EXACT AND APPROXIMATE INFERENCES FOR AN EXPONENTIAL MEAN FROM TYPE I CENSORED DATA

Abe, Takayuki
Biostatistics Division, Banyu Pharmaceutical Co. Ltd.

Iwasaki, Manabu
Department of Computer and Information Science, Seikei University

https://doi.org/10.5109/12589

出版情報：Bulletin of informatics and cybernetics. 37, pp.31-39, 2005-12. Research Association of Statistical Sciences
バージョン：
権利関係：

EXACT AND APPROXIMATE INFERENCES FOR AN EXPONENTIAL MEAN FROM TYPE I CENSORED DATA

by

Takayuki Abe and **Manabu Iwasaki**

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.37

FUKUOKA, JAPAN
2005

Bulletin of Informatics and Cybernetics, Vol. 37, 2005

EXACT AND APPROXIMATE INFERENCES FOR AN EXPONENTIAL MEAN FROM TYPE I CENSORED DATA

By

Takayuki Abe* and **Manabu Iwasaki†**

Abstract

We will consider statistical inferences for an exponential mean from Type I right censored data. Most inferential procedures developed so far are conditional in the sense that they can be used only when the number of observed lifetime data is at least one. One of our aims of the present article is to remove this restriction to allow zero observed frequency. An expression for exact unconditional probability calculus is given. Our second aim is to evaluate some approximate procedures by using the exact formula whether or not they give sensible numerical values.

Key Words and Phrases: Confidence interval, Exact calculus, Likelihood ratio, Type I censoring

1. Introduction

Let us consider lifetime data from an exponential distribution with mean lifetime θ. In usual lifetime data analyses lifetimes are not always fully observed. Some observations will be censored due, for example, to the limited study period, in which the lifetime of the subjects still alive at the end of the study will not be recorded.

It is assumed here that the lifetime is possibly right censored at the same pre-specified time point c. This is the problem called time- or Type I censoring in the literature, see, for example, Lawless (1982). This censoring scheme can arise in the following study designs.

Design (A): At the start of the study all the n subjects are included, and the study period is c.

Design (B): The including time of each subject might be different, but each subject is observed at least period c.

It is important to specify how we should do when all the subjects happen to be censored. For designs (A) above, when we have no observed lifetime by the time c we may extend the study period to $c + c'$ so that at least one lifetime is observed. On the contrary, it is difficult for design (B) to make such extension. Design would be possible, but analysis may become intractably complicated.

In this paper we call the inference conditional that requires at least one observed lifetime, because the inference is conditioned on the number of observed data to be greater than or equal to one. On the other hand, inference that allows zero observed

* Biostatistics Division, Banyu Pharmaceutical Co. Ltd. AIG Kabutocho Building, Nihonbashi-Kabutocho, Chuo-ku, Tokyo 103-0026, Japan. takayuki-abe@merck.com

† Department of Computer and Information Science, Seikei University, 3-3-1 Kichijoji-Kitamachi, Musashino-shi, Tokyo 180-8633, Japan. iwasaki@st.seikei.ac.jp

frequency is called unconditional. Almost all statistical methods for the exponential mean developed so far are conditional in this sense.

Our aim of the present article is two-fold. First we give an expression for exact unconditional probability calculus under Type I censoring, which is an extension of the conditional formula given in Bartholomew (1963). This unconditional formula enables us to give exact P-values and exact confidence intervals for the exponential mean. These exact procedures are, however, too complicated to use in practical applications. Therefore, our second aim is to seek adequate approximation formulas for inference. The exact formula is effectively used to evaluate the preciseness of the approximate procedures. The exact unconditional formula will be given in Section 2. Section 3 discusses several approximations for P-value calculation and construction of confidence intervals. Numerical illustrations are shown in Section 4. Finally we conclude the discussion in Section 5.

2. Exact Inference

Let $X_1,\ldots,X_n$ be n mutually independent random variables each having the same exponential distribution with density function

$$f(x,\theta)=\begin{cases}\dfrac{1}{\theta}\exp\left[-\dfrac{x}{\theta}\right] & (x\geq 0)\\ 0 & (x<0),\end{cases}$$

where θ is an unknown mean lifetime to be estimated. This distribution will be denoted by $\mathrm{Exp}(\theta)$ in the following. It is supposed here that the lifetime will be observed if it is less than or equal to a pre-specified constant c, and will be censored if it is greater than c. This sampling scheme is called a Type I right censoring. We consider the case that m out of n lifetimes are observed and that $n-m$ remaining lifetimes are censored. In this case the number of observed data m is a random variable which may take values $0,1,\ldots,n$.

