalent to minimizing the sum $\sum_{k=1}^{n} h(X_k)$ subject to the condition that $\widetilde{S}_n = x$.

Let us consider the case where h is a convex function (e.g., a Weibull distribution with $\alpha > 1$). Jensen's inequality implies that the minimum of the sum $\sum_{k=1}^{n} h(X_k)$ is attained at $X_1 = ... = X_n = x/n$.[14] In other words, the most likely combination of $X_1, ..., X_n$ that produces the sum $\widetilde{S}_n = x$ is the one where all components have the same value of x/n. Thus, it is most probable that each component contributes equally to the sum. In contrast, when h is a concave function (e.g., a Weibull distribution with $\alpha < 1$), the way components $X_1, ..., X_n$ generate the sum $\widetilde{S}_n = x$ differs qualitatively. The sum $\sum_{k=1}^{n} h(X_k)$ is minimized when $X_{k^*} = x$ for some $k = k^*$ and $X_1 = \dots = X_n = 0$ for $k \neq k^*$.[15] This suggests that a single component dominates the sum while other components contribute nothing. Lastly, note that the boundary case is the exponential distribution, in which h is a linear function. In this case, both types of behavior could occur.

The fact that the sum of random variables can be generated in two different ways can be illustrated by considering the ratio of the contribution of X_1 within the sum.[16] Consider two independent random variables

[13] This example is taken from Chapter 3 in Sornette (2006).

[14] Indeed, let $\widehat{X}_k$ be the deviation from x/n, i.e., $\widehat{X}_k := X_k - x/n$. Jensen's inequality states that for a real convex function φ, $\varphi\left(\frac{\sum_k x_k}{n}\right) \leq \frac{\sum_k \varphi(x_k)}{n}$. Thus,

$$\sum_k h(X_k) = h\left(\frac{x}{n} + \widehat{X}_1\right) + \ldots + h\left(\frac{x}{n} + \widehat{X}_n\right) \geq nh\left(\frac{x}{n}\right)$$

where we used $\sum_k \widehat{X}_k = 0$ by definition.

[15] This can be shown as follows: Suppose that the statement does not hold. Thus, there exist at least two k such that $0 < X_k < x$. Take such two k (denoted by k_1, k_2) so that $X_{k_1} \geq X_{k_2}$. The concavity of the function h yields that $\sum_k h(X_k)$ can be lowered by replacing X_{k_1}, X_{k_2} with $X_{k_1} - \varepsilon, X_{k_2} + \varepsilon$. This is a contradiction.

[16] This example is taken from Chapter 1 in Foss et al. (2011).

$X_1, X_2 \geq 0$ drawn from a Weibull distribution with parameter α. We examine the distribution of the ratio $X_1/(X_1+X_2)$ conditioned on the event that their sum equals x (i.e., $X_1+X_2=x$). The probability density function of the ratio given x (denoted by $g_{\alpha,x}$) is as follows:

$$g_{\alpha,x}(r) = c(r(1-r))^{\alpha-1}e^{-x^\alpha(r^\alpha+(1-r)^\alpha)} \tag{3}$$

where c is a normalizing constant independent of r.[17]

Figure 7 displays the density $g_{\alpha,x}$ for three different values of α. As observed in the figure, it is symmetric at $1/2$ for all cases, which is obvious since X_1 and X_2 are two independent and identically distributed. Let us discuss the density more closely for each value of α. In the case of $\alpha > 1$ (i.e., a light-tailed case), the density is unimodal and peaked at $1/2$. This indicates that the most probable event is for X_1 and X_2 to be of similar size (i.e., $X_1 = X_2 = x/2$). Particularly for a large value of x, the density is concentrated around $1/2$. In contrast, for the case where $\alpha < 1$ (i.e., a heavy-tailed case), the density peaks at 0 and 1, exhibiting a U-shaped curve. That is, it becomes more probable that either X_1 or X_2 (but not both) takes a large value and dominates the sum x. Furthermore, as suggested by Eq.(3), the density concentrates at 0 and 1 as $x \to \infty$. Thus, for large values of u, it is highly unlikely that both X_1 and X_2 are large and contribute equally to the sum. Finally, consider the case where $\alpha = 1$ (i.e., the exponential case). The density is uniform, and this case can be seen as a boundary case. In this way, how each component contributes to the total sum is determined by whether the tail of the distribution is heavier than an exponential.

The intuition provided above is more rigorously formalized as follows. First, we consider the case where the distribution of the growth rate X_k is light-tailed. When the growth rate distribution is light-tailed, the moment generating function can be defined (denote it by $\psi(\lambda) := \mathbb{E}e^{\lambda X_k}$). We introduce the Cramér transform of the random variable X_k, denoted X_k^β, as follows:

$$\mathbb{P}(X_k^\beta \in dx) = \frac{e^{\lambda(\beta)x}\mathbb{P}(X_k \in dx)}{\psi(\lambda(\beta))},$$

where $\lambda(\beta)$ is the value of λ that gives the supremum of $\beta\lambda - \log\psi(\lambda)$, i.e., $\lambda(\beta) := \arg\sup_\lambda(\beta\lambda - \log\psi(\lambda))$. When the sum $\widetilde{S}_n$ takes on large values, the conditional probability of the growth rate X_k is given by the following.

Theorem 3.4 (Corollary 3.1.2 in Borovkov (2020))**.** *Suppose that $\beta = x/n \to \beta_0$ as $n \to \infty$. Then, for any*

[17] This can be proved as follows: Let ξ_1, ξ_2 be random variables such that $\xi_1 := \frac{X_1}{X_1+X_2}, \xi_2 := X_1 + X_2$. Then,

$$\begin{aligned}\Pr(\xi_1 = r|\xi_2 = u) &= \frac{\Pr(\xi_1 = r, \xi_2 = u)}{\Pr(\xi_2 = u)} \\ &= \frac{\Pr(X_1 = ru, X_2 = u(1-r))}{\Pr(\xi_2 = u)}\end{aligned}$$

The numerator is calculated using the fact that X_1 and X_2 are independent of each other. The denominator is determined by u and independent of r. Setting the normalizing constant c, we obtain the result.

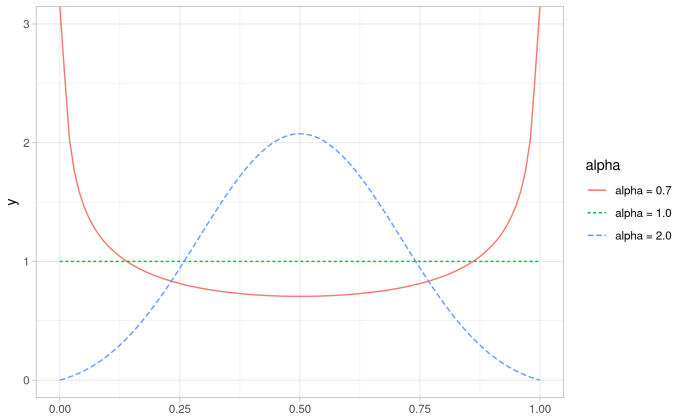


Figure 7: Plots of the density function $g_{\alpha,u}$ with $\alpha = 0.7$, 1.0, and 2.0.

Borel sets $B_1, \ldots, B_m$ *from* $\mathbb{R}$*, any* $k_1, \ldots, k_m$*,*

$$\prod_{i=1}^{m} \mathbb{P}\big(X_i^{\beta_0} \in B_i\big) = \lim_{n\to\infty} \mathbb{P}\big(X_{k_1} \in B_1, \ldots, X_{k_m} \in B_m \mid \widetilde{S}_n \in [x, x+\Delta)\big)$$

Note that Theorem 3.4 considers the case where x increases on the order of n. In such cases of large deviations of $\widetilde{S}_n$ (i.e., when the cumulative growth rate $\widetilde{S}_n$ over n periods is exceptionally high), the conditional distribution of the growth rates X_k conditional on this event is given by that of their Cramér transform, X_k^{β}. As an example, let us consider the Gaussian distribution and its Cramér transform. If X_k follows a Gaussian distribution with mean μ and variance σ^2, the moment generating function is given by $\psi(\lambda) = e^{\mu\lambda+\sigma^2\lambda^2/2}$, and $\lambda(\beta) = \frac{\beta-\mu}{\sigma^2}$. Therefore, the Cramér transform of X_k results in a Gaussian distribution with mean β and variance σ^2 (i.e., for a Gaussian distribution, the Cramér transform simply shifts the distribution horizontally along the x-axis). That is, if we assume $\mu = 0$ for simplicity, the unconditional probability distribution of the growth rate X_k (i.e., when considering all samples) has a mean of 0. In contrast, if we focus only on those samples that achieve a large deviation $\widetilde{S}_n = x = n\beta$, the average growth rate for those samples would be $\beta = x/n$. In other words, the most typical pattern that results in the large deviation $\widetilde{S}_n = x$ is one where the growth rate increases gradually by x/n each year.

The properties of the conditional probability of the growth rates differ significantly when the growth rate distribution is subexponential, compared to when it is light-tailed. The rigorous results are provided by Armendáriz and Loulakis (2011).

Theorem 3.5 (Theorem 2 in Armendáriz and Loulakis (2011))**.** *Let* μ *be the probability measure of* X_k*, i.e.,* $\mu(A) := \mathbb{P}(X_k \in A)$*. Suppose that* μ *is subexponential. Then, the conditional probability* $\mathbb{P}((X_1, ..., X_n) \in \cdot \mid \widetilde{S}_n > x)$ *converges in the total variance to a product of* $n-1$ *copies of* μ *and* ν_x*, where* $\nu_x(A) := \mathbb{P}(X_k \in A \mid X_k > x)$*.*

Theorem 3.5 states that when a large deviation $\widetilde{S}_n > x$ occurs, the $n-1$ smallest values of its components follow the distribution μ^{n-1}, while the largest component follows the distribution ν_x. Note that

μ represents the unconditional distribution of X_k. In other words, the large deviation of $\widetilde{S}_n$ is driven solely by the maximum value among $X_1, \ldots, X_n$ (i.e., a jump), while the other growth rates remain distributed according to the unconditional growth rate distribution.[18] This property is in contrast to Theorem 3.4 and can be used for empirical verification with data. Specifically, when selecting the $n-1$ smallest growth rates from $X_1, \ldots, X_n$ for each firm or individual, they should follow the same distribution as the unconditional growth rates. On the other hand, if the distribution of X_k is light-tailed, then the conditional probability distribution of X_k given $\widetilde{S}_n > x$ will follow the Cramér transform X_k^{β}. Therefore, by examining empirical data to determine which of these two cases it most closely resembles, we can identify the most typical pattern that leads to $\widetilde{S}_n > x$.

[18] While the analysis here considers the limit as $x \to \infty$, it is not necessarily the case that the properties observed in the $x \to \infty$ regime are well approximated by, e.g., the 99th percentile of $\widetilde{S}_n$ (i.e., even the 99th percentile may be too small to capture the behavior in the $x \to \infty$ regime). In practice, the principle of a single big jump for subexponential distributions holds only in a very narrow portion of the tail as n increases, making it difficult to observe this property from empirical data. Similarly, verifying the properties of Theorem 3.5 literally requires considering exceptionally large values of x, which may not be easily testable with empirical data. However, even when considering relatively moderate values of x or smaller jumps, some studies have demonstrated that large deviations of $\widetilde{S}_n$ are still driven by jumps, similar to the behavior predicted by Theorem 3.5 (see Bazhba et al. (2020)). Thus, the characteristic that large deviations of $\widetilde{S}_n$ are driven by jumps does not necessarily require interpreting $x \to \infty$ in the strictest sense.

4 Empirical results: Test of the two assumptions

This section provides empirical evidence supporting our explanation of Zipf's law. Section 4.1 provides summary statistics on the sizes and their growth rates. Section 4.2 shows that the random walk hypothesis provides a reasonable approximation of the empirical growth process, particularly in the tail region. Section 4.3 shows that the growth rate distribution is subexponential and exhibits a Weibull tail.

4.1 Data and summary statistics

In our empirical analysis, S_0 is defined as the logarithm of the firm's sales five years after its establishment in the case of firm sales, as the logarithm of income at age 25 in the case of individual income. For example, if a firm was established in the year 2000, and the log growth rate is denoted as g_n, then the size S_n in the year $2000 + n$ is expressed as follows:

$$S_n = \underbrace{\log(\text{annual sales revenue in 2000}) + g_{01} + g_{02} + g_{03} + g_{04} + g_{05}}_{S_0} + \underbrace{g_{06}}_{X_1} + \cdots + \underbrace{g_{n+5}}_{X_n}$$

Note that $\log(\text{annual sales revenue in 2000}) + g_{01} + g_{02} + g_{03} + g_{04} + g_{05}$ is equal to the logarithm of sales in 2005, and thus corresponds to S_0.

The reason we do not define S_0 as the size immediately following an agent's establishment is that the growth mechanism immediately after establishment may substantially differ from the growth mechanism in subsequent periods. For instance, in the case of firms, the establishment of a new firm may result from the restructuring of a firm group. When a new firm is created by transferring the business operations of multiple firms within the group, the transfer may be executed over several years following the firm's establishment. Treating the apparent growth caused by such pre-planned transfers as equivalent to growth in other periods (i.e., assuming the growth rates to be iid random variables) would not suit our analysis. Therefore, by including the first five years after establishment in S_0, we mitigate the effect of such issues.[19] Similarly, for individual income, it is not appropriate to treat the substantial income increase associated with transitioning from student status to employment as equivalent to income growth in other periods. Thus, we define S_0 as the logarithm of income at age 25, after the individual has completed their student years, where the random walk assumption is more applicable.

Firm sales

The data used for our analysis of firm sales is firm-level data compiled by Tokyo Shoko Research (TSR). As a credit rating agency, TSR conducts surveys based on the requests of its clients, and the data reflects

[19] Moreover, there is an additional data-related reason for defining S_0 in this manner for firm sales. The TSR data used in our analysis is based on firm surveys; therefore, newly established firms may not be included in the database if they were not surveyed by TSR immediately after their establishment. By utilizing the sales data from firms five years after their establishment as S_0, we can enhance the coverage and representativeness of the database.

Summary statistics of firm size

Data: Tokyo Shoko Research from 2010 to 2020

year	count	mean	sd	q1	median	q3
2010	1223312	11.407	1.730	10.309	11.300	12.393
2011	1244565	11.383	1.739	10.309	11.290	12.374
2012	1244671	11.392	1.740	10.309	11.290	12.384
2013	1239092	11.396	1.741	10.309	11.290	12.388
2014	1228270	11.428	1.754	10.309	11.339	12.429
2015	1241866	11.417	1.764	10.309	11.329	12.429
2016	1252042	11.410	1.766	10.309	11.327	12.414
2017	1255368	11.420	1.772	10.309	11.346	12.429
2018	1247023	11.435	1.777	10.309	11.352	12.429
2019	1223653	11.453	1.788	10.309	11.385	12.456
2020	1169440	11.431	1.816	10.309	11.361	12.461

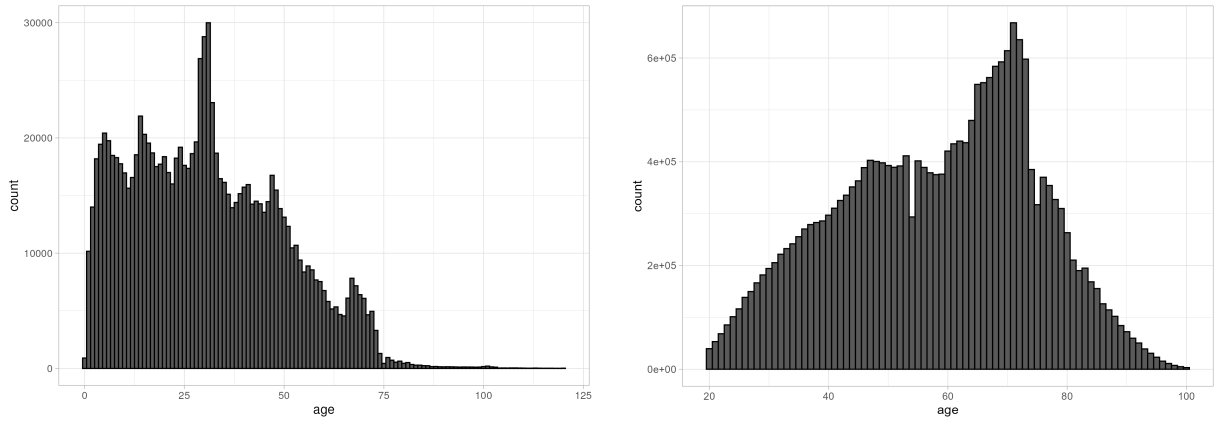
Table 1: Summary statistics of firms' sizes. The period is from 2010 to 2020. The unit is 1,000 yen. The summary statistics are calculated using the log of annual sales (i.e., $\log(\text{sale})$).

the results of these surveys. It includes both listed and unlisted firms, covering more than one million companies annually. Notably, nearly all large firms are included, and the information is frequently updated. Therefore, we can consider the tail of the firm size distribution—our primary focus in this analysis—to be comprehensively covered by the data.

