Rank Distribution and RIF Index

Brian LLinas

2024-06-20

# Title, Abstract and Keywords

### Title

1. **Modeling Rank Distribution and the Relative Importance Factor Index in Power-Law Models: Application to Social Resilience Using Scopus Databases**
2. OTRO POSIBLE TITULO: Modeling Rank Distribution and Introducing the Relative Importance Factor Index in Heavy-Tailed Distributions: Application to Social Resilience Using Scopus Databases

### Abstract

Heavy-tailed distributions, such as the power-law distribution, are prevalent in both natural and human-made phenomena, characterized by their slow-decaying tails. Previous studies have primarily focused on modeling these distributions based on frequencies, often neglecting the estimation of rank distributions. This paper addresses this gap by introducing a novel approach to model the rank distribution of frequent factors and by defining the Relative Importance Factor (RIF) index. The first objective is to develop and validate a formula that accurately models the rank distribution of frequent factors. The second objective is to introduce the RIF index, comparing the probability of a factor occupying the first position with its proba-bility of being in any other position within the range. We apply our methodology to databases downloaded from Scopus, focusing on social resilience and analyzing the number of citations of articles. Using maxi-mum likelihood estimation (MLE) to determine power-law parameters and the Kolmogorov-Smirnov (KS) statistic to estimate the optimal threshold, we ensure robust modeling. Additionally, we employ bootstrap-ping techniques to assess the uncertainty of our estimates. By modeling the rank distribution and introduc-ing the RIF index, this study enhances the understanding of heavy-tailed distributions and provides a valua-ble tool for analyzing complex systems.

### Keywords

Power-law distribution; Rank distribution; Relative Importance Factor (RIF); Heavy-tailed distributions; Social resilience; Citation analysis

# 1. Introduction

Heavy-tailed distributions are common in both natural and human-made phenomena. These probability distributions are characterized by a slower decay in their tails compared to the normal or exponential distribution. Several distributions that fall into this category include the log-normal and stretched exponential distributions, among others. In addition to these, the most well-known heavy-tailed distribution is the power-law distribution (Jiang et al., 2023). Besides its ability to describe phenomena with heavy-tailed distributions, the power-law model has captured the interest of researchers in various disciplines because it can also be applied to analyze the behavior of many complex systems. This makes it ideal for describing phenomena in diverse areas (Corral and Gonzalez, 2019; Blasius, 2020; Chen, 2021; Jiang, and de Rijke, 2021; Sethna, 2022). However, limitations in data collection often result in empirical datasets that cover a narrow observation range, complicating the clear identification of power-law behavior (Navas-Portella et al., 2019).

Power-law distributions have several characteristics. One of them is their scale-invariance property, which means that the shape of the distribution remains constant regardless of the scale at which it is observed (Corral and Gonzalez, 2019; Banshal, 2020). Another is that when the entire range is plotted on a linear axis, the curve takes the shape of a perfect “L”. Moreover, when represented on a logarithmic scale, the curve always appears straight (Chen et al., 2020; Banshal, 2020).

In computer science, the power law has been applied to the study of networks, where it has been observed that a few nodes have an extremely high number of connections while most have few (Artico et al., 2020; Devlin, 2021; Somin et al., 2022). This pattern has also been observed in social networks, where a few users have a large number of followers, and most have few (Xu et al., 2019; Arthur and Williams, 2019; Rajput et al., 2020). Additionally, in risk management, the power law is used to model the occurrence of extreme events, such as natural disasters (Pisarenko and Rodkin, 2019; Yum, 2023; Sohn et al., 2023) or financial crises (Dufrénot and Paret, 2019; Taleghani, 2020; Ghosh et al., 2021; Ben Yaala and Henchiri, 2023).

In economics, it has been used to describe the distribution of wealth, where a small fraction of the population holds most of the financial resources while another suffers from extreme poverty (Masseran, 2019; Cardoso et al., 2020; Safari et al., 2020; Puttanapong et al., 2022; Kumer, 2024). Similarly, power laws have been used to study the distribution of crime, where a few criminals commit most of the crimes while the majority commit only a few (He et al., 2023; Ng et al., 2023). In ecology, it has been employed to model the distribution of forest types (Atkins et al., 2022), freshwater fishes (Baumgartner and Peláez, 2024), hot-spring microbiomes (Li and Ma, 2019), among others. In physics, power laws are fundamental for studying the behaviors of the decay rate as a link between dissociation energy and temperature (Fischer and Schweikhard, 2020), fully developed turbulent flows in a smooth pipe (Afzal et al., 2023), and nonlinear phonon hydrodynamics (Sciacca and Jou, 2024).

