A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 255, 2011

Tomasz Jurkiewicz[*], *Ewa Wycinka*[**]

SIGNIFICANCE TESTS OF DIFFERENCES BETWEEN TWO CROSSING SURVIVAL CURVES FOR SMALL SAMPLES

Abstract. Survival analysis is concerned with studying the time between entry to a study and a subsequent event. Time-to-event is considered as a continuous variable. When the outcome of a study is the time between one event and another, a number of problems can occur, such as: the distribution of the variable tends to be unknown, observed distributions are strongly skewed, more over we lost to follow up some entities (right censoring).

The assessment of overall homogeneity of survival curves is a key element in survival analysis. Recently there have been developed several tests that compare survival at two or more cohorts e.g. most popular log-rank test and tests of Gehan, Tarone-Ware, Peto-Peto, Harrington-Fleming, Renyi-type. Nevertheless a little attention is drawn to comparison of applicability of these tests.

The main goal of this paper is to examine a small-sample characteristics of above tests. There were made a variety of situations by means of Monte Carlo simulations. With the assumption that survival curve has Weibull distribution, there were taken into consideration different share of censored observations (randomly appeared due to uniform distribution) and the ability of these test to detect overall differences between crossing survival curves.

Key words: Survival analysis, Censoring, Crossing survival curves, Statistical power, Effective sample size

I. INTRODUCTION

The comparison of time-to-event distributions, particularly with censored data is one of the most common tasks in survival analysis. There were proposed several tests for two or more survival curves. Only a few of them were implemented in statistical software [mainly log-rank and Gehan] which makes their applications routine in most problems. However the power of these tests differs in particular situations, so in some applications one of the other weight function may be more appropriate.

[*] PhD, Department of Statistics, University of Gdańsk, jurkiewicz@wzr.ug.edu.pl.

[**] PhD, Department of Statistics, University of Gdańsk, wycinka@wzr.ug.edu.pl.

The aim of this article is a comparison of the power of some two sample tests most recommended in literature. We focused on four situations that very often occur in real data studies and in which the power and other characteristics of compared tests can differ:

Problem 1: comparison of small sample properties of two sample tests (sample size)

Problem 2: the effect of increasing number of censored observations on properties of tests

Problem 3: the ability of tests to detect overall differences between non-crossing survival curves

Problem 4: the ability of tests to detect overall differences between crossing survival curves when the crossing point changes

To evaluate the performance of the compared tests, Monte Carlo simulations were carried out to study the statistical power and type I error under a variety situations.

II. LITERATURE STUDIES

It is stressed in literature that the log-rank test has optimal local power to detect differences in the hazard rates, when the hazard rates are proportional [Klein, Moeschberger 1997, p. 2001; Siu and all 2004, p. 259]. When this test is applied to samples from populations where the hazard rate cross, these tests have little power because early differences in favor of one group are canceled out by late differences in favor of the other treatment. Schumacher (1984) and Fleming (1987) conducted simulations and concluded that for non proportional or crossing hazards the Renyi test seems to perform much better than usual log-rank test for light censoring [Stablein, Koutrouvelis 1985]. Cramer von Misses test and t-test were proposed as ones with greater power to detect crossing hazard rates than Wilcoxon type tests [Klein, Moeschberger 1997, p. 214].

III. TESTS FOR TWO SURVIVAL CURVES

We compared the power of four groups of tests for two survival curves. In these tests the null and alternative hypothesis are formulated as:

H_0: $\lambda_1(t) = \lambda_2(t) = ... = \lambda_K(t)$ for all $t < \tau$

H_1: $\lambda_i(t) \neq \lambda_j(t)$ for any $t < \tau$

First group were Wilcoxon type tests (Wilcox I–X), they are based on weighted comparisons of the hazard rates in two groups: $Z^2 = \frac{Z_j^2(\tau)}{\mathrm{Var}(\tau)} \sim \chi^2_{v=1}$, where $Z_j^2(\tau) = \sum_{i=1}^{D} W(t_i)\left(\frac{d_{ij}}{l_{ij}} - \frac{d_i}{l_i}\right)$, $\mathrm{Var}(\tau) = \sum_{i=1}^{D} W(t_i)^2 \frac{l_{ij}}{l_i}\left(1 - \frac{l_{ij}}{l_i}\right)\left(\frac{l_i - d_i}{l_i - 1}\right)$.

