# Probability density functions for prediction

***Original Research Article***


## Abstract

When data are found to be realizations of a specific distribution, constructing the probability density function based on this distribution may not lead to the best prediction result. In this study, numerical simulations are conducted using data that follow a normal distribution, and we examine whether probability density functions that have shapes different from that of the normal distribution can yield larger log-likelihoods than the normal distribution in light of future data. The results indicate that fitting realizations of the normal distribution to a different probability density function produces better results from the perspective of predictive ability. Similarly, a set of simulations using the exponential distribution shows that better predictions are obtained when the corresponding realizations are fitted to a probability density function that is slightly different from the exponential distribution. These observations demonstrate that when the form of the probability density function that generates the data is known, the use of another form of the probability density function may achieve more desirable results from the standpoint of prediction.


*Keywords: exponential distribution; future data; log-likelihood; normal distribution; predictive estimator; probability density function*

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

## 1 Introduction

The derivation of a probability density function on the basis of data is one of the major subjects of interest in statistical science, and has been discussed in many papers and books (e.g., (4), (1)). In conventional theories, if we know that the data are realizations of, for example, the normal distribution, the data are fitted to the normal distribution. Our problem is then to find estimators that give parameters of the normal distribution.

When we use the normal distribution, the focus is on which estimators we should adopt when calculating the value of the variance. The unbiased variance ($\hat{\sigma}^2_{ub}$) is a typical estimator for determining

the variance of the normal distribution. The unbiased variance is defined as

$$\hat{\sigma}_{ub}^2 = \frac{RSS}{n-1}, \tag{1.1}$$

where $n$ is the number of data and $RSS$ is defined as

$$RSS = \sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{1.2}$$

in which $\{x_i\}$ $(1 \leq i \leq n)$ are the data and $\bar{x}$ is the average of the data.

The maximum likelihood variance ($\hat{\sigma}_{like}^2$) is a well-known quantity, and is defined as

$$\hat{\sigma}_{like}^2 = \frac{RSS}{n}. \tag{1.3}$$

The "third variance" ($\hat{\sigma}_{third}^2$) is another estimator for variance, and is defined as follows ((6), section $5.5$ of (7), (2), (3)):

$$\hat{\sigma}_{third}^2 = \frac{RSS}{n-\alpha}, \tag{1.4}$$

where $\alpha$ is defined as

$$\alpha = \frac{4n}{n+1} \approx 4. \tag{1.5}$$

In the simulations described below, $\frac{4n}{n+1}$ is employed as $\alpha$.

The maximum likelihood variance and unbiased variance are widely used as estimators of variance for the normal distribution. However, the third variance is a better choice than both the maximum likelihood variance and unbiased variance when the aim is to maximize the log-likelihood in light of future data (i.e., to construct a beneficial probability density function in terms of prediction). That is, when the values of the log-likelihood in light of future data are compared, the third variance provides a larger value than either the maximum likelihood variance or the unbiased variance. However, it has not yet been proved whether the third variance is the best and unique estimator from the perspective of the log-likelihood in light of future data. Hence, we cannot deny the possibility that another estimator gives a better probability density function than the third variance.

To estimate the probability density function for a given set of data, if the parametric form of the probability density function is known, the parameters of the function should be estimated using the data. For example, if we know that the data are realizations of the normal distribution, common sense tells us that the data should be fitted to the normal distribution. However, even if the form of the probability density function that generates the data is known, the probability density function that gives the best fit to future data may not necessarily be represented by the parametric form of the probability density function that generates the data. For example, when the data are known to be realizations of a seventh-order polynomial equation plus independent and identically distributed errors of the normal distribution, fitting a quadratic equation using least-squares can lead to better results than a seventh-order polynomial equation from the perspective of prediction (see page 16 in (5)). A similar situation may occur in the estimation of the probability density function.

This paper describes numerical simulations conducted to investigate the possibility of obtaining better probability density functions by fitting probability density functions that have forms other than that of the probability density function that generates the given data. These numerical simulations open up a new horizon for techniques where the aim is to derive more beneficial probability density functions than those given by conventional methods.

## 2 Procedures and results of numerical simulations

We consider the possibility of obtaining better probability density functions in terms of the log-likelihood in light of future data. The numerical simulations involve fitting normally distributed data to a probability density function other than that of the normal distribution.

