

Article

Heavy-Tailed Probability Distributions: Some Examples of Their Appearance

Lev B. Klebanov ^{1,*}, Yulia V. Kuvaeva-Gudoshnikova ² and Svetlozar T. Rachev ³¹ Department of Probability and Mathematical Statistics, Charles University, 186 75 Prague, Czech Republic² Department of Finance, Money Circulation and Credit, Ural State University of Economics, 620144 Yekaterinburg, Russia; ykuvaeva1974@mail.ru³ Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA; zari.rachev@ttu.edu

* Correspondence: klebanov@karlin.mff.cuni.cz

Abstract: We provide two examples of the appearance of heavy-tailed distributions in social sciences applications. Among these distributions are the laws of Pareto and Lotka and some new ones. The examples are illustrated through the construction of suitable toy models.

Keywords: heavy-tailed distributions; Pareto's law; Lotka's law; Zipf's law; probability-generating function

MSC: 60E05; 62A01; 60A99; 62P25



Citation: Klebanov, L.B.; Kuvaeva-Gudoshnikova, Y.V.; Rachev, S.T. Heavy-Tailed Probability Distributions: Some Examples of Their Appearance. *Mathematics* **2023**, *11*, 3094. <https://doi.org/10.3390/math11143094>

Academic Editors: Alexander Zeifman, Victor Korolev and Alexander Sipin

Received: 16 May 2023

Revised: 8 July 2023

Accepted: 12 July 2023

Published: 13 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. History of the Considered Problems

Distributions with heavy (power-like) tails have been used in the social sciences for more than one hundred years.

The relevant studies include the following:

1. The distribution of big capital was considered by Pareto in 1896 (see [1]). The density is $p(x) = \frac{\alpha}{x_0} \left(\frac{x_0}{x}\right)^{\alpha+1}$ for $x \geq x_0, \alpha > 0$.
2. Scientific production, that is, the number of scientists who published one paper, two papers, etc., (the number of papers x published by a scientist) has also been studied: Ref. [2] showed that $n(x) = n_1/x^a$, where $n_1 > 0, a \leq 2$ (in many cases, a is close to 2).
3. Lotka's law approximately holds for the number of citations of a paper by a scientist.
4. For a specific artistic text, the sequence of all words is written in descending order according to the frequency of their occurrence. Comparing the frequency of the word (x) and its place in this sequence (rank r) leads to $x = B/r$, where $B = const$ (see [3]).

Why do these patterns emerge? Probably, Laws 1–3 refer to individual human abilities, while Law 4 refers to the memory or other functions of the human brain.

We will not consider the fourth law in this paper and will focus on Laws 1 and 3; more precisely, we will focus on their qualitative explanation. This is because Zipf explained his law based on the least effort principle. Although there are no rigorous results on the existence of a mechanism related to this principle in the human brain, not wasting memory seems natural. However, the application of the least effort principle in Laws 1–3 does not seem to be related to the essence of the issues under consideration.

At first glance, everything looks quite simple. The population of a country is heterogeneous. There are people more capable in business (Law 1) or scientific work (Laws 2 and 3) and people who are not (or are less) capable of such activities.

However, how large are these differences in ability, and are all differences in “success” determined by ability?

Does chance have an effect? First, let us focus on the first law. Let us try to build a model that explains the reason for its occurrence.

The distributions of income and capital are subject to many factors that are not fully accounted for. Our interest is not in the whole mechanism of the accumulation and distribution of capital; instead, we are only interested in the roles of human talent and chance in this process. How essential are these roles? Therefore, we have to use a toy model that assumes that all people have identical abilities. If the role of chance is small, then there will not be many variations in the model between different investors. On the other hand, if we see large differences between investors, this will indicate that chance plays a significant role.

As previously noted, we want to provide examples of the possible occurrence of distributions with power tails in connection with classical empirical facts. The presentation of modern results related to the use of such distributions is not within the scope of the problems considered here. The reader interested in studying the modern use of heavy-tailed distributions in financial problems is referred to [4] and the literature cited therein.

2. A Toy Model for the Distribution of Capital

Let us consider the first toy model of the distribution of capital leading to Pareto’s law.

