doi:10.20944/preprints202204.0200.v1

Do Not Use I.F. for Scientific Value Estimation

Lev B. Klebanov,[a] Yu.V. Kuvaeva,[b] V.E. Volkovich[c]

Abstract

A toy model for the distribution of the impact factor (IF) of a journal is proposed. The model demonstrates the presence of a heavy tail for the IF distribution, which occurs due to some random non-scientific circumstances. Therefore, the use of IF as an indicator of the quality of scientific papers seems to be inadequate.

Key words: citation distribution; impact factor distribution; essential randomness; Sibuya distribution.

1 Introduction

The number of scientific citations and the influence of the IF of the respective journals is often recognized in the scientific community, especially by the relevant administrative bodies, as important indicators of scientific value. Such an approach is not based on a formal model or statistical assessment. It is only a consequence of a priori and one of a widely held opinion that seems to be consistent with common sense, but no more. However, the overall citation is affected by a vital and varied process consisting of the scientific study itself, the publication procedure, and many internal and external random factors that are usually overlooked for no apparent reason. More precisely, it is indirectly assumed that a random cause cannot significantly affect the citation process. It is assumed that the citation number mainly depends on

[a] Department of Probability and Mathematical Statistics, Charles University, Prague, Czech Republic. e-mail: levbkl@gmail.com

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

[c] Software Engineering Department, ORT Braude College, Karmiel 21982, Israel; vlvolkov@braude.ac.il

the scientific value. However, it is doubtful such a view can be considered suitable for assessing scientific results and advancing researchers. A detailed consideration of the scientific value involves processes influencing all significant circumstances and involves a large amount of statistical information, often inaccessible. However, the inconsistency of the intuitive approach can be shown by the construction of the "toy" model, including the most important factors.

The proposed paper presents a model for the role of randomness in the process of citing, publishing, and reviewing, leading to the fact that equivalent articles may have completely different numbers of citations.

The suggested outlook pursues to cover the main trends and properties of the mentioned practice to demonstrate its most significant drawbacks without claiming to be an all-encompassing description. Formally speaking, this inherent process ambiguity is expressed by a heavy tail of the citations distribution. As a result, the accepted publications' indications cannot precisely distinguish between different levels of scientific activity and the research significance. Let us emphasize such a substantial stratification may form between almost scientifically equal researchers. Obviously, it is not generally suggested that most scientists have the same qualifications and abilities, but any subgroup consisting of “equal" ones would be ranked resting upon sufficiently random causes. Again, we would like to emphasize that a seemingly significant stratification in the citation numbers may occur between scientifically equal researchers.

Many authors consider various distributions of papers citations (see, for example, [4, 8]). Almost all such studies conclude that suitable distributions have heavy power tails preventing the existence of finite variance or mean values. It is exhibited, for example, by arising in the citations modeling through the celebrated Lotka's law [2] (see also [4]), which being in good agreement with the actual data provides no idea justifying the heavy tails presence.

The rest of the test is organized as follows. Section 2 is devoted to the citation distribution model. Section 3 discusses the distribution of IF, including the peer review process, and Section 4 provides the relevant empirical data and discusses their fit with the constructed model.

2 Citations distributions

Our first supposition consists of the equivalence of all scientists under consideration:

Assumption 1.

All scientists under consideration are equal in their scientific and literary abilities.

Assumption 2.

The citations of a paper occur independently.

Each publication produces a specific citations amount. Of course, the likelihood of an article being repeatedly quoted would depend on the number of previous citations. This fact is formalized in the framework of our model as follows:

Assumption 3.

Assume the probability that an article having $k-1$ ($k \geq 1$) citations will not have any further quote is

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

where $a > 0$ and $b \geq 0$, are real numbers under the condition $a+b > 1$. Let Y be a random variable describing the number of citations of a paper during the considered period. We suggest that the random variable Y has the same distribution for different papers because the authors are supposed to be equal in their scientific abilities. Assuming independence of the quotes, the likelihood for an article to be quoted 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 that the citations number will be greater or equal than a given $m \geq 1$ has the probability equals to

$$\mathbb{P}\{Y \geq m\} = \sum_{n=m}^{\infty} \mathbb{P}\{Y=n\} = \frac{\left(\frac{a+b-1}{a}\right)_{m-1}}{\left(\frac{a+b}{a}\right)_{m-1}}. \tag{2.1}$$

Thus

$$\mathbb{P}\{Y \geq m\} \sim \frac{\Gamma\left(\frac{a+b}{a}\right)}{\Gamma\left(\frac{a+b-1}{a}\right)} \cdot \frac{1}{m^{1/a}} \quad \text{as} \quad m \to \infty. \tag{2.2}$$

The sign $\sim$ here denotes, as usual, the asymptotic equivalence, and $\Gamma(z)$ is the Euler Gamma function.