First, numerical simulations were conducted to confirm that the third variance maximizes the log-likelihood in light of future data when normally distributed data are fitted to the normal distribution. This procedure involved the following steps:

(1) Data generation: the data were taken to be $50$ realizations of the normal distribution with a mean of $0$ and variance of $4^2$.
(2) Using the mean and the variance given by the data generated in step (1), the probability density function of the normal distribution was constructed. The mean of the normal distribution is the average of the data and the variance of the normal distribution is given by multiplying the third variance by $\beta$. Thus, the following probability density function is derived:

$$p_1(x) = \left(\frac{1}{\sqrt{2\pi\beta\hat{\sigma}^2_{third}}}\right)\exp\left(-\frac{1}{2\beta\hat{\sigma}^2_{third}}(x-\bar{x})^2\right). \tag{2.1}$$

(3) To examine the validity of the probability density function constructed in step (2), the log-likelihood in light of future data was calculated. The future data in this case were $100$ realizations of the same normal distribution described above. The log-likelihood ($l^*$) in light of future data is defined as

$$l^* = \sum_{i=1}^{100} \log(p_1(x_i^*)), \tag{2.2}$$

where $\{x_i^*\}$ $(1 \leq i \leq 100)$ represent the future data.
(4) Steps (1)–(3) were conducted $20,000$ times with different initial values of pseudo-random numbers to calculate the average log-likelihood in light of future data.
(5) Using one of $\{0.90, 0.91, 0.92, \ldots, 1.1\}$ as $\beta$, steps (1)–(4) were repeated to determine the relationship between $\beta$ and the log-likelihood in light of future data.
(6) Steps (1)–(5) were performed $10$ times.

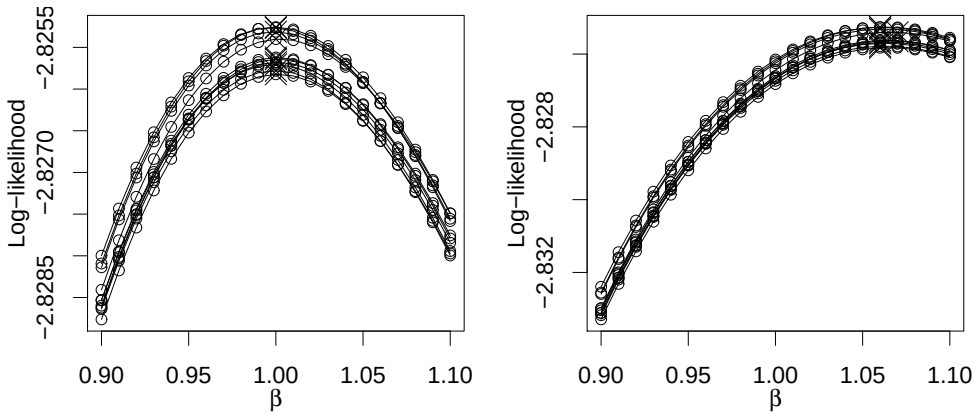


Figure 1: Relationship between $\beta$ and log-likelihood in light of future data. "$\times$" denotes the maximum point.

Figure 1(left) indicates that, from the perspective of the log-likelihood in light of future data, the third variance leads to better results than a variance that is somewhat larger or smaller than the third

variance. Figure 1(right) shows the results of simulations using the unbiased variance instead of the third variance. That is, $\hat{\sigma}^2_{third}$ (Eq. (1.4)) in Eq. (2.1) was replaced with $\hat{\sigma}^2_{ub}$ (Eq. (1.1)). This graph indicates that a variance with a slightly higher value than the unbiased variance is favorable from a viewpoint of the log-likelihood in light of future data.

Next, the log-likelihoods in light of future data were calculated using the following probability density function:

$$\begin{aligned} p_2(x) &= (1-\theta)\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2_{third}}}\right)\exp\left(-\frac{1}{2\hat{\sigma}^2_{third}}(x-\bar{x})^2\right) \\ &+ \theta\mathrm{tri}(x;\bar{x},\hat{\sigma}^2_{third},w). \end{aligned} \tag{2.3}$$

where $\bar{x}$ and $\hat{\sigma}^2_{third}$ are the average and third variance of the data ($\{x_i\}$), respectively. $\mathrm{tri}(x;\bar{x},\hat{\sigma}^2_{third},w)$ is the probability density function of the triangular distribution. The shape of the density function for the triangular distribution is illustrated in Fig. 2. $\theta$ gives the weights of the two probability density functions; $0 \leq \theta \leq 1$ is assumed.

