Communications in Statistics—Theory and Methods, 40: 4409–4416, 2011

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2010.513791

Taylor & Francis
Taylor & Francis Group

Adjusted Hazard Rate Estimator Based on a Known Censoring Probability

ÜLKÜ GÜRLER[1] AND PAUL KVAM[2]

[1]Department of Industrial Engineering, Bilkent University, Ankara, Turkey
[2]H. Milton Stewart School of Industrial Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

In most reliability studies involving censoring, one assumes that censoring probabilities are unknown. We derive a nonparametric estimator for the survival function when information regarding censoring frequency is available. The estimator is constructed by adjusting the Nelson–Aalen estimator to incorporate censoring information. Our results indicate significant improvements can be achieved if available information regarding censoring is used. We compare this model to the Koziol–Green model, which is also based on a form of proportional hazards for the lifetime and censoring distributions. Two examples of survival data help to illustrate the differences in the estimation techniques.

Keywords Hazard function; Kaplan–Meier product-limit estimator; Koziol–Green model; Nelson–Aalen estimator; Stochastic precedence.

1. Problem Description

Suppose we have a sample of potentially right-censored observations and lifetime distribution F with the paired censoring distribution G. If $X_i \sim F(\cdot)$ and $Y_i \sim G(\cdot)$ with $i = 1, \ldots, n$, suppose X_i and Y_i are independent and let $Z_i = \min(X_i, Y_i)$ represent the observed lifetime of the ith item with non censoring indicator $\delta_i = I(X_i < Y_i)$. The Kaplan and Meier (1958) product-limit estimator is asymptotically efficient for F in this case.

In many problems of survival analysis, it is known that values generated from F are stochastically smaller than those generated by G in some sense. In applications, this is evident in trials in which censoring is uncommon. With this kind of censoring, in which the censoring conveys knowledge about F, the Kaplan–Meier estimator is not necessarily asymptotically efficient.

Received November 11, 2009; Accepted July 27, 2010
Address correspondence to Paul Kvam, H. Milton Stewart School of Industrial Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA; E-mail: pkvam@isye.gatech.edu

Censored data are typical in survival and reliability studies, and there is a vast literature on the estimation and inference methods with censored data. In almost all of these studies it is assumed that the probability of censoring is unknown. To reduce the uncertainty regarding the censoring mechanism, several models have been employed by researchers. In parametric life-testing problems, for example, the relationship between F and G can be modeled by imposing constraints on the parameters of the lifetime distribution or the censoring distribution. In nonparametric problems, this constraint on the relationship between censoring time and lifetime must be modeled directly through F and G. For example, the Koziol–Green (KG) model (Koziol and Green, 1976) stipulates that $\overline{G}(t) = \overline{F}(t)^{\beta}$, where $\overline{F}(t) = 1 - F(t)$ and $\beta > 0$. This particular structure induces an ordering between F and G, depending on the value of β; if $\beta > 1$, for example, the random variable X tends to be larger than Y in a stochastic sense. One can show that $\beta > 1$ if and only if G is smaller than F in *likelihood ratio* (lr) *ordering*. For this ordering, X is less than Y in likelihood ratio ($X \leq_{lr} Y$) iff $G(F^{-1})$ is convex. Note the order between F and G are simply reversed in the case $\beta \leq 1$.

Likelihood ratio is one of many stochastic orders that can distinguish rank between the lifetime distribution and the censoring distribution when censoring is present. Other commonly applied orders are stochastic ordering (st) and hazard rate ordering (hr). See Shaked and Shanthikumar (1994) for a comprehensive discussion of stochastic orders. We have $X \leq_{st} Y$ iff $F(t) \geq G(t), \forall t$, and $X \leq_{hr} Y$ iff $\overline{F}(t)/\overline{G}(t)$ decreases in t. It is known that $X \leq_{lr} Y \Rightarrow X \leq_{hr} Y \Rightarrow X \leq_{st} Y$, so that likelihood ratio ordering is the strongest of the three.

The likelihood ratio ordering is considered extremely restrictive in many applications, and as a consequence, the Koziol–Green model can only be applied to survival data in which the censoring variable is larger than the lifetime variable in a strict stochastic sense. Csörgő (1989) showed that this assumption is insupportable in typical sets of lifetime data. Extensions have been constructed to make the KG model more applicable; e.g., Peña and Rohatgi (1987).

