INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)

THEORETICAL AND EMPIRICAL STUDIES OF THE TAILS OF SCIENTOMETRIC DISTRIBUTIONS

W. GLÄNZEL, A. SCHUBERT

Library of the Hungarian Academy of Sciences, Information Science and Scientometrics Research Unit, Budapest, POB, H-1361, Hungary

Abstract

The distributions of non-negative random variables occurring in scientometrics are said to have a proper tail if they asymptotically obey "Zipf's Law", i.e., if

$$\lim_{k\to\infty} (1-F(k))\, k^{\alpha} = \text{const}$$

for some real $\alpha > 0$ where F denotes the cumulative distribution. The tail of scientometric distributions has a particular significance because it generally contains the most "prominent" elements of the population (e.g. highest cited papers or most productive authors). In addition, the tail parameter, α, is a sensitive indicator of several fundamental features of the whole distribution. It is shown that, among others, the tail parameter governs order and rank statistics. New estimation methods of α as well as statistical tests for extreme values and ranked tail elements are developed. The methods are illustrated on empirical samples of citation rates and publication activity.

1 INTRODUCTION

A variety of statistical studies in scientometrics are based on ranked statistics. Interesting and important rules and laws have been discovered, and models have been built. Rank statistics containing the most "prominent" members of a population are closely associated with the tails of the scientometric distributions concerned. In the case of the wide class of Paretian distributions ("Zipf's Law") the "tail behaviour" is approximately determined by a single parameter, the characteristic tail parameter, α. In the following, some statistical tests based on the properties of Paretian tails are presented. Multi-sample tests can be used for testing the extreme values of sets of like samples. Moreover a test for large samples based on logarithmic transformations is developed. All the tests presented are applicable for both Gaussian and non-Gaussion distributions.

2 THE CHARACTERISTIC TAIL PARAMETERS

The present study is focussed on scientometric distributions, therefore only discrete non-negative integer valued random variables (modelling , e.g., publication activity and citation rates) are taken into consideration. The distribution of such a random variable, X, is said to have a proper tail, if it asymptotically obeys "Zipf's Law", i.e., if

$$\lim_{k\to\infty} (1-F(k))k^{\alpha} = \text{const} \qquad (2.1)$$

for some real $\alpha > 0$, where F denotes the cumulative distribution function of X. The parameter α is called the characteristic tail parameter. The distributions obeying the asymptotic form of Zipf's Law together with the analogous continuous distributions form the class of Paretian distributions. If $\alpha > 2$, a distribution of the above type belongs to the domain of attraction of the normal distribution and is therefore said to be Gaussian, else, if $\alpha < 2$ then the distribution belongs to the domain of attraction of a non-Gaussian stable distribution with the characteristic parameter α. Thus for $\alpha < 2$, but not for $\alpha > 2$, the characteristic tail parameter is at the same time the characteristic parameter of the stable limit distribution.
According to Eq. (2.1) the behaviour of a Paretian distribution function for great values is governed by the characteristic tail parameter only. Some simple but effective methods for the estimation of the parameter α derive from characterization theorems based on truncated moments (Glänzel & al. 1984, Glänzel 1987). Tail properties of a distribution are reflected by moments truncated from the left side. For a given non-negative integer valued random variable X and a real function h the truncated moments are defined as

$$(2.2) \qquad E(h(X) \mid X \geq k) = \frac{\sum_{i \geq k} h(i) p_i}{\sum_{i \geq k} p_i} \; ; \; k \geq 0,$$

provided this quantity is defined. It can be shown that if X has a Paretian distribution with a characteristic tail parameter $\alpha > 1$ then $E(X) < \infty$ and

$$(2.3) \qquad \lim_{k \to \infty} E(X/k \mid X \geq k) = \alpha/(\alpha-1).$$

The following lemma sets bounds to the approximation in the case of distributions satisfying weaker conditions :
Lemma 2.1 : Let X be a non-negative integer-valued random variable. Assume that a threshold value $k_0 \geq 0$ can be given so that

