Numerical Method for Obtaining a Predictive Estimator for the Geometric Distribution

Kunio Takezawa

Division of Informatics and Inventory, Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization

Kannondai 3-1-3, Tsukuba, Ibaraki 305-8604, Japan

Original Research Article

Abstract

An optimal estimator in the light of future data (i.e., a predictive estimator) is obtained using numerical simulations. The predictive estimator is assumed to be one of various functions of the maximum likelihood estimator. We then formulate an estimator that yields better results than the maximum likelihood estimator when the parameters are located within a specific range. Using this method, we derive a predictive estimator for the geometric distribution. This procedure leads to a predictive estimator that outperforms the maximum likelihood estimator in terms of the expected log-likelihood when the parameter is known to be located within a certain range.

Keywords: expected log-likelihood; future data; geometric distribution; maximum likelihood estimator; predictive estimator

2010 Mathematics Subject Classification: 60G25; 62F10; 62M20

1 Introduction

The main purpose of statistical inference lies in finding estimates that predict future data within the desired degree of accuracy. The maximum likelihood estimator, however, performs estimations by fitting available data (i.e., data from the past). As this method can suffer from the "overfitting" phenomenon, the resultant estimates do not always predict future data sufficiently well. To cope with this problem, we require an estimator that gives a good fit to future data. We call such an estimator a "predictive estimator" (e.g., page 35 in [1]). The exact solution of the predictive estimator for the variance of the normal distribution is already known. This is called the "third variance" ([2], section 5.5 in [3]). The exact solution of the predictive estimator for the parameter of

the exponential distribution has also been derived [4], although another predictive estimator for the exponential distribution may be available. In addition, the characteristics of predictive estimators for the binomial distribution and the geometric distribution have been examined [5, 6]. We expect that further research on the analytical characteristics of predictive estimators will be conducted from various perspectives.

However, if the analytical features of a predictive estimator are largely unexplored, an approximate predictive estimator given by numerical simulations can be a practical alternative to the maximum likelihood estimator. Such numerically derived predictive estimators are useful for analytical studies. As predictive estimators for estimating the variance of the normal distribution and the parameter of the exponential distribution are described as functions of the maximum likelihood estimators, we assume here that the predictive estimator is a function of the maximum likelihood estimator. We then consider diverse functions as predictive estimators to calculate the expected log-likelihood. The results of this calculation using R version 2.15.3 allow us to choose an estimator that realizes better predictions than the maximum likelihood estimator when the parameters are located within a specific range. Using this method, we construct the predictive estimator for the parameter of the geometric distribution. Moreover, simple numerical simulations using R language to compare the predictive estimator with the maximum likelihood estimator indicate that the predictive estimator obtained here is of great practical use.

2 Construction of a predictive estimator using a numerical method

We assume that the number of parameters is one (θ). The true value of θ is termed $\tilde{\theta}$. The probability density function is represented as $f(x|\tilde{\theta})$, and the random variable obeying this probability density function is called X. The realizations of X, that is, $\{x_1, x_2, \ldots, x_n\}$, are assumed to be available. In this setting, the likelihood function can be written as

$$L(\theta|\{x_1, x_2, \ldots, x_n\}) = \prod_{i=1}^{n} f(x_i|\theta). \tag{2.1}$$

The log-likelihood function is therefore described as

$$l(\theta|\{x_1, x_2, \ldots, x_n\}) = \log(L(\theta|\{x_1, x_2, \ldots, x_n\})) = \sum_{i=1}^{n} \log(f(x_i|\theta)). \tag{2.2}$$

The value of θ that maximizes $l(\theta|\{x_1, x_2, \ldots, x_n\})$ is represented as $\hat{\theta}$. This $\hat{\theta}$ is the maximum likelihood estimator.

The estimator $\hat{\theta}$ fits well with the data at hand ($\{x_1, x_2, \ldots, x_n\}$). However, an estimator that fits well with future data may be slightly different from $\hat{\theta}$. Hence, we assume that the maximum likelihood estimator in the light of future data (i.e., the predictive estimator, denoted as θ^+) is represented as

$$\theta^+ = g_k(\hat{\theta}) \qquad (k = 1, 2, \ldots, G). \tag{2.3}$$

Then, we consider G functions of $g_k(\cdot)$ to obtain a predictive estimator. When we assume that the future data are $\{x_1^*, x_2^*, \ldots, x_m^*\}$, the log-likelihood of θ^+ in the light of future data is written as

$$l(\theta^+|\{x_1^*, x_2^*, \ldots, x_m^*\} = \sum_{i=1}^{n} \log(f(x_i^*|\theta^+)). \tag{2.4}$$

Dividing both sides of the above equation by m and letting $m \to \infty$, we have

$$\lim_{m\to\infty} \frac{1}{m} l(\theta^+|\{x_1^*, x_2^*, \ldots, x_m^*\} = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{n} \log(f(x_i^*|\theta^+)) = E_x\big[f(x|\theta^+)\big]. \tag{2.5}$$

$E_x\big[f(x|\theta^+)\big]$ is the average of $f(x|\theta^+)$ when x is sampled an infinite number of times. That is, this function gives the expectation of $f(x|\theta^+)$. The expectation in the sense of Eq. (2.5) is used to derive AIC (Akaike Information Criterion). For example, $E_{G(\mathbf{x}_n)}$ on page 55 in [1] is the expectation in this sense.

