*Journal of Statistical*
*Theory & Methods*
*1998, Vol. 9, No. 1, pp. 37 ~ 45*

# 임의중도절단된 자료에서 생존함수의 동시신뢰대 구성

**이원기 · 송명언 · 송재기[1] · 박희주[2]**

**Abstract**

임의중도절단된 생존시간자료에서 생존함수에 대한 동시신뢰대를 근사식이나 표없이 구성하는 간단한 방법을 제안하였다. 그리고 모의실험을 통하여 기존의 동시신뢰대와 포함확률측면에서 서로 비교하고, 실제자료에 적용하여 보았다.

## 1. 서 론

질병의 발병, 사망 또는 기계의 고장 등의 임의중도절단된 생존시간자료에서 생존함수와 누적위험함수에 대해 오랫동안 많은 연구가 이루어져 오고 있다. Kaplan과 Meier(1958)는 생존함수에 대한 승극한(product-limit)추정량(이하 K-M추정량)을, Nelson(1969, 1972)과 Aalen(1975, 1978)은 누적위험함수에 대한 Nelson과 Aalen 추정량(이하 N-A추정량)을 제안하였으며, 이 추정량들은 많은 분야에서 폭넓게 사용되어지고 있다.

또한 생존함수와 누적위험함수에 대한 동시신뢰대도 많은 연구가 진행되어져 오고 있다. Hall과 Wellner(1980)는 K-M추정량을 이용하여 완전한 자료에 대한 Kolmogorov의 동시신뢰대를 임의중도절단된 자료로 확장하여 생존함수에 대한 동시신뢰대를 제안하였고(이하 HW 동시신뢰대), Nair(1984)는 생존함수에 대한 신뢰구간을 확장하여 생존함수에 대한 동시신뢰대을 구성하였다(이하 EP 동시신뢰대). 또한 Bie, Borgan, 그리고 Liestøl(1987)은 N-A추정량과 적절한 변수변환을 이용하여 누적위험함수에 대한 여러가지 동시신뢰대를 구하였으며, 그 결과 변수변환을 이용한 동시신뢰대가 소표본에서 더 좋다는 것을 보였다. Borgan과 Liestøl(1990)은 생존함수에 대한 기존의 여러 동시신뢰대들을 비교한 결과 소표본인 경우 Nair가 제안한 동시신뢰대는 포함확률(coverage probability)이 목적수준(target level)을 만족시키지 못한다는 것을 보이면서 적절한 변수변환을 통하여 동시신뢰대를 구성할 것을 권장하였다.

[1]701-702 대구광역사 북구 산격동 1370 경북대학교 통계학과
[2]712-701 경북 경산시 하양읍 부호리 33 경일대학교 컴퓨터공학과

그러나 이러한 동시신뢰대들을 구성할 때 필요한 임계치(critical point)는 복잡한 근사식을 이용하여 구하여야 하며, Chung(1986)은 많이 사용되어지는 임계치를 근사적으로 구하여 표를 만들었다.

본 논문에서는 근사식이나 특별한 표 없이 쉽게 생존함수의 동시신뢰대를 구성하는 방법을 제안하고, 모의실험을 통하여 기존의 동시신뢰대와 포함확률측면에서 서로 비교하고자 한다. 또한 실제 자료를 제안된 동시신뢰대와 기존의 여러 동시신뢰대에 적용해 보고자 한다.

## 2. 동시신뢰대의 구성

생존시간 $T_1, T_2, \cdots, T_n$은 서로 독립이고 연속인 분포함수 $F$를 따르는 확률변수이고, 임의중도절단시간 $C_1, C_2, \cdots, C_n$ 또한 서로 독립이며 연속인 분포함수 $G$를 따르는 확률변수이며, $F$와 $G$는 서로 독립이라 가정한다. 이때 우리가 관찰할 수 있는 자료는 $(X_i, \Delta_i)$이며, 여기서 $X_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$이다.

셈과정(counting process) $N(t)$는 $t$ 시각까지의 중도절단되지 않은 관찰치의 수로서

$$N(t) = \sum_{i=1}^{n} I(X_i \leq t, \Delta_i = 1)$$

이며, 승법강도모형(multiplicative intensity model) $\lambda(t) = Y(t)\alpha(t)$하에서

$$M(t) = N(t) - \int_0^t \alpha(s)Y(s)ds$$

는 지역제곱적분가능 마팅게일 (local square integrable martingale)이 된다. 여기서 $Y(t) = \sum_{i=1}^{n} I(X_i \geq t)$이며, $\alpha(t)$는 위험함수이다.

