应用概率统计 第 40 卷
第 6 期 2024 年 12 月

Chinese Journal of Applied Probability and Statistics
Dec., 2024, Vol. 40, No. 6, pp. 1000–1015
doi: 10.12460/j.issn.1001-4268.aps.2024.2023004

Parameter Interval Estimation for Yule-Simon Distribution*

DENG Wenli†[1] WANG Liming[2] WANG Jinglong[3]

[1] School of Mathematics and Statistics, Jiangxi Normal University, Nanchang, 330022

[2] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, 200433

[3] School of Statistics, East China Normal University, Shanghai, 200062

Abstract: Yule-Simon distribution has a wide range of practical applications, such as in network science, biology and humanities. A lot of work focuses on the study of how well the empirical data fits Yule-Simon distribution or how to estimate the parameter. There are still some open problems, such as the error analysis of parameter estimation, the theoretical proof of the convergence of the iterative algorithm for maximum likelihood estimation of parameters. The Yule-Simon distribution is a heavy-tailed distribution and the parameter is usually less than 2, so the variance does not exist. This makes it difficult to give an interval estimation of the parameter. Using the compression transformation, this paper proposes a method of interval estimation based on the central limit theorem. This method can be applied to many heavy-tailed distributions. The other two asymptotic confidence intervals of the parameter are obtained based on the maximum likelihood and the mode method. These estimation methods are compared in simulations and applications to empirical data.

Keywords: Yule-Simon distribution; maximum likelihood estimation; confidence interval; compression transformation

2020 Mathematics Subject Classification: 62B05; 62F10; 62F25; 62P10

Citation: DENG W L, WANG L M, WANG J L. Parameter interval estimation for Yule-Simon distribution [J]. *Chinese J Appl Probab Statist*, 2024, **40**(**6**): 1000–1015.

1 Introduction

George Udny Yule was one of the greatest statisticians from the late 19th century to the mid-20th century. He extended statistics to the fields of biology, sociology and economics, and promoted the progress of statistical history. The Yule distribution was first discovered and proposed by him in species research. Genus and species are two categories in biological classification. Species is the most basic taxonomic unit, and genus is the nearest upper layer of species. Most genera have only one species, while a few genera have many species. Yule[1] proposed a simple model to describe this phenomenon, which was

* This work was supported by the National Natural Science Foundation of China (Grant No. 11961035) and Jiangxi Provincial Natural Science Foundation (Grant No. 20224BCD41001).

† Corresponding author, E-mail: wldfudan@126.com.

Received on January 3, 2023. Revised on January 18, 2023. Accepted on May 3, 2023.

later called the Yule process and widely used. Thirty years later, American psychologist Simon[2] found similarities in some empirical distributions. The two statisticians found this statistical law from different aspects, so in many literature it was called Yule-Simon distribution. Newman[3] described some of the mechanisms by which power-law behaviour can arise and he deemed that the Yule process is one of the most convincing and widely applicable mechanisms for generating power laws. The Yule-Simon distribution was also understood as the preferential attachment processes of complex networks proposed by Barabási and Albert[4] and Bornholdt and Ebel[5]. In summary, the Yule-Simon distribution is related not only to the distribution of the number of species, the frequency distribution of words and complex network degree distribution, but also to other stochastic data such as distributions of incomes by size, distributions of cities by population, distributions of scientists by number of papers published and so on.

The distribution law of Yule-Simon distribution is

$$f(k,\rho)=\rho B(k,\rho+1)=\rho\frac{\Gamma(k)\Gamma(\rho+1)}{\Gamma(k+\rho+1)},k\in\{1,2,3,\cdots\},\rho>0. \tag{1}$$

Let $x_1,x_2,\cdots,x_n$ be a random sample from a Yule-Simon distribution, and m be the largest observation, $m=\max(x_1,x_2,\cdots,x_n)$. Let n_k be the number of the observations equal to k, $k\in\{1,2,\cdots,m\}$, $\sum_{k=1}^{m}n_k=n$. Chen and Chong[6] introduced the likelihood equation of parameter ρ, which is

$$\sum_{k=1}^{m}f(k,\rho)\frac{F_a(k,\rho)}{F(k,\rho)}=1, \tag{2}$$

where $F(k,\rho)=\sum_{s=k}^{\infty}f(s,\rho)=\rho B(k,\rho)$, $F_a(k,\rho)=\sum_{s=k}^{m}n_s/n$. If $m=1$, which means that $x_1=\cdots=x_n=1$, then the likelihood equation of the parameter ρ is simplified to $\rho/(\rho+1)=1$, which has no solution. Thus, in order to estimate ρ, it is required that m must be equal to or greater than 2.

Later, Garcia[7] obtained a likelihood equation similar to (2), and proposed a fixed-point iterative algorithm to obtain the maximum likelihood estimator of ρ. A Yule process with some known parameter was simulated to generate the random number and verify the accuracy of the algorithm. In this paper, the convergence of the iterative algorithm is proved theoretically. Based on the asymptotic normality of the maximum likelihood estimator, an asymptotic confidence interval of ρ is obtained in this paper.

The Yule-Simon distribution is a factorial distribution. Maybe it is the reason why the likelihood equation is not easy to solve. In this paper we will turn to the second place to construct a moment estimator. As we all know that the Yule-Simon distribution is a special kind of heavy-tailed distribution, which is called the long-tailed distribution. Distributions

such as these are not well expressed by their characteristic numbers, for example average. Only when parameter $\rho > 1$ and $\rho > 2$, the Yule-Simon distribution has finite mean and variance, respectively. Practical empirical data show that the parameter ρ rarely exceeds 2, although there are occasional exceptions. It is difficult to construct the moment estimator for the Yule-Simon distribution. In this paper, a reciprocal transformation is provided, and for the transformed distribution, the mean and variance are always finite. Based on transformed data, the moment estimator of ρ will be constructed. This estimator can be expressed in an explicit form. In addition, it is completely possible to construct an interval estimator of ρ based on the central limit theorem.

The structure of this paper is arranged as follows. The convergence of the iterative algorithm of the likelihood equation is proved in Section 2. Three methods to construct the interval estimations of the parameter ρ are described in Section 3. In Section 4 the feasibility of these methods is demonstrated through simulation, and their advantages and disadvantages are compared. In Section 5 the construction method of interval estimation given in this paper will be applied to the empirical data of Cerambycinv species and genera disposed by Yule[1]. In Section 6 we summarize the methods and conclusions of this paper, and propose some issues to be discussed in further research.