The relation (2.2) shows that the citation distribution has a power-like tail. Its severity depends on the value of the a parameter, which is responsible for the influence of previous citations. Therefore, the larger is the a value, the heavier the tail. In any case, the presence of such a tail allows us to conclude that the level of citation of almost identical scientists can vary significantly. This shows a significant stratification of the scientific community may be based on random circumstances that have nothing to do with research abilities. Thus, the citation number as an indicator of the scientific value of a paper seems meaningless.

It is necessary to note that the distributions involved in the current research appear in another context in [9].

3 Distributions of scientific significance indicators

The proposed scientific significance indicators distributions are composed of several components. We base their description on the stated earlier distribution of the number Y of published paper citations presented in (2.1) restricted to the case when $b = 0$ and $a > 1$. This inconsequential loss of generality allows providing closed analytical representations well illustrating the methodology. Thus

$$\mathbb{P}\{Y = n\} = \frac{p}{n} \cdot \prod_{k=1}^{n-1} \big(1 - p/k\big);\ p = 1/a \in (0, 1). \tag{3.1}$$

The generating function of this law is

$$\mathcal{P}(z) = \mathbb{E}z^Y = 1 - (1 - z)^p. \tag{3.2}$$

3.1 Impact Factor distribution

As was demonstrated in Section 2, the collection of the author's citations would not provide sufficient information on the scientific value of his/her publications. A natural question arises in this connection: How is it reasonable or feasible to use Impact Factor (IF) of journals instead, taking into account the following matters (“hops"):

A. *IF appears to be a mean-type attitude, which is generally more stable (i.e. less dependent on randomness) than an individual indicator like the number of publications and thus can provide an appropriate approach to scientific rating.*

B. *The papers published in journals with high IF were peer-reviewed. Hopefully, the reviewing process allows to reject the results without essential scientific value.*

Let us consider both items A. and B. in details.

Remind the definition of IF. Namely, IF for a given year y is the ratio of the number of citations received in this year to the number of papers issued during the two preceding years in the journal:

$$IF_y = \frac{\text{Citations}_{y-1} + \text{Citations}_{y-2}}{\text{Publications}_{y-1} + \text{Publications}_{y-2}}.$$

It is possible to consider this indicator in any desired period. For example, the Journal Citation Report *(JCR)* involves a five-year impact factor similarly calculated within a five-years stage [3]. This value is commonly used as the published articles quality metric, and, presumably, it should not be directly related to the number of the issued articles. In this consideration, for the sake of simplicity, we assume that the number of papers published in a journal for a given period is a constant m. It is clear that in this case the IF is a random variable coinciding with the average

$$\bar{Y}_m = (Y_1 + \ldots + Y_m)/m. \tag{3.3}$$

Random variable $\bar{Y}_m$ is no longer integer–valued. Therefore, to study the asymptotic behavior of its distribution we have to pass to its Laplace transform. Because of independence and identically distribution property of the numbers of citations, the Laplace transform of $\bar{Y}_m$ has, taking into account (3.2), the following form

$$\varphi(s) = \Big(1 - \big(1 - \exp\{-s/m\}\big)^p\Big)^m,$$

doi:10.20944/preprints202204.0200.v1

where $p \in (0,1)$. Asymptotic behavior of this function has the following form

$$\varphi(s) \sim e^{-s^p m^{1-p}} \quad \text{for } m \to \infty. \tag{3.4}$$

In other words, $\bar{Y}_m$ growth with m as m^{1-p} multiplied by skewed to the right stable random variable with index p.

The relation (3.4) shows that:

1. IF has a heavy tail.

2. The random variable $\bar{Y}_m$ growth with growing m. Therefore, IF of a journal depends on the number of papers published in it.

Basing on these items we may conclude the “hope" A has no scientific basis. In opposite, IF does depend on the number of published papers.

