Biometrika (1981), **68**, *pp.* 316–9
Printed in Great Britain

The asymptotic properties of nonparametric tests for comparing survival distributions

BY DAVID SCHOENFELD
Sidney Farber Cancer Institute, Harvard School of Health, Boston, Massachusetts

SUMMARY

The asymptotic distribution under alternative hypotheses is derived for a class of statistics used to test the equality of two survival distributions in the presence of arbitrary, and possibly unequal, right censoring. The test statistics include equivalents to the log rank statistic, the modified Wilcoxon statistic and the class of rank invariant test procedures introduced by Peto & Peto. When there are equal censoring distributions and the hazard functions are proportional the sample size formula for the F test used to compare exponential samples is shown to be valid for the log rank test. In certain situations the power of the log rank test falls as the amount of censoring decreases.

Some key words: Asymptotic theory; Cure rate; Hazard function; Large sample theory; Life table; Log rank test; Modified Wilcoxon test; Random censorship model.

1. INTRODUCTION

A frequently used statistical method for the analysis of clinical data is a nonparametric test to compare the survival of two patient groups. The test must allow for right censoring, the fact that when the data is analysed some of the patients will still be alive. Tests have been proposed by Gehan (1965), Mantel (1966), Efron (1967) and Cox (1972). Peto & Peto (1972) proposed a method of constructing asymptotically efficient tests for a given family of alternatives. Using the results of Latta (1977) and Morton (1978) we can show that any of the proposed tests is asymptotically equivalent to a member of the family of tests proposed by Tarone & Ware (1977). The present paper examines the asymptotic properties of Tarone & Ware's test statistics under alternatives expressed as a hazard ratio.

2. THE TEST STATISTIC

We assume that there are n patients who are allocated between two treatment groups labelled 0 and 1. We consider tests of the hypothesis that the patient groups have the same survival distribution against alternatives that the patients in group 1 have improved survival. Let $j = 1, \ldots, n$ identify the patients, let D be the set of identifiers of those patients who died, and let t_j be the death time of the jth patient. We assume that the t_j are distinct. Let $X_j = 0, 1$ be a variable indicating the treatment group of the jth patient. Define $n_i(t)$ to be the number of patients who have not died or been censored before time t in treatment group i and let $p(t) = n_1(t)/\{n_0(t)+n_1(t)\}$.

The statistics under study have the form

$$S = \sum_{j \in D} a_j \{X_j - p(t_j)\} / [\sum_{j \in D} a_j^2 \, p(t_j) \{1 - p(t_j)\}]^{\frac{1}{2}}, \tag{1}$$

where $\{a_j\}$ are a set of predetermined weights. To test whether group 1 has better survival than group 0 one rejects when S is too small.

3. The asymptotic distribution of S

Let P_i be the proportion of patients allocated to group i. Let $H_i(t)$ be the distribution function of censoring and let $\lambda_i(t)$, $f_i(t)$ and $F_i(t)$ be the hazard, density and distribution function of survival in group i. Define the functions

$$V(t) = P_0 f_0(t)\{1-H_0(t)\} + P_1 f_1(t)\{1-H_1(t)\},$$

$$\pi(t) = \frac{P_1\{1-F_1(t)\}\{1-H_1(t)\}}{P_0\{1-F_0(t)\}\{1-H_0(t)\} + P_1\{1-F_1(t)\}\{1-H_1(t)\}}$$

and suppose that a_j has been chosen so that

$$\lim_{n\to\infty} a_j = g(t_j).$$

Then S is asymptotically normal with unit variance and mean given by

$$\frac{n^{\frac{1}{2}} \int g(t) \log\{\lambda_1(t)/\lambda_0(t)\}\, \pi(t)\{1-\pi(t)\}\, V(t)\, dt}{[\int \{g(t)\}^2 \pi(t)\{1-\pi(t)\}\, V(t)\, dt]^{\frac{1}{2}}}, \tag{2}$$

where integration is over the range $0, \ldots, \infty$.

Conditionally on $n_0(t)$ and $n_1(t)$, the $\{X_j\}$ can be treated as if they were a sequence of independent Bernoulli random variables with means (Cox, 1972)

$$\mu_j = n_1(t_j)\lambda_1(t_j)/\{n_0(t_j)\lambda_0(t_j) + n_1(t_j)\lambda_1(t_j)\}.$$

To obtain an asymptotic distribution for S assume that $\log\{\lambda_1(t_j)/\lambda_0(t_j)\}$ is $O(n^{-\frac{1}{2}})$. Then one can easily show by writing μ_j in terms of $\log\{\lambda_1(t)/\lambda_0(t)\}$ and expanding in a Taylor series about zero that in probability S tends to

$$\frac{\Sigma a_j\{X_j-\mu_j\}}{\{\Sigma a_j^2 \mu_j(1-\mu_j)\}^{\frac{1}{2}}} + \frac{\Sigma a_j \log\{\lambda_1(t_j)/\lambda_0(t_j)\}\, p(t_j)\{1-p(t_j)\}}{[\Sigma a_j^2 p(t_j)\{1-p(t_j)\}]^{\frac{1}{2}}}, \tag{3}$$

where the sums are over the range $j \in D$.

The first term in (3) has a limiting unit normal distribution (Cox, 1975). To show that the limit of the second term is given by (2), replace a_j and $p(t_j)$ by their limits $g(t_j)$ and $\pi(t_j)$. Multiply numerator and denominator by $n^{-\frac{1}{2}}$. In the denominator the law of large numbers implies that the sum will converge to the denominator in (2) since $V(t)$ is the incomplete density of failure at time t. Because $\log\{\lambda_1(t)/\lambda_0(t)\}$ is $O(n^{-\frac{1}{2}})$ the same argument can be used with the numerator.

