British Journal of Mathematics & Computer Science
10(2): 1-7, 2015, Article no.BJMCS.19191
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Estimation of Binomial Distribution in the Light of Future Data

Kunio Takezawa[1*]

[1] *Agroinformatics Division, Agricultural Research Center, National Agriculture and Food Research Organization Kannondai 3-1-1, Tsukuba, Ibaraki 305-8666, Japan.*

Article Information

DOI: 10.9734/BJMCS/2015/19191
Editor(s):
(1) H. M. Srivastava, Department of Mathematics and Statistics, University of Victoria, Canada.
Reviewers:
(1) P. E. Oguntunde, Department of Mathematics, Covenant University, Nigeria.
(2) Jos Alejandro Gonzlez Campos, Department of Mathematic and Statistic, Playa Ancha University, Chile.
Complete Peer review History: http://sciencedomain.org/review-history/10077

Short Research Article

Received: 29 May 2015
Accepted: 15 June 2015
Published: 07 July 2015

Abstract

A predictive estimator for estimating the parameter of binomial distribution is suggested. This estimator aims to maximize the expectation of expected log-likelihood. The results given by this estimator are superior to those given by the maximum likelihood estimator in terms of the predictions when a little prior knowledge about the parameter is available.

Keywords: expected log-likelihood; binomial distribution; maximum likelihood estimator; optimization

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

1 Introduction

The maximum likelihood estimator and the unbiased estimator are common tools, which are used to estimate the parameters of a probability density function from data. However, the maximum likelihood estimator is known to give unacceptable estimates in some situations (e.g., page 343 in [1]). The unbiased estimator does not exist under certain conditions and, even if it exists, it may not be invariant (page 415 in [1]). In this paper, I propose a predictive estimator for estimating the parameter of binomial distribution. The estimator aims to maximize the expectation of the expected log-likelihood, and hereafter is called the "predictive estimator". Because nothing is known about the form of the predictive estimator, I assume a very simple form. Therefore, predictive estimators with other forms may yield better results.

**Corresponding author: E-mail: nonpara@gmail.com*

2 Predictive Estimator for Binomial Distribution

The probability density function of the binomial distribution is written as

$$f(\xi) = \binom{n}{\xi} \tilde{p}^{\xi}(1-\tilde{p})^{n-\xi} \qquad (\xi = 0, 1, 2, \ldots, n). \tag{2.1}$$

where $\tilde{p}$ is the true value of the parameter and ξ is the variable. The expectation of Eq.(2.1) can be written as

$$\sum_{\xi=0}^{n} \xi f(\xi) = n\tilde{p}. \tag{2.2}$$

Assuming that X is a random variable that obeys the probability density function $f(\xi)$ and its realization (i.e., data) is called x, then the log-likelihood ($l(p|x)$) of the data is given as

$$\frac{l(p|x)}{n} = \log(\binom{n}{x}) + x\log(p) + (n-x)\log(1-p). \tag{2.3}$$

To derive the p that maximizes the above value, the value is differentiated with respect to p and set equal to 0 as follows:

$$\frac{x}{p} - \frac{n-x}{1-p} = 0. \tag{2.4}$$

Therefore, the maximum likelihood estimator ($\hat{p}$) is obtained from the data at hand (x) as

$$\hat{p} = \frac{x}{n}. \tag{2.5}$$

Next, future data are named $\{x_i^*\}$ $(1 \leq i \leq m)$; future data comes from another sampling than that for data at hand (x). The log-likelihood ($l(\hat{p}|\{x_i^*\})$) of $\hat{p}$ in the light of this data is represented as

$$\frac{l(\hat{p}|\{x_i^*\})}{m} = \frac{1}{m}\sum_{i=1}^{m} \log(\binom{n}{x_i^*}) + \frac{1}{m}\sum_{i=1}^{m} x_i^* \log(\hat{p}) + \frac{1}{m}\sum_{i=1}^{m}(n-x_i^*)\log(1-\hat{p}). \tag{2.6}$$