When the maximum estimator is adopted, Eq. (2.5) is replaced by

$$\lim_{m\to\infty}\frac{1}{m}l(\hat{\theta}|\{x_1^*,x_2^*,\ldots,x_m^*\}=\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{n}\log(f(x_i^*|\hat{\theta}))=E_x\big[f(x|\hat{\theta})\big]. \tag{2.6}$$

Hence, when a specific value is set as $\tilde{\theta}$ and

$$E_x\big[f(x|\theta^+)\big]>E_x\big[f(x|\hat{\theta})\big], \tag{2.7}$$

we conclude that the predictive estimator (θ^+) is better than the maximum likelihood estimator ($\hat{\theta}$) in terms of prediction on the assumption that the parameter of the population is $\tilde{\theta}$.

However, in practice, the parameter of the population is unknown. Even so, we often have some knowledge, such as "$\tilde{\theta}$ is between θ_{min} and θ_{max}," from an experiment or survey. Therefore, when we know that $\tilde{\theta}$ is between θ_{min} and θ_{max}, θ^+ is a better estimator than $\hat{\theta}$ in terms of prediction if Eq. (2.7) holds on the assumption that $\tilde{\theta}$ is located between θ_{min} and θ_{max}.

Hence, various functions $g_k(\cdot)$ $(1\le k\le G)$ are used to calculate Eqs. (2.5) and (2.6) with $\tilde{\theta}$ between θ_{min} and θ_{max}. One or more $g_k(\cdot)$ that satisfy Eq. (2.7) are then selected. If no functions satisfy Eq. (2.7), other functions are used for $g_k(\cdot)$. If more than one $g_k(\cdot)$ are selected, the values of $E_y\big[f(y|\theta^+)\big]$ are calculated using equispaced values between θ_{min} and θ_{max} as $\tilde{\theta}$, and the resultant values are averaged. This calculation is carried out with the selected $g_k(\cdot)$. These averages are then compared with the average of $E_y\big[f(y|\hat{\theta})\big]$ with the same setting. The function $g_k(\cdot)$ that maximizes the difference between the average given by a predictive estimator and that given by the maximum likelihood estimator is regarded as the optimal $g_k(\cdot)$.

3 Construction of a predictive estimator for the geometric distribution

A predictive estimator for the geometric distribution is now constructed using the method described in the previous section. The probability density function of the geometric distribution is written as

$$f(\xi)=\tilde{p}(1-\tilde{p})^{\xi-1}, \qquad \xi=1,2,3,\ldots. \tag{3.1}$$

The expectation of Eq. (3.1) is represented as

$$\sum_{\xi=1}^{\infty}\xi f(\xi)=\frac{1}{\tilde{p}}. \tag{3.2}$$

The random variable obeying $f(\xi)$ is denoted by X. The realizations (data) of this random variable are named $\{x_i\}$ $(1\le i\le n)$. The log-likelihood ($l(p|\{x_i\})$) of this data is described as

$$l(p|\{x_i\})=n\log(p)+\log(1-p)\sum_{i=1}^{n}(x_i-1). \tag{3.3}$$

To derive the value of p that maximizes the above function, we differentiate Eq. (3.3) with respect to p and set it equal to 0. Then, we have

$$\frac{n}{p}-\frac{\sum_{i=1}^{n}(x_i-1)}{1-p}=0. \tag{3.4}$$

Hence, the following maximum likelihood estimator is obtained:

$$\hat{p} = \frac{n}{\sum_{i=1}^{n} x_i}. \tag{3.5}$$

This $\hat{p}$ is the maximum likelihood estimator in the light of the data at hand ($\{x_i\}$, $1 \leq i \leq n$).

Next, we assume that the predictive estimator p^+ is a function of the maximum likelihood estimator $\hat{p}$. Thus, Eq. (2.3) becomes

$$p^+ = g_k(\hat{p}) \qquad (k = 1, 2, \ldots, G). \tag{3.6}$$

When we represent future data as $\{x_1^*, x_2^*, \ldots, x_m^*\}$, the log-likelihood of p^+ in the light of these data is written in a similar form to Eq. (3.3):

$$\frac{l(p^+|\{x_i^*\})}{m} = \log(p^+) + \log(1-p^+)\Big(\sum_{i=1}^{m} \frac{x_i^*}{m} - 1\Big). \tag{3.7}$$

The p^+ that maximizes this value (denoted as p^*) is represented as

$$p^* = \frac{m}{\sum_{i=1}^{m} x_i^*}. \tag{3.8}$$

When we use infinite number of future data to examine the validity of an estimate, the evaluation is more reliable than that given by finite number of future data. Then, we assume that m is infinite. In this setting, when p^* is denoted as p_∞^*, Eq. (3.2) shows that the following holds:

$$p_\infty^* = \lim_{m\to\infty} \frac{m}{\sum_{i=1}^{m} x_i^*} = \tilde{p}. \tag{3.9}$$