For $m\geq 1$, if we let $x_1,\ldots,x_m$ be m observed lifetimes, then the likelihood and the log-likelihood functions for θ based on all the observed and censored lifetimes become

$$L(\theta)=\frac{1}{\theta^m}\exp[-\frac{1}{\theta}\sum_{i=1}^{m}x_i]\exp[-\frac{(n-m)c}{\theta}] \tag{1}$$

and

$$l(\theta)=\log L(\theta)=-m\log\theta-\frac{1}{\theta}\left\{\sum_{i=1}^{m}x_i+(n-m)c\right\}, \tag{2}$$

respectively. We note that these expressions are also valid for $m=0$ if we let $\sum_{i=1}^{0}x_i=0$. For $m\geq 1$, the maximum likelihood (ML) estimator of θ is given by

$$\hat{\theta}=\frac{1}{m}\left\{\sum_{i=1}^{m}X_1+(n-m)c)\right\}, \tag{3}$$

which is the total observed time for all n individuals divided by the number of actually observed lifetimes m. If $m = 0$ the likelihood function (1) is monotone increasing in θ, and so does not possess a finite maximum, see, for example, Johnson and Kotz (1970) and Lawless (1982).

Our aim is to obtain an exact expression for the unconditional probability distribution of $\hat{\theta}$. In order to get the desired result we first give the conditional distribution.

Proposition 1 (Bartholomew, 1963)

The conditional distribution of the ML estimator (3) of θ is given by

$$\Pr(y \le \hat{\theta} \mid m \ge 1)$$

$$= \frac{1}{1-\exp[-nc/\theta]} \sum_{k=1}^{n} {}_nC_k \sum_{j=0}^{k} {}_kC_j (-1)^j \exp[-c(n-k+j)/\theta] \int_a^{\infty} f_{2k}(x)dx \qquad (4)$$

where

$$a = \frac{2k}{\theta} \max\{0, y - \frac{c}{k}(n-k+j)\}, \qquad (5)$$

and $f_{2k}(x)$ is the probability density function of a chi-squared distribution with $2k$ degrees of freedom.

As noted above, when $m = 0$ the likelihood function has no finite maximum, and hence the ML estimate is indefinite. In this case, however, if we let the ML estimate being larger than any finite number, then we can obtain the following unconditional distribution of the ML estimator.

Theorem 1

The unconditional distribution of the ML estimator (3) is given by

$$\Pr(y \le \hat{\theta})$$

$$\begin{aligned} &= \exp[-nc/\theta] + \sum_{k=1}^{n} {}_nC_k \sum_{j=0}^{k} {}_kC_j (-1)^j \exp[-c(n-k+j)/\theta] \int_a^{\infty} f_{2k}(x)dx, \\ &= \sum_{k=0}^{n} {}_nC_k \sum_{j=0}^{k} {}_kC_j (-1)^j \exp[-c(n-k+j)/\theta] \int_a^{\infty} f_{2k}(x)dx, \end{aligned} \qquad (6)$$

where the constant a is the same as (5), and for the density function $f_0(x)$ with zero degree of freedom we let $\int_{\alpha}^{\infty} f_0(x)dx = 1$.

Proof

For the unconditional distribution it holds the relationship

$$\Pr(y \le \hat{\theta}) = \Pr(y \le \hat{\theta} \mid m = 0)\Pr(m = 0) + \Pr(y \le \hat{\theta} \mid m \ge 1)\Pr(m \ge 1), \tag{7}$$

for any finite y. We note that

$$\Pr(m = 0) = \Pr(c \le X_1, \ldots, X_n) = \exp(-nc/\theta)$$

and

$$\Pr(m \ge 1) = 1 - \Pr(m = 0) = 1 - \exp(-nc/\theta).$$

When $m = 0$ the fact that the ML estimator $\hat{\theta}$ tends to infinity implies that $\Pr(y \le \hat{\theta} \mid m = 0) = 1$ for any finite y, and hence the first term of the right-hand side of (7) becomes $\exp(-nc/\theta)$. Substitutions of the expression (6) and $\Pr(m \ge 1) = 1 - \exp(-nc/\theta)$ into the second term of the right-hand side of (7) yield the required result after some reordering of the summation. (QED)

It is worth noting here that Bartholomew (1963) derived the formula (4) by inverting the moment generating function (mgf) of $\hat{\theta}$. Existence of the mgf necessarily requires all the moments to be finite. Therefore it was necessary in Bartholomew (1963) to assume that m is greater than 0. Theorem 1 indicates that essentially the same expression can be used for the unconditional case.

