A NEW TEST FOR TWO SAMPLES WITH CENSORED DATA[+]

Saffa Ali Nasir [*]

Abstract

A two sampl test about distribution of censored data is proposed. The regularity conditions used for deriving asymptotic distribution of the test statistis are based mainly on assumptions about smoothness of the underlying distribution functions.

المستخلص

في هذا البحث تم اقتراح اختبار جديد لتوزيع البيانات المراقبة(censored data) لعينتين ، واشـــتقاق التوزيع التقاربي لاحصائية الاختبار التي تعتمد على فرضيات التقارب (weak convergence) للعمليات التصادفية اذ تم برهنة ان التوزيع الخاص باحصائية الاختبار يقترب من التوزيع الطبيعي القياسي (المعياري) اذ تـــتم مقارنـــة القيمة المحسوبة للاختبار مع قيمة (Z) المجدولة لاختبار الفرضية المطلوبة . كما تم اعتماد مثال تطبيقي يمثل تجربة لتوضيح هذا الاختبار ومقارنته مع الاختبارات المناظرة الاخرى .

Introduction:-

In many situations when the lifetime of the items is under the study, the true lifetimes may not be directly observable, and we have to deal with some kind of censoring.

Suppose we have two independent random samples $X_1^o, X_2^o, \ldots\ldots, X_n^o$ and $y_1^o, y_2^0, \ldots\ldots y_m^o$ with continuous distribution function $F_o(x)$ and $G_o(y)$ respectively. For the first sample, the censoring lifetime U_1, U_2, … U_n independent of $X_1^0, \ldots\ldots, X_n^0$ is independent of identically distributed (iid) random variables with a continuous distribution function $H_1(u)$ and we observe the pairs (X_i, d_i), i=1, …., n where

$$X_i = Min\ (X_i^0, U_i), d_i = I(X_i^o \le U_i) = \begin{cases} 1 & X_i \le Ui \\ 0 & X_i > Ui \end{cases}$$

i.e d_i =1 if and only if X_i is uncensored. Similarly for the second sample the censoring times $V_1, V_2, \ldots V_m$ are independent of $Y_1^0, \ldots\ldots Y_m^0$ and they are iid random variables with a continuous distribution function $H_2(u)$. We observe the pairs $(y_i, \in_i)$, i=1,…,m, where

$$y_i = Min\ (y_i^0, V_i), \in_i = I(y_i^0 \le V_i)$$

The aim of this paper is to present a test of the hypothesis Fo=Go, against Fo=Go based on observed values $(X_1, d_1)\ldots\ldots\ (X_n, d_n), (y_i, \in_i), \ldots\ldots.(y_{m,} \in_m)$. An asymptotic test of this hypothesis, which is constructed in the next section, is non-parametric and does not require validity of the proportional hazard model.

[+] Received on 5/8/2004 , Accepted on 14/4/2005
[*] Asst. Prof /Technical College of Management

<u>The proposed test</u>: -

Let $\hat{Fo}$ based on the sample $(X_1,d_1).....(X_n,d_n)$ and $\hat{G}_o$ based on the sample $(y_1,\in_1).....(y_m,\in_m)$, be the Kaplan-Meier product limit estimator of Fo and Go, respectively. We recall, that this estimator was introduced by [١], we shall assume, that the usual conditions.

$F_o(o)=G_o(o)=H_1(o)=H_2(o)=0$

hold. We denote by T_{Fo}, T_{Go}, T_{H1}, T_{H2} the uppermost support points of

1- F_o, 1-G_o, 1-H_1, 1-H_2 respectively, and put

$$T_1^* = Min\,(T_{FO},T_{H1}),\quad T_2^* = Min\,(T_{Go,},T_{H2})$$

Let us denote for $0 \le P < 1$
$x_P = \inf[t : Fo(t) \ge p, t \ge 0]$

The p-th quantile of Fo. Let Po is a fixed constant, satifying the inequalities 0<Po<1. We shall assume, that the following regularity conditions are fulfilled.

