Simultaneous Confidence Intervals for the Difference of Two Survival Functions

M. I. PARZEN
University of Chicago

L. J. WEI
Harvard University

Z. YING
Rutgers University

ABSTRACT. In comparing two treatments with failure time observations, confidence bands for the “difference” of two survival curves provide useful information about a global picture of the treatment difference over time. In this note, we propose a rather simple procedure for constructing such simultaneous confidence intervals. Our technique can also be used in the one-sample case, which has been extensively studied in the literature.

Key words: censoring; Kaplan–Meier estimate; Martingale

1. Introduction

In comparing two survival functions \(S_1(t)\) and \(S_2(t)\), one usually plots their Kaplan–Meier estimates \(\hat{S}_2(t)\) and \(\hat{S}_1(t)\), visually examines these curves, and subjectively summarizes how the difference \(D(t) = \hat{S}_2(t) - \hat{S}_1(t)\) varies over time. In this note, we show a rather simple technique for constructing purely non-parametric confidence bands of \(D(t)\). These bands provide an objective way to evaluate the “treatment” difference quantified by \(D(t)\) over time. Under the proportional hazards model assumption, asymptotic simultaneous confidence intervals for such a difference have been obtained by Dabrowska et al. (1989). Recently, with the general Cox regression model, confidence bands of the survival function of the patient with a specific set of covariates have been proposed by Lin et al. (1994).

For the one-sample case, various confidence bands for the survival function are derived, for example, by Gill (1980), Hall & Wellner (1980), Nair (1984), Csörgő & Horváth (1986), and Hollander & Peña (1989). Excellent reviews on this subject are given by Fleming & Harrington (1991, sec. 6.3) and Andersen et al. (1993, sec. IV.1.3). Our new technique can also be used to construct aforementioned confidence bands without referring to any special distribution table.

2. Simultaneous confidence intervals

Let \(X_j\) be the minimum of the failure time and censoring variable for the \(j\)th patient in the \(i\)th group; also, let \(A_i = 1\) if \(X_j\) is observed, and 0, otherwise, \(j = 1, \ldots, n_i, i = 1, 2\). Suppose that the process \(\{g(S_1(t), S_2(t))\}\) is used to quantify the “difference” between \(S_1(t)\) and \(S_2(t)\). For example, \(g(S_1(t), S_2(t)) = D(t) = S_2(t) - S_1(t)\), or is the logarithm of the ratio of two cumulative hazard functions. To obtain a confidence band for such a difference on an interval \([u_1, u_2]\), consider the following process:

\[
V(t) = n^{1/2} \tilde{v}(t) \{g(\hat{S}_2(t), \hat{S}_1(t)) - g(S_2(t), S_1(t))\},
\]

(2.1)

where \(n = n_1 + n_2\) and \(\{\tilde{v}(\cdot)\}\) is a weight function which converges uniformly to a deterministic
function \( \{v(t)\} \), in probability, on \([u_1, u_2]\). For example, \( \tilde{v}(t) \) may be the reciprocal of the estimated standard error of \( g(\tilde{S}_2(t), \tilde{S}_1(t)) \). If \( g \) is “smooth” enough, \( \{V(t)\} \) is asymptotically equivalent to the process

\[
n^{1/2}v(t)\{g_2(S_1(t), S_2(t))(\tilde{S}_2(t) - S_2(t)) + g_1(S_1(t), S_2(t))(\tilde{S}_1(t) - S_1(t))\},
\]

(2.2)

where \( g_1 \) is the partial derivative with respect to the \( i \)th argument, \( i = 1, 2 \). We assume that \( g_1(S_1(t), S_2(t)) \) converges uniformly to its theoretical counterpart, in probability, on \([u_1, u_2] \).

Now, the process: \( n^{1/2}(\tilde{S}_1(t) - S_1(t)) \), is asymptotically equivalent to the process

\[
U_i(t) = -n^{1/2}\tilde{S}_i(t)\int_0^n \left\{ \sum_{k=1}^n I(X_{ik} \geq s) \right\}^{-1} d\left\{ \sum_{j=1}^n M_j(s) \right\},
\]

(2.3)

where \( I(\cdot) \) is the indicator function,

\[
M_j(t) = I(X_{ij} \leq t; \Delta_{ij} = 1) - \int_0^n I(X_{ij} \geq s)dA_j(s),
\]

and \( A_i(\cdot) \) is the cumulative hazard function for patients in the \( i \)th group (see Fleming & Harrington, 1991, coroll. 3.2.1). Note that \( M_j(t) \) is a martingale in \( t \). Therefore, it follows from the martingale central limit theorem, (2.2) and (2.3), that the distribution of the process \( \{V(t); u_1 \leq t \leq u_2\} \) in (2.1) can be approximated by the distribution of a zero Gaussian process \( \{H(t), u_1 \leq t \leq u_2\} \). Here, we assume that \( u_2 \) satisfies the condition that \( \Pr(X_{i1} > u_2) > 0, i = 1, 2 \). Unfortunately, except for the case that \( S_1(\cdot) = S_2(\cdot) \) (see Fleming et al., 1980), distributional properties of \( \{H(t)\} \) are rather difficult to obtain analytically.