$$P(X \geq k) \sim c_1 \cdot (c_2 + k)^{-\alpha} \; ; \; k \geq k_0$$

for some real $c_1 > 0$, $c_2 \geq 0$ and $\alpha > 1$. Then the following inequality holds :

$$(2.4) \qquad E((X/k-1) \mid (X/k-1) \geq 0) = 1/(\alpha - 1) + \varepsilon_k \; ; \; k \geq k_0 \; ,$$

where

$$|\varepsilon_k| \leq c_3/k \qquad (c_3 > 0).$$

Eq. (2.3) is obviously a direct consequence of the lemma. Note that the above approximations assume a finite expectation of the random variable. If the expectation is suspected to be infinite ($\alpha \leq 1$), then Eq. (2.3) and Lemma 2.1 must be modified : the first descending factorial moment, which is finite for any non-negative integer valued random variable with Paretian distribution ($0 < E(1/(X+1)) < 1$), is to be used. The alternative Lemma 2.2 is thus valid for all Paretian distributions. Since this lemma is also a consequence of the above mentioned characterization theorems, the proof is omitted.
Lemma 2.2 : Let X be a non-negative integer valued random variable. Assume that a threshold value $k_0 \geq 0$ can be given so that

$$P(X \geq k) \sim c_1 \cdot (c_2 + k)^{-\alpha} \; ; \; k \geq k_0$$

for some real $c_1 > 0$, $c_2 \geq 0$ and $\alpha > 0$. Then the following inequality holds :

$$(2.5) \qquad E((1-k/(1+X)) \mid ((1-k/(1+X)) \geq 0) = 1/(1+\alpha) + \varepsilon_k \; ; \; k \geq k_0 \; ,$$

where

$$| \varepsilon_k | \leq c_3'/k \quad (c_3' > 0).$$

The condition $(1-k(1+X)) > 0$ in Eq. (2.5) is equivalent to the condition $X \geq k$ because X is integer valued. The following equation is an obvious consequence of Lemma 2.2 :

$$(2.6) \qquad \lim_{k \to \infty} E(k/(1+X) \mid X \geq k) = \alpha/(1+\alpha) \; ,$$

provided X has a Paretian distribution. The above lemmas and equations are the necessary tools of statistical applications. For the estimation of the parameter α, the theoretical values of the truncated moments have to be replaced by the corresponding empirical values.

2 CHARACTERIZATION OF TAILS BY RANK STATISTICS AND EXTREME VALUES

Consider a given sample $(X_i)_{i=1}^n$. Assume that the sample elements are ranked in decreasing order

$$X_1^* \geq X_2^* \geq \ldots \geq X_n^* .$$

It is obvious that a certain number $m << n$ of ranked sample elements $(X_i^*)_{i=1}^m$ form the tail of the empirical distribution. The question arises which theoretical values can be assigned to the values of the ranked "tail" elements. In the "quantile approximation" the statistic X_k^* is considered as the empirical $(1-k/n)$-quantile. This is associated with Gumbel's so-called "characteristic k-th extreme value" (Gumbel 1958). The characteristic k-th greatest value is defined by

$$(3.1) \qquad u_k = G^{-1}(k/n) = \sup \{x : G(x) > k/n\} \; ; \; k = 1,2, \ldots, n \; ,$$

where n is the sample size, $G = 1-F$ and F is the common cumulative distribution function of the random variables X_i. u_k is properly speaking the $(1-k/n)$-quantile of the distribution while X_k^* is the corresponding quantile of the sample. The definition (cf. Eq. (3.1)) makes sure that the rank of Gumbel's characteristic extremes coincides with the sample rank :

$$u_1 \geq u_2 \geq \ldots \geq u_n \; .$$

If there is a Paretian distribution underlying to the model then the values u_k have the following property which can be obtained immediately from Eqs (3.1) and (2.1):

$$(3.2) \qquad u_k \sim c. \, (n/k)^{1/\alpha} \quad ; \; c > 0$$

and

$$(3.3) \qquad u_k/u_{k+1} \sim (1 + 1/k)^{1/\alpha} \; ; \; k = 1, 2, \ldots, n \; .$$