1- The distribution function Fo has a bounded second derivative on $(0, x_{Po} + d)$ where d is a positive number, and the first and the second right-hand derivatives at x=0 exist.
2- Inf $[Fo'(x_q) : 0 \le q \le Po]$ is positive.
3- $x_{Po} < Min(T_1^*, T_2^*)$
4- $G'o$ exists and is continuous on (0,T) for some $T > x_{PO}$.

In the following theoretical section (theorem) of our test, by the weak convergence of stochastic process we understand the weak convergence of their finite dimensional distributions. [2]

<u>Our test: (theorem)</u>:-

Let $\hat{x}_P = \inf\left[t : \hat{Fo}(t) \ge p, t \ge o\right]$ be the p-th quantile of $\hat{Fo}$. If the sample sizes $m,n \longrightarrow \infty$ and m/n tend to a non-negative real number l, then the process $m^{1/2}\left[\hat{Go}(\hat{x}_{Po}) - Go(x_P)\right], o \le P \le Po$ converges weakly to Gaussian process Z(p) , $O \le P \le Po$, having zero mean and covariance structure given for $0 \le P \le q \le Po$ by

$$Cov[Z(p),Z(q)] = ((1-Go(x_P))(1-Go(x_q))\int_O^{x_P}(1-G(y))^{-2}\,d\bar{G}(y) + l\;\frac{Go'(xp)}{Fo'(xq)} - \frac{Go'(xp)}{Fo'(xq)}$$ Whe

$$(1-Fo(x_P))\left(1-Fo(\overset{O}{x}_q)\right)\int_0^{x_P}(1-F(x))^{-2}\,d\bar{F}(x)..........(1)$$

re

$$1-G(y) = (1-Go(y))\;(1-H_2(y))\;,\;\bar{G}(y) = P\{y_1 \le y, e_1 = 1\}$$

$$1-F(y) = (1-Fo(y))\;(1-H_1(y))\;,\;\bar{F}(y) = P\{X_1 \le y, d_1 = 1\}$$

We can prove that obviously as:

$$m^{1/2}\left[\hat{G}o(\hat{x}_p) - Go(x_p)\right] = m^{1/2}\left[\hat{G}o(\hat{x}_p) - Go(\hat{x}_P)\right] + m^{1/2}\left[Go(\hat{x}_p) - Go(x_p)\right]...(2)$$

By corollary 1(i) and theorem 1 from [3] the relation:

$$\hat{G}o(\hat{x}_p) - Go(\hat{x}_p) = \hat{G}o(x_p) - Go(x_p) + \left(m^{-3/4}(\log m)^{3/4}\right)...(3)$$

holds with probability 1, as $m,n \longrightarrow \infty$. According to theorem 5 from [4] the process $m^{1/2}\left[\hat{G}o(x_P) - Go\,x_P\right], 0 \le P \le P.o$ converges weakly to a Gaussian Process $Z_1(p)$ $0 \le P \le Po$, having the mean Zero and the covariance:

$$\left(1-Go(x_P)\;\;(1-Go(x_q)\right)\int_O^{x_P}(1-G(y))^{-2}\,d\bar{G}(y). \quad o \le p \le q \le po$$

Utilizing the Lagrange theorem we can write the second part of the equality (2) with probability tend to 1 as

$$m^{1/2}\left[Go(\hat{x}) - Go(x_P)\right] = m^{1/2}Go'(\hat{a}_P)\left(\hat{x}_P - x_P\right)......(4)$$

where $\hat{a}_P$ is between $\hat{x}\,p$ and x_P Since G_O' is continuous and theorem 2 from [3] implies that $\sup\left\{m^{1/2}\left|\hat{x}_P - x_P\right|; o \le p \le po\right\}$ is bounded in probability, the relation (4) can be expressed in the form

$$m^{1/2}\left[Go(\hat{x}\,p) - Go(x\;\;p)\right] = m^{1/2}Go'(x_p)(\hat{x}\,p - x\,p) + o(1)......(5)$$

But according to corollary 1(ii) from [3] the process $n^{1/2}Fo'(x_p)(\hat{x}\,p - x\,p).o \le p \le po$ converges weakly to a Gaussian process Zz(p), $o \le p \le po$, having the mean zero and the covariance.