3.2 Total number of papers and the total number of citations produced

Here we analyze the “hope" B. Let X_a be a number of manuscripts submitted by an author for publications. We consider all the coauthors of a manuscript as one unified “author". It is reasonable to presume the distributions of X_a and Y are the same in their inner structure, paying attention to their similar origins. Namely, the generating function of X_a is supposed to be

$$\mathcal{P}_o(z) = 1 - (1-z)^\gamma, \quad \gamma \in (0,1).$$

Another important assumption of the model is that a manuscript can be accepted (and published) with probability c and rejected with probability $1-c$. This is a somewhat simplified view of the peer review process that does not take into account many factors, such as the reputation of the article and the perception of the reviewer. On the other hand, this refinement takes into account many significant factors, such as reviewer workload, irrelevant likes, dislikes, etc. However, the basic supposition of our toy model was that all scientists have the same scientific potential. Thus, the assumption of reviewing process can be considered acceptable within the presented simplified model. Consequently, the generating function of the umber X_a of the author's published paper takes the form

$$\mathcal{P}_a(z) = 1 - c(1-z)^\gamma, c,\ \gamma \in (0,1) \tag{3.5}$$

According to the definition of the random variable Y, the distribution of the total citations of the papers of an author is the distribution superposition of $\mathcal{P}$ (see (3.2)) and $\mathcal{P}_a$ with the following generating function:

$$Q(z) = \mathcal{P}(\mathcal{P}_a(z)) = 1 - c^p(1-z)^{\gamma p}. \tag{3.6}$$

It is clear this distribution has a heavy tail with the index $\gamma p < \min(\gamma, p) < 1$.

The reviewing process apparently decreases the number of possible citations. Thus a manuscript rejection with probability $1 - c$ is equivalent to the total omitting of the citations with the probability $1 - c^p < 1 - c$. Several reviewers' situations may be considered as an instance corresponding to a random parameter c with the exchange of c^p in (3.6) by its mean value with another interpretation of the parameter c^p.

With the obvious analogy with the Queueing theory, we can suppose that the submission process is a Poisson one having the generating function

$$\mathcal{R}(z) = \exp\{-\lambda(1-z)\}.$$

In view of (3.5), the total number of papers published in the journal has probability generating function

$$\mathcal{R}(\mathcal{P}_a(z)) = \exp\{-\lambda c(1-z)^\gamma\}. \tag{3.7}$$

Accordingly, the number of citations produced by the papers has probability law with the generating function

$$\mathcal{R}(\mathcal{Q}(z)) = \exp\{-\lambda c^p(1-z)^{\gamma p}.\} \tag{3.8}$$

Thus, it can be seen that both variables have distributions with heavy tails. However, the citations distribution, as expected, is characterized by a much heavier tail. Obviously, their ratio (IF) has heavy tail again.

4 Consistency with observed data

Of course, any toy type model is constructed aspiring to emphasize and comprehend the main features and tendencies of the patterned problem. From this standpoint, it would be natural to expect in main a qualitative agreement between the modeled outcomes and the actual situation reflecting the main process trends. Such correspondence appears to reason the proposed method's suitability sufficiently.

doi:10.20944/preprints202204.0200.v1

4.1 Citation data

The Google Scholar site is a reliable source for citations emerging in different fields. An empirical verification can be provided bearing in mind the model format on a relatively modest dataset. We consider the 500 highest citations records in the Probability theory, valid on December 28, 2021.

For empirical verification, it is sufficient to use a relatively modest dataset. We consider the 500 highest citations records in the Probability theory, valid on December 28, 2021.

The maximum number of paper citations number is 257811, while the second one(110179) is more than two times smaller. The essential differences also occur between the fourth, fifth, and sixth observations witnessing a heavy tail presence (see, for example, [1]). The mean citations number is 4712.41 with a standard deviation of 14812.5. So, the standard deviation is three times bigger than the mean. *It evidently indicates the presence of a heavy tail of the citations number distribution.* Moreover, the maximal deviation from the mean value is more than 17 times greater than the standard deviation.

Let us also mention without any details that the modeling by means of a Bootstrap procedure allows estimating the tail index to be $p \approx 0.92 < 1$. (See Appendix for the details.) In other words, the tail is very heavy so the general mean value is infinite. These facts are in good coordination with the stated toy model. Of course, an analogous statistical analysis can be performed for the citations in other scientific fields with expected similar results. However, we do not need such an analysis because the presence of heavy tails is known since the paper by A. Lotka [2].