4. Application to the log rank test

The log rank test is the most commonly used test to compare survival curves. This test has $a_j = g(t_j) = 1$.

Suppose that $\log\{\lambda_1(t)/\lambda_0(t)\}$ is a constant. Furthermore assume that $H_1(t) = H_0(t)$. Let $d = \int V(t)\, dt$, where integration is over the range $0, \ldots, \infty$, be the combined probability of death. Then, $\pi(t) \to P_1$ since $F_1(t) \to F_0(t)$ and (2) becomes $\log(\lambda_1/\lambda_0)(nP_1 P_0 d)^{\frac{1}{2}}$. Therefore the one-sided sample size formula for a test with size α and power $(1-\beta)$ is

$$n = (Z_{1-\alpha} + Z_\beta)^2/[\{\log(\lambda_1/\lambda_0)\}^2 P_1 P_0 d],$$

where Z denotes the normal tail area. Except for the computation of d, this is the same formula used to compute the sample size for the F test used when both samples are exponential (Bernstein & Lagakos, 1978).

When $H_1(t) \neq H_0(t)$ the log rank test is not fully efficient for exponential survival distributions. Letting d_i be the probability of death on the ith treatment, the efficiency is

given by

$$e = \int_0^\infty \pi(t)\{1-\pi(t)\}\, V(t)\, dt \{(P_0 d_0)^{-1} + (P_1 d_1)^{-1}\}$$

(Hájek & Sîdák, 1970, p. 267). For instance, suppose all patients in treatment i have the same censoring time. Then

$$e = (P_0 d_0 + P_1 d_1)/\max(d_0, d_1).$$

Thus if group 0 has 50% dead and group 1 has 80% dead and $P_0 = P_1 = \frac{1}{2}$, the efficiency will be 81%. This loss of efficiency is a consequence of the fact that the log rank test discards all deaths after the minimum censoring time.

Clinical trials are often designed to take in patients during a fixed period of time. Then one waits for an additional 'follow-up' period, τ, before analysing the data. It is normally assumed that the sensitivity of the test for a treatment difference will increase with τ because more deaths will have been observed. This assumption is not necessarily true. Survival in treatment group 1 is uniformly better than in treatment group 0 if and only if for all T, $\int\{\lambda_0(t)-\lambda_1(t)\}\, dt > 0$, where integration is over the range $0, \ldots, T$, but this inequality does not imply that $\lambda_0(t) > \lambda_1(t)$ for all t. Thus at some point, τ^*, the hazard functions may cross even though the survival curves do not. Examination of (2) shows that additional follow-up beyond time τ^* will decrease the power of any test with $g(t) > 0$ such as the log rank test. The 'optimal' test for this type of alternative would require $g(t)$ to change sign at τ^*. However such a test would be biased in that it would have power less than α when $\lambda_0(t) = \lambda_1(t)$ for $t < \tau*$ and $\lambda_0(t) > \lambda_1(t)$ for $t > \tau*$.

5. Constructing optimal tests

To find the most powerful test that can be put in form (1) we note that, as a consequence of the Cauchy–Schwarz inequality, (2) is maximized when $g(t)$ is proportional to $\log\{\lambda_1(t)/\lambda_0(t)\}$. This is reasonable as we would want to put large weights on $X_j - p(t_j)$ when the hazard on treatment 0 is much greater than that of treatment 1; we are rejecting when $S < -Z_{1-\alpha}$. Conversely, if the ratio is reversed we would want negative weights. Thus the optimal choice of a_j is to let $a_j = \log\{\lambda_1(t_j)/\lambda_0(t_j)\}$. Often the ratio $\log\{\lambda_1(t_j)/\lambda_0(t_j)\}$ can be written in the form $g\{F_0(t)\}$. In this case, the optimal test can be constructed without knowledge of $\lambda_0(t)$ if we let $a_j = g\{\hat{F}(t_j)\}$, where $\hat{F}$ is the Kaplan & Meier (1958) estimate of the distribution function of the pooled sample (Peto & Peto, 1972).

The author is grateful to Richard Gelber and Stephen Lagakos for their careful review of the manuscript. The work was supported by the National Cancer Institute.

References

Bernstein, D. & Lagakos, S. W. (1978). Sample size and power determination for stratified clinical trials. *J. Statist. Comp. Simul.* **8**, 65–73.

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.

Efron, B. (1967). The two-sample problem with censored data. *Proc. 5th Berkeley Symp.* **4**, 831–53.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika* **52**, 203–23.

Hájek, J. & Sîdák, Z. (1967). *Theory of Rank Tests.* New York: Academic Press.

Kaplan, E. L. & Meier, P. (1958). Nonparametric estimation for incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.

Latta, R. (1977). Generalized Wilcoxon statistics for the two-sample problem with censored data. *Biometrika* **64**, 633–5.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemo. Rep.* **50**, 163–70.

MORTON, R. (1978). Regression analysis of life tables and related non-parametric tests. *Biometrika* **65**, 329–33.

PETO, R. & PETO, J. (1972). Asymptotically efficient rank invariant test procedures (with discussion). *J. R. Statist. Soc.* A **135**, 185–207.

TARONE, T. E. & WARE, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika* **64**, 156–60.

[*Received August* 1979. *Revised July* 1980]