Value of the statistic and subsequently the characteristics of the test depends on the kind of the chosen weight function [Blossweld, Golsch, Rohwer 2007, p. 79–81]. Selection of different weights leads to (see table 1): log-rank test (Wilcoxon 1), Gehan test (Wilcoxon 2), Tarone-Ware test (Wilcoxon 3), Peto-Peto test (Wilcoxon 4), modified Peto-Peto test (Wilcoxon 5), Harrington-Fleming test with p=0 and q=1 (Wilcoxon 6), Harrington-Fleming test with p=1 and q=0 (Wilcoxon 7), Harrington-Fleming test with p=1 and q=1 (Wilcoxon 8), Harrington-Fleming test with p=0,5 and q=0,5 (Wilcoxon 9), Harrington-Fleming test with p=0,5 and q=2 (Wilcoxon 10) [Suciu and all (2004)].

Second group of tests are Renyi type tests (Renyi 11–20) that are censored-data analogous of the Kolmogorov-Smirnov statistics. $Q = \sup\{|Z_t|,\ t \le \tau\}/\sigma(\tau)$, where $Z_t = \sum_{t_k \le \tau} W(t_k)\left(d_{kl} - l_{kl}\frac{d_k}{l_k}\right)$, $\sigma^2(\tau) = \sum_{t_k \le \tau} W(t_k)^2 \frac{l_{k1}}{l_k}\frac{l_{k2}}{l_k}\frac{l_k - d_k}{l_k - 1} d_k$.

The value of the statistics also depends on the chosen weight function. The set of the weight functions is the same as for the Wilcoxon type tests.

Table 1. Weight functions for Wilcoxon and Renyi type tests

No	Test	Weight $W(t_i)$
I XI	Log-Rank	1
II XII	Gehan	l_i
III XIII	Tarone – Ware	$(l_i)^{0.5}$
IV XVI	Peto-Peto	$\tilde{S}(t_i)$ where $\tilde{S}(t_i) = \prod_{t_i \le t}\left(1 - \frac{d_i}{l_i + 1}\right)$
V XV	Modified Peto – Peto	$\tilde{S}(t_i) \cdot l_i / (l_i + 1)$
VI XVI	Harrington-Fleming $p = 0$, $q = 1$	$1 - \hat{S}(t_{i-1})$
VII XVII	Harrington-Fleming $p = 1$, $q = 0$	$\hat{S}(t_{i-1})$
VIII XVIII	Harrington-Fleming $p = 1$, $q = 1$	$\hat{S}(t_{i-1})\,(1 - \hat{S}(t_{i-1}))$
IX XIX	Harrington-Fleming $p = 0.5$, $q = 0.5$	$\hat{S}(t_{i-1})^{0.5}\,(1 - \hat{S}(t_{i-1}))^{0.5}$
X XX	Harrington-Fleming $p = 0.5$, $q = 2$	$\hat{S}(t_{i-1})^{0.5}\,(1 - \hat{S}(t_{i-1}))^{2}$

Source: Blossweld, Golsch, Rohwer 2007, p. 79–81.