4.2 Analysis with Empirical IF data

As it was mentioned above, the presence of a heavy tail implies that IF essentially depends on the number m of the published papers. More precisely, IF growth with growing m. If so, one cannot compare different journals based on their IF without considering the numbers of papers published in each of the journals. This circumstance prevents using journals IF to estimate the scientific value of papers. Another consequence of an IF dependence on m is that IF is probably higher for a more considerable period than for the shorter one. This fact is seen while comparing the 2 and 5 years IF values presented in the following Figure 1. All the data on IF of the journals were taken from Journal Citation Reports$^{\text{TM}}$ (JCR) https://clarivate.com/

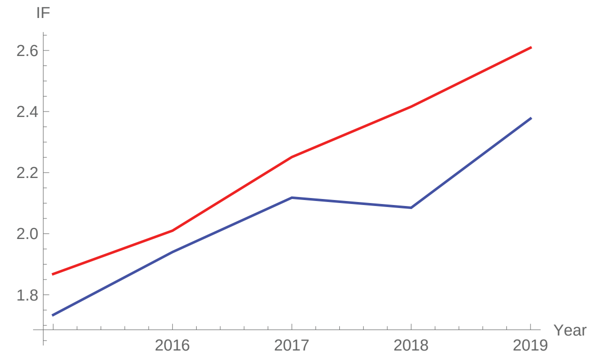

Figure 1: IF of the Annals of Probability during 2015-2019; the 5 years IF is in red, the 2 years IF is in blue.

An additional remark is that 2 (or just 5) years periods are often too short of revealing essential innovative publications citations. Therefore, the citations appearing in a short interval mostly match the papers elaborating on already proposed questions or previous ideas. To understand the genuine value of a new idea, one needs to study its connections with the previous results and consider possible future perspectives.

Let us study the relations between different types of IF and consider Fig.1, Fig. 2 and 3 containing a comparisons of two and five years IF.

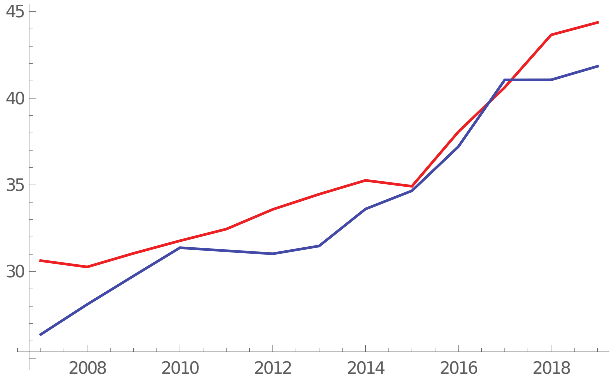

Figure 2: IF of the Science during 2015-2019; 5 years IF is in red, 2 years IF is in blue.

If the first two charts clearly show that the 5-year index dominates, then in the last one, the tendency has changed in recent years. The results presented

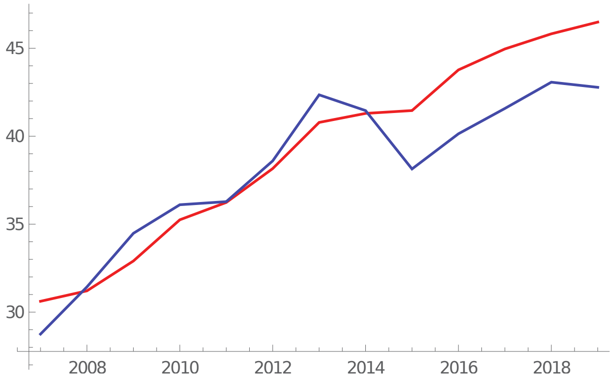

Figure 3: IF of the Nature during 2015-2019; 5 years IF is in red, 2 years IF is in blue.

in Fig. 2 (Science) are analogous to those in Fig. 3. However, the behavior of IF in Fig. 3 is different. It seems that the more recently issued information is targeted not only at long-term global objectives but also in short-term applications. This phenomenon can be partly affected by the exclusive role played by this magazine and its publishing creed.

