Methods of verification of compliance with an asymptotically power law

I A Belashova, V V Bochkarev, V A Tyurin
Kazan Federal University, Kremlevskaya str.18, Kazan 420018, Russia

E-mail: vbochkarev@mail.ru

Keywords: power-law distribution, robust criteria, runs test, Zipf's law

Abstract This paper describes methods of the power law verification using different empirical data. We do not analyse the value of deviations from the model but try to found out whether these deviations are regular or random. The suggested approach is based on the idea of finding local power approximation of the considered series for each range of ranks, after which one or another trend criterion is applied to the obtained series of local exponents. Application of the runs test is also discussed. The suggested methods were tested using 10 sets of empirical data, which are available for free. It was shown that compliance with the power law is satisfactory only in one case.

1. Introduction

It was found out that the power law manifests itself in many practical cases related to different fields of science and technology [1]. However, the conclusion about distribution of empirical data by the power law was not strictly justified for all cases in terms of statistics. A large number of data sets from previous works was analysed in [2] and it was shown that the conclusion about compliance with the power law was statistically justified only in 11 cases out of 24. In [2], the verification was performed assuming that the measured values precisely correspond to power distribution, at list in some relatively large range of values. The paper describes the case when the distribution strictly complies with the power law:

\[ F(x) = A x^{-\alpha} \]  

Here \( F(x) = P(\xi \leq x) \) is a cumulative distribution function. However, only asymptotic approximation to the power law can be expected in many practical cases. As an example, let us consider the distribution of word frequency. Zipf’s law established in 1949 was one of the first examples of the power law in empirical data. According to this law, frequencies of words \( f_r \) are characterised by power dependence on their rank \( r \) (the numbers in a list of words sorted by frequency reduction):

\[ f_r \sim r^{-\gamma} \]  

As the sample cumulative distribution function is equal to \( 1-r/n \), where \( n \) is the sample size, Zipf’s law is equivalent to the assumption that the word frequencies are distributed according to the power law (1), while the exponents in formulas (1) and (2) are related by an obvious relation \( \gamma = \alpha - 1 \).

Probability models, such as monkey model and Markov models are traditionally used to explain Zipf’s law. The monkey model assumes that a text is a random sequence of symbols (including a
white-space) and the probability of selecting the subsequent symbol doesn’t depend on the previous symbol. Dependence of the subsequent symbol on the previous one is taken into account in Markov models. In [3, 4], it is proved that for a typical case both models provide distribution which satisfies the following inequation:

$$A_1 x^{-\gamma} \leq 1 - F(x) \leq A_2 x^{-\gamma}$$  

(3)

It is clear that if we apply a statistical criterion based on model (1) to a set of empirical data with a distribution law which satisfies equations (3) (but not an exact relation (1)), the correct underlying hypothesis will certainly be rejected if a sample is relatively large.

This paper describes methods which allow us to perform correct verification for cases when the power law is satisfied only asymptotically, in other words when the distribution satisfies inequations (3). Testing is performed using 10 sets of empirical data, which were considered in [2] and are available for free.

2. Methods

Assume that the data are described by the following dependence:

$$f(r) = A(r) r^{-\gamma(r)}$$  

(4)

Functions $A(r)$ and $\gamma(r)$, which the mentioned equation contains, are considered to be slowly varying. Observing the behaviour of these functions, it can be concluded whether there is convergence to a power law or not. When the distribution asymptotically tends to the power distribution, $A(r)$ and $\gamma(r)$ should tend to constant values with increasing of the argument. If a trend is detected in these sequences, this will mean that there is a systematic deviation from the power law.

The proposed algorithm includes the following sequence of steps:

- The variation range $r$ is divided into a sufficient number of intervals. Then, data are separately fitted at each interval by a power curve, estimating the amplitude and the exponent.

- This or that trend criterion is applied to the obtained series $A_i$ and $\gamma_i$.

