UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 08-52
Statistics and Econometrics Series 15
October 2008

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

UNBALANCED GROUPS IN NONPARAMETRIC SURVIVAL TESTS

Emilio Letón and Pilar Zuluaga*

Abstract

It is fairly common to find medical examples with survival data with unequal sample size among the groups. There are several tests for those cases, but in practice, the use of one test instead of another is done without justifying the election. Sometimes, the choice of one test or another can lead to different conclusions, so it is important to have some guidelines to help to choose the suitable test in unbalanced groups. The computation of the tests is done with the statistical software (BMDP, SAS, SPSS, Stata, Statgraphics, and S-Plus). However the commercial software only covers tests for the family of the weigthed tests, none of the score tests, and the nomenclature is not unified, using different names for the same test. We perform several simulations to give some pieces of advice for picking out the right test. Due to the fact that there are situations where it is advisable to use a test from the family of the score tests against a weighted one, we have developed a new software in JavaScript for Internet that computes score and weighted tests versions (10 tests) that unifies the nomenclature (this software is available from the authors upon request). We include real examples where we apply, using the new JavaScript programs, the recommendations suggested by the simulations.

Keywords: Comparison of several survival curves, Score tests, Weighted test.

* Letón, Department of Statistics, Universidad Carlos III de Madrid, Av. Universidad 30, 28911 Leganés (Madrid), Spain, e-mail: emilio.leton@uc3m.es; Zuluaga, Department of Statistics and Operational Research I, Faculty of Medicine, Universidad Complutense de Madrid, Av. Complutense s/n, 28040 Madrid, Spain, e-mail: pilarzul@med.ucm.es.

Work supported by the Spanish Ministry of Science and Technology under grant SEJ2007-64500.

Unbalanced groups in nonparametric survival tests

E. Letón[*, 1], **and P. Zuluaga** [2]

[1] Department of Statistics, Carlos III University, 28903 Madrid, Spain

[2] Department of Statistics and Operation Research I, Faculty of Medicine, Complutense University, 28040 Madrid, Spain

Abstract

It is fairly common to find medical examples with survival data with unequal sample size among the groups. There are several tests for those cases, but in practice, the use of one test instead of another is done without justifying the election. Sometimes, the choice of one test or another can lead to different conclusions, so it is important to have some guidelines to help to choose the suitable test in unbalanced groups. The computation of the tests is done with the statistical software (BMDP, SAS, SPSS, Stata, Statgraphics, and S-Plus). However the commercial software only covers tests for the family of the weigthed tests, none of the score tests, and the nomenclature is not unified, using different names for the same test. We perform several simulations to give some pieces of advice for picking out the right test. Due to the fact that there are situations where it is advisable to use a test from the family of the score tests against a weighted one, we have developed a new software in JavaScript for Internet that computes score and weighted tests versions (10 tests) that unifies the nomenclature (this software is available from the authors upon request). We include real examples where we apply, using the new JavaScript programs, the recommendations suggested by the simulations.

Key words: Comparison of several survival curves, Score tests, Weighted test

1 Introduction

Many authors have considered the problem of comparing r survival curves for several tests (Mantel (1966), Peto and Peto (1972), Tarone-Ware (1977), Lawless (1982), Lee (1992), Letón and Zuluaga (2002), Desu and Raghavarao (2004), among others). However, there are no recommendations for which test is the best, among all of the tests considered above, in the general setting of unbalanced groups for survival data.

In this paper we perform a simulation study to address the former problem, finding that there are situations where the best test is not implemented in the statistical software. For that reason we offer JavaScript programs, that we have developed, to compute all of the tests considered in this paper.

The example that has motivated us, for writing this paper, is the data set of 137 patients from the Veteran's Administrations Lung-Cancer Trial quoted in Kalbfleisch and Prentice (1980). This example has been extensively treated in the literature. The variable that defines the groups is the type of cell that is considered as large (N_{large}=27), adeno (N_{adeno}=27), small (N_{small}=48), and squamous ($N_{squamous}$=35).

