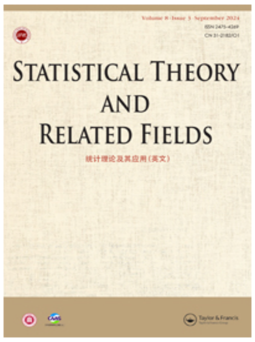

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tstf20

Goodness of fit for the Waring distribution

Yanlin Tang, Jinglong Wang, Menghan Yi & Zhongyi Zhu

To cite this article: Yanlin Tang, Jinglong Wang, Menghan Yi & Zhongyi Zhu (04 Sep 2024): Goodness of fit for the Waring distribution, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2024.2389758

To link to this article: https://doi.org/10.1080/24754269.2024.2389758

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

View supplementary material

Published online: 04 Sep 2024.

Submit your article to this journal

Article views: 109

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tstf20

STATISTICAL THEORY AND RELATED FIELDS
https://doi.org/10.1080/24754269.2024.2389758

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Goodness of fit for the Waring distribution

Yanlin Tang[a†], Jinglong Wang[a†], Menghan Yi[a†] and Zhongyi Zhu[b†]

[a]KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai, People's Republic of China; [b]Department of Statistics and Data Science, Fudan University, Shanghai, People's Republic of China

ABSTRACT
The Waring distribution is an important two-parameter discrete distribution, commonly used in fields such as ecology, linguistics, and information science, where heavy tails are often observed. In this paper, we propose a new goodness-of-fit test for the Waring distribution, which is established through the hazard rate and a linear equivalent definition of the Waring distribution. We establish an asymptotic Chi-square null distribution for the proposed test and show that it is more powerful than classical methods in simulation studies. Finally, we apply the test to analyze the authorships of published papers on computer science.

ARTICLE HISTORY
Received 5 December 2023
Revised 12 April 2024
Accepted 31 July 2024

KEYWORDS
Waring distribution; goodness-of-fit test; long-tailed distribution; hazard rate

1. Introduction

The Waring distribution is an important two-parameter long-tailed discrete distribution, which can be used to describe the cumulative advantage distribution of the 'success breeds success' mechanism. For example, Price (1965, 1976) connected the published literature with the cited literature to form a directed network, and found that the number of citations of the literature follows a special Waring distribution. Huete-Morales and Marmolejo-Martín (2020) fitted the number of organic livestock farms to a Waring distribution. Recently, there have been some works on the Waring distribution, such as the EM algorithm (Cueva-López et al., 2019), generalizations of the distribution (Cueva-López et al., 2021; Rivas & Campos, 2021) and the MLE (Tang et al., 2023). However, most of the current literature focuses on parameter estimation, while the literature on inference is quite limited. In this paper, we propose a goodness-of-fit test for the Waring distribution.

A commonly used goodness-of-fit test is the Pearson's Chi-square test. However, we need to group the empirical data for the Chi-square test, while different grouping methods may lead to different conclusions, making the inferences less convincing. Under certain conditions for the expected frequency of each group, Haberman (1988), Rempała and Wesołowski (2016) and Chang et al. (2023) explored the performance of Pearson's Chi-square, but these conditions are not always satisfied. Another commonly used distribution test is the Kolmogorov-Smirnov (KS) test. However, the standard table for the KS test is no longer valid if there are unknown parameters that must be estimated from the sample. Lilliefors (1967, 1969), Goldstein et al. (2004), and Clauset et al. (2009) constructed new tables of KS statistics for specified distributions with unknown parameters. Other literature focuses on applications of the KS test, such as mixed-censored life data (Banerjee & Pradhan, 2018), grouped data (Okamura & Dohi, 2019) and homogeneity generated variables (Otsu & Taniguchi, 2020). The Pearson's and KS tests are nonparametric and widely applicable, but they can also lead to inefficiencies for specified distributions. Our proposed test is specifically designed for the Waring distribution and thus more accurate.

Our motivation lies in the following points. Firstly, a heavy tailed distribution implies that extreme events occur more frequently than with a standard normal distribution. Therefore, we hope to accurately identify heavy-tailed distributions, which can help to assess the probability and impact of extreme events. Secondly, non-parametric tests may not accurately capture information related to specific distributions. Therefore, we specially design a goodness-of-fit test for the Waring distribution, which has stronger interpretability in certain application scenarios and higher statistical efficiency in small sample sizes. In practice, the choice of parametric and nonparametric tests should be comprehensively considered based on the research objectives and the nature of the data. Thirdly, although the Waring distribution has been applied in different fields, there is relatively little research on its goodness-of-fit tests. Our objective is to fill the gaps in existing literature.

