TESTING TREATMENT DIFFERENCES IN CENSORED SURVIVAL DATA: A SMALL SAMPLE STUDY

Enrico A. Colosimo
and
Nancy L. Garcia

RELATÓRIO TÉCNICO Nº 41/90

Abstract. Parametric, semiparametric and nonparametric tests have been used to compare two samples of censored survival times possibly in the presence of covariates. The power functions of these tests are compared by Monte Carlo simulations under certain misspecifications commonly found in real problems. The nonparametric test (logrank) is in general less powerful than the parametric and semiparametric (score) tests. The performance of these latter tests are equivalent.

Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Ciência da Computação
IMECC - UNICAMP
Caixa Postal 6065
13.081 - Campinas - SP
BRASIL

Novembro - 1990

1 Introduction

The literature on statistical tests to compare two samples of censored survival times possibly in the presence of covariates is fairly extensive. Parametric, semiparametric and nonparametric tests have been used according to personal preference and previous knowledge about the data set of interest. However, it is not rare a situation where someone is confused about which test should be used in a particular case.

Parametric regression models for the failure time T have the following form

$$logT = Z'\beta + \sigma W \tag{1}$$

where Z is a $p+1$-vector of covariates, $\beta' = (\beta_0, ..., \beta_p)$ and σ are unknown parameters to be estimated and W has a known density $f(w)$. These models are covered in detail by Kalbfleisch and Prentice (1980, chapter 3) and Lawless (1982, chapter 6). The classical tests to be used associated with these models are Wald's, score and likelihood ratio tests. They are asymptotically equivalent .

In the nonparametric set-up, several tests have been proposed for the two sample problem. The logrank test (Mantel, 1966) is the most commonly used in this situation. Gehan (1965) proposed a generalization of the Wilcoxon statistic with a variance estimate based on the permutation distribution. Other generalizatons are due to Peto and Peto (1972), Prentice (1978) and Tarone and Ware (1977) among others.

A third approach is using the proportional hazards model introducted by Cox (1972).

This model in its most popular form, specifies that the hazard function $\lambda(t|Z)$ is given by

$$\lambda(t|Z) = \lambda_0(t)e^{Z'\beta} \tag{2}$$

where $\lambda_0(t)$, the baseline hazard function, is an unknown nonnegative function of time. Here the vectors β and Z have dimension p since the constant term in the regression portion of (2) is incorporated in the baseline hazard function. This model is neither fully parametric nor nonparametric (it has been called semiparametric) and inference is based on the partial (Cox, 1975) or marginal (Kalbfleisch and Prentice, 1973) likelihood. These inferences procedures in many situations are nearly as efficient as those for a parametric model. The same classical tests for the parametric models are associated with this model.

Whereas small sample properties of these tests have been studied by some authors, comparisons among them have been difficult to obtain. Latta (1981) investigated properties of nonparametric tests concerning sample size, censoring mechanism and distribution of failure time. The performance of the partial likelihood estimators was evaluated by Johnson et al (1982) under covariate imbalance and type II censoring. Lagakos and Schoenfeld (1984) studied the power and size of the proportional hazards score test under some misspecifications of the functional form of the regression portion of the model.

This study was carried out by using Weibull distribution for the failure time T and Monte Carlo simulation. The tests mentioned above were included in the simulation but just three of them will be presented in our analyses. In the parametric and semiparametric approaches the score test performed slightly better than the other two. Among the nonparametric tests

the logrank was the most powerful that agrees with the conclusions reached by Latta (1981).

In this paper we compare the power of these tests in small samples under certain misspecifications commonly found in real problems. Section 2 presents the tests and the simulations are described in section 3. Section 4 considers the design unbalance and the omission of a covariate effects and section 5 mixture of distributions for failure times (nonproportional hazards).

2 Description of the Tests

The test statistics for the parametric and semiparametric approaches are evaluated in the setting of two treatment groups, represented by the binary covariate Z_1. In these approaches the test $H_0: \beta_1 = 0$ (β_1 is the parameter associated with Z_1) is equivalent to test the equality between those two survival functions in the nonparametric approach.

