INFORMATION
Volume 12,Number 1, pp.33-50

ISSN 1343-4500

Estimation for the size of fragile population in the trunsored and truncated models with application to the confidence interval for the case fatality ratio of SARS

Hideo Hirose *

Kyushu Institute of Technology, Department of Systems Design and Informatics, Iizuka, Fukuoka, 820-8502 Japan

Abstract

A method to obtain the estimates for parameters and the size of fragile population with their confidence intervals in mixed populations of the fragile and durable samples, i.e., in the trunsored model, along with those in the truncated model, is introduced. The confidence intervals for the estimates in the trunsored model are compared with those in the truncated model. The maximum likelihood estimates for the parameters in the underlying probability distribution in both models are exactly the same when all the samples have the same censoring time, and consequently the confidence interval for the parameters are also the same. The estimate for the number of fragile samples in the trunsored model is the same as that in the truncated model when the failure data are the same; however, the confidence interval for it in the trunsored model differs from that in the truncated model. In the truncated model, the confidence interval for the fragile samples is affected by the fluctuation effect due to the censoring time and the parameters in the underlying probability distribution. In the trunsored model, however, the confidence interval is affected by two kinds of fluctuation effect: one is the same as in the truncated model, and the other is the extra parameter which corresponds to the ratio of the number of fragile samples to the total number of samples. When the censoring time becomes large, the width of the confidence interval in the truncated model tends to zero, whereas the confidence interval in the trunsored model tends to a positive constant value, which is corresponding to the binomial case.

A typical example of the method applied to the case fatality ratio for the infectious diseases such as SARS shows different confidence intervals between the trunsored model and the truncated model. Using the truncated model we may have the paradoxical case fatality ratio; using the trunsored model, however, we can obtain the reasonable estimate for it. This indicates that we have to be cautious in selecting the appropriate model when we deal with the incomplete data models.

Key words: trunsored data, truncated data, fragile, durable, confidence interval,

case fatality ratio, SARS.

1 Introduction

Many researchers have been tackled the problems for obtaining the confidence intervals of the estimates in the underlying probability distributions provided that the observed data are complete or incomplete. However, the confidence intervals for the size of population, provided that the observed data are incomplete, have not been dealt with so much so far. This paper treats this problem, and reveals some important consequences.

Here is a typical formula in estimating the population size N using the truncated model when all the samples have the same censoring time T and r is the number of observed (failure) cases by time T, such that

$$\hat{N} = r/\hat{F}(T), \tag{1}$$

where $F(t)$ denotes the underlying probability distribution. Let's see what happens if we apply this formula to the estimation for the case fatality ratio of infectious diseases.

The case fatality ratio p_f could be estimated by

$$p_f = M_2/M_1, \tag{2}$$

if we observed the total number of infected cases M_1 and the total number of deaths M_2 simultaneously. When M_i $(i = 1, 2)$ are not fully observed, these values may be replaced by the estimated values of $\hat{N}_i$ using (1). Here, M_i is replaced by $\hat{N}_i$ for clarity. We suppose that r_1 is the number of observed infected cases by time T and r_2 is the number of observed deaths by time T; $F_1(t)$ denotes the underlying probability distribution for the infected time and $F_2(t)$ denotes the underlying probability distribution for the time of death observed. When the censoring time T is large enough, $\hat{N}_1$ and $\hat{N}_2$ go to the ultimate observed infected and fatal cases, i.e., $\hat{N}_i \to M_i$ $(T \to \infty)$, and their standard errors for $\hat{N}_1$ and $\hat{N}_2$ will become close to zeros if we adopt the truncated model, as will be shown later. Consequently, the width of the confidence interval for the case fatality ratio will also become close to zero.

* Corresponding author.
Email address: hirose@ces.kyutech.ac.jp (Hideo Hirose).

The WHO revealed the daily number of infected cases, fatal cases, and cured cases of SARS from March 17, 2003 to December 31, 2003, to the public; see WHO (2003). In Hong Kong, if we use the truncated model, the case fatality ratio would become just 299/1755, having zero width confidence interval, provided that no new patients, deaths, and recoveries were observed after December 31, 2003; similarly in Taiwan, just 37/346 would be expected; in Singapore, only 33/238; and in Canada, a mere 43/251. It would, however, be reasonable that the case fatality ratio of SARS is assumed to be a probabilistic value. Then, it becomes paradoxical that there are various estimates having zero width confidence intervals. This is caused by adopting the truncated model in estimating the case fatality ratio. In this paper, we show that a reasonable width of the confidence interval for the case fatality ratio can be obtained by using the trunsored model.