CONTACT Yanlin Tang ✉ yltang@fem.ecnu.edu.cn KLATASDS-MOE, School of Statistics, East China Normal University, Shanghai 200062, People's Republic of China

[†] All the authors contribute equally to the paper.

Supplemental data for this article can be accessed online at http://dx.doi.org/10.1080/24754269.2024.2389758.

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

In this paper, we first reformulate the goodness-of-fit problem as a multiple linear test problem, and then construct an approximate Chi-square statistic. The proposed test faces two challenges. First, we need to find an equivalent definition of the Waring distribution, which requires appropriate modification of the Waring's hazard rate. Second, we need to find the orthogonal matrix to construct mutually independent and approximately standard normal statistics in order to establish the Chi-square statistic.

The rest of the paper is organized as follows. In Section 2, we describe the theoretical properties of the Waring distribution, which help us to propose a new goodness-of-fit test in Section 3. In Section 4, we conduct simulation studies to compare the performance of the proposed and classical tests in terms of size and power studies. In Section 5, we apply the proposed goodness-of-fit test to analyze the authorships of published papers on computer science. All technical proofs are provided in the online Supplementary Materials.

2. A characterization of the Waring distribution

In this section, we present the characteristics of the Waring distribution, including its tail probabilities and an equivalent definition.

A discrete random variable X has a Waring distribution if its probability mass function is

$$p_k := P(X=k) = \alpha\frac{\Gamma(\alpha+\beta)\Gamma(\beta+k-1)}{\Gamma(\beta)\Gamma(\alpha+\beta+k)}, \quad \alpha>0,\ \beta>0,\ k=1,2,\ldots, \tag{1}$$

where $\Gamma(\cdot)$ is the Gamma function. The Waring distribution is a highly skewed distribution with a long right tail whose tail probabilities have the following properties.

Lemma 2.1: *For the Waring probability distribution* (1), *its right tail probability is*

$$\sum_{i=k}^{\infty} p_i = \frac{\Gamma(\alpha+\beta)\Gamma(\beta+k-1)}{\Gamma(\beta)\Gamma(\alpha+\beta+k-1)}.$$

To further understand the tail properties of the Waring distribution, we characterize its convergence speed through the hazard rate. In discrete distributions, the hazard rate (Barlow et al., 1963) is defined as the conditional probability of $X=k$ given $X\geq k$, that is,

$$q_k := P(X=k\,|\,X\geq k) = p_k\Big/\left(\sum_{i=k}^{\infty} p_i\right),$$

with integers k satisfying $\sum_{i=k}^{\infty} p_i > 0$. It refers to the conditional probability that an event, which has not yet occurred, will occur at the next moment, under certain time or conditions. Therefore, the risk rate can be used to measure the risk of future events occurring. According to Lemma 2.1, we can calculate the hazard rate of the Waring distribution as $q_k = \alpha/(\alpha+\beta+k-1)$, which decreases as k increases. This implies that although both p_k and $\sum_{i=k}^{\infty} p_i$ decrease with k, $\sum_{i=k}^{\infty} p_i$ decreases more slowly than p_k. This is not surprising in long-tailed distributions.

Another key role of the hazard ratio is that it motivates our goodness-of-fit tests. In order to transform the Waring distribution into a valid and easily testable form, we define a transformation of the hazard rate as

$$a_k = \left(\sum_{i=k+1}^{\infty} p_i\right)\Big/p_k, \quad k=1,2,\ldots. \tag{2}$$

Then, under the Waring distribution, we can calculate $q_k = 1/(1+a_k)$ and $a_k = (\beta+k-1)/\alpha$. Therefore, a distribution over the positive integer is a Waring distribution only if $a_k = bk+c, k=1,2,\ldots$, where $b=1/\alpha$ and $c=(\beta-1)/\alpha$. In the following Theorem 2.1, the key feature of the Waring distribution tells us that the converse is also true.

Theorem 2.1 (Equivalent definition of the Waring distribution): *A distribution over the positive integers is a Waring distribution if and only if*

$$a_k = bk+c, \quad k=1,2,\ldots, \tag{3}$$

where b and c are constants satisfying $b>0$, $b+c>0$.

Similar to q_k, a_k increases with k, which indicates that $\sum_{i=k+1}^{\infty} p_i$ decreases more slowly than p_k, reflecting the long-tailed characteristics of the Waring distribution.