We thus assume that intervals can be divided in such a way that the deviations from the average values $\gamma_i$ (and $A_i$) will be uncorrelated series. It is clear that to do this, the intervals should not be less than the characteristic scale of variations $A(r)$, $\gamma(r)$. The choice of dividing into intervals is left to a researcher in each case. Of course, there are some cases when this method doesn’t give the desired result. In particular, these are all cases when $A(r)$ and $\gamma(r)$ monotonically approach the limit value (from above or below). Nevertheless, the proposed approach is completely correct for a wide range of cases when $A(r)$ and $\gamma(r)$ are around the limiting value.

If the number of sample values is small, it can be difficult to perform fitting for some intervals. In this case, the following method can be used. The value $-\frac{\partial \log f}{\partial \log r}$ gives the value of the exponent for the exact power-law dependence. Calculating the derivative using the method of finite differences we obtain

$$\gamma_i = - \frac{\log f_{i+1}/f_i}{\log r_{i+1}/r_i}$$  

(5)

To verify the presence of a trend in the series $A_i$ and $\gamma_i$, it is preferable to use nonparametric criteria, for example, the runs test. Run are sequences of points for which the error has the same sign. If the deviations of the empirical data from the approximating curve are random and weakly correlated, the probability of occurrence of long runs is small. Thus, the presence of long runs indicates systematic deviations of empirical data from the model. The Ramachandran-Ranganatan criterion [5] is used in this paper. According to this criterion, the decision is made depending on the value of the sum of the squares of the lengths of the runs:

$$R = \sum_i l_i^2$$  

(6)
Here \( l_i \) is the length of the \( i \)-th run. P-values, corresponding to different values of \( R \), are found by statistical modelling.

3. Data analysis

The following data sets are considered [2]:

1) The frequency of occurrence of unique words in the novel Moby Dick by Herman Melville.
2) The severity of terrorist attacks worldwide from February 1968 to June 2006, measured as the number of deaths directly resulting.
3) The number of species per genus of "recent" mammals.
4) The numbers of customers affected in electrical blackouts in the United States between 1984 and 2002.
5) The human populations of US cities in the 2000 US Census.
6) The sizes in acres of wildfires occurring on US federal land between 1986 and 1996.
7) Peak gamma-ray intensity of solar flares between 1980 and 1989.
8) The intensities of earthquakes occurring in California between 1910 and 1992, measured as the maximum amplitude of motion during the quake.
9) The frequencies of occurrence of US family names in the 1990 US Census.
10) The number of links to web sites found in a 1997 web crawl of about 200 million web pages, represented as a simple histogram.

All these data sets and brief information about these cases can be found on the page of one of the authors of the work [2] (http://tuvalu.santafe.edu/~aaronc/powerlaws//data.htm). Due to the limited size, we present the results of the tests only for the series of local indices \( \gamma_i \). The range of rank variations was divided into uniform intervals in a logarithmic scale. The value \( \log r_{i+1}/r_i \) was chosen as 0.25 for examples 2-10, the value 0.45 was chosen for the first example. The calculation results for the series of local exponents are given in Table 1. The sample size, the number of intervals for which the values \( \gamma_i \) were found, the value of the statistics \( R \) and the corresponding p-value are given for each example.

| The number of samples | The number of intervals | \( R \) | p-value |
|-----------------------|------------------------|---------|---------|
| 1                     | 18855                  | 20      | 182     | 9.64·10^{-4} |
| 2                     | 9101                   | 32      | 266     | 5.43·10^{-4} |
| 3                     | 5731                   | 33      | 311     | 1.81·10^{-4} |
| 4                     | 211                    | 21      | 91      | 0.0661    |
| 5                     | 19447                  | 39      | 303     | 5.11·10^{-4} |
| 6                     | 203785                 | 46      | 974     | < 10^{-7} |
| 7                     | 12773                  | 37      | 423     | 1.96·10^{-4} |
| 8                     | 19302                  | 39      | 305     | 4.75·10^{-4} |
| 9                     | 2753                   | 30      | 248     | 6.82·10^{-4} |
| 10                    | 14480                  | 36      | 182     | 0.017     |

As can be seen from the table, the hypothesis about the power law is confirmed only in one case (example 4). It should be noted that we have the smallest sample for this example. In another case (example 10), the p-value is equal to 0.017 and the solution will depend on the accepted level of significance.

