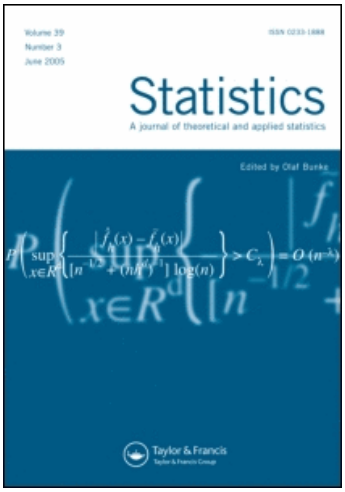
This article was downloaded by: *[Jammalamadaka, S. Rao]*
On: *22 October 2010*
Access details: *Access Details: [subscription number 928527631]*
Publisher *Taylor & Francis*
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713682269

Estimation of Survival Functions and Failure Rate

JAMMALAMADAKA S. Rao[a]; R. C. Tiwari[a]
[a] University of California, Santa Barbara

To cite this Article Rao, JAMMALAMADAKA S. and Tiwari, R. C.(1985) 'Estimation of Survival Functions and Failure Rate', Statistics, 16: 4, 535 — 540
To link to this Article: DOI: 10.1080/02331888508801887
URL: http://dx.doi.org/10.1080/02331888508801887

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.informaworld.com/terms-and-conditions-of-access.pdf

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

statistics 16 (1985) 4, 535–540

Estimation of Survival Function and Failure Rate

JAMMALAMADAKA S. RAO and R. C. TIWARI*

University of California, Santa Barbara

Summary. The BAYESian estimation of the survival function and failure rate in the uncensored case has been treated in PROSCHAN and SINGPURWALLA [5]. In this paper the extension of estimation to randomly censored data is considered. The time interval is partitioned into fixed class intervals. Assuming constant failure rate on these intervals and using a DIRICHLET distribution as the prior, the resulting estimates of survival function and failure rate have nice and simple forms. If instead of the fixed time intervals, one uses the "natural" intervals formed by the observed failure times, this gives essentially the same result as in FERGUSON and PHADIA [3], SUSARLA and VAN RYZIN [7], but in a much simpler form. In this situation the limiting estimates are the KAPLAN-MEIER analog for the discrete situation (not the KAPLAN-MEIER product limit estimator (KAPLAN and MEIER [4]).

Key words: BAYESian inference, DIRICHLET distribution, survival function, failure rate.

Downloaded By: [Jammalamadaka, S. Rao] At: 23:12 22 October 2010

1. Introduction

Let T be a nonnegative random variable representing the failure time, with the distribution function F. The survival function $\bar{F}$ is given by

$$\bar{F}(t) = \mathsf{P}\,(T \geqq t) = 1 - F\,(t-0)$$

and the failure rate is given by

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h}\,\mathsf{P}\,(t \leqq T < t+h \mid T \geqq t)\,,$$

if the limit exists. If F is continuous with the density function f, then

$$\lambda(t) = \frac{f(t)}{\bar{F}(t)}\,.$$

Notice in the literature $\lambda(t)$ is also called the hazard rate, force of mortality, and intensity rate. For a reference to the origin of these definitions see BARLOW and PROSCHAN [1].

Given a continuous failure time distribution F, consider a partition $\{(t_i, t_{i+1}]\}_{i=0}^{k}$ of $(0,\infty)$ with $t_0 = 0$ and $t_{k+1} = \infty$. Consider a set of N individuals. We take observations at the k time points $t_1 \leqq t_2 \leqq \ldots \leqq t_k$ (say at the end of each hour, day or

* Now at Indian Institute of Technology, Bombay, India.

similar period, not necessarily of equal lengths) and observe how many individuals failed and how many were censored (e.g., left the study) in each of these intervals. Let n_i denote the number of failures and m_i the number censored in the interval $(t_i, t_{i+1}]$, $i=0, 1, \ldots, k-1$. For definiteness, we shall assume that those censored in the interval $(t_i, t_{i+1}]$ survived past t_{i+1}. Without loss of generality, we can and shall assume that t_k is sufficiently large that no deaths or censorings occur beyond t_k so that $n_k=0$ and $m_k=0$. Let $n=\sum_{i=0}^{k-1} n_i$, $m=\sum_{i=0}^{k-1} m_i$ so that $N=n+m$. Define $n_{(i)}=\sum_{j=i}^{k-1} n_j$ and $m_{(i)}=\sum_{j=i}^{k-1} m_j$. Then $N_{(i)}=n_{(i)}+m_{(i)}$ is the number of individuals at risk at time t_i+0; that is, whose failure or censoring time is at least t_i, $i=0, 1, \ldots, k-1$.