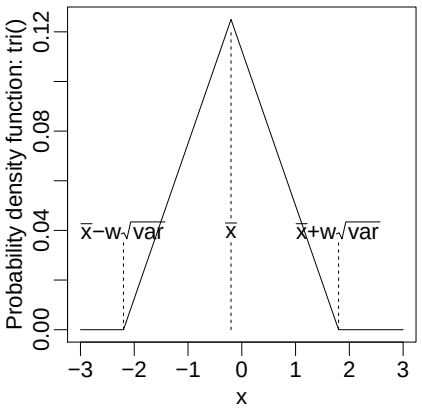


Figure 2: Probability density function of the triangular distribution. "$var$" denotes the third variance of the data.

The data are realizations of $N(0,16)$ (the normal distribution with mean $0$ and variance $16$). The number of data is fixed at $10$. The values of $l^*$ (Eq. (2.2)) in which $p_1(x^*_i)$ is replaced by $p_2(x^*_i)$ were calculated using $100$ future data, which are other realizations of $N(0,16)$. This calculation was conducted $20,000$ times with different initial values of pseudo-random numbers to derive the average of the $20,000$ log-likelihoods in light of future data. The value of $w$ was set to one of $\{0.2, 0.4, 0.6, \ldots, 2.2\}$ and the value of $\theta$ was set to one of $\{0.0, 0.02, 0.04, \ldots, 0.4\}$. The procedure described above enables us to calculate the average log-likelihood in light of future data.

The results of these calculations are shown in Fig. 3(left), and indicate that $\theta = 0.14, w = 1.2$ yields the maximum value of the mean log-likelihood in light of future data. The value of $p_2(x)$ (Eq. (2.3)) based on $\theta = 0.14, w = 1.2$ is illustrated in Fig. 3(right); a set of example data is superimposed. The discrepancy between $p_2(x)$ and the normal distribution is substantial. This indicates that a probability distribution function with a shape somewhat different from the normal distribution leads to better prediction performance.

Next, we set the number of data to 5. This setting leads to Fig. 4(left), which indicates that $\theta = 0.28, w = 1$ yields the maximum value of the mean log-likelihood in light of future data. The value of $p_2(x)$ (Eq. (2.3)) based on $\theta = 0.28, w = 1$ is illustrated in Fig. 4(right); a set of example data is superimposed. A comparison with the results when the number of data is 10 (Fig. 3(left)) shows that using fewer data increases the discrepancy between the shape of the optimized probability density function and that of the normal distribution.

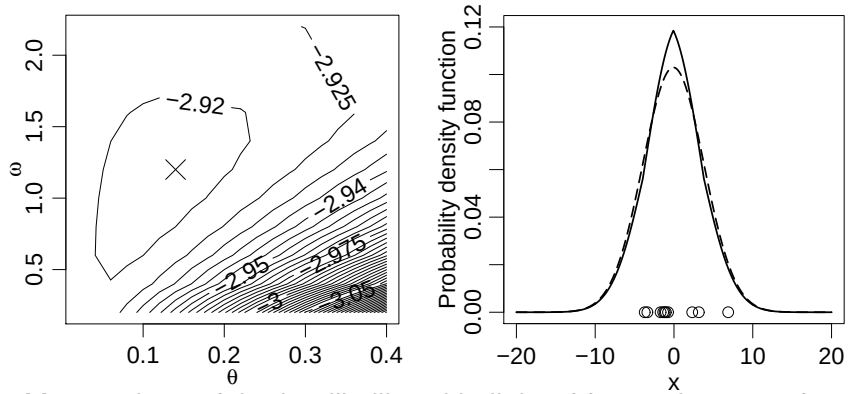


Figure 3: Mean values of the log-likelihood in light of future data as a function of $\theta$ and $w$. The number of data is $10$. The optimal parameter values are $\theta = 0.14, w = 1.2$ (left). The dashed line represents the probability density function of the normal distribution given by the third variance. The solid line represents the probability density function of the optimized weighted average of the normal distribution and the triangular distribution. "◯" denotes a set of example data (right).