Let us consider Example 7 in more detail (the intensity of gamma rays during solar flares). Figure 1.a shows (in a log-log scale) the dependence of empirical values sorted in a list in a descending order on the rank. The shape of this graph suggests that the power law describes the empirical data well, at least in a relatively wide range of ranks (approximately from 20 to 2000). But Figure 1.b, which shows change of local exponents \( \gamma_i \), doesn’t confirm this information. It should be noted that no horizontal
sections can be seen in the graph (otherwise, the presence of horizontal sections would indicate that the power-law is realised in the selected range), on the contrary, significant systematic variations of exponents are observed. Selecting any range, we obtain 2-3 long runs and a large value of statistics (6), which results in the rejection of the hypothesis about the power-law. The above-mentioned interval 20-2000 which is characterised by systematic growth of $\gamma(r)$ is the most interesting. This means that in the given case it is necessary to choose a model distribution that decreases more rapidly than the power distribution.

Fig. 1 a) Distribution of peak gamma-ray intensity of solar flares between 1980 and 1989 as a rank function. b) Local exponents depending on the rank. The dotted line shows the median level, the dashed line shows a trend schematically.

Fig. 2 a) Distribution of the frequency of occurrence of unique words in the novel “Moby Dick” by Herman Melville. b) Local exponents depending on the rank. The dotted line shows the median level, the dashed line shows a trend schematically.

Let us consider example 1. Figure 2.a shows dependence of the frequency of words in the novel "Moby Dick" on their rank, Figure 2.b shows the corresponding dependence for local exponents. Variations of local exponents are more significant in this case than in the previous one. However, a
systematic increase of exponents is observed with increase of a rank. However, there are no indicators that the exponent $\gamma(r)$ can tend to some limit with a further increase in rank. Thus, the assumption about the power-law (even in the expanded treatment (3)) should be rejected.

A similar behaviour is observed for other examples considered (except for example 4 and, possibly, example 10): it is a systematic increase of the exponent $\gamma(r)$ with increase of the rank.

4. Conclusion
In this paper, we proposed and tested a criterion that allows one to verify the correspondence of empirical data to an asymptotically power-law distribution which satisfies inequalities of the form (3). It was shown that compliance with the power law can be regarded satisfactory only in one of the ten cases considered. For three cases (examples 1, 2, 5), our conclusions differ from the conclusions of [2]. In these cases, deviations from the distribution (1) may seem small, but they are of a regular nature, which results in rejection of the hypothesis of a power law. It should also be noted that the analysis of the series of local exponents $\gamma(r)$, both quantitative and visual, allows us to make more valid conclusions about the correspondence of empirical data to the power law.

This work was supported by the Russian Foundation for Basic Research, Grant no. 15-06-07402. The research of the second author was supported by the Russian Government Program of Competitive Growth of Kazan Federal University.

References
[1] M. E. J. Newman Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323-351 (2005)
[2] A. Clauset, C. R. Shalizi, M. E. J. Newman Power-law distributions in empirical data. SIAM Review 51, 661-703 (2009)
[3] Bochkarev, V.V., Lerner, E.Y. The Zipf law for random texts with unequal probabilities of occurrence of letters and the Pascal pyramid. Russian Mathematics (2012) 56: 25. doi:10.3103/S1066369X12120031.
[4] Bochkarev, Vladimir V. and Lerner, Eduard Yu.. (2014), "Strong power and subexponential laws for an ordered list of trajectories of a Markov chain", Electronic Journal of Linear Algebra, Volume 27. pp. 534-556. doi: 10.13001/1081-3810.1917
[5] David M. Himmelblau Process Analysis by Statistical Methods. John Wiley & Sons Inc, 1970.