For $0 \leqq i \leqq k$ define

$$p_i=\int_{t_i}^{t_{i+1}} \mathrm{d}F(t) \tag{1.1}$$

$$q_i=\frac{p_i}{(1-p_0-p_1-\ldots-p_{i-1})}$$

with $q_0=p_0$. Clearly, $q_k=1$. Holding $f(t)=p_i/(t_{i+1}-t_i)$ constant over the interval $(t_i, t_{i+1}]$, the survival function and failure rate are given by

$$\bar{F}(t)=\left\{(1-p_0-\ldots-p_{i-1})-\frac{(t-t_i)}{(t_{i+1}-t_i)}\,p_i\right\} \tag{1.2}$$

$$=\left\{1-\frac{(t-t_i)}{(t_{i+1}-t_i)}\,q_i\right\}\prod_{j=0}^{i-1}(1-q_j)\,,$$

$$\lambda(t)=\frac{q_i}{\{(t_{i+1}-t_i)-(t-t_i)\,q_i\}}\,, \quad t_i<t\leqq t_{i+1}\,, \quad i=0, 1, \ldots, k\,. \tag{1.3}$$

Our problem is to estimate the survival function $\bar{F}(t)$ and the failure rate $\lambda(t)$ defined in (1.2) and (1.3) respectively using the data on $\boldsymbol{n}=\{n_i\}_{i=0}^{k-1}$ and $\boldsymbol{m}=\{m_i\}_{i=0}^{k-1}$. We will denote this data by $\boldsymbol{d}$.

2. Bayes estimation of survival function and failure rate

A Bayesian considers $\boldsymbol{q}=\{q_i\}_{i=0}^{k}$ as a random vector and represents his/her opinion about $\boldsymbol{q}$ by a probability distribution, called a prior distribution, or simply, a prior. After observing the failure data $\boldsymbol{d}=(\boldsymbol{n}, \boldsymbol{m})$, his/her opinion about $\boldsymbol{q}$, given the data $\boldsymbol{d}$, is called the posterior distribution. Let g be the prior density function of $\boldsymbol{q}$. The posterior density function of $\boldsymbol{q}$ given $\boldsymbol{d}$, namely $g(\boldsymbol{q} \mid \boldsymbol{d})$, is given by the relation

$$g(\boldsymbol{q} \mid \boldsymbol{d}) \propto g(\boldsymbol{q})\, L(\boldsymbol{q} \mid \boldsymbol{d})\,, \tag{2.1}$$

where $L(\boldsymbol{q} \mid \boldsymbol{d})$ is the likelihood function of $\boldsymbol{q}$ at the point $\boldsymbol{d}$, and the constant of proportionality which does not depend on $\boldsymbol{q}$ is $\{\int L(\boldsymbol{q} \mid \boldsymbol{d})\, g(\boldsymbol{q})\, \mathrm{d}\boldsymbol{q}\}^{-1}$.

Downloaded By: [Jammalamadaka, S. Rao] At: 23:12 22 October 2010

One usually employs a prior within a family G of distributions, which is large enough to accommodate various shades of opinion about the parameter $\boldsymbol{q}$. Further, if $g \in G$ is a prior for $\boldsymbol{q}$, then the posterior $g(\boldsymbol{q} \mid \boldsymbol{d})$ ought to be in a simple computable form. If $g(\boldsymbol{q} \mid \boldsymbol{d}) \in G$ for all $g \in G$ and for all data $\boldsymbol{d}$, then G is called a *conjugate family* of priors.

Observing that there is one-to-one correspondence between $\boldsymbol{p} = \{p_i\}_{i=0}^{k}$ and $\boldsymbol{q} = \{q_i\}_{i=0}^{k}$, we define a conjugate family of priors for $\boldsymbol{p}$ as follows. Let $\boldsymbol{\alpha} = \{\alpha_i\}_{i=0}^{k}$ be a sequence of finite non-negative numbers. We say the random vector $\boldsymbol{p}$ has a k-dimensional DIRICHLET distribution with parameter $\boldsymbol{\alpha}$, and denote it by $\boldsymbol{p} \sim D(\boldsymbol{\alpha})$, if the distribution of $(p_0, p_1, \ldots, p_{k-1})$ is $D(\alpha_0, \alpha_1, \ldots, \alpha_{k-1}; \alpha_k)$ as defined in WILKS [8], Section 7.7.