2.1 Parametric and Semiparametric Approaches

The score test has the same form for these two approaches and is given by

$$S = U'(\hat{\beta}_0)\, I^{-1}(\hat{\beta}_0)\, U(\hat{\beta}_0) \qquad (3)$$

where $U(\hat{\beta}_0)$ and $I(\hat{\beta}_0)$ are the score function and observed information matrix evaluated at $\hat{\beta}_0$.

The elements of the score function are

$$[U(\beta)]_j = \frac{\partial \ln L(\beta)}{\partial \beta_j} \quad , (j = 0, ..., p)$$

and those of the observed information matrix are

$$[I(\beta)]_{jk} = \frac{-\partial^2 \ln L(\beta)}{\partial \beta_j \partial \beta_k}, \quad (j = 0, ..., p; k = 0, ..., p)$$

In the parametric model, L is the usual likelihood derived from the model (1) taking into account the censoring observations (see Kalbfleisch and Prentice, 1980, p.54-55). In the semiparametric model (2), L is the partial likelihood given by

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{exp(Z_i'\beta)}{\sum_{l \in R_i} exp(Z_l'\beta)} \right]^{\delta_i} \tag{4}$$

where R_i is the risk set at time t_i and δ_i is the failure indicator.

In the expression (3) of the score test, $\tilde{\beta}_0$ is the maximum (partial) likelihood estimator under the null hypothesis $H_0: \beta_1 = 0$.

Under H_0 and certain regularity conditions, $S \xrightarrow{D} \chi_1^2$ for both models. See Andersen and Gill (1982) for the large sample properties of the maximum partial estimators.

2.2 Nonparamentric Approach

When using nonparamentric models, we are not assuming any covariate structure for the survival functions. The observations from the sample i are $(min(T_{ij}, U_{ij}),\ \delta_{ij})$, $j = 1, .., n_i$, $i = 1, 2$, where U_{ij} is the corresponding censoring variable.

Let $t_1 < t_2 < ... < t_m$ be the distinct failure times in the combined sample and at each time t_k, $k = 1, 2, .., m$, define n_{ik} : number of individuals at risk from sample i and d_{ik} : number of events (death or censoring) in sample i.

A general class of nonparametric statistic tests for the equality of the two survival functions $(H_0 : S_1 = S_2)$ is given by

$$T_w = \frac{[\sum_{k=1}^{m} w_k(d_{2k} - \frac{n_{2k}}{n_k} d_k)]^2}{\sum_{k=1}^{m} w_k^2 var(d_{2k})} \tag{5}$$

where $var(d_{2k}) = \frac{n_{1k} n_{2k} d_k (n_k - d_k)}{n_k^2 (n_k - 1)}$, $d_k = d_{1k} + d_{2k}$ and $n_k = n_{1k} + n_{2k}$.

Under H_0, $T_w \xrightarrow{\mathcal{D}} \chi_1^2$, for suitable choices of w's. The logrank test is determined by taking $w_k = 1$ for $k = 1, .., m$.

3 Monte Carlo Simulation

The performance of these tests was evaluated via Monte Carlo simulations. The random samples were obtained by applying the appropriate inverse cumulative transformation to pseudorandom $U(0, 1)$ deviates. These deviates have been formed from a mixture of numbers from independents multiplicative and shift generators.

Two independent sets of independent random variables $T' = (T_1, ..., T_{100})$ and $U' = (U_1, ..., U_{100})$ were generated for each repetition and the lifetime $min(T_i, U_i)$ and $\delta_i = I(T_i < U_i)$ were recorded. T is a vector of realizations of a Weibull distribution $(\rho, exp(Z_i'\beta))$ and U_i, corresponding to the random censoring mechanism, has a uniform distribution in $(0, \theta)$. This means that W in the model (1) has a extreme value distribution.