That is, the maximum likelihood estimator using infinite future data is the true parameter of the population, $\tilde{p}$. Thus, the log-likelihood in the light of an infinite number of future data is written as

$$\lim_{m\to\infty} \frac{l^*(p^+)}{m} = \log(p^+) + \log(1-p^+)\Big(\frac{1}{\tilde{p}} - 1\Big). \tag{3.10}$$

Next, we consider the average of $\lim_{m\to\infty} \frac{l^*(p^+)}{m}$ given by sampling x an infinite number of times. Equivalently, we consider the expectation of $l^*(p^+)$. However, the sampling must be conducted again if all of the data $\{x_i\}$ are equal to 1. The expectation is written as

$$E_x\Big[\lim_{m\to\infty} \frac{l^*(p^+)}{m}\Big] = E_x\Big[\log(p^+) + \log(1-p^+)\Big(\frac{1}{\tilde{p}} - 1\Big)\Big]. \tag{3.11}$$

When X_i obeys the geometric distribution, $\sum_{i=1}^{n} X_i$ obeys the negative binomial distribution (e.g., 87 page in [7]). Hence, Eq. (3.11) is transformed into

$$E_x\Big[\lim_{m\to\infty} \frac{l^*(p^+)}{m}\Big] = \frac{\sum_{j=n+1}^{\infty} \Big(\log(p^+) + \Big(\frac{1}{\tilde{p}} - 1\Big)\log(1-p^+)\Big)\,{}_{j-1}C_{n-1}\tilde{p}^n(1-\tilde{p})^{j-n}}{\sum_{j=n+1}^{\infty} {}_{j-1}C_{n-1}\tilde{p}^n(1-\tilde{p})^{j-n}}. \tag{3.12}$$

${}_{j-1}C_{n-1}$ is defined as ${}_{j-1}C_{n-1} = \frac{(j-1)!}{(j-n)!(n-1)!}$ ("!" represents a factorial.). The summation in the numerator of the right-hand side of Eq. (3.12) is $\sum_{j=n+1}^{\infty}$ instead of $\sum_{j=n}^{\infty}$. Moreover, the right-hand

side is normalized with respect to $\sum_{j=n+1}^{\infty} {}_{j-1}C_{n-1}\tilde{p}^n(1-\tilde{p})^{j-n}$. These modifications account for the resampling when all $\{x_i\}$ are equal to 1. When p^+ in Eq. (3.12) is replaced by $\hat{p}$, we have

$$E_x\left[\lim_{m\to\infty}\frac{l^*(\hat{p})}{m}\right] = \frac{\sum_{j=n+1}^{\infty}\left(\log(\hat{p}) + \left(\frac{1}{\tilde{p}}-1\right)\log(1-\hat{p})\right){}_{j-1}C_{n-1}\tilde{p}^n(1-\tilde{p})^{j-n}}{\sum_{j=n+1}^{\infty} {}_{j-1}C_{n-1}\tilde{p}^n(1-\tilde{p})^{j-n}}. \tag{3.13}$$

We represent the difference between the value of Eq. (3.12) and that of Eq. (3.13) as Δ:

$$\Delta = E_x\left[\lim_{m\to\infty}\frac{l^*(p^+)}{m}\right] - E_x\left[\lim_{m\to\infty}\frac{l^*(\hat{p})}{m}\right]. \tag{3.14}$$

If a specific value is given as $\tilde{p}$ and Δ is positive, $p^+(= g_k(\hat{p}))$ is a predictive estimator that outperforms the maximum likelihood estimator. However, in practice, the exact value of $\tilde{p}$ is unknown. Instead, we often have information such as "$\tilde{p}$ is between p_{min} and p_{max}." Therefore, p^+ is a better estimator than $\hat{p}$ in terms of prediction if Δ is always positive (i.e., Eq. (2.7) holds) on the assumption that $\tilde{\theta}$ is between θ_{min} and θ_{max}.

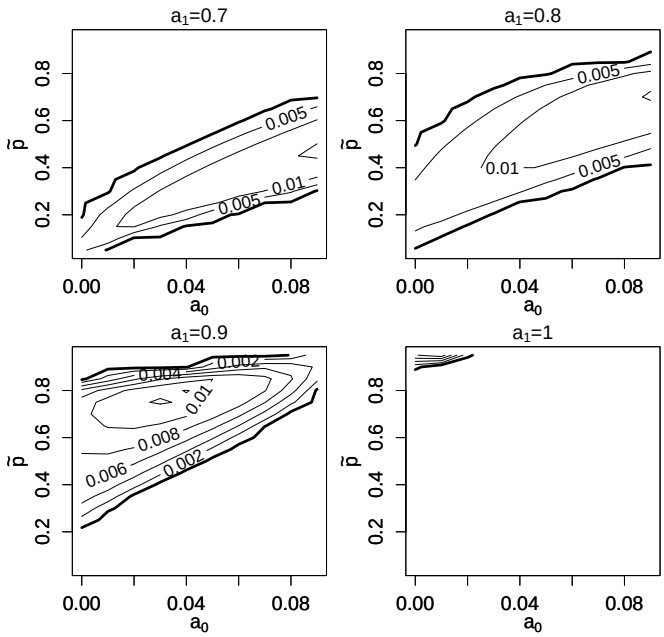

Figure 1: The value of Δ for various values of a_1 for each value of $\tilde{p}$. $p^+ = (a_0+a_1\hat{p})^{0.9}$ is adopted. $a_1 = 0.7$ (top-left), $a_1 = 0.8$ (top-right), $a_1 = 0.9$ (bottom-left), $a_1 = 1$ (bottom-right). The contour is drawn in the region of positive Δ. The contour is not drawn in the region of negative and incomputable Δ.