2 Likelihood equation and iterative algorithm

For the distribution law in (1) and its random sample $x_1, \cdots, x_n$, fixed-point iterative algorithm formula given by Garcia[7], i.e., the likelihood equation of parameters ρ can be simplified to

$$\rho = \frac{1}{\frac{1}{\rho+1} + \frac{\sum_{i=2}^{m} p_i}{\rho+2} + \frac{\sum_{i=3}^{m} p_i}{\rho+3} + \cdots + \frac{p_m}{\rho+m}}, \tag{3}$$

where $p_k = n_k/n$, $k \in \{1, \cdots, m\}$, $n_m \geqslant 1$, $\sum_{k=1}^{m} n_k = n$, $\sum_{k=1}^{m} p_k = 1$. To analyze the convergence of the iterative algorithm, the following function is investigated

$$g(x) = \frac{1}{\frac{1}{x+1} + \frac{\sum_{i=2}^{m} p_i}{x+2} + \frac{\sum_{i=3}^{m} p_i}{x+3} + \cdots + \frac{p_m}{x+m}}, \tag{4}$$

where $x > 0$. It's easy to calculate

$$\begin{aligned} g'(x) =& \frac{\frac{1}{(x+1)^2} + \frac{\sum_{i=2}^{m} p_i}{(x+2)^2} + \frac{\sum_{i=3}^{m} p_i}{(x+3)^2} + \cdots + \frac{p_m}{(x+m)^2}}{\left(\frac{1}{x+1} + \frac{\sum_{i=2}^{m} p_i}{x+2} + \frac{\sum_{i=3}^{m} p_i}{x+3} + \cdots + \frac{p_m}{x+m}\right)^2} \\ =& g^2(x) \left[\frac{1}{(x+1)^2} + \frac{\sum_{i=2}^{m} p_i}{(x+2)^2} + \frac{\sum_{i=3}^{m} p_i}{(x+3)^2} + \cdots + \frac{p_m}{(x+m)^2}\right] > 0, \end{aligned} \tag{5}$$

$$\begin{aligned} g''(x) =& 2g^3(x)\left\{\left[\frac{1}{(x+1)^2}+\frac{\sum_{i=2}^m p_i}{(x+2)^2}+\frac{\sum_{i=3}^m p_i}{(x+3)^2}+\cdots+\frac{p_m}{(x+m)^2}\right]^2\right. \\ & \left.-\left[\frac{1}{x+1}+\frac{\sum_{i=2}^m p_i}{x+2}+\frac{\sum_{i=3}^m p_i}{x+3}+\cdots+\frac{p_m}{x+m}\right]\sum_{s=1}^m\frac{\sum_{i=s}^m p_i}{(x+s)^3}\right\} \\ =& -2g^3(x)\left[\sum_{1\leqslant t<s\leqslant m}\frac{\sum_{i=t}^m p_i\sum_{i=s}^m p_i}{(x+t)(x+s)}\left(\frac{1}{x+t}-\frac{1}{x+s}\right)^2\right]<0. \end{aligned} \tag{6}$$

Thus, $g(x)$ is a monotone increasing and upper convex function. Obviously,

$$g(0)=\frac{1}{1+\frac{\sum_{i=2}^m p_i}{2}+\frac{\sum_{i=3}^m p_i}{3}+\cdots+\frac{p_m}{m}}>0.$$

Since $g(x)=P_1(x)/P_2(x)$, $P_1(x)$ and $P_2(x)$ are the m-degree and $(m-1)$-degree polynomials of x, respectively. $P_1(x)=nx^m+\cdots$, $P_2(x)=(1+a)x^m+\cdots$, where $a=\sum_{t=2}^m\sum_{i=t}^m p_i>0$, therefore

$$g(x)\sim c\cdot x,\quad 0<c=1/(1+a)<1,\ \text{as } x\to\infty.$$

Thus, when $x>0$ the curve $y=g(x)$ must intersect the line $y=x$, which means that there exists a solution to the likelihood equation (3).

Assuming that x_1 is a solution of likelihood equation (3) and there is no solution in the interval $(0,x_1)$. Provided that $x_2(>x_1)$ is a solution too and there is no solution in the interval (x_1,x_2). Let $h(x)=x-g(x)$. Since $h(x_1)=h(x_2)=0$, and $h(x)>0$ when $x\in(x_1,x_2)$, there exists $x_0\in(x_1,x_2)$ that $h(x)$ has a local maxima at point x_0 and $h''(x_0)=-g''(x_0)<0$, which obviously contradicts equation (6). This proves that the likelihood equation (3) has a unique solution.

Let $\hat{\rho}$ be the solution of the likelihood equation (3), and then $\hat{\rho}$ is the maximum likelihood estimator of ρ and $\hat{\rho}=g(\hat{\rho})$. Garcia[7] suggested the fixed-point iterative algorithm to estimate the parameter ρ and verified the accuracy of the estimation by simulation. Here, the convergence of the iterative algorithm will be proved theoretically.

Theorem 1 Assume that $x_1,\cdots,x_n$ is a random sample from a Yule-Simon distribution expressed in (1). $\hat{\rho}$ is the maximum likelihood estimator of ρ. $\rho_0>0$ is an arbitrary given initial value. $g(\cdot)$ is a function defined in (4). A sequence is defined as follows: $\rho_1=g(\rho_0)$, $\rho_i=g(\rho_{i-1})$, $i\in\{1,2,\cdots\}$, then $\lim\limits_{i\to\infty}\rho_i=\hat{\rho}$.

Proof It can be seen that when $x\in(0,\hat{\rho})$, the curve $y=g(x)$ is above the line $y=x$, so that $g(x)>x$, and when $x\in(\hat{\rho},\infty)$, the curve $y=g(x)$ is below the line $y=x$, so that $g(x)<x$. Thus, if $\rho_0<\hat{\rho}$, then $\rho_1=g(\rho_0)>\rho_0$, and $g(x)$ is a monotone increasing function and $\rho_0<\hat{\rho}$, so $\rho_1=g(\rho_0)<g(\hat{\rho})=\hat{\rho}$. Thus, when taking initial value $\rho_0<\hat{\rho}$, an increasing sequence $\rho_0<\rho_1<\rho_2<\cdots<\hat{\rho}$ will be obtained. Similarly, if $\rho_0>\hat{\rho}$, then a

decreasing sequence $\rho_0 > \rho_1 > \rho_2 > \cdots > \hat{\rho}$ will be obtained.