When our observed ML estimate is θ^*, we can calculate the unconditional exact P-value for testing $H_0 : \theta = \theta_0$ by using the formula $\Pr(\theta^* \le \hat{\theta} \mid \theta_0)$ given in Theorem 1. Confidence intervals for θ are obtained by inverting this test to find the set of values θ which are not rejected by the test based on the observed estimate θ^*. Specifically, the lower limit θ_L and the upper limit θ_U of the $100(1-\alpha)$ confidence interval will be given as the values that satisfy $\Pr(\theta^* \le \hat{\theta} \mid \theta_L) = \alpha/2$ and $\Pr(\theta^* \ge \hat{\theta} \mid \theta_U) = \alpha/2$, respectively. In order to get such limits, however, we need time-consuming trial and error computation.

When $m = 0$, that is all lifetimes happen to be greater than c and censored, the ML estimate tends to infinity. However, we can calculate one-sided P-value for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$ as P-value$= \Pr(m = 0 \mid \theta_0) = \exp(-nc/\theta_0)$. The lower limit θ_L of a one-sided confidence interval (θ_L, ∞) for θ can also be obtained by $\theta_L = -nc/\log(\alpha/2)$, which is derived from inverting the equation $\exp(-nc/\theta_L) = \alpha/2$. Although it might be arguable here whether we should use $\alpha/2$ or α in this "one-sided" confidence interval, we choose $\alpha/2$ in accordance with the $m \ge 1$ cases. For related discussions in this respect, see also Iwasaki and Hidaka (2001) and the references therein.

3. Approximate Inference

The exact results obtained in the previous section are theoretically important but too time-consuming for practical use. Therefore we need to develop good approximate

methods that can be used in practical applications. First, we consider some approximations to the distribution of $\hat{\theta}$. It is argued in Section 3.2 of Lawless (1982) that the procedure based on the large-sample normal approximation to the ML estimator performs poor for small and moderate sample sizes, and hence it will not be discussed further in this paper. We will deal with two chi-squared approximations which are discussed in Lawless (1982).

From (1) and (3) the likelihood ratio for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ becomes

$$\lambda_1 = \frac{L(\theta_0)}{L(\hat{\theta})} = \left(\frac{\hat{\theta}}{\theta_0}\right)^m \exp\left[-m\left(\frac{\hat{\theta}}{\theta_0} - 1\right)\right]. \tag{8}$$

Hence, the quantity

$$-2\log\lambda_1 = -2m\left\{\log\left(\frac{\hat{\theta}}{\theta_0}\right) - \frac{\hat{\theta}}{\theta_0} + 1\right\} \tag{9}$$

approximately follows a chi-squared distribution with one degree of freedom, which will be denoted by χ^2_1 for short. If we let y^* be the observed value of (9) then the two-sided P-value for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ can be easily calculated as

$$P - value = \text{CHIDIST}(y^*, 1), \tag{10}$$

where CHIDIST is the built-in function of MS EXCEL.

Different from the P-value calculation, we need more computation to obtain a confidence interval for θ. Let $\chi^2_1(\alpha)$ be the upper 100α point of the χ^2_1 distribution. For example $\chi^2_1(0.05) \approx 3.84$. Then, solving the equation

$$-2m\left\{\log\left(\frac{\hat{\theta}}{\theta_0}\right) - \frac{\hat{\theta}}{\theta_0} + 1\right\} = \chi^2_1(\alpha)$$

for θ_0 we have two expressions

$$\theta_0 = \hat{\theta} / \left(\frac{\chi^2_1(\alpha)}{2m} + 1 + \log\hat{\theta} - \log\theta_0\right)$$

and

$$\theta_0 = \exp\left[\frac{\chi^2_1(\alpha)}{2m} + \log\hat{\theta} - \frac{\hat{\theta}}{\theta_0} + 1\right].$$