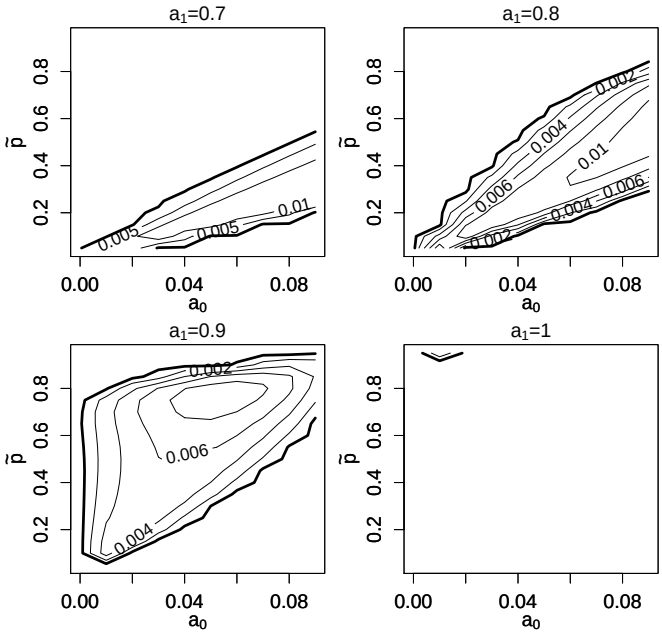

Figure 2: The value of Δ for various values of a_1 for each value of $\tilde{p}$. $p^+ = a_0 + a_1\hat{p}$ is adopted. $a_1 = 0.7$ (top-left), $a_1 = 0.8$ (top-right), $a_1 = 0.9$ (bottom-left), $a_1 = 1$ (bottom-right). The contour is drawn in the region of positive Δ. The contour is not drawn in the region of negative and incomputable Δ.

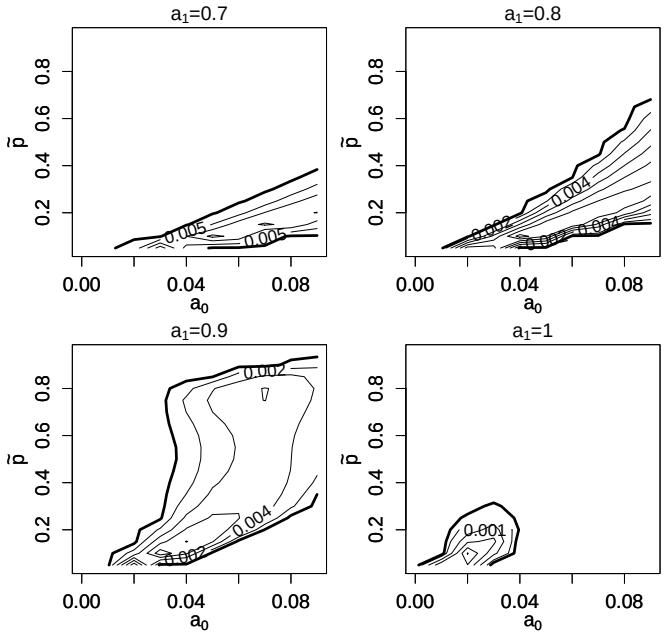

Figure 3: The value of Δ for various values of a_1 for each value of $\tilde{p}$. $p^+ = (a_0 + a_1\hat{p})^{1.1}$ is adopted. $a_1 = 0.7$ (top-left), $a_1 = 0.8$ (top-right), $a_1 = 0.9$ (bottom-left), $a_1 = 1$ (bottom-right). The contour is drawn in the region of positive Δ. The contour is not drawn in the region of negative and incomputable Δ.

Numerical simulations were conducted by setting the number of data n to 20 and using the three equations below as $g_k(\hat{p})$.

$$g_1(\hat{p}|a_0, a_1) = (a_0 + a_1\hat{p})^{0.9}, \tag{3.15}$$

$$g_2(\hat{p}|a_0, a_1) = a_0 + a_1\hat{p}, \tag{3.16}$$

$$g_3(\hat{p}|a_0, a_1) = (a_0 + a_1\hat{p})^{1.1}. \tag{3.17}$$

The forms of the equations above are not based on deep insight into predictive estimators. They are mere attempts to derive a predictive estimator by simple algorithm. However, if such rudimentary derivation of a predictive estimator is successful, it implies that numerical methods are promising tools to obtain more useful estimators than the maximum likelihood estimator.