2 Main nonparametric tests for comparing r survival curves

The nonparametric tests to compare r survival curves perform the test:

$$H_0 \equiv S_1(t) = S_2(t) = ... = S_r(t)$$

$$H_1 \equiv \exists m, m' \quad with \quad S_m(t) \neq S_{m'}(t)$$

being $S_m(t)$ the theoretical survival curve for group m, with *m=1,...,r.*

[*] Corresponding author: e-mail: emilio.leton@uc3m.es

These tests use the fact that $t_1 < t_2 < ... < t_k$ are the k exact times considering all of the groups, and that in the group m there are N_m individuals at the beginning of the study, d_{mj} individuals that die at t_j , n_{mj} individuals at risk just before t_j and l_{mj} censored individuals in $\left[t_j, t_{j+1}\right)$. Additionally, we define:

$$n_j = \sum_{m=1}^{r} n_{mj} \quad , N = \sum_{m=1}^{r} N_m \quad , d_j = \sum_{m=1}^{r} d_{mj} \quad , D_j = \sum_{i=1}^{j} d_i \quad , l_j = \sum_{m=1}^{r} l_{mj} \quad , L_j = \sum_{i=1}^{j} l_i$$

The possibility of ties among the exact survival times, among censored survival times and among exact and censored survival times ($k \le N, d_j \ge 1, l_j \ge 1$) is assumed.

In the usual statistical software for the nonparametric comparison of r survival curves, the tests that appear belong to the family of weighted tests for r groups.

Table 1 Statistical software and the JavaScript program for the comparison of survival curves

Software	$w_j^{(r)} = n_j$	$w_j^{(r)} = S_{j-1}^{KM}$	$w_j^{(r)} = S_j^{PREN^*}$	$w_j^{(r)} = 1$	$w_j^{(r)} = \sqrt{n_j}$
BMDP(PC90)	Generalized Wilcoxon (Breslow)		Generalized Wilcoxon (PetoPrentice)*	Generalized Savage(Mantel-Cox)	Tarone-Ware
SAS (9)	Wilcoxon			Log-Rank	
S-Plus (2000)		Harrington-Fleming rho=1 Peto-Peto		Harrington-Fleming rho=0 Mantel-Haenszel	
SPSS (15.0)	Breslow			Log-Rank	TaroneWare
Stata (9)	Wilcoxon (Breslow)		Peto-Peto*	Log-Rank	Tarone-Ware
Statgraphics (Centurion)	Wilcoxon			Log-Rank	

The weighted tests assign weights, $w_j^{(r)}$, in each exact survival time t_j , to the differences, in each group m, between the observed deaths d_{mj} and the expected deaths $e_{mj} = d_j \frac{n_{mj}}{n_j}$ under H_0.
Their complete expression is given by:

$$U_m = \sum_{j=1}^{k} w_j^{(r)} \left(d_{mj} - d_j \frac{n_{mj}}{n_j} \right) \quad m=1,....,r \ , \quad U^T = (U_2, U_3, ..., U_r)$$

$$V[U] = \left[Cov(U_m, U_{m'}) \right] = \sum_{j=1}^{k} w_j^{(r)^2} \frac{n_{m_j} d_j (n_j - d_j)}{n_j (n_j - 1)} \left(\delta_{mm'} - \frac{n_{m'j}}{n_j} \right), \ m, \ m' = 2,...,r$$

with $\chi_u^2 = U^T V(U)^{-1} U \ \approx \chi_{r-1}^2$ under H_0, with $\delta_{mm'} = 1$ if m=m' and $\delta_{mm'} = 0$ if $m \neq m'$.

