Statistical Analysis of Competing Risks With Missing Causes of Failure

Isha Dewan\textsuperscript{1,3} and Uttara V. Naik-Nimbalkar \textsuperscript{2}
\textsuperscript{1}Indian Statistical Institute, New Delhi, INDIA
\textsuperscript{2}Department of Statistics University of Pune, Pune, INDIA
\textsuperscript{3}Corresponding Author : Isha Dewan isha@isid.ac.in

April 15, 2013

Abstract

In the competing risks model, a unit is exposed to several risks at the same time, but it is assumed that the eventual failure of the unit is due to only one of these risks, which is called a ‘cause of failure’. Thus the competing risks data consist of failure time and the cause of failure of each unit on test. Statistical inference procedures when the time to failure and the cause of failure are observed for each unit are well documented. In this paper we address the problem when the cause of failure may be unknown for some units. Several articles have proposed estimation of the survival or the sub-survival function in this situation. However the problem of testing whether the risks are equal or some risk dominates the other has not received much attention. We review some the estimation procedures and propose tests for the equality of the risks based on the sub-distribution, sub-survival functions and cause-specific hazard rates.

Key Words : Failure time, missing failure causes, Kaplan-Meier, cause-specific failure rate

1 Introduction

Consider a series system of $k$ components. When the system fails we know the system failure time $T$ and a random variable $\delta$ where $\delta = j$, $j = 1, 2, \ldots, k$ if the failure of the $j$th component led to system failure. Apart from reliability, such competing risks data also arise in survival analysis where $T$ is the failure time and $\delta$ the cause of failure of the patient, in economics where $T$ is the time spent in an unemployment register and $\delta$ indicating the type of job an individual gets, in social sciences where $T$ could be time of marriage and $\delta$ indicating the cause of end of marriage - say death or divorce. For other examples see Crowder (2001).
The joint distribution of \((T, \delta)\) can be expressed in terms of the sub-distribution function \(F(j, t)\) or the sub-survival function \(S(j, t)\) of the risk \(j, j = 1, \ldots, k\) which are defined, respectively, as

\[
F(j, t) = P[T \leq t, \delta = j], \quad (1.1)
\]
\[
S(j, t) = P[T > t, \delta = j]. \quad (1.2)
\]

The overall distribution function and the survivor function of the lifetime \(T\) are given by \(F(t)\) and \(S(t)\), respectively. Let \(f(j, t)\) denote the sub-density function corresponding to risk \(j\) and the density function of \(T\) be \(f(t)\). We know that

\[
F(t) = \sum_{j=1}^{k} F(j, t), \quad S(t) = \sum_{j=1}^{k} S(j, t), \quad f(t) = \sum_{j=1}^{k} f(j, t).
\]

The cause-specific hazard rates for \(j = 1, \ldots, k\) are defined as

\[
\lambda(j, t) = \frac{f(j, t)}{S(t)}.
\]

This is the probability of instantaneous failure of the unit due to \(j\)th cause given that the unit has survived time \(t\). One would want to know whether all "risks" are equally important for the unit. For example, if a series system fails due to the same component repeatedly, then one could improve on the reliability of that component so as to ensure better functioning of the system. In mathematical terms such hypotheses can be written in terms of the sub-distribution, sub-survival functions or cause-specific hazards as follows.

\[
H_{01} : F(1, t) = \ldots = F(k, t),
\]
\[
H_{02} : S(1, t) = \ldots = S(k, t),
\]
\[
H_{03} : \lambda(1, t) = \ldots = \lambda(k, t).
\]

Dewan and Deshpande (2005) noted the following result when there are 2 risks of failure.

**Theorem 1** : Under the null hypothesis of bivariate symmetry of hypothetical failure times due to two risks we have

(i) \(F(1, t) = F(2, t)\) for all \(t\),

(ii) \(S(1, t) = S(2, t)\) for all \(t\),

(iii) \(\lambda(1, t) = \lambda(2, t)\) for all \(t\),

(iv) \(P[\delta = 1] = P[\delta = 2]\),

(v) \(T\) and \(\delta\) are independent.