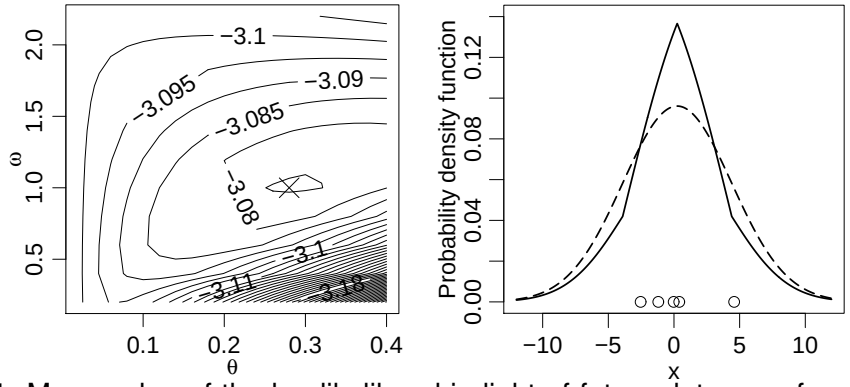


Figure 4: Mean value of the log-likelihood in light of future data as a function of $\theta$ and $w$. The number of data is $5$. The optimal parameter values are $\theta = 0.28, w = 1$ (left). The dashed line represents the probability density function of the normal distribution given by the third variance. The solid line represents the probability density function of the optimized weighted average of the normal distribution and the triangular distribution. "◯" denotes a set of example data (right).

Next, we replaced $p_2(x)$ (Eq. (2.3)) with

$$\begin{aligned} p_3(x) &= (1-\theta)\left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2_{third}}}\right)\exp\left(-\frac{1}{2\hat{\sigma}^2_{third}}(x-\bar{x})^2\right) \\ &+ \theta\mathrm{uni}(x;\bar{x},\hat{\sigma}^2_{third},w). \end{aligned} \tag{2.4}$$

Here, $\mathrm{uni}(x;\bar{x},\hat{\sigma}^2_{third},w)$ is the probability density function of the uniform distribution. The shape of the density function of the uniform distribution is illustrated in Fig. 5. $\theta$ gives the weights of the two probability density functions; $0 \leq \theta \leq 1$ is assumed.

Numerical simulations were conducted using the same data as with $p_2(x)$ (Eq. (2.3)). The value of $w$ was set to one of $\{0.1, 0.2, 0.3, \ldots, 1.1\}$ and the value of $\theta$ was set to one of $\{0.0, 0.01, 0.015, \ldots, 0.095\}$. With these settings, the log-likelihoods in light of future data were calculated.

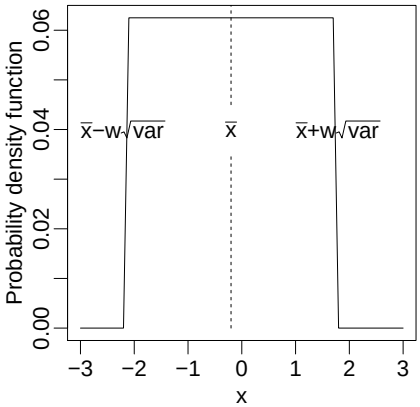


Figure 5: Probability density function of the uniform distribution. "$var$" denotes the third variance of the data.

The results of these numerical simulations are presented in Fig. 6(left). This graph indicates that $\theta = 0.07, w = 0.7$ yields the maximum value of the mean log-likelihood in light of future data. The value of $p_3(x)$ (Eq. (2.4)) based on $\theta = 0.07, w = 0.7$ is illustrated in Fig. 6(right); a set of example data is superimposed. The discrepancy between $p_3(x)$ and the normal distribution is not negligible. This result shows that a probability distribution function with a somewhat different shape to that of the normal distribution provides a better prediction result.

Next, we set the number of data to 5. The value of $w$ was set to one of $\{0.1, 0.2, 0.3, \ldots, 1.1\}$ and the value of $\theta$ was set to one of $\{0.0, 0.02, 0.04, \ldots, 0.38\}$. The log-likelihoods in light of future data were calculated with these settings, and the results are shown in Fig. 7(left). This graph indicates that $\theta = 0.18, w = 0.6$ yields the maximum value of the mean log-likelihood in light of future data. The value of $p_3(x)$ (Eq. (2.4)) given by $\theta = 0.18, w = 0.6$ is illustrated in Fig. 7(right); a set of example data is superimposed. The optimized probability density function differs considerably from that of the normal distribution.

# 3 Simulations using exponential distribution

The probability density function ($f(x)$) of the exponential distribution is

$$f(x) = \begin{cases} \tilde{\lambda}\exp(-\tilde{\lambda}x) & if \quad x \geq 0 \\ 0 & if \quad x < 0. \end{cases} \tag{3.1}$$