It is resulted from equations (5) and (6) that for $x > 0$, $g'(x)$ is a decreasing function with positive values. Therefore,

$$\text{for } \rho_0 < \hat{\rho},\ 0 < \hat{\rho} - \rho_i = g(\hat{\rho}) - g(\rho_{i-1}) \leqslant g'(\rho_0) \cdot (\hat{\rho} - \rho_{i-1}), i \in \{1, 2, \cdots\};$$

$$\text{for } \rho_0 > \hat{\rho},\ 0 < \rho_i - \hat{\rho} = g(\rho_{i-1}) - g(\hat{\rho}) \leqslant g'(\hat{\rho}) \cdot (\rho_{i-1} - \hat{\rho}), i \in \{1, 2, \cdots\}.$$

Since

$$\begin{aligned}\left(\frac{1}{x+1} + \frac{\sum_{i=2}^m p_i}{x+2} + \cdots + \frac{p_m}{x+m}\right)^2 &= \sum_{t=1}^m \frac{\sum_{i=t}^m p_i}{x+t}\left(\frac{1}{x+1} + \frac{\sum_{i=2}^m p_i}{x+2} + \cdots + \frac{p_m}{x+m}\right) \\ &> \sum_{t=1}^m \frac{\sum_{i=t}^m p_i}{x+t} \cdot \frac{1}{x+1} \geqslant \sum_{t=1}^m \frac{\sum_{i=t}^m p_i}{(x+t)^2},\end{aligned}$$

$g'(x) < 1$ when $x > 0$, according to equation (5). It can be seen that there is a positive number $\delta < 1$, $|\rho_n - \hat{\rho}| \leqslant \delta^n \cdot |\rho_0 - \hat{\rho}|$. Therefore, $\rho_n \to \hat{\rho}$ when $n \to \infty$. □

So far, the convergence of the fixed-point iterative algorithm has been proved.

3 Confidence interval estimator

3.1 Confidence interval based on First frequency mode estimator

To estimate ρ, in addition to maximum likelihood, the following two methods are usually used:

(i) First frequency mode estimation

For the Yule-Simon distribution, $f(1, \rho) > f(2, \rho) > f(3, \rho) > \cdots$, and $X = 1$ is its mode. The first frequency is important. Due to $p_1 = f(1, \rho) = \rho/(\rho + 1)$, then $\rho = \frac{p_1}{1-p_1}$. $\hat{p}_1 = n_1/n$ is the frequency corresponding to the probability p_1, so the parameter ρ can be estimated by

$$\hat{\rho}_{\text{mod}} = \frac{\hat{p}_1}{1 - \hat{p}_1}. \tag{7}$$

Obviously, this method divides the data into two parts: $X = 1$ and $X > 1$, which combines the information of $X = 2$, $X = 3$ and so on. This method only uses the information of a part of the samples. From this perspective, it is a method with obvious defects.

Doing so produces the Bernoulli variable. According to the central limit theorem, $\sqrt{n}(\hat{p}_1 - p_1) \xrightarrow{L} N(0, p_1(1 - p_1))$. Since ρ is a differentiable function of p_1 and $\partial\rho/\partial p_1 = \frac{1}{(1-p_1)^2}$, $\sqrt{n}(\hat{\rho}_{\text{mod}} - \rho) \xrightarrow{L} N(0, p_1/(1 - p_1)^3)$. This formula can be expressed equivalently as $\sqrt{n}(\hat{\rho}_{\text{mod}} - \rho) \xrightarrow{L} N(0, \rho(1 + \rho)^2)$. Thus, a $100(1 - \alpha)$ percent approximate confidence

interval of ρ is

$$\left(\hat{\rho}_{\mathrm{mod}} - U_{1-\alpha/2}\sqrt{\frac{\hat{\rho}_{\mathrm{mod}}(1+\hat{\rho}_{\mathrm{mod}})^2}{n}}, \hat{\rho}_{\mathrm{mod}} + U_{1-\alpha/2}\sqrt{\frac{\hat{\rho}_{\mathrm{mod}}(1+\hat{\rho}_{\mathrm{mod}})^2}{n}}\right). \tag{8}$$

(ii) Expectation Estimation

The mean of the Yule-Simon distribution is $\mu = \frac{\rho}{\rho-1}$ for $\rho > 1$. Thus, ρ can be estimated by

$$\hat{\rho}_{\mathrm{mea}} = \frac{\bar{x}}{\bar{x}-1} > 1 \tag{9}$$

where $\bar{x} = \sum_{k=1}^{m}(k \cdot n_k)/n$ is the sample mean. Obviously, this method is not credible unless you know in advance $\rho > 1$. Furthermore, in most practical applications, the parameter ρ is less than 2 and the variance does not exist, so it is impossible to construct a confidence interval based on the sample mean.

Although the role of the sample mean in the parameter estimation of Yule distribution is limited, it can be used to speculate the value of the parameter. If $\rho \in (1, 2]$, then $\mu = \frac{\rho}{\rho-1} \geqslant 2$; if $\rho \geqslant 2$, then $\mu = \frac{\rho}{\rho-1} \in (1, 2]$. A larger sample mean corresponds to a smaller parameter. If the sample mean value is less than 2, it can be inferred that the parameter value is likely to be greater than 2.