$$(1-Fo(x_P))(1-Fo(x_q))\int_0^{x_p}(1-F(x))^2\,d\bar{F}(x)\;\; o \le p \le q \le po$$

and it follows from (5) that the process $m^{1/2}\left[Go(\hat{x}_P) - Go(\hat{x}_P)\right]$ $o \le p \le po$, converges weakly to the process

$$l^{\frac{1}{2}}\frac{Go'(x_P)}{Fo'(x_P)}Z_2(p) \ , \ o \le p \le po$$

According to (2), (3) and independence of the processes

$$m^{\frac{1}{2}}\left[\hat{Go}(x_P)-Go(x_P)\right] \ , \ m^{\frac{1}{2}}\left[\hat{Go}\left(\hat{x}_P\right)-Go(x_P)\right] \ , \ o \le p \le po$$

together with the weak convergence results yield that the process $m^{\frac{1}{2}}\left[\hat{Go}(\hat{x}_P)-Go(x_P)\right], o \le p \le po$ converges weakly to a Gaussian process Z(p), $o \le p \le po$ with the mean Zero and the covariance (1) and the proof completed.

Let us assume, that Fo=Go since

$Sup\left\{\left|P-\hat{Fo}(\hat{x}_P)\right| : o \le p \le po\right\} = O(\frac{1}{n})$ with probability 1.

The process $m^{\frac{1}{2}}\left[\hat{Go}(\hat{x}_P)-\hat{Fo}(x_P)\right]$ has the same weak limit as the process $m^{\frac{1}{2}}\left[\hat{Go}(\hat{x}_P)-P\right], o \le p \le po$.

Hence if o<p< … <P_K≤Po and

$$T(P_I,........P_K) = \sum_{i=1}^{K} m^{\frac{1}{2}}\left[\hat{G}(\hat{x}_{Pi})-\hat{Fo}(x_{Pi})\right]......(6)$$

Then in the notation (1) and

$$V(P_I,........P_K) = \sum_{i=1}^{K}\sum_{j=1}^{K} COV\left(Z(P_i),Z(P_j)\right)......(7)$$

the statistic $T(P_1,......,P_K)/V(P_1,......,P_K)^{\frac{1}{2}}$ according to theorem converges in distribution to the standard normal distribution. N(0.1). But if we denote by $d_{(j)}$ and $e_{(j)}$ the $d_{(j)}$ and $e_{(j)}$ associated with jth order statistics $X^{(j)}$ and $y^{(j)}$, respectively, then for $o < p \le po$

$$\lim_{m\to\infty}\frac{1}{m}\sum_{y^{(j)}\le \hat{x}_P}\frac{e(j)}{(1-j/m)^2} = \int_O^{x_P}(1-G(y))^{-2}d\bar{G}(y)$$

with probability 1, and

$$\hat{V}(P_1,......,P_K) = \sum_{i=1}^{K}(1-P_i)^2\left\{\sum_{y^{(j)}\le \hat{x}p_i}\frac{e(j)/m}{(1-j/m)^2}+\frac{m}{n}\sum_{y(j)\le \hat{x}p_i}\frac{d(j)/n}{(1-j/n)^2}\right\}+$$

$$2\sum_{r<s}\sum(1-P_r)(1-P_s)\left\{\sum_{y^{(j)}\le \hat{x}p_r}\frac{e(j)/m}{(1-j/m)^2}+\frac{m}{n}\sum_{y(j)\le \hat{x}P_r}\frac{d(j)/n}{(1-j/n)^2}\right\}.......(8)$$

is a consistent estimator of (7) thus distribution of the our proposed test statistic.

$$Q(P_1,......,P_K) = T(P_1,......,P_K)/\hat{V}(P_1,......,P_K)^{\frac{1}{2}}.......(9)$$

can be, under the null hypothesis Fo=Go and under the imposed regularity conditions, approximated by the standard normal distribution. Since the test statistic (9) tends to ∞ if

$Fo(\chi_{Pi}) \le Go(\chi_{Pi}), i = 1,......., k$ and
$Fo(\chi_{Pi}) < Go(\chi_{Pji})$ for some j , …….(10)

the test based on (9) rejects the null hypothesis Fo=Go and accepts the alternative hypothesis (10) if $Q(P_1,........,P_K) > Z_a$ where Z_a is the percentage point of the standard normal distribution. Testing of the null hypothesis Fo=Go against the alternative $Fo(\chi_{Pi}) \ge G(\chi_{Pi})$, i=1,….,K and $Fo(\chi_{Pji}) > Go(\chi_{Pj})$ for some j. can be obviously performed in an analogous way.