Under the prior $D(\boldsymbol{\alpha})$ for $\boldsymbol{p}$ the coordinates of random vector $\boldsymbol{q}$ are independent, with q_i having a Beta distribution with parameters α_i and $\alpha_{(i+1)}$ denoted by Beta $(\alpha_i, \alpha_{(i+1)})$, where $\alpha_{(i)} = \sum_{j=i}^{k} \alpha_j$, $i = 0, 1, \ldots, k-1$. After observing the data $\boldsymbol{d}$ the likelihood function of $\boldsymbol{q}$ at the point $\boldsymbol{d}$ is

$$L(\boldsymbol{q} \mid \boldsymbol{d}) = \prod_{i=0}^{k-1} p_i^{n_i} (1 - p_0 - p_1 - \ldots - p_i)^{m_i} \tag{2.2}$$
$$= \prod_{i=0}^{k-1} q_i^{n_i} (1 - q_i)^{n_{(i+1)} + m_{(i)}}$$

Using (2.1) and (2.2) the posterior distribution of $\boldsymbol{q}$ given the data $\boldsymbol{d}$ is given by the relation

$$g(\boldsymbol{q} \mid \boldsymbol{d}) \propto \prod_{i=0}^{k-1} q_i^{\alpha_i + n_i - 1} (1 - q_i)^{\alpha_{(i+1)} + n_{(i+1)} + m_{(i)} - 1}. \tag{2.3}$$

The BAYES estimate of q_i under prior $D(\boldsymbol{\alpha})$ (with respect to the squared error loss function) based on no sample, called the prior BAYES estimate of q_i, is

$$\hat{q}_{i,\alpha}^{0} = \mathsf{E}_{D(\alpha)}(q_i) = \frac{\alpha_i}{\alpha_{(i)}}, \tag{2.4}$$

and based on the data $\boldsymbol{d}$ is

$$\hat{q}_{i,\alpha}^{N} = \mathsf{E}_{D(\alpha)}(q_i \mid \boldsymbol{d}) = \frac{\alpha_i + n_i}{\alpha_{(i)} + n_{(i)} + m_{(i)}} = \frac{\alpha_i + n_i}{\alpha_{(i)} + N_{(i)}}, \quad i = 0, 1, \ldots, k-1. \tag{2.5}$$

Clearly, from (2.4) and (2.5) we have

$$\hat{q}_{i,\alpha}^{N} = w_i \hat{q}_{i,\alpha}^{0} + (1 - w_i)\, \hat{\lambda}_i, \tag{2.6}$$

where $w_i = \alpha_{(i)} / (\alpha_{(i)} + N_{(i)})$. Thus $\hat{q}_{i,\alpha}^{N}$ is a weighted mean of the prior BAYES estimate $\hat{q}_{i,\alpha}^{0}$ and the empirical estimate $\hat{\lambda}_i = n_i / N_{(i)}$ with weights w_i and $(1 - w_i)$ respectively. Substituting $\hat{q}_{i,\alpha}^{0}$ and $\hat{q}_{i,\alpha}^{N}$ in equations (1.2) and (1.3), the BAYES estimates of $\bar{F}$ and λ for no sample size and with the data $\boldsymbol{d}$ are respectively

$$\hat{\bar{F}}_{\alpha}^{0}(t) = \left\{1 - \frac{(t - t_i)}{(t_{i+1} - t_i)} \frac{\alpha_i}{\alpha_{(i)}}\right\} \prod_{j=0}^{i-1} \left(1 - \frac{\alpha_j}{\alpha_{(j)}}\right) \tag{2.7}$$
$$\hat{\bar{F}}_{\alpha}^{N}(t) = \left\{1 - \frac{(t - t_i)}{(t_{i+1} - t_i)} \frac{\alpha_i + n_i}{\alpha_{(i)} + N_{(i)}}\right\} \prod_{j=0}^{i-1} \left(1 - \frac{\alpha_j + n_j}{\alpha_{(j)} + N_{(j)}}\right), \quad t_i < t \leqq t_{i+1},$$