A summary for the possibilities for $w_j^{(r)}$, found in the statistical software is given in Table 1 where $S_j^{KM} = \prod_{i=1}^{j} \frac{n_i - d_i}{n_i}$ is the Kaplan-Meier (1958) estimation for the pooled survival function, $S_j^{ALT} = \prod_{i=1}^{j} \exp\left(\frac{-d_i}{n_i} \right)$ the Altshuler (1970) estimation, and $S_j^{PREN^*} = \prod_{i=1}^{j} \frac{n_i - d_i + 1}{n_i + 1}$ the Prentice and Mareck (1979) estimation. We point out, that instead of using $w_j = S_j^{PREN^*}$, we should use $w_j = S_j^{PREN}$, with $S_j^{PREN} = \prod_{i=1}^{j} \frac{n_i}{n_i + d_i}$ being the estimation given in Moreau *and others* (1992) and that verifies the regularity condition given in León and Zuluaga (2001). Nevertheless, there are others tests that belong to the family of score tests for r groups (see, for example, Peto and Peto (1972), Lawless (1982), Lee (1992), León and Zuluaga (2002), Desu and Raghavarao (2004) which are not included in

the statistical software. The score tests and the weighted tests for r groups can be considered to be equivalent, although with a different kind of estimation of the variance-covariance used (see Letón and Zuluaga (2002)) for the proof of this fact in the general case of ties.

3 Simulations for the size and power of tests

In this paper we have performed additional simulations to the ones given in Letón and Zuluaga (2002) to cover the general situation of unbalanced sample sizes among the groups. The scenario for the size of the tests is given by several unit exponential survival distributions, and the scenarios considered for the power of the tests are: "proportional hazard" (PH), "early hazard differences" (EHD), "late hazard differences" (LHD) and "middle hazard differences" (MHD). These scenarios are shown in Figure 1 for the survival functions.

Table 2 Size of the tests: N_1=75, N_2=75, N_3=75(left); N_1=50, N_2=50, N_3=50 (middle), N_1=20, N_2=20, N_3=20 (right)

U(0,τ)	Gehan	Peto-Peto	Prentice	LRAltshu	Tar.-Ware
τ=0.5 79	0.048 0.047 0.042 0.049 0.049 0.047	0.045 0.049 0.043 0.047 0.053 0.050	0.045 0.049 0.045 0.047 0.052 0.051	0.046 0.053 0.042 0.042 0.054 0.047	0.053 0.051 0.048 0.052 0.051 0.050
τ=1.0 63	0.058 0.056 0.049 0.057 0.059 0.053	0.045 0.055 0.054 0.047 0.056 0.059	0.047 0.055 0.054 0.046 0.056 0.058	0.054 0.051 0.053 0.051 0.055 0.062	0.046 0.056 0.056 0.045 0.057 0.057
τ=2.0 43	0.045 0.050 0.046 0.046 0.050 0.051	0.049 0.052 0.048 0.045 0.052 0.052	0.049 0.052 0.048 0.046 0.052 0.051	0.041 0.053 0.044 0.044 0.056 0.054	0.047 0.051 0.048 0.045 0.053 0.052
τ=4.0 24	0.047 0.048 0.054 0.047 0.048 0.059	0.046 0.046 0.058 0.049 0.048 0.063	0.046 0.046 0.058 0.048 0.048 0.063	0.043 0.050 0.051 0.047 0.052 0.065	0.046 0.040 0.062 0.043 0.045 0.071

In the simulations the censoring distribution used is U(0,τ) and we have generated 1000 samples of the data to evaluate the performance of each test in each scenario. The more frequent case of r=3 groups has been considered. In each scenario there are two lines of results, the first one being for the score tests and the second for the weighted tests.