3. A test of fit for the Waring distribution

3.1. Reformulation of hypothesis testing

Let X be a random variable distributed over positive integers. According to Theorem 2.1, the hypothesis

$$H_0 : X \text{ is distributed as a Waring distribution}$$

is equivalent to the multiple comparison test with two hypotheses

$$H_{01}: \ a_k = bk + c, \ k = 1, 2, \dots, \tag{4}$$

$$H_{02}: \ b > 0. \tag{5}$$

If these two hypotheses are accepted, then $b + c > 0$ can be deduced from $a_1 = (1 - p_1)/p_1 = b + c$. Next, we use Lemma 3.1 to show that hypothesis (4) does not imply (5), so it is necessary to test hypothesis (5).

Lemma 3.1: *There exist discrete random variables whose distribution satisfies* (4) *but* $b \le 0$. *The details are as follows.*

- *When* $b = 0$, *a distribution on positive integers is a geometric distribution* $p_k = \frac{1}{c+1} \cdot (\frac{c}{c+1})^{k-1}$ *if and only if* $a_k = c, k = 1, 2, \dots$, *where* $c > 0$.
- *When* $b < 0$ *and there is a positive integer* k' *such that* $bk' + c = 0$, *then there is a random variable* X *with probability masses*

$$p_1 = \frac{1}{b + c + 1}, \quad p_i = \frac{b(i-1) + c}{bi + c + 1} p_{i-1}, \quad i = 2, 3, \dots, k',$$
$$p_i = 0, \quad i = k' + 1, \dots,$$

  so that (4) *holds.*
- *When* $b < 0$ *and there is no positive integer* k' *satisfying* $bk' + c = 0$, *then there is no random variable* X *on the positive integers, which makes Equation* (4) *hold.*

3.2. Asymptotic Chi-square test

Suppose that $X_1, \dots, X_n$ is a random sample from the discrete random variable X, whose value range is $\{1, 2, \dots\}$. Let $m = \max\{X_1, \dots, X_n\}$ be the largest observed value, n_k be the number of sample observations equal to $k = 1, 2, \dots, m$, such that $\sum_{k=1}^{m} n_k = n$. Therefore, the probability p_k can be consistently estimated by frequency n_k/n, and a_k can be consistently estimated by

$$\widehat{a}_k = \frac{1 - \sum_{i=1}^{k} n_i/n}{n_k/n} = \frac{n - \sum_{i=1}^{k} n_i}{n_k}, \quad k = 1, 2, \dots, m, \tag{6}$$

for $n_k > 0$. If the sample comes from a Waring distribution, a scatterplot of $\{(k, \widehat{a}_k), \ k = 1, 2, \dots, m\}$ can be fitted to a straight line with a positive slope. This step can be used as a preliminary data check. Next, we construct the asymptotic Chi-square statistic for the hypotheses.

Without losing generality, we assume that $n_k \ge 1$ for $k = 1, 2, \dots, m$. Since $\sum_{k=1}^{m} n_k = n$ makes $\widehat{a}_m = 0$, we give $k \le m - 1$ and construct a statistic in the form of $T_n(\widehat{a}_1, \widehat{a}_2, \dots, \widehat{a}_k)$ based on the data $\{n_1, \dots, n_k, n_{k+1}\}$ with $n_{k+1} = n - \sum_{i=1}^{k} n_i$. Note that $\{n_1, \dots, n_k, n_{k+1}\}$ is distributed as the multinomial distribution Multi$(p_1, \dots, p_k, p_{k+1})$ with $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$, then $(\widehat{a}_1, \widehat{a}_2, \dots, \widehat{a}_k)$ can be derived to be asymptotically normal, as shown in Lemma 3.2.

Lemma 3.2: *For every* $1 \le k \le m - 1$, *let* $\boldsymbol{a} = (a_1, a_2, \dots, a_k)^\top$ *and* $\widehat{\boldsymbol{a}} = (\widehat{a}_1, \widehat{a}_2, \dots, \widehat{a}_k)^\top$ *be defined in* (2) *and* (6), *respectively. Then we have*

$$\sqrt{n}(\widehat{\boldsymbol{a}} - \boldsymbol{a}) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{\Sigma}_k), \tag{7}$$

where $\boldsymbol{\Sigma}_k = diag(\sigma_{11}, \dots, \sigma_{kk})$ *is a diagonal matrix with*

$$\sigma_{11} = \frac{1 - p_1}{p_1^3}, \quad \sigma_{ii} = \frac{(1 - \sum_{j=1}^{i-1} p_j)(1 - \sum_{j=1}^{i} p_j)}{p_i^3}, \quad i = 2, 3, \dots, k.$$

Table 1. Type I errors for 1000 MC simulations at a significance level of 0.05, for $\alpha = 0.5$ and 1.