Several conditions are imposed on the main sample. First, the sales of non-consolidated firms are used as the definition of firm size. Firms for which sales data are unavailable are excluded from the sample. Some firms in the TSR dataset report their financial results more than once within a single year, meaning their fiscal period is less than 12 months. In our analysis, we include only firms with a fiscal period of 12 months. Firms in the government and financial sectors are excluded from the sample. The sample period covers data from 2010 to 2020 (11 years), chosen to avoid the effects of both the global financial crisis and the COVID-19 pandemic.[20] As a result of these criteria, the sample size for the 2020 data is $1,169,440$. Summary statistics on firm size, including other years, are provided in **Table 1**.

Another important variable in our analysis is the growth rate of firm sales. The sample used for growth rate analysis adds two additional conditions to the sample used for the analysis of firm size. The first condition is that the initial firm size (i.e., the firm's sales in 2010) must be at least $10^{7.5}$ (i.e., about 30 million) yen. The reason for this is that if firm sales are too small, the fluctuations in the growth rate become large, deviating from the iid assumption (or Gibrat's law) underlying our theoretical analysis. The second condition pertains to firm age. As stated in the definition of S_0, the growth rate within the first five years of a

[20] In our analysis, *year* is based on the fiscal year-end date. For example, if a firm reports its sales for the period from April 1, 2010, to March 31, 2011, with a fiscal year-end of March 31, 2011, we treat these sales as the firm's sales for 2011.



(a) Firm sales data

(b) Individual income data

Figure 8: Age distributions in 2020. The samples used here are the same as those considered in **Table 1** and **Table 3**.

firm's establishment is included in S_0, so in the analysis of firm growth rates X_k, only firms that are at least five years old are considered.

The summary statistics of growth rates for these samples are provided in **Table 2**. The table presents summary statistics for one-year growth rates across different years, as well as statistics for growth rates over longer periods, with 2010 as the initial year. As evident from the table, the fluctuations in one-year growth rates remain remarkably stable throughout the sample period. Another notable point is that the fluctuations in growth rates become larger as the period lengthens. While the latter point may seem obvious, it will be analyzed in detail in the following section.

Individual income

For individual income data, we use tax return data provided by the National Tax College. In Japan, individuals file a tax return to report their income and calculate the taxes owed for the year (January 1 to December 31). This process determines the amount of income tax and local taxes to be paid. While not mandatory for everyone, filing a tax return is required, for example, if one's salary income exceeds 20 million yen. Since our analysis focuses on high-income individuals in the tail of the distribution, tax return data is well-suited to our study. Additionally, Japan has a system called the medical expense deduction, where individuals can deduct medical expenses exceeding a certain amount from their taxable income, thereby reducing their tax by filing a tax return. For these reasons, more than 20 million individuals file tax returns each year. These tax returns are panelized, with each individual assigned a unique ID, and include information on the individual's age, providing all the data necessary for our theoretical analysis.

One of the important characteristics of the tax return data for our analysis is that it provides not only the total amount of an individual's income but also details on the types of income. Since our analysis assumes that income shocks are persistent (i.e., the assumption of iid random variables), we define total

Summary statistics of firm growth rates

Data: Tokyo Shoko Research from 2010 to 2020

year	count	mean	sd	q1	median	q3
g_11_10	749436	−0.022	0.292	−0.087	0.000	0.067
g_12_10	709218	−0.021	0.372	−0.123	0.000	0.114
g_13_10	682375	−0.021	0.422	−0.154	0.000	0.147
g_14_10	660473	0.009	0.468	−0.154	0.001	0.213
g_15_10	649458	−0.002	0.514	−0.185	0.000	0.225
g_16_10	636295	−0.011	0.549	−0.219	0.000	0.241
g_17_10	622511	−0.012	0.578	−0.234	0.000	0.260
g_18_10	605523	−0.003	0.607	−0.246	0.000	0.289
g_19_10	586576	0.004	0.635	−0.255	0.008	0.318
g_20_10	561839	−0.038	0.676	−0.321	−0.005	0.305
g_12_11	750417	−0.008	0.286	−0.070	0.000	0.074
g_13_12	756783	−0.009	0.275	−0.069	0.000	0.067
g_14_13	755528	0.019	0.272	−0.044	0.000	0.097
g_15_14	761418	−0.016	0.272	−0.076	0.000	0.061
g_16_15	769384	−0.016	0.269	−0.074	0.000	0.059
g_17_16	777785	−0.010	0.267	−0.062	0.000	0.061
g_18_17	776427	−0.001	0.266	−0.051	0.000	0.071
g_19_18	770767	−0.001	0.261	−0.051	0.000	0.065
g_20_19	755077	−0.051	0.284	−0.120	−0.005	0.032

Table 2: Summary statistics of firm growth rates. Here, I present the summary statistics for the one-year firm growth rates for different years, as well as the growth rate statistics for longer periods with 2010 as the base year. For example, g_20_10 represents the summary statistics for the growth rate from 2010 to 2020.

Summary statistics of individual income

Data: Tax return data from National Tax College from 2014 to 2020

year	count	mean	sd	q1	median	q3
2014	22275479	15.190	1.015	14.659	15.155	15.764
2015	22442153	15.198	1.017	14.666	15.165	15.776
2016	22611792	15.206	1.015	14.674	15.176	15.784
2017	22856952	15.213	1.013	14.680	15.186	15.790
2018	22993167	15.219	1.011	14.684	15.193	15.797
2019	23117452	15.217	1.013	14.678	15.193	15.799
2020	22567730	15.224	1.016	14.691	15.208	15.806

Table 3: Summary statistics of individual income. The period covers from 2014 to 2020. The summary statistics are calculated using the log of individual income (i.e., $\log(\text{income})$).

income excluding temporary income sources. Specifically, we define income as the sum of business income, agricultural income, real estate income, interest income, dividend income, salary income, public pension income, and miscellaneous business income. Additionally, individuals with zero income are excluded from the sample. The sample size, for example, is $22,567,730$ in 2020. Summary statistics for individual income, including other years, are provided in **Table 1**.

The sample used for analyzing individual income growth rates imposes two additional conditions, similar to the case of firm sales, on the sample of income size. The first condition is that the individual's initial income must be at least 4 million yen. The second condition pertains to individual's age. Only individuals who were between 25 and 60 years old as of 2014 are included in the sample.

The summary statistics of growth rates for these samples are provided in **Table 2**. This table shows the summary statistics for one-year growth rates across different years, as well as statistics for growth rates over longer periods, with 2014 as the initial year. Similar to the case of firm sales growth rates, the fluctuations in one-year growth rates remain remarkably stable during the sample period from 2014 to 2020. Furthermore, as the length of the period increases, the fluctuations in growth rates become larger. The similarity between the distribution of firm sales growth rates and individual income growth rates will be analyzed in detail in the following section.

Summary statistics of the growth rates of individual income

Data: Tax return data from National Tax College from 2014 to 2020

year	count	mean	sd	q1	median	q3
g_15_14	5074654	−0.018	0.328	−0.045	0.011	0.068
g_16_14	4829912	−0.029	0.426	−0.080	0.019	0.106
g_17_14	4678281	−0.039	0.493	−0.114	0.024	0.137
g_18_14	4555980	−0.047	0.544	−0.153	0.029	0.165
g_19_14	4435731	−0.056	0.590	−0.197	0.031	0.189
g_20_14	4302956	−0.082	0.634	−0.265	0.018	0.202
g_16_15	5220246	−0.022	0.332	−0.050	0.009	0.065
g_17_16	5381767	−0.019	0.325	−0.045	0.007	0.065
g_18_17	5517381	−0.017	0.322	−0.043	0.010	0.066
g_19_18	5583172	−0.018	0.326	−0.044	0.008	0.066
g_20_19	5579968	−0.038	0.341	−0.074	0.000	0.059

Table 4: Summary statistics of the growth rates of individual's income. Here, I present the summary statistics for the one-year firm growth rates for different years, as well as the growth rate statistics for longer periods with 2014 as the initial year. For example, g_20_14 represents the summary statistics for the growth rate from 2014 to 2020.

4.2 Random walk assumption

In this section, we examine the assumption in our explanation that growth rates are independent random variables. A widely used measure for analyzing dependencies between variables is Pearson's correlation coefficient; however, we do not use it here (see the Appendix). Instead, we use Spearman's ρ_S as an alternative measure. Spearman's ρ_S ranges from -1 to 1, and it equals 0 if the two variables are independent.

While Spearman's ρ_S provides insights into dependencies between growth rates, this coefficient is strongly influenced by dependencies in regions with more abundant samples, namely the central region. It may not adequately capture dependencies in the tail region, specifically in cases of extreme values. To address this concern, we use the tail dependence coefficient. This coefficient measures the likelihood that one variable takes an extreme value given that another variable also takes an extreme value and is defined as follows:

$$\lambda_U := \lim_{q \to 1} \mathbb{P}(X_2 > F_2^{-1}(q) \mid X_1 > F_1^{-1}(q))$$

Intuitively, this coefficient can be interpreted as the probability of a large deviation in the second period, given a large deviation in the first period. When λ_U takes a positive value (i.e., the conditional probability above converges to a positive value), the two variables are said to exhibit tail dependence. If $\lambda_U = 0$, the variables are said to exhibit tail independence. $\lambda_U = 0$ suggests that large deviations do not occur consecutively.

Furthermore, as another measure of dependence in the tail region, we introduce the conditional tail expectation:

$$\mathbb{E}[X_2 | X_1 > t]$$

We examine the behavior of this function as $t \to \infty$. For example, if X_1 and X_2 represent growth rates in consecutive periods, then the conditional tail expectation can be viewed as the expected growth rate in the second period for firms that achieved high growth in the first period. If the two random variables are completely independent, this function would remain constant regardless of t. In contrast, if there is tail dependence (i.e., $\lambda_U > 0$), this function becomes a linear increasing function of t (i.e., $\mathbb{E}[X_2 | X_1 > t] \sim O(t)$ as $t \to \infty$) (see Section 2.20 in Joe (2014)).

Finally, to examine the extent to which dependence in the tail region leads to consecutive extreme values, we consider the frequency of occurrences where $X_k > F_k^{-1}(p)$ within the sample period for each agent, where p is a value close to 1. If the random walk assumption holds, the frequency of occurrences should follow a binomial distribution:

$$\mathbb{P}(\text{\# of occurrences} = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

We also calculate the frequency of the longest consecutive occurrences of $X_k > F^{-1}(p)$ (i.e., longest success run) within a sample period. These results are compared with those obtained from computer simulations under the random walk assumption.

Matrix of correlation coefficients

Spearman's rank correlation coefficient

term	g_20_19	g_19_18	g_18_17	g_17_16	g_16_15	g_15_14	g_14_13	g_13_12	g_12_11	g_11_10
g_20_19	NA	−0.056	0.052	0.054	0.061	0.044	0.050	0.046	0.032	0.021
g_19_18	−0.056	NA	−0.073	0.059	0.055	0.055	0.059	0.046	0.049	0.043
g_18_17	0.052	−0.073	NA	−0.078	0.046	0.060	0.071	0.038	0.042	0.058
g_17_16	0.054	0.059	−0.078	NA	−0.090	0.046	0.063	0.050	0.039	0.040
g_16_15	0.061	0.055	0.046	−0.090	NA	−0.076	0.046	0.046	0.043	0.036
g_15_14	0.044	0.055	0.060	0.046	−0.076	NA	−0.080	0.046	0.050	0.049
g_14_13	0.050	0.059	0.071	0.063	0.046	−0.080	NA	−0.066	0.060	0.066
g_13_12	0.046	0.046	0.038	0.050	0.046	0.046	−0.066	NA	−0.076	0.035
g_12_11	0.032	0.049	0.042	0.039	0.043	0.050	0.060	−0.076	NA	−0.075
g_11_10	0.021	0.043	0.058	0.040	0.036	0.049	0.066	0.035	−0.075	NA

Table 5: Matrix of Spearman's ρ_S. Samples are the same as in **Table 2**.

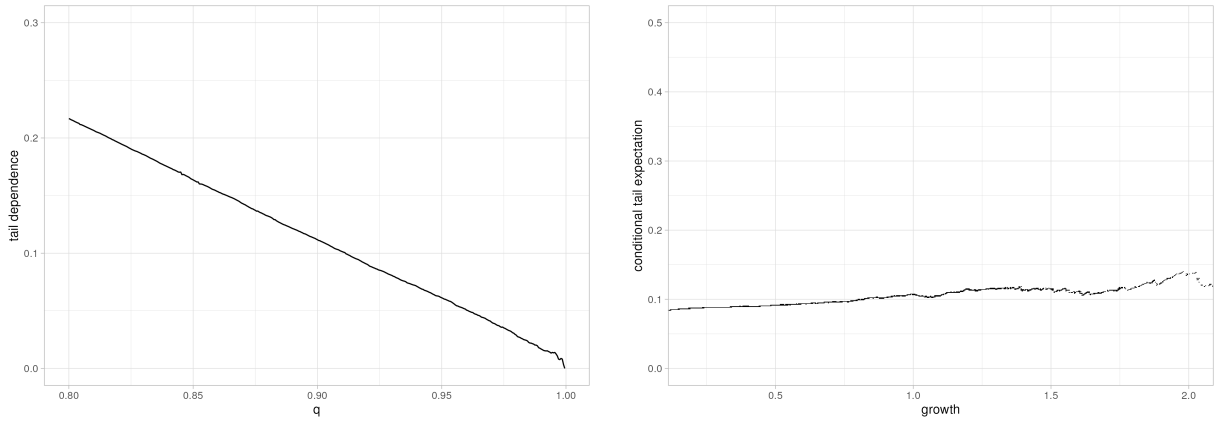
Firm sales

We apply the above method to the growth rates of firm sales. The correlation matrix of growth rates for different years, calculated using Spearman's ρ_S, is shown in **Table 5**. As expected, the coefficients approach 0 as the time interval between the two growth rates increases. Additionally, as indicated in the table, even for consecutive periods, the absolute value of the coefficients is below 0.1 and close to 0.[21] In the following sections, we focus on consecutive periods and analyze them in greater detail.

Figure 9(a) shows the tail dependence measure for growth rates in 2010 and 2011 (more precisely, the conditional probability within the definition of λ_U) and how it changes as $q \to 1$. As $q \to 1$, the tail dependence measure decreases and approaches 0. This suggests that growth rates in consecutive periods become closer to tail independence as higher growth rates are considered. The results for the conditional tail expectation are shown in **Figure 9**(b). Here, considering that growth rates can take both positive and negative values, we calculate the conditional tail expectations for $X_k \mathbf{1}_{X_k>0}$. As t increases, the conditional tail expectation does not behave as an increasing function; rather, it remains nearly constant with respect to t. In other words, achieving high growth in the preceding period does not increase the probability of continued high growth in subsequent periods. This observation aligns with the results of the tail dependence measure λ_U, further indicating that the dependence between growth rates in the positive tail region (i.e., high-high relationships) is weak. Thus, this provides evidence supporting the validity of using the random walk assumption in our analysis.