2 The trunsored and truncated models

In many lifetime estimation problems, it is usually assumed that the underlying probability distribution is a single homogeneous population, and that all samples will eventually fail (or die). When data are incomplete, we estimate the parameters that determine the characteristics of the population in parametric models by regarding the data as censored or truncated. However, we possibly encounter the cases that some portion of the samples will fail by certain time whereas the rest of them still continue to work (or survive). In such a case we may assume heterogeneous populations.

If the left endpoint, T_0, of the underlying distribution is extremely large, the failures will not be observed at all within the prescribed time T ($<< T_0$) even if T is relatively large. Such a population is defined as *durable* population, whereas the failures may or may not be observed within time T are defined as *fragile* population; see Hirose (2005). Here, we suppose that we are dealing with an incomplete data case problem with the following conditions: 1) the fragile and durable populations may be mixed, and 2) the population size of the mixed populations, n, is known. We define the fragile and durable probability distributions by $F(t;\theta)$ and $G(t;\phi)$ respectively, and also define their linear combination by

$$H(t;\psi) = sF(t;\theta) + (1-s)G(t;\phi), \quad (t \geq 0), \tag{3}$$

with a combination parameter s; θ and ϕ are the parameters in the underlying probability distribution functions. Since $G(t) = 0$ $(t < T_0)$, the likelihood function, $L(\psi)$, for this model becomes $L(\psi) \to L_{ts}(\theta, s)$ when T is large,

where

$$L_{ts}(\theta, s) = \{1 - sF(T;\theta)\}^{n-r} \cdot \prod_{i=1}^{r} \{sf(t_i;\theta)\}, \tag{4}$$

under the condition that the number of the censoring times is just one and the samples are right censored. Here, t_i is the observed failure time; r is the number of observed failures; $f(t)$ is the density function. Hirose calls this model the *trunsored* model (Hirose (2005)); this kind of the model is often used in many fields; see, e.g., Boag (1948), Maltz and MaCleary (1977), and Meeker (1987). Our primary interest here is to estimate the number of fragile samples, m, and its standard error, $se(m)$.

In the truncated model, the likelihood function $L_t(\theta)$ is expressed by

$$L_t(\theta) = \prod_{i=1}^{r} \{f(t_i;\theta)/F(T;\theta)\}. \tag{5}$$

In Hirose (2005), the estimation for parameter θ is mainly discussed; however, we will deal with the estimation for m, the number of fragile samples, in this paper.

3 Errors for the parameters in the trunsored and truncated models

3.1 In the trunsored model

The likelihood equations corresponding to (4) are

$$\frac{\partial \log L_{ts}}{\partial s} = (n-r)\frac{\partial \log\{1 - sF(T)\}}{\partial s} + \sum_{i=1}^{r} \frac{\partial \log\{sf(t_i)\}}{\partial s} = 0, \tag{6}$$

$$\frac{\partial \log L_{ts}}{\partial \theta} = (n-r)\frac{\partial \log\{1 - sF(T)\}}{\partial \theta} + \sum_{i=1}^{r} \frac{\partial \log\{sf(t_i)\}}{\partial \theta} = 0. \tag{7}$$

From (6), we have

$$\hat{s} = \frac{r}{n\hat{F}(T)}, \quad \text{or} \quad r = n\hat{s}\hat{F}(T). \tag{8}$$

By substituting (8) into (7), we have the equation,

$$\frac{\partial \log L_{ts}}{\partial \theta} = -\frac{r}{F(T)}\frac{\partial F(T)}{\partial \theta} + \sum_{i=1}^{r}\frac{\partial \log f(t_i)}{\partial \theta} = 0, \tag{9}$$

which should have the same solution $\hat{\theta}$ as that in the likelihood equation for the truncated model,

$$\frac{\partial \log L_t}{\partial \theta} = -r\frac{\partial \log F(T)}{\partial \theta} + \sum_{i=1}^{r}\frac{\partial \log f(t_i)}{\partial \theta} = 0. \tag{10}$$

This fact also proves that the width of the confidence interval for θ in the trunsored model and that in the truncated model are exactly the same. This is obvious; imagine that any other failure data case $\{t_1, t_2, \ldots, t_k, \ldots\}$ gives exactly the same solution in the trunsored and truncated models; the bootstrap method will provide this situation.