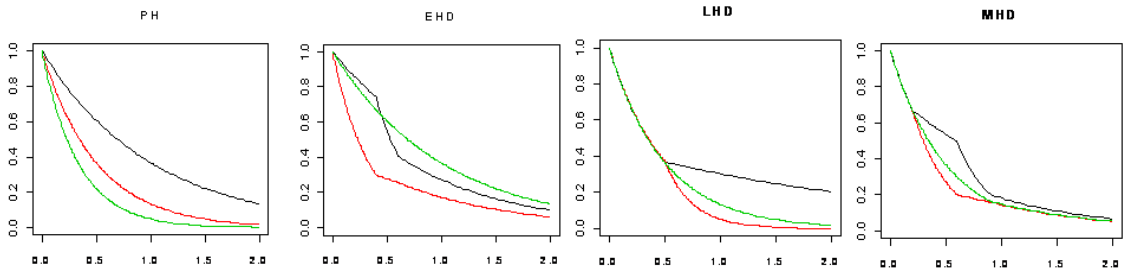

Fig. 1. Survival Functions (time t is x-axis and S(t) is y-axis)

The configurations for different sample sizes and censoring mechanism that we have considered, for the study of the size of the tests, are described in Tables 2 and 3 where τ = 0.5, 1, 2 or 4, (under the scenario, the percentages of censoring are shown) therefore we can conclude that:

-All of the values of the score tests belong to the acceptation region for the two-sided tests for α=0.05, $(0.036,\ 0.063)=(0.05\pm 1.96\sqrt{\frac{0.05*0.95}{1000}})$ and that some of the weighted tests give sizes out of that interval, with the worst performance for the weighted Log-Rank Altshuler test and for the weighted Tarone-Ware test and in unbalanced situation.

- In general, there is an "anticonservative" approach for the weighted tests, due to the fact that they reject the null hypothesis (being true) more times that the score tests. So, if the censoring mechanism can be assumed to be the same, we should use the score tests instead of the weighted tests.

Table 3. Size of the tests. In rows 1 and 2: N_1=50, N_2=50 , N_3=20 (left), N_1=50, N_2=20 , N_3=20 (right).In rows 3 and 4: N_1=75, N_2=75 , N_3=20 (left), N_1=75, N_2=20 , N_3=20 (right).

U(0,τ)	Gehan		Peto-Peto		Prentice		LRAltshu		Tar.-Ware	
	0.041	0.042	0.045	0.049	0.046	0.046	0.049	0.051	0.049	0.049
	0.045	0.049	0.050	0.054	0.051	0.053	0.052	0.058	0.051	0.050
τ=0.5										
79	0.046	0.046	0.042	0.052	0.040	0.051	0.046	0.052	0.039	0.053
	0.046	0.046	0.045	0.053	0.044	0.053	0.045	0.057	0.040	0.053
	0.055	0.051	0.054	0.057	0.052	0.059	0.049	0.056	0.049	0.057
τ=1.0	0.057	0.058	0.057	0.061	0.058	0.061	0.059	0.060	0.054	0.062
63										
	0.047	0.052	0.049	0.054	0.049	0.055	0.052	0.047	0.050	0.053
	0.047	0.060	0.050	0.058	0.050	0.058	0.052	0.060	0.048	0.061
	0.050	0.048	0.049	0.057	0.049	0.056	0.051	0.056	0.047	0.059
	0.054	0.052	0.057	0.066	0.057	0.065	0.057	0.068	0.057	0.070
τ=2.0										
43	0.044	0.047	0.037	0.049	0.037	0.049	0.046	0.055	0.040	0.051
	0.047	0.059	0.046	0.061	0.046	0.061	0.055	0.066	0.049	0.061
	0.051	0.049	0.051	0.052	0.051	0.051	0.050	0.056	0.053	0.056
	0.055	0.057	0.057	0.062	0.057	0.062	0.068	0.069	0.060	0.063
τ=4.0										
24	0.041	0.053	0.044	0.056	0.044	0.056	0.036	0.048	0.040	0.054
	0.044	0.053	0.045	0.054	0.045	0.055	0.041	0.052	0.042	0.056

In Tables 4 and 5 we give the results for the power of the tests (the censoring distribution used is U(0,2)). In each table six different sample sizes are considered. Under the scenario, the percentages of censoring for each group are shown. We conclude that:

Table 4 Power of the tests. In rows 1 and 2 : N_1=50, N_2=50, N_3=20 (left); N_1=50, N_2=20, N_3=50(middle); N_1=20, N_2=50, N_3=50 (right). In rows 3 and 4 rows : N_1=75, N_2=75, N_3=20 (left); N_1=75, N_2=20, N_3=75(middle); N_1=20, N_2=75, N_3=75 (right).