In this study two covariates were used. A binary covariate Z_1 indicating treatment group and a continuous covariate Z_2 with uniform distribution in $(0, 1)$. Z_2 was generated once and it was maintained the same in all analyses. The parameters β_0, for the parametric model, and β_2 were set to be equal zero and one respectively.

The parameter ρ of the Weibull distribution was set to be equal 0.5. This means survival times with a monotone decreasing hazard function. Some populations are well known for this kind of pattern, such as human and manufactured items life populations in their initial period.

The amount of censoring, $P(U_i < T_i)$, was maintained constant around 0.25 by controlling the value of the parameter θ in terms of β_1.

One thousand repetitions of this process were used for various values of β_1 and different situations. For each repetition the tests described in the previous section were calculated. The power of these tests were evaluated comparing the statistic tests with the 95^{th} quantile of the chisquare distribution with one degree of freedom (3.84) for each value of β_1.

4 Unbalance and Covariate Omission Effects

The vector Z_1 of binary covariate was generated as a sequence of independent bernoulli random variables with probability of sucess p. Figure 1 presents the power functions of the tests for three values of p.

The three power functions in Figure 1 are almost equivalent when Z_1 is balanced ($p = 0.5$). As the imbalance increases the three tests lose power. The parametric and semiparametric score tests are still comparable but they differ from the logrank test when p is 0.4 and 0.3.

The logrank test is more powerful for negative values of β_1 compared with the power for the positive values when $p = 0.4$. The power function for this test seems to be shifted to the right. Looking at the design matrix it can be noticed that although the covariates Z_1 and Z_2 were generated independently, by chance they have a correlation of -0.178 for this value of p. Since the continuous covariate Z_2 is not taken into account by the nonparametric test the results were affected.

The correlation between Z_1 and Z_2 are basically null for the curves corresponding to $p = 0.5$ and 0.3. However all power functions are not completely symmetric when $p = 0.3$. The curve for the logrank test is shifted to the right and the curves for the other tests are slightly shifted to the left.

The same conditions as above were considered but now omitting the covariate Z_2 in the parametric and semiparametric models. In other words, the generation was made with both covariates and the tests derived under the model with just Z_1. Figure 2 presents the power functions of the tests for the same three values of p. The power functions for the logrank

