Paper 269-28

Beyond Proc Lifetest: Alternative Linear Rank Tests for Comparing Survival Distributions

Alan B. Cantor, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL

ABSTRACT

Those who use SAS® Proc Lifetest know that it performs the two best known non-parametric tests to compare survival distributions – the Log Rank Test and the Wilcoxon-Gehan Test. In fact, however, these two tests are actually two members of a large class of linear rank tests. These tests are all equally valid, but will have different power. The specific linear rank test having greatest power will depend upon the actual survival distributions of the populations being compared. If the ratio of the hazards is constant, then the Log Rank Test has greatest power. This fact undoubtedly accounts for its popularity. However, this property often does not hold; and, when it does not, it could be beneficial to consider other linear rank tests instead. In this paper, I will discuss these tests as well as a method, originally described by Lakatos, which allows the user to estimate the power of a variety of linear rank tests under various scenarios.

INTRODUCTION

This paper will describe the class of linear rank tests that include both the log rank test and the Gehan test. Suggestions will be made to guide the reader as to the choice of a test of optimal power. A macro that implements a large number of these tests is described, as is a macro to compute their power under various scenarios. Both macros are described in the author's BBU book. They are available for download from the SAS web site.

THE LOG RANK TEST

Suppose that the two groups being compared are indexed by 1 and 2 and assume that they have unknown survival functions $S_1(t)$ and $S_2(t)$ respectively. Suppose you have samples of sizes N_1 and N_2 from these groups. Let $N = N_1 + N_2$ and let $t_1 < t_2 < . . . < t_M$ be the distinct ordered censored or complete times for the combined sample. There may be ties so that $M \le N$. For each i from 1 to M and for j = 1 and 2, let d_{ij} be the number of deaths in group j at time t_i and let $d_i = d_{i1} + d_{i2}$. That is, d_{i1} is the number of deaths among those in group 1, d_{i2} is the number of deaths among those in group 2, and d_i is the total number of deaths at time t_i. Since we are allowing ties, d_{i1}, d_{i2}, and d_i may be greater than 1. Let R_{ij} be the number at risk in group j just prior to time t_i and let $R_i = R_{i1} + R_{i2}$. If we let $E_{i1} = d_iR_{i1}/R_i$, then E_{i1} is the proportion of those at risk time t_i who are members of group 1 times the number of deaths at that time. Under the null hypothesis of equivalent survival distributions, it is, therefore, the expected number of deaths in group 1 at time t_i conditioned on the fact that there were a total of d_i deaths at that time. Then $E_{i1} – d_{i1}$ compares the actual number of deaths in group 1 at time t_i to the number expected under the null hypothesis. Summing over i gives us the Log Rank statistic that compares the total number of deaths in group 1 to the number expected under the null hypothesis. Dividing that sum by its standard deviation gives us a statistic that is asymptotically standard normal.

OTHER LINEAR RANK TESTS

Now let's associate with each of the times, t_i, a weight, w_i. Then the sum, over i, of the $w_i(E_{i1} – d_{i1})$ is also a statistic that tests the null hypothesis of equivalent mortality in the two groups. In fact, it can be shown that the Wilcoxon-Gehan test is the special case where $w_i = R_i$. Since the w_i's are decreasing for this test, it will give greater weight to earlier deaths than the Log Rank Test. Several authors have proposed other variations based on alternative ways of defining the weights w_i. Tarone and Ware (1977) discuss weights defined by $w_i = R_i^{1/2}$. Still another choice, suggested by Harrington and Fleming (1982), is to assign weights equal to $[KM(t_i)]^\rho$ where KM(t) is the Kaplan-Meier estimate based on the combined sample and ρ is a fixed non-negative constant.

THE LINRANK MACRO

The macro, linrank, can perform a variety of linear rank tests. It is invoked by filling in the values in the following template:

```
%linrank(dataset=  , time=  ,cens=   , censval=
groupvar= , method=  ,rho=   )
```

The parameters needed are:

Dataset: The name of the dataset to be used
Time: The name of the time variable.
Cens: The name of the censoring variable.
Censval: A list of values for censvar that indicate that a survival time is censored.
Groupvar: The name of the grouping variable for the two groups being compared.
Method: This must be one of the following:

1) logrank
2) gehan
3) tarone
4) harrington

rho: The value of ρ in the Harrington method. Needed only when method = harrington.

Here is an example that uses the linrank macro to perform two tests: the Log Rank Test, and the Tarone/Ware Test. The reader is cautioned that the small number of events in this example makes the asymptotic behavior of the statistics questionable.

```
Data x;
input group time cens @@;
datalines;
1 5.3 1 1 6.2 1 1 6.8 0 1 7.8 1 1 8.4 0 1 9.0 1
1 10.1 1
2 5.3 0 2 7.1 1 2 8.2 0 2 9.1 0 2 11.0 1 2 12.1
0 2 12.5 1
;
%linrank(dataset= x, time=time, cens= cens,
censval= 0  ,groupvar=  group, method =logrank);
%linrank(dataset= x, time=time, cens= cens,
censval= 0  ,groupvar= group, method =tarone);
```

The output is displayed as Figure 1.