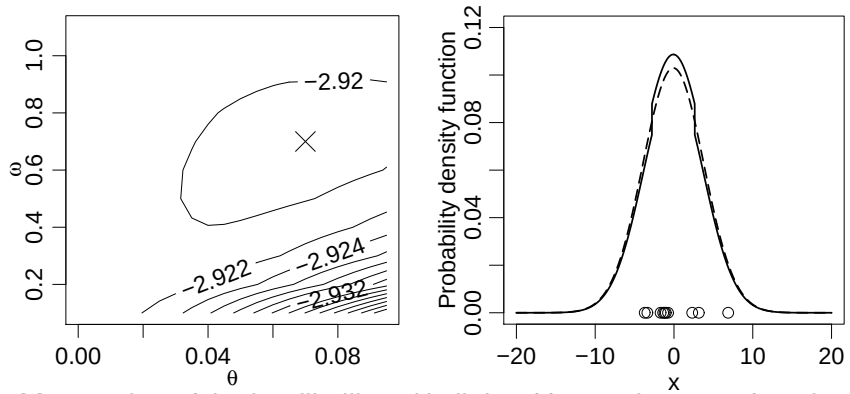


Figure 6: Mean value of the log-likelihood in light of future data as a function of $\theta$ and $w$. The number of data is $10$. The optimal parameter values are $\theta = 0.07, w = 0.7$ (left). The dashed line represents the probability density function of the normal distribution given by the third variance. The solid line represents the probability density function of the optimized weighted average of the normal distribution and uniform distribution. "◯" denotes a set of example data (right).

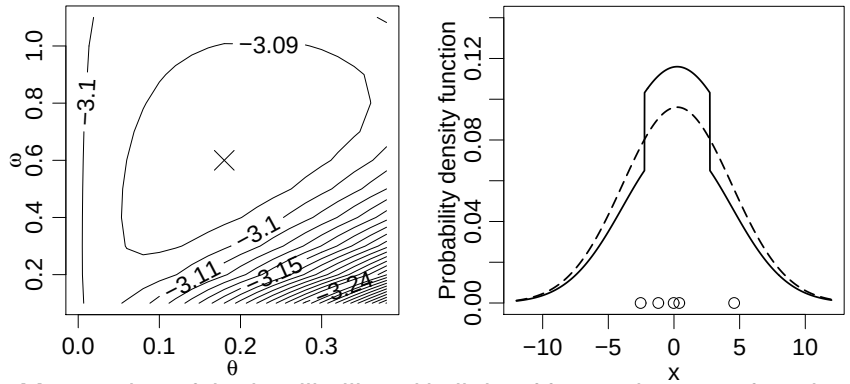


Figure 7: Mean value of the log-likelihood in light of future data as a function of $\theta$ and $w$. The number of data is $5$. The optimal parameter values are $\theta = 0.18, w = 0.6$ (left). The dashed line represents the probability density function of the normal distribution given by the third variance. The solid line represents the probability density function of the optimized weighted average of the normal distribution and uniform distribution. "◯" denotes a set of example data (right).

Let us assume that the realizations of the random variable obeying this distribution are $\{x_i\}$ $(1 \leq i \leq n)$. Then, the log-likelihood ($l(\lambda|\{x_i\})$) of $\lambda$ is

$$\frac{l(\lambda|\{x_i\})}{n} = \log(\lambda) - \frac{\lambda}{n}\sum_{i=1}^{n} x_i. \tag{3.2}$$

To derive the value of $\lambda$ that maximizes the log-likelihood, we differentiate this equation with respect to $\lambda$ and set the result equal to $0$. Then, we have

$$\frac{1}{\lambda} - \frac{1}{n}\sum_{i=1}^{n} x_i = 0. \tag{3.3}$$

This leads to the maximum likelihood estimator:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}. \tag{3.4}$$

This $\hat{\lambda}$ is the maximum likelihood estimator in light of the data at hand ($\{x_i\}$). Next, we assume that $\{x_i^*\}$ $(1 \leq i \leq m)$ are the future data. The log-likelihood ($l(\hat{\lambda}|\{x_i\})$) of $\hat{\lambda}$ in light of this future data is

$$\frac{l(\hat{\lambda}|\{x_i^*\})}{m} = \log(\hat{\lambda}) - \frac{\hat{\lambda}}{m}\sum_{i=1}^{m} x_i^*. \tag{3.5}$$

By maximizing the expectation of this value with respect to $\{x_i^*\}$, the predictive estimator of the exponential distribution is obtained as follows ((8)):

$$\lambda^+ = \Big(1 - \frac{1}{n}\Big)\hat{\lambda} = \Big(1 - \frac{1}{n}\Big)\frac{n}{\sum_{i=1}^{n} x_i}. \tag{3.6}$$

Numerical simulations were carried out to confirm that $\lambda^+$ as defined above maximizes the log-likelihood in light of future data when exponentially distributed data are fitted to the exponential distribution. This procedure involved the following steps:

(1) Data generation: the data in this case were $50$ realizations of the exponential distribution with $\lambda$ = $5$.
(2) Using $\hat{\lambda}$ (Eq. (3.4)) given by the data generated in step (1), the probability density function of the exponential distribution was constructed. The value of $\lambda$ was calculated by multiplying $\lambda^+$ (Eq. (3.6)) by $\beta$. That is, the following probability density function was derived:

$$p_4(x) = \beta\lambda^+ \exp(-\beta\lambda^+ x). \tag{3.7}$$

(3) To examine the validity of the probability density function constructed in step (2), the log-likelihood in light of future data was calculated. The future data in this case were $100$ realizations of the same exponential distribution defined above. The log-likelihood ($l^*$) in light of future data is defined as

$$l^* = \sum_{i=1}^{100} \log(p_4(x_i^*)), \tag{3.8}$$

where $\{x_i^*\}(1 \leq i \leq 100)$ represent the future data.
(4) Steps (1)–(3) were conducted $20,000$ times with different initial values of pseudo-random numbers and the average log-likelihood in light of future data was calculated.
(5) Using one of $\{0.90, 0.91, 0.92, \ldots, 1.1\}$ as $\beta$, steps (1)–(4) were repeated to determine the relationship between $\beta$ and the log-likelihood in light of future data.
(6) Steps (1)–(5) were carried out $10$ times.

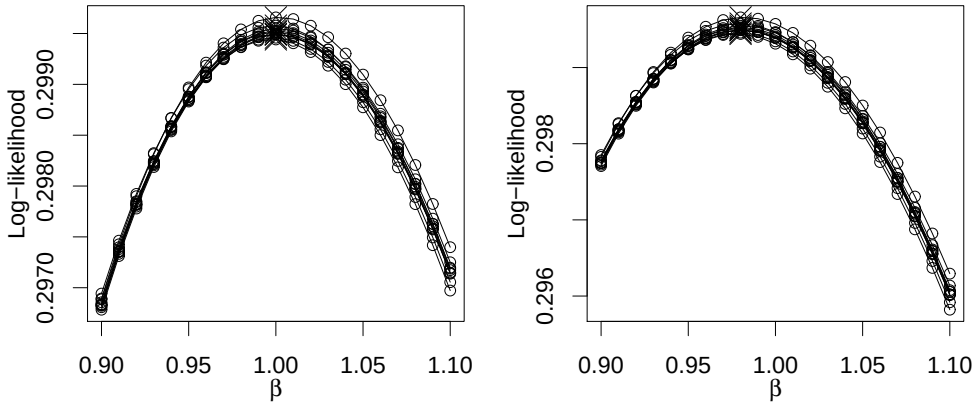


Figure 8: Relationship between $\beta$ and log-likelihood in light of future data. "$\times$" denotes the maximum point.

Figure 8(left) indicates that, from the perspective of the log-likelihood in light of future data, $\lambda^+$ (predictive estimator) leads to better results than estimators that take somewhat larger or smaller values than $\lambda^+$. Figure 8(right) shows the results of simulations using $\hat{\lambda}$ instead of $\lambda^+$. That is, $\lambda^+$ in Eq. (3.7) was replaced with $\hat{\lambda}$. This graph indicates that a slightly smaller estimator than $\hat{\lambda}$ is favorable in terms of the mean log-likelihood in light of future data.

Next, the log-likelihood in light of future data was calculated and the relationship with $\theta$ was investigated using the following probability density function:

$$p_5(x) = (1-\theta)\Big(\lambda^+ \mathrm{exp}(-\lambda^+ x)\Big) + \theta\Big(\beta\lambda^+ \mathrm{exp}(-\beta\lambda^+ x)\Big). \tag{3.9}$$

First, we set $\beta = 1.5$ and used one of $\{0, 0.002, 0.004, \ldots, 0.038\}$ as $\theta$. Therefore, if the exponential distribution based on $\lambda^+$ was the best probability density function in light of future data, the setting $\theta = 0$ would maximize the log-likelihood in light of future data. If $\theta > 0$ maximizes the log-likelihood in light of future data, then we have a counter-example against the conventional belief that realizations of the exponential distribution should be fitted to the exponential distribution.