Thus, the following testing problems can be looked at:

\[
H_0 : F(1, t) = F(2, t) \text{ for all } t \quad \text{against} \quad H_1 : F(1, t) \leq F(2, t) \text{ with strict inequality for some } t
\]

\[
H_0 : S(1, t) = S(2, t) \text{ for all } t \quad \text{against} \quad H_2 : S(1, t) \leq S(2, t) \text{ with strict inequality for some } t
\]

\[
H_0 : \lambda(1, t) = \lambda(2, t) \text{ for all } t \quad \text{against} \quad H_3 : \lambda(1, t) \leq \lambda(2, t) \text{ with strict inequality for some } t.
\]

(1.3)
Under each of the three alternatives risk 2 is more ”effective” than risk 1 stochastically. The three hypotheses mentioned above are not equivalent, but $H_3$ implies both $H_1$ and $H_2$. Sometimes the sub-distribution functions cross may cross but the sub-survival functions could be ordered or vice-versa. These hypotheses are discussed by Carriere and Kochar (2000).

The experimenter may know only the failure time for the units under consideration. The cause of failure of some units may not be available. Kundu and Lapidus et al. (1994) observed that 40 percent of the death certificates of people who had died in motor accidents had no information on causes Dinsle (1982), Dewanji(1992), Goetghebeur and Ryan (1995), Dewanji and Sengupta (2003), Lu and Tsatis (2005) considered likelihood based estimation of the cause specific failure rates when causes of failure are unknown. Goetghebeur and Ryan (1990) considered a modified log rank test for competing risks with missing failure type. Miyawaka (1984) obtained MLE’s and MVUE’s of the parameters of exponential distribution for the missing case. Kundu and Basu (2000) discussed approximate and asymptotic properties of these estimators and obtained confidence intervals. For recent work on competing risks with missing data see Hyun et al (2012), Wang and Yu (2012), Yu and Li (2012), Sun et al (2012), Datta et al (2010) etc.

In what follows we consider a nonparametric test for testing $H_0$ against $H_3$ when we have information on time to failure $T_1, T_2, \ldots, T_n$ for all $n$ units but the cause of failure $\delta_1, \delta_2, \ldots, \delta_n$ may not be known. Let $O_i$ be an indicator variable which takes value one if $\delta_i$ is observed and zero if $\delta_i$ is missing. The indicator variables $O_1, \ldots, O_n$ are observed.

## 2 Test $H_0$ against $H_3$

We base the test on the following counting processes.

$$N_1^{(n)}(t) = \sum_{i=0}^{n} I[T_i \leq t, \delta_i = 1, O_i = 1], \quad N_2^{(n)}(t) = \sum_{i=0}^{n} I[T_i \leq t, \delta_i = 2, O_i = 1],$$

$$N_3^{(n)}(t) = \sum_{i=0}^{n} I[T_i \leq t, \delta_i = 1, O_i = 0], \quad N_4^{(n)}(t) = \sum_{i=0}^{n} I[T_i \leq t, \delta_i = 2, O_i = 0].$$

The corresponding intensity functions are $\alpha(j; t)Y^{(n)}(t)$, $i = 1, 2, 3, 4$, where $Y^{(n)}(t) = \sum_{i=1}^{n} I[T_i > t]$ and $\alpha(j; t)$ are the cause specific hazard functions. Let $A_1(t), A_2(t), A_{1,3}(t), A_{2,4}(t)$ and $A_{3,4}(t)$ denote the cumulative hazard functions corresponding to the counting processes $N_1^{(n)}, N_2^{(n)}, N_3^{(n)} = N_1^{(n)} + N_3^{(n)}, N_2^{(n)} = N_2^{(n)} + N_4^{(n)}$ and $N_3^{(n)} = N_3^{(n)} + N_4^{(n)}$. The processes $N_3^{(n)}$ and $N_4^{(n)}$ are not observable but their sum $N_{3,4}^{(n)}$ is.