Suppose for simplicity that there exists only one business. All possible investors are equal in terms of their talent and initial capital. Consider the case in which each investor invests one unit of capital in the business. After one unit of time, the business outcome is X_1 , where X_1 is a random variable. Suppose that the investor leaves this entire sum in the business, and the conditions of the market remain the same during the following time interval. Then, the outcome after the second time interval is $X_1 \cdot X_2$, where X_1 and X_2 are independent identically distributed (i.i.d.) random variables. In the same way, the outcome after the n -th time interval is $\prod_{j=1}^n X_j$, where X_1, X_2, \dots, X_n are i.i.d. random variables. Let us suppose that the conditions of the market will change radically at a random moment ν_p so that investing in the business becomes unprofitable. Therefore, the final outcome is $\prod_{j=1}^{\nu_p} X_j$. We are interested in the outcome behavior for large values of ν_p . More precisely, we suppose the following:

1. $\mathbf{X} = \{X_1, X_2, \dots, X_n, \dots\}$ is a sequence of i.i.d. positive random variables, and $a = \mathbb{E} \log X_1$.
2. $\nu = \{\nu_p, p \in \Delta \subset (0, 1)\}$ is a family of positive-integer-valued random variables independent of the sequence \mathbf{X} , and $\mathbb{E}\nu_p = 1/p$.

Generally, no information on the ν -family is available. We shall consider a few cases, starting with a simple one:

3. $\mathbb{P}\{\nu_p = k\} = p \cdot (1 - p)^{k-1}, k = 1, 2, \dots$, i.e., ν_p has a geometric distribution.

Define $Z_p = \prod_{j=1}^{\nu_p} X_j^p$.

Theorem 1. *Suppose that cases 1–3 hold. Let $a \neq 0$. Then,*

$$\lim_{p \rightarrow 0} \mathbb{P}\{Z_p < x\} = 1 - x^{-1/a}, \text{ for } x \geq 1, a > 0,$$

and

$$\lim_{p \rightarrow 0} \mathbb{P}\{Z_p < x\} = x^{1/a}, \text{ for } x \leq 1, a < 0.$$

In the case of $a > 0$ (a profitable business), we have a Pareto distribution, which Pareto proposed on the basis of an empirical study (see [1]). For the proof of Theorem 1, see [5]. In [5], this result is obtained for $a = 0$. For this case, Z_p must be changed to $Z'_p = \prod_{j=1}^{\nu_p} X_j^{\sqrt{p}}$. Under the condition of the existence of the logarithmic second moment of X_1 , the product Z'_p converges in distribution to a mixture of the distributions provided in Theorem 1. It is well-known that the Pareto distribution has heavy tails. This implies that capital belongs to

a relatively small number of people. Now, we see that the Pareto distribution appears in a very natural way, described as a limit distribution for the product of a random number ν_p of random variables X_j . The value of ν_p , $p \in (0, 1)$, in case 3 has a geometric distribution. What will happen with other (“natural”) distributions? Below, we consider two additional cases leading to the existence of analogs of stable distributions (see [6]):

4. ν_p has a probability-generating function

$$\mathcal{P}(z, p, m) = \frac{p^{1/m}z}{(1 - (1 - p)z^m)^{1/m}}, \quad p \in (0, 1), \quad m \in \mathbb{N}.$$

5. ν_p has a probability-generating function

$$\mathcal{P}(z, n) = \frac{1}{T_n(1/z)},$$

where $T_n(u)$ is a Chebyshev polynomial of the first kind and $n = 1/\sqrt{p}$ is its degree. $\mathbb{E}\nu_p = 1/p$.

Let us consider case 4. The following result holds.

Theorem 2. *Suppose that cases 1, 2, and 4 hold. Let $a \neq 0$. Then,*

$$\lim_{p \rightarrow 0} \mathbb{P}\{Z_p < x\} = \int_1^x \frac{1}{b^{1/m}\Gamma(1/m)u^{1+1/m} \log^{1-1/m}(u)} du, \quad \text{for } x \geq 1,$$

where $b > 0$ is a parameter.

