Proc. Nat. Acad. Sci. USA
Vol. 73, No. 1, pp. 11–13, January 1976
Statistics

Bounds for a joint distribution function with fixed sub-distribution functions: Application to competing risks

(survival functions/censored data)

ARTHUR V. PETERSON, JR.*

Department of Statistics, Stanford University, Stanford, California 94305

Communicated by Jerzy Neyman, October 10, 1975

ABSTRACT **This paper gives sharp bounds for the joint survival function $G(t_1, t_2, \ldots, t_r) \equiv P(X_1 > t_1, X_2 > t_2, \ldots, X_r > t_r)$, and for the marginal survival functions $S_j(t) \equiv P(X_j > t)$, $j = 1,2, \ldots, r$, when the sub-survival functions $S_j^*(t) \equiv P(X_j > t, X_j = \min_{k=1,2,\ldots,r} X_k)$ are fixed. Theorem 1 gives the bounds for $r = 2$, and Theorem 2 gives the bounds for general r. Theorem 3 applies the result to the competing risks problem, and presents empirical bounds based on the observations. Finally, an example illustrates the bounds.**

1. Introduction

As recently pointed out by Tsiatis (1) and Peterson (2), serious errors can be made in estimating the (potential) survival functions in the competing risks problem if the risks are assumed to be independent when in fact they are not. Furthermore, there is no way of knowing from the data in the competing risks problem whether or not an error is being made, since the data contain no information on whether or not the risks are independent. This paper shows that, lacking any assumption about whether or not the risks are independent, the data can give only bounds for the survival functions.

Section 2 finds bounds for the joint survival function $G(\cdot,\cdot)$ of a bivariate random vector (X_1, X_2) with fixed sub-survival functions (sometimes called crude survival functions) $S_1^*(t) \equiv P(X_1 > t, X_1 < X_2)$ and $S_2^*(t) \equiv P(X_2 > t, X_2 < X_1)$. [This problem is similar to the one solved by Fréchet (3), and also by Hoeffding (4), except that here the sub-distribution functions are fixed, where in the problem considered by Fréchet and Hoeffding the marginal distribution functions are fixed.] For convenience it is assumed throughout that $P(X_1 = X_2) = 0$; hence $S_1^*(-\infty) + S_2^*(-\infty) = 1$.

Section 3 extends the results of Section 2 to the multivariate case. Section 4 applies the results of Section 3 to the problem of estimating survival functions in the competing risks problem. Section 5 illustrates the bounds with an example.

2. Bounds for bivariate distributions

Fix the sub-survival functions $S_1^*(t) \equiv P(X_1 > t, X_1 < X_2)$ and $S_2^*(t) \equiv P(X_2 > t, X_2 < X_1)$. The main theorem, which follows, gives bounds for the joint survival function $G(t_1, t_2) \equiv P(X_1 > t_1, X_2 > t_2)$ and bounds for the marginal survival functions $S_1(t) \equiv P(X_1 > t)$ and $S_2(t) \equiv P(X_2 > t)$.

THEOREM 1

1. A joint survival function $G(\cdot,\cdot)$ with fixed sub-survival functions $S_1^*(\cdot)$ and $S_2^*(\cdot)$ is bounded above and below as follows:

$$[S_1^* + S_2^*][\max(t_1, t_2)] \le G(t_1, t_2) \le S_1^*(t_1) + S_2^*(t_2). \quad [2.1]$$

The bounds are sharp.

2. The marginal survival functions $S_1(\cdot)$ and $S_2(\cdot)$ are bounded above and below as follows:

$$S_1^*(t) + S_2^*(t) \le S_1(t) \le S_1^*(t) + (1 - p_1), \quad [2.2]$$
$$S_1^*(t) + S_2^*(t) \le S_2(t) \le S_2^*(t) + (1 - p_2), \quad [2.3]$$

where

$$p_1 \equiv P(X_1 < X_2),$$
$$p_2 \equiv P(X_2 < X_1) = 1 - p_1.$$

The bounds are sharp.