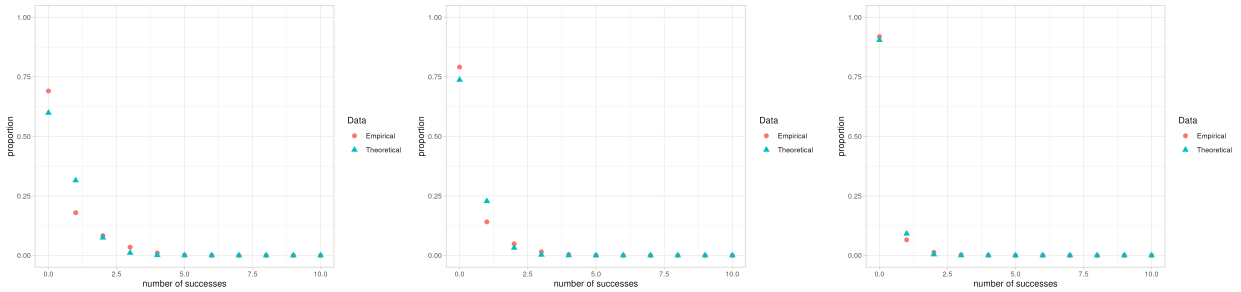
The results for the frequency of high growth $X_k > F_k^{-1}(p)$ are shown in **Figure 10**. Here, we consider

[21] Spearman's $\rho_S = 0$ does not necessarily imply the independence of the two random variables across all regions of the distribution. As discussed in the Appendix, the semi-correlation coefficient, which measures dependency in the positive region (i.e., $X_1 > 0, X_2 > 0$), for growth rates in 2019 and 2020 is $\rho_N^+ = 0.278$ and $\rho_N^- = 0.274$, respectively. This suggests that when the range is restricted to the positive region, the dependency is stronger than when considering the entire distribution.



(a) Tail dependence measure **(b)** Conditional tail expectation

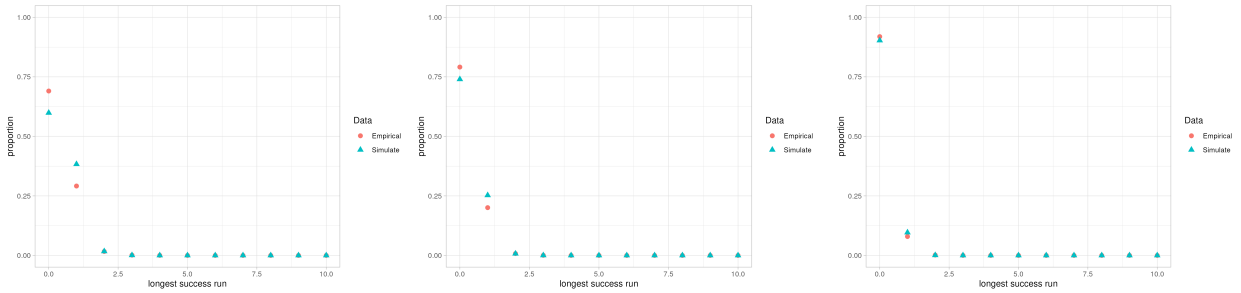
Figure 9: Tail dependence measure and conditional tail expectation. In Panel(a), the conditional tail expectation of $X_k \mathbf{1}_{X_k>0}$ (on the y-axis) are plotted as a function of q (on the x-axis).



(a) $p = 0.95$ **(b)** $p = 0.97$ **(c)** $p = 0.99$

Figure 10: Histogram of the occurrence frequency of growth rates greater than $F^{-1}(p)$ and the binomial distribution.

high growth thresholds at $p = 0.95, 0.97, 0.99$. As the figure shows, the histogram of the frequency of $X_k > F_k^{-1}(p)$ calculated from the empirical data is very close to the theoretical values of a binomial distribution. Therefore, the dependence does not have an significant impact on the tail region for $p = 0.95, 0.97, 0.99$. In other words, there is no evidence that high growth $X_k > F_k^{-1}(p)$ is disproportionally concentrated in specific firms. A similar result can be observed for the longest success run. **Figure 11** compares the histogram of the frequency of the longest success runs with that of an iid simulation. As the figure shows, when considering high growth thresholds such as $p = 0.95, 0.97, 0.99$, the two histograms are very close. This indicates that in the tail region, high growth is not driven by dependence between growth rates, and the growth rates can be well-approximated as iid. These results suggest that, when focusing on the tail region, which is of primary interest, the random walk assumption is empirically reasonable.



(a) $p = 0.95$ **(b)** $p = 0.97$ **(c)** $p = 0.99$

Figure 11: Histograms of the frequency of the longest success runs and for the i.i.d. case calculated from simulations.

Matrix of correlation coefficients

Spearman's rank correlation coefficient

term	g_20_19	g_19_18	g_18_17	g_17_16	g_16_15	g_15_14
g_20_19	NA	0.013	−0.006	0.004	0.020	0.012
g_19_18	0.013	NA	0.021	0.008	0.030	0.033
g_18_17	−0.006	0.021	NA	0.036	0.014	0.028
g_17_16	0.004	0.008	0.036	NA	0.024	0.012
g_16_15	0.020	0.030	0.014	0.024	NA	0.035
g_15_14	0.012	0.033	0.028	0.012	0.035	NA

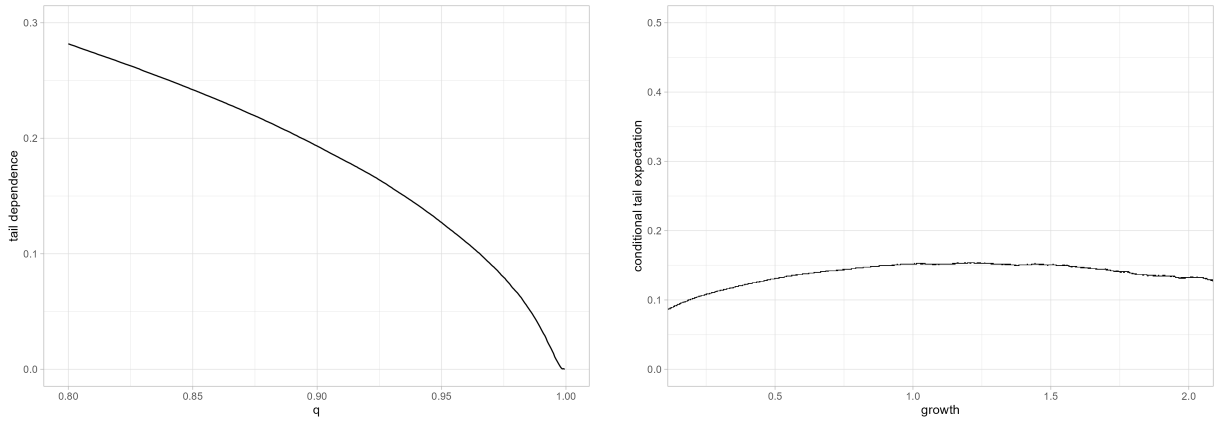
Table 6: Matrix of Spearman's ρ. Samples are the same as in **Table 2**.

Individual income

We describe the results for individual income growth rates. The correlation matrix of growth rates for different years, calculated using Spearman's ρ_S, is shown in **Table 6**. As the time interval between two growth rates increases, the coefficients approach 0. Even for consecutive periods, the coefficients are below 0.1 and close to 0. Similar to the case of firm sales, when considering dependence across the entire distribution, the dependence between growth rates is weak.

Figure 12(a) illustrates how the tail dependence measure for growth rates in 2015 and 2016 changes as $q \to 1$. As $q \to 1$, the tail dependence measure decreases and approaches 0. The results for the conditional tail expectation of the non-negative variable $X_k \mathbf{1}_{X_k>0}$ are shown in **Figure 12**(b). As t increases, the conditional tail expectation is not an increasing function but rather remains nearly constant with respect to t. These results indicate that the dependency between growth rates in the positive tail region (i.e., high-high relationships) is weak. Since our focus is on dependency in the tail region, these findings support the validity of using the random walk assumption in our theoretical explanation.

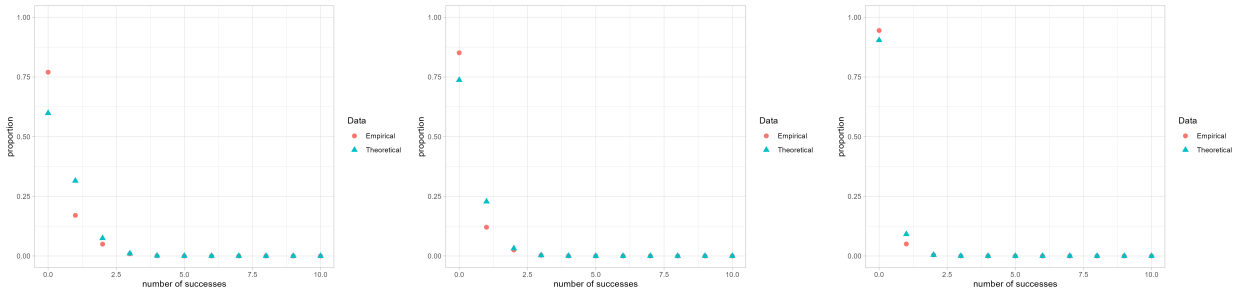
Finally, to examine dependence in the positive tail region (i.e., cases where high growth occurs consecutively), we calculate the frequency of occurrences where $X_k > F_k^{-1}(p)$ and the frequency of the longest



(a) Tail dependence measure

(b) Conditional tail expectation

Figure 12: Tail dependence measure and conditional tail expectation. In Panel(a), the conditional tail expectation of $X_k \mathbf{1} X_k > 0$ (on the y-axis) are plotted as a function of q (on the x-axis).



(a) $p = 0.95$

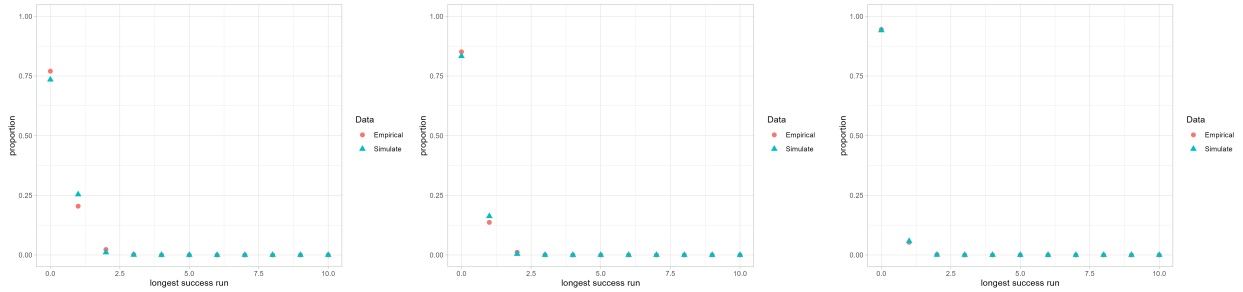
(b) $p = 0.97$

(c) $p = 0.99$

Figure 13: Histogram of the occurrence frequency of growth rates greater than $F^{-1}(p)$ and the binomial distribution.

success run for each individual during the sample period. The results are presented in **Figure 13** and **Figure 14**. When considering high growth at levels of $p = 0.95, 0.97, 0.99$, the histogram of high growth occurrences is close to the theoretical prediction of a binomial distribution under the assumption of iid growth rates. The histogram of the longest success run frequency is also very similar to the histogram obtained from simulations under the i.i.d. assumption. This suggests that in the tail regions, such as $p = 0.95, 0.97, 0.99$, high growth is not concentrated within certain groups, and the i.i.d. case (i.e., our random walk assumption) provides a reasonable approximation.

As demonstrated, the dependency structures of firm sales growth rates and individual income growth rates are quite similar to each other. Specifically, the dependence in the tail regions can be reasonably approximated by independence. Given that our theoretical explanation in Section 3 relies solely on such statistical properties, it is natural that Zipf's law applies similarly to both phenomena when the underlying statistical characteristics are alike. In the next section, we will test the other key assumption of our model, namely that the growth rate distribution follows a Weibull tail.



(a) $p = 0.95$ **(b)** $p = 0.97$ **(c)** $p = 0.99$

Figure 14: Histograms of the frequency of the longest success runs and for the i.i.d. case calculated from simulations.

4.3 Growth rate distribution

In this section, we estimate whether the growth rates follow subexponential distributions, specifically Weibull-tail distributions. We analyze the growth rate distribution using the following methods: density function estimation, statistical tests for exponentiality, the mean excess function, and tail shape estimation as proposed by Gardes et al. (2011) and El Methni et al. (2012). These methods are explained below.

The mean excess function for the threshold u (denoted as $e(u)$) is defined as the conditional expected value of the overshoot $X_k - u$, given that $X_k > u$ (see, e.g., Embrechts et al. (1997)):

$$e(u) := \mathbb{E}[X - u \mid X > u] \quad \text{for } u > 0.$$

The reason for using the mean excess function is that its shape reflects the heaviness of the tail of the distribution of X. In particular, our analysis focuses on the following three cases: When the distribution has an exponential tail, $e(u)$ remains constant regardless of u. When the distribution has a Pareto tail, $e(u)$ is a linearly increasing function of u. When the distribution has a Weibull tail, it lies between the two cases above. Specifically, as u increases, it is known that $e(u)$ can be written as follows:

$$e(u) = \frac{u^{1-\alpha}}{c\alpha}(1 + o(1))$$

as $u \to \infty$ (cf. Beirlant et al. (1995)). By estimating the empirical mean excess function using growth rates, we can characterize the heaviness of the distribution's tail.

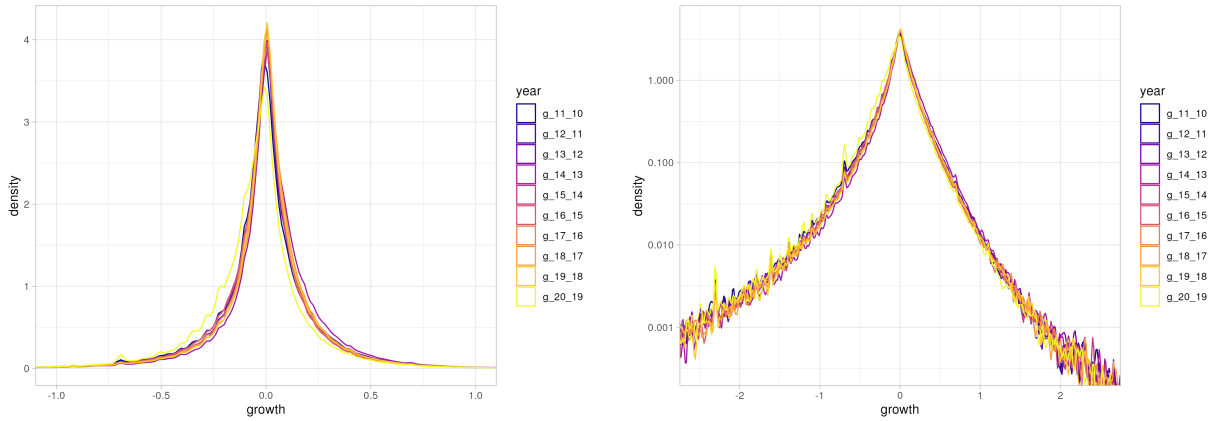
To statistically show that the tail of the distribution is subexponential, we set the null hypothesis that the tail follows an exponential function and test whether this hypothesis can be statistically rejected. While various statistical tests for exponentiality have been proposed (see Ascher (1990), Henze and Meintanis (2005) for surveys), Ascher (1990) suggests that the Cox-Oakes test, proposed by Cox and Oakes (1984), possesses the strongest statistical power. We apply the Cox-Oakes test to the following two subsamples of growth rates: samples with $X_k \geq 0.1$ and samples with $X_k \geq 0.2$.