In the trunsored model, we have parameter s which does not exist in the truncated model. To obtain the confidence intervals for s and θ, we compute the Fisher information matrix. For simplicity, we first assume that θ is scalar, but this restriction could easily be extended to vector cases. The elements for the observed Fisher information matrix are corresponding to

$$\frac{\partial \log L_{ts}}{\partial s^2} = -\frac{r}{s^2} - \frac{(n-r)F(T)^2}{\{1-sF(T)\}^2} \tag{11}$$

$$\frac{\partial^2 \log L_{ts}}{\partial s \partial \theta} = \frac{\partial^2 \log L_{ts}}{\partial \theta \partial s} = -(n-r) \times \frac{F_\theta(T)}{(1-sF(T))^2}, \tag{12}$$

$$\frac{\partial^2 \log L_{ts}}{\partial \theta^2} = \times \frac{sF_\theta(T)^2 + \{1-sF(T)\}F_{\theta\theta}(T)}{(1-sF(T))^2} - \sum_{i=1}^{r}\frac{f_\theta(t_i)^2 - f(t_i)f_{\theta\theta}(t_i)}{f(t_i)^2}. \tag{13}$$

By using $E[r] = nsF(T)$, (11)-(13) become

$$E[\frac{\partial^2 \log L_{ts}}{\partial s^2}] = -n\frac{F(T)}{s\{1-sF(T)\}}, \tag{14}$$

$$E[\frac{\partial^2 \log L_{ts}}{\partial s \partial \theta}] = E[\frac{\partial^2 \log L_{ts}}{\partial \theta \partial s}] = -n\frac{F_\theta(T)}{1-sF(T)}, \tag{15}$$

$$E\ [\ \frac{\partial^2 \log L_{ts}}{\partial \theta^2}] = -ns\frac{sF_\theta(T)^2 + \{1 - sF(T)\}F_{\theta\theta}(T)}{1 - sF(T)}$$
$$-n\int_0^T \frac{f_\theta(t)^2 - f(t)f_{\theta\theta}(t)}{f(t)^2} sf(t)dt. \tag{16}$$

The approximate variance for s and θ can be given by the diagonal elements of I_n^{-1}, where

$$I_n = -\begin{pmatrix} E[\frac{\partial^2 \log L_{ts}}{\partial s^2}] & E[\frac{\partial^2 \log L_{ts}}{\partial s \partial \theta}] \\ E[\frac{\partial^2 \log L_{ts}}{\partial \theta \partial s}] & E[\frac{\partial^2 \log L_{ts}}{\partial \theta^2}] \end{pmatrix} \tag{17}$$

3.1.1 When $T \to \infty$

We next consider the case when $T \to \infty$ in addition to $T << T_0$. In (14)-(16), the expectations of the second derivatives will converge to

$$E[\frac{\partial^2 \log L_{ts}}{\partial s^2}] \to -n\frac{1}{s\{1-s\}}, \tag{18}$$

$$E[\frac{\partial^2 \log L_{ts}}{\partial s \partial \theta}] = E[\frac{\partial^2 \log L_{ts}}{\partial \theta \partial s}] \to 0, \tag{19}$$

$$E\,[\frac{\partial^2 \log L_{ts}}{\partial \theta^2}] \to -ns\int_0^\infty \frac{f_\theta(t)^2 - f(t)f_{\theta\theta}(t)}{f(t)^2} f(t)dt. \tag{20}$$

In (19) and (20), we used that $F_\theta(T) \to 0$ and $F_{\theta\theta}(T) \to 0$ since $F(T) \to 1$. Taking into account that

$$\int_0^\infty \frac{f_\theta(t)^2 - f(t)f_{\theta\theta}(t)}{f(t)^2} f(t)dt \approx 1/(nVar(\theta_c)), \tag{21}$$

where $Var(\theta_c)$ is the asymptotic variance for θ in the complete data case with n samples, we have

$$I_n \to n\begin{pmatrix} 1/(s(1-s)) & 0 \\ 0 & s/(nVar(\theta_c)) \end{pmatrix}, \tag{22}$$

and its inverse matrix of

$$I_n^{-1} \to \begin{pmatrix} s(1-s)/n & 0 \\ 0 & Var(\theta_c)/s \end{pmatrix}. \tag{23}$$

Thus, $Var(s)$ goes to $s(1-s)/n$ and $Var(\theta)$ goes to $Var(\theta_c)/s$. Considering that the number of fragile samples is ns, the latter is obvious. We can see that the standard error error for s is the same as that in the binomial distribution. When s is close to zero, $Var(s)$ is simply obtained by

$$Var(s) \approx \sqrt{s/n}. \tag{24}$$

On the contrary, when s is close to one,

$$Var(s) \approx \sqrt{(1-s)/n}. \tag{25}$$

When $s \to 1$, which may be interpreted as the truncated case,

$$Var(s) \to 0. \tag{26}$$

3.1.2 When the underlying probability distribution is exponential

Let's take a look at a more specific case for simplicity to find the difference between the trunsored model and the truncated model. We imagine a model where the underlying probability distribution is exponential, although we know that the actual cases such as SARS are not so simply described. When $F(t)$ is the exponential distribution function with parameter λ,

$$E[\frac{\partial^2 \log L_{ts}}{\partial s^2}] = -n\frac{1-e^{-\lambda T}}{s\{1-s(1-e^{-\lambda T})\}}, \tag{27}$$