이제 신뢰대의 구성을 위하여 다음과 같은 과정 $U(t)$을 고려하자.

$$U(t) = \sqrt{n}\left(\widehat{A}(t) - A(t)\right)$$

여기서 $A(t) = \int_0^t \alpha(s)ds$는 누적위험함수이고, $\widehat{A}(t)$는 N-A추정량으로

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y(s)} dN(s),$$

이며, 여기서 $J(s) = I(Y(s) > 0)$이다. 그러면 정칙가정(regularity conditions)하에서 과정 $U(t)$는 평균이 0인 가우스과정으로 분포수렴하며, 또한 다음의 $\widetilde{U}(t)$와 같은 극한분포를 갖는다는 것이 알려져 있다.

$$\begin{aligned} \widetilde{U}(t) &= \sqrt{n}(\widehat{A}(t) - A^*(t)) \\ &= \sqrt{n}\int_0^t \frac{J(s)}{Y(s)} dM(s) \end{aligned}$$

여기서 $A^*(t) = \int_0^t \alpha(s)J(s)ds$이다. 그러므로 과정 $\widetilde{U}(t)$의 극한분포로부터 생존함수의 동시신뢰대의 구성에 필요한 임계치를 구하고자 한다. 그러나 과정 $\widetilde{U}(t)$는 마팅게일을 포함하고 있으므로 $\widetilde{U}(t)$의 극한분포로부터 직접적으로 임계치를 구할 수 없다. 그래서 본 논문에서는 Lin, Wei와 Ying(1993)의 방법을 이용하여 다음과 같이 임계치를 구하고자 한다. 마팅게일 $M(t)$는 임의의 $t$에 대하여 $E[M(t)] = 0$, $Var[M(t)] = E[N(t)]$이므로 $\widetilde{U}(t)$에서 $M(t)$를 평균과 분산이 같은 $N(t)$ $G$로 교체하면 다음과 같은 과정 $\widehat{U}(t)$을 얻을 수 있다.

$$\begin{aligned}\widehat{U}(t) &= \sqrt{n}\int_0^t \frac{J(s)}{Y(s)} dN(s)\ G \\ &= \sqrt{n}\sum_{i=1}^{n} \frac{J(X_i)}{Y(X_i)}\ I(X_i \le t)\ \Delta_i\ G_i\end{aligned}$$

여기서$\{G_i\ ;\ i = 1, \cdots, n\}$는 표준정규 모집단에서 생성한 난수이다. 그러면 관찰자료 $\{X_i, \Delta_i\}$가 주어졌을 때 과정 $\widehat{U}(t)$에서 $G_i$만 유일한 확률변수이므로 과정 $\widehat{U}(t)$는 서로독립인 확률변수의 합으로 표현되어 진다. 그러므로 $U(t)$의 극한분포는 관찰자료 $\{X_i, \Delta_i\}$가 주어졌을때 $\widehat{U}(t)$의 조건부 극한분포와 같게 된다(Lin 등(1993), Lin 등(1994), Song 등(1997) 등 참조). 또한 표준정규 모집단으로부터 $G_i$를 반복하여 생성하면 $\widehat{U}(t)$의 극한분포를 근사적으로 구할 수 있으며, 이를 이용하여 동시신뢰대의 구성에 필요한 임계치를 구할 수 있다.

생존함수 $S(t)$의 $100(1-\alpha)\%$ 동시신뢰대의 구성을 위해 다음과 같은 과정 $B(t)$을 고려하자.

$$B(t) = \sqrt{n}\ g(t)\ \left[\phi\{\widehat{A}(t)\} - \phi\{A(t)\}\right]$$

여기서 $g(t)$는 비음의 유계함수(bounded function)로 일양적으로 수렴하는 가중함수(weight function)이며, 평활함수 $\phi(t)$는 구간 $[t_1,\ t_2]$에서 일차도함수 $\phi'(t)$가 연속이고 0이 아닌 값을 가지는 알려진 함수이다. 델타-방법(delta-method)에 의해 과정 $B(t)$는 과정 $\widetilde{B}(t)$와 근사적으로 같게 된다.

$$\widetilde{B}(t) = g(t)\ \phi'\ \{A(t)\}\ U(t)$$

또한 앞의 결과로부터 $A(t)$, $U(t)$를 $\widehat{A}(t)$, $\widehat{U}(t)$로 교체한 $\widehat{B}(t)$는 $\widetilde{B}(t)$와 근사적으로 같은 분포를 같게 된다.