Proof:
To prove [2.1], express the joint survival function $G(t_1, t_2)$ as the sum of two terms:

$$G(t_1, t_2) = P(X_1 > t_1, X_2 > t_2, X_1 < X_2) + P(X_1 > t_1, X_2 > t_2, X_2 < X_1). \quad [2.4]$$

Upper bounds for these terms are:

$$P(X_1 > t_1, X_2 > t_2, X_1 < X_2) \le P(X_1 > t_1, X_1 < X_2) \equiv S_1^*(t_1); \quad [2.5]$$
$$P(X_1 > t_1, X_2 > t_2, X_2 < X_1) \le P(X_2 > t_2, X_2 < X_1) \equiv S_2^*(t_2). \quad [2.6]$$

Lower bounds for the two terms of [2.4] are:

$$P(X_1 > t_1, X_2 > t_2, X_1 < X_2) \ge P[X_1 > \max(t_1, t_2), X_2 > \max(t_1, t_2), X_1 < X_2] = P[X_1 > \max(t_1, t_2), X_1 < X_2] \equiv S_1^*[\max(t_1, t_2)]; \quad [2.7]$$

$$P(X_1 > t_1, X_2 > t_2, X_2 < X_1) \ge P[X_1 > \max(t_1, t_2), X_2 > \max(t_1, t_2), X_2 < X_1] = P[X_2 > \max(t_1, t_2), X_2 < X_1] \equiv S_2^*[\max(t_1, t_2)]. \quad [2.8]$$

Using the bounds [2.5] through [2.8] in [2.4] yields the bounds stated in [2.1].

The bounds [2.2] and [2.3] for the marginals $S_1(t)$ and $S_2(t)$ follow immediately from [2.1] by setting t_2 and t_1, respectively, equal to $-\infty$.

* Present address: Department of Biostatistics, University of Washington, Seattle, Wash. 98195.

To show that the bounds [2.1] for the joint survival function $G(t_1, t_2)$ are sharp (and hence that the bounds [2.2] and [2.3] for the marginals are sharp), it suffices to show that for each bound there exists a joint probability distribution, having the specified sub-survival functions $S_1^*(\cdot)$ and $S_2^*(\cdot)$, that has (probability) mass arbitrarily close to the mass of the bound.

Such a joint probability distribution for the lower bound in [2.1] is the sum of the two sub-distributions that follow: (1) the sub-distribution that places mass at $(X_1, X_2) = (s, s + \epsilon)$ according to $S_1^*(s)$, and (2) the sub-distribution that places mass at $(X_1, X_2) = (s + \epsilon, s)$ according to $S_2^*(s)$, where $\epsilon > 0$ is arbitrarily close to zero.

Such a joint probability distribution for the upper bound in [2.1] is the sum of the two sub-distributions that follow: (1) the distribution that places mass at $(X_1, X_2) = (s, s + a)$ according to the distribution $S_1^*(s)$, and (2) the sub-distribution that places mass at $(X_1, X_2) = (s + a, s)$ according to the distribution $S_2^*(s)$, where $a > 0$ is arbitrarily large.

This completes the proof of *Theorem 1*.

The bounds [2.1] can be regarded as the most extreme models of dependency corresponding to fixed sub-survival functions. For a discussion of some other models of dependency that are less extreme, see ref. 5.

We make the following observations about the bounds [2.1] for the joint survival function $G(t_1, t_2)$: (1) For no $S_1^*(\cdot)$, $S_2^*(\cdot)$ do the bounds collapse for all t_1, t_2. (2) The bounds for $G(t_1, t_2)$ coincide for $t_1 = t_2$. (3) The lower bound is a joint probability distribution that places all mass on the line $t_1 = t_2$. (4) The upper bound is a joint probability distribution that places all mass on $(s, +\infty)$ and $(+\infty, s)$.

Also, observe the following about the bounds [2.2] for the marginal survival function $S_1(\cdot)$. [Similar statements can be made about the bounds [2.3] for $S_2(\cdot)$, since the problem is symmetric in X_1, X_2.] (1) The bounds [2.2] for $S_1(\cdot)$ collapse iff $p_1 \equiv P(X_1 < X_2) = 1$. (2) The difference function between the upper and lower bounds for $S_1(t)$ is $1 - p_1 - S_2^*(t)$. It is nondecreasing in t, and tends toward the zero function as $p_1 \to 1$. (3) The lower bound in [2.2] is the survival function of $\min(X_1, X_2)$. (4) The upper bound in [2.2] is a survival function that places mass according to $S_1^*(t)$, and mass equal to $1 - p_1$ at $t = +\infty$.

3. Extension to multivariate distributions

The extension of the bounds in *Theorem 1* to multivariate distributions is given in the next theorem. The proof uses the same methods as that for *Theorem 1*, and hence is omitted.

Define the joint r-variate survival function $G(t_1, t_2, \ldots, t_r)$ by $P(X_1 > t_1, X_2 > t_2, \ldots, X_r > t_r)$, and define for $j = 1,2, \ldots, r$ the sub-survival functions $S_j^*(t)$ by $P(X_j > t, X_j = \min_{k=1,2,\ldots,r} X_k)$.