We assumed that $a_0 = 0.0, 0.01, 0.02, 0.03, \ldots, 0.19$ and $a_1 = 0.7, 0.8, 0.9, 1.0$. That is, 228 $(3 \times 19 \times 4)$ functions were used to construct predictive estimators by transforming $\hat{p}$, and the values of Eqs. (3.12) and (3.13) were calculated to obtain Δ. Furthermore, because we set $\tilde{p} = 0.05, 0.1, 0.15, \ldots, 0.95$, Δ was computed for a total of $4,332$ $(3 \times 19 \times 4 \times 19)$ settings. When the value of Δ was negative, we set $\Delta = 0$. When Δ could not be computed because the argument of the logarithm was 0 or negative, we again set $\Delta = 0$. Figures 1, 2, and 3 show the results for $g_1(\hat{p}|a_0, a_1)$, $g_2(\hat{p}|a_0, a_1)$, and $g_3(\hat{p}|a_0, a_1)$, respectively. The four graphs in each figure correspond to the settings of $a_1 = 0.7, 0.8, 0.9, 1.0$. In each graph, the values of Δ are drawn for $a_0 = 0.0, 0.01, 0.02, 0.03, \ldots, 0.19$ and $\tilde{p} = 0.05, 0.1, 0.15, \ldots, 0.95$.

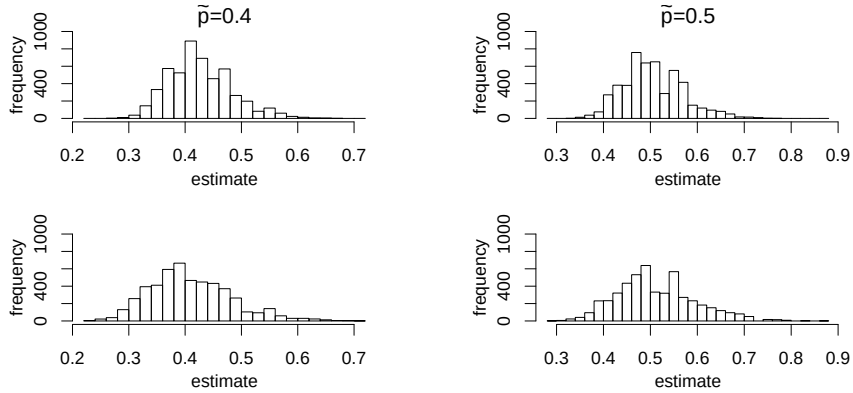

Figure 4: Estimation of p when we have information: "$\tilde{p}$ is between 0.4 and 0.7." (Top-left) The distribution of the predictive estimator and (bottom-left) maximum likelihood estimator when $\tilde{p} = 0.4$. (Top-right) The distribution of the predictive estimator and (bottom-right) maximum likelihood estimator when $\tilde{p} = 0.5$.

We assumed that we have information: "$\tilde{p}$ is between 0.4 and 0.7." We also assumed that $\tilde{p} = 0.4$ was the real parameter in the population. Then, using Figs. 1, 2, and 3, we selected the equations that ensured Δ (Eq. (3.15)) was positive when $\tilde{p}$ took values of $\{0.4, 0.45, 0.5, \ldots, 0.7\}$. Among the selected equations, we adopted the optimal equation that maximized the average value of Δ given by setting $\tilde{p}$ to $\{0.4, 0.45, 0.5, \ldots, 0.7\}$. This procedure found that the following equation was the optimal predictive estimator.

$$g_1(\hat{p}|a_0, a_1) = (0.06 + 0.8\hat{p})^{0.9}. \tag{3.18}$$

Next, $5,000$ datasets (each consisting of 20 data) were generated according to the geometric distribution with $\tilde{p} = 0.4$. Figure 4 (top-left and bottom-left) compares the distributions of the

estimates given by Eq. (3.5) and the estimates given by Eq. (3.18). The mean of the predictive estimator is 0.426764, and its unbiased variance is 3.237333×10^{-3}. In contrast, the mean of the maximum likelihood estimator is 0.4108221, and its unbiased variance is 5.200498×10^{-3}. That is, the maximum likelihood estimator is closer to an unbiased estimator. The variance of the predictive estimator, however, is smaller. Hence, to determine the preferred estimator, we must examine the goodness of fit to future data (i.e., expected log-likelihood).

Thus, we derived the values of the expected log-likelihood using the following two equations based on Eq. (3.7).

$$\frac{l(p_j^+|\{x_i^*\})}{m} = \log(p_j^+) + \log(1-p_j^+)\Big(\sum_{i=1}^{m}\frac{x_i^*}{m} - 1\Big). \tag{3.19}$$

$$\frac{l(\hat{p}_j|\{x_i^*\})}{m} = \log(\hat{p}_j) + \log(1-\hat{p}_j)\Big(\sum_{i=1}^{m}\frac{x_i^*}{m} - 1\Big), \tag{3.20}$$

where p_j^+ is the estimate obtained using the j-th simulation datum and the predictive estimator, and $\hat{p}_j$ is the estimate obtained using the j-th simulation datum and the maximum likelihood estimator.