The numerical simulations were performed as follows:
(1) Data generation: the data in this case were $50$ realizations of the exponential distribution with $\lambda$ = $5$.
(2) Using $\hat{\lambda}$ (Eq. (3.4)) given by the data in step (1), a probability density function in the form of Eq. (3.9) was produced.
(3) To examine the validity of the probability density function constructed in step (2), the log-likelihood in light of future data was calculated. The future data in this case were $100$ realizations of the same exponential distribution as defined above.
(4) Steps (1)–(3) were conducted $20,000$ times with different initial values of pseudo-random numbers to obtain the average log-likelihood in light of future data.
(5) Using one of $\{0, 0.002, 0.004, \ldots, 0.038\}$ as $\theta$ in Eq. (3.9), steps (1)–(4) were repeated to investigate the relationship between $\theta$ and the log-likelihood in light of future data.
(6) Steps (1)–(5) were repeated $10$ times.

The results of these procedures are shown in Fig. 9(left). The log-likelihood in light of future data is not maximized when $\theta = 0$. Rather, a positive value of $\theta$ maximizes the log-likelihood. Furthermore, numerical simulations comparing the results given by $\theta = 0$ with those given by $\theta = 0.015$ were conducted $100$ times. Figure 9(right) shows the values obtained by subtracting the log-likelihood in

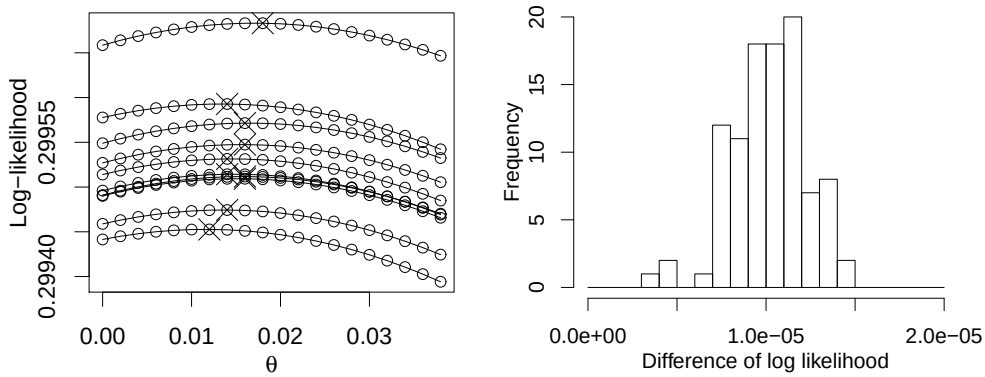


Figure 9: Relationship between $\theta$ and log-likelihood in light of future data when $\beta = 1.5$ (left). Values obtained by subtracting log-likelihood in light of future data with $\beta = 1.5$ and $\theta = 0$ from that with $\beta = 1.5$ and $\theta = 0.015$ (right).

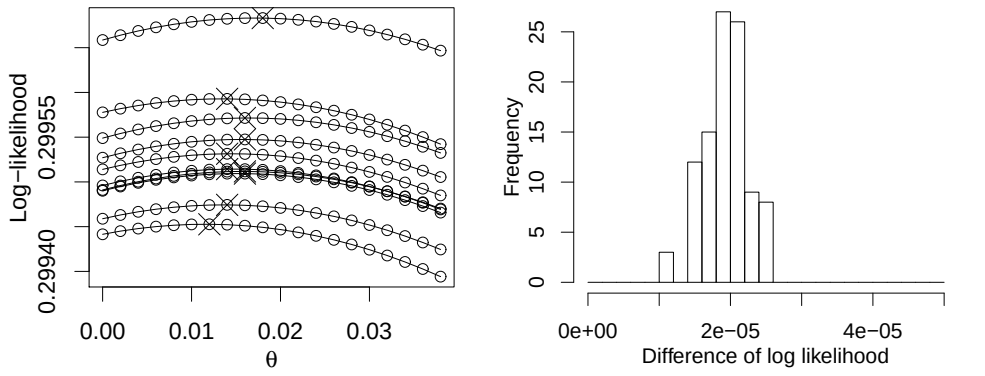


Figure 10: Relationship between $\theta$ and log-likelihood in light of future data when $\beta = 2$ (left). Values obtained by subtracting log-likelihood in light of future data with $\beta = 2$ and $\theta = 0.015$ from that with $\beta = 2$ and $\theta = 0$ (right).