(α, β)	n	50	100	200	500	1000	2000
(0.5, 0.5)	T	0.001	0.000	0.005	0.080	0.038	0.032
	PS	0.035	0.030	0.032	0.025	0.037	0.046
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(0.5, 1)	T	0.038	0.005	0.012	0.044	0.018	0.030
	PS	0.024	0.019	0.027	0.040	0.044	0.056
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(0.5, 1.5)	T	0.092	0.047	0.018	0.033	0.020	0.022
	PS	0.018	0.019	0.024	0.039	0.046	0.056
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(0.5, 2)	T	0.156	0.099	0.053	0.033	0.021	0.027
	PS	0.016	0.019	0.035	0.038	0.048	0.042
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(1, 0.5)	T	0.002	0.000	0.006	0.022	0.025	0.029
	PS	0.054	0.059	0.043	0.038	0.033	0.034
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(1, 1)	T	0.007	0.000	0.003	0.016	0.020	0.030
	PS	0.039	0.043	0.030	0.033	0.038	0.042
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(1, 1.5)	T	0.037	0.009	0.006	0.024	0.018	0.012
	PS	0.028	0.044	0.031	0.031	0.038	0.051
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(1, 2)	T	0.104	0.058	0.025	0.015	0.019	0.021
	PS	0.023	0.024	0.025	0.023	0.038	0.058
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000

The proof steps are similar to Theorem 3.1 of Wang et al. (2023). By normalizing (7), we can get

$$\sqrt{n}(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}) \xrightarrow{D} N_k(\mathbf{0}, \boldsymbol{I}_k),$$

where $\widehat{\mathbf{\Lambda}} = (\widehat{a}_1/\sqrt{\sigma_{11}}, \dots, \widehat{a}_k/\sqrt{\sigma_{kk}})^\top$, and $\mathbf{\Lambda} = (a_1/\sqrt{\sigma_{11}}, \dots, a_k/\sqrt{\sigma_{kk}})^\top$. Now we construct the Chi-square test statistic based on the orthogonal transformation of $\widehat{\mathbf{\Lambda}}$, whose explicit form and asymptotic distribution are given by the following Theorem 3.1.

Theorem 3.1: *For every $k = 3, 4, \dots, m-1$, the statistic*

$$T(\widehat{\boldsymbol{a}}) = n\left\{\sum_{i=1}^{k} \widehat{a}_i^2/\widehat{\sigma}_{ii} - (\widehat{y}_{1k}^2 + \widehat{y}_{2k}^2)\right\} \xrightarrow{D} \chi^2(k-2),$$

under the null hypothesis (4), *where*

$$\widehat{\sigma}_{11} = \frac{n^2(n-n_1)}{n_1^3},$$

$$\widehat{\sigma}_{ii} = \frac{n(n - \sum_{j=1}^{i-1} n_j)(n - \sum_{j=1}^{i} n_j)}{n_i^3}, \quad i = 2, 3, \dots, k,$$

$$\widehat{y}_{1k} = \frac{1}{\widehat{s}_{1k}} \sum_{i=1}^{k} \frac{\widehat{a}_i}{\widehat{\sigma}_{ii}}, \quad \widehat{y}_{2k} = \frac{1}{\widehat{s}_{3k}} \sum_{i=1}^{k} \frac{\widehat{a}_i}{\widehat{\sigma}_{ii}}\left(i - \frac{\widehat{s}_{2k}^2}{\widehat{s}_{1k}^2}\right),$$

$$\widehat{s}_{1k}^2 = \sum_{i=1}^{k} \frac{1}{\widehat{\sigma}_{ii}}, \quad \widehat{s}_{2k}^2 = \sum_{i=1}^{k} \frac{i}{\widehat{\sigma}_{ii}}, \quad \widehat{s}_{3k}^2 = \sum_{i=1}^{k} \frac{i^2}{\widehat{\sigma}_{ii}} - \frac{\widehat{s}_{2k}^4}{\widehat{s}_{1k}^2}.$$

Furthermore, under the null hypothesis (4), $\widehat{y}_{1k}$ *and* $\widehat{y}_{2k}$ *asymptotically converge to the standard normal distribution:*

$$\sqrt{n}\{\widehat{y}_{1k} - (c \cdot \widehat{s}_{1k} + b \cdot \widehat{s}_{2k}^2/\widehat{s}_{1k})\} \xrightarrow{D} N(0,1),$$

$$\sqrt{n}(\widehat{y}_{2k} - b \cdot \widehat{s}_{3k}) \xrightarrow{D} N(0,1).$$

Table 2. Type I errors for 1000 MC simulations at a significance level of 0.05, for $\alpha = 1.5$ and 2.