As discussed in Section 3.2, within the family of subexponential distributions, there are two groups: those with Pareto tails and Weibull tails. In particular, our analysis needs to confirm that the growth rate distribution can be approximated by a Weibull tail. As a graphical method, we use the following linear relationship, which holds if the distribution has a Weibull tail. For $0 < u < v < 1$,

$$\log(-\log u) - \log(-\log v) \approx \alpha(\log x_u - \log x_v)$$

Here, x_u and x_v are the u-quantile and v-quantile values of the growth rates, respectively. Specifically, letting N be the sample size and X_{N-i+1} the i-th largest growth rate value, if the growth rate distribution follows a Weibull tail, the points $(\log\log(N/i), \log X_{N-i+1})$ should plot as a straight line. Using this property, we verify whether the empirical distribution of growth rates can be approximated by a Weibull tail.

To statistically determine which of these two groups the empirical distribution of growth rates is closer to, we use the methods proposed by Gardes et al. (2011) and El Methni et al. (2012). Define the following



(a) Normal scale

(b) Log scale

Figure 15: The distributions of annual growth rates. Here, we provide the density function estimation for the distribution of annual growth rates during the sample period (2010 to 2020). In panel (a), the y-axis is on a normal scale, while in panel (b), the y-axis is on a log scale.

family of distributions that encompass both Pareto-type and Weibull-type tails:

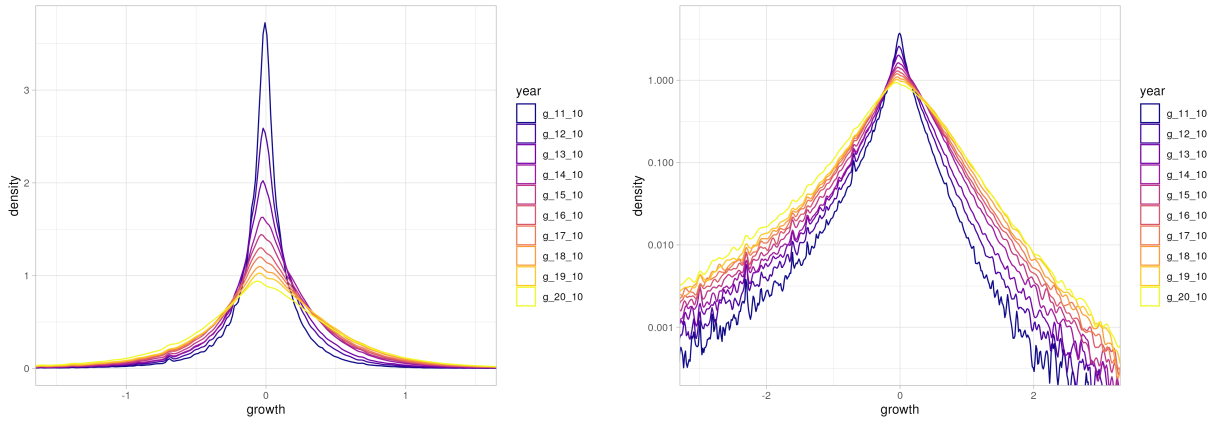
$$\overline{F}(x) = \exp(-K_\tau^{\leftarrow}(\log H(x))) \quad \text{for } x \geq x_* > 0, \text{ with } \tau \in [0,1] \tag{4}$$

Here, H is a function such that its inverse $H^{\leftarrow}$ is given by $H^{\leftarrow}(t) = t^\theta L(t),\ \theta > 0$ (where $L(t)$ is a slowly varying function) and $K_\tau(x) = \int_1^x u^{\tau-1} du$. A larger value of τ means a heavier tail of the distribution, with $\tau = 0$ corresponding to a Weibull-type tail and $\tau = 1$ corresponding to a Pareto-type tail. The parameter $\alpha = \theta^{-1}$ is the shape parameter for each distribution tail corresponding to τ. Therefore, by estimating the value of τ, we can assess how well the Weibull tail assumption approximates the empirical data.

Firm sales

Here, we apply the above methods to the growth rates of firm sales. The results of the density estimation are presented in **Figure 15**. As shown in **Figure 15**(a), consistent with previous studies, the distribution deviates from a normal distribution, exhibiting a sharp peak in the center and heavier tails. Additionally, **Figure 15**(b) shows that the tail of the distribution on a log scale is curved rather than linear, indicating that the tail is heavier than that of an exponential function. This provides evidence that the distribution is subexponential. Next, we consider the distribution of longer-term growth rates. **Figure 16** shows the distribution of growth rates over n periods ($n = 1, 2, \ldots, 10$). **Figure 16**(a) illustrates that as n increases, the density around 0 approaches a bell-shaped curve. As shown in **Figure 16**(b), the density in the right tail region appears nearly linear on a log scale. Additionally, as n increases, the linear portion of the tail in the density function appears to shift upward in parallel. These characteristics are consistent with the shape described in Section 3.2.

The mean excess function for growth rates is shown in **Figure 17**. In all years, the estimated $e(u)$ is



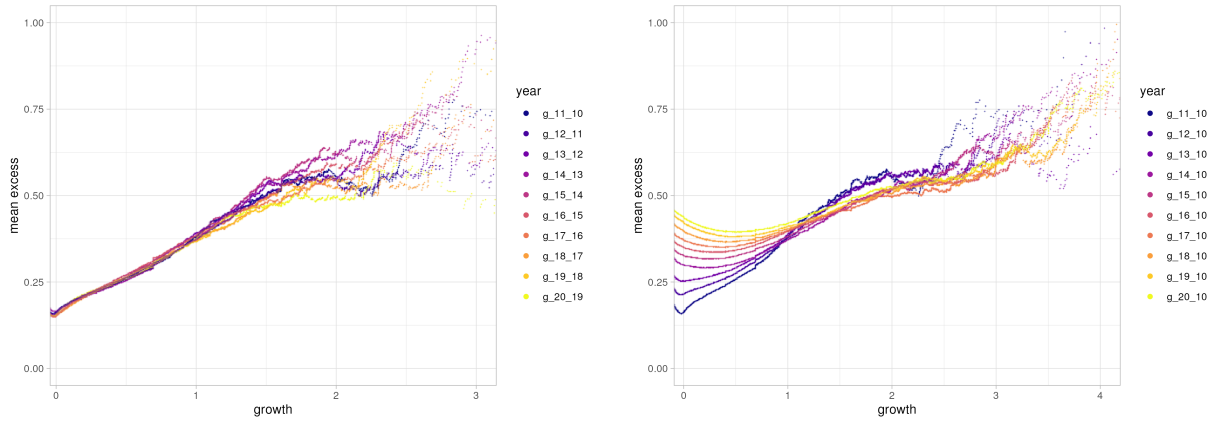
(a) Normal scale

(b) Log scale

Figure 16: The distributions of long-term growth rates. Here, we provide the density function estimation for the distribution of n-year growth rates during the sample period from 2010 to 2020 ($n = 1, 2, ..., 10$). The initial period for the k-year growth rates is 2010 for all n. In panel (a), the y-axis is on a normal scale, while in panel (b), the y-axis is on a log scale.

an increasing function of u, indicating that the distribution has heavier tails than an exponential distribution. Moreover, statistical testing using the Cox-Oakes test shows that exponentiality is rejected in all cases with p-values below 1%. In particular, **Figure 17**(b) shows that the shape of $e(u)$, particularly in the tail region, appears to be largely independent of n. This aligns with the property of subexponential distributions, where the tail probabilities of n-year growth rates match those of one-year growth rates except, apart from a multiplier factor. This provides further evidence that the growth rate distribution is subexponential. Furthermore, in all the figures, $e(u)$ is an increasing function of u, but its slope decreases as u becomes larger. This suggests that the growth rate distribution is closer to a Weibull tail than a Pareto tail. We examine in more detail below whether the distribution is closer to a Pareto tail or a Weibull tail.

Finally, we examine whether a Weibull tail can approximate the tail of the growth rate distribution. In **Figure 18**(a), we plot $(\log\log(N/i), \log X_{N-i+1})$, considering the top 5% of the growth rate samples from one-year growth rates (i.e., $i = 1, \ldots, 0.05N$). As shown in the figure, the plot of $(\log\log(N/i), \log X_{N-i+1})$ aligns closely with a straight line, indicating that the tail of the growth rate distribution is well approximated by a Weibull tail. The estimation results for τ in Eq.(4) are shown in **Figure 18**(b). Since the estimation of τ uses only the top k_n samples, the x-axis represents the number of samples k_n, while the y-axis shows the estimated values of τ for each k_n. **Figure 18**(b) shows that the estimates are close to zero, further supporting that a Weibull tail provides a better approximation within the class of subexponential distributions. Furthermore, using the estimated τ, we calculate the other parameter α, and based on Eq.(4), compare the estimated tail probabilities with the counter cumulative distribution function (i.e., $1 - F_n(x)$, where F_n is the empirical distribution). This comparison is presented in **Figure 18**(c). The estimated tail probabilities closely approximate the tail probabilities of the growth rates.



(a) One-year growth rates

(b) k-year growth rates

Figure 17: Mean excess function over threshold u. Here, we present the mean excess function of the one-year growth rate for the sample period from 2010 to 2020 in panel (a), and the mean excess function of the n-year growth rate ($n = 1, 2, ..., 10$) in panel (b). The initial period for the n-year growth rates is set to 2010 for all n. The standardized Cox-Oakes test statistic is -54.4 for $X_k > 0.1$ and -38.9 for $X_k > 0.2$, both of which allow the null hypothesis of exponentiality to be rejected at a p-value below 0.01.

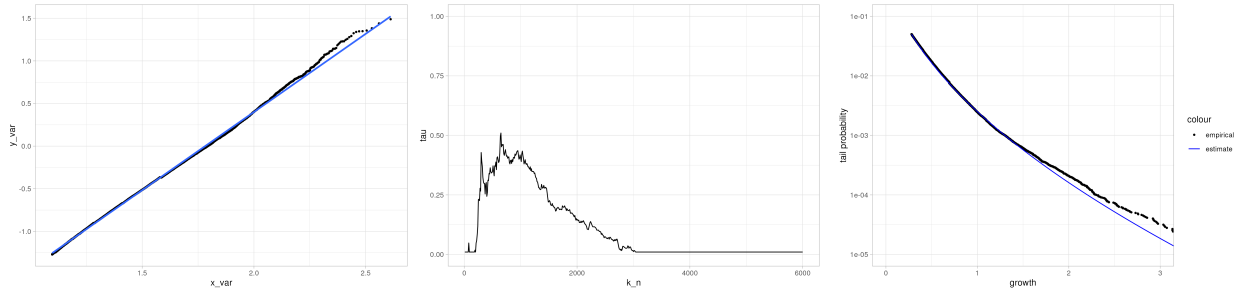
These results, consistent with the findings of the mean excess function, support our assumption that the growth rate distribution of firm sales follows a Weibull tail.

Individual income

Here, we examine the tail of the growth rate distribution for individual incomes. **Figure 19** and **Figure 20** present the density estimates for one-year growth rates and n-year growth rates ($n = 1, 2, \ldots, 6$), respectively. As in the case of firm sales, it is evident that the growth rate distribution deviates from a Gaussian distribution and has heavier tails than an exponential function. Moreover, as n increases, the density around 0 approaches the shape of a Gaussian distribution, while its tail retains a nearly linear shape on a logarithmic scale. These characteristics indicate that the growth rate distribution is subexponential.

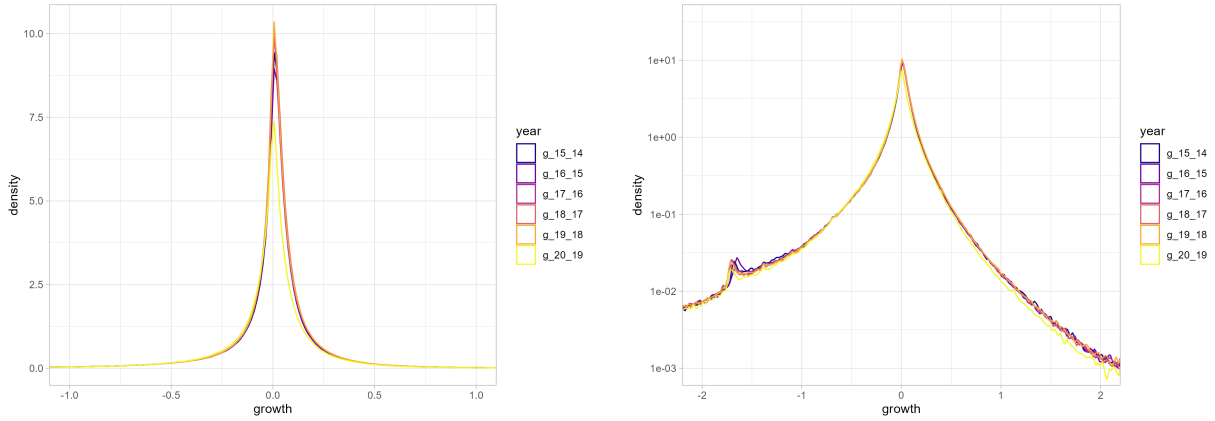
The estimation results for the mean excess function of growth rates are shown in **Figure 21**. Similar to the case of firm sales, $e(u)$ is an increasing function of u in all years, indicating that the distribution has heavier tails than an exponential function. Furthermore, for the n-year growth rates ($n = 1, 2, \ldots, 6$) shown in **Figure 21**(b), the shape of $e(u)$ in the tail region does not depend on n. Indeed, statistical tests for exponentiality reject the null hypothesis with p-values below 1%. These results indicate that the tail of the growth rate distribution is not exponential but heavier, confirming that the distribution is subexponential. Moreover, the slope of $e(u)$ decreases as u becomes larger, suggesting that the growth rate distribution is closer to a Weibull tail than a Pareto tail.

Figure 22 presents the plot of $(\log\log(N/i), \log X_{N-i+1})$ for the growth rates of individual incomes. The plot aligns closely with a straight line, indicating that the tail of the growth rate distribution follows a



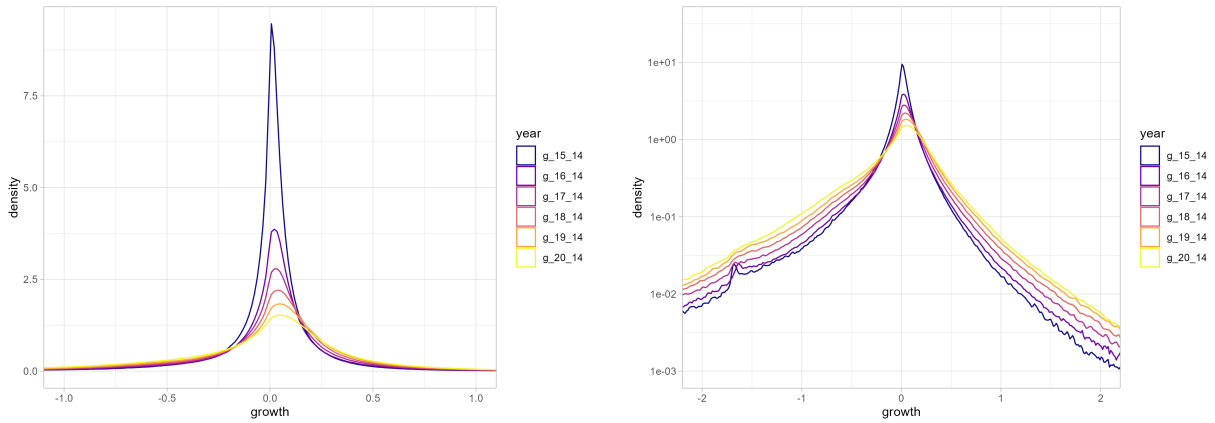
(a) Linear relation **(b)** Estimate of τ **(c)** Estimate of the tail probability

Figure 18: Parameter estimates of the tail probability of growth rates. Here, we use the one-year growth rate for 2020. We consider the top 5% of growth rates in our samples. In Panel (a), $(\log\log(n/i), \log X_{n-i+1})$ is plotted. In Panel (c), we calculate Eq.(4) based on the estimated τ and α. For comparison, we also plot the counter cumulative distribution function of growth rates.