The $\hat{p}$ that maximizes this value is termed $\hat{p}^*$ and is depicted as

$$\hat{p}^* = \frac{\sum_{i=1}^{m} x_i^*}{mn}. \tag{2.7}$$

Because the number of future data is infinite, m is set to be infinite. In this situation, the value of $\hat{p}^*$ is called $\hat{p}_{\infty}^*$ and Eq.(2.2) leads to

$$\hat{p}_{\infty}^* = \lim_{m\to\infty} \frac{\sum_{i=1}^{m} x_i^*}{mn} = \tilde{p}. \tag{2.8}$$

Here, the maximum likelihood estimator for an infinite number of future data is the true value of the parameter (i.e., $\tilde{p}$). Substitution of Eq.(2.8) into Eq.(2.6) yields

$$\lim_{m\to\infty} \frac{l(\hat{p}|\{x_i^*\})}{m} = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \log(\binom{n}{x_i^*}) + n\tilde{p}\log(\hat{p}) + n(1-\tilde{p})\log(1-\hat{p}). \tag{2.9}$$

When the above equation is taken into account and the log-likelihood of $\hat{p}$ for an infinite number of future data is named $l^*(\hat{p})$, we have

$$l^*(\hat{p}) = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \log(\binom{n}{x_i^*}) + n\tilde{p}\log(\hat{p}) + n(1-\tilde{p})\log(1-\hat{p}). \tag{2.10}$$

Because the first term of the right-hand side does not affect derivation of $\hat{p}$, this term can be omitted and the log-likelihood, named $l^{*\prime}(\hat{p})$), can be written as

$$l^{*\prime}(\hat{p}) = n\tilde{p}\log(\hat{p}) + n(1-\tilde{p})\log(1-\hat{p}). \tag{2.11}$$

The $\hat{p}$ given by Eq.(2.5) may not be the optimal estimator in the light of future data; therefore, the optimal estimator for future data (i.e., the predictive estimator), here called $\hat{p}^+$, can be represented as

$$\hat{p}^+ = \frac{x}{n+a}, \tag{2.12}$$

where a is a constant. Hence, the log-likelihood of $\hat{p}^+$ for an infinite number of future data is

$$l^{*\prime}(\hat{p}^+) = n\tilde{p}\log\left(\frac{x}{n+a}\right) + n(1-\tilde{p})\log\left(1-\frac{x}{n+a}\right). \tag{2.13}$$

By differentiating this equation with respect to a and setting it equal to 0, we obtain

$$a = \frac{x}{\tilde{p}} - n. \tag{2.14}$$

The resulting equation can be written as

$$\hat{p}^+ = \tilde{p}, \tag{2.15}$$

which is an obvious relationship.

Next, let us consider the averaged value of $l^{*\prime}(\hat{p}^+)$ (i.e., expectation of $l^{*\prime}(\hat{p}^+)$), which is given by sampling x infinite times. This expectation is represented as

$$\begin{aligned} E_x\left[l^{*\prime}(\hat{p}^+)\right] &= E_x\left[n\tilde{p}\log\left(\frac{x}{n+a}\right) + n(1-\tilde{p})\log\left(1-\frac{x}{n+a}\right)\right] \\ &= \frac{\sum_{\xi=1}^{n-1}\left(n\tilde{p}\log\left(\frac{\xi}{n+a}\right) + n(1-\tilde{p})\log\left(1-\frac{\xi}{n+a}\right)\right)\binom{n}{\xi}\tilde{p}^r(1-\tilde{p})^{n-\xi}}{\sum_{\xi=1}^{n-1}\binom{n}{\xi}\tilde{p}^{\xi}(1-\tilde{p})^{n-\xi}}. \end{aligned} \tag{2.16}$$

The range of summation of the right-hand side of the second line of this equation is from $r = 1$ through $r = n-1$ instead of $r = 0$ through $r = n$. This is because the possibility that x takes the value of 0 or n has to be excluded. As a result of this measure, the expectation is standardized by the constant (i.e., the denominator). Moreover, if a is set to be negative, the argument of $\log\left(1-\frac{x}{n+a}\right)$ can be negative; hence, a has to be 0 or positive. It should be noted that the type of expectation used in Eq.(2.16) is found in the derivation of AIC (Akaike's information criterion). For example, $E_{G(\mathbf{x}_n)}$ on page 55 in [2] is a similar expectation.