3.2 Confidence interval based on likelihood estimator

Under certain regular conditions, we can derive asymptotic normality of the maximum likelihood estimator $\hat{\rho}$,

$$\sqrt{n}(\hat{\rho} - \rho) \xrightarrow{L} N(0, I^{-1}(\rho)),$$

where $I(\rho)$ is the Fisher information:

$$I(\rho) = \frac{1}{\rho^2} - \frac{1}{\rho}\sum_{k=1}^{\infty}\frac{f(k,\rho)}{\rho+k}.$$

Selecting a large enough integer t, we can take $I_1(\rho) = \frac{1}{\rho^2} - \frac{1}{\rho}\sum_{k=1}^{t}\frac{f(k,\rho)}{\rho+k}$ as an approximation of the Fisher information $I(\rho)$, since

$$|I(\rho) - I_1(\rho)| = \frac{t}{\rho(\rho+t)}f(t,\rho)\sum_{k=1}^{\infty}\frac{1}{\rho+k}f(k,\rho+t) < \frac{t}{\rho(\rho+t)}f(t,\rho)\frac{1}{\rho+1} = O(t^{-(\rho+1)}).$$

The calculation process of the accuracy is complex and is put in Appendix I. Under the given accuracy, if t is large enough, $I_1(\rho)$ will meet the accuracy requirements. In addition, the larger the parameter ρ, the higher the accuracy. For example, if $\rho = 1$ and $t = 100$, the accuracy will be 0.0001; if $\rho = 2$ and $t = 100$, the accuracy will be 0.000001.

$\hat{\rho}$ is a consistent estimator for ρ, and then based on the asymptotic normality of the maximum likelihood estimator, a $100(1-\alpha)$ percent approximate confidence interval of ρ is

$$\left(\hat{\rho} - U_{1-\alpha/2} \cdot \frac{1}{\sqrt{nI_1(\hat{\rho})}}, \hat{\rho} + U_{1-\alpha/2} \cdot \frac{1}{\sqrt{nI_1(\hat{\rho})}}\right). \tag{10}$$

3.3 Reciprocal Transformation and Interval Estimation

Let X follow a Yule-Simon distribution. Considering that it is a factorial distribution, the reciprocal transformation will be used. Providing that the reciprocal transformation $Y = 1/(X-1)$ is taken, then X cannot be equal to 1. Therefore, the reciprocal transformation used in this paper is modified as

$$Y = \begin{cases} c, & X = 1 \\ 1/(X-1), & X = 2, 3, \cdots \end{cases} \tag{11}$$

where c is a constant, and its value remains to be determined. This reciprocal transformation compresses the heavy-tailed Yule-Simon distribution into the following distribution

$$\mathsf{P}(Y = y) = \begin{cases} \rho/(\rho+1), & y = c \\ \rho B(k, \rho+1), & y = 1/(k-1), k = 2, 3, \cdots \end{cases}$$

For $\forall\, m > 0$ and $\rho > 0$,

$$\left(\frac{1}{k-1}\right)^m \cdot \rho \frac{\Gamma(k)\Gamma(\rho+1)}{\Gamma(k+\rho+1)} \sim k^{-(m+\rho+1)}, \text{ as } k \to \infty,$$

then

$$\mathsf{E}(Y^m) = c^m \cdot \frac{\rho}{\rho+1} + \sum_{k=2}^{\infty} \left(\frac{1}{k-1}\right)^m \cdot \rho \frac{\Gamma(k)\Gamma(\rho+1)}{\Gamma(k+\rho+1)} < +\infty.$$

Therefore, the mean $\mu_{_Y} = \mathsf{E}(Y)$ and the variance $\sigma_{_Y}^2 = \mathsf{Var}\,(Y)$ are limited for $\rho > 0$, and

$$\begin{aligned} \mu_{_Y} &= c \cdot \frac{\rho}{\rho+1} + \sum_{k=2}^{\infty} \frac{1}{k-1} \cdot \rho \frac{\Gamma(k)\Gamma(\rho+1)}{\Gamma(k+\rho+1)} \\ &= \frac{c\rho}{\rho+1} + \frac{\rho}{(\rho+1)^2} \sum_{s=1}^{\infty} (\rho+1) \frac{\Gamma(s)\Gamma(\rho+2)}{\Gamma(s+\rho+2)} = \frac{c\rho}{\rho+1} + \frac{\rho}{(\rho+1)^2}, \end{aligned} \tag{12}$$

$$\sigma_{_Y}^2 = \frac{\rho}{(\rho+1)^2} \left(c - \frac{\rho}{\rho+1}\right)^2 - \frac{\rho^2}{(\rho+1)^3} + \frac{\rho}{\rho+1} \sum_{s=1}^{\infty} \frac{1}{(\rho+s+1)^2}. \tag{13}$$

The derivation process of $\sigma_{_Y}^2$ is complex, so we put it in Appendix II.

The reciprocal transformation, described in (11), transforms $x_1, x_2, \cdots, x_n$, the sam-

ple from the Yule-Simon distribution, into the sample $y_1, y_2, \cdots, y_n$ from Y. The moment estimator of μ is the sample mean

$$\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{cn_1}{n} + \sum_{k=2}^{m} \frac{n_k}{n} \cdot \frac{1}{k-1}.$$

Since ρ is a function of $\mu_{_Y}$ and

$$\rho(\mu_{_Y}) = \frac{(2\mu_{_Y} - c - 1) + \sqrt{(c+1)^2 - 4\mu_{_Y}}}{2(c - \mu_{_Y})},$$

the moment estimator of ρ is

$$\hat{\rho}_{\mathrm{re}} = \frac{(2\bar{y}_n - c - 1) + \sqrt{(c+1)^2 - 4\bar{y}_n}}{2(c - \bar{y}_n)}.$$

According to the central limit theorem,

$$\sqrt{n}(\bar{y}_n - \mu_{_Y}) \xrightarrow{L} N\left(0, \sigma_{_Y}^2\right).$$

For the the moment estimation of ρ,

$$\sqrt{n}(\hat{\rho}_{\mathrm{re}} - \rho) \xrightarrow{L} N\left(0, \left[\frac{\partial \rho(\mu_{_Y})}{\partial \mu_{_Y}}\right]^2 \sigma_{_Y}^2\right).$$

Here, the asymptotic variance of $\hat{\rho}_{\mathrm{re}}$ is a function of c and ρ. When

$$c = \frac{\rho}{\rho+1} - \rho + (\rho+1)^2 \sum_{s=1}^{\infty} \frac{1}{(\rho+s+1)^2},$$

the asymptotic variance of $\hat{\rho}_{\mathrm{re}}$ takes the minimum value. The detailed calculation process is given in Appendix III. $c = 1$ is a suitable choice for any $\rho > 0$.