Arcones et al. (2002) introduced *stochastic precedence* between X and Y ($X \leq_{sp} Y$), which occurs if $P(X \leq Y) \geq 1/2$. It is known that stochastic precedence (sp) is implied by stochastic ordering, and is thus the weakest ordering of the four mentioned. Unlike the censoring constraints generated by the Koziol–Green model, the sp-constraint is relatively flexible and a wide variety of distributions can be considered for modeling lifetime and censoring. Arcones et al. (2002) discussed applications where the sp-constraint makes a difference in developmental testing, robust estimation of location parameters, and tolerance-limit problems to name a few.

Although such restrictive models have been considered to link the censoring and lifetime distributions to obtain more efficient estimators, to the best of our knowledge, there is no study that assumes a known censoring probability. Hence, how the estimators should be modified and what the value of this information would be in terms of the estimators' quality have not been discussed in the literature. In this short note, we aim to fill this gap. Motivated by the idea of stochastic precedence for linking F and G, we assume that rather than an available bound, exact information regarding the censoring proportion is available from external resources. In particular, we assume that $P(X \leq Y) = \alpha$, where $0 \leq \alpha \leq 1$ is specified. This assumption may be realistic in applications where there has been sufficient data accumulation from similar studies.

In the following section, we use an adjusted hazard rate estimator based on the Kaplan–Meier product limit estimator of F under the constraint that $P(X \leq Y) = \alpha$

(for some specified value of $0 \leq \alpha \leq 1$). The estimation of the censoring distribution G is considered secondary. The estimator derived is illustrated with the stage-IV prostate cancer data referenced by Koziol and Green (1976) to motivate the KG model.

2. Adjusted Hazard Estimator

If we define the counting process $N(t) = \sum I(Z_i \leq t, \delta_i = 1)$ and $Y(t) = \sum I(Z_i \geq t)$, the Kaplan–Meier estimator for right-censored data is

$$F_{KM}(t) = 1 - \prod_{Z_i \leq t} \left(1 - \frac{dN(Z_i)}{Y(Z_i)}\right)$$

and the (cumulative) hazard of F, defined as $R(t) = -\log(\overline{F}(t))$, can be expressed in convenient Nelson–Aalen form: $R_{KM}(t) = \int_0^t dN(u)/Y(u)$. The Nelson–Aalen estimator does not perfectly match up with the product limit estimator, especially after the last observation, so here we assume t such that $Y(t) > 0$. Because the two estimators are asymptotically equivalent, we focus on the Kaplan–Meier estimator to illustrate asymptotic properties. Assume that $m = \sum \delta_i$, so that $n - m$ of the n observations are censored.

F and G are two distributions such that $P(X \leq Y) = \alpha$, for some fixed non censoring probability of $\alpha \in [0, 1]$. Equivalently, $\int \overline{G}(u)dF(u) = \alpha$ and $\int \overline{F}(u)dG(u) = 1 - \alpha$. Let F_n be the empirical distribution function (EDF) based on the m observed failure times, and G_n be the empirical distribution based on the $n - m$ censored observations. Along with F_n and G_n, define H_n as the EDF of the combined data, i.e., $H_n(t) = n^{-1} \sum I(Z_i \leq t)$. Under the assumption that $P(X \leq Y) = \alpha$, it's easy to show that:

1. $\overline{F}_n(t) \to \overline{F}^*(t) \equiv \frac{1}{\alpha} \int_t^\infty \overline{G}(u)dF(u)$;
2. $\overline{G}_n(t) \to \overline{G}^*(t) \equiv \frac{1}{1-\alpha} \int_t^\infty \overline{F}(u)dG(u)$;
3. $\overline{H}_n(t) = n^{-1} \sum I(z_i > t) \to \overline{H}^*(t) \equiv \overline{F}(t)\overline{G}(t)$.

Note that $dF^*(t) = \overline{G}(t)dF(t)/\alpha$, $dG^*(t) = \overline{F}(t)dG(t)/(1 - \alpha)$. From this, R can be expressed as

$$R(t) = \alpha \int_0^t \frac{dF^*(u)}{\overline{H}^*(u)}. \tag{1}$$

An intuitive estimator for the hazard, then, can be constructed from (1) as a function of $\hat{\alpha} = m/n$ and the Kaplan–Meier hazard function $R_{KM}(t)$:

$$\begin{aligned} \widehat{R}(t) &= \alpha \int_0^t \frac{dF_n(u)}{\overline{H}_n(u)} = \alpha \sum_{i=1}^n \frac{m^{-1} I(\delta_i = 1)}{n^{-1} \sum_{j=1}^n I(z_j \geq z_i)} \\ &= \alpha \left(n^{-1} \sum_i I(x_i \leq y_i) \right)^{-1} \int_0^t \frac{dN(u)}{Y(u)} \\ &= \frac{\alpha}{\hat{\alpha}} R_{KM}(t). \end{aligned} \tag{2}$$

We call $\widehat{R}(t)$ the *adjusted hazard rate* (AHR) estimator.

