Article

Not peer-reviewed version

Heavy-Tailed Probability Distributions: Some Examples of Their Appearance

Lev b. Klebanov *, Yulia V. Kuvaeva , Svetlozar T Rachev

Posted Date: 17 May 2023

doi: 10.20944/preprints202305.1198.v1

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

Article

Heavy-Tailed Probability Distributions: Some Examples of Their Appearance

Lev B. Klebanov *, Yulia V. Kuvaeva † and Svetlozar T. Rachev ‡

* Department of Probability and Mathematical Statistics, Charles University, Prague, 186 75, Czech Republic; e-mail: klebanov@karlin.mff.cuni.cz

† Department of Finance, Money Circulation and Credit, Ural State University of Economics, 620144 Yekaterinburg, Russia; ykuvaeva1974@mail.ru

‡ Department of Mathematics and Statistics, Texas Tech University; zari.rachev@ttu.edu

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

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

---

1. History of the problems

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

Relevant studies include the following:

1. Distribution of big capital. Pareto, 1896 (see [7])). The density is $p(x) = \frac{\alpha}{x_o}\left(\frac{x_o}{x}\right)^{\alpha+1}$, for $x \geq x_o$, $\alpha > 0$.
2. Scientific production. The number of scientists who published one, two and so on x papers (the number x published by scientist papers). Lotka (1926) (see [5]) 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 and the place in this sequence (rank) leads to $x = B/r$, $B = const$ (see [8]).

Why do these patterns emerge? Probably, Laws 1–3 refer to some 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 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 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 (for Law 1) or scientific work (for Laws 2 and 3) and people who are not (or are less capable) of such activities.

But how big are differences in ability, and are all differences in 'success' determined by ability?

Is there an effect of chance? First, let's focus on the first law. Let's try to build a model that explains the reason for its occurrence.

However, the distributions of income and capital are subject to many factors not fully accounted for. Our interest is not in the whole mechanism of accumulation and distribution of capital but only in the roles of human talent and chance in this process. How essential are these roles? Therefore, we have to use a toy model which assumes all people have identical abilities. If the role of chance is small, then

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2023 doi:10.20944/preprints202305.1198.v1

there will not be many variations in the model between different investors. On the other hand, if we see a large difference between investors, this will indicate a significant role of chance.

As noted, we want to give 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 does not belong to 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 there.

2. A toy model for the distribution of capital

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

Suppose for simplicity that there exists only one business. All possible investors are equal in their talents and initial capital. Consider the case when 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 the investor leaves all this sum in the business, and the conditions on 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 $\prod_{i=1}^{n} X_j$, where $X_1, X_2, \ldots, X_n$ are i.i.d. random variables. Let us suppose that the conditions on the market will change radically at a random moment ν_p so that investing in that business becomes not profitable. 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 that

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

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

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

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

Theorem 2.1. *Suppose that 1–3 hold. Let $a \neq 0$. Then*

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

and

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

In the case of $a > 0$ (a profitable business), we have a Pareto distribution, which Pareto had proposed on the basis of empirical study (see [7]). For the proof of Theorem 2.1 see [2]. In [2], 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 given in Theorem 2.1. 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 the Pareto distribution appears in a very natural way, described as a limit distribution for a product of a random number ν_p of random variables X_j. The value of ν_p, $p \in (0,1)$ in 3 had a geometric distribution. What will happen with other ('natural') distributions? Below we consider two additional cases:

4. ν_p has a probability generating function

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2023 doi:10.20944/preprints202305.1198.v1

5. ν_p has a probability generating function

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

where $T_n(u)$ is 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 2.2. *Suppose that the 1, 2, and 4 hold. Let $a \neq 0$. Then*

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

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 [6] it follows that the limit distribution of $\log Z_p$ as $p \to 0$ has the density $\exp\{-u/b\}/\left(u^{1-1/m}b^{1/m}\Gamma(1/m)\right)$, $u > 0$. Now it is sufficient to pass to the limit distribution of Z_p from its logarithmic density. □

Theorem 2.3. *Suppose that 1, 2, and 5 hold. Let $a = 0$ and suppose that the second logarithmic moment of X_1 exists. Then*

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

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 [3], and go back to the limit distribution for the initial random variables. □

None of the three models constructed above take into account any abilities of the people investing in the given enterprise, but lead to heavy-tailed distributions. The difference between investors is only in the occurrence of some unfavorable event for them (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. So 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.

3. Distribution of the number of citations

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

Assumption 1.
All scientists under consideration are equal in 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. It is increasing in the number of citations. More precisely,
Assuming the probability that an article having $k-1 (k \geq 1)$ citations will have no further citations is

$$p_k = \frac{a}{(k+b)}, \tag{3.1}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2023 doi:10.20944/preprints202305.1198.v1

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

$$\mathbb{P}\{Y \geq m\} = \frac{\left(\frac{a+b-1}{a}\right)_{m-1}}{\left(\frac{a+b}{a}\right)_{m-1}} \underset{m\to\infty}{\sim} \frac{\Gamma((a+b)/a)}{\Gamma((a+b-1)/a)} \frac{1}{m^{1/a}}\,. \tag{3.2}$$

The distribution of the number of citations.