Scenario	Gehan			Peto-Peto			Prentice			LR Altshu			Tar-Ware		
	0.865	0.955	0.764	0.895	0.969	0.786	0.892	0.969	0.786	0.927	0.982	0.886	0.914	0.976	0.834
	0.876	0.956	0.743	0.909	0.970	0.777	0.906	0.970	0.777	0.951	0.982	0.866	0.926	0.977	0.816
I: PH															
43, 25, 16	0.936	0.994	0.836	0.952	0.996	0.868	0.953	0.996	0.868	0.970	0.999	0.943	0.964	0.996	0.903
	0.937	0.994	0.821	0.960	0.996	0.852	0.960	0.996	0.851	0.979	0.999	0.929	0.964	0.996	0.885
	0.934	0.755	0.939	0.923	0.722	0.922	0.923	0.722	0.922	0.769	0.538	0.828	0.908	0.699	0.902
	0.938	0.804	0.940	0.925	0.783	0.931	0.925	0.783	0.931	0.786	0.658	0.839	0.912	0.760	0.911
II: EHD															
38, 24, 44	0.992	0.774	0.989	0.989	0.733	0.990	0.990	0.732	0.989	0.905	0.543	0.948	0.977	0.705	0.987
	0.992	0.816	0.991	0.990	0.792	0.991	0.990	0.792	0.991	0.911	0.668	0.953	0.980	0.772	0.988
	0.057	0.058	0.065	0.077	0.066	0.080	0.076	0.066	0.080	0.252	0.172	0.219	0.111	0.082	0.100
III: LHD	0.062	0.064	0.064	0.083	0.073	0.075	0.081	0.071	0.072	0.272	0.192	0.208	0.117	0.092	0.098
36, 20, 24															
	0.082	0.062	0.067	0.110	0.076	0.082	0.108	0.075	0.081	0.380	0.226	0.298	0.156	0.090	0.117
	0.088	0.066	0.065	0.110	0.086	0.080	0.107	0.085	0.080	0.406	0.250	0.264	0.158	0.102	0.112
	0.196	0.142	0.115	0.223	0.150	0.125	0.223	0.150	0.125	0.243	0.175	0.155	0.247	0.168	0.143
	0.205	0.153	0.105	0.226	0.167	0.112	0.225	0.167	0.112	0.260	0.212	0.132	0.261	0.195	0.130
IV: MHD															
31, 24, 26	0.259	0.175	0.138	0.294	0.190	0.153	0.293	0.190	0.153	0.332	0.192	0.190	0.327	0.206	0.171
	0.264	0.190	0.130	0.299	0.205	0.143	0.299	0.205	0.143	0.347	0.240	0.172	0.334	0.231	0.156

- Similar power between score and weighted tests, although sometimes the power for the score tests is better than for the weighted tests.
- The differences in power are greater between score and weighted tests if the sample sizes are different.
- The greater sample sizes give greater power for the score and weighted tests.
- There is a great variability in the power for each test in different scenarios. The worse power is observed in LHD scenario.
- In unbalanced groups, it is observed that power depends on the scenario, hazard of the groups and sample sizes.