$$\widehat{B}(t) = g(t)\ \phi'\ \{\widehat{A}(t)\}\ \widehat{U}(t)$$

그러므로 $\sup\{B(t)\ ;\ t_1 \le t \le t_2\}$ 분포의 임계치 $q_\alpha$는 다음의 식을 만족하는 값으로 근사적으로 구할 수 있다.

$$Pr\left\{\max_{t_1 \le X_j \le t_2} \left|\widehat{B}(X_j)\right| > q_\alpha\right\} = \alpha$$

이것으로부터 시간구간 $[t_1,\ t_2]$에서 $\phi\{A(t)\}$의 $100(1-\alpha)\%$ 동시신뢰대는

$$\phi\{\widehat{A}(t)\} \mp \frac{1}{\sqrt{n}} \frac{q_\alpha}{g(t)}$$

이 되며, 여기서 평활함수 $\phi(x) = logx$라 두고 $HW$형태와 $EP$형태의 동시신뢰대을 구성하기 위하여 아래와 같은 가중함수 $g_j(t)$를 고려하였다.

$$g_1(t) = \frac{\widehat{A}(t)}{1+\widehat{\sigma}^2(t)},\ \ g_2(t) = \frac{\widehat{A}(t)}{\widehat{\sigma}(t)}$$

이들 $g_j(t)$를 대입하여 각 형태의 신뢰대 구성에 필요한 임계치 $q_{j,\alpha}$는 아래와 같은 $\widehat{B_j}(t)$로부터 얻을 수 있다.

$$\widehat{B_1}(t) = \frac{\widehat{U}(t)}{1+\widehat{\sigma}^2(t)},\ \ \widehat{B_2}(t) = \frac{\widehat{U}(t)}{\widehat{\sigma}(t)}$$

이렇게 얻은 임계치와 $g_j(t)$를 이용하여 생존함수 $S(t)$의 동시신뢰대들을 아래와 같이 구성할 수 있다. 먼저 $g = g_1$인 경우

$$\widehat{S}(t)^{exp\{\mp \frac{1}{\sqrt{n}}\ q_{1,\alpha}\widehat{S}(t)(1+\widehat{\sigma}^2(t))\}}$$

으로 $log$ 변환된 $HW$ 형태의 동시신뢰대가 되며, $g = g_2$인 경우

$$\widehat{S}(t)^{exp\{\mp \frac{1}{\sqrt{n}}\ q_{2,\alpha}\widehat{S}(t)\widehat{\sigma}^2(t)\}}$$

으로 $log$ 변환된 $EP$ 형태의 동시신뢰대가 된다. 또한 변환이 없는 $HW$형태와 $EP$형태의 신뢰대를 구성하기 위해서는 $\phi(x) = e^{-x},\ \ g(t) = \frac{1}{1+\widehat{\sigma}^2(t)}$ 그리고 $\phi(x) = e^{-x},\ \ g(t) = \frac{1}{\widehat{\sigma}^2(t)}$로 주면 되며 그 결과는 각각 아래와 같다.

$$\widehat{S}(t) \mp \frac{1}{\sqrt{n}}\ q_{1,\alpha}\ \widehat{S}(t)\ \{1+\widehat{\sigma}^2(t)\},$$

$$\widehat{S}(t) \mp \frac{1}{\sqrt{n}}\ q_{2,\alpha}\widehat{S}(t)\widehat{\sigma}^2(t)$$

## 3. 모의실험 및 예제

이 절에서는 제안된 동시신뢰대를 기존의 다른 방법들과 비교하기 위하여 몬테칼로 모의실험을 하였다. 모의실험에서 생존시간의 분포는 와이블분포, 중도절단시간의 분포는 지수분포로 하여 중도절단율을 10%, 30%가 되도록 모수를 조정하였고 표본의 수는 소표본에서의 성적을 보기위하여 20, 30, 50, 100을 선택하였으며 반복수는 1000번을 하였다. 또