THEOREM 2

1. A joint r-variate survival function $G(t_1, t_2, \ldots, t_r)$ with fixed sub-survival functions $S_j^*(\cdot)$, $j = 1,2, \ldots, r$, is bounded, sharply, above and below as follows:

$$[S_1^* + S_2^* + \cdots + S_r^*][\max(t_1, t_2, \ldots, t_r)] \leq G(t_1, t_2, \ldots, t_r) \leq S_1^*(t_1) + S_2^*(t_2) + \cdots + S_r^*(t_r). \quad [3.1]$$

2. The marginal survival functions $S_j(\cdot)$, $j = 1,2, \ldots, r$, are bounded, sharply, above and below as follows:

$$S_1^*(t) + S_2^*(t) + \cdots + S_r^*(t) \leq S_j(t) \leq S_j^*(t) + (1 - p_j) \quad [3.2]$$

where

$$p_j \equiv S_j(-\infty) = P(X_j = \min_{k=1,2,\ldots,r} X_k).$$

4. Application to the competing risks problem

In the competing risks problem we have n independent samples of the r-variate random vector $(X_1, X_2, \ldots, X_r)$, denoted here by $(X_{i1}, X_{i2}, \ldots, X_{ir})$, $i = 1,2, \ldots, n$. Observed for each $i = 1,2, \ldots, n$ are

$$Y_i \equiv \min(X_{i1}, X_{i2}, \ldots, X_{ir})$$

and

$$\delta_i \equiv j \text{ for which } Y_i = X_{ij}.$$

The problem is to estimate the survival functions $S_j(t) \equiv P(X_{ij} > t)$, $j = 1,2, \ldots, r$.

As pointed out by Tsiatis (1) and Peterson (2), the distribution of the observations (Y_i, δ_i), $i = 1,2, \ldots, n$ is determined by the sub-survival functions $S_j^*(\cdot)$, $j = 1,2, \ldots, n$, since $P(Y_i > t, \delta_i = j) \equiv S_j^*(t)$. The observations provide consistent estimators of $S_j^*(\cdot)$, $j = 1,2, \ldots, r$:

$$\hat{S}_j^*(t) \equiv \frac{1}{n}\sum_{i=1}^{n} I[Y_i > t, \delta_i = j].$$

Hence, consistent estimators for the bounds [3.1] for the joint survival function $G(\cdot, \cdot, \ldots, \cdot)$, and for the bounds [3.2] for the marginal survival functions $S_j(\cdot)$, $j = 1,2, \ldots, r$, can be obtained by substituting $\hat{S}_j^*(\cdot)$ for $S_j^*(\cdot)$ in [3.1] and [3.2]. This result is stated in the next theorem.

THEOREM 3

1. The following empirical bounds are consistent estimators of the bounds for the joint survival function $G(t_1, t_2, \ldots, t_r)$: Lower empirical bound: $[\hat{S}_1^* + \hat{S}_2^* + \cdots + \hat{S}_r^*][\max(t_1, t_2, \ldots, t_r)]$, Upper empirical bound: $\hat{S}_1^*(t_1) + \hat{S}_2^*(t_2) + \cdots + \hat{S}_r^*(t_r)$.
2. The following empirical bounds are consistent estimators of the bounds for the marginal survival functions $S_j(t)$, $j = 1,2, \ldots, r$: Lower empirical bound: $\hat{S}_1^*(t) + \hat{S}_2^*(t) + \cdots + \hat{S}_r^*(t)$, Upper empirical bound: $\hat{S}_j^*(t) + (1 - \hat{p}_j)$, where $\hat{p}_j \equiv \hat{S}_j^*(-\infty)$.

As a final note, the empirical bounds above are nonparametric maximum likelihood estimators of the bounds. This follows from the invariance principle for maximum likelihood estimators (e.g., ref. 6, pp. 222–223), and because the $\hat{S}_j^*(\cdot)$ are nonparametric maximum likelihood estimators of the $S_j^*(\cdot)$ (ref. 2).