For simplicity, take $c = 1$,

$$\bar{y}_n = \frac{n_1}{n} + \sum_{k=2}^{m} \frac{n_k}{n} \cdot \frac{1}{k-1} < \frac{n_1}{n} + \sum_{k=2}^{m} \frac{n_k}{n} = 1. \tag{14}$$

According to (12) and (13),

$$\mu = \mathsf{E}(Y) = \frac{\rho^2 + 2\rho}{(\rho+1)^2},$$

$$\sigma_{_Y}^2 = \frac{\rho}{(\rho+1)^4} - \frac{\rho^2}{(\rho+1)^3} + \frac{\rho}{\rho+1} \sum_{s=1}^{\infty} \frac{1}{(\rho+s+1)^2}.$$

Considering $\bar{y}_n < 1$, the moment estimator of ρ is

$$\hat{\rho}_{\mathrm{re}} = \frac{1}{\sqrt{1-\bar{y}_n}} - 1, \tag{15}$$

where $\bar{y}_n$ is defined in (14). The asymptotic variance of $\hat{\rho}_{\mathrm{re}}$ is

$$\sigma^2_{\mathrm{re}}(\rho) = \frac{\rho(\rho+1)^5}{4n}\left[\frac{1}{(\rho+1)^3} - \frac{\rho}{(\rho+1)^2} + \sum_{s=1}^{\infty}\frac{1}{(\rho+s+1)^2}\right].$$

Therefore,

$$(\hat{\rho}_{\mathrm{re}} - U_{1-\alpha/2}\sigma_{\mathrm{re}}(\hat{\rho}_{\mathrm{re}}), \hat{\rho}_{\mathrm{re}} + U_{1-\alpha/2}\sigma_{\mathrm{re}}(\hat{\rho}_{\mathrm{re}})) \tag{16}$$

is a $100(1-\alpha)$ percent approximate confidence interval of ρ.

4 Simulation

On the basis of the likelihood method, mode method and reciprocal transformation method, three approximate confidence intervals are given above. Usually, different methods have their own advantages and disadvantages. The coverage and interval length of these interval estimates will be compared under different values of ρ.

Yule-Simion distribution is a heavy-tailed distribution. Its random variable will get a very large value with a nonnegligible probability. The traditional random number generation method of discrete random variables is not suitable. Garcia[7] proposed a modified Polya urn process to get the random numbers from a Yule-Simon distribution. Random numbers are generated through an experiment test with some bins each containing one ball. Additional balls arrive one at a time. With probability α, a new ball is placed in a new bin; with probability $1-\alpha$, the new ball is placed in an existing bin and the probability of the ball being placed in a particular bin is proportional to the number of balls in that bin. Here, $\alpha = 1 - 1/\rho$.

This random number generation method has some limitations. First, it is only suitable for occasions with parameter ρ bigger than 1. As the examples listed in [7], there are a large number of occasions where the parameters are smaller than 1. Second, if the total number of pitches is fixed, with the increase of parameters, the number of samples that can be obtained from the test will decrease. For example, if 5000 balls are thrown, when $\rho = 1.1$, $\alpha = 0.091$, about 450 samples can be obtained; when $\rho = 2$, $\alpha = 0.5$, about 84 samples can be obtained. Therefore, in order to obtain the random samples with a given size, as the parameters become larger, the number of pitches must increase significantly. This will result in a longer time to generate random numbers. If 10000 data sets need to be generated, the increase of parameter ρ will significantly increase the running time of the program. Third, the random number obtained by this method is essentially the random number from a truncated Yule-Simon distribution, and the maximum value of the sample obtained is far less than the number of balls in the test.

Considering the accuracy of statistical analysis and the running time of the program, the random number from a truncated Yule-Simon distribution is used for analysis in this paper. In order to make the random samples and the population distribution as close as possible, the cutoff point is set to 10000. Set the sample size to 1000 and the simulation times to 10000, and carry out simulation calculation for different ρ. The comparison results of the three methods are shown in Table 1.

Table 1 Comparison of estimations for different ρ

ρ	method	estimator	bias	sd	MSE	width	coverage
0.6	MLE	0.6161	0.0161	0.0213	0.0007	0.0474	0.9415
0.6	Mode	0.6048	0.0048	0.0398	0.0016	0.0774	0.9501
0.6	Reciprocal	0.6053	0.0053	0.0277	0.0008	0.0542	0.9481
0.8	MLE	0.8056	0.0056	0.0301	0.0009	0.0624	0.9601
0.8	Mode	0.8012	0.0012	0.0508	0.0026	0.1000	0.9531
0.8	Reciprocal	0.8015	0.0015	0.0367	0.0014	0.0717	0.9489
1	MLE	1.0028	0.0028	0.0393	0.0016	0.0774	0.9507
1	Mode	1.0009	0.0009	0.0626	0.0039	0.1241	0.9526
1	Reciprocal	1.0010	0.0010	0.0463	0.0021	0.0912	0.9517
2	MLE	2.0055	0.0055	0.0977	0.0096	0.1475	0.8686
2	Mode	2.0061	0.0061	0.1349	0.0182	0.2641	0.9517
2	Reciprocal	2.0057	0.0057	0.1103	0.0122	0.2154	0.9499
3	MLE	3.0129	0.0129	0.1715	0.0296	0.2130	0.7896
3	Mode	3.0136	0.0136	0.2192	0.0482	0.4324	0.9576
3	Reciprocal	3.0143	0.0143	0.1952	0.0383	0.3847	0.9515

It can be seen that all three methods get good estimators with small deviations.

The estimator obtained by the MLE method has the smallest standard deviation (sd), mean square error (MSE) and half interval length (width). However, with the increase of the parameter ρ, its coverage has changed from the best to the worst. The low coverage may be caused by the high accuracy of the confidence interval, since for a given sample size, the two quantities will restrict each other.

In comparison, the mode method performs the worst in standard deviation, mean square error and half interval length. This can be explained as that the mode method only uses the information of a part of samples, and the loss of information leads to the increase of variance and the width of confidence interval.

Among the three methods, all the results of the reciprocal transformation method are very good and stable.

5 Real data analysis

The data on longicorn disposed by Yule[1] contains $n = 1024$ genera, see Table 2, where n_k is the number of genera which have k species.