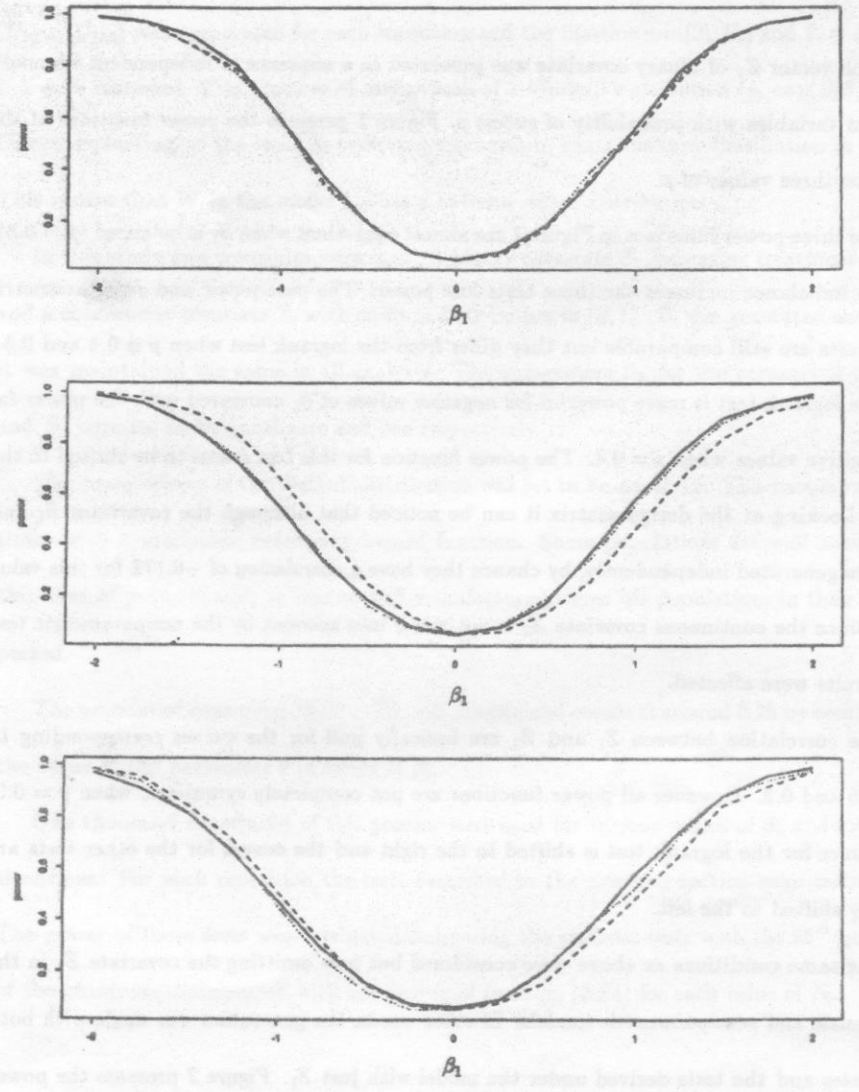

Figure 1: Power Functions for Imbalance Effect - $p = 0.5$ top, $p = 0.4$ middle and $p = 0.3$ bottom - Tests: ... parametric, — semiparametric, - - - nonparametric.