(α, β)	n	50	100	200	500	1000	2000
(1.5, 0.5)	T	0.019	0.001	0.003	0.012	0.015	0.017
	PS	0.074	0.067	0.053	0.048	0.040	0.055
	KS	0.999	1.000	1.000	1.000	1.000	1.000
	CVM	0.999	1.000	1.000	1.000	1.000	1.000
(1.5, 1)	T	0.002	0.000	0.005	0.010	0.017	0.022
	PS	0.053	0.054	0.039	0.031	0.032	0.040
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(1.5, 1.5)	T	0.023	0.006	0.002	0.009	0.026	0.029
	PS	0.046	0.046	0.034	0.023	0.029	0.041
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(1.5, 2)	T	0.050	0.020	0.007	0.017	0.016	0.022
	PS	0.039	0.032	0.029	0.031	0.031	0.042
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(2, 0.5)	T	0.069	0.004	0.001	0.010	0.012	0.018
	PS	0.079	0.070	0.064	0.052	0.051	0.048
	KS	0.989	1.000	1.000	1.000	1.000	1.000
	CVM	0.989	1.000	1.000	1.000	1.000	1.000
(2, 1)	T	0.007	0.000	0.003	0.006	0.018	0.024
	PS	0.063	0.063	0.056	0.048	0.040	0.044
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(2, 1.5)	T	0.012	0.003	0.006	0.009	0.020	0.020
	PS	0.055	0.055	0.055	0.042	0.042	0.049
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000
(2, 2)	T	0.040	0.014	0.003	0.006	0.030	0.030
	PS	0.050	0.047	0.033	0.039	0.030	0.033
	KS	1.000	1.000	1.000	1.000	1.000	1.000
	CVM	1.000	1.000	1.000	1.000	1.000	1.000

It is easy to see that, a larger $T(\widehat{\boldsymbol{a}})$ indicates a larger deviation from (4), and a smaller $\widehat{y}_2$ indicates a larger deviation from (5). For the multiple comparison test with two hypotheses, we take the significance level γ. Then according to the Bonferroni inequality, the hypotheses (4) and (5) are accepted when $T(\widehat{\boldsymbol{a}}) < \chi^2_{1-\gamma/2}(k-2)$ and $\widehat{y}_{2k} > -Z_{1-\gamma/2}/\sqrt{n}$. Furthermore, the $1-\gamma$ confidence intervals for b and parameter $\alpha = 1/b$ can be taken as

$$\left(\frac{\widehat{y}_{2k} - Z_{1-\gamma/2}/\sqrt{n}}{\widehat{s}_{3k}}, \frac{\widehat{y}_{2k} + Z_{1-\gamma/2}/\sqrt{n}}{\widehat{s}_{3k}}\right),$$

$$\left(\frac{\widehat{s}_{3k}}{\widehat{y}_{2k} + Z_{1-\gamma/2}/\sqrt{n}}, \frac{\widehat{s}_{3k}}{\widehat{y}_{2k} - Z_{1-\gamma/2}/\sqrt{n}}\right).$$

It can be seen that the test statistic $T(\widehat{\boldsymbol{a}})$ is actually used to test the hypothesis

$$H_{01} : a_i = bi + c, \quad i = 1, 2, \ldots, k, \tag{8}$$

which is weaker than hypothesis (4). Therefore, the rejection of (8) implies the rejection of (4). Theoretically, the choice of k can be any value between 3 and $m-1$. However, in practice, the choice of k will be affected by tail outliers. This is because the right tail of the Waring distribution has a smaller probability mass, and outliers can introduce significant bias to the test. Therefore, in practice, we usually first roughly determine the position of the outliers in the tail. The value of k should then be chosen before this position to ensure that the test is not influenced by tail outliers. Another simpler approach is to let k take different values between 3 and $m-1$, comprehensively analyze the results of this series of tests, and then decide whether to reject or accept the hypothesis of the Waring distribution. For instance, in Section 5, we initially set k to range from 3 to 60 for a series of tests. However, we finally set $k = 3 \sim 24$ and deleted the tail, because the presence of outliers in the tail at $k > 24$ caused a reversal of the test results.