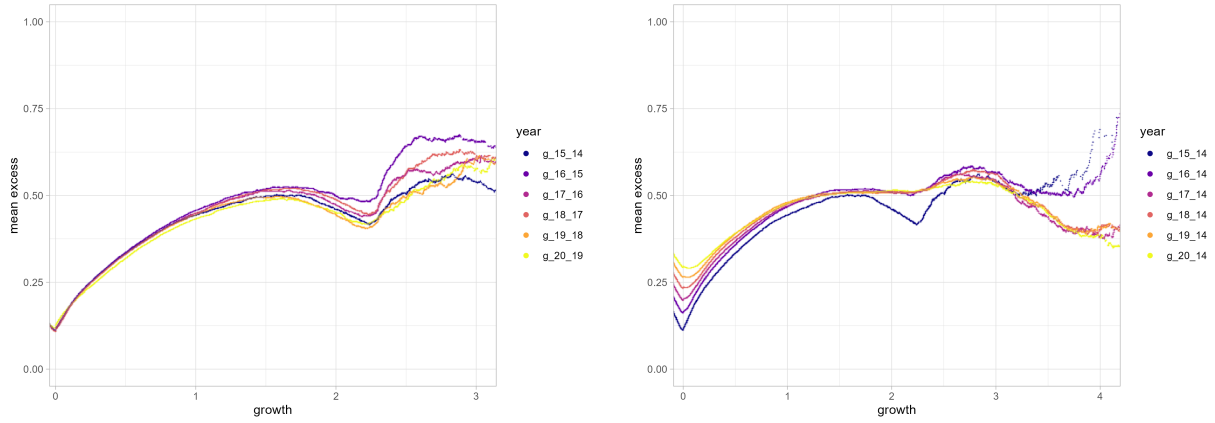
(a) Normal scale **(b)** Log scale

Figure 19: The distributions of annual growth rates. Here, we provide the density function estimation for the distribution of annual growth rates during the sample period (2014 to 2020). In panel (a), the y-axis is on a normal scale, while in panel (b), the y-axis is on a log scale.



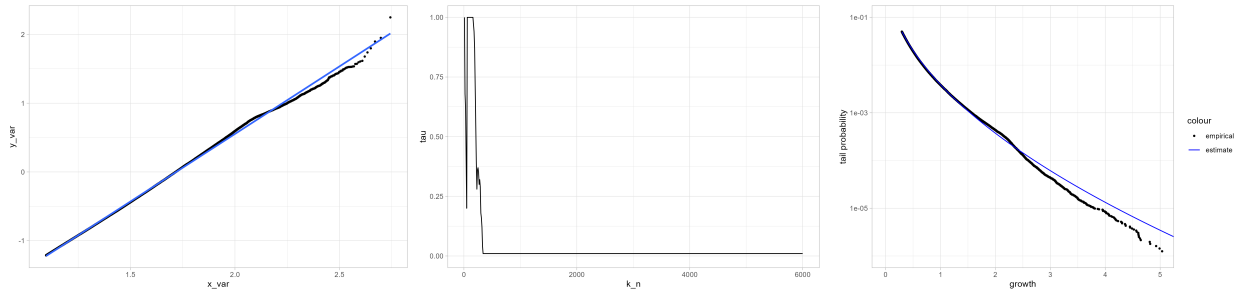
(a) Normal scale **(b)** Log scale

Figure 20: The distributions of long-term growth rates. Here, we provide the density function estimation for the distribution of n-year growth rates during the sample period from 2014 to 2020 ($n = 1, 2, ..., 6$). The initial period for the n-year growth rates is 2014 for all n. In Panel (a), the y-axis is on a normal scale, while in panel (b), the y-axis is on a log scale.



(a) One-year growth rates **(b)** k-year growth rates

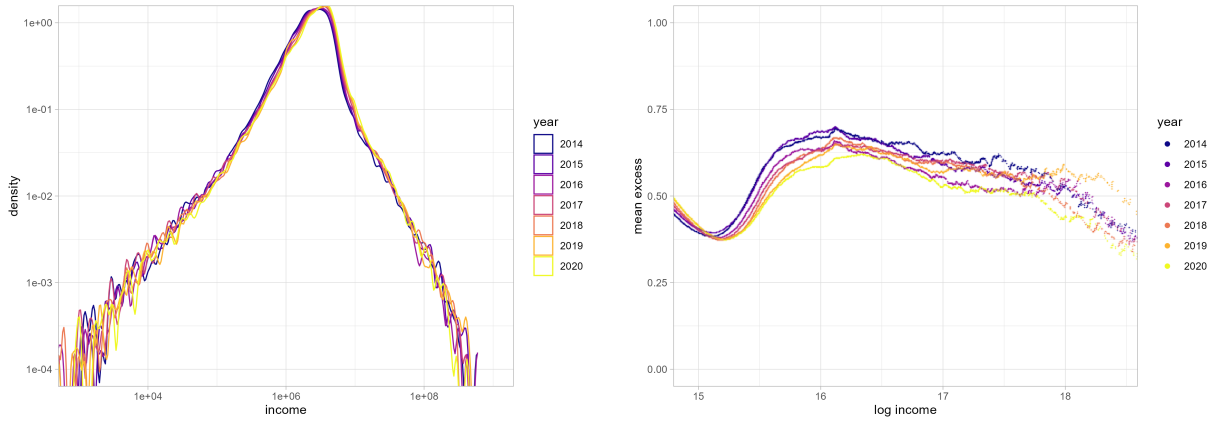
Figure 21: Mean excess function over threshold u. Here, we present the mean excess function of the one-year growth rate for the sample period from 2014 to 2020 in Panel (a), and the mean excess function of the n-year growth rate ($k = 1, 2, ..., 6$) in Panel (b). The initial period for the n-year growth rates is set to 2014 for all n. The standardized Cox-Oakes test statistic is -344.8 for $X_k > 0.1$ and -182.6 for $X_k > 0.2$, both of which allow the null hypothesis of exponentiality to be rejected at a p-value below 0.01.



(a) Linear relation **(b)** Estimate of τ **(c)** Estimate of the tail probability

Figure 22: Parameter estimates of the tail probability of growth rates. Here, we use the one-year growth rate for 2020. We consider the top 5% of growth rates in our samples. In Panel (a), $(\log\log(n/i), \log X_{n-i+1})$ is plotted. In Panel (c), we calculate Eq.(4) based on the estimated τ and α. For comparison, we also plot the counter cumulative distribution function of growth rates.

Weibull tail. Furthermore, the estimation results for τ in Eq.(4), shown in **Figure 18**(b), are close to $\tau = 0$. As shown in **Figure 18**(c), the tail probabilities calculated using the estimated values of τ and α in Eq.(4) closely approximate the counter cumulative distribution function. These results, consistent with the case of firm sales, indicate that the growth rate distribution follows a Weibull tail.



(a) Density estimates

(b) Mean excess functions

Figure 23: The distribution of the initial size S_0. Here, we consider the logarithm of the income S_0 of individuals that are 25 years old during the sample period (2014 to 2020). Panel (a) provides a density estimate of these samples, with the y-axis presented on a log scale. Panel (b) presents the mean excess function of S_0.

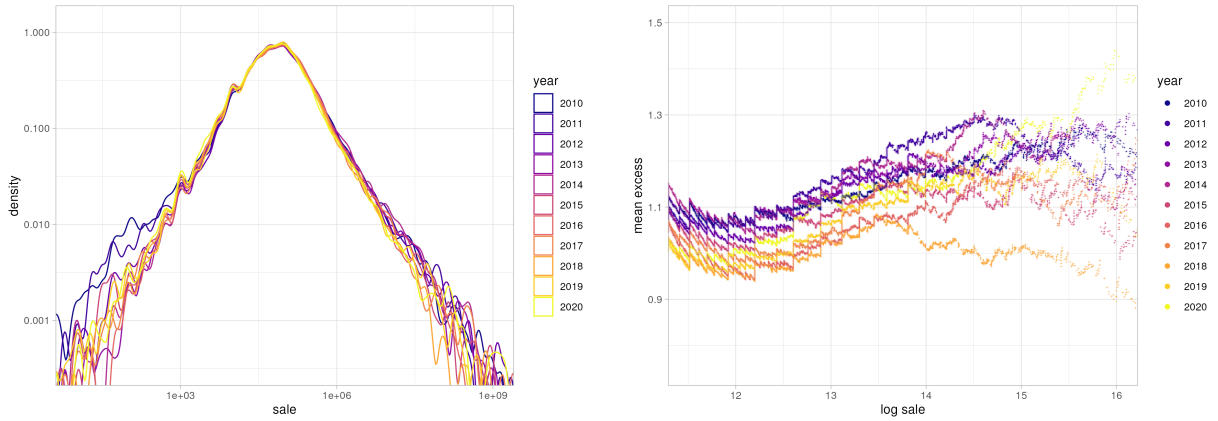
5 Empirical results: implications

In this section, we empirically evaluate three implications to determine whether the existing models or our theoretical explanation aligns more closely with the data.

5.1 Tail exponent

As discussed in Section 3.3, the slope of the tail of the size distribution on a log scale is determined by the weighted average of the tail slopes of the initial size S_0 distribution and the growth rate distribution. In particular, when the tail slopes of the growth rate distribution and S_0 distribution are equal, the size distribution shares the common slope, and an increase in n shifts the straight line upward in parallel. When the tail slopes of the growth rate distribution and S_0 distribution differ, the slope of the size distribution approaches the tail slope of the growth rate distribution as n increases. In this section, we use data to estimate the tail slopes of the size distribution, S_0 distribution, and the growth rate distribution to verify whether the above properties hold. In the previous literature, it has been shown that the tail slopes of firm sales distributions and individual income distributions differ, and we examine whether this difference aligns with our theoretical explanation.

Consider the distribution of S_0 for individual income. For individual income, S_0 is defined as the income of individuals who are 25 years old in 2014. The density estimate and mean excess function of the distribution of S_0 are provided in **Figure 23**. As evident from the figure, the observed heterogeneity in S_0 is significant. The tail slope of the distribution of S_0, estimated using the Hill estimator, is 1.746, while the tail slope of the distribution of S in 2020 is 1.80. This suggests that the distributions of S_0 and S have similar tail exponents.



(a) Density estimates

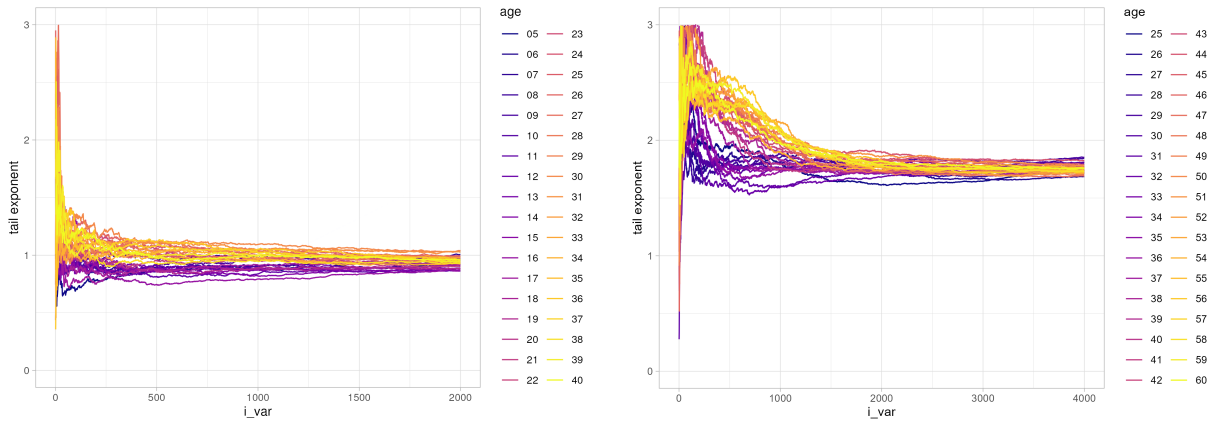
(b) Mean excess functions

Figure 24: Here, we consider the logarithm of the sales S_0 of firms that are five years old during the sample period (2010 to 2020). Panel (a) provides a density estimate of these samples, with the y-axis presented on a log scale. Panel (b) presents the mean excess function of S_0.

To estimate the tail slope of the growth rate distribution, we use the distribution of six-year growth rates $\widetilde{S}_6$, starting from 2014. The tail slope of this distribution, estimated using Hill's method, is 1.892. The difference between the tail slopes of the distribution of S_0 and the growth rate distribution is small. Therefore, according to Section 3.3, the tail exponent of the distribution for each age group should also be close to the tail exponents of the S_0 distribution and the growth rate distribution. As shown in Fig. (fig, tail expo)(b), the tail exponents of the distribution of S_n for each age group are clustered around 1.8, regardless of n. This result is consistent with our explanation that the tail exponent of the distribution of S_n is determined by a weighted average of the tail exponents of the distribution of S_0 and the growth rate distribution.

Next, consider the distribution of S_0 for firm sales. Here, S_0 is defined as the logarithmic value of sales in 2010 for firms established in 2005. The density estimate and mean excess function of the S_0 distribution are shown in **Figure 24**. As evident from the figure, the heterogeneity of S_0 is significant, with some firms positioned in the tail of the overall sales distribution despite $n = 0$. Furthermore, the tail slope of the S_0 distribution on a log scale, estimated using Hill's method, is 0.860. The tail slope of the S distribution in 2020 is 1.104, indicating that the S_0 distribution has a heavier tail compared to the S distribution.

To estimate the tail slope of the growth rate distribution, we use the distribution of 10-year growth rates $\widetilde{S}_{10}$, starting from 2010. The tail slope of this distribution, estimated using Hill's method, is 1.913. This indicates that the distribution of S_0 has a heavier tail than the growth rate distribution when comparing their tail slopes. According to Section 3.3, if the tail slopes of the distribution of S_0 and the growth rate distribution remain constant over time, the slope of the distribution of S_n should gradually converge to that of the growth rate distribution as n increases. In other words, the value of the tail exponent a should increase. This is demonstrated in Fig (fig, tail expo)(a). As evident from the figure, older groups exhibit larger tail exponents, indicating that the tails of the distributions of S_n become lighter as n increases. Thus,



(a) Firm sales **(b)** Individual incomes

Figure 25: Estimates of tail exponents. Panel (a) provides the Hill estimates for S_n ($n = 11, ..., 40$) as of 2020, based on the number of samples used for the estimation (i.e., the x-axis represents the number of samples used in descending order).

our theoretical explanation, which posits that the slope of the distribution of S_n is determined by a weighted average of these two tail exponents, is consistent with this result.