Now, we present a simple approximation to the distribution of \( \{V(t)\} \). Note that the mean and variance of \( M_j(t) \) are 0 and \( E\{I(X_{ij} \leq t \); \( \Delta_{ij} = 1\} \), respectively. Consider a process \( \tilde{U}_i(t) \) which is obtained from (2.3) by replacing \( \tilde{S}_i, X_{ij} \) and \( \Delta_{ij} \) with their observed values \( \tilde{S}_i, x_{ij} \) and \( \tilde{\delta}_{ij} \), and replacing \( M_j(s) \) with \( Z_j\{I(X_{ij} \leq s; \tilde{\delta}_{ij} = 1)\} \), where \( \{Z_{ij}, j = 1, \ldots, n_i, i = 1, 2\} \) is a random sample from the standard normal population. This results in

\[
\tilde{U}_i(t) = -n^{1/2}\tilde{S}_i(t)\sum_{j=1}^{n_i} \left\{ \sum_{k=1}^{n_i} I(X_{ik} \geq x_{ij}) \right\}^{-1} I(x_{ij} \leq t)\tilde{\delta}_{ij}Z_j).
\]

Note that the only random quantities in \( \tilde{U}_i \) are those \( Zs \). Although \( U_i \) and \( \tilde{U}_i \) are defined in different sample spaces, asymptotically they have the same first two moments. Furthermore, since both processes are tight, they have the same limiting distribution on the interval \([0, u_2]\), where \( \tilde{S}(u_2) > 0 \). It follows from (2.2) that the distribution of \( V(t) \) can be approximated by that of the process:

\[
\tilde{V}(t) = \tilde{v}(t)\{g_2(\tilde{S}_1(t), \tilde{S}_2(t))\tilde{U}_2(t) + g_1(\tilde{S}_1(t), \tilde{S}_2(t))\tilde{U}_1(t)\},
\]

on \([u_1, u_2]\), where \( \tilde{v} \) is the observed \( \tilde{v}(u_1) \) and \( g_i(\tilde{S}_1(u_1), \tilde{S}_2(u_2)) \) are bounded. Moreover, the distribution of \( \tilde{G} = \sup_{u_1 \leq u_2} |\tilde{V}(t)| \) can be approximated by that of \( G = \sup_{u_1 \leq u_2} |V(t)| \).

To derive a \( 1 - \alpha \) simultaneous confidence band \( B \) for \( g(S_1(t), S_2(t)) \), we first obtain a constant \( c(\alpha) \) such that \( \Pr(G > c(\alpha)) = \alpha \). Then,

\[
B = \left\{ g(\tilde{S}_1(t), \tilde{S}_2(t)) \pm n^{-1/2}c(\alpha)\tilde{v}^{-1}(t), u_1 \leq t \leq u_2 \right\}.
\]

(2.4)

In practice, \( c(\alpha) \) can be approximated by generating \( N \) realizations from \( \tilde{G} \). Each realization of \( \tilde{G} \) is obtained by generating a realization from \( \{Z_{ij}, i = 1, 2; j = 1, \ldots, n_i\} \) and computing the corresponding \( \sup_{u_1 \leq u_2} |\tilde{V}(t)| \). Note that the width of the simultaneous confidence band \( B \) depends on the interval \([u_1, u_2]\), which may be chosen not only based on statistical, but also on some biological considerations.
3. Examples

First, we use a recent AIDS clinical trial which evaluates the efficacy and safety of a reduced dose of AZT to illustrate our method (see Fischl et al., 1990). For this study, 524 patients who had an episode of Pneumocystis carinii pneumonia prior to their entry were randomly assigned to receive AZT in either a dose of 250 mg every four hours (standard group, Treatment 1) or a dose of 200 mg every four hours for four weeks and thereafter 100 mg taken every four hours (low-dose group, Treatment 2). Patients were enrolled into the trial between 2 December 1986 and 12 November 1987. A total of 429 deaths were observed during the study (1259 days). The corresponding Kaplan–Meier curves are given in Fig. 1. Suppose that we are interested in obtaining a confidence band of $D(t) = S_2(t) - S_1(t)$. Here, $n_1 = n_2 = 262$. To mimic the usual confidence interval procedure for $D(t)$ at a fixed time point $t$, we let $\hat{v}(t)$ be the reciprocal of the estimated standard error of $\{\hat{S}_2(t) - \hat{S}_1(t)\}$. To approximate the distribution of $\tilde{G}$, we generated $N = 500$ sets of $(Z_{ij}, i = 1, 2, j = 1, \ldots, 262)$. The resulting 0.95 confidence band for $D(t)$ is given in Fig. 2 with $u_1 = \inf \{t: \tilde{S}_1(t) < 1 \text{ or } \tilde{S}_2(t) < 1\} = 7$ and $u_2 = \sup \{t: \tilde{S}_1(t) > 0 \text{ and } \tilde{S}_2(t) > 0\} = 1036$. Note that for this illustration, we choose the largest possible interval: $[7, 1036]$, on which the confidence band $B$ is statistically valid. For this simultaneous interval, the estimated cutoff point $c(0.05)$ is 3.37. For comparison, we also plot the usual confidence interval for $D(t)$ in Fig. 2. That is, we replace $c(0.05)$ in (2.4) with 1.96. Based on Fig. 2, one may provide simultaneous confidence intervals for the differences of the two survival functions at any set of time points, which does not have to be prespecified in the analysis. For example, a 0.95 confidence band for $D(200), D(400), D(600)$ and $D(800)$ are $(-0.057, 0.088), (-0.049, 0.197), (-0.011, 0.276)$ and $(-0.107, 0.153)$, respectively. This indicates that the low dose of AZT is at least as good as the standard one with respect to patient’s survival. Based on this study, this specific low dosing scheme has now become the standard AZT mono-therapy for treating AIDS patients.