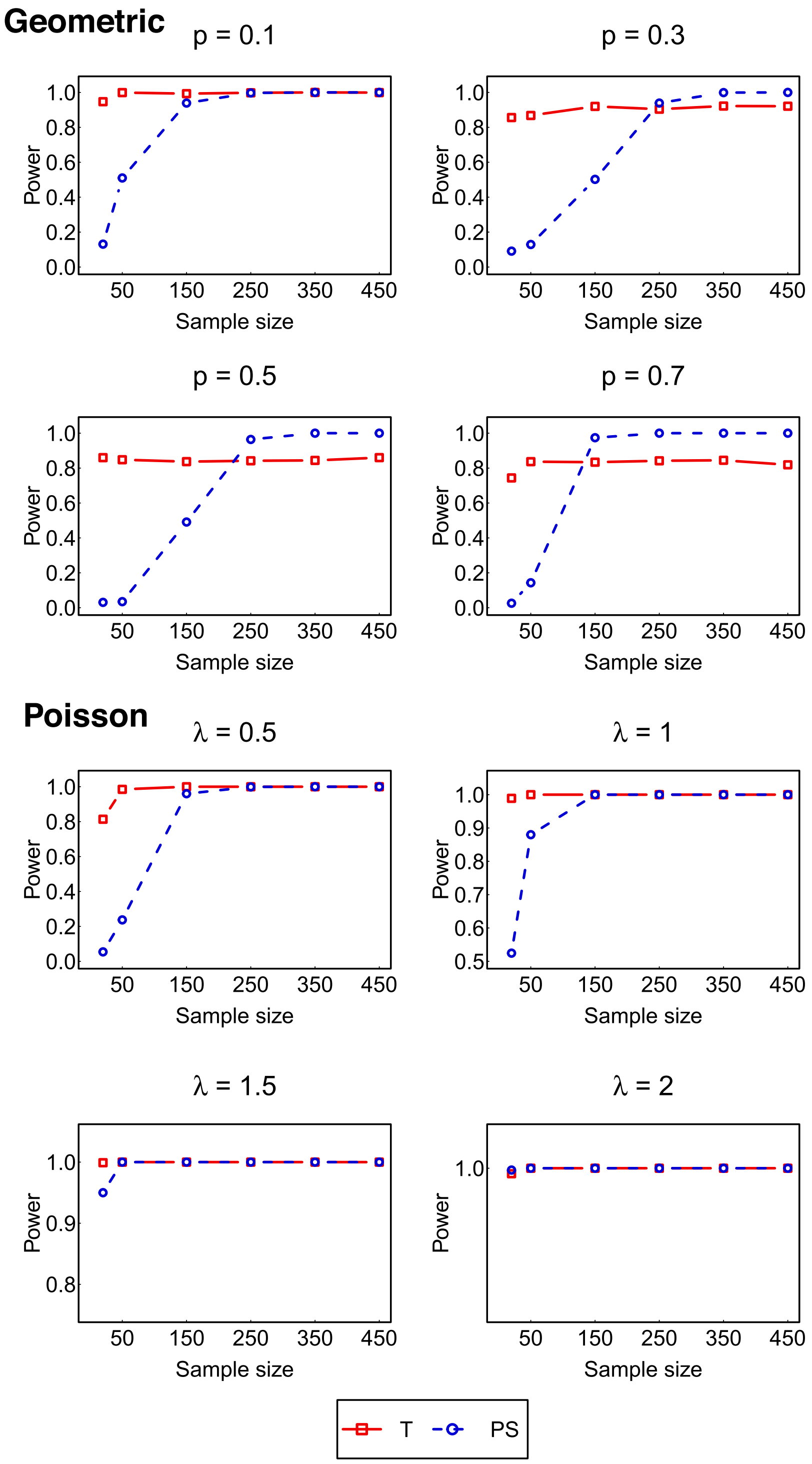

Figure 1. Powers of 1000 MC simulations at a significance level of 0.05 for the proposed (T) and Pearson's Chi-square (PS) tests.

4. Simulation studies

In this section, we compare the proposed test with the well-known Pearson's Chi-square (PS), Kolmogorov-Smirnov (KS), Cramér-von Mises (CVM) defined as follows,

$$\mathrm{PS} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

$$\mathrm{KS} = \sqrt{n} \max\left\{ \max_{1\le i\le n} \left| \frac{i}{n} - F_0(x_{(i)}) \right|, \max_{1\le i\le n} \left| F_0(x_{(i)}) - \frac{i-1}{n} \right| \right\},$$

$$\mathrm{CVM} = \sum_{i=1}^{n} \left\{ F_0(x_{(i)}) - \frac{2i-1}{2n} \right\}^2 + \frac{1}{12n},$$

where n_i is the observed frequency of point i, p_i is the probability of the Waring distribution at i, $x_{(i)}$ is the ith order statistic of the observed value, and F_0 is the distribution function of the Waring distribution. The null distributions of PS, KS, and CVM are $\chi^2(k-1)$, Kolmogorov-Smirnov distribution, and Cramer-von Mises distribution, respectively. We compare the Type I error probability and power for the four statistics under 1000 Monte Carlo repetitions at a critical value $\gamma\ = 0.05$ and different sample sizes n.