5.2 Age composition in the tail of the size distribution

The proportion of the tail of the size distribution occupied by agents of different ages serves as a key distinction between existing models and our theoretical explanation. As shown in Section 5.1, the tail of the size distribution S_n for each generation can be approximated on a logarithmic scale as follows:

$$\log \mathbb{P}(S_n > x \mid \text{age} = n) \approx -a_n x + b_n.$$

In particular, suppose that the slopes are common across generations, i.e., $a_1 = a_2 = \cdots =: a$. If $p_n := \mathbb{P}(\text{age} = n)$ represents the proportion of agents of age n in the overall population, then the distribution of size S, aggregated across different generations, can be expressed as:

$$\log \mathbb{P}(S > x) = \log \sum_n p_n \mathbb{P}(S_n > x \mid \text{age} = n) \approx \log \left(\left(\sum_n p_n e^{b_n} \right) e^{-ax} \right) = -ax + b$$

where $b := \log \sum_n p_n e^{b_n}$. In other words, if the tail of the size distribution of each generation shares a common slope a, the tail of the overall size distribution also share the same slope a. In this case, note that $\log \mathbb{P}(S_n > x \mid \text{age} = n) - \log \mathbb{P}(S > x) = \text{const.}$, meaning that the ratio $\frac{\mathbb{P}(S_n > x \mid \text{age}=n)}{\mathbb{P}(S>x)}$ is a constant, independent of x. Thus, the proportion of agents of age n occupying the tail of the size distribution S is given by:

$$\frac{\mathbb{P}(S_n > x, \text{age} = n)}{\mathbb{P}(S > x)} = p_n \frac{\mathbb{P}(S_n > x \mid \text{age} = n)}{\mathbb{P}(S > x)} = \text{const.}$$

That is, the proportion of each generation occupying the tail of the overall size distribution depends only on n and is independent of x. This explains why Zipf's law holds for the overall size distribution in our theoretical explanation.

The above property is confirmed using individual income data, as shown in **Figure 5**(b). As evident from the figure, the proportion of each generation contributing to the tail probability remains stable across a wide range of x. This contrasts with the predictions of existing models, which suggest that the tail of the size distribution becomes increasingly dominated by older agents as x increases. Therefore, this result supports our theoretical explanation that the Zipf's law for the overall size distribution does not arise from the superposition of distributions from different generations, but rather that Zipf's law already holds within the size distribution of each generation.

The results for firm sales are shown in **Figure 3**(a). This figure suggests that the proportion of younger firms increases as x grows. Strictly speaking, this result deviates from Zipf's law, but it aligns with the findings in Section 5.1 regarding how such deviations occur. Specifically, in the case of firm sales, the tail slope of the initial size distribution S_0 is heavier than that of the growth rate distribution. As a result, when n is small, the tail slope of the size distribution tends to be closer to the tail slope of S_0's distribution. Thus, in the tail of the size distribution S, the proportion of younger firms increases with x when n is small. This result is consistent with the findings in Section 5.1 and supports our theoretical explanation.

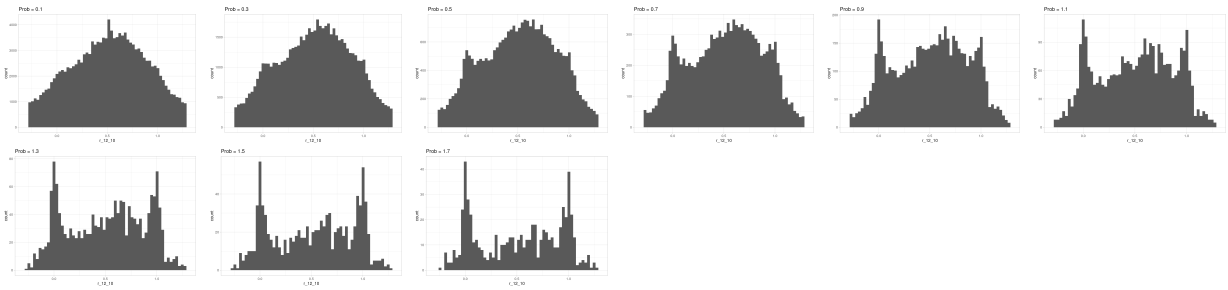


Figure 26: Histogram of r for the subsample satisfying the condition $X_{10} + X_{11} > u$. The value of u increases from 0.1 (top left) to 1.9 (bottom right) in increments of 0.2.

5.3 Jump-driven growth processes

Here, we verify the implications discussed in Section 3.4 that the large deviation of $\widetilde{S}_n$ is determined not by the average growth over n periods but by rapid growth in a specific period, i.e., a jump. First, considering the growth rates X_k, X_{k+1} and their sum $X_k + X_{k+1}$, let us define the following ratio:

$$r := \frac{X_k}{X_k + X_{k+1}}$$

By definition, the sum $X_k + X_{k+1}$ represents the agent's growth rate over two years, and r indicates the contribution of the first year's growth rate to the total growth rate over these two years. For instance, if the growth rates are 3% in the first year and 3% in the second year, r becomes $1/2$, meaning that both years equally contribute to the two-year growth rate. Since X_k is assumed to be an i.i.d. random variable in our analysis, the distribution of r is symmetric around $1/2$. We test whether the event $r = 1/2$, indicating equal contributions from the first and second years to the total growth rate, is the most likely scenario. According to Section 3.4, if the growth rate distribution is subexponential, then for large $X_k + X_{k+1}$, the event $r = 1/2$ should be the least likely. Using our empirical data, we calculate r for each agent and examine the histogram of r.

Using the growth rates of firm sales for 2010 and 2011, we compute the histogram of r, as shown in **Figure 26**. Here, we construct the histogram of r for samples satisfying the condition $X_{10} + X_{11} > u$ and analyze how the histogram changes as u increases. The value of u ranges from 0.1 to 1.9 in increments of 0.2. As shown in **Figure 26**, when u is small, the histogram of r peaks at $1/2$. This indicates that when the two-year growth rate is low, the growth rates of both 2010 and 2011 contribute almost equally to the total two-year growth rate, making this the most likely event. In contrast, as u increases (e.g., $u \geq 0.9$), the histogram of r exhibits peaks near 0 and 1. This suggests that high two-year growth rates are driven by an exceptionally large growth rate in either 2010 or 2011. In other words, instead of both years contributing equally to the high growth rate, rapid growth—or a jump—occurs in one of the years, making this the most likely event leading to the high overall growth rate.

A similar result is observed for the growth rates of individual incomes. The results for individual income growth rates are shown in **Figure 27**. Here, we consider the contribution of the first three-year growth rate

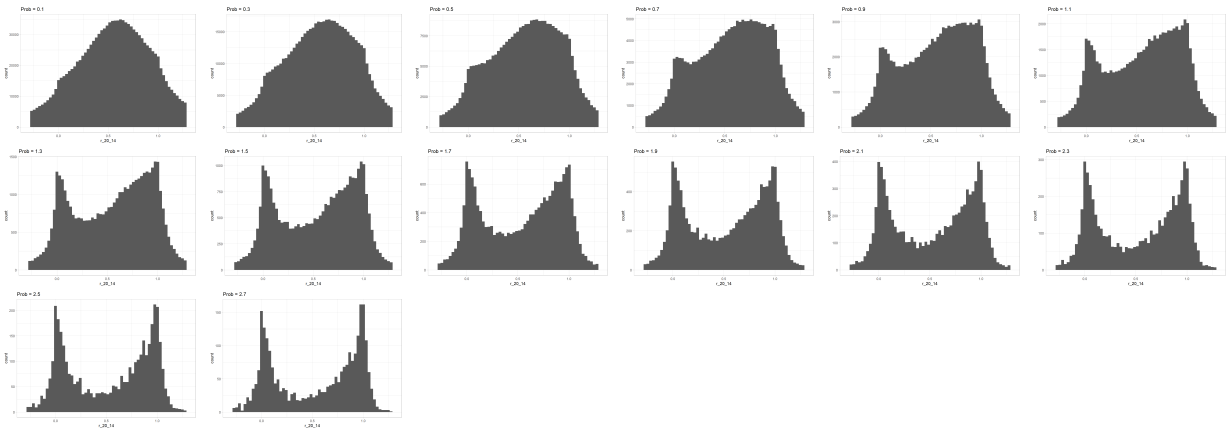
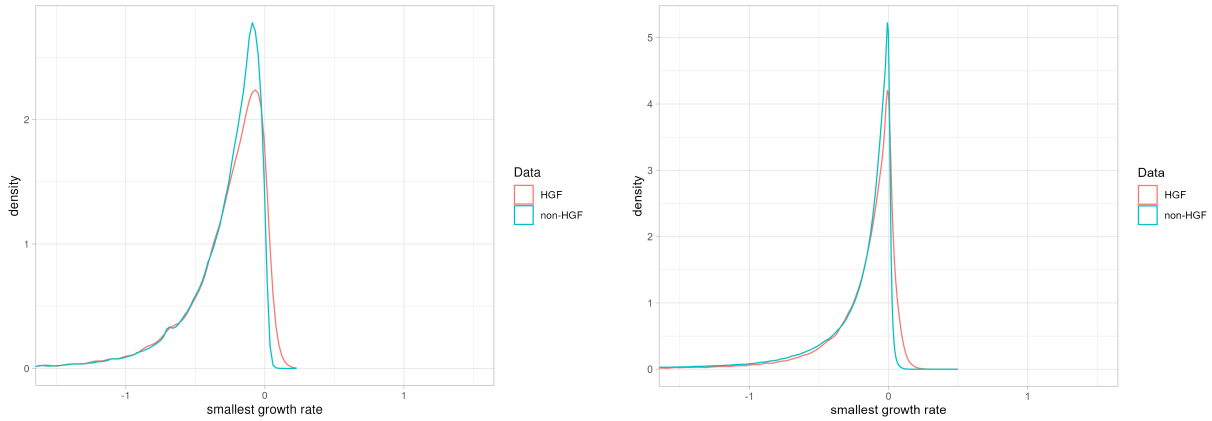


Figure 27: Histogram of r for the subsample satisfying the condition $\sum_{k=2015}^{2020} X_k > u$. The value of u increases from 0.1 (top left) to 2.7 (bottom right) in increments of 0.2.

to the total six-year growth rate, expressed as the ratio r, for the period from 2014 to 2020. The value of u ranges from 0.1 to 2.7 in increments of 0.2. As with the growth rates of firm sales, when u is small, the histogram of r peaks at $1/2$. This indicates that when the six-year growth rate is low, the growth rates of both the first and second three-year periods contribute almost equally to the total growth rate, making this the most likely event. However, as u increases (e.g., $u \geq 0.9$), the histogram exhibits peaks near 0 and 1. This suggests that instead of both three-year periods contributing equally to the high growth rate, rapid growth occurs in one of the periods, resulting in the overall high growth rate being most likely driven by a single period.[22] These results are consistent with the subexponential property discussed in Section 3.4 and directly demonstrate the principle of a single big jump.

Next, we examine the conditional probability of growth rates given a large deviation in $\widetilde{S}_n$. Specifically, according to our theoretical explanation, high growth over n periods ($\widetilde{S}_n > x$) is realized through rapid growth in a single period (i.e., a jump), while the distribution of the remaining $n-1$ smallest growth rates should match the unconditional growth rate distribution. Therefore, the distribution of the minimum value of the $n-1$ smallest growth rates for each agent satisfying this condition should match the minimum value of $n-1$ growth rates drawn from the unconditional growth rate distribution. In contrast, if the growth rate distribution is light-tailed, the conditional probability of growth rates given $\widetilde{S}_n > x$ would differ from the unconditional distribution. In particular, if the growth rate distribution is Gaussian, the conditional probability of growth rates given $\widetilde{S}_n > x$ would shift the mean by x/n in the positive direction. To determine which of these two cases the data aligns with, we divide the samples into agents who satisfy $\widetilde{S}_n > x$ and those who do not. We then compare the histograms of the minimum values of $n-1$ growth rates for each

[22] Here, we are essentially determining the distribution of the variable r for a very small subsample that satisfies $\widetilde{S}_n > u$. As such, the sample size plays a critical role in observing the shape of the distribution of r. Compared to the firm sales data, the individual income data has a larger sample size, making it sufficient to observe the shape of the distribution of r under the condition $\widetilde{S}_n > u$. Indeed, compared to **Figure 26**, **Figure 27** more clearly demonstrates the characteristic shape of the subexponential property.



(a) Firm sales growth

(b) Individual income growth

Figure 28: Comparison of histograms of the smallest value among n individual growth rates. The histograms are compared between agents that achieve $\widetilde{S}_n > x$ and those that do not.

agent. By examining whether the histograms of these two samples differ by approximately x/n or are almost identical, we can distinguish between the two cases described above.

For the growth rates of firm sales, we use the 97th percentile value as the threshold for large deviations in $\widetilde{S}_n$ over $n = 10$ periods. From the data, the 97% percentile value of $\widetilde{S}_n$ is 0.926. If the growth rate distribution were Gaussian, the conditional probability distribution would shift approximately $0.926/10 = 0.0926$ to the right. **Figure 28**(a) compares the histograms of the minimum growth rate for each firm between the group that satisfies $\widetilde{S}_n > 0.926$ and the group that does not. As shown in the figure, the distributions of the minimum growth rates for the two groups are very similar. In fact, the medians of the minimum growth rates are -0.198 for the group satisfying $\widetilde{S}_n > x$ and -0.201 for the group that does not, with a difference of only 0.003. Given that the expected difference of 0.0926 for a Gaussian distribution, the observed difference in the histograms is negligible.

Similarly, the results for the growth rates of individual incomes are shown in **Figure 28**(b). Here, with $n = 6$, the 97% percentile value of the growth rate $\widetilde{S}_n$ over this period is 0.891. Thus, if the growth rate distribution were Gaussian, the conditional probability distribution would shift approximately $0.891/10 = 0.0891$ to the right. As shown in **Figure 28**(b), the histograms of the minimum growth rates for individuals in the group satisfying $\widetilde{S}_n > x$ and those not satisfying it are very similar. The medians of the minimum growth rates for these groups are -0.0947 and -0.112, respectively, with a difference of 0.017. Compared to the expected difference of 0.0891 for a Gaussian distribution, the observed difference in these distributions is small.

These results on the ratio r and the conditional probability demonstrate that the large deviation in $\widetilde{S}_n$ is driven by a jump occurring in some (but unspecified) period. Therefore, these findings provide evidence that our theoretical explanation aligns better with the data compared to existing models.

6 Conclusion

Zipf's law is a characteristic feature that commonly appears in the tail of various distributions, such as firm sales and individual incomes, and has long attracted the attention of many economists. However, recent theoretical and empirical studies have pointed out discrepancies between existing models and data, particularly regarding the time required for growth and convergence to a steady state. We proposed an alternative explanation for Zipf's law to resolve discrepancies with the data. Using Japanese firm-level data and individual income data, we verified that our theoretical explanation aligns with the observed data.

The main idea of this paper is that there are two distinct patterns that generate giant firms or super-rich individuals. The first pattern assumes a light-tailed growth rate distribution, where moderate growth rates sustained over a long period result in the emergence of giant firms or super-rich individuals. Existing models are based on this growth pattern, leading to discrepancies with data, such as the excessively long time required for firms to become giant or individuals to become super-rich, as well as for convergence to a stationary distribution. In contrast, the second pattern assumes a heavy-tailed growth rate distribution, where a single large deviation in growth rate—a jump—can lead to giant firms or super-rich individuals. Our explanation is based on this growth pattern, which resolves the issues pointed out in existing models.

It is important to note that our conclusions are derived from statistical properties observed in empirical data. Rather than relying on economic models of firm behavior or individual decision-making, our explanation is grounded in the statistical properties of distributions, allowing it to be applied to both firm sales and individual incomes. Our explanation is an example that demonstrates how the logic of probability forms the basis of the statistical universality observed in economic phenomena.

7 Appendix: Copula Theory

7.1 Why is copula theory needed?

Examining the dependence between growth rates, particularly the dependence between consecutive growth rates X_t and X_{t+1}, is crucial for understanding the growth dynamics of firm sales and individual incomes. In previous studies, the measure most commonly used to quantify this type of dependence has been Pearson's correlation coefficient:

$$\text{Corr}(X_t, X_{t+1}) = \mathbb{E}[(X_t - \mu_t)(X_{t+1} - \mu_{t+1})]$$

Here, μ_t and μ_{t+1} denote the mean growth rates for each period, respectively.