Properties of the corresponding estimator for the lifetime distribution, $\widehat{F}(t) = 1 - \exp\{-\widehat{R}(t)\}$ are given in the theorems that follow.

Theorem 2.1. *In* $\{t : \overline{H}(t) > 0\}$, *if* $\widehat{\overline{F}}(t) = \exp\{-\widehat{R}(t)\}$ *where* $\widehat{R} = (\alpha/\hat{\alpha})R_{KM}(t)$ *is the AHR estimator and* R_{KM} *is the Kaplan–Meier* (*cumulative*) *hazard function, then with probability* 1,

$$\sup_t |\widehat{F}(t) - F(t)| \to 0.$$

Theorem 2.2. *If* $\widehat{\overline{F}}(t) = \exp\{-\widehat{R}(t)\}$ (*as in Theorem* 2.1), *then*

$$\sqrt{n}\left(\widehat{F} - F\right) \Rightarrow \mathscr{W} \tag{3}$$

where $\mathscr{W}$ *is a zero-mean Gaussian process with covariance function*

$$\sigma^2(s, t) = \overline{F}(t)\overline{F}(s) \int_0^{s\vee t} \frac{dF(u)}{\overline{F}(u)\overline{G}(u)}. \tag{4}$$

Theorem 2.1 follows from the strong consistency of the KM estimator and the strong law of large numbers for $\hat{\alpha}$. The asymptotic variance in (4) is the familiar covariance function of the Kaplan Meier estimator for right-censored data. Because $\hat{\alpha} \xrightarrow{P} \alpha$, by Slutsky's Theorem, Theorem 2.2 follows.

Comparisons made between estimators based on the KG model and the KM estimator are synonymous if we substitute the AHR estimator for KM. The nonparametric MLE for the KG model, derived by Cheng and Lin (1987) can be expressed as $\overline{F}_{KG}(x) = \overline{H}_n(x)^{\hat{\alpha}}$. Unlike the KG estimator, $\widehat{F}$ and F_{KM} assign probability mass only on non censored observations.

Cheng and Lin showed that $\overline{F}_{KG}(x) = \overline{H}_n(x)^{\hat{\alpha}}$ is more efficient than the AHR estimator in the case the KG model holds. Otherwise, the AHR estimator is more efficient. The arguments in Csörgő (1988) hold for both cases. Both estimators adjust the Kaplan–Meier estimator via proportional hazards. Compared to (1), the nonparametric MLE for F in the KG model can be expressed in terms of its hazard function (R_{KG}) as

$$R_{KG}(t) = \hat{\alpha} R_{H_n}(t),$$

where R_{H_n} is the cumulative hazard function for H_n. With $R_{H_n} \to R_F + R_G$, we see how the role of the censoring distribution in the KG estimator is clearly more primary for the KG estimator than the AHR estimator.

3. Examples

We consider below two examples that motivated past research using censored data and the Koziol–Green model. The first set (prostate cancer data), referenced by Koziol and Green (1976), does not actually fit the KG model well. The second set (retirement center data) was found to be more suited in a comparative study by Csörgő (1989). In neither set of historic data can we informatively select a probability that accurately reflects the true nature of the censoring that is

expected. Furthermore, later studies have shown that the observed censoring rate is high because it includes death by other causes. Still, the examples are helpful in illustrating the applicability of the AHR estimator.

3.1. *Prostate Cancer*

The model proposed by Koziol and Green (1976) was inspired, in part, by a set of data based on a clinical trial of 211 individuals who had Stage IV prostate cancer. An updated version of the data are listed in Table 2 in Hollander and Proschan (1979). Of the 211 individuals who were treated with estrogen, 90 died of prostate cancer, 105 died of other diseases, and 16 were still alive at the end of the study. These $105 + 16 = 121$ observations were treated as right censored.