The relation (3.2) 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 intensity of almost identical scientists can differ significantly, which leads to a significant stratification of the scientific community through 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 (3.1) seems to be not 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,

$$\mathbb{P}\{Y \geq n\} = \frac{C}{n^\alpha}(1 + \varkappa(n)), \tag{3.3}$$

where $\varkappa(n) \xrightarrow[n\to\infty]{} 0$ and has "regular" behavior in a sense. The symbol C is used for constants, possibly different. From (3.3) it follws that

$$\mathbb{P}\{Y = n\} = \mathbb{P}\{Y \geq n\} - \mathbb{P}\{Y \geq n+1\} =$$

$$= \frac{C}{n^\alpha}\left(1 - \left(\frac{n}{n+1}\right)^\alpha\right) + \frac{C\varkappa(n)}{n^\alpha}\left(1 - \frac{\varkappa(n+1)}{\varkappa(n)}\left(\frac{n}{n+1}\right)^\alpha\right). \tag{3.4}$$

Suppose that $\frac{\varkappa(n+1)}{\varkappa(n)}$ is bounded from above. Then the equality (3.4) implies that

$$\mathbb{P}\{Y = n\} = \frac{\alpha C}{n^{\alpha+1}}\Big(1 + \varkappa_1(n)\Big), \tag{3.5}$$

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

If

$$\mathbb{P}\{Y = n\} = p_n \prod_{k-1}^{n-1} p_k, \tag{3.6}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 May 2023 doi:10.20944/preprints202305.1198.v1

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 $k \to \infty$, we have

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

The symbol $\sim$ is used here for asymptotic equivalence as $n \to \infty$. Therefore, from (3.3) we must have

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

Taking logarithms of the 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 the corresponding constant C in (3.3).

There remains the question of *how many distributions may be represented in the form (3.6)*? Suppose that Y is a random variable taking positive integer values and such that $\mathbb{P}\{Y = n\} > 0$ for any $n \in \mathbb{N}$. Then there are probabilities p_n such that (3.6) holds. Indeed, write $\kappa_n = \mathbb{P}\{Y = n\}$ and

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

Then (3.6) holds.

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

From the considerations given above it follows that, under mild restrictions, the distribution of a positive integer random variable possessing power tails has a representation (3.6) with p_k asymptotically equivalent to that of (3.1). 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 for considerations of some classes of applied problems.

We now make some remarks on the Impact Factor distribution.

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 themselves. 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, this is not true.

We mention that there exists a rather large literature stating 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 has been noted. However, these circumstances have not been analyzed from a theoretical point of view, and only comments are made on the advisability of replacing the arithmetic mean with some other statistics for the purpose of statistical data analysis. We

 doi:10.20944/preprints202305.1198.v1

note one of the typical works of this kind: [1]. True, the author notes the similarity of the distribution of some 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, but only its 'naive' form. Below we will try to clarify the appearance of heavy tails of 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 fixed $\lambda > 0$ and $p \in (0,1)$. Clearly, this distribution has a heavy tail with 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 again and cannot be used as an indicator of scientific significance.

4. Conclusions

I. It is shown that distributions with heavy tails can arise in some manifestations of social inequality (the distribution of capital, the number of citations, the impact factor) due to purely random reasons. In this case, the spread in the magnitude of inequality is significant.
II. The circumstance specified in 1 makes it impossible to use such indices 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.
III. We do not need any proof of the existence of heavy tails for the distributions under consideration. Their presence follows from the mentioned papers by Lotka, Pareto, and Zipf published many years ago and has withstood the test of time.

References

1. Blanford C. F. (2016) Impact factors, citation distributions and journal stratification. Journal of Materials Science volume 51, 10319–10322.
2. Klebanov L.B., Melamed J.A., Rachev S.T. (1987) On the products of a random number of random variables in connection with a problem from mathematical economics. In: Stability Problems for Stochastic Models, Lecture Notes in Mathematics, 1412, 103–109.
3. Klebanov L.B., Kakosyan A.V., Rachev S.T., Temnov G. (2012) On a class of distributions stable under random summations. Journal of Applied Probability, 49, 303–318.
4. Lindquist W.B., Rachev S. T., Hu Y., Shirvani A. (2022) Advanced REIT Portfolio Optimization. Innovative Tools for Risk Management, Springer.
5. Lotka A. J. (1926). "The frequency distribution of scientific productivity". Journal of the Washington Academy of Sciences. 16 (12): 317–324.
6. Melamed, J.A. (1989). 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, 1412, 194–228.
7. Pareto V. (1964) Cours d'Économie Politique: Nouvelle édition par G.-H. Bousquet et G. Busino, Librairie Droz, Geneva, pp. 299–345.
8. Zipf G.K. (1949) Human Behavior and the Principle of Least Effort. Cambridge. Addison–Wesley.