While u_k depends on the sample size, and the rate of divergence for increasing sample size ($n \to \infty$) is n, the ratios u_k / u_{k+1} are asymptotically independent of n, and thus $u_k/u_{k+1} = (1 + 1/k)^{1/\alpha}$ is an exact limiting formula for Paretian distributions if $n \to \infty$ ($k << n$). In this context the case $\alpha = 1$ (Price distribution) was studied by Glänzel & Schubert (1985).

3 A MULTI-SAMPLE TEST FOR RANKED TAIL ELEMENTS

A multi-sample test examines a set of samples $\{\{X_i^{(j)}\}_{i=1}^{n_j}\}_{j=1}^{m}$ simultaneously. The test is used pointwise, i.e., ranked sample elements (tail elements) such as the greatest, the second greatest etc. sample elements are tested separately. A multi-sample test based on Gumbel's extreme values was introduced and applied to Waring-type distributions by Schubert and Telcs (1987). The test can be essentially described as follows.
We proceed on one unique sample $\{X_i\}_{i=1}^{n}$. If $n > \lambda_k > 0$ is a given real number and n the size of the sample, then it can be shown (e.g. Rényi, 1962) that

$$(4.1) \qquad \lim_{n\to\infty} P(X_k^* < G^{-1}(\lambda_k/n)) = \exp\{-\lambda_k\} \sum_{i=1}^{k-1} \lambda_k^i / i! = F^*(\lambda_k); \; k << n,$$

where 1-G = F is the common distribution function of the random variables X_i. If the values λ_k are determined so that $F^*(\lambda_k) = 0.5$ then we have

$$(4.2) \qquad P(X_k^* < G^{-1}(\lambda_k/n)) \sim P(X_k^* \geq G^{-1}(\lambda_k/n)) \sim 0.5 \; ; \quad n >> 1 \; .$$

According to Eqs (4.1) and (4.2), λ_k's are determined as the solutions of the equations

$$(4.3) \qquad \exp\{-\lambda_k\} \sum_{i=1}^{k-1} \lambda_k^i / i! = 0.5 \; ,$$

i.e., $(2\lambda_k)$ can be considered as the median of a χ^2-distribution with 2k degrees of freedom. On the other hand, the theoretical values for the comparison with the ranked tail elements can be deteremined according to Eq. (4.1) as

$$(4.4) \qquad u_k^* = G^{-1}(\lambda_k/n) \; .$$

Since $(k-1) < \lambda_k < k$ (i.e. $\lambda_k \sim k-0.3$) u_k^* can be interpreted as a modification of Gumbel's characteristic extreme values. Table 1 contains more precise approximations of the values λ_k ($k = 1,2, \ldots, 10$).

k	1	2	3	4	5	6	7	8	9	10
λ_k	0.6931	1.6783	2.6741	3.6720	4.6709	5.6701	6.6696	7.6692	8.6689	9.6687

Table 1 The values λ_k for $k = 1, \ldots, n$.

In the case of Paretian distributions the preceding equation has the following approximate solution (cf. Eq. (3.2)) :

$$(4.5) \qquad u_k^* \sim c(n/\lambda_k)^{\alpha}, \; c > 0.$$

The values u_k^* can be calculated based on a particular distribution, or, if Eq. (4.5) is used, the factor c must be estimated, too.
If we have to answer the question whether the *maxima of a set of m samples* correspond to theoretical values expected from the Pareto property, we determine the modified characteristic extreme values $\{u_1^{*(j)}\}_{j=1}^{m}$, and if our hypothesis is true, then the test statistic (see Eq. (4.2))

$$(4.6) \qquad (2\nu_m - m)/\sqrt{m} = (2\sum_{i=1}^{m} \chi(X_1^{*(j)} < u_1^{*(j)}) - m)/\sqrt{m}$$

has approximately a standard normal distribution. If ν_m significantly differs from m/2 then the hypothesis must be rejected. If needed, the test can be pointwise repeated for the second greatest sample elements $\{X_2^{*(j)}\}_{i=1}^{m}$, the third greatest sample elements $\{X_3^{*(j)}\}_{i=1}^{m}$ and so on.