These expressions can be used for iterative calculations. Specifically, starting from an appropriate initial value $\theta^{(0)}$, we obtain an iterative algorithm

$$\theta^{(t+1)} = \hat{\theta} / \left(\frac{\chi^2_1(\alpha)}{2m} + 1 + \log\hat{\theta} - \log\theta^{(t)}\right) \tag{11}$$

and

$$\theta^{(t+1)} = \exp\left[\frac{\chi_1^2(\alpha)}{2m} + \log\hat{\theta} - \frac{\hat{\theta}}{\theta^t} + 1\right]. \tag{12}$$

For details of such functional algorithms, see, for example, Lange (1999) and Iwasaki (2001). From (11) the lower limit θ_L of the $100(1-\alpha)$ confidence interval is obtained, whereas (12) yields the upper limit θ_U. These limits can be easily calculated by EXCEL. Lawless (1982) calculated the same interval in his example by trial and error. The simple iterative algorithms (11) and (12) are preferable to such time-consuming ad hoc trial-and-error computation.

Another approximation makes use of the fact that $2m\hat{\theta}/\theta$ approximately distributes as a x^2_{2m+1} distribution. This result is sometimes attributed to Cox (1953), see Section 3.2.1 of Lawless (1982) and also Regal (1980). This approximation implies another approximate $100(1-\alpha)$ confidence interval given by

$$\left(\frac{2m\hat{\theta}}{\chi^2_{2m+1}(\alpha/2)}, \frac{2m\hat{\theta}}{\chi^2_{2m+1}(1-\alpha/2)}\right), \tag{13}$$

where $\chi^2_{2m+1}(\alpha/2)$ and $\chi^2_{2m+1}(1-\alpha/2)$ are the upper and lower $100\alpha/2$ points of the χ^2_{2m+1} distribution, respectively. It should be noted that the interval (13) can be obtained without any iterative computation. This χ^2_{2m+1} approximation practically works well as will be shown in the following example and also in the numerical illustrations in Section 4. However, this approximation has the following logical difficulty. As was noted above, the number of observed lifetimes m is a random variable in Type I censoring scheme. In fact, the exact probability calculus (6) sums up the probabilities for all possible values of m. In the χ^2_{2m+1} approximation, however, m is treated as a fixed value. This is contradicted to the Neyman-Pearson principle that the probability calculus should be performed for all possible values of the random variables considered.

Example 2.1 (Lawless (1982) Example 3.3.3)

In a clinical trial to investigate the duration of remission achieved by a drug used in the treatment of leukemia, 20 patients were recruited in which 10 remission times were observed and 10 patients were Type I censored. The sum of remission and censoring times was 700 weeks, and which gives the ML estimate $\hat{\theta} = 70$ weeks. Next we will obtain a 95 confidence interval for the mean remission time θ. Since the censoring point c is not explicitly specified here, the exact probability calculus (6) is not possible, whereas chi-squared approximations can be applied. The likelihood- ratio approximation procedures (11) and (12) converge to $\theta_L = 39.91$ and $\theta_U = 139.70$, respectively, both from starting the same initial value $\theta^{(0)} = 70$. The χ^2_{2m+1} approximation (13) gives $\theta_L = 39.46$ and $\theta_U = 136.15$. These estimates are close to each other.

4. Numerical illustration

For illustration, we will give numerical results using an artificial but typical dataset. Although only a particular numerical result is shown here, we have obtained but not shown here almost the same results in our extensive computations for various settings

of similar sample sizes. We will give exact results that use Theorem 1 and also some approximations.

Let $n = 10$, and suppose that the complete dataset is given by Table 1. We set three censoring time points as $c = 0.3$ (CASE-1), 1.5 (CASE-2) and 3.0 (CASE-3). ML estimates of θ and the exact one-sided P-values for testing

$$H_0 : \theta = 0.6 \text{ vs. } H_1\theta > 0.6$$

are given in Table 2, in which two P-values for the case C (conditional on $m \geq 1$, Proposition 1) and U (unconditional, Theorem 1) are shown. Furthermore, several 95 confidence intervals calculated from four procedures of the previous section are given in Table 3. Figure 1 shows the likelihood and log-likelihood functions of θ for Case-2 ($c = 1.5$).

We observe in Table 2 and Table 3 that when m is not so small numerical values calculated by the conditional and the unconditional methods are almost the same. If m is very small such as CASE-1 there exist some differences between the two distributions, although it would be negligible in actual applications.

For two approximations considered here, the likelihood-ratio method based on the chi-squared distribution of one degree of freedom provides better result in the sense that the confidence intervals are closer to the exact unconditional counterparts. The approximation based on $2m + 1$ degrees of freedom is also attractive because it requires no iterative computation to get the confidence intervals, although it has logical difficulty given in the previous section.