Moreover we took into consideration Cramer von Misses test (CramMises) and t-Student generalized test (t-Stud). Cramer von Misses test is based on the integrated, squared difference between the two empirical survival functions:

$$Q_2 = n\sum_{t_i \le \tau}\left[\frac{\hat{\Lambda}_1(t_i)-\hat{\Lambda}_2(t_i)}{1+nS^2(t_i)}\right]^2\left[A(t_i)-A(t_{i-1})\right],$$

where: $\hat{\Lambda}_j(t_i) = \sum_{t_i \le \tau}\frac{d_{ij}}{l_{ij}}$, $S_j^2(t_i) = \sum_{t_i < t}\frac{d_{ij}}{l_{ij}(l_{ij}-1)}$, $A(t_i) = nS^2(t_i)/(1 + nS^2(t_i))$,
$S^2(t_i) = S_1^2(t_i) + S_2^2(t_i)$.

The t-Student test is a censored data version of the test for the difference in sample means between the two populations:

$$Z = \frac{W_{KM}}{\sigma_{1,2}},$$

where:

$$W_{KM} = \sqrt{\frac{n_1 n_2}{n}}\sum_{i=1}^{D-1}\left[t_{i+1}-t_i\right]w(t_i)\left[\hat{S}_1(t_i)-\hat{S}_2(t_i)\right],$$

$$\sigma_{1,2}^2 = \sum_{i=1}^{D-1}\frac{A_i^2}{\hat{S}_{1,2}(t_i)\hat{S}_{1,2}(t_{i-1})}\frac{n_1\hat{G}_1(t_{i-1})+n_2\hat{G}_2(t_{i-1})}{n\hat{G}_1(t_{i-1})\hat{G}_2(t_{i-1})}\left[\hat{S}_{1,2}(t_{i-1})-\hat{S}_{1,2}(t_i)\right],$$

$$A_i = \sum_{k=i}^{D-1}\left[t_{k+1}-t_k\right]w(t_k)\hat{S}_{1,2}(t_k).$$

This test is based on Kaplan-Meier estimators in two samples, the population means are calculated by the area under the Kaplan-Meier curves [Klein, Moeschberger 1997, p. 191–220].

IV. PROCEDURE OF THE MONTE CARLO ANALYSIS

The performance of the compared tests was evaluated on the basis of pseudorandom data (PRNGs) generated by inversion method from Weibull distribution with different parameters. In each of 100.000 simulation experiments two random independent samples of size $n_1 = n_2$ were generated. On the basis of a generated samples values of test statistics and p-values for

statistics ware calculated. The estimated statistical power was calculated as the fraction of samples in which we reject the null hypothesis at 0.01, 0.05 and 0.10 significance level.

We use following sets of parameters in simulations:

For problem 1: distributions – 1st Weibull(1;1), 2nd Weibull(1;1) (null hypothesis true); samples size: 15, 20, 30, 40, ..., 100 completed + 20% censored.

For problem 2: distributions – 1st Weibull(1;1), 2nd Weibull(2; {.6, 1, 1.4}) (null hypothesis false); samples size 30 completed + 20%, 40%, 60%, 80%, 100%, 150%, 200% censored.

For problem 3: distributions – 1st Weibull(1;1), 2nd Weibull({1, 1.1, ..., 2.5}; 1) (null hypothesis from true to false); samples size 30 completed + 20% censored.

For problem 4: distributions – 1st Weibull(1;1), 2nd Weibull(2; {.3, .5, ..., 3.1}) (null hypothesis false); samples size 30 completed + 20% censored.

V. EMPIRICAL STUDY RESULTS

Firstly we generated observations for two samples with the same distribution (Weibull(1,1)) and observed the changes in the difference between observed and expected p-value in a situation that sample size grows with the fixed share of censored data. The best characteristics for small samples had Wilcoxon tests (with weights 1–5), Wilcoxon test with weights 6–10 were quickly improving their characteristics with the growth of sample size. Other tests revealed to have much worse characteristics. In this matter our observations were consistent with other empirical studies. Subsequently we examined the power of tests in a situation that we drown samples from two different populations (Weibull (1,1) and Weibull (2,0.6) where real survival curves cross in their beginnings. The greatest power for small samples had both Wilcoxon and Renyi tests with weights number 6 and 10, so we can conclude that in this situation most important factor in the test statistic is kind of weight function. Wilcoxon tests II–V do not improve their power even if the number of items in the sample increases. Rest of the examined tests were improving their characteristics as the sample size were growing only by addition of censored observations (Figure 1).