When $a = 0$ is assumed in Eq.(2.16), it yields the expectation of expected log-likelihood. Therefore, if the value of Eq.(2.16) for $a \neq 0$ is larger than its value for $a = 0$, an a that holds $a \neq 0$ will give a better estimator. Then, $\Delta l_f(a)$ is defined as

$$\begin{aligned} \Delta l_f(a) &= E_x\left[l^{*\prime}(\hat{p}^+)\right] - E_x\left[l^{*\prime}(\hat{p})\right] \\ &= E_x\left[n\tilde{p}\log\left(\frac{x}{n+a}\right) + n(1-\tilde{p})\log\left(1-\frac{x}{n+a}\right)\right] \\ &\quad -E_x\left[n\tilde{p}\log\left(\frac{x}{n}\right) + n(1-\tilde{p})\log\left(1-\frac{x}{n}\right)\right]. \end{aligned} \tag{2.17}$$

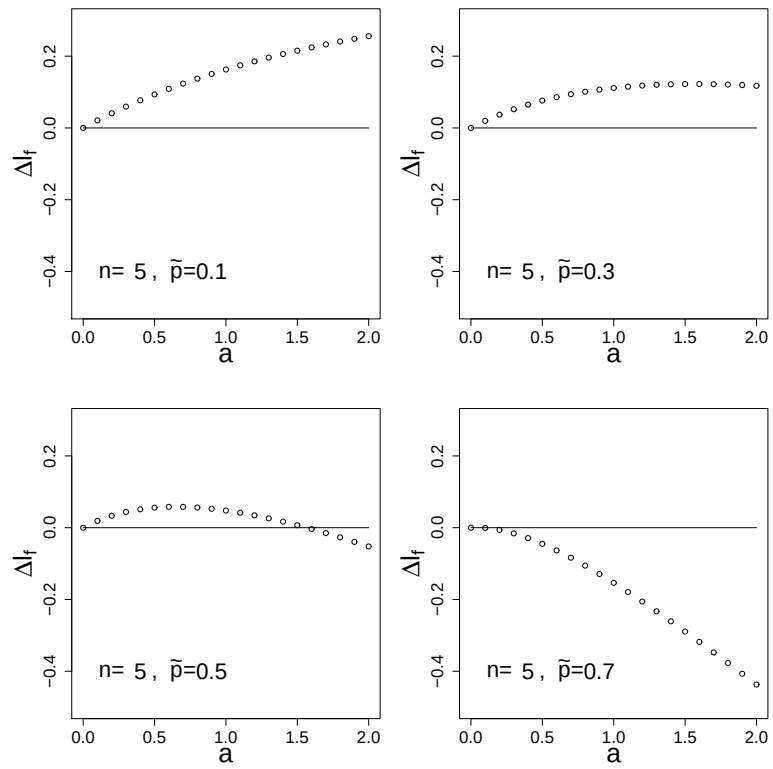

Figure 1: $\Delta l_f(a)$ given by Eq.(2.17) when $n = 5$. $\tilde{p} = 0.1$ (top left); $\tilde{p} = 0.3$ (top right); $\tilde{p} = 0.5$ (bottom left); and $\tilde{p} = 0.7$ (bottom right).

This $\Delta l_f(a)$ is calculated using Eq.(2.16). When $\Delta l_f(a)$ is calculated assuming $n = 5$, $\tilde{p} = \{0.1, 0.3, 0.5, 0.7\}$, and $a = \{0.1, 0.2, 0.3, \ldots, 2.0\}$ the results are shown in Figure1. When $n = 10$ is assumed and the other parameters remain the same as those used in Figure 1, the results are shown in Figure2. For example, if we know from the phenomenon generating the data that $n = 10$ and $0.1 \leq \tilde{p} \leq 0.5$, a should be set at 0 to 1.5. In this situation, when a is set at 0, it gives the maximum likelihood estimator. Hence, when $a = 0.5$ or $a = 1$ is used for example, the averaged expectation of expected log-likelihood yielded by repeated samplings is larger than the averaged expectation given by the maximum likelihood estimator (Figure 2), which indicates that this predictive estimator is a more desirable estimator. To optimize the value of a exactly, detailed prior knowledge about the parameters or settings of the data are required.