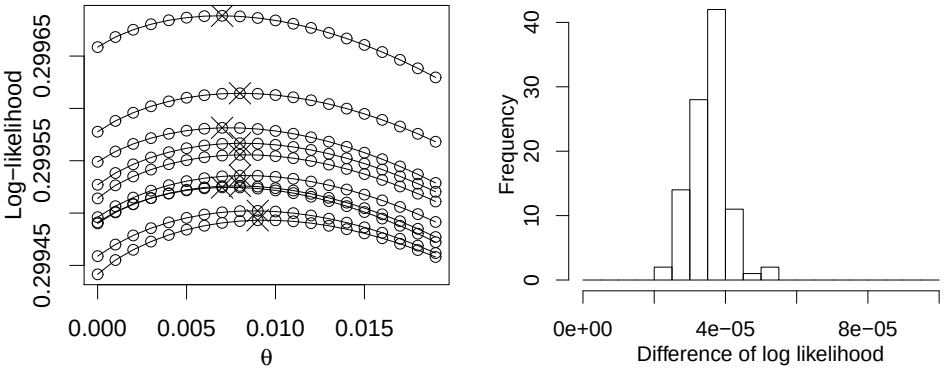


Figure 11: Relationship between $\theta$ and log-likelihood in light of future data when $\beta = 0.5$ (left). Values obtained by subtracting log-likelihood in light of future data with $\beta = 0.5$ and $\theta = 0.0$ from that with $\beta = 0.5$ and $\theta = 0.008$ (right).

light of future data with $\theta = 0$ from that with $\theta = 0.015$. These values are the averages of $20,000$ log-likelihood values, and $100$ future data were used to calculate each log-likelihood. Because all values in this graph are positive, we conclude that $\theta = 0.008$ yields larger values of the log-likelihood in light of future data in all $100$ numerical simulations.

Next, steps (1)–(4) were performed with $\beta = 2$ and one of $\{0, 0.002, 0.04, \ldots, 0.038\}$ as $\theta$ in Eq. (3.9). The results are presented in Fig. 10(left). Numerical simulations comparing the setting of $\theta = 0$ with that of $\theta = 0.015$ were conducted $100$ times. Figure 10(right) illustrates the values obtained by subtracting the log-likelihood in light of future data with $\theta = 0$ from that with $\theta = 0.015$.

Furthermore, steps (1)–(4) were performed with $\beta = 0.5$ and one of $\{0, 0.001, 0.002, \ldots, 0.019\}$ as $\theta$ in Eq. (3.9). The results are illustrated in Fig. 11(left). Numerical simulations comparing the results given by $\theta = 0$ with those given by $\theta = 0.008$ were repeated $100$ times. Figure 11(right) shows the log-likelihood in light of future data given by $\theta = 0.008$ subtracted from that given by $\theta = 0$.

# 4 Conclusions

The numerical simulations performed in this study show that, from the perspective of the log-likelihood in light of future data, it can be preferable to fit data generated by a specific probability density function to a probability density function other than the original probability density function. Common sense tells us that the original probability density function should be used for fitting if the data are considered to be realizations of that specific probability density function. In this instance, the estimators for calculating the parameters of the probability density function have been widely discussed. However, when we derive a probability density function obeyed by realizations, the derivation of estimators for calculating parameters should not be restricted by the assumption that the same probability density function must be used to fit the data. This paper has demonstrated this for cases where we wish to obtain the optimal probability density function from the standpoint of prediction. Instead, we should take account of the possibility that a slightly different probability density function from the original one that generated the data may lead to larger values of the log-likelihood in light of future data.

In future work, an analytical investigation of the probability density function that should be fitted to the data is required. For example, we should obtain the analytical form of the best probability density function, in the sense of prediction, for fitting data generated from the normal distribution

or exponential distribution. Such research will change our understanding of the construction of probability density functions on the basis of data.

# References

[1] Bickel PJ, Doksum KA. Mathematical Statistics: Basic Ideas and Selected Topics, Vol. I. 2nd ed. Upper Saddle River, USA: Prentice Hall; 2000.

[2] Ogasawara H. A family of the adjusted estimators maximizing the asymptotic predictive expected log-likelihood. Behaviormetrika. 2017;44:57-95.

[3] Ogasawara H. Predictive estimation of a covariance matrix and its structural parameters. J Jpn Soc Comput Stat. 2017;30:45-63.

[4] Silverman BW. Density estimation for statistics and data analysis. London, UK: Chapman & Hall/CRC; 1986.

[5] Takezawa K. Introduction to Nonparametric Regression. Hoboken, USA: Wiley; 2005.

[6] Takezawa K. A Revision of AIC for Normal Error Models. Open J Stat. 2012;2(3):309-12.

[7] Takezawa K. Learning regression analysis by simulation. Tokyo, Japan: Springer; 2014.

[8] Takezawa K. Estimation of the exponential distribution in the light of future data. Br J Math Comput Sci. 2015;5(1):128-32.