In the health sector, power laws have been used to model the distribution of epidemics, where a few outbreaks can infect a large number of people, while most outbreaks affect a much smaller number (Blasius, 2020; Neipel et al., 2020; Jha, 2020). This model has also been applied to analyze the distribution of health resources, such as the availability of hospital beds or the allocation of medications, where a few hospitals or health centers receive the majority of resources (Srivastav et al., 2021; Dong et al., 2021).

In the field of academic publications, the academic influence of articles, journals, authors, etc., has been studied for several decades. Currently, one of the most widely used metrics and long recognized as an important indicator for evaluating the impact of a journal or author is the number of citations (Zhao et al., 2019). Numerous studies have revealed that the citation distribution of scientific articles follows a power law (Thelwall and Nevill, 2018; Arroyo-Machado et al., 2020; Benatti et al., 2023). In particular, the number of articles with a specific number of citations is proportional to raised to a negative scale exponent (Popescu, 2003; Banshal, 2020).

It is emphasized that previous studies model the power-law distribution based solely on frequencies but do not estimate the distribution of the rank. Therefore, firstly, in this work, we present and demonstrate a formula that allows us to model the distribution of the rank of the most frequent factors. Secondly, we introduce a parameter, which we will call

**(BRIAN. SON DOS OPCIONES DE NOMBRE. Para continuar, sigo con el de preferencia, el primero):**

1. Relative Importance Factor index (RIF index).
2. Relative Prominence Factor index (RPF index).

The RIF index compares the probability of a factor occupying the first position with the probability of it being in any other position within the range. This innovation provides a new perspective for evaluating the relative prominence and importance of factors, an aspect that has not been considered in previous studies.

The proposed theory will be applied using databases downloaded from Scopus in the area of social resilience. In particular, we will focus on the number of citations of articles published in this area. This approach allows us to identify citation distribution patterns and evaluate how academic attention is concentrated on certain articles and specific topics but based on the RIF index. The choice of social resilience as a study area responds to the growing importance of this topic in the context of social sciences and its relevance for public policy formulation and the implementation of resilient practices in communities and organizations. According to Dagdeviren et al. (2020), social resilience depends on power relations, rules/institutions, and resource distribution, which are interconnected. Without favorable conditions in these three elements, individuals may be overwhelmed by crises or survive through harmful mechanisms.

The article is organized as follows. Section 1 contains the introduction. In Section 2, we explain the power-law model and the different methods for estimating the scale parameter and the optimal threshold of the model. In Section 3, we introduce the distribution of the rank and the so-called Relative Importance Factor index. In Section 4, we present an application. Finally, in Section 5, we present the conclusions.

# 2. Power-law for data frequencies

### 2.1 The power-law model

Power-law distributions can appear in two forms: continuous distributions, which govern continuous real numbers, and discrete distributions, where the quantity of interest can only take a discrete set of values, typically positive integers (Clauset et al., 2009). Variables of the problem are defined as follows: (1) denotes the variable of interest; (2) is composed of several factors , ; and (3) is the frequency of occurrence of one factor within . Considering that our data capture the frequency with which each occurs, our study deals with a discrete problem.

Based on that, the discrete power-law model to estimate the probability that the factor within appears with frequency can be defined as

where represents the random variable corresponding to the frequency of occurrences, is the exponent or scaling parameter for , indicating the steepness of the distribution, and is a normalization constant to ensure the probabilities sum up to 1 for . Both and depend on the distribution and can be found in Clauset et al. (2009).

This distribution diverges to zero when , so there must be a lower bound for the behavior of the power-law.The normalizing constant is calculated as follows:

where is the generalized or Hurwitz zeta function (Zaghloul, 2019):

The cumulative distribution function of a power law is given by

### 2.2 Estimations of the parameters

To ensure accurate estimation of the power-law exponent and determine the range over which the power-law behavior is applicable, it is necessary to decide on the lower bound . This estimation helps identify the specific part of the media coverage distribution where the power-law model is valid. Moreover, obtaining an estimate is crucial for deriving an unbiased estimate of the power-law exponent .