Note that our method can also be used to approximate the null distribution of the modified two-sample Smirnov test statistic $\sup_t \hat{v}(t)|\tilde{S}_2(t) - \tilde{S}_1(t)|$ studied by Fleming et al. (1980) for testing the equality of two survival functions. We find that generally our approximation to the

![Fig. 1. Kaplan–Meier estimates for standard vs. low dose treatment groups for AIDS example.](image)

above null distribution is quite close to the one provided by Fleming et al. We illustrate this point with the data from a Mayo Clinic study of patients having limited Stage II or IIA ovarian carcinoma. The ovarian cancer study is presented in Fleming et al. Its goal was to determine whether or not grade of disease was associated with time to progression of disease. The Kaplan–Meier estimates with respect to disease progression for patients with low-grade and high-grade ovarian carcinoma are displayed in Fig. 3. As in the first example, the resulting 0.95 confidence band for $D(.)$ (based on 500 resamples) is given in Fig. 4 with $u_1 = 28$ days and $u_2 = 462$ days.

---

**Fig. 2.** Pointwise and simultaneous confidence bands for the difference of survival functions between low and standard doses.

**Fig. 3.** Kaplan–Meier estimates of time to progression of disease for patients with low grade vs high grade ovarian carcinoma.
To estimate the p-value for the above Smirnov test, one may calculate the percentage of the 500 generated $\tilde{G}$s which are larger than the observed $\sup_{t \in [u_1, u_2]} \tilde{v}(t) \mid \tilde{S}_1(t) - \tilde{S}_2(t) \mid$. This results in a p-value of 0.003 while Fleming et al. (1980) reported a value of 0.002.

4. Remarks

An important question is how to choose an appropriate $N$ in simulating the random variable $\tilde{G}$ to obtain an estimate of the critical value $c(\alpha)$. In practice, one may try several values for $N$ until a stable estimate of $c(\alpha)$ is reached. For example, in the above AIDS example, we have used $N = 500$, 1000, and 5000 realizations to estimate the distribution of $\tilde{G}$. The resulting estimates for $c(\alpha)$ are all about 3.37. Regarding computing time, our method was programmed in S$^+$ and requires only a few seconds to compute the cut-off value on a 100 Mhz Pentium computer.

The proposed method can be easily used to construct a confidence band for a single survival function based on the process $n^{1/2} \tilde{v}(t) (\bar{S}(t) - S(t))$, where $S(\cdot)$ is the survival function, $\bar{S}(\cdot)$ is the corresponding Kaplan–Meier estimate, and $\tilde{v}(t)$ is an arbitrary weight function. For example, we apply our technique to the famous data set from a Mayo Clinic trial in primary biliary cirrhosis (PBC) of liver (see Fleming & Harrington, 1991, app. D) with $N = 500$ and $\tilde{v}(t)$ being the Hall–Wellner’s weight (Hall & Wellner, 1980). The resulting band is identical to the one given in Fig. 6.3.5(a) of Fleming & Harrington (1991).

According to the simulations done by Bie et al. (1987), pointwise and simultaneous confidence interval procedures for the survival function $S(\cdot)$ based on the large sample theory of the process $(\bar{S}(\cdot) - S(\cdot))$ are not satisfactory for small sample sizes. Transformations for $\bar{S}$ may be considered to improve the accuracy of the Gaussian approximation. Although there are no systematic studies on the best choice of such transformations in the literature, Bie et al. (1987) recommended two ad hoc transformations: the arcsin-square-root and the log-minus-log, for the censored survival data situation. For the two-sample problem considered here, it is not clear,
however, if there are simple transformations for the difference \( \left( \hat{S}_2(t) - \hat{S}_1(t) \right) \), which may improve the accuracy of the proposed confidence bands for small samples.

References


Received March 1996, in final form September 1996

M. I. Parzen, Graduate School of Business, University of Chicago, Chicago, IL 60637, USA.