Of course the preceding discussion raises the important question of which test, i.e. which set of weights, to use. Although all are valid, one should not compute more than one statistic and choose the one "most significant." You may, however, specify the test to be done based upon the way you expect the survival distributions to differ from the null hypothesis. For two groups, if the ratio of the hazards is constant over time and the censoring distributions are the same, then the Log Rank Test will have maximal power in the class of all linear rank tests (Peto and Peto, 1972). Perhaps for this reason, this test is the most frequently used. Lee et al (1975) and Tarone and Ware (1977) show that when the proportional hazards assumption does not hold, other tests may have greater power.

Figure 1

```
                        Summary of Events vs Expected
                             Method = logrank

                      Percent of
         Frequency       Total                                         Weighted
group      Count       Frequency    Events    Expected      Diff        Diff

  1           7           50           5       2.42778     2.57222     2.57222
  2           7           50           3       5.57222    -2.57222    -2.57222
         =========                  ======
             14                        8

                              Covariance Matrix
                               Method = logrank

                    group          1               2

                      1         1.39664        -1.39664
                      2        -1.39664         1.39664

                             Method = logrank

                                   RESULTS
                        CHISQUARE        DF   P_VALUE

                        4.7373318         1 0.0295148

                        Summary of Events vs Expected
                               Method = tarone

                      Percent of
         Frequency       Total                                         Weighted
group      Count       Frequency    Events    Expected      Diff        Diff

  1           7           50           5       2.42778     2.57222     7.13763
  2           7           50           3       5.57222    -2.57222    -7.13763
         =========                  ======
             14                        8


                              Covariance Matrix
                               Method = tarone

                    group          1               2

                      1         13.2056        -13.2056
                      2        -13.2056         13.2056

                             Method = tarone

                                   RESULTS
                        CHISQUARE        DF   P_VALUE

                        3.8579019         1 0.0495122
```

Only the Wilcoxon-Gehan and Log Rank tests are available in SAS. If the proportional hazards assumption seems tenable, then the log rank test is probably the best choice. Otherwise, it would be reasonable to perform power calculations for the type of alternative considered likely using a variety of test statistics and to choose the statistic that provides the greatest power or the desired power most efficiently.

Many approaches to this problem of calculating power for the comparison of survival distributions assume that survival distributions of the two groups being compared are exponential (George and Desu, 1974 and Rubenstein, Gail, and Santner, 1981). Others allow for other distributions, but require that the proportional hazards assumption holds (Shuster,1990 and Cantor, 1992). In these cases the asymptotic efficiency of the log rank test can be shown to lead to a fairly simple result for the power of that test.

The Survpow Macro

Experience with actual clinical trials suggests, however, that such assumptions are not necessarily realistic. For this reason, the method presented here, which is due to Lakatos (1988), requires no assumption concerning the survival distributions of the groups being compared. His formulation allows for rather complex designs and permits specifying the effects of crossovers and non-compliers and losses to follow-up. Shih (1995) discusses a program that implements the Lakatos method, maintaining the functionality described by Lakatos. In this paper, a SAS macro, survpow, that also implements the Lakatos method is presented. While survpow does not allow for some of the complexity of the Lakatos approach as Shih does, it may be easier for the user to invoke. Space does not permit an explanation of Lakatos' method here. The reader is referred to the Lakatos (1988) paper or to Cantor (2003). However, we will discuss its use below.

This macro survpow can be implemented by filling in values for the parameters in the following template:

```
%survpow(s1=     ,s2 =    ,actime=     ,futime=
         ,rate=    ,w=      ,siglevel=       );
```

The macro, survpow, takes the following parameters:

s1 and s2: These are the names of datasets that describe the survival distributions in groups 1 and 2, respectively. They must contain the two variables t (for time) and s (for survival probability at time t). If a dataset contains only one observation, (t, s), the macro assumes that the group has exponential survival with S(t) = s. This equation determines the constant hazard for that group. Otherwise it is assumed that the survival curve is given by the piecewise linear curve formed by connecting the points (t, s) for each observation. The values of t and s must satisfy the following conditions:

a) For the first observation, t = 0 and s = 1.
b) The values of t must be increasing.
c) The values of s must be decreasing.
d) The last value of t must be at least the total time of the study (the accrual time plus the follow-up time).

actime and futime: The accrual time and post-accrual follow-up time respectively.

rate: The accrual rate.

w: This is the formula for the weights that define the linear rank test planned. The variable n, representing the total number at risk, may be used. Thus w = 1 gives the Log Rank test statistic (the default), w = n gives the Wilcoxon-Gehan test statistic, and w = (n**.5) gives the test statistic discussed by Tarone and Ware (1977). The parentheses are needed in the last example to assure that the expression w**2 is evaluated as intended. That's because SAS will evaluate n**.5**2 as n**.25 instead of n.

siglevel: The (two-sided) significance level of the test.