4 A TEST FOR LARGE SAMPLES

In this section a large sample test for Paretian tail properties is presented. Consider a given sample $\{X_i\}_{i=1}^{n}$ with $n \gg 1$. The distribution of the *transformed ranked sample elements* $\alpha \log X_k^*$ ($k \ll n$) can then be approximated by an exponential distribution. It was shown that the differences

$$k(\log X_k^* - \log X_{k-1}^*) = k.\log(X_k^*/X_{k+1}^*) \; ; \; k = 1,2, \ldots, k_0 \ll n$$

are approximately independent identically distributed random variables, where the common distribution is exponential with parameter α (cf. Glänzel & al., 1984), i.e.,

$$(5.1) \qquad P(k.\log(X_k^*/X_{k+1}^*) < x) \sim 1 - \exp\{-\alpha x\} \; ; \; k \leq k_0.$$

So, if the characteristic parameter, α, is known, the set $\{k.\log X_k^*/X_{k+1}^*\}$ can be tested for having an *exponential distribution with parameter* α. The threshold k_0 can be estimated for instance with the help of Gumbel's characteristic extremes, or if there is no other way, k_0 can roughly be approximated by log n, where *n is the sample size*. A Kolmogorov-Smirnov test can be applied. We define the maximum deviation $D_{k_0}(x)$ from the hypothetical exponential distribution as follows :

$$D_{k_0}(x) = \max | F_{k_0}(x) + \exp(-\alpha x) - 1 | ,$$

where the empirical distribution function F_{k_0} is defined by

$$F_{k_0}(x) = 1/k_0 \sum_{k=1}^{k_0} \chi(k.\log X_k^*/X_{k+1}^* < x) .$$

According to the Kolmogorov-Smirnov test the hypothesis can be accepted if $D_{k_0}(x)$ does not exceed the critical value $K(k_0, x)$ belonging to a given confidence level. Table 2 presents the values $K(k_0, x)$ for three confidence levels $\varepsilon = 0.90$, 0.95 and 0.99.
If the characteristic tail parameter, α, is unknown one can apply some standard methods for parameter estimation based on a hypothetical distribution. The parameter estimation can then be followed by a standard goodness-of-fit test.

Table 2. The values of the function $K(k_0, x)$ for three confidence levels.

k	$\varepsilon = 0.90$	$\varepsilon = 0.95$	$\varepsilon = 0.99$
1	0.950	0.975	0.995
2	0.776	0.842	0.929
3	0.636	0.708	0.829
4	0.565	0.624	0.734
5	0.509	0.563	0.669
6	0.468	0.519	0.617
7	0.436	0.483	0.576
8	0.410	0.454	0.542
9	0.387	0.430	0.513
10	0.369	0.409	0.489
> 10	$\sim 1.22/\sqrt{k_0}$	$\sim 1.36/\sqrt{k_0}$	$\sim 1.63/\sqrt{k_0}$

5 APPLICATIONS

The first example, a multi-sample test, illustrates the theoretical considerations of section 4 on the frequency distribution of scientific productivity of 51 US states in the two-years period 1978-79. This example is a part of a study on publication potential (Schubert &Telcs, 1987). The data were obtained from the Science Citation Index (SCI) of the ISI. The underlying model assumed the publication activity to have a Waring distribution, i.e.,

$$P(X=k) = \frac{\alpha}{N+\alpha} \frac{N}{N+\alpha+1} \cdots \frac{N+k-1}{N+\alpha+k}; \quad k = 0,1,2, \ldots ,$$