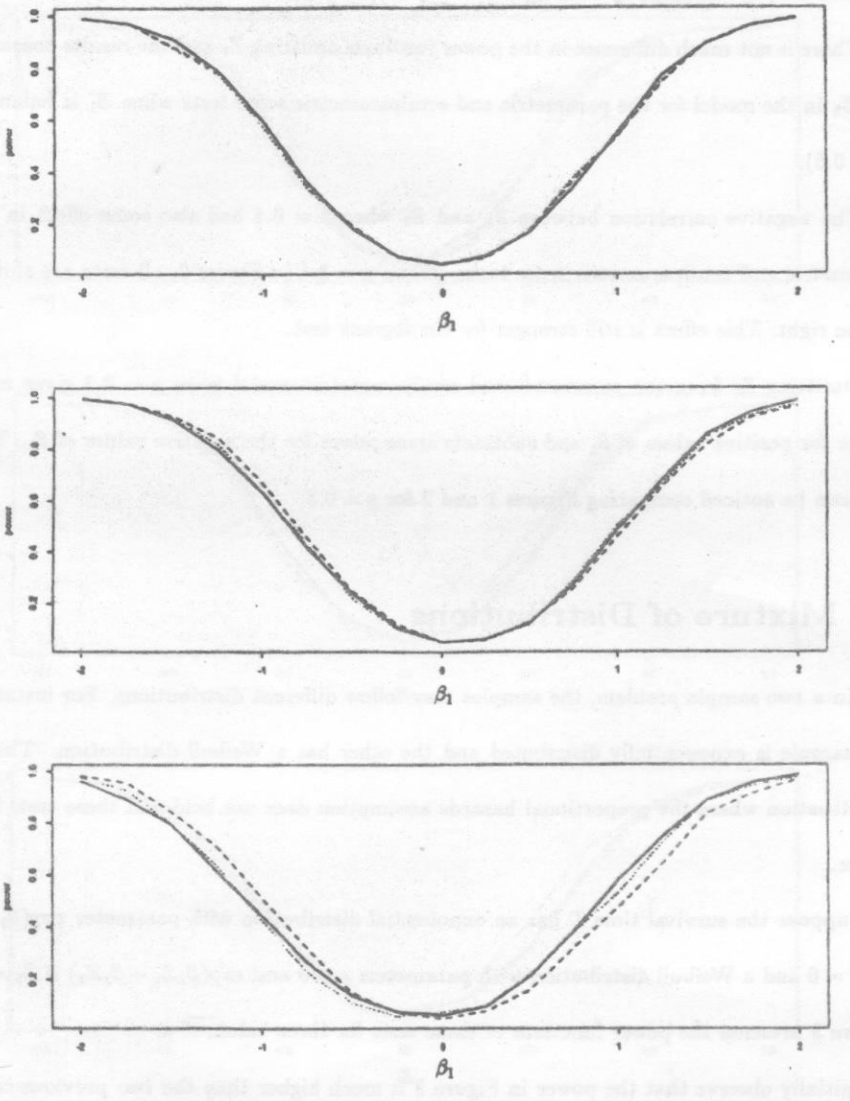

Figure 2: Power Functions for Covariate Omission Effect - $p = 0.5$ top, $p = 0.4$ middle and $p = 0.3$ bottom - Tests: ... parametric, — semiparametric, - - - nonparametric.

test are exactly the same in both figures.

There is not much difference in the power functions omitting Z_2 and the results considering Z_2 in the model for the parametric and semiparametric score tests when Z_1 is balanced ($p = 0.5$).

The negative correlation between Z_1 and Z_2 when $p = 0.4$ had also some effect in the parametric and semiparametric score tests. When $p = 0.4$ in Figure 2, all tests are shifted to the right. This effect is still stronger for the logrank test.

Omitting Z_2 from the parametric and semiparametric model when $p = 0.3$ gives more power for positive values of β_1 and subtracts some power for the negative values of β_1. This fact can be noticed comparing Figures 1 and 2 for $p = 0.3$.

5 Mixture of Distributions

In a two sample problem, the samples may follow different distributions. For instance, one sample is exponentially distributed and the other has a Weibull distribution. This is the situation where the proportional hazards assumption does not hold and these tests lose power.

Suppose the survival time T has an exponential distribution with parameter $exp(\beta_2 Z_2)$ if $Z_1 = 0$ and a Weibull distribution with parameters $\rho > 0$ and $exp(\beta_1 Z_1 + \beta_2 Z_2)$ if $Z_1 = 1$. Figure 3 presents the power functions of these tests for three values of ρ.

Initially observe that the power in Figure 3 is much higher than the two previous cases studied in the last section. The power of these tests is known to be higher when the distri-