For the Type I error, we generate data from the Waring distribution, where the parameters α and β are taken from the set $(0.5, 1, 1, 5, 2)$, and the maximum observed value is 30. Let k in the proposed test $T(\widehat{\boldsymbol{a}})$ be $\max(3, 0.5m)$, where m is the maximum value of the generated data. Under the sample size of $n = 50 \sim 2000$, the results of the Type I errors of the four methods are shown in Tables 1 and 2, in which T is the proposed method. PS, KS and CVM are Pearson's Chi-square, Kolmogorov-Smirnov, Cramér-von Mises tests respectively. The two tables show the results for different (α, β), and it can be seen that the proposed test T and PS test can guarantee a reasonable Type I error probability. However, the KS and CVM tests are invalid with results exceeding 0.99, because their statistics are constructed based on continuous distribution function.

For the power, we generate data from four discrete distributions: (i) Geometric distribution with success probabilities $p = (0.1, 0.3, 0.5, 0.7)$; (ii) Poisson distribution with parameters $\lambda = (0.5, 1.0, 1.5, 2.0)$; (iii) Binomial distribution with total number of trials 5 and success probabilities $p = (0.1, 0.3, 0.5, 0.7)$; (iv) Discrete uniform distribution over $[L, R]$, that is, generating integer points on $[L, R]$ with equal probability, where we set $L = 1$ and $R = (3, 4, 5, 6)$. Under the sample size of $n = 50 \sim 2000$, the power results are shown in Figures 1 and B.1 of the online Supplementary Materials. The powers of KS and CVM are not included due to their highly inflated Type I errors. It can be seen that the proposed test T is more powerful than PS under small samples. For example, in Geometric distribution data, when $p = 0.7$, $n = 50$, the power of the proposed test is 0.848, while that of PS is only 0.035. Therefore, the proposed test is the most powerful method while guaranteeing a reasonable Type I error.

5. A real data application

Kang et al. (2007) conducted an extensive analysis of five core Chinese computer science journals between 1993 and 2002, and extracted the empirical data in Table B.1 in the online Supplementary Material. A total of 12,509 papers were published by 5798 authors. The scatter plot of $(k, \widehat{\alpha}_k)$ is roughly a straight line, as shown in Figure B.2 of the supplementary materials. It is preliminarily judged that the Waring distribution may be a good fit for the empirical data.

The results of goodness-of-fit are shown in Table 3, in which k is the number of papers; T is the proposed test; $\chi^2_{0.025}$ is the critical value. It can be seen that the data of papers published from 1 to 24 can be fitted with the Waring distribution, but the rest parts are not. We speculate that the reversal of the statistical inference conclusions may be due to the presence of outliers at the right tails, such as $n_{27} = 3$, $n_{60} = n_{62} = 1$, or excessively large gaps, such as $n_{32} = \cdots = n_{41} = 0$ and $n_{53} = \cdots = n_{59} = 0$. We generally do not use these outlier data to make statistical inferences. In short, according to the specific analysis of specific problems, we think that the assumption (4) is acceptable in terms of the value of T on $k = 3 \sim 24$.

Table 3. Waring distribution test for hypothesis (4).

k	T	$\chi^2_{0.025}$	p-value	k	T	$\chi^2_{0.025}$	p-value
3	0.007	5.024	0.936	19	16.178	30.191	0.511
4	1.570	7.378	0.456	20	16.342	31.526	0.569
5	4.724	9.348	0.193	22	23.574	34.170	0.213
6	5.180	11.143	0.269	23	23.811	35.479	0.251
7	6.183	12.833	0.289	24	25.537	36.781	0.225
8	6.761	14.449	0.349	25	39.777	38.076	0.011
9	7.576	16.013	0.371	26	40.982	39.364	0.012
10	8.214	17.535	0.413	27	56.925	40.646	0.000
11	8.267	19.023	0.508	28	57.044	41.923	0.000
12	8.308	20.483	0.599	29	67.406	43.195	0.000
13	9.131	21.920	0.610	31	68.631	45.722	0.000
14	10.588	23.337	0.565	42	73.927	59.342	0.000
15	10.698	24.736	0.636	44	83.815	61.777	0.000
16	10.948	26.117	0.690	47	103.979	65.410	0.000
17	10.950	27.488	0.765	52	151.135	71.420	0.000
18	12.261	28.845	0.726	60	276.173	80.936	0.000

Table 4. Waring distribution test for hypothesis (5).