Table 1: Complete Lifetime Data

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
0.02	0.17	0.29	0.38	0.48	1.24	1.30	1.36	1.67	2.66

Table 2: ML Estimates and P-values for Testing $H_0 : \theta = 0.6$

	c	m	MLE	P-value(C)	P-value(U)
CASE-1	0.3	3	0.860	0.24560	0.25068
CASE-2	1.5	8	1.030	0.04526	0.04526
CASE-3	3.0	10	0.957	0.04864	0.04864

Table 3: Several 95% Confidence Intervals

CASE-1	Lower	Upper
Exact(C)	0.33199	4.92522
Exact(U)	0.33172	3.65668
LR	0.33165	3.45814
ChiSq(7)	0.32224	3.05350

CASE-2	Lower	Upper
Exact(C)	0.55453	2.32875
Exact(U)	0.55453	2.32801
LR	0.55333	2.25391
ChiSq(15)	0.54586	2.17869

CASE-3	Lower	Upper
Exact(C)	0.55291	2.00634
Exact(U)	0.55291	2.00634
LR	0.54563	1.90989
ChiSq(21)	0.53948	1.86134

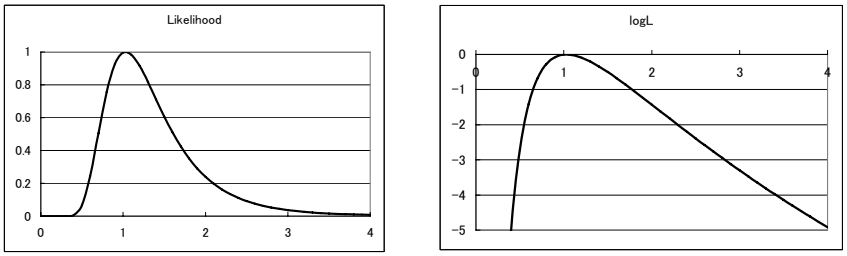

Figure 1. Likelihood and Log-Likelihood Functions for CASE-2 (c=1.5)

5. Discussion

We observed in the example of Section 3 and in the numerical illustrations of Section 4 that conditional and unconditional inferences gave almost the same results for moderate m. When m is very small such as three or less, exact and approximate P-values and confidence intervals can be different. In such cases the exact unconditional method should be used.

It can be concluded from the argument of preceding sections that the most promising inferential strategy for the exponential mean from Type I censored data is as follows:

A point estimate of θ can be obtained from (3) when $m \geq 1$. When $m = 0$ we have no sensible point estimate. The P-value calculus for hypothesis testing we can use either the exact method (6) or the approximation (10). Lower and upper limits of a confidence interval can be obtained from the iterative method (11) and (12), respectively. The approximation (13) also gives numerically sensible results.

A SAS program for the exact probability calculus (6) will be provided from the authors upon request.

Acknowledgement

An early version of this article was orally presented at the joint annual meeting of the Biometric Society of Japan and the Japanese Society of Applied Statistics held in 2001. We thank Professor Takashi Yanagawa for his valuable comments given to our talk at that time. We also thank the referees for their valuable comments which are very helpful to revise the manuscript.

This research was partially supported by Grant-in-Aid for Scientific Research No. 16200022.

References

Bartholomew, D. J. (1963). The sampling distribution of an estimate arising in life testing, *Technometrics*, **5**, 361–374.

Cox, D. R. (1953). Some simple approximate tests for Poisson variates, *Biometrika*, **40**, 354–360.

Iwasaki, M. (2001). A simple iterative method for solving non-linear equations and its statistical applications, *Japanese Journal of Applied Statistics*, **30**, 107–118 (in Japanese).

Iwasaki, M. and Hidaka, N. (2001). Notes on the central and shortest confidence intervals for a binomial parameter, *Japanese Journal of Biometrics*, **22**, 1–13.

Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions*-1, John Wiley & Sons, New York.

Lange, K. (1999). *Numerical Analysis for Statisticians*, Springer, New York.

Lawless, J. F. (1982). *Statistical Models & Methods for Lifetime Data*, John Wiley & Sons, New York.

Regal, R. (1980). The F test with time-censored exponential data, *Biometrika*, **67**, 479–481.

Received October 14, 2003
Revised May 8, 2004