We then calculated $\left\{\frac{l(p_j^+|\{x_i^*\})}{m}\right\}$ and $\left\{\frac{l(\hat{p}_j|\{x_i^*\})}{m}\right\}$ using the $5,000$ sets of simulation data with $m = 500$, and averaged the $5,000$ values of $\left\{\frac{l(p_j^+|\{x_i^*\})}{m}\right\}$ and the $5,000$ values of $\left\{\frac{l(\hat{p}_j|\{x_i^*\})}{m}\right\}$. Accordingly, the average of $\left\{\frac{l(p_j^+|\{x_i^*\})}{m}\right\}$ was found to be -1.702428, and that of $\left\{\frac{l(\hat{p}_j|\{x_i^*\})}{m}\right\}$ was found to be $= -1.709975$. Hence, the predictive estimator is better in terms of the expected log-likelihood.

However, it is difficult to gain an intuitive feel for how the discrepancy in the expected log-likelihood is associated with any practical advantages. Thus, we compared the maximum likelihood estimator and the predictive estimator by calculating the sum of the squared differences between the true value and the predictive estimator, because this squared difference is a more intuitive metric. The sum of squared differences between the predictive estimator and the true value is defined as

$$d(p^+) = \frac{1}{5,000}\sum_{j=1}^{5,000}(p_j^+ - \tilde{p})^2, \tag{3.21}$$

and that between the maximum likelihood estimator and the true value is defined as

$$d(\hat{p}) = \frac{1}{5,000}\sum_{j=1}^{5,000}(\hat{p}_j - \tilde{p})^2. \tag{3.22}$$

Using these expressions, we obtain values of $d(p^+) = 3.952998 \times 10^{-3}$ and $d(\hat{p}) = 5.316575 \times 10^{-3}$. Therefore, we can conclude that the difference between the predictive estimator and the maximum likelihood estimator is not negligible.

Next, $5,000$ datasets (each consisting of 20 data) were generated with $\tilde{p} = 0.5$. Using these data, a simulation was conducted under the same conditions as the previous simulation. The results are shown in Fig. 4 (top-right and bottom-right). The mean of the predictive estimator is 0.5055996, and its unbiased variance is 3.974303×10^{-3}. In contrast, the mean of the maximum likelihood estimator is 0.5114333, and its unbiased variance is 6.619982×10^{-3}. That is, the predictive estimator is not only closer to the unbiased estimator, but its variance is also smaller. Therefore, the superiority of the predictive estimator is intuitively clear. In fact, a comparison of the expected log-likelihood shows that the average of $\left\{\frac{l(p_j^+|\{x_i^*\})}{m}\right\}$ is -1.402638, whereas that of $\left\{\frac{l(\hat{p}_j|\{x_i^*\})}{m}\right\}$ is -1.414519. Calculating the sums of the squared errors between the true value and the two estimators, we obtain $d(p^+) = 4.004863 \times 10^{-3}$ and $d(\hat{p}) = 6.749378 \times 10^{-3}$. These results are as expected.

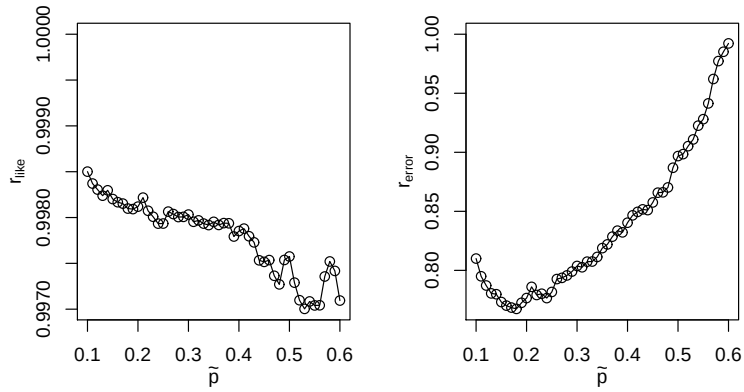

Figure 5: It is assumed that we have information: "$\tilde{p}$ is between 0.1 and 0.6." $r_{like}(p^+, \hat{p})$ (left) and $r_{error}(p^+, \hat{p})$ (right) are shown for $\tilde{p}$ equal to $\{0.1, 0.11, 0.12, \ldots, 0.6\}$.

Next, we assumed that we have information: "$\tilde{p}$ is between 0.1 and 0.6." We set $\tilde{p} = 0.1, 0.11, 0.12, \ldots, 0.6$, and examined whether the predictive estimator performed better than the maximum likelihood estimator in terms of the expected log-likelihood. We obtained the following predictive estimator:

$$g_2(\hat{p}|a_0, a_1) = 0.01 + 0.9\hat{p}. \tag{3.23}$$

To compare the expected log-likelihood of the predictive estimator and that of the maximum likelihood estimator, we calculated

$$r_{like}(p^+, \hat{p}) = \frac{l(\hat{p}|\{x_i^*\})}{l(p^+|\{x_i^*\})}. \tag{3.24}$$

The following expression was also derived to compare the sum of the squared errors between the true value and the predictive estimator and that between the true value and the maximum likelihood estimator:

$$r_{error}(p^+, \hat{p}) = \frac{d(p^+)}{d(\hat{p})}. \tag{3.25}$$

The results of Eqs. (3.24) and (3.25) are illustrated in Fig. 5.

We obtained the predictive estimator

$$g_2(\hat{p}|a_0, a_1) = 0.06 + 0.9\hat{p}. \tag{3.26}$$

Because we defined the predictive estimator as that which is better than the maximum likelihood estimator in terms of the expected log-likelihood, the predictive estimator is inferior to the maximum likelihood estimator in terms of the sum of squared errors based on the true value when $\tilde{p}$ is close to 0.9.