Table 5 Power of the Tests. In rows 1 and 2: N_1=20, N_2=20, N_3=50 (left); N_1=20, N_2=50, N_3=20(middle); N_1=50, N_2=20, N_3=20 (right). In rows 3 and 4: N_1=20, N_2=20, N_3=75 (left); N_1=20, N_2=75, N_3=20(middle); N_1=75, N_2=20, N_3=20 (right).

Scenario	Gehan			Peto-Peto			Prentice			LR Altshu			Tar-Ware		
	0.746	0.588	0.817	0.785	0.632	0.836	0.785	0.630	0.836	0.890	0.715	0.874	0.842	0.672	0.853
	0.728	0.584	0.837	0.760	0.625	0.854	0.758	0.624	0.855	0.864	0.719	0.911	0.813	0.666	0.886
I: PH															
43, 25, 16	0.795	0.589	0.895	0.840	0.627	0.909	0.838	0.627	0.909	0.932	0.731	0.912	0.879	0.679	0.914
	0.769	0.586	0.909	0.802	0.617	0.925	0.800	0.617	0.925	0.895	0.721	0.952	0.853	0.656	0.931
	0.701	0.851	0.745	0.655	0.832	0.704	0.656	0.831	0.703	0.484	0.674	0.490	0.631	0.798	0.670
	0.732	0.843	0.786	0.716	0.813	0.757	0.719	0.816	0.757	0.578	0.661	0.601	0.701	0.779	0.725
II: EHD															
38, 24, 44	0.726	0.907	0.769	0.691	0.898	0.705	0.691	0.898	0.705	0.517	0.778	0.488	0.668	0.895	0.653
	0.776	0.897	0.805	0.746	0.887	0.777	0.747	0.887	0.776	0.637	0.726	0.585	0.739	0.869	0.736
	0.055	0.057	0.060	0.065	0.075	0.065	0.064	0.075	0.065	0.143	0.214	0.119	0.080	0.090	0.071
	0.057	0.060	0.066	0.064	0.070	0.074	0.063	0.070	0.075	0.140	0.196	0.164	0.079	0.090	0.083
III: LHD															
36, 20, 24	0.054	0.067	0.047	0.061	0.084	0.053	0.061	0.082	0.053	0.154	0.275	0.116	0.080	0.112	0.061
	0.061	0.069	0.064	0.068	0.077	0.069	0.066	0.077	0.069	0.152	0.235	0.168	0.081	0.103	0.086
	0.111	0.116	0.127	0.118	0.125	0.139	0.117	0.125	0.137	0.131	0.155	0.148	0.134	0.144	0.148
	0.111	0.108	0.144	0.114	0.115	0.166	0.114	0.114	0.167	0.144	0.136	0.206	0.138	0.131	0.184
IV: MHD															
31, 24, 26	0.104	0.138	0.129	0.114	0.147	0.143	0.114	0.148	0.142	0.131	0.183	0.139	0.120	0.171	0.159
	0.104	0.120	0.158	0.112	0.135	0.189	0.112	0.136	0.188	0.133	0.153	0.227	0.127	0.143	0.205

Considering all of the above, we can give some recommendations for the use of each test. Scenario I (PH): Log-Rank Altshuler, Scenario II (EHD): Gehan, Peto-Peto, Prentice, Scenario III (LHD): Log-Rank Altshuler, Scenario IV (MHD): Tarone-Ware. Tarone-Ware is an intermediate test in all of the scenarios.

4 Javascript software

Only a part of the tests of this paper can be found in the statistical software. On the other hand, the majority of the software only facilitates the value of the test and its p-value. For that, we have developed JavaScript software that covers all of the tests mentioned in this paper (including score and weighted versions), giving a more homogeneous notation and more information in the output.

This software can read ASCII files and it is incorporated in a web page because it also uses HTML code. The JavaScript software has a data zone, "buttons zone" (that include some help with the recommendation of the former section), and an output zone.