. Application[5] :-

To illustrate the use of our proposed test and compare its value with the other tests, the data given in the following table represent two samples, The first sample consists of lengths of remission in weeks for 21 patients with acut leukemia, maintained on the drug 6-MP , sample2 consists of lengths of unmaintained remission for 21 control patients. The plus sign denotes censored observations.

Table (1) Length of remissions (weeks)

Sample (1) :	6 , 6 , 6 , 7 , 10 ,10, 13 , 16 , 22 , 23 , 6+ , 9+ , 10+ 11+ , 17+ , 20+ , 25+,32+, 32+ , 34+ , 35+
Sample (2) :	1 , 1 , 2 , 2 , 3 , 4 , 4 , 5 , 5 , 8 , 8 , 8 , 8 , 8 ,11 , 11 , 12 , 12 , 15 , 17 , 22 , 23

Let p1 = 0.25 , p2 = 0.50 , p3 = 0.75 . The test statistics

Q(p1,p2,p3) calculated in two methods . In method one the quantile $\hat{X}pi = (i = 1,2,3)$ is calculated from the first sample and $\hat{G}_0$ from the second sample. In method two , the $\hat{X}pi$ is calculated from the second sample and $\hat{G}_0$ from the first sample . Values of test statistics and important terms in the test statistics are given in table (2)

Table (2) Values of important terms in the proposed test statistics for two methods

	method One			**Method Two**		
	P1	P2	P3	P1	P2	P3
$\hat{x}_{pi}$	10	22	23	3	8	12
$\hat{F}_0(\hat{x}_{pi})$	0.2471	0.4622	0.5518	0.2381	0.4762	0.7619
$\sum \frac{d(i)/n}{(1-j/n)^2}$	0.36471	1.3315	2.1715	0.3302	1.7716	4.9361
$\hat{G}_0(\hat{x}_{pi})$	0.6667	0.9524	1	0	0.1933	0.3091
$\sum \frac{E(j)/m}{(1-j/m)^2}$	1.7717	33.5193	33.567	0	0.2575	0.3641
Q(P1,P2,P3)	3.9043			-2.8720		

The data are also analyzed by the three known tests mentioned above. The results are given in the table below

Table (3) Results of known tests

Gehan [6]	Mant.-Haen..[6]	Taron-Ware[6]	Q(p1,p2,p3)	
			1	2
4.7	5	3.9	3.9043	-2.872

Q(p1,p2,p3) test is significant at the (0.05) level for both methods (Z 0.05= 1.96) . In fact we may conclude that $G_0(t) > F_0(t)$ that is the drug 6-MP is effective. Gehan test , Mantel-Haenszel test , Tarone – Ware test gave the same conclusion.

References

1-Cox,D.R .” Regression models and life table”. *Journal of the Royal Statistical Society* , series B , 34, pp (187-202) (1972).

2- Jorge , L.R. “ Censored data “ *START* vol.(11),No. 3 (2004) .

3- Chen , Kuang-Fu. “ On almost sure representation for quantities of product limit estimator with application “ . *The Indian Journal of Statistics* , 46 , series A , pp(426-443) (1984).

4- Breslow , N and Crowley , J “ A large sample study of lifetable and product limit estimates under random censorshiop “ *The Annals of Statistics* ,2 , (437- 453) (1974) .

5- Cox,D.R & OAKES.D " *Analysis of survival data* " Chapman & Hall , New York 1988) .

6- Miller , R.G .”*Survival analysis* “ . Wiley New York (1981)