$$E[\frac{\partial^2 \log L_{ts}}{\partial s \partial \lambda}] = E[\frac{\partial^2 \log L_{ts}}{\partial \lambda \partial s}] = -n\frac{Te^{-\lambda T}}{1-s(1-e^{-\lambda T})}, \tag{28}$$

$$E[\frac{\partial^2 \log L_{ts}}{\partial \lambda^2}]$$
$$= -ns\frac{s(Te^{-\lambda T})^2 - \{1-s(1-e^{-\lambda T})\}T^2e^{-\lambda T}}{1-s(1-e^{-\lambda T})}$$

$$-n\int_0^T \frac{\{(1-\lambda t)e^{-\lambda t}\}^2 - \lambda e^{-\lambda t}\{t(\lambda t-2)e^{-\lambda t})\}}{(\lambda e^{-\lambda t})^2} \times s\lambda e^{-\lambda t}dt$$
$$= -ns\Big(\frac{(-1+s)T^2e^{-\lambda T}}{1-s(1-e^{-\lambda T})} + \frac{(1-e^{-\lambda T})}{\lambda^2}\Big). \tag{29}$$

More specifically, when $\lambda = 1, s = 1/2$, then,

$$I_n^{-1} \approx \frac{1}{n}\begin{pmatrix} 0.53361 & -1.312 \\ -1.312 & 8.38244 \end{pmatrix}, \quad (T=2),$$
$$I_n^{-1} \approx \frac{1}{n}\begin{pmatrix} 0.250023 & -0.0004561 \\ -0.0004561 & 2.00921 \end{pmatrix},$$
$$(T=10),$$
$$I_n^{-1} \approx \frac{1}{n}\begin{pmatrix} 0.25 & -3.720\times 10^{-42} \\ -3.720\times 10^{-42} & 2 \end{pmatrix},$$
$$(T=100). \tag{30}$$

Then, the standard errors for s and λ are

$$se(s) \approx 0.730486/\sqrt{n}, \quad se(\lambda) \approx 2.89524/\sqrt{n},$$
$$(T=2),$$
$$se(s) \approx 0.500023/\sqrt{n}, \quad se(\lambda) \approx 1.41747/\sqrt{n},$$
$$(T=10),$$
$$se(s) \approx 0.5/\sqrt{n}, \quad se(\lambda) \approx 1.41421/\sqrt{n},$$
$$(T=100). \tag{31}$$

3.2 In the truncated model

In the truncated model, the approximate variance for θ can be given by the inverse of $-E[\partial^2 \log L_t/\partial\theta^2]$. From (5), the second derivative of $\log L_t$ with respect to θ is

$$\frac{\partial^2 \log L_t}{\partial\theta^2} = r\frac{F_\theta(T)^2 - F(T)F_{\theta\theta}(T)}{F(T)^2} - \sum_{i=1}^{r}\frac{f_\theta(t_i)^2 - f(t_i)f_{\theta\theta}(t_i)}{f(t_i)^2}, \tag{32}$$

and its expectation can be computed by

$$E[\frac{\partial^2 \log L_t}{\partial \theta^2}] = E[N]\Big(\frac{F_\theta(T)^2 - F(T)F_{\theta\theta}(T)}{F(T)} - \int_0^T \frac{f_\theta(t)^2 - f(t)f_{\theta\theta}(t)}{f(t)^2} f(t)dt\Big), \tag{33}$$

where $E[N]$ denotes the expected size of population in the truncated model.

3.2.1 When $T \to \infty$

When $T \to \infty$, the expectation of the second derivative in (33), will converge to a constant value,

$$\begin{aligned} & E[\frac{\partial^2 \log L_t}{\partial \theta^2}] \\ \to & -E[N] \int_0^\infty \frac{f_\theta(t)^2 - f(t)f_{\theta\theta}(t)}{f(t)^2} f(t)dt \\ \approx & -1/Var(\theta_c), \end{aligned} \tag{34}$$

which is corresponding to the value in the complete data case.