Now, we consider the typical examples given in Fig. 2 and Fig. 3. The aim is to formally justify a reasonably evident fact that calculated over a more extended period IF is expected to be higher. Denote by IF 2y, IF 4y and IF 5y two, four, and five years IF s calculated at the moment y. By cy and 3̆bay denote the citations number and the publications number at the moment y, respectively. Then

$$IF2_{y+1} - IF4_{y+1} = \frac{c_{y_{-1}} + c_y}{\kappa_{y_{-1}} + \kappa_y} - \frac{c_{y-3} + c_{y-2} + c_{y-1} + c_y}{\kappa_{y-3} + \kappa_{y-2} + \kappa_{y-1} + \kappa_y} = \tag{4.1}$$

$$= \frac{(c_{y-1} + c_y)(\kappa_{y-3} + \kappa_{y-2}) - (c_{y-3} + c_{y-2})(\kappa_{y-1} + \kappa_y)}{(\kappa_{y-1} + \kappa_y)(\kappa_{y-3} + \kappa_{y-2} + \kappa_{y-1} + \kappa_y)}.$$

It shows that if $IF2_y$ increases (decreases) in y with step 2 then $IF4_{y+1}$ is smaller (greater) than $IF2_{y+1}$. Of course, $IF5_{y+1}$ differs from $IF4_{y+1}$. However, if c_{y-4} and κ_{y-4} are small, then the difference between $IF5_{y+1}$ and $IF4_{y+1}$ is small and the difference between $IF2$ and $IF5$ may have the same sing as that between $IF2$ and $IF4$. These considerations show $IF2$, $IF4$ and $IF5$ are strictly dependent. Decreasing of $IF2$ between 2012 and 2015 explains further increasing of $IF5$ on the Fig. 3.

doi:10.20944/preprints202204.0200.v1

he sign of the difference $IF2_{y+1} - IF4_{y+1}$ is identical to the sign of the numerator in the right hand side of (4.1). However, the later is the same as the sign of $IF2_{y+1} - IF2_{y-1}$. It shows that if $IF2_y$ increases (decreases) in y with step 2 then $IF4_{y+1}$ is smaller (greater) than $IF2_{y+1}$. Of course, $IF5_{y+1}$ differs from $IF4_{y+1}$. However, if c_{y-4} and κ_{y-4} are small, then the difference between $IF5_{y+1}$ and $IF4_{y+1}$ is small and the difference between $IF2$ and $IF5$ may have the same sing as that between $IF2$ and $IF4$. These considerations show $IF2$, $IF4$ and $IF5$ are strictly dependent. Decreasing of $IF2$ between 2012 and 2015 explains further increasing of $IF5$ on the Fig. 3.

However, we see an additional drawback of IF. As it was mentioned above, the relation (4.1) shows that the increasing of $IF2$ with the time leads to the fact that $IF4 < IF2$ at the same moment. Of course, it is not the same that $IF5 < IT2$. However, the difference between $IF5$ and $IF4$ seems to be not too large. On the other hand, this difference may be large because of the dependence of IF on the number m of published papers. Secondary, presence of a heavy tail implies that IF essentially depends on the number m of the published papers. This statement seems rather strange. Although, if the finite first moment (i.e. expected value) exits, then according to the law of large numbers, the ratio of the total number of citations to the total number of publications would be approximately equal to the mean value. More precisely, this ratio would converge to the mean value with an increase in the publications number. However, due to a heavy distribution tail, the first moment does not exist.

5 On the use of asymptotic methods

Our critics of the use of citation number and IF are based on the tail behavior of the corresponding distributions. A reader may object that the total number of publications in the world is not only finite but limited, and therefore conclusions based on the presence of heavy tails seem doubtful. In this connection, let us note our derivations require prelimit versions of the classical limit theorems. Suitable methods were proposed in paper [7]. Of course, the scope and topic of the proposed publication do not imply a detailed presentation of these methods. However, we will outline their main idea.

The applicability of our prelimit theorem relies not on the tail but on the 'central section' ('body') of the distributions and as a result, instead of a limiting behavior (when n, the number of i.i.d. observations tends to infinity), the prelimit theorem should provide an approximation for distribution functions in case n is 'large' but not too 'large'. Prelimiting approach seems to be more realistic for practical applications.