Next, we assumed that we have information: "$\tilde{p}$ is between 0.2 and 0.4," and set $\tilde{p} = 0.2, 0.21, 0.22, \ldots, 0.4$. The results are shown in Fig. 7. We also obtained the predictive estimator

$$g_2(\hat{p}|a_0, a_1) = 0.08 + 0.7\hat{p}. \tag{3.27}$$

Furthermore, we assumed that we have information: "$\tilde{p}$ is between 0.6 and 0.8," and set $\tilde{p} = 0.6, 0.61, 0.62, \ldots, 0.8$. The results are shown in Fig. 8. Finally, we obtained the predictive estimator

$$g_1(\hat{p}|a_0, a_1) = (0.09 + 0.8\hat{p})^{0.9}. \tag{3.28}$$

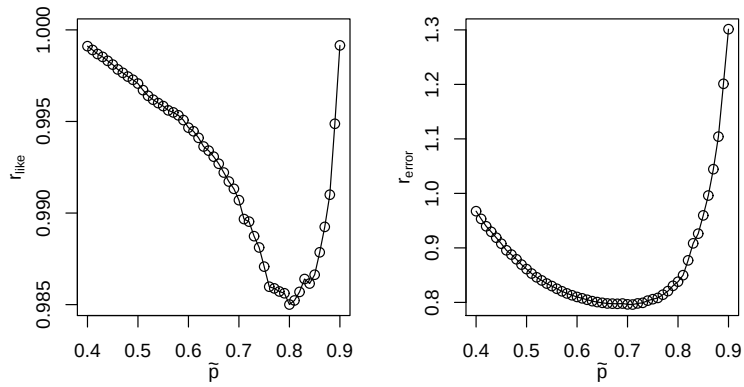

Figure 6: It is assumed that we have information: "$\tilde{p}$ is between 0.4 and 0.9." $r_{like}(p^+, \hat{p})$ (left) and $r_{error}(p^+, \hat{p})$ (right) are shown for $\tilde{p}$ equal to $\{0.4, 0.41, 0.42, \ldots, 0.9\}$.

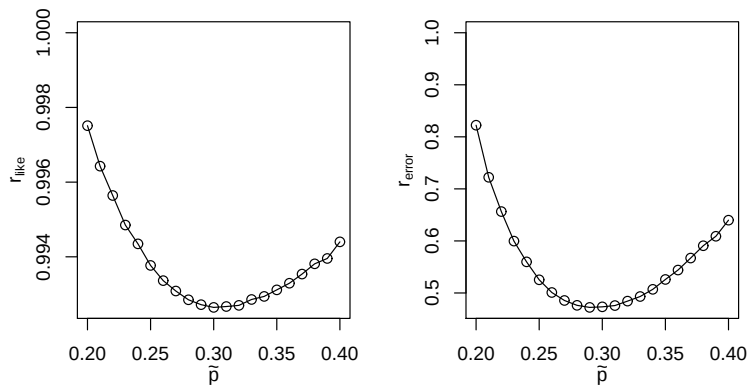

Figure 7: It is assumed that we have information: "$\tilde{p}$ is between 0.2 and 0.4." $r_{like}(p^+, \hat{p})$ (left) and $r_{error}(p^+, \hat{p})$ (right) are shown for $\tilde{p}$ equal to $\{0.2, 0.21, 0.22, \ldots, 0.4\}$.

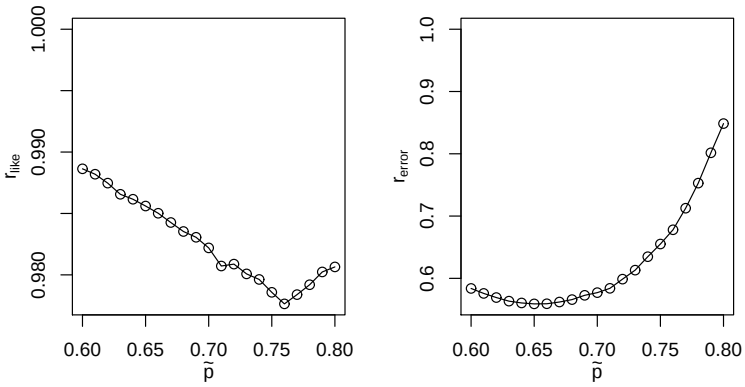

Figure 8: It is assumed that we have information: "$\tilde{p}$ is between 0.6 and 0.8." $r_{like}(p^+, \hat{p})$ (left) and $r_{error}(p^+, \hat{p})$ (right) are shown for $\tilde{p}$ equal to $\{0.6, 0.61, 0.62, \ldots, 0.8\}$.

When the possible range of $\tilde{p}$ is narrow, the distance to the true value tends to be shorter. If we know that the possible range of $\tilde{p}$ is narrow, then we have more information on the true value. Hence, the above results are intuitive.