3.2.2 When the underlying probability distribution is exponential

In the specific exponential model,

$$E[\frac{\partial^2 \log L_t}{\partial \lambda^2}] = E[N]\Big(\frac{T^2 e^{-\lambda T}}{1 - e^{-\lambda T}} - \frac{1 - e^{-\lambda T}}{\lambda^2}\Big). \tag{35}$$

More specifically ,when $\lambda = 1, s = 1/2$,

$$\begin{aligned} E[\frac{\partial^2 \log L_t}{\partial \lambda^2}] &\approx -0.238594E[N], \quad (T = 2), \\ E[\frac{\partial^2 \log L_t}{\partial \lambda^2}] &\approx -0.995414E[N], \quad (T = 10), \\ E[\frac{\partial^2 \log L_t}{\partial \lambda^2}] &\approx -E[N], \quad (T = 100), \end{aligned} \tag{36}$$

and consequently,

$$
\begin{aligned}
se(\lambda) &\approx 2.04725/\sqrt{E[N]}, \quad (T = 2), \\
se(\lambda) &\approx 1.00230/\sqrt{E[N]}, \quad (T = 10), \\
se(\lambda) &\approx 1/\sqrt{E[N]}, \quad (T = 100).
\end{aligned} \tag{37}
$$

Considering that $s = 1/2$, it is expected that $n = 2E[N]$, and we can see that $se(\lambda)$ in the trunsored model in this specific case is the same as that in the truncated model as mentioned earlier in more general cases, because $2.89524/\sqrt{2} \approx 2.04725$, for example.

4 Errors for the size of fragile population in the trunsored and truncated models

4.1 In the trunsored model

In the trunsored model, the size of fragile population is estimated by

$$\hat{m} = n\hat{s}, \tag{38}$$

thus, the error for $\hat{m}$ is obtained by

$$se(\hat{m}) = n \times se(\hat{s}). \tag{39}$$

4.1.1 When $T \to \infty$

As for the standard error for the size of fragile population, when $T \to \infty$,

$$
\begin{aligned}
se(m) = n \times se(s) &\to n\sqrt{s(1-s)/n} \\
&= \sqrt{n}\sqrt{s(1-s)}.
\end{aligned} \tag{40}
$$

This standard error is the same as that in the binomial distribution for the sample size, when $T \to \infty$.

When s is close to zero, $se(m)$ is simply obtained by

$$se(m) \approx \sqrt{ns} = \sqrt{m}. \tag{41}$$

On the contrary, when s is close to one,

$$se(m) \approx \sqrt{n(1-s)}. \tag{42}$$

When $s \to 1$, which may be interpreted as the truncated case,

$$m \to n \ (\approx N), \quad se(m) \to 0. \tag{43}$$

4.1.2 When the underlying probability distribution is exponential

When $F(t)$ is the exponential distribution function with parameter $\lambda = 1$, and $s = 1/2$, then, $E[se(m)]$ is obtained as shown in Table 1 from (31).

4.2 In the truncated model

In the truncated model, the size of the fragile population is naturally estimated by

$$\hat{N} = r/\hat{F}(T). \tag{44}$$

The variance for $\hat{N}$ is obtained by using

$$\begin{aligned} &Var(r/F(T)) = \\ &r^2 \sum_{i,j} \frac{\partial}{\partial \theta_i}\Big(\frac{1}{F(T)}\Big) \cdot \frac{\partial}{\partial \theta_j}\Big(\frac{1}{F(T)}\Big) \\ &\cdot Cov(\theta_i, \theta_j). \end{aligned} \tag{45}$$

4.2.1 When $T \to \infty$

When $T \to \infty$, considering that $\partial(1/F(T))/\partial\theta_i \to 0$ since $F(T) \to 1$ $(T \to \infty)$, the variance for $E[N]$ is finally

$$Var(E[N]) \to 0, \ (T \to \infty), \tag{46}$$

by (44) and (45).

Table 1
Comparison of the approximate standard errors for the size of the fragile population.

	trunsored model	truncated model
	$E[se(m)]$	$E[se(N)]$
$T=2$	$0.730486\sqrt{n}$	$0.741166\sqrt{E[N]}$
$T=10$	$0.500023\sqrt{n}$	$0.000455085\sqrt{E[N]}$
$T=100$	$0.5\sqrt{n}$	$3.72008\times10^{-42}\sqrt{E[N]}$
$T\to\infty$	$0.5\sqrt{n}$	0

$F(t)=1-\exp(-t/\lambda),\quad(\lambda=1,\ s=1/2)$

4.2.2 When the underlying probability distribution is exponential

When $F(t)$ is the exponential distribution function with parameter λ,

$$E[\frac{\partial}{\partial\lambda}(\frac{1}{F(T)})]=\frac{-Te^{-\lambda T}}{(1-e^{-\lambda T})^2}, \tag{47}$$

then, the standard error for N is obtained as shown in Table 1. The tendency of $E[se(N)]$ shrinking for large T is totally different from the tendency of $E[se(m)]$ becoming a positive constant value as shown in the table.