where X denotes the random variable "publication activity", and α, $N > 0$ are real parameters. The Waring distribution is Paretian with characteristic parameter α. The characteristic maximum u_1^* can be calculated after estimation of the parameters α and N according to the following equation (cf. Eq. (4.4)) :

$$u_1^* = G^{-1}(\lambda_1/n) = G^{-1}((\log 2)/n) .$$

The probabilities G_k can be explicitly expressed by the formula

$$G_k = \frac{N \ \ldots \ (N+k-1)}{(N+\alpha) \ldots (N+\alpha+k-1)} \quad ; \quad k \geq 0,$$

and thus u_1^* can easily be calculated. Table 3 contains the characteristic parameters, α, the values of the observed maximum productivity and the characteristic greatest values u_1^* for all the 51 US states.
We examine the hypothesis whether the maxima of the 51 samples fit the expected extreme values resulting from the sample size and the parameters of the assumed distribution. According to Eq. (4.6) we have $\nu_{51} = 21$. (Half of

Table 3. The characteristic parameters, α, the sample sizes, n, the observed maxima, X_1^*, and the characteristic greatest values, u_1^*, for 51 US states.

State	n	α	X_1^*	u_1^*	$\mathrm{sgn}(X_1^* - u_1^*)$
AK	167	3.58	14	9	+
AL	1316	5.02	15	16	-
AR	520	3.90	14	14	0
AZ	1620	5.90	13	15	-
CA	20414	5.81	21	28	-
CO	3091	4.13	25	23	+
CT	2996	5.49	15	18	-
DC	3680	6.62	12	15	-
DE	592	13.15	8	8	0
FL	3221	5.07	23	19	+
GA	2470	4.85	22	20	+
HI	695	3.78	17	16	+
IA	1986	4.84	21	18	+
ID	3410	10.60	6	7	-
IL	7741	5.44	19	23	-
IN	2720	3.50	38	25	+
KS	1222	5.31	13	14	-
KY	1167	4.87	23	15	+
LA	1505	3.66	27	20	+
MA	9505	4.31	39	31	+
MD	7344	6.34	32	23	+
ME	365	14.20	7	7	0
MI	4724	4.35	26	23	+
MN	3372	5.08	16	21	+
MO	2762	6.75	11	16	-
MS	696	8.85	9	10	-
MT	329	11.83	7	7	0
NC	3572	5.21	25	19	+
ND	336	4.27	16	11	+
NE	771	3.90	18	16	+
NH	496	7.66	9	9	0
NJ	4911	4.83	19	23	-
NM	1569	5.37	15	17	-
NV	176	2.80	15	12	+
NY	16571	5.05	33	30	+
OH	5149	4.96	28	20	+
OK	1143	4.76	15	13	+
OR	1630	7.55	11	12	-
PA	7336	4.06	27	30	+
RI	886	14.39	8	9	-
SC	994	3.33	34	19	+
SD	177	8.39	6	6	0
TN	2474	8.02	12	15	-
TX	7450	4.53	45	28	+
UT	1241	5.99	14	13	+
VA	2829	6.34	15	15	0
VT	334	8.52	9	8	+
WA	2983	6.48	13	16	-
WI	3204	3.95	35	24	+
WV	538	5.32	13	10	+
WY	217	33.04	6	6	0

the cases $X_1^* = u_1^*$ is considered as $X_1^* < u_1^*$). Hence $(2v_{51} - 51)/\sqrt{51} = -1.26$ follows, and thus our hypothesis can be accepted at a level of $p = 0.95$ (the corresponding critical value is 1.96).
The second example, a citation rate, is given for large sample tests described in the preceding section. Data were obtained from the same source as above. The considered papers were published in 22 analytical chemistry journals in the two years period 1981/82. Citations received by them were counted in 1983. As in the above example, a Waring distribution was assumed. The estimated parameters are $\hat{a} = 4.299$ and $\hat{N} = 4.721$. Table 4 presents observed and estimated frequencies of the distribution.