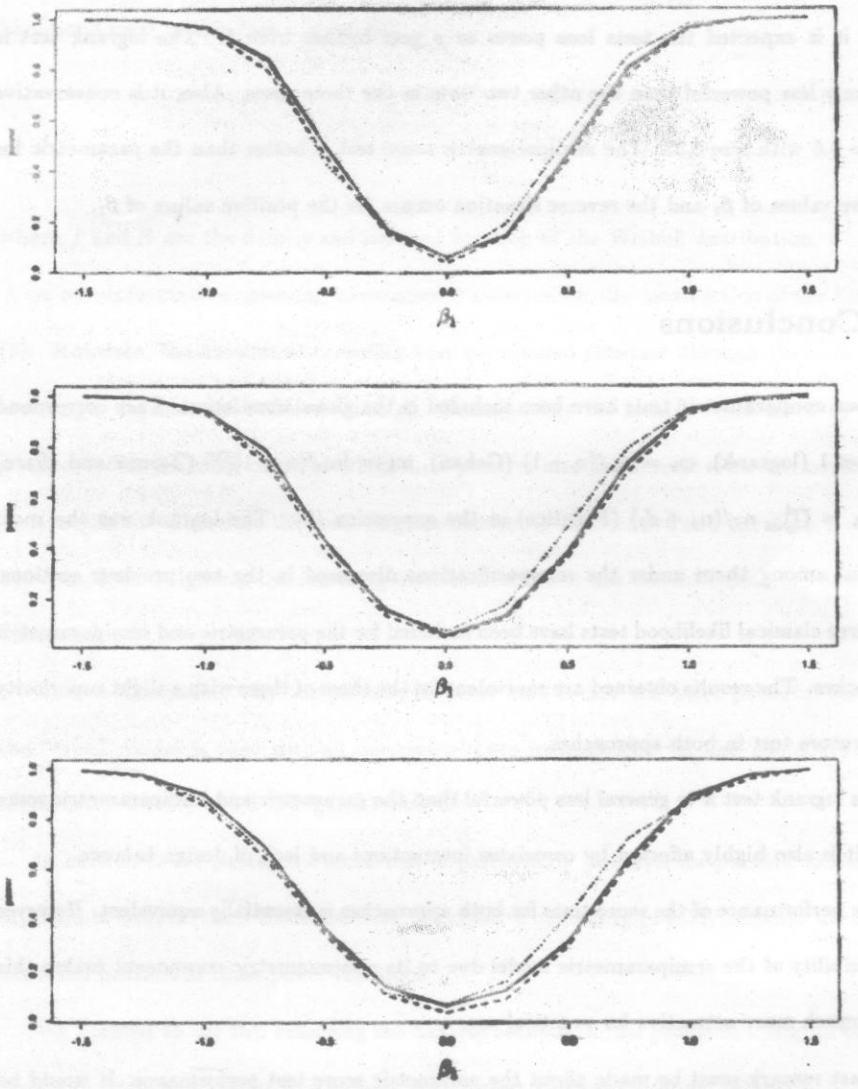

Figure 3: Power Functions for Mixture of Distributions - $\rho = 0.8$ top, $\rho = 0.7$ middle and $\rho = 0.6$ bottom - Tests: ... parametric, — semiparametric, - - - nonparametric.

bution of the survival times approach the exponential.

As it is expected the tests lose power as ρ gets further from 1. The logrank test is uniformly less powerful than the other two tests in the three cases. Also, it is conservative for $\rho = 0.6$ with size 0.33. The semiparametric score test is better than the parametric for negative values of β_1 and the reverse situation occurs for the positive values of β_1.

6 Conclusions

Four nonparametric tests have been included in the global simulation. They correspond to $w_k = 1$ (logrank), $w_k = n_k/(n+1)$ (Gehan), $w_k = [n_k/(n+1)]^{0.5}$ (Tarone and Ware) and $w_k = \prod_{j=1}^{k} n_j/(n_j + d_j)$ (Prentice) in the expression (5). The logrank was the most powerful among them under the misspecifications discussed in the two previous sections. The three classical likelihood tests have been included for the parametric and semiparametric approaches. The results obtained are equivalent for the three of them with a slight superiority for the score test in both approaches.

The logrank test is in general less powerful than the parametric and semiparametric score tests. It is also highly affected by covariates interactions and lack of design balance.

The performance of the score tests for both approaches is essentially equivalent. However the flexibility of the semiparametric model due to its nonparametric component makes this model much more attractive for practical use.

A last remark must be made about the parametric score test performance. It would be expected a better performance of this test specially in the situations in section 4 where a

generated Weibull data set was fitted using a completely parametric Weibull model. The likelihood used in this case is given by

$$L(\beta) = \prod_{i=1}^{n} [f(t_i, \beta)]^{\delta_i} \, [S(t_i, \beta)^{1-\delta_i}] \tag{6}$$

where f and S are the density and survival function of the Weibull distribution.

A non-informative censoring mechanism is assumed for the construction of the likelihood (6). However, the amount of censoring was maintained constant through the simulations, creating a dependence between θ and β_1. Consequently the sample represented by $Z_1 = 1$ is expected to be more censored when β_1 is positive and the reverse situation occurs when β_1 is negative. Therefore the censoring mechanism is informative and the likelihood (6) is not complete. It ignores the censoring distribution and it has the interpretation of a partial likelihood.

This fact explains why the parametric score test is not the more powerful test when the "true" model is used with all covariates taken into account. It also could explain the asymmetry when $p = 0.3$ in Figures 1 and 2.