Proof. Consider $\log Z_p = p \sum_{j=1}^{\nu_p} Y_j$, where $Y_j = \log X_j$. From the result of [7], it follows that the limit distribution of $\log Z_p$ as $p \rightarrow 0$ has the density $\exp\{-u/b\} / (u^{1-1/m}b^{1/m}\Gamma(1/m))$, $u > 0$. Now, it is sufficient to pass to the limit distribution of Z_p from its logarithmic density. \square

Theorem 3. *Suppose that cases 1, 2, and 5 hold. Let $a = 0$ and suppose that the second logarithmic moment of X_1 exists. Then,*

$$\lim_{p \rightarrow 0} \mathbb{P}\{Z'_p < x\} = \frac{2}{\pi} \arctan(x^b), \quad \text{for } x > 0,$$

where $b > 0$ is a parameter.

Proof. Similarly to the proof of the previous theorem, we have to pass from Z'_p to its logarithm, apply the corresponding result from [6], and go back to the limit distribution for the initial random variables. \square

None of the three models constructed above take into account the abilities of the people investing in the given enterprise, but they lead to heavy-tailed distributions. The investors differ only in terms of the occurrence of some unfavorable event (the moment ν_p). An objection is that this moment is the same for the whole store; i.e., it is insolvent for all investors at once because the investors invested in the business at different times. Therefore, the period for which the investment was made is different for each investor. Thus, we see that the dependence on the moment and the case are really very high. We do not deny that the dependence on the talent of the investor is indeed significant, but it would be very difficult to separate this component from random factors.

Here and later, we provide only the simplest examples of schemes leading to heavy-tailed distributions. Namely, the case of independent identically distributed random variables is considered. The reason for this is to show that the occurrence of heavy tails can be associated exclusively with randomness. Involving dependent variables can lead to

the incorrect opinion that certain types of dependence are necessary for the appearance of such distributions. However, suppose that the reader wishes to see more general results. In that case, it suffices to note that results on limiting distributions for products of random variables can be obtained by applying the corresponding theorems for the sums of the logarithms of the original variables. The corresponding theory of the summation of a random number of random variables is quite well-developed (see [5,8]).

It is well-known that limit theorems are examples of ill-posed problems. This may raise doubts about the practical significance of the conclusions based on these theorems. So-called pre-limit theorems allow us to justify these conclusions in many situations (see [5]).

3. Distribution of the Number of Citations

A similar situation occurs when the distribution of the number of citations of scientific publications is studied. Let us make some assumptions.

Assumption 1. *All scientists under consideration are equal in terms of their scientific and literary abilities.*

Assumption 2. *The citations of a paper occur independently.*

Assumption 3. *The probability that an article will be repeatedly cited depends on the number of previous citations. This probability is increasing in the number of citations. More precisely, the probability that an article with $k - 1 (k \geq 1)$ citations will have no further citations is*

$$p_k = \frac{a}{(k + b)}, \tag{1}$$

where $a > 0$ and $b > a - 1$.

Let Y be a random variable describing the number of citations during the considered period. Assumption 1 implicitly de facto implies that Y has the same distribution for different papers because the scientific abilities of the authors are supposed to be the same.

In view of the independence of the citations, the probability that a paper is cited exactly n times is

$$\mathbb{P}\{Y = n\} = p_n \prod_{k=1}^{n-1} (1 - p_k) = \frac{\left(\frac{a+b-1}{a}\right)_{n-1}}{(a n + b) \left(\frac{a+b}{a}\right)_{n-1}},$$

where $(a)_n = a(a + 1) \dots (a + n - 1)$ is the Pochhammer symbol.

It is not difficult to calculate this probability:

$$\mathbb{P}\{Y \geq m\} = \frac{\left(\frac{a+b-1}{a}\right)_{m-1}}{\left(\frac{a+b}{a}\right)_{m-1}} \underset{m \rightarrow \infty}{\sim} \frac{\Gamma((a + b)/a)}{\Gamma((a + b - 1)/a)} \frac{1}{m^{1/a}}. \tag{2}$$

The relation (2) shows that the distribution of the number of citations has a heavy tail, the severity of which depends on the value of the parameter a responsible for the degree of influence of previous citations. Therefore, a larger value of a corresponds to a heavier tail. In any case, the presence of such a tail makes it possible to conclude that the citation intensities of almost identical scientists can differ significantly, which leads to a significant stratification of the scientific community due to various random circumstances that have nothing to do with research abilities. Thus, the number of citations seems meaningless as an indicator of scientific value.