The order restriction inherent with $\widehat{F}$ is specified by the experimenter. Any specification of $\alpha = P(X \leq Y)$ pulls $\widehat{F}$ over or under the regular Kaplan–Meier estimator F_{KM}. Figure 1 shows the order restricted estimators based on $\alpha = 0.50$ alongside the KG estimator. Survival time was measured in months. The magnitude of difference between the curves is not strongly evident in the figure; the mean square distance $(\int [F_1(x) - F_2(x)]^2 dx)$ between the KM estimator and the KG estimator is more than twice that between the KM and the adjusted hazard estimator $(\widehat{F})$. The AHR estimator makes a lesser augmentation on the KM estimate, and since its hazard is proportional to that of the KM, the shape remains the same. The KG estimator features a proportional hazard, but it is not the hazard of the KM estimator, and Fig. 1 shows how the KG estimator changes the shape to subscribe to the Koziol–Green constraint.

In this example, $\alpha = 0.50$ was somewhat arbitrarily chosen without any knowledge of the lifetime distribution's relationship to the censoring distribution. In fact, the data showed more-than-expected censoring; since $\hat{\alpha} = 0.42654$, the stochastic precedence constraint of $\alpha = 0.5$ actually pulls the AHR distribution *under* the KM distribution. At $\alpha = \hat{\alpha}$, we have a "break-even point" where F_{KM} and $\widehat{F}$ are coincidental.

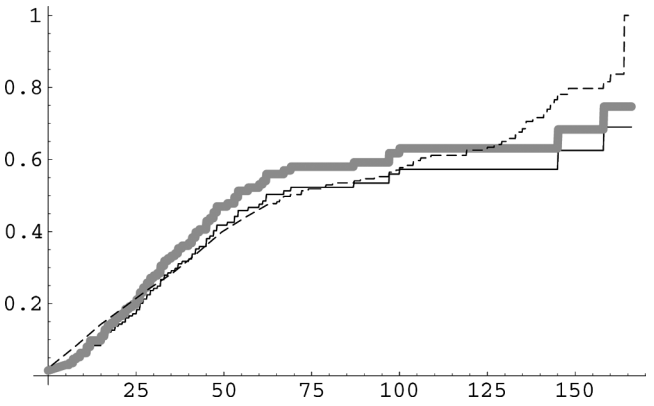

Figure 1. MLE of $F(t)$ for prostate data with $\alpha = 0.50$ (solid line), the Kaplan–Meier estimator (gray line), and the KG estimator (dashed line). Time is measured in months.

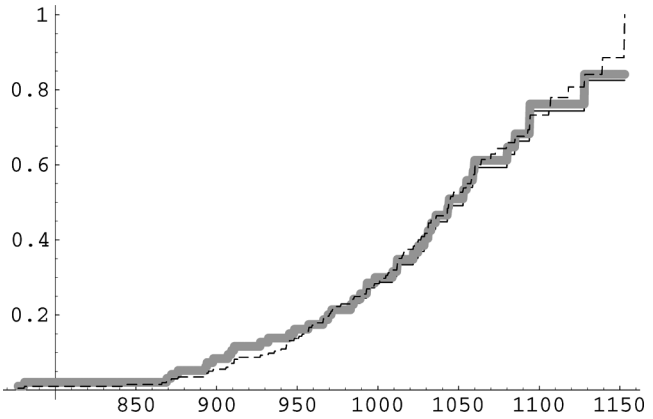

Figure 2. MLE of $F(t)$ for Channing House retirement data with $\alpha = 0.50$ (solid line), the Kaplan–Meier estimator (gray line), and the KG estimator (dashed line). Time (x-axis) is measured in months.

3.2. *Retirement Center Data*

In contrast to the last example, we consider a set of survival data that actually fits the KG model well. Csörgő (1989) presents a test for the proportional hazard found in the KG model and considered several published sets of survival data to illustrate the test, including the example above. The prostate survival data, in fact, does not fit the KG model adequately. This fact has unforseen consequences on Koziol and Green's test for exponentiality because it is based on the assumption of the proportional hazard in the KG model.

Csörgő (1989) examined the well-known Stanford heart transplant data by Miller and Halpern (1982), censored recurrence times of myocardial infarction from Chen (1981), pacemaker failure data described in Csörgő and Horváth (1986) and survival data for male residents of a retirement center featured in Efron (1999). Of these six sets of censored survival data, only the retirement center data can be modeled well with the proportional hazard of Koziol and Green.