In the situation that two curves cross close to the median value all of the tests (except Cramer von Misses test) had the comparable power and react with the same rate for increase in number of censored observations (figure 2).

When two survival curves cross close to the end of the curve the best predictive power have Wilcoxon tests and Renyi test with weights 1–5 and 7, the weakest abilities had tests with weight 6 and 10 (Figure 3).

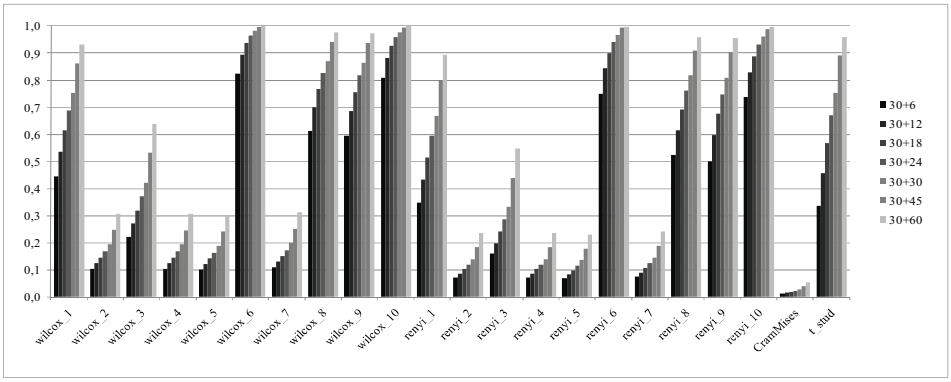

Figure 1. Power of the tests for comparison of the two samples drew from Weibull(1,1) and Weibull(2, 0.6) (α=0,01)

Source: own elaboration.

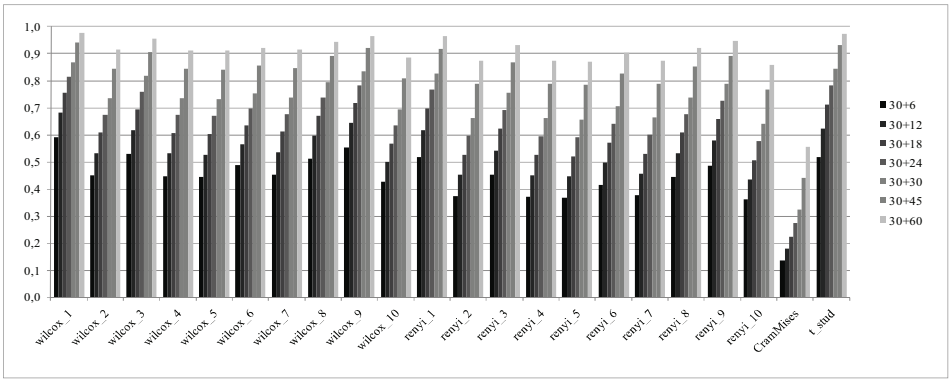

Figure 2. Power of the tests for comparison of the two samples drew from Weibull(1,1) and Weibull(2, 1) (α=0,01)

Source: own elaboration.