Comments on Assumption 3. At first glance, the relation (1) does not seem to be too natural. However, it seems almost unique asymptotically, leading to a heavy-tailed

distribution. We will consider this in more detail but without complete proofs (obtaining general mathematical results is not an aim of this paper).

Let Y be the (random) number of citations of a paper. Suppose that the distribution of Y has a power tail. In other words,

$$\mathbb{P}\{Y \geq n\} = \frac{C}{n^\alpha}(1 + \varkappa(n)), \tag{3}$$

where $\varkappa(n) \xrightarrow[n \rightarrow \infty]{} 0$ and has “regular” behavior in a sense. The symbol C is used for constants, which are possibly different. From (3), it follows that

$$\begin{aligned} \mathbb{P}\{Y = n\} &= \mathbb{P}\{Y \geq n\} - \mathbb{P}\{Y \geq n + 1\} \\ &= \frac{C}{n^\alpha} \left(1 - \left(\frac{n}{n+1}\right)^\alpha\right) + \frac{C\varkappa(n)}{n^\alpha} \left(1 - \frac{\varkappa(n+1)}{\varkappa(n)} \left(\frac{n}{n+1}\right)^\alpha\right). \end{aligned} \tag{4}$$

Suppose that $\frac{\varkappa(n+1)}{\varkappa(n)}$ is bounded from above. Then, the equality (4) implies that

$$\mathbb{P}\{Y = n\} = \frac{\alpha C}{n^{\alpha+1}}(1 + \varkappa_1(n)), \tag{5}$$

where $\varkappa_1(n)$ possesses the same properties as $\varkappa(n)$.

If

$$\mathbb{P}\{Y = n\} = p_n \prod_{k=1}^{n-1} p_k, \tag{6}$$

where p_k is the probability of the termination of citations, then

$$\mathbb{P}\{Y \geq n\} = \prod_{k=1}^{n-1} (1 - p_k).$$

Under some restrictions on the behavior of p_k as $k \rightarrow \infty$, we have

$$\prod_{k=1}^{n-1} (1 - p_k) \sim \exp\left\{-\sum_{k=1}^{n-1} p_k\right\}.$$

The symbol \sim is used here for asymptotic equivalence as $n \rightarrow \infty$. Therefore, from (3), we must have

$$\exp\left\{-\sum_{k=1}^{n-1} p_k\right\} \sim \frac{C}{n^\alpha} \text{ as } n \rightarrow \infty.$$

Taking the logarithm of both sides of the last relation yields

$$\sum_{k=1}^{n-1} p_k \sim \alpha \log(n - 1)$$

and

$$p_n \sim \alpha \log(n - 1) - \alpha \log(n - 2) \sim \frac{\alpha}{n}.$$

It is clear that Assumption 3 leads to the same asymptotic behavior. However, the presence of the parameter b may make the asymptotics more precise if we fix not only the tail index α but also the corresponding constant C in (3).

There remains the question of *how many distributions may be represented in the form (6)*. Suppose that Y is a random variable taking positive integer values such that $\mathbb{P}\{Y = n\} > 0$

for any $n \in \mathbb{N}$. Then, there are probabilities p_n such that (6) holds. Indeed, we can write $\kappa_n = \mathbb{P}\{Y = n\}$ and

$$p_n = \frac{\kappa_n}{1 - \sum_{k=1}^{n-1} \kappa_k}.$$

Then, (6) holds.

Note that p_n represents the intensity rate for the distribution of Y .

From the considerations provided above, it follows that, under mild restrictions, the distribution of a positive integer random variable possessing power tails has a representation (6), with p_k being asymptotically equivalent to that of (1). The indicated method of the occurrence of heavy-tailed distributions on the set of positive integers turns out to be quite universal and probably can be applied to some classes of applied problems.

We now make some remarks on the distribution of the impact factor.