However, it is known in statistics that Pearson's correlation coefficient does not exclusively represent the dependence between two random variables (cf. Embrechts et al. (2002)). This is because Pearson's correlation coefficient is influenced not only by the dependence between the variables but also by their marginal distributions. In other words, Pearson's correlation coefficient can change with changes in the marginal distributions, even if the dependence structure between the two random variables remains the same. For example, if h_1 and h_2 are strictly increasing functions, the Pearson correlation coefficient of X_t and X_{t+1} does not necessarily equal that of $h_1(X_t)$ and $h_2(X_{t+1})$. This property of Pearson's correlation coefficient makes it difficult to discern whether it reflects the strength of dependence or merely the marginal distributions of X_t and X_{t+1}. This issue is especially relevant in ours case, where the growth rate distributions deviate significantly from a Gaussian distribution and exhibits heavy tails, making this property of Pearson's correlation coefficient problematic.

This issue with Pearson's correlation coefficient has already been pointed out in the literature on firm growth dynamics. For example, Bottazzi et al. (2023) addresses this problem by employing a quantile transition matrix. That is, rather than analyzing the transition probabilities of growth rates X_t and X_{t+1} directly, they focus on the transition probabilities of quantiles $F_t(X_t)$ and $F_{t-1}(X_{t+1})$, where F_t and F_{t-1} represent the distributions of growth rates in the current and previous periods, respectively. They show that, along with the dependence in the tails (i.e., extreme values are likely to occur consecutively), there is also a bouncing effect where an extreme value is likely to be followed by an extreme value in the opposite direction. Such analyses can be rigorously discussed using copulas, which are introduced below.

7.2 Introduction to copula theory

The main idea of copula theory is that any bivariate distribution function F can be uniquely decomposed as follows:

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2))$$

Here, C is called a copula function, and F_1 and F_2 are the marginal distributions for X_1, X_2, respectively. In other words, C is independent of the marginal distributions, and the dependence structure is uniquely determined by C. In the following, we analyze the properties of the dependence structure by focusing on the characteristics of C.

A useful alternative measure for assessing dependence between two random variables is Spearman's ρ_S, which is also employed in Section 4.2:

$$\rho_S := \text{Cor}\left[F_1\left(X_1\right), F_2\left(X_2\right)\right]$$

This measure is advantageous because it is known to be determined by the copula C as follows:

$$\rho_S = 12 \int_{[0,1]^2} C(u, v) du dv - 3$$

Since Spearman's ρ is independent of the marginal distributions, it avoids the issues associated with Pearson's correlation coefficient.

Now, what is the typical form of a copula function C? While the copula for two independent variables (i.e., the independence copula) is uniform with respect to variables U_1, U_2, what form does the copula take in cases beyond this simplest case? For instance, the findings of Bottazzi et al. (2023), when expressed in terms of a copula, imply that function C has a higher density in the tail region compared to the independence copula. Is this characteristic unusual? Below, we describe the structure of two commonly used copulas: the Gaussian copula and the Student-t copula.

First, as a non-parametric method to characterize the copula function C, we consider the empirical copula (and its density function).

$$\widehat{C}_n(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} I\left(\left[\frac{r_{ij} - \frac{1}{2}}{n} \leq u_j,\ j = 1, 2\right]\right)$$

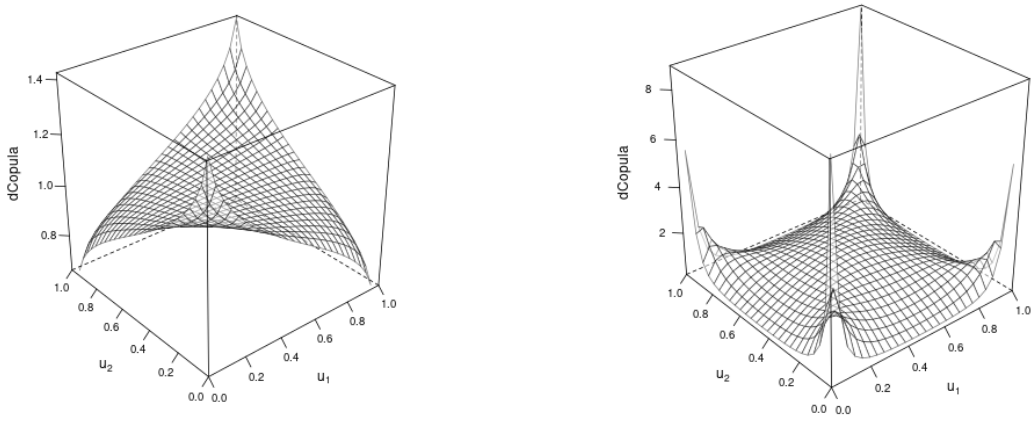
Here, $r_{1j}, \ldots, r_{nj}$ are the ranks of the jth variable in increasing order. As seen from the definition of the copula, the copula is a function of $F_1(X_1)$ and $F_2(X_2)$, rather than the random variables X_1 and X_2 themselves. The empirical copula is based on the realizations of $F_1(X_1)$ and $F_2(X_2)$.

Gaussian copulas are derived from bivariate Gaussian distributions, expressed as follows:

$$C(u, v; \rho) = \Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v); \rho\right)$$

Here, Φ_2 represents the bivariate Gaussian distribution with parameter ρ, and Φ denotes the univariate Gaussian distribution. Its density function is given by:

$$c(u, v; \rho) = \left(1 - \rho^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(x^2 + y^2 - 2\rho xy\right)\left(1 - \rho^2\right)\right\} \cdot \exp\left\{\frac{1}{2}\left(x^2 + y^2\right)\right\}$$



(a) Gaussian copula

(b) Student-t copula

Figure 29: Density plots of the copula function C. In both copulas, the parameter ρ set to 0.1.

The density function of the Gaussian copula is provided in **Figure 29**(a). As can be seen from this graph, the density values in the tail regions (i.e., around $(0,0)$ and $(1,1)$) become infinite. Note that even copulas with weak tail dependence, such as the Gaussian copula, exhibit spikes at $(0,0)$ and $(1,1)$ (i.e., spikes in the density plot at $(0,0)$ and $(1,1)$ do not necessarily indicate strong tail dependence for that copula).

The second example is the Student-t copula, derived from the Student-t distribution. Its density function is given by:

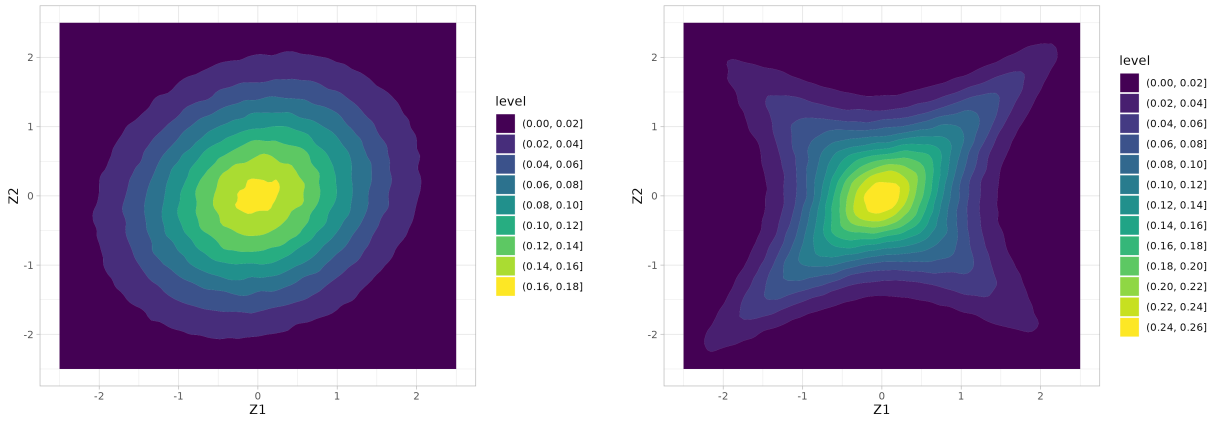
$$t_{2,\nu}(y;\rho) = \left(1-\rho^2\right)^{-1/2} \frac{\Gamma((\nu+2)/2)}{\Gamma(\nu/2)[\pi\nu]} \left(1 + \frac{y_1^2 + y_2^2 - 2\rho y_1 y_2}{\nu\left(1-\rho^2\right)}\right)^{-(\nu+2)/2}$$

This density function is provided in **Figure 29**(b). A characteristic feature of this Student-t copula is that, for $\rho > 0$, it not only shows dependence around the $(0,0)$ and $(1,1)$ regions but also indicates dependence in the $(0,1)$ and $(1,0)$ regions. In other words, even with a positive parameter ρ, the Student-t copula, compared to the Gaussian copula, has the characteristic that extreme values are more likely to be followed by opposite extreme values.

While plotting the empirical copula and analyzing its shape is often used in the literature, the density function can be infinite near $(0,0)$ and $(1,1)$. To avoid this issue, Joe (2014) recommends transforming the variables to normal scores and analyzing their plot. Specifically, transforming X_j to $Z_j = \Phi^{-1} \circ F_j(X_j)$, where Φ is the standard normal distribution function, and consider the bivariate distribution F_N with standard normal marginals:

$$F_N(Z_1, Z_2) := C(\Phi(Z_1), \Phi(Z_2))$$

If the copula C is a Gaussian copula, then F_N is a bivariate Gaussian distribution. Thus, if the plot of Z_1 and Z_2 calculated from the empirical data deviates from a bivariate Gaussian, it indicates a deviation from the Gaussian copula. Additionally, the Pearson's correlation coefficient of these transformed variables Z_1 and Z_2 (denoted as ρ_N) is also used as a measure of dependence between the variables. Similar to Spearman's



(a) Gaussian copula

(b) Student-t copula

Figure 30: Density plots of normal scores. The parameters used are the same as those in **Figure 29**.

ρ, ρ_N is determined by the copula and is independent of the marginal distributions.

Figure 30 presents Gaussian and Student-t copulas, not in the density plot of U_1 and U_2, but transformed into normal scores, Z_1 and Z_2 (see also Section 4.2). As shown in the figure, the Gaussian copula takes on an elliptical shape, while the Student-t copula extends more towards the four corners, resembling a diamond shape. This reflects the stronger tail dependence of the Student-t copula compared to the Gaussian copula.

Additionally, we consider another measure of dependence using normal scores, known as the semi-correlation coefficients:

$$\rho_N^+ = \text{Corr}\left[Z_1, Z_2 \mid Z_1 > 0, Z_2 > 0\right],$$
$$\rho_N^- = \text{Corr}\left[Z_1, Z_2 \mid Z_1 < 0, Z_2 < 0\right]$$

These are referred to as the upper and lower semi-correlation coefficients, respectively. These coefficients represent the correlation in the upper-right and lower-left quadrants of a plot of the sample data.[23]

The parameters of the copula are estimated using the pseudo-maximum likelihood method proposed by Genest et al. (1995). This method involves two steps: (1) non-parametrically estimating the parameters of the marginal distributions (2) estimating the parameters of the copula. By separating the estimation process into two parts, this method offers the advantage of reducing the overall computational cost involved in estimation. Furthermore, to verify whether the Student-t copula fits the data significantly better than the Gaussian copula, we employ Vuong's method (Vuong (1989)). This method compares the maximum likelihood of the two models (specifically, the log-likelihood ratio) to test if the maximum likelihoods of the

[23] When the copula C is a Gaussian copula parameterized by ρ (i.e., when Z_1, Z_2 follow a Gaussian distribution), the ρ_N^+ (and by symmetry ρ_N^-) can be derived precisely.

$$\text{Corr}\left[Z_1, Z_2 \mid Z_1 > 0, Z_2 > 0; \rho\right] = \frac{v_{1,1}(\rho) - v_{1,0}^2(\rho)}{v_{2,0}(\rho) - v_{1,0}^2(\rho)}$$

Therefore, by comparing this theoretical solution with the actual values obtained from the data, deviations from the Gaussian copula can be identified.

two models are statistically significantly different. Here, the two models considered are the Student-t copula and the Gaussian copula.

Matrix of correlation coefficients

Pearson's correlation coefficient of normal scores

term	n_20_18	n_19_18	n_18_17	n_17_16	n_16_15	n_15_14	n_14_13	n_13_12	n_12_11	n_11_10
n_20_18	NA	−0.054	0.056	0.058	0.061	0.048	0.050	0.044	0.034	0.033
n_19_18	−0.054	NA	−0.082	0.062	0.060	0.061	0.052	0.045	0.047	0.045
n_18_17	0.056	−0.082	NA	−0.086	0.053	0.069	0.066	0.043	0.043	0.058
n_17_16	0.058	0.062	−0.086	NA	−0.092	0.056	0.062	0.052	0.044	0.046
n_16_15	0.061	0.060	0.053	−0.092	NA	−0.078	0.049	0.050	0.049	0.048
n_15_14	0.048	0.061	0.069	0.056	−0.078	NA	−0.083	0.052	0.056	0.058
n_14_13	0.050	0.052	0.066	0.062	0.049	−0.083	NA	−0.076	0.056	0.064
n_13_12	0.044	0.045	0.043	0.052	0.050	0.052	−0.076	NA	−0.081	0.041
n_12_11	0.034	0.047	0.043	0.044	0.049	0.056	0.056	−0.081	NA	−0.077
n_11_10	0.033	0.045	0.058	0.046	0.048	0.058	0.064	0.041	−0.077	NA

Table 7: Matrix of the correlation coefficients of normal scores ρ_N. Samples are the same as in **Table 2**.

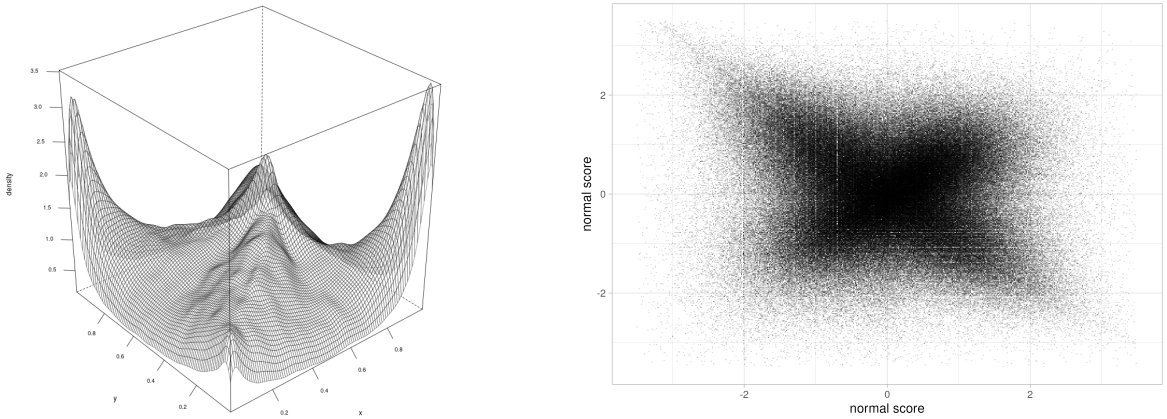
7.3 Empirical Results

Firm sales

We apply the above method using growth rate data for firm sales from 2011 and 2012. **Figure 31** presents the empirical copula and the plot of normal scores. The density function of the empirical copula exhibits spikes not only near $(1, 1)$ and $(-1, -1)$, but also near $(1, -1)$ and $(-1, 1)$, which is a characteristic of the Student-t copula. Additionally, the normal scores plot does not resemble the elliptical shape typical of a Gaussian copula but instead takes on a diamond shape, characteristic of a Student-t copula.

The fact that the empirical copula is closer to a Student-t copula than a Gaussian copula is also reflected in the correlation coefficients. As shown in **Table 5** and **Table 7**, both ρ_S and ρ_N are close to 0. However, the semi-correlation coefficients are $\rho_N^+ = 0.240$ and $\rho_N^- = 0.304$, respectively. This indicates that while no strong positive or negative correlation is observed across the entire distribution, relatively strong correlations emerge when the range is restricted.