Table 2 Longicorn

k	n_k	k	n_k	k	n_k	k	n_k	k	n_k	k	n_k
1	469	11	11	21	2	32	1	46	1	69	1
2	152	12	4	22	5	34	3	47	1	89	1
3	82	13	10	23	1	35	2	49	1	95	1
4	61	14	9	24	3	36	1	50	1	104	1
5	33	15	8	25	3	37	1	52	1	107	1
6	36	16	7	26	3	39	2	53	1	120	1
7	18	17	11	27	1	40	2	57	1	125	1
8	17	18	6	28	1	42	1	59	1	total	1024
9	14	19	51	30	2	43	2	66	1		
10	11	20	3	31	1	44	1	67	1		

Using mode estimation, by (7) and (8), the estimator of ρ is $\hat{\rho}_{\text{mod}} = 0.8450$ and the 95% confidence interval estimator for ρ is $(0.7412, 0.9488)$. For longicorn data, the mode method is feasible. But if using mean method in (9), the estimator of ρ is $\hat{\rho}_{\text{mea}} = 1.2181 > 1$, so this method is not credible for longicorn data.

According to (14) and (15), the calculation result of reciprocal transformation method can be obtained. The sample mean $\bar{y}_n = 0.6967$, and by equation (15), the moment estimator of ρ is $\hat{\rho}_{\text{re}} = 0.8157$. The estimation of asymptotic variance of $\hat{\rho}_{\text{re}}$ is 0.00136, and by equation (16), the 95% confidence interval estimator for ρ is $(0.7435, 0.8879)$.

Taking any positive value as the initial value of ρ and using the fixed-point algorithm of Garcia[7], we can quickly get the maximum likelihood estimate of $\hat{\rho}_L = 0.8995$. After calculation, the approximate value of Fisher information $I_1(\rho) = 0.8081$. According to (10), the 95% confidence interval estimator for is $(0.8307, 0.9683)$.

6 Conclusions

For the Yule-Simon distribution, more attention is focused on the application fields. Its heavy tail causes many problems worthy of theoretical study. In most cases, the variance of the Yule-Simon distribution does not exist, which increases the difficulty of research on the variance of the estimator and interval estimation of the parameter. In this paper, the asymptotic variances of three estimators and the asymptotic confidence intervals of the parameter are studied.

There are two issues related to interval estimation. One is the hypothesis test of the parameter, and the other is the distribution test. These two issues deserve further study.

Appendix

I. The Fisher information $I(\rho)$

$$\begin{aligned} I(\rho) &= -\mathsf{E}\left\{\frac{\partial^2\left[\ln(\rho)-\sum_{s=1}^{k}\ln(\rho+s)\right]}{\partial\rho^2}\right\} = \frac{1}{\rho^2}-\sum_{k=1}^{\infty}\sum_{s=1}^{k}\frac{1}{(\rho+s)^2}f(k,\rho)\\ &= \frac{1}{\rho^2}-\sum_{k=1}^{\infty}\frac{1}{(\rho+k)^2}\sum_{s=k}^{\infty}f(s,\rho) = \frac{1}{\rho^2}-\sum_{k=1}^{\infty}\frac{1}{(\rho+k)^2}[\rho B(k,\rho)]\\ &= \frac{1}{\rho^2}-\frac{1}{\rho}\sum_{k=1}^{\infty}\frac{1}{\rho+k}f(k,\rho). \end{aligned}$$

In order to calculate the series, we express the fraction $\frac{1}{\rho+k}$ in a different way:

$$\begin{aligned} \frac{1}{\rho+k} &= \frac{1}{\rho+k+1}+\frac{1}{(\rho+k)(\rho+k+1)}\\ &= \frac{1}{\rho+k+1}+\frac{1}{(\rho+k+1)(\rho+k+2)}+\frac{2}{(\rho+k)(\rho+k+1)(\rho+k+2)}\\ &= \frac{1}{\rho+k+1}+\frac{1}{(\rho+k+1)(\rho+k+2)}+\frac{2}{(\rho+k+1)(\rho+k+2)(\rho+k+3)}\\ &\quad+\frac{6}{(\rho+k)(\rho+k+1)(\rho+k+2)(\rho+k+3)}\\ &= \cdots = \frac{1}{\rho+k+1}+\cdots+\frac{(t-1)!}{(\rho+k+1)(\rho+k+2)\cdots(\rho+k+t)}\\ &\quad+\frac{t!}{(\rho+k)(\rho+k+1)\cdots(\rho+k+t)}\\ &= \sum_{u=1}^{t}\frac{\Gamma(u)\Gamma(\rho+k+1)}{\Gamma(\rho+k+u+1)}+\frac{\Gamma(t+1)\Gamma(\rho+k)}{\Gamma(\rho+k+t+1)}. \end{aligned}$$

This equation holds for $t\in\{1,2,\cdots\}$. So, given an arbitrary $t\in\{1,2,\cdots\}$,

$$\begin{aligned} I(\rho) &= \frac{1}{\rho^2}-\frac{1}{\rho}\sum_{k=1}^{\infty}\left[\sum_{u=1}^{t}\frac{\Gamma(u)\Gamma(\rho+k+1)}{\Gamma(\rho+k+u+1)}+\frac{\Gamma(t+1)\Gamma(\rho+k)}{\Gamma(\rho+k+t+1)}\right]\cdot\rho\cdot\frac{\Gamma(k)\Gamma(\rho+1)}{\Gamma(k+\rho+1)}\\ &= \frac{1}{\rho^2}-\sum_{k=1}^{\infty}\left[\sum_{u=1}^{t}\frac{\Gamma(u)\Gamma(k)\Gamma(\rho+1)}{\Gamma(k+\rho+u+1)}+\frac{\Gamma(t+1)\Gamma(k)\Gamma(\rho+1)}{(k+\rho)\Gamma(k+\rho+t+1)}\right]\\ &= \frac{1}{\rho^2}-\sum_{u=1}^{t}\frac{\Gamma(u)\Gamma(\rho+1)}{\Gamma(\rho+u+1)}\sum_{k=1}^{\infty}\frac{\Gamma(k)\Gamma(\rho+u+1)}{\Gamma(k+\rho+u+1)}\\ &\quad-\frac{\Gamma(t+1)\Gamma(\rho+1)}{\Gamma(\rho+t+1)}\sum_{k=1}^{\infty}\frac{\Gamma(k)\Gamma(\rho+t+1)}{(\rho+k)\Gamma(k+\rho+t+1)} \end{aligned}$$

$$= \frac{1}{\rho^2} - \sum_{u=1}^{t} \frac{\Gamma(u)\Gamma(\rho+1)}{\Gamma(\rho+u+1)} \cdot \frac{1}{\rho+u} - \frac{\Gamma(t+1)\Gamma(\rho+1)}{(\rho+t)\Gamma(\rho+t+1)} \sum_{k=1}^{\infty} \frac{1}{\rho+k} f(k, \rho+t)$$
$$= \frac{1}{\rho^2} - \frac{1}{\rho} \sum_{u=1}^{t} \frac{f(u,\rho)}{\rho+u} - \frac{t}{\rho(\rho+t)} f(t,\rho) \sum_{k=1}^{\infty} \frac{1}{\rho+k} f(k, \rho+t).$$