The JavaScript computes the weighted tests for r groups given by $w_j^{(r)} = n_j$, $w_j^{(r)} = S_{j-1}^{KM}$, $w_j^{(r)} = S_j^{PREN}$, $w_j^{(r)} = 1$ and $w_j^{(r)} = \sqrt{n_j}$, and their corresponding score tests.

The names used for the 10 tests of the JavaScript software are based on the relationships proved in Letón and Zuluaga (2005).

These JavaScript programs can be requested from the authors for those interested.

5 Application

In this section we include some applications using a real example, where we try to apply the recommendations given in this paper, using the JavaScript software presented in former sections.

The example considered is the one data set of 137 patients from the Veteran´s Administrations Lung cancer Trial cited in the text of Kalbfleisch and Prentice (1980, pages 223-224). This example has been extensively treated in the literature.

One of the studied variables is the type of cell that they consider to be large (N_{large}=27), adeno (N_{adeno}=27), small (N_{small}=48), and squamous ($N_{squamous}$=35). For pedagogical purposes in some texts those type of cells have been grouped. For example, Kleinbaum and Klein (2005, page 77) consider two groups: large (N_1=27) vs. others (N_2=110) and perform the weighted Log-Rank Altshuler test.

Table 6 p-values for several examples.

JavaScript	Large-others		Adeno-others		Adeno-squamous-others	
	Weighted	Score	Weighted	Score	Weighted	Score
Gehan	0.0053	0.0028	0.0453	0.0549	0.0321	0.0340
Peto-Peto	0.0059	0.0031	0.0398	0.0501	0.0265	0.0284
Prentice	0.0061	0.0032	0.0400	0.0498	0.0267	0.0286
LR Altshuler	0.0822	0.0524	0.0042	0.0194	0.0005	0.0010
Tarone-Ware	0.0121	0.0061	0.0168	0.0275	0.0061	0.0071

In Figure 2 we show the estimated survival curves for this case. If we compare this figure with the case of differences at the beginning, and due to that, it seems that the censoring mechanism is the same in the two groups, the Peto-Peto score test (p-value=0.0031) should be used (according to our recommendations).

For didactical purposes in Table 6 the p-value for the 10 tests of Table 2 is shown with the help of our JavaScript software. From Table 6 (Large-others) we observe that a great difference may result in the conclusion (including rejecting vs. accepting the null hypothesis) if we do not use the proper test. For example, in the case mentioned before, the weighted Log-Rank Altshuler test as the usual default gives a p-value = 0.0822, accepting the null hypothesis of equality between the survival curves.

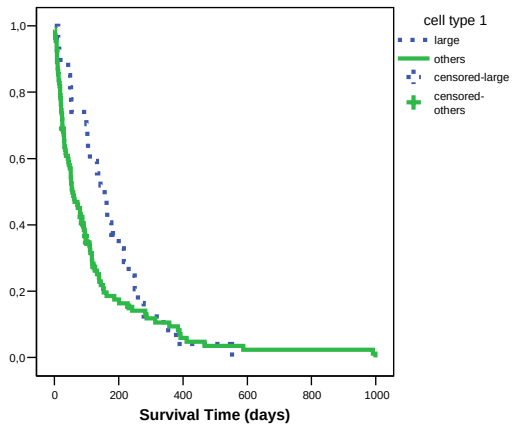

Fig. 2. Survival curves for the groups large vs. others

Using the same lung cancer data set we can illustrate other scenarios, for example, if we consider the groups adeno (N_1=27) vs other (N_2=110), we observe from Figure 3 and Figure 1 that we are in the scenario of late hazard differences. From our recommendations, the test to be used is the score Log-rank Altshuler test that gives a p-value=0.0194, rejecting the equality between the survival curves. In Table 6 (Adeno-others) we give the 10 results for the 10 tests to illustrate the differences that can be found in practice.