4.3 The difference of the errors between the two models

We have shown that the error for the size of the fragile population in the trunsored model is different from that in the truncated model. In the truncated model, the error is affected by the censoring time T and $\hat{\theta}$, that is, the fluctuation of the parameters in the probability distribution. However, in the trunsored model, the error is affected by the two kinds of fluctuation effect: one is the same as in the truncated model, and the other is $\hat{s}$; in Fig.1, type A fluctuation due to the probability distribution and type B fluctuation due to $\hat{s}$ is illustrated. In the truncated model, since the fluctuation depends only on $\hat{\theta}$, the error by this effect will vanish when $T\to\infty$. In the trunsored model, however, the fluctuation effect by $\hat{s}$ remains even if $T\to\infty$, although the fluctuation effect by $\hat{\theta}$ vanishes. This means that the two kinds of fluctuation effect due to θ and s will be reduced to only one fluctuation effect due to s when $T\to\infty$ in the trunsored model. We should be cautious in selecting the appropriate model between the truncated and trunsored models. The next example illustrates the discrepancy between the two models clearly.

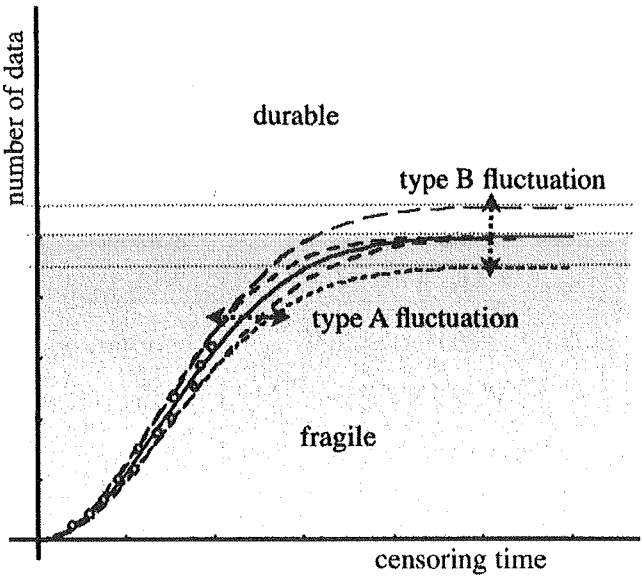

Fig. 1. Two kinds of fluctuation effect in the trunsored model.

5 Example

To show the difference between the truncated and trunsored models, 20 simulated data, which are generated by an exponential distribution with scale parameter $\lambda = 1$, are used as a typical example. The generated data are shown in Table 2.

The maximum likelihood estimate for the complete data model is $\hat{\lambda} = 0.84495$ and its approximate standard error is $se(\lambda) = 0.18894$ using the observed Fisher information. Fig.2 shows the estimates and their errors using the observed Fisher information (matrix) in the censored, truncated, and trunsored models; the horizontal axes mean the censoring time. In the trunsored model, $s = 1/100$ is assumed here.

We can see that 1) the estimates for λ both in the truncated and trunsored models are exactly the same, 2) the estimates for λ are all converged to the estimate in the complete data model as the censoring time becomes large, 3) the errors in the truncated and trunsored models are substantially larger than the error in the censored model when the censoring time is not so large, 4) the errors for λ are all converged to the error in the complete data model as the censoring time becomes large, 5) the estimates for the size of fragile population are exactly the same and they converge to the ultimate value as the censoring time becomes large, however, 6) the error for the size of fragile population in the truncated model quickly converges to zero, whereas that in the trunsored model converges to a positive constant value of $\sqrt{20}$ as mentioned in 4.1.

Table 2
Simulated random samples in the exponential distribution.

0.01	0.03	0.1	0.16	0.26
0.5	0.53	0.56	0.71	0.92
0.96	1.06	1.34	1.54	1.62
1.77	1.95	2.41	3.11	4.13