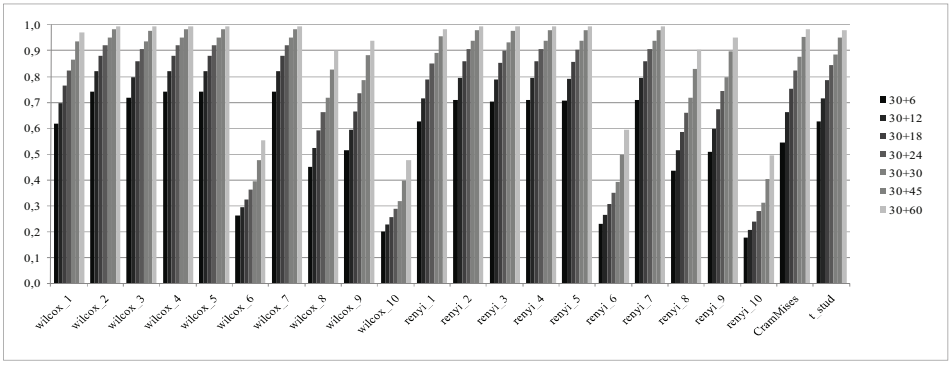

Figure 3. Power of the tests for comparison of the two samples drew from Weibull(1,1) and Weibull(2, 1.4) (α=0,01)

Source: own elaboration.

Finally we compared tests for the ability of detection of differences between non crossing survival functions. To obtain this result we drew samples from two Weibull distributions with the same shape parameter and different scale parameter. In this situation the greatest power had Wilcoxon test with weight 1 and the poorest one Cramer von Misses test (figure 4).

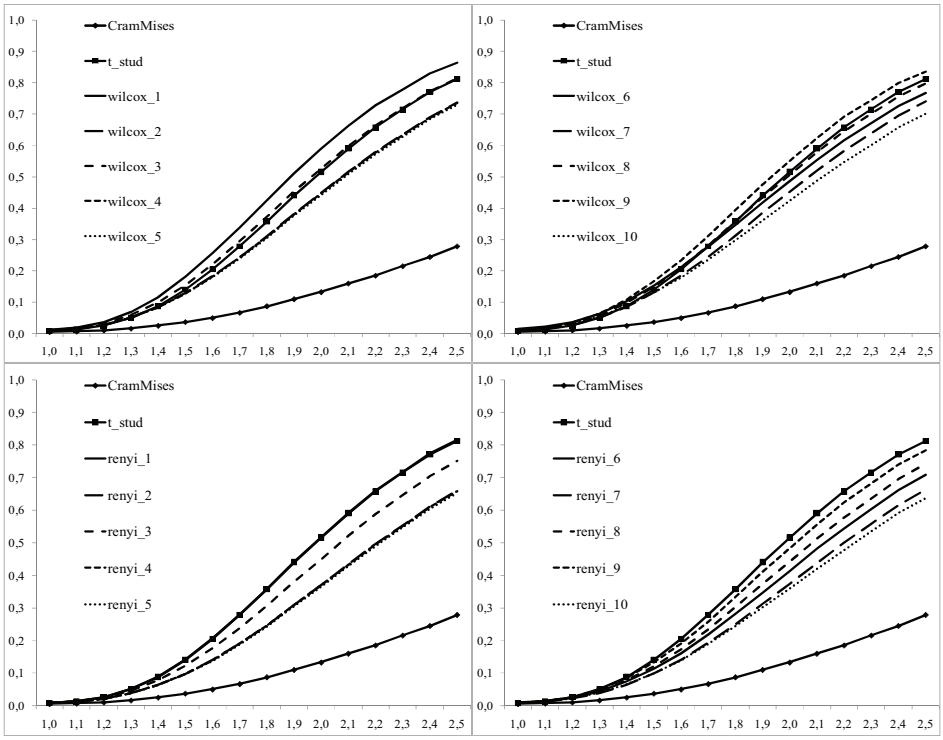

Figure 4. Power of the tests for comparison of the two samples drew from Weibull(1,1) and Weibull(x, 1) (α=0,01)

Source: own elaboration.