Next, we consider the situation when the probability of success obeys the binomial distribution and the probability of failure also obeys the binomial distribution. Assuming, $\tilde{p}_s$ is the probability of success, $\tilde{p}_f$ is the probability of failure, and $\tilde{p}_s + \tilde{p}_f = 1$. If we know that $n = 10$ and $0.1 \leq \tilde{p}_s \leq 0.5$, then setting a in (2.12) at 0.5 or 1 gives a larger value of the expectation of expected log-likelihood than is given by the maximum likelihood estimator. By contrast, because we know $0.5 \leq \tilde{p}_f \leq 0.9$, setting of a in Eq.(2.12) at 0 (i.e., use of the maximum likelihood estimator) is a reasonable choice, and $\hat{p}_s^+ + \hat{p}_f^+ \neq 1$. Conventionally, we believe that when $\tilde{p}_s + \tilde{p}_f = 1$ holds, $\hat{p}_s + \hat{p}_f = 1$ is satisfied.

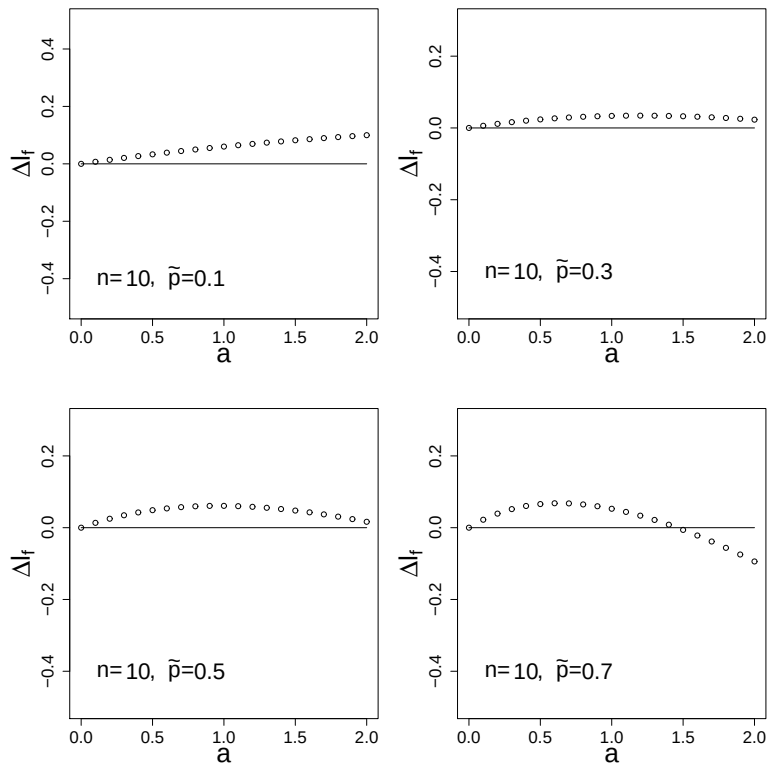

Figure 2: $\Delta l_f(a)$ given by Eq.(2.17) when $n = 10$. $\tilde{p} = 0.1$ (top left); $\tilde{p} = 0.3$ (top right); $\tilde{p} = 0.5$ (bottom left); and $\tilde{p} = 0.7$ (bottom right).

This inference is not obviously valid. When the purpose of the estimation is maximization of the expectation of expected log-likelihood, this relationship is no longer satisfied.