**표 1.** 각 분포와 중도절단률에 따른 동시신뢰대의 포함확률($\alpha = 0.05$)

| 분포 | 표본크기 | CR(%) | HW | LWY HW | THW | LWY THW | EP | LWY EP | TEP | LWY TEP |
|---|---|---|---|---|---|---|---|---|---|---|
| Weibull (1, 0.5) | 20 | 10% | .936 | .913 | .918 | .939 | .862 | .900 | .902 | .895 |
| | | 30% | .935 | .906 | .918 | .936 | .868 | .883 | .907 | .881 |
| | 30 | 10% | .958 | .914 | .940 | .941 | .903 | .924 | .924 | .909 |
| | | 30% | .947 | .906 | .939 | .919 | .900 | .884 | .918 | .905 |
| | 50 | 10% | .960 | .924 | .940 | .946 | .928 | .928 | .933 | .917 |
| | | 30% | .959 | .924 | .951 | .945 | .942 | .921 | .933 | .910 |
| | 100 | 10% | .971 | .948 | .963 | .954 | .961 | .949 | .958 | .934 |
| | | 30% | .963 | .946 | .956 | .951 | .952 | .931 | .949 | .936 |
| Weibull (1, 1.0) | 20 | 10% | .942 | .911 | .919 | .945 | .854 | .903 | .905 | .897 |
| | | 30% | .945 | .909 | .910 | .929 | .835 | .851 | .905 | .877 |
| | 30 | 10% | .960 | .918 | .944 | .947 | .887 | .905 | .928 | .918 |
| | | 30% | .945 | .917 | .936 | .933 | .877 | .867 | .918 | .900 |
| | 50 | 10% | .963 | .927 | .950 | .952 | .922 | .922 | .935 | .916 |
| | | 30% | .959 | .932 | .946 | .949 | .922 | .894 | .936 | .918 |
| | 100 | 10% | .971 | .952 | .954 | .954 | .953 | .941 | .955 | .938 |
| | | 30% | .959 | .947 | .948 | .951 | .937 | .910 | .946 | .927 |
| Weibull (1, 2.0) | 20 | 10% | .939 | .916 | .915 | .948 | .846 | .891 | .902 | .889 |
| | | 30% | .951 | .918 | .924 | .940 | .843 | .868 | .902 | .880 |
| | 30 | 10% | .956 | .915 | .940 | .940 | .883 | .891 | .919 | .907 |
| | | 30% | .956 | .912 | .933 | .935 | .864 | .873 | .905 | .891 |
| | 50 | 10% | .959 | .923 | .948 | .948 | .913 | .908 | .936 | .923 |
| | | 30% | .964 | .932 | .952 | .946 | .902 | .885 | .933 | .916 |
| | 100 | 10% | .966 | .954 | .952 | .957 | .953 | .934 | .948 | .935 |
| | | 30% | .970 | .951 | .958 | .951 | .944 | .921 | .943 | .927 |

한 제안된 방법에서 $\sup\{B(t)\ ;\ t_1 \le t \le t_2\}$의 임계치를 얻기 위하여 표준정규 모집단의 임의표본 $\{G_i,\ i = 1, 2, \cdots, n\}$는 1000번 반복하여 생성하였으며 기존의 방법을 위해서는 Chung(1986)이 구해놓은 표를 이용하였다. 동시신뢰대의 하한치 $t_1$은 임의표본 중 10백분위수를, 상한치 $t_2$는 90백분위수를 사용하였으며, 모의실험 결과는 표1에 주어져 있다.

표1을 살펴보면 $HW$형태의 신뢰대에서는 제안된 방법의 신뢰대가 기존의 신뢰대보다 포함확률 측면에서 못하나 $log$변환한 $HW$형태의 신뢰대에서는 더 낫다는 것을 알 수 있었으며, $EP$형태의 신뢰대에서는 제안된 방법의 신뢰대가 기존의 신뢰대보다 포함확률에서는 좋았으나 $log$변환한 $EP$형태의 신뢰대에서는 약간 못함을 알 수 있다.

실제 자료의 적용을 위한 예제자료는 Copelan 등(1991)이 Ohio주립대학등 네 개 대학병원에서 1984년 3월 1일부터 1989년 6월 30일까지 치료를 받은 백혈병환자 137명의 자료 중 일부를 이용하였다. 137명의 전체 환자중 38명은 ALL(Acute Lymphoblastic Leukemia)환자로 이들에 대하여 각 병원은 Busulfan 16mg/kg복용과 Cyclophasphamide 120mg/kg 정맥주사 치료법을 사용하였으며 약 37%의 자료가 중도절단되었다. 이들 38명의 자료를 이용하여 생존함수의 동시신뢰대를 구성한 후 그림1 -그림4에 나타내었다. 이때 하한치 $t_1$은

4번째 자료를, 상한치 $t_2$는 26번째 자료를 사용하였다. 그림1 - 그림4를 살펴보면 $HW$형태의 신뢰대는 기존의 신뢰대와 비교하여 다소 다른 형태를 띄고 있으나 $EP$형태는 거의 유사하게 보인다.