k	97.5% interval for b	k	97.5% interval for b
3	(1.585, 2.077)	14	(1.839, 2.199)
4	(1.692, 2.138)	15	(1.844, 2.201)
5	(1.796, 2.229)	16	(1.849, 2.206)
6	(1.787, 2.192)	17	(1.850, 2.206)
7	(1.772, 2.159)	18	(1.864, 2.221)
8	(1.794, 2.174)	19	(1.893, 2.256)
9	(1.784, 2.155)	20	(1.892, 2.254)
10	(1.802, 2.169)	22	(1.936, 2.308)
11	(1.801, 2.163)	23	(1.935, 2.307)
12	(1.806, 2.164)	24	(1.949, 2.324)
13	(1.820, 2.178)		

Next, we test the hypothesis (5), and the 97.5% confidence intervals of b corresponding to the data of published papers from 1 to $3 \sim 24$ are shown in Table 4, in which k is the number of papers. These interval estimates suggest that the empirical data from the distribution of authors can be better fitted to the Waring distribution.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is partially supported by the National Natural Science Foundation of China 12371265, Natural Science Foundation of Shanghai 21ZR1420700, and Fundamental Research Funds for the Central Universities (2022QKT001).

Supplementary Materials

The online Supplementary Materials include all technical proofs and some additional numerical results from the simulation studies and real data analysis.

References

Banerjee, B., & Pradhan, B. (2018). Kolmogorov–Smirnov test for life test data with hybrid censoring. *Communications in Statistics-Theory and Methods*, *47*(11), 2590–2604. https://doi.org/10.1080/03610926.2016.1205616

Barlow, R. E., Marshall, A. W., & Proschan, F. (1963). Properties of probability distributions with monotone hazard rate. *Annals of Mathematical Statistics*, *34*(2), 375–389. https://doi.org/10.1214/aoms/1177704147

Chang, S., Li, D., & Qi, Y. (2023). Pearson's goodness-of-fit tests for sparse distributions. *Journal of Applied Statistics*, *50*(5), 1078–1093. https://doi.org/10.1080/02664763.2021.2017413

Clauset, A., Shalizi, C., & Newman, M. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. https://doi.org/10.1137/070710111

Cueva-López, V., Olmo-Jiménez, M. J., & Rodríguez-Avi, J. (2019). EM algorithm for an extension of the Waring distribution. *Computational and Mathematical Methods*, *1*(5), e1046.

Cueva-López, V., Olmo-Jiménez, M. J., & Rodríguez-Avi, J. (2021). An over and underdispersed biparametric extension of the Waring distribution. *Mathematics*, *9*(2), 170. https://doi.org/10.3390/math9020170

Goldstein, M., Morris, S., & Yen, G. (2004). Problems with fitting to the power-law distribution. *The European Physical Journal B*, *41*(2), 255–258. https://doi.org/10.1140/epjb/e2004-00316-5

Haberman, S. J. (1988). A warning on the use of Chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association*, *83*(402), 555–560. https://doi.org/10.1080/01621459.1988.10478632

Huete-Morales, M. D., & Marmolejo-Martín, J. A. (2020). The Waring distribution as a low-frequency prediction model: A study of organic livestock farms in Andalusia. *Mathematics*, *8*(11), 2025. https://doi.org/10.3390/math8112025

Kang, L., Xu, W., & Jiang, L. (2007). Waring distribution and the application of its parameter estimation method (in Chinese). *Statistics and Decision-making*, *9*, 138–139.

Lilliefors, H. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*(318), 399–402. https://doi.org/10.1080/01621459.1967.10482916

Lilliefors, H. (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, *64*, 387–389. https://doi.org/10.1080/01621459.1969.10500983

Okamura, H., & Dohi, T. (2019). On Kolmogorov-Smirnov test for software reliability models with grouped data. In *2019 IEEE 19th International Conference On Software Quality, Reliability And Security (QRS), Sofia, Bulgaria* (pp. 77–82). IEEE.

Otsu, T., & Taniguchi, G. (2020). Kolmogorov–Smirnov type test for generated variables. *Economics Letters*, *195*, 109401. https://doi.org/10.1016/j.econlet.2020.109401

Price, D. (1965). Network of scientific papers. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510

Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/10.1002/asi.v27:5

Rempała, G. A., & Wesołowski, J. (2016). Double asymptotics for the Chi-square statistic. *Statistics & Probability Letters*, *119*, 317–325. https://doi.org/10.1016/j.spl.2016.09.004

Rivas, L., & Campos, F. (2021). Zero inflated Waring distribution. *Communications in Statistics – Simulation and Computation*, *52*, 1–16.

Tang, Y., Wang, J., & Zhu, Z. (2023). On the MLE of the Waring distribution. *Statistical Theory and Related Fields*, *7*(2), 144–158. https://doi.org/10.1080/24754269.2023.2176608

Wang, J., Wu, X., & Xi, H. (2023). *A novel test on yule distributions* [Tech. Rep.]. East China Normal University.