Downloaded By: [Jammalamadaka, S. Rao] At: 23:12 22 October 2010

and

$$\hat{\lambda}^0_\alpha(t) = \frac{\alpha_i}{\{(t_{i+1}-t_i)\,\alpha_{(i)} - (t-t_i)\,\alpha_i\}} \tag{2.8}$$

$$\hat{\lambda}^N_\alpha(t) = \frac{\alpha_i + n_i}{\{(t_{i+1}-t_i)\,(\alpha_{(i)}+N_{(i)}) - (t-t_i)\,(\alpha_i+n_i)\}}, \quad t_i < t \leqq t_{i+1}\,,$$

$$i = 0, 1, \ldots, k-1\,.$$

3. Limiting Bayes estimates

It appears that $\alpha_{(i)}$, the sum of the parameters α_i and $\alpha_{(i+1)}$ of the distribution Beta $(\alpha_i, \alpha_{(i+1)})$ of q_i, enters into the expressions (2.4) and (2.5) as the 'prior sample size.' This has given rise to the general feeling that allowing $\alpha_{(i)}$ to become small not only makes the 'prior sample size' small but also it corresponds to no prior information (see, for example, FERGUSON [2], in the context of DIRICHLET processes). By investigating the limit of the BAYES estimate of q_i when $\alpha_{(i)}$ is allowed to converge to zero, we show below that it is misleading to think of $\alpha_{(i)}$ as the prior sample size and the smallness of $\alpha_{(i)}$ as having no information. (This result is discussed in SETHURAMAN and TIWARI [6] in the context of DIRICHLET processes.)

Consider the convergent sequences of non-negative numbers $\boldsymbol{\alpha} = \{\alpha_i^r\}_{i=0}^k$, $r = 0, 1, 2, \ldots$ If $\alpha_j^r \to \alpha_j^0$ as $r \to \infty$, then $\boldsymbol{q} \xrightarrow{\mathcal{D}} \boldsymbol{q}^0$, where for each r, $\boldsymbol{q}^r = \{q_i^r\}$ is defined in (1.1) and $\xrightarrow{\mathcal{D}}$ represents convergence in distribution, and

$$\begin{aligned} &\hat{q}^0_{i,\alpha^r} \to \hat{q}^0_{i,\alpha^0} \\ &\hat{q}^N_{i,\alpha^r} \to \hat{q}^N_{i,\alpha^0}\,, \quad i = 0, 1, \ldots, k-1\,, \end{aligned} \tag{2.9}$$

as $r \to \infty$. Further, if we let $\alpha_{(i)}^r$ converge to zero such that $\alpha_i^r/\alpha_{(i)}^r$ converges to a constant ϑ_i, $0 < \vartheta_i < 1$, then in the limit as r tends to infinity q_i^r is the Binomial $(1, \vartheta_i)$ random variable; and

$$\hat{q}^N_{i,\alpha^r} \to \frac{n_i}{N_{(i)}}, \tag{2.10}$$

which is the empirical estimate λ_i, $i = 0, 1, \ldots, k-1$. Also, $\hat{\bar{F}}^N_{\alpha^r}$ (which is as defined in (2.7) with $\boldsymbol{\alpha}$ replaced by $\boldsymbol{\alpha}^r$) in the limit is the empirical estimate

$$\begin{aligned} &\hat{\bar{F}}(t) = \left\{1 - \frac{(t-t_i)}{(t_{i+1}-t_i)}\,\frac{n_i}{N_{(i)}}\right\} \prod_{j=0}^{i-1} \left(1 - \frac{n_j}{N_{(j)}}\right), \\ &t_i < t \leqq t_{i+1}, \quad i = 0, 1, \ldots, k-1\,. \end{aligned} \tag{2.11}$$

Let $\{t_i\}_{i=1}^k$ denote the distinct observed failure times, again with $t_0 = 0$ and $t_{k+1} = \infty$. Let n_i denote the number of failures at t_{i+1}, $i = 0, 1, \ldots, k-1$, and let m_i, $n_{(i)}$, $m_{(i)}$ and $N_{(i)}$ remain as before. We wish to partition the time-interval using these t_i's, $i = 1, 2, \ldots, k$, as before. This procedure is clearly justified when T is a discrete random variable and $\{t_i\}_{i=1}^k$ is its support so that failures occur