As an example, suppose we are planning a study to compare an experimental treatment to a standard with respect to survival. Experience with the standard suggests about 60% three-year and 25% six-year survival with very little mortality after six years. We expect the experimental treatment to extend early survival somewhat but leave the six year rate unchanged. Specifically, it might provide 80% three year survival.

The following program estimates the power for the Log Rank Test and the Wilcoxon- Gehan Test at the two-sided 0.05 significance level if we accrue 80 patients per year for five years and follow this with three years of additional follow-up.

```
data s1;
input t s;
datalines;
0 1.0
3 .6
6 .25
7 .24
8 .23
;

data s2;
input t s;
datalines;
0 1.0
3 .8
6 .25
7 .24
8 .23
;

%survpow(s1= s1, s2=s2,  actime= 5, futime= 3,
rate= 80, w=1, siglevel=.05) ;

%survpow(s1= s1, s2=s2,  actime= 5, futime= 3,
rate= 80, w=n, siglevel=.05) ;
```

The results are presented as Figure 2.

It's interesting to note the rather dramatic superiority of the Wilcoxon-Gehan Test in this case. With a bit of reflection, we see why this is so. The survival distributions being compared differ most during the first few years when the number at risk is greatest. The Wilcoxon-Gehan Test gives greatest weight to deaths at these times.

Discussion

When planning clinical trials involving the comparison of survival distributions, the Log Rank Test is almost always specified and power calculations are almost always based on the assumption of exponential survival distributions. This is undoubtedly due to the

Figure 2

```
 Accrual    Followup    Accrual
  Time        Time        Rate        N      alpha     Weights      Power

   5           3           80        400      .05         1        0.55060


 Accrual    Followup    Accrual
  Time        Time        Rate        N      alpha     Weights      Power

   5           3           80        400      .05         n        0.91127
```

fact that the Log Rank Test is widely implemented by statistical software such as SAS and formulas and programs to perform power calculations for this test assuming exponential survival distributions are widely available in the literature and on the Internet. But slavish acceptance of this practice is unfortunate. First of all, the Log Rank Test may not be the most powerful test for the type of alternative anticipated. Secondly, the power calculations may be wrong due to the survival distributions' departure from exponentiality. It is hoped that this discussion, and the macros that are available from the author will encourage readers to think a bit more deeply about the nature of the survival distributions being compared and to perform more appropriate tests and power calculations.

References

Tarone, R. and Ware, J. (1977) On Distribution-free Tests for Equality of Survival Distributions, *Biometrika*, 64, 156-160.

Harrington, D. P. and Fleming, T. R.(1982) A Class of Rank Test Procedures for Censored Survival Data, *Biometrika*, 69, 553-566

Peto, R. and Peto, J. (1972) Asymptotically Efficient Rank Invariant Test Procedures (with Discussion), *Journal of the Royal Statistical Society, A,* 135, 185-206.

Lee, E. T., Desu, M. M., and Gehan, E. A. (1975) A Monte Carlo Study of the Power of Some Two-sample Tests, *Biometrika*, 62, 425-432

George, S. L. and Desu, M. M. (1974) Planning the Size and Duration of a Clinical Trial Studying the Time to Some Critical Event, *Journal of Chronic Diseases, 27, 15-24*

Rubenstein, L. V., Gail, M. H., and Santner, T. J. (1981) Planning the Duration of a Comparative Clinical Trial with Loss to Follow-up and a Period of Continued Observation, *Journal of Chronic Diseases,* 34, 469-479

Lakatos, E. (1988) Sample Sizes Based on the Log-Rank Statistic in Complex Clinical Trials, *Biometrics*, 44 229-241

Cantor, A. B. (1992) Sample Size Calculations for the Log Rank Test: A Gompertz Model Approach, *Journal of Clinical Epidemiology,* 45, 1131-1136

Shuster , J. J. (1993), *Handbook of Sample Size Guidelines for Clinical Trials*, Boca Raton: CRC Press, Inc

Shih, J. H. (1995) Sample Size Calculation for Complex Clinical Trials with Survival Endpoints, *Controlled Clinical Trials*, 16:395-407

Cantor, A. B. (2003) Extending SAS Survival Analysis Techniques for Medical Research, 2nd ed. Cary; SAS Institute, Inc.

Author Information

The author can be contacted at
Moffitt Cancer Center
12902 Magnolia Drive
Tampa, FL 33612
abcantor@moffitt.usf.edu