To verify the characteristics of the copula observed above, we now consider the Gaussian copula and the Student-t copula, and examine which of these better approximates the empirical data. The estimation results for the parameters of each copula are presented in **Table 8**. Additionally, the Akaike Information Criterion (AIC) calculated from the pseudo-maximum likelihood method is provided for each copula. Based on the AIC, the copula that best fits the data is the Student-t copula with degrees of freedom $\nu = 2$. Furthermore, we test whether the difference between the Gaussian copula and the Student-t copula is statistically significant using Vuong's method. The null hypothesis is rejected with a p-value of less than 0.01. This means that the Student-t copula provides a statistically significantly better fit compared to the Gaussian copula. This result also indicates that, when the region is restricted to, e.g., the first quadrant, the actual copula exhibits stronger dependence than that predicted by a Gaussian copula.



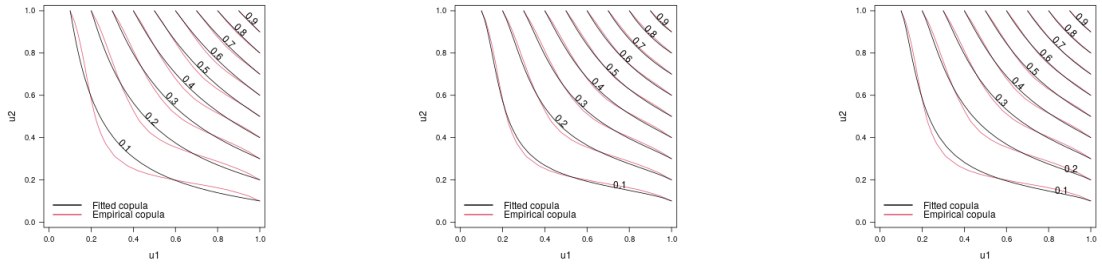
(a) Pseudo observations

(b) Normal scores

Figure 31: Plot of pseudo observations and normal scores of growth rates X_t, X_{t+1}.

Estimates for parametric copulas Data: Tokyo Shoko Research from 2010 to 2020				
copula	shape	negative LL	Vuong	semi-correlation coefficient
Gaussian	–0.163	1344	-	–0.055
Student-t with df 1	–0.052	1054	-0.00290(0.306)	0.604
Student-t with df 2	–0.104	7072	0.0573(0.151)	0.367
Student-t with df 3	–0.129	6810	-	0.246
Student-t with df 4	–0.143	6158	-	0.175
Student-t with df 5	–0.151	5583	-	0.130
Student-t with df 10	–0.163	3952	-	0.036

Table 8: Results of pseudo-maximum likelihood of copula families. In this analysis, since the number of parameters for all considered copula families is equal, the choice based on the Akaike Information Criterion (AIC) is equivalent to the choice based on likelihood. The statistic used in Vuong's method is $\widehat{D} := \log(f^S(y_i;\widehat{\theta}^S)/f^G(y_i;\widehat{\theta}^G))$, where f^S, θ^S represent the likelihood and maximum likelihood estimate for the Student-t copula, and f^G, θ^G represent those for the Gaussian copula.



(a) Gaussian copula **(b)** Student-t copula with $\nu = 1$ **(c)** Student-t copula with $\nu = 2$

Figure 32: Comparison between empirical and model-estimated copulas. The contour plots of the copula $C(u, v)$ are drawn.

We directly assess how well the estimated Student-t copula fits the data by comparing it to the empirical copula. **Figure 32** presents contour plots of both the empirical copula and the estimated copula. In addition to the best-fitting Student-t copula, we also provide the contour plot of the estimated Gaussian copula for comparison. It is evident that the Student-t copula closely approximates the empirical copula. We also verify whether the estimated Student's t-copula replicates the observed dependence measures. The correlation coefficients for the normal scores from the estimated Student's t-copula with degrees of freedom $\nu = 2$ are $\rho_N^+ = \rho_N^- = 0.373$. These results suggest that the Student's t-copula captures the characteristics of the empirical copula well, including the dependence measure.

Individual incomes

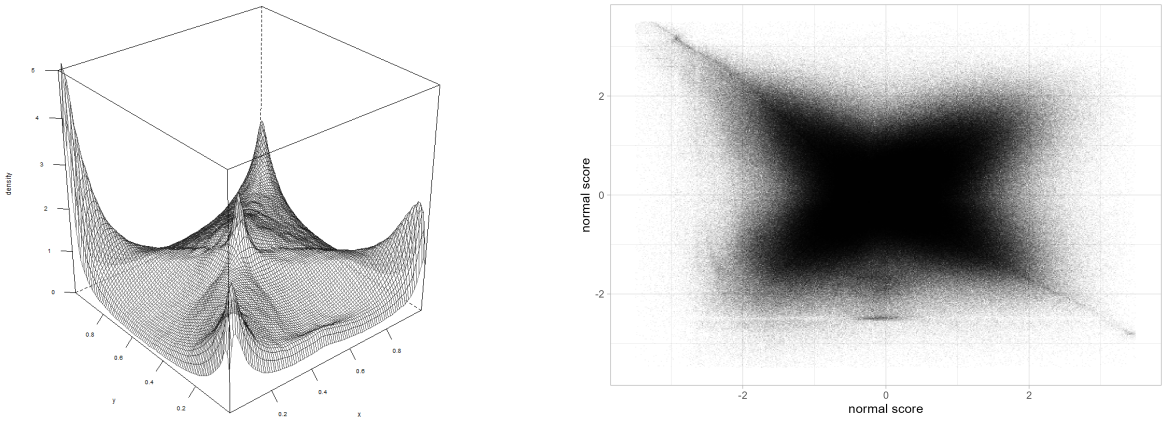
We investigate the dependency between individual income growth rates for 2015 and 2016. **Figure 33** presents the empirical copula and the plot of normal scores. As evident from the figure, the density function of the empirical copula exhibits spikes at the four corners, a characteristic of the Student-t copula. The normal scores plot also takes on a diamond shape, resembling the Student-t copula. Furthermore, the semi-correlation coefficients are $\rho_N^+ = 0.387$ and $\rho_N^- = 0.375$, indicating stronger correlations in the first and third quadrants compared to a Gaussian copula.

This property is also confirmed through parameter estimation for the Gaussian copula and the Student-t copula. The estimation results are presented in **Table 10**. Based on the AIC criterion, as in the case of firm sales, the copula that best fits the data is the Student-t copula with degrees of freedom $\nu = 2$. Additionally, using Vuong's method, we find that the difference between the Student-t copula and the Gaussian copula is statistically significant at a p-value of 0.01.

The contour plots of the empirical copula and the estimated copulas are provided in **Figure 34**. For comparison, the contour plot of the Gaussian copula is also included alongside the Student-t copula. As evident from the figure, the Student-t copula closely approximates the empirical copula. Furthermore, the semi-correlation coefficients estimated from the Student-t copula, $\rho_N^+ = \rho_N^- = 0.396$, closely align with the

Matrix of correlation coefficients						
Pearson's correlation coefficient of normal scores						
term	n_20_19	n_19_18	n_18_17	n_17_16	n_16_15	n_15_14
n_20_19	NA	0.010	−0.011	0.000	0.014	0.007
n_19_18	0.010	NA	0.015	−0.002	0.021	0.022
n_18_17	−0.011	0.015	NA	0.024	0.005	0.020
n_17_16	0.024	−0.002	0.024	NA	0.019	0.003
n_16_15	0.014	0.021	0.005	0.019	NA	0.026
n_15_14	0.007	0.022	0.020	0.003	0.026	NA

Table 9: Matrix of ρ_N. Samples are the same as in **Table 2**.



(a) Pseudo observations

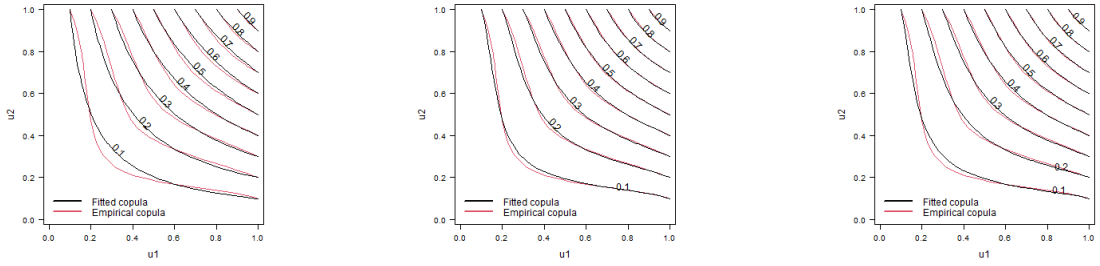
(b) Normal scores

Figure 33: Plot of pseudo observations and normal scores of growth rates X_t, X_{t+1}.

Estimates for parametric copulas Data: Tax return data from National Tax College from 2014 to 2020				
copula	shape	negative LL	Vuong	semi-correlation coefficient
Gaussian	−0.014	9.323	-	−0.005
Student-t with df 1	0.037	3575.000	0.03566(0.0009)	0.616
Student-t with df 2	0.032	7876.000	0.1477(0.0009)	0.396
Student-t with df 3	0.026	6980.000	-	0.287
Student-t with df 4	0.021	5989.000	-	0.222
Student-t with df 5	0.017	5201.000	-	0.180
Student-t with df 10	0.006	3121.000	-	0.091

Table 10: Results of pseudo-maximum likelihood of copula families. In this analysis, since the number of parameters for all considered copula families is equal, the choice based on the Akaike Information Criterion (AIC) is equivalent to the choice based on likelihood. The statistic used in Vuong's method is $\widehat{D} := \log(f^S(y_i;\widehat{\theta}^S)/f^G(y_i;\widehat{\theta}^G))$, where f^S, θ^S represent the likelihood and maximum likelihood estimate for the Student-t copula, and f^G, θ^G represent those for the Gaussian copula.

empirical values. These findings suggest that the Student-t copula effectively captures the dependence in growth rates.



(a) Gaussian copula **(b)** Student-t copula with $\nu = 1$ **(c)** Student-t copula with $\nu = 2$

Figure 34: Comparison between empirical and model-estimated copulas. The contour plots of the copula $C(u, v)$ are drawn.

References

Arata, Y. (2019). Firm growth and laplace distribution: The importance of large jumps. *Journal of Economic Dynamics and Control*, 103:63–82.

Arata, Y., Miyakawa, D., and Mori, K. (2023). The U-shaped law of high-growth firms. *NTC Discussion Paper*, 230201-01HJ(2).

Armendáriz, I. and Loulakis, M. (2011). Conditional distribution of heavy tailed random variables on large deviations of their sum. *Stochastic processes and their applications*, 121(5):1138–1147.

Ascher, S. (1990). A survey of tests for exponentiality. *Communications in Statistics-Theory and Methods*, 19(5):1811–1825.

Axtell, R. L. (2001). Zipf distribution of us firm sizes. *science*, 293(5536):1818–1820.

Bazhba, M., Blanchet, J., Rhee, C. H., and Zwart, B. (2020). Sample path large deviations for lévy processes and random walks with weibull increments. *Annals of Applied Probability*, 30(6):2695–2739.

Beare, B. K. and Toda, A. A. (2022). Determination of pareto exponents in economic models driven by markov multiplicative processes. *Econometrica*, 90(4):1811–1833.

Beirlant, J., Broniatowski, M., Teugels, J. L., and Vynckier, P. (1995). The mean residual life function at great age: Applications to tail estimation. *Journal of Statistical Planning and Inference*, 45(1-2):21–48.

Borovkov, A. A. (2020). *Asymptotic analysis of random walks: light-tailed distributions*, volume 176. Cambridge University Press.

Borovkov, A. A. and Borovkov, K. (2008). *Asymptotic analysis of random walks*, volume 118. Cambridge University Press.

Bottazzi, G., Coad, A., Jacoby, N., and Secchi, A. (2011). Corporate growth and industrial dynamics: Evidence from french manufacturing. *Applied Economics*, 43(1):103–116.

Bottazzi, G., Kang, T., and Tamagni, F. (2023). Persistence in firm growth: inference from conditional quantile transition matrices. *Small Business Economics*, 61(2):745–770.

Bottazzi, G. and Secchi, A. (2006). Explaining the distribution of firm growth rates. *The RAND Journal of Economics*, 37(2):235–256.

Champernowne, D. G. (1953). A model of income distribution. *The Economic Journal*, 63(250):318–351.

Coad, A. (2009). *The growth of firms: A survey of theories and empirical evidence*. Edward Elgar Publishing.

Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*. CRC press.

Dosi, G., Grazzi, M., Moschella, D., Pisano, G., and Tamagni, F. (2020). Long-term firm growth: an empirical analysis of us manufacturers 1959–2015. *Industrial and Corporate Change*, 29(2):309–332.

Dosi, G., Pereira, M. C., and Virgillito, M. E. (2017). The footprint of evolutionary processes of learning and selection upon the statistical properties of industrial dynamics. *Industrial and Corporate Change*, 26(2):187–210.

El Methni, J., Gardes, L., Girard, S., and Guillou, A. (2012). Estimation of extreme quantiles from heavy and light tailed distributions. *Journal of Statistical Planning and Inference*, 142(10):2735–2747.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: properties and pitfalls. *Risk management: value at risk and beyond*, 1:176–223.

Foss, S., Korshunov, D., and Zachary, S. (2011). *An Introduction to Heavy-Tailed and Subexponential Distributions*.

Gabaix, X. (1999). Zipf's law for cities: an explanation. *The Quarterly journal of economics*, 114(3):739–767.

Gabaix, X. (2009). Power laws in economics and finance. *Annu. Rev. Econ.*, 1(1):255–294.

Gabaix, X., Lasry, J.-M., Lions, P.-L., and Moll, B. (2016). The dynamics of inequality. *Econometrica*, 84(6):2071–2111.

Gardes, L., Girard, S., and Guillou, A. (2011). Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of Statistical Planning and Inference*, 141(1):429–444.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552.

Guvenen, F., Karahan, F., Ozkan, S., and Song, J. (2021). What do data on millions of us workers reveal about lifecycle earnings dynamics? *Econometrica*, 89(5):2303–2339.

Guvenen, F., Pistaferri, L., and Violante, G. L. (2022). Global trends in income inequality and income dynamics: New insights from grid. *Quantitative Economics*, 13(4):1321–1360.

Henze, N. and Meintanis, S. G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons. *Metrika*, 61:29–45.

Ijiri, Y. and Simon, H. A. (1964). Business firm growth and size. *The American Economic Review*, 54(2):77–89.

Ijiri, Y. and Simon, H. A. (1977). *Skew distributions and the sizes of business firms*, volume 24. North-Holland.

Joe, H. (2014). *Dependence modeling with copulas*. CRC press.

Kotz, S., Kozubowski, T., and Podgórski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media.

Luttmer, E. G. (2007). Selection, growth, and the size distribution of firms. *The Quarterly Journal of Economics*, 122(3):1103–1144.

Luttmer, E. G. (2010). Models of growth and firm heterogeneity. *Annu. Rev. Econ.*, 2(1):547–576.

Luttmer, E. G. (2011). On the mechanics of firm growth. *The Review of Economic Studies*, 78(3):1042–1068.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Clarendon Press.

Reed, W. J. (2001). The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19.

Simon, H. A. and Bonini, C. P. (1958). The size distribution of business firms. *The American economic review*, 48(4):607–617.

Sornette, D. (2006). *Critical phenomena in natural sciences: chaos, fractals, selforganization and disorder: concepts and tools*. Springer Science & Business Media.

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

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: journal of the Econometric Society*, pages 307–333.

Wold, H. O. and Whittle, P. (1957). A model explaining the pareto distribution of wealth. *Econometrica, Journal of the Econometric Society*, pages 591–595.