Downloaded By: [Jammalamadaka, S. Rao] At: 23:12 22 October 2010

only at these points. On the other hand, one may look at the case of continuous failure times as being approximated by a discrete situation like this where, because of observational restrictions, one observes the process at t_{i+1} and makes the approximation that the n_i failures in the interval actually occurred at t_{i+1} instead of over the period $(t_i, t_{i+1}]$, $i=0, 1, \ldots, k-1$. This is analogous to the assumptions one makes in computing statistics like the mean and variance from continuous data that has been grouped into class intervals. It should be remarked that the procedure used by SUSARLA and VAN RYZIN [7] amounts implicitly to such a partitioning of the timeinterval using the distinct observed failure times and using it to find the censoring numbers (cf. their section 3). In this case, the survival function and the failure rate are given by (cf. equations (1.2) and (1.3))

$$\bar{F}(t)=\prod_{j=0}^{i}(1-q_j), \quad t_i<t\leqq t_{i+1}\,, \tag{2.12}$$

and

$$\lambda(t)=\begin{cases} q_i & \text{at } t=t_{i+1} \\ 0, & t_i<t<t_{i+1} \end{cases}, \quad i=0, 1, \ldots, k\,. \tag{2.13}$$

Also, the BAYES estimates of $\bar{F}$ and λ are given by (cf. equations (2.7) and (2.8))

$$\hat{\bar{F}}^N_\alpha(t)=\prod_{j=0}^{i}\left\{1-\frac{\alpha_j+n_j}{\alpha_{(j)}+N_{(j)}}\right\}, \quad t_i<t\leqq t_{i+1}\,, \tag{2.14}$$

and

$$\hat{\lambda}^N_\alpha(t)=\begin{cases} \dfrac{\alpha_i+n_i}{\alpha_{(i)}+N_{(i)}} & \text{at } t=t_{i+1}, \\ 0, & t_i<t<t_{i+1} \end{cases} \quad i=0, 1, \ldots, k-1\,. \tag{2.15}$$

Now an equation similar to (2.6) holds with λ_i replaced by the KAPLAN-MEIER estimate $\hat{\lambda}_{KM}(t)=\dfrac{n_i}{N_{(i)}}$ at $t=t_{i+1}$. (Note that this not the usual KAPLAN-MEIER product limit estimate at $t=t_{i+1}$). Again, if $\alpha_{(i)}\to 0$ and $\dfrac{\alpha_i}{\alpha_{(i)}}\to\vartheta_i$, $0<\vartheta_i<1$, in the sense discussed above, then the estimates (2.14) and (2.15) converge to

$$\hat{\bar{F}}_{KM}(t)=\prod_{j=0}^{i}\left\{1-\frac{n_j}{N_{(j)}}\right\}, \quad t_i<t\leqq t_{i+1}\,, \tag{2.16}$$

and

$$\hat{\lambda}_{KM}(t)=\begin{cases} \dfrac{n_i}{N_{(i)}} & \text{at } t=t_{i+1}, \\ 0, & t_i<t<t_{i+1} \end{cases} \quad i=0, 1, \ldots, k-1\,. \tag{2.17}$$

References

[1] BARLOW, R. E. and PROSCHAN, F. (1975). *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, Inc., New York.

[2] FERGUSON, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Ann. Statist.*, **1**, 209–230.

[3] FERGUSON, T. S. and PHADIA, E. G. (1977). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **3**, 736–740.

Downloaded By: [Jammalamadaka, S. Rao] At: 23:12 22 October 2010

[4] KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53**, 457–481.
[5] PROSCHAN, F. and SINGPURWALLA, N. D. (1980). A new approach to inference from accelerated life tests. *IEEE Transactions on Reliability* **29**, 2, 98–102.
[6] SETHURAMAN, J. and TIWARI, RAM C. (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Statistical Decision Theory and Related Topics III, Vol. 2, eds.* SHANTI S., GUPTA and JAMES, O. BERGER, Academic Press. Inc. 305–315.
[7] SUSARLA, V. and VAN RYZIN, J. (1976). Nonparametric estimation of survival curves from incomplete observations. *J. Amer. Statist. Assoc.* **71**, 897–902.
[8] WILKS, S. S. (1962). *Mathematical Statistics*. John Wiley and Sons, New York.

Received October 1982; revised December 1983.

Jammalamadaka S. RAO
Department of Mathematics
University of California
Santa Barbara, CA. 93106
U.S.A.

R. C. TIWARI
Indian Institute of Technology
Bombay
India 400076

Downloaded By: [Jammalamadaka, S. Rao] At: 23:12 22 October 2010