VI. CONCLUSIONS

As we could expect for proportional hazard populations best performance had Wilcoxon type tests, especially those with weights 1–5. These tests loose their power in the situation of crossing survival curves. Crossing point has great influence on the power of particular tests. Wilcoxon and Renyi test (4,6,10) have the greatest power to detect early time crossings and Cramer, Renyi (3,7) have the greatest power to detect late crossings. All of the analyzed tests had comparable power in a situation of crossing point around the median value. From results of simulations it can be drawn the conclusion that the choice of

weight function has the great influence on the test power in particular situation. Appropriate use of tests for two survival curves require prior recognition of a type of differences, i.e. by plotting Kaplan-Meier survival curves [compare: Suciu and all, 2004., p. 2600–261].

REFFERENCES

Balicki A. (2006) *Survival analysis and life tables*, PWE, Warszawa (in Polish)
Benedetti J. and all (1982) *Effective sample size for tests of censored survival data*, Biometrics vol. 69
Blossweld H, Golsch K., Rohwer G. (2007) *Event history analysis*, Lawrence Erlbaum Associates Publishers, London
Klein i Moeschberger M. (1997) *Survival Analysis*, Springer-Verlag, New York
Lin X, Wang H. (2004) *A New Testing Approach for Comparing the Overall Homogeneity of Survival Curves*, Biometrical Journal 46
Logan B., Klein J., Zhang M. (2008) *Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation*, Biometrics 64
Stablein D., Koutrouvelis I. (1985) *A Two-Sample test Sensitive to Crossing Hazards in Uncensored and Singly Censored Data*, Biometrics 41
Suciu G., Lemeshow S, Moeschberger M (2004) *Statistical Tests of the Equality of Survival Curves: Reconsidering and Options*, at: Belakrishnan N., Rao C.R. *Advances in Survival Analysis*, Elsevier

Tomasz Jurkiewicz, *Ewa Wycinka*

TESTY ISTOTNOŚCI RÓŻNIC DWÓCH KRZYŻUJĄCYCH SIĘ KRZYWYCH PRZEŻYCIA W MAŁYCH PRÓBACH

Analiza przeżycia to zespół metod służących do modelowania czasu trwania kohorty, której jednostki są obserwowane od zdefiniowanego momentu początkowego do zdefiniowanego zdarzenia końcowego. Czas trwania jest traktowany jako zmienna losowa ciągła. Specyfika metod analizy przeżycia związana jest z występowaniem obserwacji cenzurowanych (uciętych) oraz tym, iż funkcje gęstości obserwowanej zmiennej są często nieznane, a rozkłady silnie asymetryczne, co uniemożliwia stosowanie metod klasycznej statystyki.

Podstawową funkcją stosowaną w analizie przeżycia jest funkcja dalszego trwania wyrażająca prawdopodobieństwo, że jednostka nie doświadczy zdarzenia końcowego przed czasem *t*. Metodą oceny, czy pewne zmienne mają wpływ na zróżnicowanie czasu trwania jednostek, jest przeprowadzanie testów porównujących krzywe przeżycia na podstawie dwóch (lub więcej prób). Znaczna liczba tych testów została zaproponowana w ostatnich latach, w tym testy: Log-rank, będący jednym z lub najpopularniejszym, test Gehana, Tarone-Ware, Peto-Peto, Harringtona-Fleminga, testy typu Renyi. W literaturze mało uwagi poświęca się jednakże porównaniu własności tych testów.

W poniższym opracowaniu przeprowadzono, przy wykorzystaniu metody Monte Carlo, analizę porównawczą mocy predykcyjnej testów dla dwóch krzywych przeżycia w małych próbach z różnym udziałem jednostek cenzurowanych. Losowano próby z populacji o założonym rozkładzie Weibulla przy różnych proporcjach jednostek kompletnych i cenzurowanych (o losowej kolejności pojawiania się ustalanej w oparciu o rozkład jednostajny) w celu określenia efektywnej wielkości prób dla poszczególnych testów. Szczególną wagę poświęcono problemowi krzyżowania się krzywych przeżycia i zdolności testów do wykrywania różnic między krzywymi przeżycia w takiej sytuacji.