In real calculation, take $I_1(\rho) = \frac{1}{\rho^2} - \frac{1}{\rho} \sum_{u=1}^{t} \frac{f(u,\rho)}{\rho+u}$ as an approximation of the Fisher information $I(\rho)$.

$$|I(\rho) - I_1(\rho)| = \frac{t}{\rho(\rho+t)} f(t,\rho) \sum_{k=1}^{\infty} \frac{1}{\rho+k} f(k, \rho+t) < \frac{t}{\rho(\rho+t)} f(t,\rho) \frac{1}{\rho+1} = O(t^{-(\rho+1)}).$$

II. The variance of Y

$$\mathsf{E}(Y^2) = c^2 \cdot \frac{\rho}{\rho+1} + \sum_{k=2}^{\infty} \frac{1}{(k-1)^2} \cdot \rho \frac{\Gamma(\rho+1)\Gamma(k)}{\Gamma(\rho+k+1)} = c^2 \cdot \frac{\rho}{\rho+1} + \sum_{k=2}^{\infty} \frac{1}{k-1} \cdot \rho \frac{\Gamma(\rho+1)\Gamma(k-1)}{\Gamma(\rho+k+1)}.$$

In order to calculate the series, we express the fraction $\frac{1}{k-1}$ in a different way.

$$\begin{aligned}
\frac{1}{k-1} &= \frac{1}{k+\rho+1} + \frac{\rho+2}{(k-1)(k+\rho+1)} \\
&= \frac{1}{k+\rho+1} + \frac{\rho+2}{k+\rho+1} \cdot \left[\frac{1}{k+\rho+2} + \frac{\rho+3}{(k-1)(k+\rho+2)} \right] \\
&= \frac{1}{k+\rho+1} + \frac{\rho+2}{(k+\rho+1)(k+\rho+2)} + \frac{(\rho+2)(\rho+3)}{(k+\rho+1)(k+\rho+2)(k+\rho+3)} \\
&\quad + \frac{(\rho+2)(\rho+3)(\rho+4)}{(k-1)(k+\rho+1)(k+\rho+2)(k+\rho+3)} \\
&= \cdots = \frac{1}{k+\rho+1} + \sum_{s=2}^{t} \frac{\prod_{i=2}^{s}(\rho+i)}{\prod_{i=1}^{s}(k+\rho+i)} + \frac{\prod_{i=2}^{t+1}(\rho+i)}{(k-1)\prod_{i=1}^{t}(k+\rho+i)},
\end{aligned}$$

This equation holds for $t \in \{2, 3, \cdots\}$. So, given an arbitrary $t \in \{2, 3, \cdots\}$,

$$\begin{aligned}
\mathsf{E}(Y^2) &= \frac{c^2\rho}{\rho+1} + \sum_{k=2}^{\infty} \left[\frac{1}{k+\rho+1} + \sum_{s=2}^{t} \frac{\prod_{i=2}^{s}(\rho+i)}{\prod_{i=1}^{s}(k+\rho+i)} + \frac{\prod_{i=2}^{t+1}(\rho+i)}{(k-1)\prod_{i=1}^{t}(k+\rho+i)} \right] \\
&\quad \times \rho \frac{\Gamma(\rho+1)\Gamma(k-1)}{\Gamma(\rho+k+1)} \\
&= \frac{c^2\rho}{\rho+1} + \sum_{k=2}^{\infty} \rho \frac{\Gamma(\rho+1)\Gamma(k-1)}{\Gamma(\rho+k+2)} + \sum_{k=2}^{\infty} \sum_{s=2}^{t} \frac{\rho\Gamma(\rho+1)\Gamma(k-1)\prod_{i=2}^{s}(\rho+i)}{\Gamma(\rho+k+s+1)} \\
&\quad + \sum_{k=2}^{\infty} \frac{\rho\Gamma(\rho+1)\Gamma(k-1)\prod_{i=2}^{t+1}(\rho+i)}{(k-1)\Gamma(\rho+k+t+1)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{c^2\rho}{\rho+1} + \sum_{k=1}^{\infty} \rho \frac{\Gamma(\rho+1)\Gamma(k)}{\Gamma(\rho+k+3)} + \sum_{s=2}^{t}\sum_{k=2}^{\infty} \frac{\rho\Gamma(\rho+s+1)\Gamma(k-1)}{(\rho+1)\Gamma(\rho+k+s+1)} \\
&\quad + \sum_{k=2}^{\infty} \frac{\rho\Gamma(\rho+t+2)\Gamma(k-1)}{(\rho+1)(k-1)\Gamma(\rho+k+t+1)} \\
&= \frac{c^2\rho}{\rho+1} + \frac{\rho}{\rho+1}\sum_{s=1}^{t}\frac{1}{(\rho+s+1)^2} + \frac{\rho}{\rho+1}\cdot\frac{1}{(\rho+t+2)^2}\sum_{k=1}^{\infty}\frac{1}{k}f(k,\rho+t+1) \\
&= \frac{c^2\rho}{\rho+1} + \frac{\rho}{\rho+1}\sum_{s=1}^{\infty}\frac{1}{(\rho+s+1)^2}.
\end{aligned}
$$