Figure 2 shows the estimators for the lifetime distribution based on 97 men from the Channing House retirement center in Palo Alto, California. Lifetime is measured in calendar months. The study kept track of resident lifetimes from the center's opening in 1964 until the study finished in 1975. In that time, 46 of the 97 residents died at the Channing House, 5 moved elsewhere, and 46 were alive at the end of the study. Unlike the distributions in Fig. 1, there are really no remarkable differences in the three plots in Fig. 2: neither the KG estimator or the AHR estimator ($\hat{\alpha} = 0.4742$) augment the Kaplan–Meier estimator to fit the hypothesized model constraints, as the original data reflects those constraints naturally.

4. Simulation and Discussion

For the case when the censoring information is available, the adjusted hazard rate estimator derived earlier has important advantages over estimators based on the Koziol–Green model. Although the sp – constraint is weaker than the more commonly used stochastic orderings, the choice of α in $P(X \leq Y) = \alpha$ can still be a

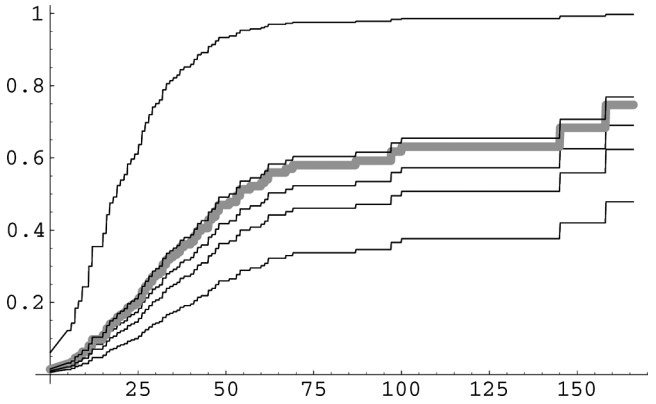

Figure 3. Order restricted MLE of F for Prostate data with $\alpha = \{0.10, 0.40, 0.50, 0.60, 0.90\}$ (solid lines from top to bottom) along with the Kaplan–Meier estimator (gray line).

crucial assumption. We have not considered the consequences of misspecifying α, for example.

In the first example, with $\alpha = 0.5$ decreased the estimated distribution function (relative to the Kaplan–Meier estimator) because there was actually more than 50% censoring ($\hat{\alpha} = 0.4665$). The difference between $\hat{\alpha}$ and 0.50 was smaller in the second example, and the plots of the two estimators are nearly coincidental.

Figure 3 shows the AHR estimator for the prostate data again, but this time various levels of α are used. While the plots for $\alpha = 0.40$ or 0.50 are close to the KM estimator, the heavier constraints using $\alpha = .90$ (bottom CDF) or $\alpha = 0.10$ (top cdf) cause a dramatic change in the estimator.

We compared relative mean squared error (MSE) in terms of the MSE for the Kaplan–Meier estimator. With $\mathrm{MSE}(\widehat{F}, F) = \int (F - \widehat{F})^2 dF$, the relative MSE for both the AHR estimator and the Koziol–Green estimator are defined as

$$\Lambda(\widehat{F}, F_{KM}) = \frac{\mathrm{MSE}(\widehat{F}, F)}{\mathrm{MSE}(F_{KM}, F)}, \quad \Lambda(F_{KG}, F_{KM}) = \frac{\mathrm{MSE}(F_{KG}, F)}{\mathrm{MSE}(F_{KM}, F)}.$$

As a function of $F(x)$, a smoothed version of Λ is plotted in Fig. 4 based on 1,000 simulations in which $n = 200$ lifetimes are generated from a Gamma(γ, 1) distribution, with γ representing the shape parameter. The censoring distribution is Exponential with the mean adapted to achieve the desired $\alpha = P(X < Y)$ value, which is either $\alpha = 0.5$ (in plots a, b, c) or $\alpha = 0.7$ (in plot d). Figure 4a has $\gamma = 1$, which actually fits the Koziol–Green model. Not surprisingly, this is the only setting for which F_{KG} performs uniformly better than the Kaplan–Meier estimator.

With $\gamma = 2$ or 4, in Figs. 4b and 4c, respectively, the MSE for F_{KG} is much larger in the tails compared to the other estimators; near $F(x) = 0.10$, $\Lambda(F_{KG}, F_{KM})$ is between 4 and 20. This is also the case for Fig. 4d, where $\gamma = 4$ but α is changed to 0.7. Perhaps most importantly, the AHR estimator performs slightly better than the Kaplan–Meier estimator in all four settings, and is unquestionably better than F_{KG} in the cases $\gamma > 1$.