According to Clauset et al. (2009), if we assume that our data are sampled from a power-law distribution for values of greater than or equal to , the maximum likelihood estimator (MLE) for in the discrete case is defined as

where is the number of occurrences by the -th factor () and are the observed values of such that , where is the estimated value of . We estimate using the maximum likelihood (ML) method. The corresponding log-likelihood function for this estimation process is derived from the data for media coverage factors related to frustrations in Colombia and Greece.

Note in the above formula that, to calculate , we must first estimate . To estimate the lower bound on the power-law , we used a metric known as the Kolmogorov-Smirnov (KS) statistic (Ramos et al., 2024), which is defined as the maximum difference between , the cumulative distribution function (CDF) of the observations, and , the CDF of the power law that optimally fits the data . The KS statistic is defined as

The fitting process is sometimes performed by linear regression using logarithmically transformed variables. This approach is used because applying the logarithm to the power law function results in

Thus, a power law appears as a straight line with slope in a logarithmic plot. It is important to note that changes in the scale parameter can affect the slope of the curve in the log-log plot, resulting in changes in the shape and behavior of the distribution represented by the power law.

The bootstrapping procedure is used to analyze the uncertainty associated with exponent estimation. It consists of randomly selecting data samples with replacement and then applying the MLE procedure with a KS cutoff on that sample. This process is repeated several times to evaluate the uncertainty. In this study, 1000 iterations were performed on all data sets. In addition, we perform a particular hypothesis test and provide the corresponding p-value to test : one power-law distribution fits adequately versus : another distribution might fit better. We will use the R 4.3.1 package poweRlaw (Gillespie, 2014) to perform all the above analyses.

# 3 Power-law for rank and RIF index

## 3.1 Relation between the Frequencies and the Rank

For the variable , which is composed of several factors , , we will suppose that the frequency of occurrences of each factor is a descending ordered series. Let denote the frequency of occurrence of the factor at rank , where is the rank of the factor within the variable . That is,

This means that the frequency of occurrence decreases as the rank increases. According to Popescu (2003), Zipf’s law of rank-frequency states that, in a generally ordered set of data, the frequency of occurrence of an element is inversely proportional to its rank :

where is the rank of the element (1 for the most frequent, 2 for the second most frequent, etc.); is the frequency of the element at rank for ; is the power law exponent, and is a constant of proportionality. This implies that

To estimate the parameters and , we only need to fit the regression model with the response variable and the explanatory variable . But the method of estimation for the parameters and can vary depending on the size of the data, i.e., the number of factors within . For datasets with a large number of factors, methods such as maximum likelihood estimation (MLE) are preferred due to their robustness and efficiency. MLE can provide precise estimates even with large sample sizes.

For smaller datasets, ordinary least squares (OLS) regression can be used to estimate the parameters. OLS is simpler and computationally less intensive, making it suitable for smaller samples where computational resources may be limited.

Additionally, bootstrapping techniques can be employed to assess the variability and confidence intervals of the estimated parameters, especially in cases where the sample size is small or the data exhibits significant variability. This approach involves repeatedly resampling the data with replacement and recalculating the estimates to build a distribution of the parameter estimates.

In summary, depending on the number of factors in , different estimation techniques can be employed to accurately determine the parameters and , ensuring the reliability of the rank-frequency relationship described by Zipf’s law. Based on the above, once these estimates have been found, we can write

### 3.2 Estimation of the Distribution of the Rank

In this section, we will demonstrate one of the primary objectives mentioned in the introduction: estimating the distribution of the rank of the most frequent factors.

**3.2.1. Theorem.** Let be a variable composed of several factors , , with descending ordered frequencies of occurrence for each factor at rank , where is the rank of the factor within . That is, for all . Suppose that follows a power-law distribution with estimated scale parameter and threshold , and normalization constant , ensuring that the probabilities sum to 1 for . Then, for all ranks such that , the estimated probability that a factor occupies rank is given by:

1. .
2. .