In section 5 the parametric score test is particularly powerful for positive values of β_1. Mixture of distributions combined with informative censoring possibly respond for this unexpected pattern of these power functions.

We decided to use this censoring mechanism because in real problems a similar behavior generally occurs. Patients in a sample with longer expected survival times ($\beta_1 > 0$) have larger probability to be censored than the patients in the other sample and the reverse

situation occurs for smaller expected survival times ($\beta_1 < 0$).

REFERENCES

ANDERSEN,P.K. and GILL,R.D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100-20.

COX,D.R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B* **34**, 187-220.

COX,D.R. (1975). Partial likelihood. *Biometrika* **62**, 269-76.

GEHAN,E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203-223.

JOHNSON,M.E., TOLLEY,H.D., BRYSON,M.C. and GOLDMAN,A.S. (1982). Covariate analysis of survival data: a small-sample study of Cox's model. *Biometrics* **38**, 685-698.

KALBFLEISCH,J.D. and PRENTICE,R.A. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **72**, 27-36.

KALBFLEISCH,J.D. and PRENTICE,R.A. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

LAGAKOS,S.W. and SCHOENFELD,D.A. (1984). Properties of proportional hazards score tests under misspecified regression models. *Biometrics* 40, 1037-1048.

LATTA,R.B. (1981). A Monte Carlo study of some two-sample rank tests eith censored data. *J. Am. Statist. Assoc.* **76**, 713-719.

LAWLESS,J.F. (1982). *Statistical Models and Methods for lifetime Data.* New York: Wiley.

MANTEL,N. (1966). Evaluation of survival data and two new rank order statistics arising

in its consideration. *Cancer Chemotherapy Reports* **50**, 163-170.

PETO,R. and PETO,J. (1972). **Asymptotically efficient rank invariant test procedures.** *J. R. Statist. Soc. A* **135**, 185-206.

PRENTICE,R.L. (1978). **Linear rank tests with right censored data.** *Biometrika* **65**, 167-179.

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

RELATÓRIOS TÉCNICOS — 1990

01/90 **Harmonic Maps Into Periodic Flag Manifolds and Into Loop Groups** — *Caio J. C. Negreiros.*

02/90 **On Jacobi Expansions** — *E. Capelas de Oliveira.*

03/90 **On a Superlinear Sturm–Liouville Equation and a Related Bouncing Problem** — *D. G. Figueiredo* and *B. Ruf.*

04/90 **F– Quotients and Envelope of F–Holomorphy** — *Luiza A. Moraes, Otília W. Paques* and *M. Carmelina F. Zaine.*

05/90 **S-Rationally Convex Domains and The Approximation of Silva-Holomorphic Functions by S–Rational Functions** — *Otília W. Paques* and *M. Carmelina F. Zaine.*

06/90 **Linearization of Holomorphic Mappings On Locally Convex Spaces** — *Jorge Mujica* and *Leopoldo Nachbin.*

07/90 **On Kummer Expansions** — *E. Capelas de Oliveira.*

08/90 **On the Convergence of SOR and JOR Type Methods for Convex Linear Complementarity Problems** — *Alvaro R. De Pierro* and *Alfredo N. Iusem.*

09/90 **A Curvilinear Search Using Tridiagonal Secant Updates for Unconstrained Optimization** — *J. E. Dennis Jr., N. Echebest, M. T. Guardarucci, J. M. Martínez, H. D. Scolnik* and *C. Vacchino.*

10/90 **The Hypebolic Model of the Mean × Standard Deviation "Plane"** — *Sueli I. R. Costa* and *Sandra A. Santos.*

11/90 **A Condition for Positivity of Curvature** — *A. Derdzinski* and *A. Rigas.*

12/90 **On Generating Functions** — *E. Capelas de Oliveira.*

13/90 **An Introduction to the Conceptual Difficulties in the Foundations of Quantum Mechanics a Personal View** — *V. Buonomano.*