Let us now consider the possibility of using the impact factor of a journal as an indicator of the scientific significance of a paper published in it. The impact factor of a journal is calculated as the ratio of the number of citations of papers published over a certain period to the number of these papers. The idea of considering such an average value is connected with the idea that, according to the law of large numbers, the influence of chance will be leveled. However, we shall show that this is not true.

We mention that there exists a rather large body of literature stating that a scientific journal's impact factor has essential value. Based on the observed data, the presence of asymmetry in the distribution of the impact factor and the presence of a heavy tail have been noted. However, these circumstances have not been analyzed from a theoretical point of view; comments have only been made on the advisability of replacing the arithmetic mean with other statistics for the purpose of statistical data analysis. We note one of the typical works of this kind: [9]. True, the author notes the similarity of the distribution of some of the data with the Pareto distribution, but a mathematical analysis of the reasons for this is not carried out. In addition, the mathematically strict definition of a distribution is not considered; only its "naive" form is examined. Below, we will try to clarify the appearance of heavy tails in the impact factor distribution.

We assume that the number of papers submitted to the journal has a Poisson distribution. For simplicity, let us assume that the number of citations for each of the submitted papers has a Sibuya distribution. Then, the citation distribution for all papers has a probability-generating function that is a superposition of the generating functions of the Sibuya and Poisson laws. The probability-generating function of this superposition is $\mathcal{P}(z) = e^{-\lambda(1-z)^p}$ for a fixed $\lambda > 0$ and $p \in (0, 1)$. Clearly, this distribution has a heavy tail with an index p . In view of the fact that $p < 1$, the law of large numbers is inapplicable in this situation. Moreover, in this case, the impact factor increases with the number of publications without increasing their scientific significance. The observed increase (over time) in the impact factors of leading journals confirms this circumstance.

Now, we can conclude that the impact factor distribution has a heavy tail and cannot be used as an indicator of scientific significance.

4. Conclusions

1. It is shown that distributions with heavy tails can arise in some manifestations of social inequality (the distribution of capital, the number of citations, or the impact factor) due to purely random reasons. In this case, the spread in the magnitude of inequality is significant.
2. The circumstance specified in the previous item makes it impossible to use indices such as the number of citations and/or the impact factor of a journal as an indicator of the scientific significance (scientific quality) of a published work.
3. We do not need any proof of the existence of heavy tails for the distributions under consideration. Their presence follows from the previously mentioned papers that Lotka, Pareto, and Zipf published many years ago, which have withstood the test of time.

Author Contributions: Conceptualization, L.B.K.; Investigation, L.B.K., Y.V.K.-G. and S.T.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pareto, V. *Cours d'Économie Politique: Nouvelle édition par G.-H. Bousquet et G. Busino*; Librairie Droz: Geneva, Switzerland, 1964; pp. 299–345.
2. Lotka, A.J. The frequency distribution of scientific productivity. *J. Wash. Acad. Sci.* **1926**, *16*, 317–324.
3. Zipf, G.K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley: Cambridge, UK, 1949.
4. Lindquist, W.B.; Rachev, S.T.; Hu, Y.; Shirvani, A. *Advanced REIT Portfolio Optimization. Innovative Tools for Risk Management*; Springer: New York, NY, USA, 2022.
5. Klebanov, L.B.; Kozubowski, T.J.; Rachev, S.T. *Ill-Posed Problems in Probability and Stability of Random Sums*; Nova Science Publishers, Inc.: New York, NY, USA, 2006.
6. Klebanov, L.B.; Kakosyan, A.V.; Rachev, S.T.; Temnov, G. On a class of distributions stable under random summations. *J. Appl. Probab.* **2012**, *49*, 303–318. [[CrossRef](#)]
7. Melamed, J.A. Limit theorems in the set-up of summation of a random number of independent and identically distributed random variables. In *Stability Problems for Stochastic Models*; Lecture Notes in Mathematics; Springer: New York, NY, USA, 1989; Volume 1412, pp. 194–228.
8. Gnedenko, B.V.; Korolev, V.Y. *Random Summation. Limit Theorems and Applications*; CRC Press: Boca Raton, FL, USA, 1996.
9. Blanford, C.F. Impact factors, citation distributions and journal stratification. *J. Mater. Sci.* **2016**, *51*, 10319–10322. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.