$F(t) = 1 - \exp(-t/\lambda), \quad (\lambda = 1)$

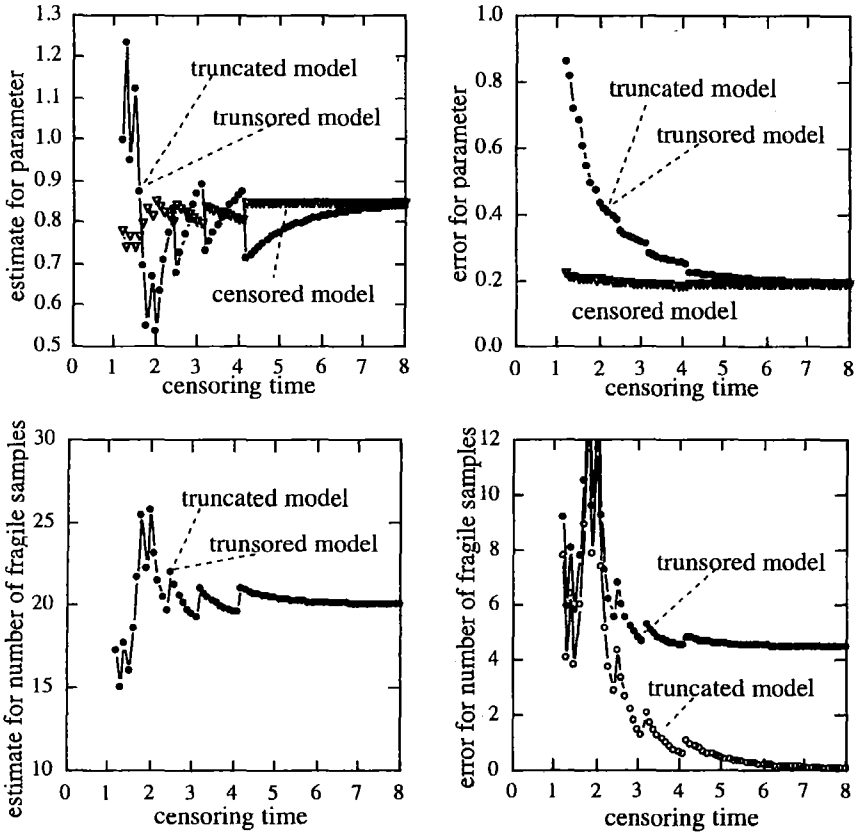

Fig. 2. Estimates and their errors in an example.

6 Simulation study

Assuming that $F(t)$ is the exponential distribution function with parameter $\lambda = 1$ and the censoring time $T = 2$. By a simulation study with $n = 20000$, $s = 1/2$ (thus, $m = 10000$), $N = 10000$, and the number of trial cases are all 1000, Then, mean, μ, and standard deviation, sd, for λ, s, and m, in the trunsored model, also N in the truncated model, are obtained as shown in Table 3. The simulation results agree well with the theoretical results, i.e., with the asymptotic standard errors, shown in sections 3 and 4.

Table 3
Simulation results.

	λ	s	m, N
trunsored model			
mean μ	1.0005	0.4999	9998.1
standard deviation sd	0.0217	0.0053	106.0
asymtotoic sd	0.0205	0.0052	103.3
truncated model			
mean μ	1.0005	-	10000.7
standard deviation sd	0.0214	-	78.4
asymtotoic sd	0.0205	-	74.1

$F(t) = 1 - \exp(-t/\lambda)$, $(\lambda = 1,\ s = 1/2,\ T = 2)$

In the table, it is seen that the standard deviation for the parameter λ in the trunsored model is different from that in the truncated model, although it is mentioned that the estimates in the two models are exactly the same provided that the sample data are the same. In generating the random variables in the trunsored model, however, the number of fragile samples is generated according to the binomial distribution $B(n, s)$; some case has 9995 fragile samples, and some has 10002 samples. This explains a small discrepancy of the simulation results between the trunsored model and the truncated model. However, once the failure data are given and the number of samples is the same, the estimates in both the models become exactly the same, as mentioned earlier.

7 Application to the case fatality ratio of SARS

If the case fatality ratio p_f is estimated by using the truncated model,

$$\hat{p}_f = \hat{N}_2/\hat{N}_1, \tag{48}$$

may be used, where, N_1 denotes the number of infected cases and N_2 denotes the number of deaths. As the censoring time T becomes large, the width of the confidence interval for $\hat{N}_i$ goes to zero, as mentioned in section 4. This is not restricted to the exponential model, and this is true if the underlying probability distribution is smooth and the regularity condition holds. Consequently, the width of the confidence interval for p_f also goes to zero; this is easily understood that $\hat{N}_1$ and $\hat{N}_2$ will not fluctuate so much then p_f also

will not. Although the case fatality ratio is considered to be a probabilistic value according to the situation of time and place, the definite value obtained by using the truncated model causes a paradox. It would be natural that the case fatality ratio of SARS in Hong Kong may be expressed as $17\% \pm 1\%$, for example, rather than just 17.037% (= 299/1755).

On the contrary, we may obtain a positive width of the confidence interval for the case fatality ratio if we use the two trunsored models for the infected cases and fatal cases together such that

$$L_{ts}(\theta_1, \theta_2, s_1, s_2) = L_{ts}^1(\theta_1, s_1) \cdot L_{ts}^2(\theta_2, s_2), \tag{49}$$

where

$$\begin{aligned} L_{ts}^j(\theta_j, s_j) = &\{1 - s_j F_j(T; \theta_j)\}^{n_j - r_j} \\ &\prod_{i=1}^{r_j} \{s_j f_j(t_i; \theta_j)\}, \ (j = 1, 2), \end{aligned} \tag{50}$$

because some durable populations are required in the trunsored models. Here, L_{ts}^1 represents the likelihood for the infected cases, and L_{ts}^2 represents the likelihood for the fatal cases. The case fatality ratio $\hat{p}_f$ is obtained by $\hat{s}_2/\hat{s}_1$. Even though n_j $(j = 1, 2)$ are unknown, they may be set to the actual population in Hong Kong (roughly 6,810,000 persons in 2003), for example.

