

Permutation Tests to Compare Survival Curves for Matched Data

*Test di Permutazione per il Confronto
di Curve di Sopravvivenza per Dati Appaiati*

Stefania Galimberti, Maria Grazia Valsecchi

Dipartimento di Medicina Clinica, Prevenzione e Biotechnologie Sanitarie,

Università di Milano-Bicocca

Via Cadore 48, 20052 – Monza (Mi)

stefania.galimberti@unimib.it

grazia.valsecchi@unimib.it

Riassunto: L'obiettivo di questo lavoro è sviluppare un test per il confronto della sopravvivenza tra gruppi di soggetti appaiati in proporzione non fissa. L'appaiamento produce dati stratificati, con strati spesso di dimensione esigua e con forme di dipendenza entro strato difficili da identificare. Si è pertanto esplorato, in questo contesto, l'uso di test di permutazione quale possibile alternativa agli usuali test non parametrici per dati non appaiati. Sono stati proposti due test ottenuti seguendo un approccio multidimensionale che combina un numero finito di test di permutazione basati sulle differenze di sopravvivenza relative a tempi distinti. Il loro comportamento, studiato per mezzo di simulazioni, sembra essere buono in presenza o meno di azzardi proporzionali.

Keywords: Matched Survival Data, Permutation Tests

1. Introduction

When data from observational studies are used to explore the role of different treatments, the main problem is to limit the biases due to the lack of randomization. Matching on relevant baseline features (confounding factors) can be used in order to increase the comparability between subjects treated with an experimental therapy or the standard treatment. In settings where matching is done with a variable number of controls, highly stratified data are produced, with strata containing a few, possibly censored, observations. For this reason, the statistical comparison of survival in the two groups cannot be directly addressed by means of the usual non parametric procedures, such as the log-rank test. The stratified version of these tests, which should account for matching, is inefficient when the number of strata increases and the stratum size is small [1]. Furthermore, these methods are less sensitive when proportional hazard is not satisfied and this might often be the case, especially in the clinical setting that motivated this paper, i.e. the comparison of bone marrow transplantation and chemotherapy in the treatment of leukemia [2].

The comparison of the survival curves for highly stratified data, due to non-paired matching, was addressed by Galimberti et al.[3], who proposed non parametric tests on the survival difference at one pre-fixed time point. The aim of this work is to extend the

comparison of survival allowing for different times points using a permutation approach.

2. Methods

A non parametric combination of a finite number of dependent permutation tests performed at different time points is proposed here. This work explores the multidimensional approach introduced by Pesarin [4], in the context of survival analysis. The permutation approach allows specific assumptions on the complex form of dependence induced by matching to be avoided.

The following hypotheses on the survival distributions $S_k(t)$ ($k=1, 2$) are considered:

$$\begin{aligned} H_0: S_1(t) &= S_2(t) \\ H_1: S_1(t) &\neq S_2(t) \quad \text{or} \quad H_1: S_1(t) > S_2(t) \end{aligned}$$

and are rephrased into q sub-hypotheses H_{0i}, H_{1i} , ($i=1, \dots, q$):

$$\begin{aligned} H_{0i}: S_1(t_i) &= S_2(t_i) \quad \rightarrow \quad H_0: \bigcap_{i=1}^q H_{0i} \\ H_{1i}: S_1(t_i) &\neq S_2(t_i) \quad \rightarrow \quad H_1: \bigcup_{i=1}^q H_{1i} \end{aligned}$$

We consider q partial univariate test statistics T_i' based on the distance of the two survival curves, i.e. $\hat{S}_1(t_i) - \hat{S}_2^w(t_i)$, estimated at q pre-fixed times points $t_1, \dots, t_i, \dots, t_q$. If

J is the number of cases (experimental group), then the control group size is of $\sum_{j=1}^J m_j$,

where m_j is the number of matched controls for case j . In this group, a weighted version of the Kaplan-Meier estimator is used, under the usual assumption of independent censoring and the additional assumption of independence of censoring conditionally on strata [5]:

$$\hat{S}_2^w(t) = \prod_{u: u \leq t} \left(1 - \frac{\sum_{j=1}^J w_j d_j(u)}{\sum_{j=1}^J w_j r_j(u)} \right)$$

where $d_j(u)$ and $r_j(u)$ are the number of events and the number of subjects at risk at time u in stratum j , respectively; while $w_j = 1/m_j$ is the weight applied to the m_j controls in stratum j .