There represents the random variable corresponding to the frequency of occurrences, is a proportionality constant, is the normalising constant defined by:

and (real number) and are estimates such that , and is the power law exponent for the rank .

**Proof Theorem.** Given that for all , and assuming follows a power-law distribution, we have , where is a constant of proportionality and is the power law exponent. To estimate the distribution of the rank , we start with the relationship between the probabilities and . Since these probabilities are inversely proportional, we have:

where is a proportionality constant. This result demonstrates the part (a). Substituting into the probability , and using the normalization constant to ensure the probabilities sum to 1, we get:

where is the power law exponent for the occurrence frequency . Defining the estimation , we get , where is the normalizing constant defined as indicated by the statement of the theorem. Thus, we have shown that the probability of a factor being at rank follows the power-law distribution with exponent . This result demonstrates the part (b).

## 3.3. The Relative Importance Factor (RIF) Index

**3.3.1. Definition.** For all ranks such that , the estimated Relative Importance Factor (RIF) Index of a factor in rank , denoted as , is defined as:

where is the estimated probability of the factor occupying the first position, and is the estimated probability of the factor being in rank .

**Properties.**

## 3.4 Analysis of the RIF index and Probability

1. This estimate will help us explain how the probability of observing the first element in the rank (rank 1) compares to the probability of observing the element in the ranking . We know that . Then, in this case, is given by
2. This ratio tells us how many times more likely it is to observe the first element in the rank (rank 1) compared to the element in rank . When , then . This is trivially 1, since we are comparing the probability of the first rank with itself. Suppose that . Then, the probability of observing the first element in the rank is times higher than the probability of observing the element in the rank . In other words,
3. For different values of , Figure 1(a) and Figure 1(b) show how the rank relates to both the RIF index and the probability . Given that is positive, the graph illustrates how the RIF index and the probability change as both and are increased.

**OJO. FALTAN LAS FIGURAS.**

Each line, color-coded according to the legend, represents the distribution for a particular value of . The dark line shows a steep increase (decrease) in RIF index (probability) with higher ranks, indicating rapid growth (decline), while other lines exhibit more gradual or stable trends in the odds. This suggests significant variability in how RIF index (probability) changes with rank across different values of , highlighting the unique impact of rank on RIF index (probability) for each type. As increases and becomes more positive, the RIF index increase (the probability decreases) exponentially. The higher the value of , the steeper the increase (decrease) in the RIF index (probability) with respect to . This indicates that for larger values of , the disparity between the probability of the highest-ranked event and lower-ranked events becomes much more pronounced. As increases, the RIF index diverge significantly and the probability converges to 0, especially for higher values, highlighting an exponential growth pattern.

Figure 1(c) explores the relationship between the probability and the RIF index for different values of . As increases, the RIF index decrease sharply initially and then level off. The gradient color represents different values of , indicating that higher values of correspond to higher RIF index for a given .

Figure 1(d) examines the relationship between the probability and the RIF index for different values of . As increases, the RIF index increase linearly. The gradient color represents different values of , showing that higher values of lead to higher RIF index for a given .

## 3.5 Domain and Ranges of the RIF index

The domain of is the set of all positive integer representing the rank of the factor with the conditon that . The range of is . Since for any , the value of will always be 1 or greater. The range of values that the RIF index can take allows for several interpretations regarding the importance of factors:

* **Near 1:** When , it indicates that the probability of the factor being at rank is nearly the same as being at rank 1. This suggests factors with relatively stable importance.
* **High Values:** When , it indicates that the factor is much more likely to be at rank 1 than at rank . This suggests a sharp decline in importance as the rank increases.

To provide clear and impactful terminology for discussing the importance levels of different factors within our analysis, we propose the following names for the ranges of , shown in the following definition.