14/90 **Quasi-Invariance of product measures Under Lie Group Perturbations: Fisher Information And L^2-Differentiability** — *Mauro S. de F. Marques* and *Luiz San Martin.*

15/90 **On Cyclic Quartic Extensions with Normal Basis** — *Miguel Ferrero, Antonio Paques* and *Andrzej Solecki.*

16/90 **Semilinear Elliptic Equations with the Primitive of the Nonlinearity Away from the Spectrum** — *Djairo G. de Figueiredo* and *Olimpio H. Miyagaki.*

17/90 **On a Conjugate Orbit of G_2** — *Lucas M. Chaves* and *A. Rigas.*

18/90 **Convergence Properties of Iterative Methods for Symmetric Positive Semidefinite Linear Complementarity Problems** — *Álvaro R. de Pierro* and *Alfredo N. Iusem.*

19/90 **The Status of the Principle of Relativity** — *W. A. Rodrigues Jr.* and *Q. A. Gomes de Souza.*

20/90 **Geração de Gerenciadores de Sistemas Reativos** — *Antonio G. Figueiredo Filho* e *Hans K. E. Liesenberg.*

21/90 **Um Modelo Linear Geral Multivariado Não–Paramétrico** — *Belmer Garcia Negrillo.*

22/90 **A Method to Solve Matricial Equations of the Type** $\sum_{i=1}^{p} A_i X B_i = C$ — *Vera Lúcia Rocha Lopes* and *José Vitório Zago.*

23/90 $\mathbb{Z}_2$**–Fixed Sets of Stationary Point Free** $\mathbb{Z}_4$**–Actions** — *Claudina Izepe Rodrigues.*

24/90 **The m–Ordered Real Free Pro–2–Group Cohomological Characterizations** — *Antonio José Engler.*

25/90 **On Open Arrays and Variable Number of Parameters** — *Claudio Sergio Da Rós de Carvalho* and *Tomasz Kowaltowski.*

26/90 **Bordism Ring of Complex Involutions** — *J. Carlos S. Kiihl.*

27/90 **Approximation of Continuous Convex–Cone–Valued Functions by Monotone Operators** — *João B. Prolla.*

28/90 **On Complete Digraphs Which Are Associated to Spheres** — *Davide C. Demaria* and *J. Carlos S. Kiihl.*

29/90 **Deriving Ampère's Law from Weber's Law** — *A. K. T. Assis.*

30/90 **Testes Não Paramétricos para Experimentos Completamente Casualizados para Três Fatores, com Interação** — *Belmer Garcia Negrillo.*

31/90 **On the Velocity which appears in Lorentz Force Law: An Illuminating Puzzle** — *A. K. T. Assis.*

32/90 **Embeddings of Fréchet Spaces in Uniform Fréchet Algebras** — *Jorge Mujica.*

33/90 **The Welerstrass Stone Theorem for Convex–Cone–Valued Functions** — *João B. Prolla.*

34/90 **Open-Nucleus Theory for Beef Cattle Breeding Systems: A Revisitation** — *E. Recami* and *I.U. Packer* and *M. Tenorio-Vasconselos.*

35/90 **On Fixed Points Sets of Involutions** — *Claudina Izepe Rodrigues.*

36/90 **The mod 2 Homology of BSO** — *Claudina Izepe Rodrigues.*

37/90 **Involutions and Stationary Point Free** $\mathbb{Z}_4$**-Actions** — *Claudina Izepe Rodrigues.*

38/90 **Analysis of Bivariate Dichotomous Data from a Stratified Two-Stage Cluster Sample** — *Eliana H. de F. Marques* and *Gary G. Koch.*

39/90 **Multiplicative Iterative Methods in Computed Tomography** — *Alvaro R. De Pierro.*

40/90 **A Clifford Bundle Approach to Gravitational Theory** — *Waldyr A. Rodrigues Jr.* and *Quintino A.G. de Souza*