3 Conclusions

Bayes estimator ([3-8]) is conventionally the most common tool for using prior knowledge to estimate parameters of a probability density function. The predictive estimator method proposed here uses prior knowledge such as $0.1 \leq \tilde{p}_s \leq 0.5$ to make the expectation of expected log-likelihood larger than that given by the maximum likelihood estimator. In the case of the predictive estimator for the exponential distribution, the estimator given by multiplying $\left(1 - \frac{1}{n}\right)$ (n is the number of data) with the maximum likelihood estimator yields a larger value for the expected log-likelihood than simply using the maximum likelihood estimator [9]. Therefore, even if no prior information about the true parameter is available, a predictive estimator that performs better than the maximum likelihood estimator is available. Conversely, our predictive estimator of the binomial distribution requires some prior knowledge about the parameters. However, a predictive estimator that does not need any prior knowledge of the parameters is still possible. Furthermore, if we develop a method of constructing an optimal predictive estimator depending on the prior knowledge on the parameters, we expect that such a predictive estimator will replace the maximum likelihood estimator. The

same is true for the estimation of parameters of distributions other than the binomial distribution.

The MSE (mean squared error) is defined as below (page 330 in [1]).

$$MSE = E_\theta(W - \theta)^2, \tag{3.1}$$

where θ is the true parameter, W is the estimator of the parameter, and E_θ is the expectation given by averaging the results of repeated samplings. As noted on page 332 of [1], MSE depends upon the true parameter; hence, some prior knowledge about the parameters is needed to estimate the parameters by minimizing MSE. In this regard, minimizing MSE is similar to maximizing the expectation of expected log-likelihood. Estimation of variance of data by minimizing MSE, however, does not need such prior knowledge. This situation should be handled as an exception in the same way that the exponential distribution has an exceptional predictive estimator and the normal distribution has "third variance" ([10-11]).

Moreover, methodologies based on predictive estimators have something in common with smoothing splines (e.g., sections 2 and 3 of [12], and section 3 of [13]). Currently, no definitive theory on the desirable form of the roughness penalty in smoothing splines is available; hence, we could not confirm that the most commonly-used roughness penalty is the optimal one. Diverse theories and numerical simulations, however, indicate that our common roughness penalty is valid from various perspectives. Further, we have not confirmed that the use of Eq.(2.12) as a predictive estimator is appropriate or optimal. However, Fig.1 and Fig.2 show that when a is set in a valid range, the predictive estimator defined in Eq.(2.12) outperforms the maximum likelihood estimator. Therefore, although we cannot deny the possibility that better predictive estimators exist, tentatively Eq.(2.12) with an appropriate value of a can be used. That is, setting an appropriate value of a corresponds to setting appropriate values for the smoothing parameters in smoothing splines. Thus, the use of a predictive estimator such as Eq.(2.12) can be justified in the same sense that the use of smoothing splines is justified, even though it has not been proven that smoothing splines leads to better results than those obtained by all the other estimation methods.

The characteristics of the predictive estimator should now be evaluated for various estimations; in particular, the predictive estimator should be compared with the maximum likelihood estimator and unbiased estimator. Such studies will help establish major techniques for using data more efficiently.

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

[2] Konishi S, Kitagawa G. Information criteria and statistical modeling. New York: Springer; 2008.

[3] Robert CP. The bayesian choice: From decision-theoretic foundations to computational implementation. 2nd edition. Springer; 2007.

[4] Carlin BP, Louis TA. Bayesian methods for data analysis. Third Edition. Chapman & Hall CRC; 2008.

[5] O'Hagan A, Forster J. Kendall's advanced theory of statistics. Bayesian Inference, 1 edition. Wiley. 2010;2B.

[6] Lee PM. Bayesian statistics: An introduction. Wiley; 2012.

[7] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian data analysis, third edition. Chapman & Hall CRC; 2013.

[8] Bessiere P, Mazer E, Ahuactzin JM, Mekhnacha K. Bayesian programming. Chapman & Hall/CRC; 2013.

[9] Takezawa K. Estimation of the exponential distribution in the light of future data. British Journal of Mathematics & Computer Science. 2015;5(1):128-132.

[10] Takezawa K. A revision of aic for normal error models. Open Journal of Statistics. 2012;2(3):309-312.

[11] Takezawa K. Learning regression analysis by simulation. Springer; 2013.

[12] Green PJ, Silverman BW. Nonparametric regression and generalized linear models: A roughness penalty approach. Chapman & Hall CRC; 1993.

[13] Takezawa K. Introduction to nonparametric regression. Wiley; 2005.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/10077