**3.5.1. Definition.** Let be the estimated RIF index for a factor of at rank . Then, the factor (in rank 1) receives the name (with respecto to in rank ) indicated below when the RIF meets the following conditions:

**BRIAN. SON RANGOS Y NOMBRES SUGERIDOS. CONSIDERO ESCOGER UNO SOLO.**

1. (*Equilibrium*, *Balanced*, or *Stable*): Indicates that the factor’s importance is nearly the same across different ranks. This suggests a stable distribution of importance and reflects minimal change in importance as the rank changes.
2. (*Moderate*, *Intermediate*, or *Notable*): The factor has a slightly higher importance at rank 1 compared to other ranks. The difference in importance is noticeable but not extreme, representing a moderate decline in importance with increasing rank.
3. (*Significant*, *Elevated*, or *Enhanced*): The factor is significantly more important at rank 1 than at higher ranks. There is a clear elevation in importance at rank 1, indicating an enhanced level of importance at the top rank.
4. (*Major*, *Critical*, or *High*): The factor’s importance at rank 1 is majorly higher than at other ranks. This reflects a critical difference in importance between the top rank and others, with a steep decline as rank increases.
5. (*Dominant*, *Supreme*, or *Peak*): The factor is overwhelmingly more important at rank 1 than at any other rank. This represents a supreme level of importance at the top rank, with the importance peaking at rank 1 and dropping sharply at higher ranks.

These categories can be used to effectively communicate the significance of our findings in terms of the relative importance of various factors, compared with factor 1.

## 3.6 The global RIF index

Let be the estimated Relative Importance Factor (RIF) index comparing the ranks and , where . Then, the factor (in rank ) receives the name (with respecto to in rank ) indicated below when the RIF meets the following conditions:

**3.6.1. Definition.**

The global Relative Importance Factor Index (estimated) between two ranks and is defined as:

where is the estimated probability of the factor occupying the rank , and is the estimated probability of the factor occupying the rank . Observe that .

**Observation.**

The domain of is the set of all positive integer representing the rank of the factor with the conditon that and . The range of is . Since for any , the value of will always be 1 or greater. In this case, the factor (in rank ) receives the name indicated below when the RIF meets the following conditions:

1. (*Equilibrium*, *Balanced*, or *Stable*): Indicates that the importance of the factor is nearly the same across both ranks. This suggests a stable distribution of importance and reflects minimal change in importance between the ranks.
2. (*Moderate*, *Intermediate*, or *Notable*): The factor has slightly higher importance at rank compared to rank . The difference in importance is noticeable but not extreme, representing a moderate decline in importance with increasing rank.
3. (*Significant*, *Elevated*, or *Enhanced*): The factor is significantly more important at rank than at rank . There is a clear elevation in importance at rank , indicating an enhanced level of importance.
4. (*Major*, *Critical*, or *High*): The factor’s importance at rank is majorly higher than at rank . This reflects a critical difference in importance between the two ranks, with a steep decline as the rank increases.
5. (*Dominant*, *Supreme*, or *Peak*): The factor is overwhelmingly more important at rank than at rank . This represents a supreme level of importance at the top rank, with the importance peaking at rank and dropping sharply at higher ranks.

These categories can be used to effectively communicate the significance of our findings in terms of the relative importance of various factors.

### 3.7 Values of for each range of

The value of will affect the values of . Here are the approximate values of for each range:

* : indicates that the importance of the factor is nearly the same at both ranks and .
* : suggests a moderate decline in importance with increasing rank.
* : indicates a significant decline in importance with increasing rank.
* : reflects a major decline in importance with increasing rank.
* : indicates a dominant decline in importance with increasing rank.

### Demostracion

To demonstrate how these values of affect , consider the following:

Assume that the probability follows a power-law distribution:

For ranks and where , the estimated Relative Importance Factor index is:

Given the properties of the power-law distribution, we can derive that:

Since , and thus .

This confirms that will always be 1 or greater. The specific value of depends on the value of : - For , as the factor’s importance is nearly the same at both ranks. - As increases, grows larger, indicating a steeper decline in importance with increasing rank.

Thus, the range of values for provided earlier gives us a practical guide for interpreting in terms of the relative importance of factors across different ranks.

### sEGUNDO CASO

To demonstrate the case where , consider the following:

Assume that the probability follows a power-law distribution:

For ranks and where , the estimated Relative Importance Factor index is:

Given that , we have:

Taking the natural logarithm on both sides, we get:

Since , and . Therefore:

This implies:

Finally, we solve for :

To ensure , the value of must lie within the range:

For typical values of and where , this generally implies:

Thus, the range of that corresponds to is approximately , indicating a moderate decline in importance with increasing rank.