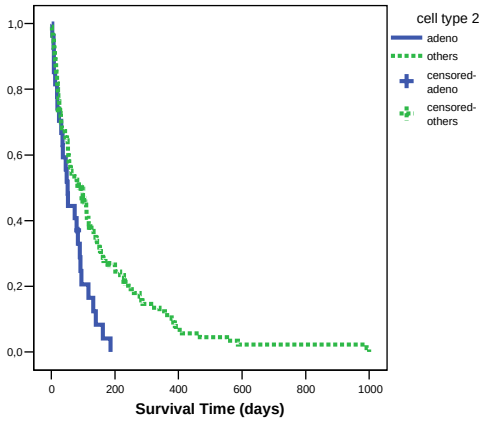

Fig. 3. Survival curves for the groups adeno vs. others

Due to the fact that the simulations given in section 3 are for three groups, we will use the lung cancer data set considering the groups: adeno, squamous and others. In Figure 4 we see that we are in PH scenario, so the test to be used is the score Log-rank Altshuler test. In Table 6 (Adeno-squamous-others) we show the results for the 10 tests in this situation: adeno (N_1=27), squamous (N_2=35) and others (N_3=75).

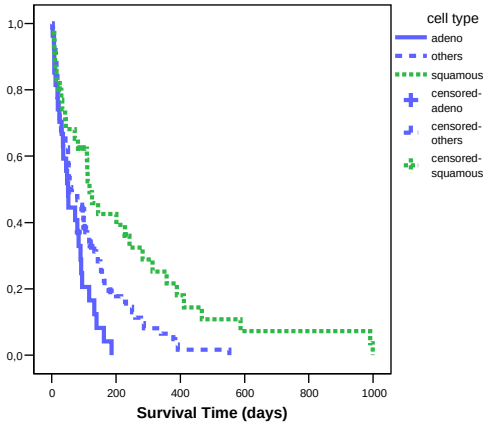

Fig. 4. Survival curves for the groups adeno, squamous vs. others

6 Conclusions

The choice of one test or another in unbalanced groups in survival data should be done with care, due to the fact that different tests can lead to different conclusions. Using simulations we give some pieces of advice that can be useful for picking out the right test.

The recommendations we suggest are:

-to draw the survival curves.

-to identify the scenario to which those survival curves belong.

-to choose the test suggested in the simulations for that scenario:

"proportional hazard": Log-Rank Altshuler,

"early hazard differences": Gehan, Peto-Peto, Prentice,

"late hazard differences": Log-Rank Altshuler,

"middle hazard differences": Tarone-Ware.

Acknowledgements Work supported by the Spanish Ministry of Science and Technology under grant SEJ2007-64500.

References

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences* **6**,1-11

Desu, M.M. and Raghavarao D. (2004). *Nonparametric statistical methods for complete and censored data*. Chapman and Hall.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data*. John Wiley and Sons, Inc., New York.

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

Kleinbaum D.G. and Klein M. (2005). *Survival Analysis*. Springer. New York.

Lawless, J.F. (1982). *Statistical models and methods for lifetime data*. John Wiley and Son, Inc., New York.

Lee, E.T. (1992). *Statistical methods for survival data analysis*. John Wiley and Sons, Inc., New York.

Letón, E. and Zuluaga, P. (2001). Equivalence between score and weighted tests for survival curves. *Communications in Statistics: Theory and Methods* **30**, 591-608.

Letón, E and Zuluaga, P. (2002). Survival tests for r groups. *Biometrical Journal* **44**,15-27.

Letón, E and Zuluaga, P. (2005). Relationships among tests for censored data. *Biometrical Journal* **47**, 377-387.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Rep*. **50**, 163-170.

Moreau, T., Maccario, J. Lellouch, J. and Huber, C. (1992). Weighted log rank statistics for comparing two distributions. *Biometrika* **79**, 195-198.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society A* **135,** 185-207.

Prentice, R.L.and Marek, P. (1979). A qualitative discrepancy between censored data rank tests. *Biometrics* **35,** 861-867.

Tarone, R.E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156-160.