Further suppose that $\beta(t) = P[\delta = 1|O = 0, T = t]$ is a known function. Our interest is in comparing $\lambda(1; t)$ and $\lambda(2; t)$, which are the cause specific hazard functions corresponding respectively to the processes $N_{1,3}^{(n)}$ and $N_{2,4}^{(n)}$. Then

$$\lambda(1; t) = \alpha(1; t) + \alpha(3; t), \quad \text{and} \quad \lambda(2; t) = \alpha(2; t) + \alpha(4; t).$$
Suppose $\alpha_{3,4}(t)Y^{(n)}(t)$ denotes the intensity function of the process $N_{3,4}^{(n)}(t)$, then we have,

$$\alpha(3; t) = \beta(t)\alpha_{3,4}(t) \quad \text{and} \quad \alpha(4; t) = (1 - \beta(t))\alpha_{3,4}(t).$$

Then the Nelson-Aalen estimators of $A_{1,3}(t)$ and $A_{2,4}(t)$ are, respectively,

$$\hat{A}_{1,3}(t) = \int_0^t \frac{1}{Y^n(s)}dN_1^n(s) + \int_0^t \frac{\beta(s)}{Y^n(s)}dN_{3,4}^n(s),$$
$$\hat{A}_{2,4}(t) = \int_0^t \frac{1}{Y^n(s)}dN_2^n(s) + \int_0^t \frac{1 - \beta(s)}{Y^n(s)}dN_{3,4}^n(s).$$

A test for the hypothesis $H_0 : \lambda(1; t) = \lambda(2; t)$ for all $t$ against $H_3$, is based on the statistics

$$S_{2n}(\tau) = \hat{A}_{2,4}(\tau) - \hat{A}_{1,3}(\tau),$$

where $\tau$ is some fixed time point. Under the null $S_{2n}(\tau)$ has zero mean. Large positive values of $S_{2n}$ show evidence against $H_0$.

Using the Doob-Meyer decompositions of the counting processes and the Rebelledo martingale central limit theorem (Andersen et al., 1993, p. 83) we obtain the following theorem.

**Theorem 2:** Suppose the subdistribution functions are absolutely continuous. As $n \to \infty$ and under $H_0 : \lambda(1, t) = \lambda(2, t)$, the process $\{\sqrt{n}S_{2n}(t), \ 0 \leq t \leq \tau\}$ converges weakly to a process $U = \{U(t), \ 0 \leq t \leq \tau\}$ in $D[0, \tau]$ with the Skorohod topology, where $\{U(t), 0 \leq t \leq \tau\}$ is a continuous zero-mean martingale and $\text{cov}(U(s), U(t)) = \sigma^2\min(t, s)$ with

$$\sigma^2(t) = \int_0^t \frac{\alpha(1, s) + \alpha(2, s) + (1 - 2\beta(s))^2\alpha_{3,4}(s)}{S(t)}ds.$$

Here $D[0, \tau]$ denotes the space of real valued functions on $[0, \tau]$ that are continuous from the right and have limits from the left everywhere. A consistent estimator for $\sigma^2(t)$ is given by

$$\sigma_{n}^2(t) = \int_0^t \frac{dN_1^n(s)}{Y^n(s)^2} + \int_0^t \frac{dN_2^n(s)}{Y^n(s)^2} + \int_0^t (1 - 2\beta(s))^2 \frac{dN_{3,4}^n(s)}{Y^n(s)^2}.$$ 

From the above results we have that

$$\frac{\sqrt{n}S_{2n}(\tau)}{\sigma_n(\tau)} \to N(0, 1), \quad \text{as} \quad n \to \infty.$$

Thus a value of $\frac{\sqrt{n}S_{2n}(\tau)}{\sigma_n(\tau)}$ larger than 1.64 shows evidence in favor of $H_3$ against $H_0$ at 5% level of significance.
3 Conclusion

We propose a test for testing the equality of two risks against the dominance of one of them. We have assumed that function $\beta(t)$ is known. In practice a form will have to be assumed for it.

A Kolmogorov-Smirnov type test for $H_0$ against $H_1$ can be obtained with the estimators of the sub-distribution functions based on the counting processes defined in the previous section.

Two other ways of handling missing causes of failure are to either ignore the missing data and use the available tests with reduced sample size or to use imputation procedures, that is generate the missing data. For the first case, the size of the sample is random and for the second case some distributional assumptions are needed.

4 References


