Statistical Analysis of Competing Risks With Missing Causes of Failure

Isha Dewan[1,3] and Uttara V. Naik-Nimbalkar [2]
[1]Indian Statistical Institute, New Delhi, INDIA
[2] Department of Statistics University of Pune, Pune, INDIA
[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 δ where $\delta = j,\ j = 1, 2, \ldots, k$ if the failure of the jth component led to system failure. Apart from reliability, such competing risks data also arise in survival analysis where T is the failure time and δ the cause of failure of the patient, in economics where T is the time spent in an unemployment register and δ indicating the type of job an individual gets, in social sciences where T could be time of marriage and δ indicating the cause of 'end" of marriage - say death or divorce. For other examples see Crowder (2001).

The joint distribution of (T, δ) 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

$$F(j,t) = P[T \leq t, \delta = j], \tag{1.1}$$

$$S(j,t) = P[T > t, \delta = j]. \tag{1.2}$$

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

$$\lambda(j,t) = f(j,t)/S(t).$$

This is the probability of instantaneous failure of the unit due to jth 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.

$$\begin{aligned} H_{01}: &\quad F(1,t) = \ldots = F(k,t), \\ H_{02}: &\quad S(1,t) = \ldots = S(k,t), \\ H_{03}: &\quad \lambda(1,t) = \ldots = \lambda(k,t). \end{aligned}$$

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 δ are independent.

Thus, the following testing problems can be looked at:

$$\begin{aligned} &H_0 : F(1,t) = F(2,t) \text{ for all t} \quad \text{against} \quad H_1 : F(1,t) \leq F(2,t) \text{ with strict inequality for some t} \\ &H_0 : S(1,t) = S(2,t) \text{ for all t} \quad \text{against} \quad H_2 : S(1,t) \leq S(2,t) \text{ with strict inequality for some t} \\ &H_0 : \lambda(1,t) = \lambda(2,t) \text{ for all t} \quad \text{against} \quad H_3 : \lambda(1,t) \leq \lambda(2,t) \text{ with strict inequality for some t.} \end{aligned} \tag{1.3}$$

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. individuals. The cause of failure of some units may not be available. Kodel and Chen (1987) considered an example from animal bioanalysis where all causes were not available. 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 Dinse (1982), Dewanji(1992), Goetghebeur and Ryan (1995) , Dewanji and Sengupta (2003) , Lu and Tsiatis (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 δ_i is observed and zero if δ_i is missing. The indicator variables $O_1, \cdots O_n$ are observed.

2 Test H_0 against H_3

We base the test on the following counting processes.

$$N_1^{(n)}(t) = \sum_{i=0}^{n} I[T_i \le t, \delta_i = 1, O_i = 1],\ N_2^{(n)}(t) = \sum_{i=0}^{n} I[T_i \le t, \delta_i = 2, O_i = 1],$$
$$N_3^{(n)}(t) = \sum_{i=0}^{n} I[T_i \le t, \delta_i = 1, O_i = 0],\ N_4^{(n)}(t) = \sum_{i=0}^{n} I[T_i \le t, \delta_i = 2, O_i = 0].$$

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 $N_1^{(n)}, N_2^{(n)}, N_{1,3}^{(n)} = N_1^{(n)} + N_3^{(n)}, N_{2,4}^{(n)} = N_2^{(n)} + N_4^{(n)}$ and $N_{3,4}^{(n)} = N_3^{(n)} + N_4^{(n)}$. 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

$$\lambda(1;t) = \alpha(1;t) + \alpha(3;t),\ \text{and}\ \lambda(2;\mathrm{t}) = \alpha(2;\mathrm{t}) + \alpha(4;\mathrm{t}).$$

Suppose $\alpha_{3,4}(t)Y^{(n)}(t)$ denotes the intensity function of the process $N_{3,4}^{(n)}(t)$, then we have,

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

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 τ 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 $\{\sqrt{n}S_{2n}(t),\ 0 \le t \le \tau\}$ converges weakly to a process $U = \{U(t),\ 0 \le t \le \tau\}$ in $D[0,\tau]$ with the Skorohod topology, where $\{U(t), 0 \le t \le \tau\}$ is a continuous zero-mean martingale and $cov(U(s), U(t)) = \sigma^2 min(t,s)$ 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

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

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

Anderson, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

Carriere, K. C. and Kochar, S.C. (2000). "Cmparing sub-survival functions in a competing risks model." *Lifetime Data Anal.*, **6**, 85-97.

Crowder, M.J. (2001). *Classical Competing Risks*. Chapman and Hall.

Datta, S., Bandyopadhyay , D. and Satten , G. A. (2010), "Inverse probability ofcensoring: weighted U-statistics for right-censored data with an application to testing hypotheses." *Scandinavian Journal of Statistics*, **37**, 680-700.

Dewan, I. and Deshpande, J.V. (2005)." Tests for some statistical hypotheses for dependent competing risks - A review."*Modern Statistical Methods in Reliability* eds Wilson, A. et al, 137-152, World Scientific, New Jersey.

Dewanji, A. (1992)." A note on a test for competing risks with general missing pattern in failure types." *Biometrics*, **79**, 855-857.

Dewanji, A.and Sengupta, D. (2003)." Estimation of competing risks with general missing pattern in failure types." *Biometrics*, **59** , 1063-1070.

Dinse, G.E. (1986). "Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data." *J. Amer. Statist. Assoc.*, **81**, 328-336.

Goetghebeur, E. and Ryan, L. (1990). "A modified log rank test for competing risks with missing failure type." *Biometrika* , **77**, 207-211.

Goetghebeur,E. and Ryan, L. (1995). "Analysis of competing risks survival data when some failure types are missing." *Biometrika* , **82**, 821–833.

Hyun, S., Lee, J. and Sun, Y. (2012) " Proportional hazards model for competing risks data with missing cause of failure". *J. Statist. Plann. Inference* , **142**, 1767-1779.

Kochar, S.C. (1995). "A review of some distribution-free tests for the equality of cause specific hazard rates." *Analysis of Censored Data* , ed Koul, H.L. and Deshpande, J.V., IMS, Hayward, 147-162.

Kodel, R.K. and Chen, J.J. (1987). Handling cause of death in equivocal cases using the EM algorithm." *Comm. Stats. - Theory and Methods*, **16**, 2565-2603.

Kundu, D. and Basu, S. (2000). "Analysis of incomplete data in presence of competing risks. " *J. Statist. Plann. Inference* , **87**, 221-239.

Lapidus, G. Braddock, M. , Schwartz , R. Banco, L. and Jacobs, L. (1994)." Accuracy of fatal motorcycle injury reporting on death certificates. *Accident Anal. Prevention* , **26**, 535-542.

Lu, K. and Tsiatis, A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. " *Lifetime Data Anal.*, **11**, 29-40.

Miyakawa, M. (1984). "Analysis of incomplete data in competing risks model. " *IEEE Transactions in Reliability*, **33**, 293-296.

Wang, Jiaping; Yu, Q. (2012) " Consistency of the generalized MLE with interval-censored and masked competing risks data." *Comm. Statist. Theory Methods*, **41**, 4360-4377.

Yu, Q. and Li, J. (2012). "The NPMLE of the joint distribution function with right-censored and masked competing risks data." *J. Nonparametr. Stat.*, **24**, 753-764.

Sun, Y., Wang, H. J. and ; Gilbert, P. B. (2012). "Quantile regression for competing risks data with missing cause of failure." *Statist. Sinica*, **22**, 703-728.