The permutation distribution of the test statistics are approximated following a strategy that considers B random samples from the population of all possible permutations. Each permutation sample is obtained by assigning one subject to the experimental treatment and the remaining ones (m_j) to the standard treatment, within each observed stratum of

m_j+1 subjects. Firstly, each p-value $\hat{l}_i$ is obtained from the permutation distribution. Subsequently, the q p-values are combined in a unidimensional second-order test statistic T'' , in order to summarize the overall information obtained in the partial tests. Two alternatives are considered:

$$T_F'' = -2 \sum_{i=1}^q \log(\hat{\lambda}_i) \quad \text{Fisher transform}$$

$$T_T'' = \max_{1 \leq i \leq q} (1 - \hat{\lambda}_i) \quad \text{Tippett transform}$$

Different ways of defining the time points involved in the comparison are explored: the time axis is divided into equally spaced intervals or into time intervals defined according to the distribution of the events in the entire sample.

The performances of these tests are evaluated through simulations. The size and the power are calculated, for one-sided and two-sided alternatives, under proportional hazard and for different scenarios of non-proportionality (e.g. survival curves showing an early difference or reaching a plateau). The behavior of the tests in the presence of a low and a high percentage of censoring, and with small and large number of strata, is also explored.

3. Results

We report here the results of simulations based on 2000 samples of 50 strata, generated under different models, in the presence of 25% censoring. The comparisons were performed on 8 and 16 equally spaced time points and the tests were one-sided, with $B=3000$. Table 1 shows the coverage of the tests, while the power in detecting a survival difference of 20%, in at least one sub-hypothesis, is reported in Table 2.

Table 1: *Performance under the Null .*

Model Under H_0	p	8 time points		16 time points	
		Fisher	Tippett	Fisher	Tippett
Proportional Hazard	0.01	0.009	0.011	0.011	0.013
	0.025	0.028	0.029	0.028	0.030
	0.05	0.055	0.049	0.056	0.057
	0.10	0.100	0.101	0.113	0.104
Early Difference	0.01	0.013	0.015	0.012	0.018
	0.025	0.030	0.032	0.025	0.035
	0.05	0.056	0.052	0.048	0.051
	0.10	0.095	0.097	0.101	0.087
Cure Model	0.01	0.012	0.012	0.011	0.010
	0.025	0.027	0.027	0.026	0.024
	0.05	0.050	0.054	0.051	0.052
	0.10	0.102	0.102	0.103	0.101

Table 2: *Performance under the alternative*

Model Under H_1	p	8 time points		16 time points	
		Fisher	Tippett	Fisher	Tippett
Proportional Hazard	0.01	0.777	0.697	0.811	0.732
	0.025	0.868	0.788	0.891	0.816
	0.05	0.914	0.869	0.929	0.886
	0.10	0.959	0.925	0.960	0.932
Early Difference	0.01	0.410	0.461	0.422	0.526
	0.025	0.556	0.605	0.575	0.634
	0.05	0.675	0.727	0.692	0.733
	0.10	0.801	0.834	0.809	0.835
Cure Model	0.01	0.475	0.439	0.467	0.405
	0.025	0.609	0.569	0.613	0.533
	0.05	0.727	0.688	0.718	0.648
	0.10	0.828	0.793	0.828	0.761

4. Conclusions

This paper aimed to provide a test for the comparison of survival in the context of highly stratified data produced by matching on a non-fixed proportion.

The simulation study suggests that the proposed approach is promising. More specifically, a relatively good coverage and power were found both for the Fisher and the Tippet transforms. Moreover, the number of sub-hypotheses involved does not seem to strongly influence the results, provided that a reasonable minimum number of time points for survival comparison is defined. Similar results were obtained when this approach was applied to the standard context of non-matched data.

5. References

- [1] Schoenfeld D.A., Tsiatis A.A. (1987) A modified log-rank test for highly stratified data, *Biometrika*; 74 (1), 167-175.
- [2] Marubini E., Valsecchi M.G. (1995) *Analysis of Survival Data in Controlled and Observational Studies*, Wiley: Chichester.
- [3] Galimberti S., Sasieni P., Valsecchi M.G. (2000) Non-parametric procedures for the comparison of survival curves for matched data, in Proceedings of the International Society of Clinical Biostatistics (ISCB) Meeting, Trento.
- [4] Pesarin F. (2001) *Multivariate Permutation Tests: With Applications in Biostatistics*, Wiley.
- [5] Galimberti S., Sasieni P., Valsecchi M.G. (2001) A weighted Kaplan-Meier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukemia, submitted for publication.