Then

$$
\begin{aligned}
\sigma_Y^2 &= \frac{c^2\rho}{\rho+1} + \frac{\rho}{\rho+1}\sum_{s=1}^{\infty}\frac{1}{(\rho+s+1)^2} - \left[\frac{c\rho}{\rho+1} + \frac{\rho}{(\rho+1)^2}\right]^2 \\
&= \frac{\rho}{(\rho+1)^2}\left(c - \frac{\rho}{\rho+1}\right)^2 - \frac{\rho^2}{(\rho+1)^3} + \frac{\rho}{\rho+1}\sum_{s=1}^{\infty}\frac{1}{(\rho+s+1)^2}.
\end{aligned}
$$

III. The asymptotic variance of $\hat{\rho}_{\mathrm{mo}}$

For simplicity, denote $c_1 = c - \frac{\rho}{\rho+1}$, $c_2(\rho) = -\frac{\rho^2}{(\rho+1)^3} + \frac{\rho}{\rho+1}\sum_{s=1}^{\infty}\frac{1}{(\rho+s+1)^2}$, and then $\sigma_Y^2 = \frac{\rho}{(\rho+1)^2}c_1^2 + c_2(\rho)$,

$$
\frac{\partial\rho(\mu_Y)}{\partial\mu_Y} = \frac{1}{\partial\mu_Y/\partial\rho} = \frac{(\rho+1)^3}{(\rho+1)\left(c - \frac{\rho}{\rho+1}\right)+1} = \frac{(\rho+1)^3}{(\rho+1)c_1+1}.
$$

The asymptotic variance of $\hat{\rho}_{\mathrm{mo}}$ can be expressed as the following:

$$
h(c_1) \hat{=} \left[\frac{\partial\rho(\mu_Y)}{\partial\mu_Y}\right]^2 \cdot \frac{\sigma_Y^2}{n} = \frac{1}{n}\frac{(\rho+1)^6}{\left[(\rho+1)c_1+1\right]^2}\left[\frac{\rho}{(\rho+1)^2}c_1^2 + c_2(\rho)\right].
$$

Next, find a suitable c to minimize the asymptotic variance.

$$
\begin{aligned}
h'(c_1) &= \frac{(\rho+1)^6}{n}\frac{\frac{2\rho c_1}{(\rho+1)^2}\left[(\rho+1)c_1+1\right]^2 - \left[\frac{\rho c_1^2}{(\rho+1)^2} + c_2(\rho)\right]\times 2\left[(\rho+1)c_1+1\right](\rho+1)}{\left[(\rho+1)c_1+1\right]^4} \\
&= \frac{(\rho+1)^6}{n}\frac{2\rho c_1\left[c_1+\frac{1}{\rho+1}\right]^2 - \left[\rho c_1^2 + c_2(\rho)(\rho+1)^2\right]\times 2\left[c_1+\frac{1}{\rho+1}\right]}{\left[(\rho+1)c_1+1\right]^4} \\
&= \frac{(\rho+1)^6}{n}\frac{2\left[c_1+\frac{1}{\rho+1}\right]}{\left[(\rho+1)c_1+1\right]^4}\left[\frac{\rho}{\rho+1}c_1 - c_2(\rho)(\rho+1)^2\right] \\
&= \frac{\rho(\rho+1)}{n}\frac{2}{\left[c_1+\frac{1}{\rho+1}\right]^3}\left[c_1 - c_2(\rho)(\rho+1)^3/\rho\right].
\end{aligned}
$$

When $c_1 = c_2(\rho)(\rho+1)^3/\rho$, $h(c_1)$ takes the minimum value. In other words, when

$$c = \frac{\rho}{\rho+1} - \rho + (\rho+1)^2 \sum_{s=1}^{\infty} \frac{1}{(\rho+s+1)^2},$$

the asymptotic variance of $\hat{\rho}_{\mathrm{mo}}$ takes the minimum value. By simple derivation, it can be concluded that

$$1 - \frac{1}{(\rho+1)(\rho+2)} < c < 2 - \frac{1}{(\rho+1)}.$$

Therefore, $c = 1$ is a suitable choice for any $\rho > 0$.

References

[1] YULE G U. II.—A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S [J]. *Philos T Roy Soc B*, 1925, **213**(**402-410**): 21–87.

[2] SIMON H A. On a class of skew distribution functions [J]. *Biometrika*, 1955, **42**(**3/4**): 425–440.

[3] NEWMAN M E J. Power laws, Pareto distributions and Zipf's law [J]. *Contemp Phys*, 2005, **46**(**5**): 323–351.

[4] BARABÁSI A L, ALBERT R. Emergence of scaling in random networks [J]. *Science*, 1999, **286**(**5439**): 509–512.

[5] BORNHOLDT S, EBEL H. World Wide Web scaling exponent from Simon's 1955 model [J]. *Phys Rev E*, 2001, **64**(**3**): 035104.

[6] CHEN Y S, CHONG P. Mathematical modeling of empirical laws in computer applications: A case study [J]. *Comput Math Appl*, 1992, **24**(**7**): 77–87.

[7] GARCIA J M G. A fixed-point algorithm to estimate the Yule-Simon distribution parameter [J]. *Appl Math Comput*, 2011, **217**(**21**): 8560–8566.

尤尔分布中参数的区间估计

邓文丽 [1] 王黎明 [2] 王静龙 [3]

[1] 江西师范大学数学与统计学院, 南昌, 330027

[2] 上海财经大学统计与管理学院, 上海, 200433

[3] 华东师范大学统计学院, 上海, 200062

摘 要: 尤尔分布在网络科学、生物学和人文科学中有着广泛的应用. 相关的研究工作主要集中在经验数据与尤尔分布的拟合程度分析或参数估计问题, 所以仍存在一些尚未解决的问题, 比如参数估计的误差分析, 参数极大似然估计的迭代算法收敛性的理论证明等. 尤尔分布是一个重尾分布, 在很多应用场合, 该分布的参数小于 2 从而导致方差不存在, 这使得参数的区间估计的构建存在一些困难. 利用压缩变换, 本文给出了一种基于中心极限定理的区间估计方法。该方法适用于许多重尾分布的区间估计. 另外, 本文还基于最大似然法和众数方法, 分别得到了参数的渐近置信区间. 文中通过模拟计算和实际数据分析, 对这三种区间估计方法进行了比较.

关键词: 尤尔分布; 极大似然估计; 置信区间; 压缩变换

中图分类号: O212.1