5. An example with $r = 2$

Consider the bounds for $S_1(t)$ when the sub-survival functions are specified to be:

$$S_1^*(t) = \frac{\lambda}{\lambda + \mu} \exp\{-(\lambda + \mu)t\}, \; t \geq 0,$$

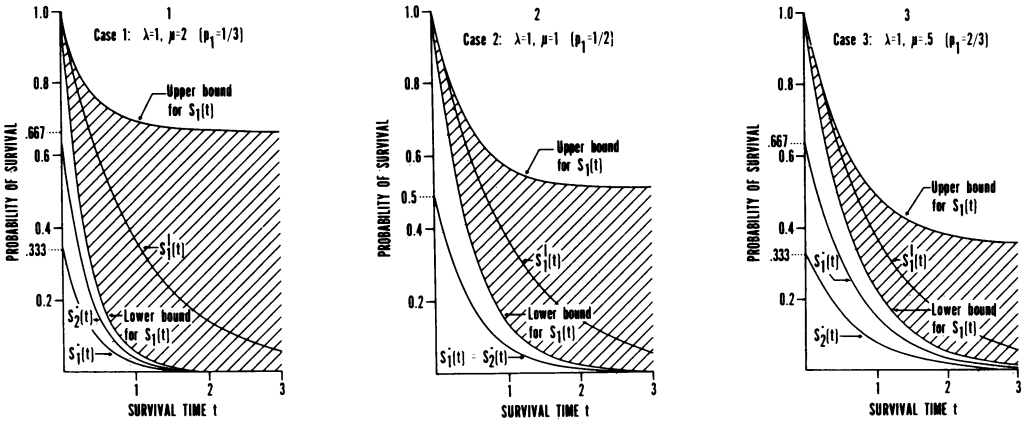
$$S_2^*(t) = \frac{\mu}{\lambda + \mu} \exp\{-(\lambda + \mu)t\}, \; t \geq 0.$$

Consider three cases:

Case 1: $\lambda = 1, \quad \mu = 2 \quad (p_1 = 1/3)$.

Case 2: $\lambda = 1, \quad \mu = 1 \quad (p_1 = 1/2)$.

Case 3: $\lambda = 1, \quad \mu = 0.5 \quad (p_1 = 2/3)$.



FIGS. 1, 2, AND 3. Bounds for the marginal survival function $S_1(t)$ with fixed sub-survival functions $S_1^*(t)$ and $S_2^*(t)$. Shown are the lower bound for $S_1(t)$, the upper bound for $S_1(t)$, and the survival function $[S_1^I(t)]$ under the assumption that the risks are independent.

The example was chosen so that the survival function of X_1 characterized by $S_1^*(\cdot)$ and $S_2^*(\cdot)$ under the assumption of independence, call it $S_1^I(\cdot)$, is the same $[S_1^I(t) = \exp\{-\lambda t\} = \exp\{-t\}, t \geq 0]$ for all three cases.

Figs. 1, 2, and 3 show the specified sub-survival functions $S_1^*(\cdot)$, $S_2^*(\cdot)$, the upper and lower bounds for the survival function $S_1(\cdot)$, and the survival function $S_1^I(\cdot)$ obtained if independence is assumed.

Note that the bounds are far apart when p_1 is small (Fig. 1) and are close together when p_1 is large (Fig. 3).

6. Concluding remarks

The bounds [3.1] and [3.2] show how misled one might be, at worst, in erroneously assuming independence of the risks. These bounds are applicable in the situation where the sub-survival functions are known, but no other information is known. To discover where between the bounds reality lies, additional information is needed. What additional information can be obtained in practical situations, and how much such information can contribute toward better bounds, are questions for further research.

The contents of this paper are a part of the author's Ph.D. dissertation, submitted at the Department of Statistics, Stanford University, Stanford, Calif. I thank Prof. Bradley Efron for his guidance and helpful advice. This research was supported by Public Health Service Grant 1 R01 GM21215-01.

1. Tsiatis, A. (1975) "A nonidentifiability aspect of the problem of competing risks," *Proc. Nat. Acad. Sci. USA* **72**, 20–22.
2. Peterson, A. V. (1975) "Nonparametric estimation in the competing risks problem," *Stanford Univ. Tech. Rep.*, No. 13.
3. Fréchet, M. (1951) "Sur les tableaux de correlation dont les margées sont données," *Ann. Univ. Lyon. Sect. A., Ser. 3*, **14**, 53–77.
4. Hoeffding, W. (1940) "Masstabinvariante Korrelations theorie," *Schr. Math. Inst. Univ. Berlin* **5**, 181–233.
5. David, H. A. (1974) "Parametric approaches to the theory of competing risks," in *Reliability and Biometry, Statistical Analysis of Lifelength*, eds. Proschan, F. & Serfling, R. J. (SIAM, Philadelphia, Pa.), pp. 275–290.
6. Zacks, S. (1971) *The Theory of Statistical Inference* (Wiley, New York).