Table 4. Frequency distribution of analytical chemistry papers published in 1981/82.

k	frequency	
	observed	estimated
0	5965	6010.5
1	2810	2831.9
2	1549	1470.2
3	846	822.0
4	498	487.5
5	312	303.2
6	177	196.3
7	140	131.3
8	79	90.4
9	63	63.9
10	43	46.1
11	22	33.9
12	27	25.3
13	5	19.2
14	16	14.8
15	14	11.5
16	9	9.1
17	2	7.2
18	9	5.8
19	4	4.7
20	4	3.9
21	1	3.2
22	0	2.6
23	3	2.2
24	1	1.8
25	1	1.6
> 25	11	10.8

The sample size ($n = 12611$) seems to be large enough for applying a large sample test to this citation rate distribution. We want to answer the question whether the observed extreme citation rates of the sample (cf. Table 5) are really in accordance with the applied distribution model. Since 10 is the smallest integer not less than $\log 12611 \sim 9.4$ the test is based on a set of 10 sample elements.
The application of the Kolmogorov-Smirnov test to the comparison of the empirical distribution function F_{10} with the exponential distribution function with parameter $\hat{a} = 4.299$ has the result $K(10,x) = 0.71$. According to Table 2 the hypothesis must be rejected at any reasonable confidence level. In order to detect the tendency of deviation from the model an additional test is presented. Since the standard deviation of the estimated parameter $\hat{a}$ is unknown and the

sample size $k_0 = 10$ is small, instead of the sample mean the median is used, and, consequently, a sign test is applied. Data are presented in Table 6.

Table 5. The 10 greatest observations of the citation rate sample (n = 12611)

k	1	2	3	4	5	6	7	8	9	10
k^*	186	174	121	92	58	44	40	36	36	33

Table 6.

k	1	2	3	4	5	6	7	8	9	10
$k\log X_k^*/X_{k+1}^*$	0.067	0.727	0.822	1.845	1.206	0.572	0.738	0.000	0.783	2.007

The median of the exponential distributions is $m = \log 2/\hat{\alpha} \approx 0.161$. Because of the above test result, the counterhypothesis

$$H_1 : \nu = 0.1 \sum_{k=1}^{10} \chi\,(k.\log X_k^*/X_{k=1}^* > m) > 0.5$$

is examined. The critical value corresponding to a confidence level $p \leq 0.99$ is $t_p \leq 8$. Since $\nu = 8 \geq t_p = 8$ the deviation of the sample median must be considered as significantly greater than 0.161. The test results suggest a tail parameter signficantly less then $\hat{\alpha}$ for characterizing the tail of the distribution.

REFERENCES

Galambos, J., Kotz, S., Characterizations of Probability Distributions, (Springer-Verlag, Berlin, Heidelberg, New York, 1978).

Glänzel, W., Telcs, A., Schubert, A., Characterization of Truncated Moments and Its Application to Pearson-type Distributions. Z. Wahrscheinlichkeitstheorie verw. Gebiete, 66 (1984) p. 173-183.

Glänzel, W., Schubert, A., Telcs, A., Goodness of Fit Test for the Tail of Distributions, Bolyai *Colloquium on Goodness of fit* (Held in Debrecen on June 25-28, 1984).

Glänzel, W., Schubert, A., Price Distribution. An Exact Formulation of Price's "Square Root Law", Scientometrics, 7(3-6) (1985) p. 211-219.

Glänzel, W., A Characterization Theorem Based on Truncated Moments and Its Application to some Distribution Families. In : Wertz, W. & al. (Eds.), Proceedings of the Sixth Pannonian Symposium on Mathematical Statistics, (Reidel, Dordrecht, Holland) (to be published in 1987).

Gumbel, E.J., Statistics of Extremes, (Columbia University Press, New York, 1958).

Renyi, A., Wahrscheinlichkeitsberechnung, (VEB Deutscher Verlag der Wissenschaften, Berlin 1962).

Schubert, A., Telcs, A., Estimation of Publication Potential in 51 US States Based on the Frequency Distribution of Scientific Productivity, Journal of the American Society for Information Science (to be published).