4 Conclusions

The main estimators treated in conventional statistics are the maximum estimator and the unbiased estimator. The reason for the emphasis on these two estimators is not only their practical utility, but also their mathematical tractability, because they allow a comparatively easy analytical treatment. However, when we consider more diverse estimators, more reliable estimations are realized by extracting the maximum information from the available data. In fact, quite a few studies fucus on better estimators than the maximum likelihood estimator (e.g., [8], page 414 in [9], [10], page 332 in [11], page 431 in [12], and [13]).

To show one more example of this attempt, we derived the predictive estimator for the geometric distribution, which aims to increase the value of the expected log-likelihood. This was motivated by the goal of obtaining an estimator that provides a good fit to future data, rather than available data. Fortunately, this policy yields a predictive estimator with high predictability. This finding will hopefully advance the construction and use of predictive estimators, and encourage further studies on the analytical treatment of predictive estimators. For this purpose, parametric equations such as Eqs. (3.15), (3.16), and (3.17) should not be the only forms of predictive estimators. Functions with many parameters should be considered, as these are often used in nonparametric regression. Moreover, the form of predictive estimators should not be limited to that of Eq. (2.3), which expresses the transformation of the maximum likelihood estimator. Functions that treat the available data directly should also be considered.

Construction of the predictive estimator for the geometric distribution is also of practical value. For example, estimation of the parameter of the geometric distribution plays a critical role in evaluation of system failure rate. We obtain more useful estimates of the parameter of the geometric distribution using the predictive estimator because the predictive estimator developed here provides a better fit to future data than the maximum likelihood estimator in many circumstances; we are

typically more interested in the system in the future than that in the past. However, the parameter estimation of the geometric distribution uses truncated data in most of practical situations because of limited observation time. Hence, we should develop the idea of the predictive estimator toward more flexible ones to treat diverse problems. The simplicity of the algorithm suggested here implies that it is not very difficult to extend this method for more general use.

The construction of predictive estimators by means of analytical or numerical methods has gathered momentum, and so we should consider purpose-oriented estimators instead of limiting our scope to the maximum likelihood estimator and the unbiased estimator. As the predictive estimator is based on the derivation of estimates that fit well with future data, it directly pursues the goal of prediction, unlike the maximum likelihood estimator and the unbiased estimator. Therefore, the predictive estimator is expected to give better results than conventional methods in most practical situations. However, even if our purpose is prediction, we sometimes hope that estimators should be close to unbiased, with some slight variance. That is, the average error of many predictions should be close to 0 in some situations if the estimator largely enhances the reliability of prediction. From another perspective, it is crucial that the absolute value of the error should not depart from a specific range, so a suitable estimator for this purpose is needed. Furthermore, another estimator is required for situations in which a certain decision is made on the grounds of predicted results and the probability of making a wrong decision must be small.

There are infinite possibilities for the form and method of creating an estimator. We should select an appropriate form and method to construct an estimator. Then, we should create an estimator that suits our goal using both analytical and numerical methods. This procedure allows us to extract information efficiently from data in a certain direction; this is essential to make the best use of opportunities and resources. We are quietly confident that our future studies will develop reliable estimators that deal flexibly with diverse needs.

Acknowledgements

The author is very grateful to the referees for carefully reading the paper and for their comments and suggestions which have improved the paper.

Competing Interests

The author declares that no competing interests exist.

References

[1] Konishi S. and Kitagawa G. Information Criteria and Statistical Modeling. new York: Springer; 2008.

[2] Takezawa K. A Revision of AIC for Normal Error Models. Open Journal of Statistics. 2012;**2**(3): 309–312.

[3] Takezawa K. Learning Regression Analysis by Simulation. Tokyo: Springer; 2013.

[4] Takezawa, K. Estimation of the exponential distribution in the light of duture data. British Journal of Mathematics & Computer Science.(2015);**5**(1):128–132. `http://www.sciencedomain.org/issue.php?iid=707&id=6`

[5] Takezawa K. Estimation of Binomial Distribution in the Light of Future Data. British Journal of Mathematics & Computer Science. 2015;**10**(2):1-7.
`http://sciencedomain.org/abstract/10077`

[6] Takezawa K. Estimation of the Geometric Distribution in the Light of Future Data. British Journal of Mathematics & Computer Science. (2015);**11**(4):1-11.
`http://sciencedomain.org/abstract/11298`

[7] Bickel P.J. & Doksum K.A. Mathematical Statistics: Basic Ideas and Selected Topics, Vol I (2nd Edition). Prentice Hall; 2000.

[8] Stein C. Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. Annals of the Institute of Statistical Mathematics. 1964;**16**(1): 155–160.

[9] DeGroot MH. Probability and statistics (2nd edition). Boston: Addison–Wesley; 1986.

[10] Shinozaki N. Some modifications of improved estimators of a normal variance. Annals of the Institute of Statistical Mathematics. 1995;**47**(1): 273-286.

[11] Casella G, Berger RL. Statistical inference (2nd edition). Pacific Grove (CA): Duxbury; 2001.

[12] DeGroot MH., Schervish MJ. Probability and statistics (3rd ed.). Boston: Addison–Wesley; 2002.

[13] Ogasawara H. Bias adjustment minimizing the asymptotic mean square error. Communications in Statistics, Theory and Methods. 2015;**44**(16): 3503-3522.