모의실험의 결과와 실제자료에 적용해 본 결과 $EP$형태의 신뢰대 구성에서는 제안된 방법에 의한 신뢰대의 구성이 기존의 방법보다 포함확률과 신뢰대 영역, 그리고 신뢰대 구성의 편리함을 고려해볼 때 좋다는 것을 알 수 있었다.

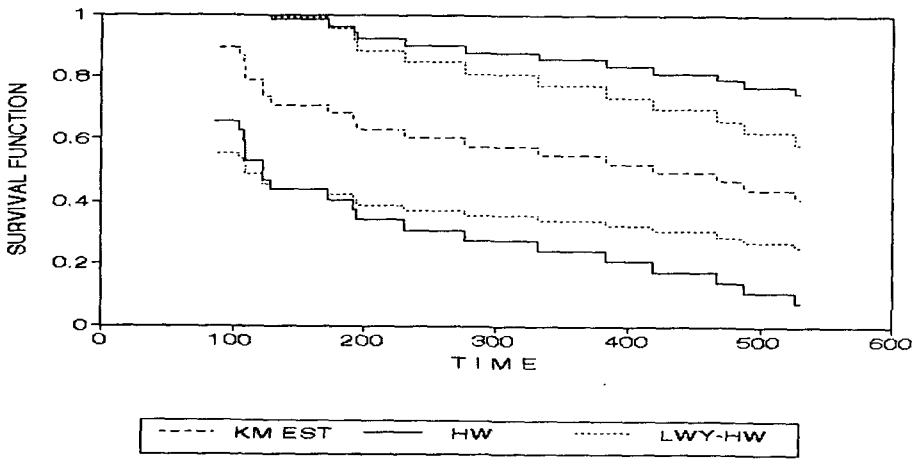


**그림 1. HW 신뢰대와 LWY 방법의 HW 신뢰대**

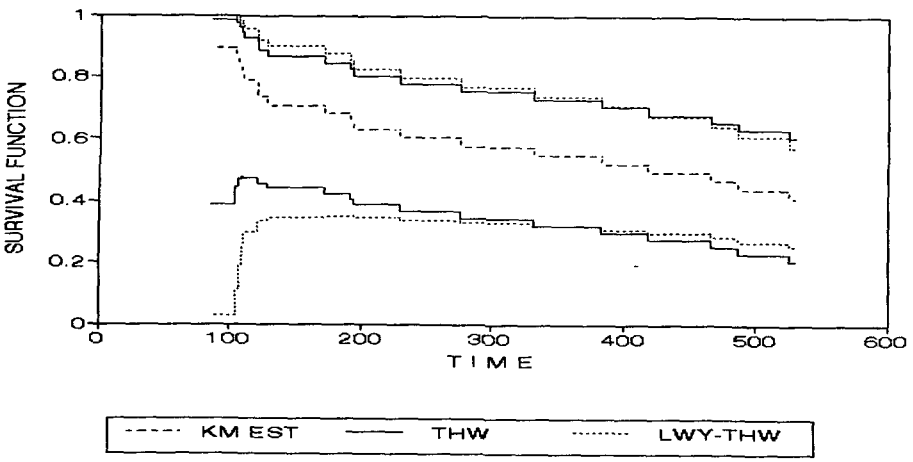


**그림 2. Log 변환한 HW 신뢰대와 Log 변환한 LWY 방법의 HW 신뢰대**

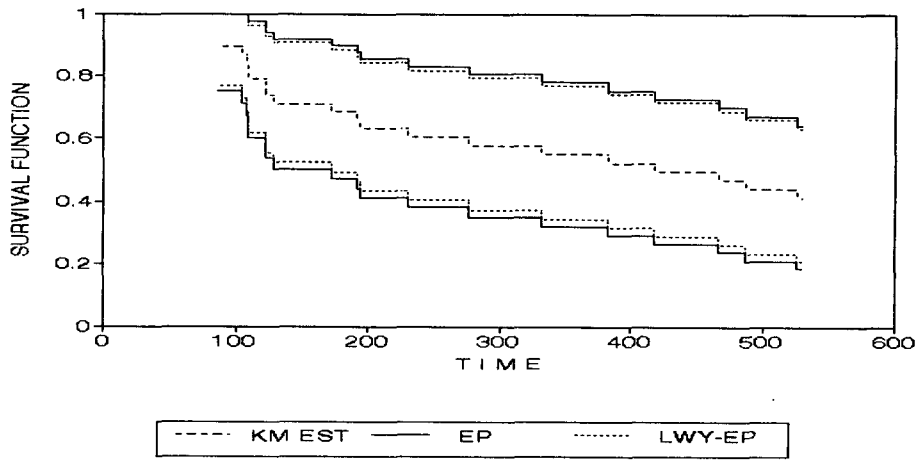


**그림 3. EP 신뢰대와 LWY 방법의 EP 신뢰대**

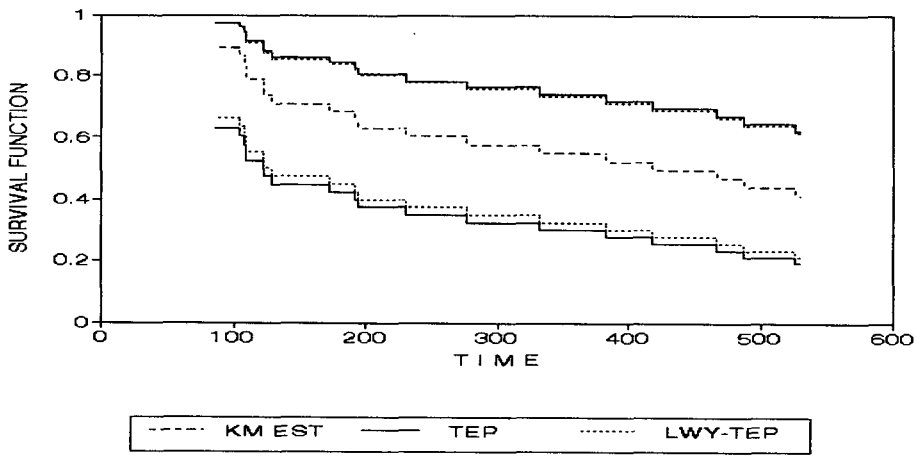


**그림 4. Log 변환한 EP 신뢰대와 Log 변환한 LWY 방법의 EP 신뢰대**

## 참 고 문 헌

1. Aalen, O.O., (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, **6**, 701-726.

2. Bie, O., Borgan,Ø., and Liestøl, K. (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, **14**, 221-233.

3. Borgan,Ø., and Liestøl, K. (1990). A note on confidence intervals and bands for the survival curve based on transformations. *Scandinavian Journal of Statistics*, **17**, 35-41.

4. Chung, C.F. (1986). Formulae for probabilities associated with Wiener and Brownian bridge processes. Technical Report **79**, Laboratory for Research in Statistics and Probability, Carleton University. Ottawa, Canada.

5. Copelan, E.A., Biggs, J.c., Thompson, J.M., Crilley, P., Szer, J., Klein, J.P., Kapoor, N., Avalos, B.R., Cunningham, I., Atkinson, K., Downs, K., Harmon, G.S., Daly, M.B., Brodsky, I., Bulova, S.I., and Tutschka, P.J. (1991). Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with Bu/Cy. *Blood*, **78**, 838-843.

6. Hall, W.J., and Wellner, J.A. (1980). Cofidence bands for a survival curve from censored data. *Biometrika*, **67**, 113-143.

7. Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457-481.

8. Lin, D.Y., Wei, L.J., and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, **80**, 557-572.

9. Lin, D.Y., Fleming, T.R., and Wei, L.J. (1994). Confidence Bands for survival curve under the proportional hazards model. *Biometrika*, **81**, 73-81.

10. Nair, V.N. (1984). Cofidence bands for survival functions with censored data: A comparative study. *Technometrics*, **14**, 265-275.

11. Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, **14**, 945-965.

12. Song, M.U., Jeong, D.M., and Song, J.K. (1997). Confidence Bands for survival curve under the additive risk model. *Journal of the Korean Statistical Society*, **26**, 429-443.

# The Confidence Bands for the Survival Function in Random Censorship Model

Won Kee Lee · Myung Unn Song · Jae Kee Song · Hee Joo Park

**Abstract**

We consider the problem of obtaining the confidence bands for the survival function with incomplete data. It is a rather simple procedure for constructing confidence bands of survival function. This method uses the weak convergence of normalized cumulative hazard estimator to a mean zero Gaussian process whose distribution can be easily approximated through simulation. Finally, we compare the performance of the proposed confidence bands through Monte Carlo simulation and are applied to construct the proposed bands with the Leukemia patient data.