Given i.i.d. r.v.'s X_j, $j \geq 1$, the limiting behavior of the normalized partial sums $S_n = n^{1/\alpha}\big(X_1 + \ldots + X_n\big)$ depends on the tail behavior of X. Both the proper normalization, $n^{1/\alpha}$, in S_n and the corresponding limiting law are extremely sensitive to a tail truncation. However, it appears that the behavior of the S_n distribution is described in the following way. For small values of n, the distribution of S_n is almost arbitrary. With larger but not too large values of n, it becomes closer to the corresponding α-stable distribution. Maximal closeness to the stable law depends on the length of the tails part of the distribution of X similar to the corresponding one of the "limit" law. For the large values of n, the distribution S_n already deviates strongly from the limit one. A precise result is given by Theorem 2.1 from [7]. It gives (in some sense) a well-posed version of the Central Limit Theorem. Similar results are true for other limit theorems used above.

6 Conclusion

We have shown that such quantities as the number of citations of publications and the impact factor of a journal have distributions with heavy tails and, therefore, can lead to a significant stratification of a homogeneous group of authors. In a heterogeneous group of scientists, such a stratification may be deeper. Thus, for a correct assessment of the scientific significance of publications, indicators of a different kind should be involved. The impact of chance on them should be minimal. Unfortunately, the authors cannot offer indicators with the required properties and doubt the very possibility of such a proposal.

Appendix

The method used to estimate the tail index is based on a result of the paper [10]. Let $\xi_1, \ldots, \xi_n$ be i.i.d. random variables with regularly varying probability distribution function $G(x)$, $G(+0) = 0$. Suppose that $\xi_{1;n}, \ldots, \xi_{n;n}$ are corresponding order statistics. Define

$$\mathcal{H}_n(\lambda) = \mathbb{P}\{\xi_{n;n} < \lambda\, \xi_{n-1;n}\}.$$

Then for any fixed $\lambda > 1$

$$\lim_{n\to\infty} \mathcal{H}_n(\lambda) = \lambda^{-a},$$

where a *is the tail index of* $G(x)$.

As was mentioned, we considered the 500 highest citations records in the Probability theory. The general number of records was 10000 approximately. In other words, the number $n = 10000$ is large enough to use the limit theorem from [10]. And we had the values $\xi_{n-499;n}, \dots, \xi_{n;n}$. To estimate the function $\mathcal{H}_n(\lambda)$ we took $\lambda = 5/2$. Select at random 100 observations from 500 records. Ordered them and calculate the frequency the maximal of selected observations is more than $5/2$ times greater than the previous observation. The procedure had been repeated 15000 times. This frequency was used as an estimator of the probability we need. Finally, we took logarithm base $5/2$ of the calculated estimator.

Acknowledgment

The work of Lev B. Klebanov was partially supported by Grant 19-28231X GA ČR.

References

[1] Lev B. Klebanov (2017) One look at the rating of scientific publications and corresponding toy-model. arXiv:1706.01238v1, 1–17.

[2] Lotka, Alfred J. (1926). "The frequency distribution of scientific productivity". Journal of the Washington Academy of Sciences. 16 (12): 317Ӑ32Ӑ02“324.

[3] Garfield, Eugene. (1955). "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas". American Association for the Advancement of Science. v. 122 (3159), 108–111.

[4] Ewen Callaway (2016) Beat it, impact factor! Publishing elite turns against impact factor, Nature, v. 535, 210-211.

[5] Anthony F. J. Van Raan (2004) Sleeping Beauties in science, Scientometrics, Vol. 59, No. 3, 461 Ă432Ă402“ 466.

[6] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini (2015) Defining and identifying Sleeping Beauties in science, PNAS, 112 (24), 7426 Ă432Ă402“ 7431.

[7] L. B. Klebanov, S. T. Rachev, G. J. Szekely (1999). Pre–limit theorems and their applications, Acta Applicandae Mathematicae 58: 159–174.

[8] Lev B. Klebanov, Yulia V. Kuvaeva and Zeev E. Volkovich (2020). Statistical Indicators of the Scientific Publications Importance: A Stochastic Model and Critical Look, Mathematics, 8, 713.

[9] Kozubowski, Tomasz and Podgorski, Krzysztof (2017). A generalized Sibuya distribution, Annals of the Institute of Statistical Mathematics, 6, 855–887, doi = 10.1007/s10463-017-0611-3

[10] Volchenkova I. V. and Klebanov L. B. (2019) Characterizations of Pareto distribution by the properties of neighboring order statistics. Zapiski Naichnih Seminarov POMI v. 486, 63–70 (in Russian).