If we use the three trunsored models for the infected cases, fatal cases, and cured cases together, such that

$$\begin{aligned} L_{ts}^j(\theta_j, s_j) = &\{1 - s_j F_j(T; \theta_j)\}^{n_j - r_j} \\ &\prod_{i=1}^{r_j} \{s_j f_j(t_i; \theta_j)\}, \ (j = 1, 2, 3), \\ &s_1 = s_2 + s_3, \end{aligned} \tag{51}$$

where $j = 1, 2, 3$ correspond to the infected, fatal, and cured cases, we can estimate a more accurate case fatality ratio. By setting $n_j = 6,810,000$ $(j = 1, 2, 3)$, and assuming that the underlying probability distributions in three cases are the three-parameter generalized logistic distribution, which is also known as Richards ' curve (Richards (1959)),

$$\begin{aligned} &F_{GL}(x; \sigma, \mu, \beta) = \frac{1}{\{1 + \exp(-(x - \mu)/\sigma)\}^{\beta}}, \\ &(-\infty < x < \infty; -\infty < \mu < \infty; \sigma, \beta > 0), \end{aligned} \tag{52}$$

the estimate of $\hat{p}_f = \hat{s}_2/\hat{s}_1 = 17.16\%$ and its standard error of 1.35% are obtained by using the bootstrap method, when we set the censoring time on May 25, 2003; see Hirose (2007). These values agree well with the other results, e.g., Donnelly at al. (2003) and Jewell et. al. (2007). Even if the censoring time goes to infinity, the standard error for $\hat{p}_f$ remains a positive value larger than 0.58% when we use the method in Bishop et. al. (1975).

8 Concluding remarks

The confidence intervals for the parameters and the number of fragile samples in the trunsored model are compared with those in the truncated model. The maximum likelihood estimates of the parameters for the underlying probability distribution in both models are exactly the same under the condition that all the samples are censored at only one censoring time; consequently, the confidence intervals of the parameters in the trunsored model and those in the truncated model are also the same.

Even if the estimate for the number of fragile samples in the trunsored model is the same as that in the truncated model, however, the confidence interval for it in the trunsored model differs from that in the truncated model because the trunsored model has two fluctuation effects by the distribution parameter and by an extra parameter s whereas the truncated model has the former one only. When the censoring time goes to infinity, the width of the confidence interval of the number of fragile samples in the truncated model converges to zero, while that in the trunsored model converges to a positive constant value, which is the same value in the binomial distribution. The validity for these theoretical consequences is shown by a typical example and a simulation study.

As an application, we have shown the confidence intervals for the case fatality ratio of SARS both in the truncated and trunsored models. Contrary to the paradoxical result that the width of the confidence interval for the case fatality ratio goes to zero when the censoring time becomes large in the truncated model, the trunsored model provides a reasonable confidence interval for the case fatality ratio. This indicates that we have to be cautious in selecting the appropriate model when we deal with the incomplete data models.

References

[1] Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. Discrete multivariate analysis, theory and practice. MIT Press; 1975.

[2] Boag, J.W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy, Journal of the Royal Statistical Society - Series B 1948;11:11–53.
[3] Donnelly, C.A., Ghani, A.C., Leung, G.M., et al. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong, Lancet 2003;361:1761–1766.
[4] Hirose H. The trunsored model and its applications to lifetime analysis: unified censored and truncated model, IEEE Transactions on Reliability 2005;54:11–21.
[5] Hirose, H. The mixed trunsored model with applications to SARS, Mathematics and Computers in Simulation 2007;74:443–453.
[6] Jewell, N.P., Lei, X., Ghani, A.C., Donnelly, C.A., Leung, G.M., Ho L-M., Cowling, B.J., Hedley A.J. Non-parametric estimation of the case fatality ratio with competing risks data: an application to Severe Acute Respiratory Syndrome (SARS) Statistics in Medicine 2007;26:1982–1998.
[7] Maltz, M.D., Mccleary,R. The mathematics of behavioral change, recidivism and construct validity, Evaluation Quarterly 1977;1:421–438.
[8] Meeker, W.Q. Limited failure population life tests, application to integrated circuit reliability, Technometrics 1987;29:51–65.
[9] Richards, F.J. A flexible growth function for empirical use, Journal of Experimental Botany 1959;10:290–300.
[10] WHO (2003), http://www.who.int/csr/sars/country/en/
[11] Wong, W.K. & Bian, G. Estimating parameters in autoregressive models with asym-metric innovations, Statistics and Probability Letters 2005;71:61–70.