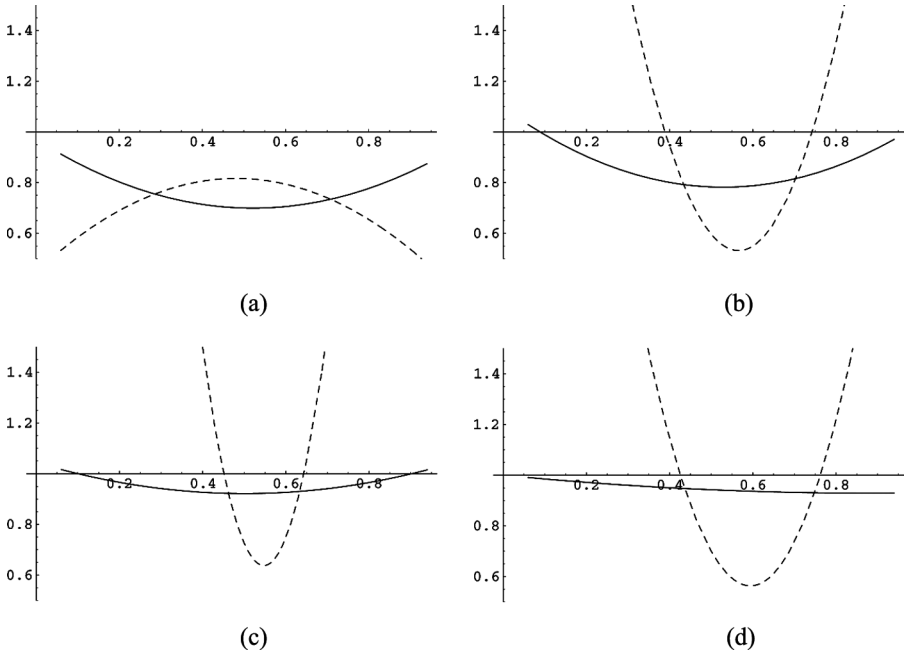

Figure 4. Plot of relative MSE (wrt KM estimator) vs. $0 < F(x) < 1$, where $\Lambda(\widehat{F}, F_{KM})$ is the solid line and $\Lambda(F_{KG}, F_{KM})$ is the dashed line. $n = 200$ data are generated from $\Gamma(\gamma, 1)$ with (a) $\alpha = 0.5$, $\gamma = 1$, (b) $\alpha = 0.5$, $\gamma = 2$, (c) $\alpha = 0.5$, $\gamma = 4$, (d) $\alpha = 0.7$, $\gamma = 4$.

References

Arcones, M. A., Kvam, P. H., Samaniego, F. J. (2002). Nonparametric estimation of a distribution subject to a stochastic precedence constraint. *J. Amer. Statist. Assoc.* 97(457):170–182.

Chen, C. H. (1981). Correlation-type goodness-of-fit tests for randomly censored data. *Technical Report No. 73*. Division of Biostatistics, Stanford University, Stanford, CA.

Cheng, P. E., Lin, G. D. (1987). Maximum likelihood estimation of a survival function under the Koziol–Green proportional hazards model. *Statist. Probab. Lett.* 5:75–80.

Csörgő, S. (1988). Estimation in the proportional hazards model of random censorship. *Statistics* 19:437–463.

Csörgő, S. (1989). Testing for the proportional hazards model of random censorship. In: Mandel, P., Hušková, M., eds. *Proc. Fourth Prague Sympo. Asymptotic Statist.* Prague: Charles University Press.

Csörgő, S., Horváth, L. (1986). Confidence bands from censored samples. *Canad. J. Statist.* 14:131–144.

Efron, B. (1999). Censored data and the bootstrap. *J. Amer. Statist. Assoc.* 76:312–319.

Hollander, M., Proschan, F. (1979). Testing to determine the underlying distribution using randomly censored data. *Biometrics* 35:393–401.

Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* 53:457–481.

Koziol, J. A., Green, S. B. (1976). A Cramer-von Mises statistic for randomly censored data. *Biometrika* 63:465–474.

Miller, R., Halpern, J. (1982). Regression with censored data. *Biometrika* 69:521–531.

Peña, E., Rohatgi, V. T. (1987). Survival function estimation for a general proportional hazards model of random censorship. *J. Statist. Plann. Infer.* 22:371–389.

Shaked, M., Shanthikumar, J. G. (1994). *Stochastic Orders and Their Applications*. Boston: Academic Press, Inc.